



**Universidad Nacional Autónoma de México**  
**Programa de Posgrado en Ciencias de la**  
**Administración**

**Ciencia de datos para la optimización del comportamiento del  
usuario en un sitio en la web**

**T e s i s**

Que para optar por el grado de:

**Maestro en Administración**  
**Campo de conocimiento: Administración de**  
**la Tecnología**

Presenta:

**Raúl Alejandro Ojeda Ramírez**

Tutor:

**Dra. Nadima Simón Domínguez**  
**Facultad de Contaduría y Administración**

**Ciudad Universitaria, CD.MX. Enero de 2024**



Universidad Nacional  
Autónoma de México

Dirección General de Bibliotecas de la UNAM

**Biblioteca Central**



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

# Dedicatoria

Esta tesis de maestría está dedicada a Dios por ser una guía en mi vida y por bendecirme y darme la fuerza para nunca rendirme y seguir persiguiendo mis metas, permitiéndome llegar a este momento tan importante en mi formación profesional.

Gracias a mi mamá por ser mi principal apoyo y que siempre me brindó su amor y apoyo incondicional a pesar de nuestras diferencias. Gracias a mi padre que me enseñó que el mejor conocimiento que puede tener una persona es el que aprende por sí mismo, gracias por el ejemplo de inculcarme el estudio, el gran hábito a la lectura, el trabajo y a no tener miedo a las adversidades porque Dios siempre está conmigo.

Un agradecimiento de manera especial a la Dra. Nadima Simón Domínguez, quien me acompañó no sólo en la elaboración de este proyecto de investigación, sino a lo largo de mi desarrollo universitario y por haberme apoyado en mi crecimiento profesional y en el cultivo continuo de mis valores.

A la Universidad Nacional Autónoma de México, de manera especial a la Facultad de Contaduría y Administración, muchas gracias por darme la oportunidad de crecer tanto académicamente como docente en ésta, la Máxima Casa de Estudios. Considero un privilegio haber podido obtener mi título en esta prestigiosa institución.

## Índice

|  |           |
|--|-----------|
| <b>INTRODUCCIÓN.....</b>   | <b>1</b>  |
| <b>A) JUSTIFICACIÓN DE LA INVESTIGACIÓN.....</b>   | <b>2</b>  |
| <b>B) PLANTEAMIENTO DEL PROBLEMA.....</b>  | <b>4</b>  |
| <b>C) PREGUNTA GENERAL DE INVESTIGACIÓN.....</b>   | <b>4</b>  |
| <b>D) PREGUNTAS ESPECÍFICAS DE INVESTIGACIÓN .....</b>   | <b>4</b>  |
| <b>E) OBJETIVO GENERAL.....</b>  | <b>6</b>  |
| <b>F) OBJETIVOS ESPECÍFICOS.....</b>   | <b>6</b>  |
| <b>G) HIPÓTESIS GENERAL.....</b>   | <b>7</b>  |
| <b>H) HIPÓTESIS ESPECÍFICAS .....</b>  | <b>7</b>  |
| <b>G) METODOLOGÍA .....</b>  | <b>8</b>  |
| <b>CAPÍTULO 1. FUNDAMENTOS TEÓRICOS.....</b>   | <b>10</b> |
| 1.1 ANTECEDENTES:.....   | 10        |
| 1.2 ESCALABILIDAD EN LA WEB .....  | 11        |
| 1.2.1 <i>Utilización del caché.....</i>  | 12        |
| 1.2.2 <i>Protocolo de Transferencia de Hipertexto.....</i>   | 14        |
| 1.2.3 <i>Compresión de datos .....</i>   | 14        |
| 1.3 INTERNET .....   | 15        |
| 1.3.1 <i>Progreso del Internet.....</i>  | 15        |
| 1.3.1.1 <i>La web 1.0 .....</i>  | 16        |
| 1.3.1.2 <i>La web 2.0 .....</i>  | 16        |
| 1.3.1.3 <i>La web 3.0 .....</i>  | 17        |
| 1.3.1.4 <i>La web 4.0 .....</i>  | 18        |
| 1.3.2 <i>Localizador Universal de Recursos.....</i>  | 19        |
| 1.3.3 <i>Navegador web .....</i>   | 21        |
| 1.3.4 <i>Internet .....</i>  | 22        |
| 1.3.5 <i>World Wide Web .....</i>  | 23        |
| 1.3.6 <i>Servidor web .....</i>  | 25        |
| 1.3.7 <i>Sistema Intranet .....</i>  | 27        |
| 1.4 MINERÍA WEB.....   | 29        |
| 1.4.1 <i>Proceso de minería web para obtener conocimiento útil y significativo .....</i>             | 30        |
| 1.4.1.1 <i>Minería de utilización: .....</i>   | 32        |
| 1.4.1.2 <i>Minería de estructura: .....</i>  | 32        |
| 1.4.1.3 <i>Minería de contenido: .....</i>   | 32        |
| 1.4.2 <i>Técnicas de minería web .....</i>   | 33        |
| 1.4.3 <i>Desafíos y limitación de la minería web.....</i>  | 35        |
| <b>CAPÍTULO 2. COMPORTAMIENTO DEL USUARIO EN LA BÚSQUEDA DE CONTENIDOS<br/>EN EL SITIO WEB .....</b> | <b>37</b> |
| 2.1 EL USUARIO Y LA WEB .....  | 37        |
| 2.2 FACTORES DEL COMPORTAMIENTO DEL USUARIO .....  | 39        |
| 2.3 ACCESIBILIDAD .....  | 40        |
| 2.4 FUNCIONALIDAD.....   | 41        |
| 2.5 ENCONTRABILIDAD.....   | 42        |
| 2.6 UTILIDAD .....   | 43        |
| 2.7 ESTÉTICA .....   | 44        |
| 2.8 CREDIBILIDAD .....   | 45        |
| 2.9 USABILIDAD COMO FACTOR ESENCIAL .....  | 46        |
| <b>CAPÍTULO 3. CIENCIA DE DATOS .....</b>  | <b>48</b> |
| 3.1 LA CIENCIA DE DATOS EN LA ERA DE LOS DATOS MASIVOS.....  | 48        |
| 3.2 EVOLUCIÓN DEL ANÁLISIS DE DATOS .....  | 50        |
| 3.3 ANÁLISIS CON CIENCIA DE DATOS O PREDICCIONES .....   | 53        |
| 3.3.1 <i>Analítica casual predictiva.....</i>  | 53        |

|  |            |
|--|------------|
| 3.3.2 <i>Análisis prescriptivo</i> .....   | 53         |
| 3.4 APRENDIZAJE AUTOMÁTICO .....   | 54         |
| 3.4.1 <i>Hacer predicciones</i> .....  | 54         |
| 3.4.2 <i>Descubrimiento de Patrones</i> .....                                    | 54         |
| 3.5 USO DE LA CIENCIA DE DATOS.....  | 55         |
| 3.5.1 <i>Servicio al cliente</i> .....   | 56         |
| 3.5.2 <i>Coches sin conductor</i> .....  | 56         |
| 3.5.3 <i>Predicciones</i> .....  | 56         |
| 3.6 DATOS .....  | 56         |
| 3.6.1 <i>Datos estructurados frente a no estructurados</i> .....                 | 56         |
| 3.6.2 <i>Datos cuantitativos frente a categóricos</i> .....                      | 57         |
| 3.6.3 <i>Clasificación y regresión</i> .....                                     | 59         |
| 3.7 PROCESO CRISP-DM .....   | 60         |
| 3.8 ALGORITMOS GENÉTICOS.....  | 65         |
| 3.8.1 <i>Introducción</i> .....  | 65         |
| 3.8.2 <i>Evolución Biológica</i> .....   | 68         |
| 3.8.3 <i>Métodos de representación en algoritmos genéticos</i> .....             | 71         |
| 3.8.4 <i>Creación de la población inicial</i> .....                              | 72         |
| 3.8.5 <i>Operadores Genéticos</i> .....  | 73         |
| A) <i>La selección</i> .....   | 74         |
| B) <i>Cruzamiento</i> .....  | 76         |
| C) <i>La mutación</i> .....  | 77         |
| 3.8.6 <i>Parámetros de control</i> .....   | 80         |
| 3.8.7 <i>Función de evaluación de aptitud</i> .....                              | 81         |
| 3.8.8 <i>Aplicaciones</i> .....  | 82         |
| 3.8.9 <i>Comparación con respecto a los métodos convencionales</i> .....         | 83         |
| 3.8.10 <i>Factores relevantes para algoritmos genéticos</i> .....                | 84         |
| <b>CAPÍTULO 4. ELABORACIÓN DE UN MODELO DE ALGORITMO .....</b>                   | <b>86</b>  |
| 4.1 OPTIMIZACIÓN WEB Y USABILIDAD .....  | 86         |
| 4.2 PARÁMETROS DEL ALGORITMO .....   | 88         |
| 4.3 REPRESENTACIÓN DEL INDIVIDUO .....   | 89         |
| 4.4 VALOR ÓPTIMO.....  | 93         |
| 4.5 MODELO: CAMINOS MÍNIMOS.....   | 95         |
| Consideraciones Finales.....   | 99         |
| <b>CAPÍTULO 5. VALIDACIÓN DEL MODELO .....</b>                                   | <b>101</b> |
| 5.1 MARKETING DIGITAL: UN ESTUDIO DE ANÁLISIS DE COMPORTAMIENTO DE USUARIOS..... | 101        |
| 5.2 ANÁLISIS DEL COMPORTAMIENTO DEL USUARIO.....                                 | 102        |
| 5.2.1 <i>Comprensión del análisis del comportamiento del usuario</i> .....       | 106        |
| 5.2.2 <i>El propósito del análisis del comportamiento del usuario</i> .....      | 106        |
| 5.2.3 <i>Indicadores clave para analizar el comportamiento del usuario</i> ..... | 107        |
| 5.2.4 <i>Implementar análisis de comportamiento del usuario</i> .....            | 108        |
| 5.2.4.1 <i>Análisis de eventos de comportamiento</i> .....                       | 108        |
| 5.2.4.2 <i>Análisis de retención de usuarios</i> .....                           | 109        |
| 5.2.4.3 <i>Análisis del modelo de embudo</i> .....                               | 109        |
| 5.2.4.4 <i>Análisis de la ruta de comportamiento</i> .....                       | 110        |
| 5.2.4.5 <i>Análisis del modelo de Fogg</i> .....                                 | 110        |
| 5.2.5 <i>MODELO AISAS</i> .....  | 111        |
| 5.2.5.1 <i>Attention</i> .....   | 111        |
| 5.2.5.2 <i>Interest</i> .....  | 111        |
| 5.2.5.3 <i>Search</i> .....  | 111        |
| 5.2.5.4 <i>Action</i> .....  | 111        |
| 5.2.5.5 <i>Share</i> .....   | 112        |
| 5.3 REPRESENTACIÓN DE LOS DATOS DE ENTRADA.....                                  | 112        |
| 5.4 MODELO DEL ALGORITMO.....  | 116        |
| 5.4.1 <i>Población inicial o primera generación</i> .....                        | 116        |
| 5.4.2 <i>Conservación de las soluciones óptimas o de mayor calidad</i> .....     | 116        |
| 5.4.3 <i>La elección de progenitores o padres</i> .....                          | 117        |
| 5.4.4 <i>Técnica de cruce destructiva con mutación aleatoria</i> .....           | 117        |
| 5.4.5 <i>Ajustes del algoritmo</i> .....   | 118        |

## CAPÍTULO 6. APLICACIÓN DEL MODELO DE ALGORITMO EN UN ESTUDIO DE CASO.

120

|   |            |
|---|------------|
| 6.1 ALGORITMO .....   | 122        |
| 6.1.1 Población.....  | 122        |
| 6.1.2 Aptitud .....   | 123        |
| 6.1.3 Selección de individuos.....                                  | 123        |
| 6.1.4 Cruce de dos individuos (recombinación).....                  | 125        |
| 6.5 Mutar individuo .....   | 127        |
| 6.3 IMPLEMENTACIÓN DEL ALGORITMO .....                              | 128        |
| 6.3.1 Evaluar la aptitud del comportamiento del usuario .....       | 128        |
| 6.3.2 Función de selección por ruleta .....                         | 130        |
| 6.3.3 Función de cruce y mutación .....                             | 132        |
| 6.3.4 Reemplazar la población anterior con la nueva población ..... | 133        |
| 6.3.5 Iteración del algoritmo .....                                 | 134        |
| 6.4 IMPLEMENTACIÓN EN CIENCIA DE DATOS.....                         | 137        |
| 6.5 RESULTADOS OBTENIDOS .....                                      | 147        |
| 6.6 ANÁLISIS DE RESULTADOS .....                                    | 154        |
| CONCLUSIONES.....   | 157        |
| LIMITACIONES DEL TRABAJO .....                                      | 160        |
| TRABAJOS FUTUROS .....  | 161        |
| <b>BIBLIOGRAFÍA.....</b>  | <b>162</b> |
| <b>ANEXOS .....</b>   | <b>166</b> |

### Índice de figuras

|   |     |
|---|-----|
| Figura 1 Metodología CRISP-DM.....  | 9   |
| Figura 2 La Arpanet original .....  | 16  |
| Figura 3 Comunicaciones en el servidor / cliente utilizando protocolo de .....  | 20  |
| Figura 4 Protocolo de Transferencia de Hipertexto .....   | 21  |
| Figura 5 World Wide Web primer navegador de la historia .....   | 22  |
| Figura 6 Internet Global .....  | 23  |
| Figura 7 www “Telaraña alrededor del mundo” .....   | 24  |
| Figura 8 Servidor web.....  | 27  |
| Figura 9 Estructura de intranet .....   | 28  |
| Figura 10 Representación jerárquica de las áreas de la minería web.....   | 32  |
| Figura 11 Un ciclo de vida genérico, variable e iterativo de desarrollo de sitios web<br>ilustra los puntos en los que la ingeniería de usabilidad es más beneficiosa.....                          | 38  |
| Figura 12 Factores de diseño enfocado a la satisfacción-no frustración de uso.....  | 40  |
| Figura 13 Disciplinas de Ciencia de Datos .....   | 49  |
| Figura 14 Evolución del análisis de datos .....   | 51  |
| Figura 15 Crecimiento de la ciencia de datos.....   | 52  |
| Figura 16 Pirámide de la ciencia de datos.....  | 60  |
| Figura 17 Las fases y tareas del proceso CRISP-DM .....   | 65  |
| Figura 18 Charles Darwin .....  | 70  |
| Figura 19 Diagrama de flujo de para un algoritmo genético sencillo .....  | 74  |
| Figura 20 Ilustración de cruce de un punto que divide el genoma de dos soluciones en<br>un punto arbitrario (aquí en el medio) y las vuelve a ensamblar para obtener dos<br>soluciones nuevas ..... | 77  |
| Figura 21 La distribución gaussiana es la base del operador de mutación gaussiana<br>que añade ruido a cada componente del cromosoma.....   | 79  |
| Figura 22 Ejemplo hipervínculos entre páginas.....  | 99  |
| Figura 23 Ejemplo del análisis del comportamiento del usuario.....  | 107 |
| Figura 24 Modelo AISAS .....  | 112 |

|   |     |
|---|-----|
| Figura 25 Ruleta selección ejemplo.....   | 124 |
| Figura 26 Ejemplo de selección por torneo con un tamaño de torneo de tres ..... | 125 |
| Figura 27 Punto único Transversal ejemplo .....                                 | 126 |
| Figura 28 Dos puntos Transversal ejemplo .....                                  | 126 |
| Figura 29 Uniforme Transversal ejemplo.....                                     | 127 |

### Índice de gráficas

|   |     |
|---|-----|
| Gráfica 1 Edad.....   | 114 |
| Gráfica 2 Edad y tiempo que pasan en el sitio .....   | 115 |
| Gráfica 3 Profundidad de las páginas principales sin enlaces duplicados de la<br>www.fca.unam.mx .....      | 151 |
| Gráfica 4 Profundidad de las páginas principales sin enlaces duplicados de la<br>www.fca.unam.mx en 3D..... | 152 |
| Gráfica 5 Nivel de profundidad en 3D con las URLs de la pagina de www.fca.unam.mx<br>.....                  | 171 |
| Gráfica 6 Cantidad de URLs totales de los sitios principales de la página<br>www.fca.unam.mx .....          | 174 |

### Índice de tablas

|  |     |
|--|-----|
| Tabla 1 Criterios para la clasificación higiénico-motivadora de los factores de diseño   | 39  |
| Tabla 2 Datos entrantes del algoritmo.....   | 87  |
| Tabla 3 Ejemplificación de la presentación de los elementos genéticos con m igual a<br>tres páginas .....                            | 91  |
| Tabla 4 Generalización de elementos .....  | 91  |
| Tabla 5 Características de Acceso y Navegación en el Sitio Web .....   | 93  |
| Tabla 6 Actividades Principales y nivel de profundidad de la FCA 2019-2020 .....   | 103 |
| Tabla 7 Actividades Principales y nivel de profundidad de la FCA 2020-2021 .....   | 105 |
| Tabla 8 Datos de entrada del algoritmo.....  | 113 |
| Tabla 9 Rango de datos del algoritmo.....  | 113 |
| Tabla 10 Casos de examen de comprensión de lectura de textos en ingles 2019-2021<br>.....  | 145 |
| Tabla 11 Resultados de URLs principales y su nivel de profundidad de la página de la<br>www.fca.unam.mx sin enlaces duplicados ..... | 149 |
| Tabla 12 Resultados de URLs principales y su nivel de profundidad de la página de la<br>www.fca.unam.mx .....                        | 168 |

### Índice de ecuaciones

|   |    |
|---|----|
| Ecuación 1 Enlace entre páginas y posición de los elementos lineales..... | 92 |
| Ecuación 2 Función cajón inferior.....                                    | 92 |
| Ecuación 3 Función residuo .....  | 92 |
| Ecuación 4 Factor tiempo .....  | 93 |
| Ecuación 5 Periodicidad de cada hipervínculo .....                        | 95 |
| Ecuación 6 Peso del hipervínculo existente xy.....                        | 96 |
| Ecuación 7 Camino.....  | 97 |
| Ecuación 8 Probabilidad mediante caminos posibles.....                    | 97 |
| Ecuación 9 Peso potencial camino mínimo .....                             | 98 |

## Índice de códigos

|  |     |
|--|-----|
| Código 1 Captura de código de la función evaluar la aptitud del comportamiento del usuario en el editor de Sublime Text .....                  | 129 |
| Código 2 Captura de código de la función de selección por ruleta en el editor de Sublime Text.....   | 131 |
| Código 3 Captura de código de la función de cruce y mutación en el editor de Sublime Text.....   | 132 |
| Código 4 Captura de código de la función de nueva población en el editor de Sublime Text.....  | 134 |
| Código 5 Captura de código de la iteración del algoritmo genético en el editor de Sublime Text.....  | 135 |
| Código 6 Programa en Python que nos permitió ver los enlaces ligados a la página de <a href="http://www.fca.unam.mx">www.fca.unam.mx</a> ..... | 166 |



---

## *Introducción*

---

En la actual competencia del mundo actual, una prioridad es conseguir la mejora del comportamiento del usuario mediante la evaluación de un sitio web, proporcionándole experiencias que le permitan neutralizar amenazas y explotar oportunidades. Por lo tanto, un objetivo ambicioso de la investigación de mercados es lograr y/o medir el comportamiento de los usuarios en los sitios web, por ejemplo: poder predecir qué les gusta a los usuarios, qué cualidades valoran más y qué les parece más interesante, qué contenido les llama más el interés, necesidades en el momento de la compra, facilitar objetivos de búsqueda, entre otros. Con la llegada del auge de la tecnología y la gestión de datos inteligente en la web, existe la necesidad de seguir el camino de cómo adaptar los sitios web a su público objetivo.

Surge así la inquietud de si es posible **construir un modelo de comportamiento del usuario en la web**, a través del registro web (capturando así la ruta de navegación del usuario), y poder determinar una mejor estructura web. El propósito de este modelo es controlar los datos generados en la web para atraer más usuarios y aumentar gradualmente su lealtad a una empresa o institución representativa.

---

## *a) Justificación de la investigación.*

---

Este trabajo de investigación se centra en la evaluación de un sitio web a través del análisis de patrones de comportamiento de los usuarios utilizando un modelo de algoritmos genéticos. Este modelo debe ser ajustado con cambios y detalles repetidos en el algoritmo para poder alcanzar dicho objetivo. **El objetivo es identificar la mejor estructura para un sitio web que permita a los usuarios navegar de manera cómoda y eficiente, reduciendo el tiempo necesario para acceder a una página específica de su interés<sup>1</sup>.** En resumen, se requiere establecer una disposición de hipervínculos entre páginas que disminuya la cantidad total de páginas y el costo de navegar a través de muchas de ellas para acceder a una página en particular, especialmente cuando no hay hipervínculos directos a ella. Es importante señalar que sólo se considera la estructura de los hipervínculos, lo cual no se compara con otros algoritmos que puedan resolver los modelos planteados, y que sólo se lleva a cabo una prueba de validación del modelo en un solo punto específico.

Fue propuesta la idea de cuantificar el comportamiento de un gran número de usuarios potenciales a través de parámetros de investigación, registros web y algoritmos genéticos para determinar la forma más óptima de construir un sitio donde recabar información utilizando la visualización universal. El objetivo es obtener información precisa para la toma de decisiones informadas en el proceso de construcción del sitio. (Goldberg, 1989).

El nombre "algoritmos genéticos" se debe a su inspiración en la selección natural y la genética molecular propuestas por Darwin, las cuales son la base de estos métodos (Darwin, 1998), cuyo objetivo es buscar un espacio de soluciones candidatas y elegir las más óptimas, tomando como referencia el comportamiento de ciertos organismos celulares en el proceso evolutivo. El proceso consiste en generar varias soluciones potenciales aleatorias (como el caso de una estructura de web óptima), luego crear nuevas soluciones (llamadas "hijas") combinando soluciones previas. Se les da mayor importancia y significado a las soluciones que han demostrado ser las mejores hasta el

---

<sup>1</sup> Una manera simple de interpretar esto es realizando menos clics, es decir que el usuario encuentre rápido lo que está buscando.

momento para seguir evolucionando en busca de la mejor solución posible dentro del conjunto de opciones posibles (Schmelkes & Schmelkes, 2010).

Este tema es particularmente útil para las empresas que tienen un sitio web institucional y desean fusionar transacciones de esta manera (**especialmente para mejorar la experiencia de los usuarios y minimizar el tiempo de búsqueda**), lo que será muy apreciado a la hora de determinar la mejor opción de estructura de red. Con este fin, comprender el uso de Algoritmos Genéticos, su utilidad, su proceso y alcance permitirá que estas aplicaciones se asocien con los datos y la web. En este caso especial, trabajaremos en una plataforma experimental para lograr la mejor estructura.

---

### *b) Planteamiento del problema.*

---

Si bien existen algunas herramientas para la búsqueda de contenidos en el sitio web, no se conoce cuál de ellas optimiza el comportamiento del usuario. Con el uso de la tecnología y en la implementación de las plataformas web, podemos acceder a cualquier sitio y desde cualquier parte, sin embargo, el usuario tiene que estar recorriendo varios lugares para poder llegar al lugar adecuado, esto muchas veces es complicado para el usuario. En este sentido, se identifica que uno de los principales ejes de una página web es la búsqueda de información, lo cual implica la necesidad de proporcionar a los usuarios una experiencia eficiente y satisfactoria al buscar y acceder a la información relevante dentro del sitio web. La estructura de hipervínculos y la organización de las páginas desempeñan un papel fundamental en la facilitación de esta tarea, y es aquí donde surge la problemática de encontrar la mejor forma de estructurar y diseñar un sitio web que optimice el comportamiento del usuario en la búsqueda de información (Hernández Sampieri, Zapata Salazar, & Mendoza Torres, 2013).

---

### *c) Pregunta general de Investigación.*

---

¿De qué manera se logra optimizar el comportamiento de los usuarios en la búsqueda del contenido deseado en el sitio web?

---

### *d) Preguntas específicas de Investigación*

---

1. ¿Cómo a través de los web logs se puede obtener/optimizar la información de los usuarios?
2. ¿Cómo se deberá retroalimentar la experiencia en la web?

Para poder llegar a contestar estas preguntas vamos a utilizar diversas herramientas de algoritmos genéticos y herramientas de minería de datos Y sobre todo la visión general que conlleva la administración de la tecnología de la información.

---

*e) Objetivo general.*

---

Elaborar un modelo de ciencia de datos para optimizar el comportamiento de los usuarios en la búsqueda del contenido deseado en el sitio web.

---

*f) Objetivos específicos*

---

- 1.- Construir un modelo de ciencia de datos usando web logs para representar el comportamiento de los usuarios en la web.
- 2.- Descubrir la mejor estructura para un sitio web
- 3.- Formar un mecanismo para extender el modelo a distintos sitios web.
- 4.- Aplicar técnicas de **minería web** y **algoritmos genéticos** a las decisiones de *marketing*, se propone una nueva metodología para estudiar el comportamiento de los usuarios en la web.

---

### *g) Hipótesis general*

---

Mediante un modelo de ciencia de datos se logra una búsqueda que optimiza el comportamiento de los usuarios en el sitio web, el cual permite cuantificar la importancia de cada hipervínculo (existente y no existente) y evaluar su mantenimiento o crearlo por separado.

---

### *h) Hipótesis específicas*

---

1. Mediante las bitácoras que generan los sitios webs es posible encontrar caminos para apoyar al usuario a obtener/optimizar la información de los usuarios, es decir, a encontrar más rápido lo que ellos necesitan.

2. La retroalimentación de la experiencia en la web se realiza con una herramienta en el lenguaje de programación Python, con la cual se obtendrán todas las ligas que tiene un sitio web para ver su incidencia en el usuario, así como mediante la creación del mapa del sitio a través de la experiencia del usuario y las actividades preestablecidas (anuales).

---

## g) Metodología

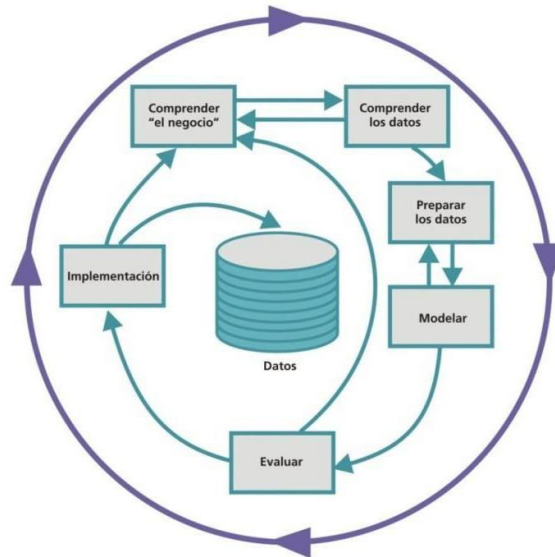
---

Se elaboró un modelo tomando como estudio de caso la FCA, en particular. Las etapas seguidas fueron:

1. Verificación de los estudios especializados en el contenido.
2. Identificación de los aspectos importantes del comportamiento de los usuarios en la web, incluyendo sus posibles manifestaciones e indagaciones relacionadas.
3. Utilización del método CRISP-DM (que contiene las siguientes etapas: 1. Comprensión del negocio, 2. Comprensión de los datos, 3. Preparación de los datos, 4. Modelado, 5. Evaluación, 6. Implementación) para la elaboración del modelo de algoritmos genéticos. CRISP-DM es un método estándar y un procedimiento típico de minería de datos que abarca desde la extracción de datos hasta la gestión y considera las cuestiones comerciales (Schmelkes & Schmelkes, 2010).
4. Elaboración de bocetos conceptuales necesarios para construir el modelo propuesto, que permitieron conocer los instrumentos necesarios para su diseño, validación y soporte.
5. Evaluación del número suficiente de hipervínculos en la ubicación correcta y la dirección de destino para lograr una estructura óptima del sitio web. Esto mejora la interactividad entre las páginas y se basa en el análisis de las búsquedas y comportamientos de los usuarios.
6. Utilización del preprocesamiento de datos y construcción, evaluación e implementación de modelos, siguiendo el enfoque propuesto por (Markov & Larose, 2007). Aunque se centra principalmente en temas de negocios, puede aplicarse de manera equivalente y ampliarse a temas de investigación, considerando las fases y relaciones mostradas en la figura 1.



Figura 1 Metodología CRISP-DM



Fuente: [https://www.researchgate.net/figure/Fases-del-proceso-de-CRISP-DM-Adaptado-de-10\\_fig2\\_306959832](https://www.researchgate.net/figure/Fases-del-proceso-de-CRISP-DM-Adaptado-de-10_fig2_306959832)

---

# Capítulo 1. Fundamentos Teóricos

---

## Introducción

En este capítulo, se establecen los fundamentos teóricos que sentarán las bases para comprender el resto de la tesis. Se abordan temas como la escalabilidad en la web, incluyendo el uso del caché, el Protocolo de Transferencia de Hipertexto y la compresión de datos. Además, se explora la evolución de Internet, desde la web 1.0 hasta la web 4.0, y se presentan conceptos clave como el localizador universal de recursos, el navegador web, el servidor web y la *World Wide Web*. También se introduce el campo de la minería web, incluyendo el proceso de obtener conocimiento útil y significativo de la web, así como las técnicas utilizadas en este ámbito. Este capítulo sienta las bases conceptuales necesarias para comprender la optimización del comportamiento del usuario en un sitio web

### 1.1 Antecedentes:

Hoy en día, las empresas deben ser conscientes de que generar valor en la experiencia de un sitio web es fundamental para mantener a los usuarios interesados y comprometidos con su contenido. Esto significa que el sitio web debe ser capaz de proporcionar una búsqueda rápida y agradable, ofreciendo contenido relevante y de interés para cada usuario, así como ser fácil de navegar y de usar. Además, las empresas buscan constantemente conocer mejor a sus clientes y entender su comportamiento en la web, utilizando herramientas de marketing y tecnología para identificar sus tendencias y preferencias en línea. De esta manera, pueden adaptar sus estrategias y mejorar la experiencia de compra en su sitio web, logrando un mayor número de ventas y una mayor satisfacción del cliente. La clave está en conocer a fondo a los consumidores y sus necesidades, para así poder ofrecerles productos y servicios que realmente les interesen y cubran sus expectativas. En resumen, la experiencia de usuario y el conocimiento profundo del comportamiento de los consumidores son

elementos clave para el éxito de cualquier sitio web, y las empresas deben asegurarse de ofrecer un sitio web que se adapte a las necesidades de cada usuario y que les permita conocerlo (Kim, Kim, & LEE, 2002).

Mediante el uso de la tecnología de minería web, es posible suponer que las variables del usuario requieren un considerable tiempo para ser estudiadas. Por esta razón, se pueden examinar las sesiones de navegación de los usuarios (registros de actividad en el sitio web) con el objetivo de identificar las rutas y páginas más interesantes y frecuentadas. De esta manera, se puede obtener información valiosa acerca del comportamiento de los usuarios en el sitio web. Google es un ejemplo destacado de un método de gestión de alto nivel, ya que nos permite no sólo dar nuestros primeros pasos en la web, sino también sobresalir en ella. Esto se debe a que el motor de búsqueda de Google se ha convertido en una herramienta esencial para encontrar información en línea y ha establecido altos estándares en cuanto a la calidad y relevancia de los resultados de búsqueda. De esta manera, Google ha contribuido significativamente a la mejora de la experiencia de usuario en Internet. (Harford, 2008).

Durante la revisión de la literatura, se halló una investigación relevante que tuvo por objetivo elaborar un algoritmo genético para hacer más eficiente el modelo de programación lineal para optimizar la asignación de personal a diversos horarios de trabajo en un centro de atención telefónica que tiene más de 3000 operadoras y alrededor de 5000 horarios de trabajo diferentes. así como una metodología eficaz para validar el desempeño del modelo de representación del problema de asignación de personal a diferentes horarios de trabajo, junto con sus distintas variantes. (Ojeda Villagómez, 2008)

## 1.2 Escalabilidad en la web

“La escalabilidad se refiere a la facilidad que tiene la red en poder crecer, contrapuesto a un modelo cliente servidor clásico, como puede tener Facebook o Google. Un gigante informático como Google puede soportar la carga de trabajo que le genera cada aplicación que se conecta a sus servicios con centros de cómputos monstruosamente grandes, la famosa computación en la nube. Pero de no contar con esa capacidad informática, se hace muy difícil poder soportar los picos de trabajo para un sistema centralizado. Al contrario, un

sistema P2P se hace más fuerte con más usuarios conectados, ya que cada uno de los usuarios actúa ayudando con una proporción de la carga de trabajo total de toda la red” (Ocariz B., 2019, p. 43)

A continuación, se examinarán los métodos y técnicas que facilitan el escalado en la web. En primer lugar, identificamos técnicas para mejorar la escalabilidad y soluciones que mejoran el ancho de banda y el rendimiento de la red para un servidor determinado. Una de estas técnicas es: usar la utilización del caché; los navegadores web se construyen utilizando el caché, llamado el caché del servidor *proxy*, lo que da como resultado mejores clientes, mejores protocolos, mejor compresión en la web, de tal forma que esto permite que la red sea escalable. (T. Kwan, E. McGrath, & A. Reed, 1995).

### 1.2.1 Utilización del caché

Caché: “Espacio del disco duro que el navegador reserva para registrar las páginas visitadas” (Virga & Menning, 2000, p. 42). Es más rápido acceder a esta copia almacenada que al original. Por ejemplo, cuando se ejecuta un proceso, las instrucciones se almacenan en varios niveles del caché, incluyendo el caché del disco duro y el caché de la CPU. A diferencia de un búfer, que puede almacenar la única copia de un elemento de datos, el caché sólo almacena una copia de un elemento que ya está almacenado en otro lugar, pero en un dispositivo de almacenamiento más rápido (Silberschatz, Baer Galvin, & Gagne, 2006).

Cuando hablamos de utilizar el caché, nos referimos a la técnica de almacenar datos en un lugar más rápido y acceder a ellos desde ahí para ser más eficientes y rápidos. De esta manera, evitamos tener que acceder constantemente a la fuente original de los datos. En resumen, se trata de aprovechar la capacidad de la memoria caché para mejorar el rendimiento y la velocidad en la manipulación de datos.

Es posible utilizar el caché en los siguientes elementos:

1. Servidor
2. Red
3. Usuario.

Se puede lograr que el caché del servidor refleje el interés global del contenido del servidor mediante la replicación del sistema de archivos y del servidor HTTP y conectando el servidor del reproductor a una red de alta velocidad (T. Kwan, E. McGrath, & A. Reed, 1995) . De tal forma que se parece a un servidor espejo que replica datos de un servidor a otros para disminuir la cantidad de datos en la red en el servidor de origen. No obstante, el servidor que es replicado en otros servidores no se encuentra en la misma ubicación y puede estar distante o remoto. El objetivo principal del caché del servidor es reducir la carga del servidor y optimizar el tiempo de respuesta, lo que mejora su rendimiento en general.

La memoria caché del lado del cliente se adapta a los intereses de los usuarios y, por lo tanto, el contenido almacenado en ella puede variar en función de la forma en que los usuarios acceden a la información y de la capacidad de la memoria caché. La memoria caché de red funciona en función de las especificaciones de acceso de un grupo de usuarios que comparten dicha memoria caché. Para mejorar la eficiencia de la red del caché, se puede colocar en ubicaciones con un alto grado de afinidad, interés o relevancia para dicho grupo de usuarios, lo que lleva a implementar una memoria caché jerárquica en un servidor proxy (proxy: “se trata de un programa que trabaja con servicios externos en nombre de clientes internos: éstos se comunican con los servidores que a su vez transmiten las solicitudes aprobadas a cada uno de ellos para después enviar las respuestas al servidor y de éste a los clientes” (Pérez Terán , 2018, p. 51)) o una implementación múltiple de los elementos. En un servidor proxy, existen varias memorias caché que son compartidas por muchos usuarios, y además hay cachés adicionales que pueden consultar otras cachés. Los cachés de segundo nivel, que se conectan a múltiples redes y a cachés de servidores web separados o del caché de primer nivel, tienen una capacidad mayor. Si un documento no está almacenado en la primera capa del caché, se buscará en la segunda capa del caché en busca del documento faltante. Es decir, se buscará en un lugar donde es posible que el documento ya esté almacenado (Wessels, 2001).

La memoria caché es una herramienta que crea un almacenamiento temporal de datos para los usuarios. Esto ayuda a solucionar el problema de la escalabilidad, aunque no es una solución perfecta. Se puede combinar con otras técnicas, como la compresión de datos y la replicación de servidores, para mejorar la red

y preparar la arquitectura web para el futuro. De esta manera, se pueden manejar grandes cantidades de datos y hacer que el acceso a la información sea más rápido y eficiente.

### 1.2.2 Protocolo de Transferencia de Hipertexto

El Protocolo de Transferencia de Hipertexto (HTTP) es el medio a través del cual se envían las solicitudes para acceder a páginas web en Internet y respuestas de ese sitio web que brindan información para ser visualizada en la pantalla de una computadora (Pollard, 2019).

El protocolo HTTP es un modelo que se utiliza para distribuir y asignar las direcciones de los usuarios o clientes a través de las computadoras o redes informáticas. Esto ayuda a manejar la ubicación de los usuarios o clientes. Esta técnica se aplica cada vez que un usuario solicita un documento que se extrae del servidor original. De esta manera, se puede garantizar que el usuario obtenga la información solicitada de manera eficiente y segura (Pollard, 2019).

Además, el caché del servidor *proxy* se utiliza para los documentos entrantes; sin embargo, todavía es necesario informar los cambios en los documentos reflejados o almacenados en caché.

### 1.2.3 Compresión de datos

La compresión es una técnica que se utiliza para disminuir el tamaño de un archivo, de manera que ocupe menos espacio de almacenamiento y sea más eficiente al transmitir datos. Es decir, se reduce el tamaño del archivo para que sea más fácil de manejar y almacenar. Esta técnica se utiliza con frecuencia en la transmisión de datos, ya que permite enviar archivos más grandes en menos tiempo y con menos recursos. La reducción se logra mediante la eliminación de datos redundantes y la optimización de la representación de caracteres mediante la asignación de un número variable de bits según su frecuencia de transmisión. La compresión de archivos elimina datos redundantes y no críticos para permitir un almacenamiento y transferencia más eficiente. Pero la efectividad de la compresión varía según el tipo de archivo, ya que la compresión funciona mejor para documentos de texto que para archivos binarios como JPEG o GIF. Para

determinar cuánto ancho de banda se puede ahorrar mediante la compresión, se puede utilizar el porcentaje de archivos comprimibles y la entropía promedio de cada tipo de archivo. (Pollard, 2019)

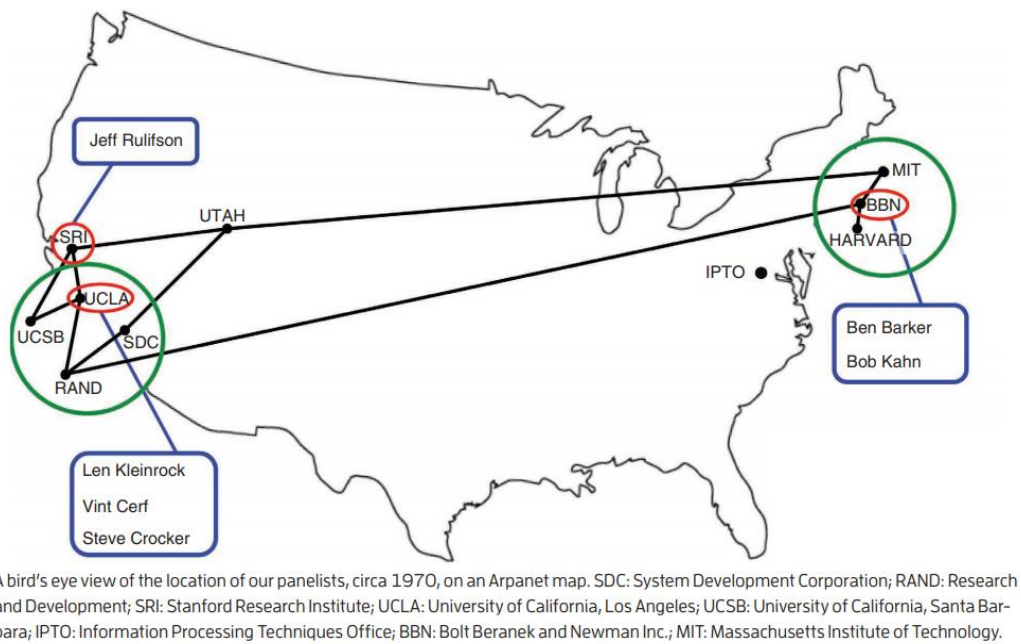
## 1.3 Internet

### 1.3.1 Progreso del Internet

A finales de los años de 1960, comenzó el desarrollo de lo que hoy conocemos como Internet. Este proyecto surgió en el contexto de la Guerra Fría, ya que Estados Unidos (EE. UU.) quería mantener su liderazgo en tecnología militar después del lanzamiento del Sputnik por parte de la Unión Soviética en 1957. El Departamento de Defensa de los Estados Unidos (DoD) se dio cuenta de que la tecnología de conmutación de circuitos utilizada por las redes telefónicas era demasiado vulnerable ante cualquier tipo de ataque, incluyendo el temor a una posible guerra nuclear. Considerando de que, si se destruye el enlace entre los dos principales intercambios, o si uno de los intercambios queda fuera de servicio, algunas partes de las comunicaciones de defensa nacional pueden volverse inutilizables (Poe, 2011).

El gobierno de EE. UU. respaldó el desarrollo de Internet y las redes de conmutación de paquetes con su financiamiento y patrocinio, como lo ha hecho con muchas otras tecnologías. La Agencia de Proyectos de Investigación Avanzada (ARPA) adoptó la idea de la conmutación de paquetes y creó Arpanet, una red principal de computadoras gubernamentales que podía resistir interrupciones en la red causadas por conflictos bélicos y desastres naturales. En colaboración con varias empresas y universidades, los esfuerzos de ARPA culminaron con el envío de la minicomputadora Honeywell 516 a la Universidad de California, Los Ángeles (UCLA) en septiembre de 1969. Esta computadora se convirtió en el primer conmutador de cuatro en total, también conocido como *Interface Message Processor* (IMP). Otros conmutadores de paquetes fueron instalados en el Instituto de Investigación de Stanford, la Universidad de California en Santa Bárbara y en la Universidad de Utah. Pronto, estas computadoras comenzaron a intercambiar paquetes de datos entre sí utilizando líneas telefónicas, lo que llevó al surgimiento del "Arpanet", considerado la "madre de Internet". El Arpanet de 1970 se puede ver en la figura 2. (Crocker, 2019).

Figura 2 La Arpanet original



Fuente: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8848151>

### 1.3.1.1 La web 1.0

La web 1.0 se inició en los años de 1960 y existió en su forma más simple con navegadores de texto sin formato, por ejemplo: ELISA, luego se desarrolló el HTML, que hacía que las páginas fueran más amenas en su apariencia, y se desarrollaron los primeros navegadores gráficos como Internet Explorer y Netscape.

Los usuarios de este sitio web sólo pueden leer el contenido y no tienen la capacidad de interactuar con él. Sólo el administrador del sitio puede cargar contenido en el sitio web.

### 1.3.1.2 La web 2.0

En 2004, O'Reilly introdujo el término "web 2.0" para describir una nueva generación de tecnologías web que incluyen redes sociales, wikis, foros, blogs y presentaciones en línea. Estas herramientas permiten una mayor interacción,



colaboración y conexión entre las personas. La web 2.0 es una web más dinámica y participativa que fomenta la colaboración y el intercambio de información a través de plataformas y redes sociales. La web 2.0 también se conoce como la red social debido a su enfoque colaborativo y social (O'Reilly, 2005).

Esto significa nuevas filosofías de navegación, nuevas formas de participar en redes y servicios de Internet que pueden cambiar el contenido de las bases de datos, los formatos o ambos. Los usuarios ya no se limitan a acceder a la información, sino a crearla. También, homogeniza lenguajes para mejorar la reutilización de código y mejora la compatibilidad entre aplicaciones y dispositivos (hardware- software)) (O'Reilly, 2005).

### 1.3.1.3 La web 3.0

“El término ‘**web 3.0**’ apareció por primera vez en 2006, en un artículo del diseñador de páginas web estadounidense Jeffrey Zeldman, crítico de la web 2.0 y fundador de la empresa *Happy-Cog* para el desarrollo de páginas web” (Argonza, 2011, p. 4). La evolución tecnológica hacia la web 3.0 ha generado un impacto notable en los usuarios que utilizan internet.

Web 3.0, también conocida como la web semántica, implica la interconexión de aplicaciones web para mejorar la experiencia del usuario, así como la conciencia del contexto geoespacial en la web y la capacidad del navegador para ser autosuficiente. Además, esta nueva web permite el uso óptimo de los datos, denominados "web de datos", y es interoperable, lo que significa que los usuarios pueden modificar directamente la base de datos. La web semántica utiliza metadatos semánticos y ontológicos para describir las relaciones entre los datos y el contenido, lo que permite que los sistemas de procesamiento los rastreen adecuadamente. (Argonza, 2011).

La intención de la web 3.0 es hacer que la información y las herramientas de Internet estén disponibles para todos, sin importar el dispositivo utilizado para conectarse, mediante la búsqueda de la flexibilidad y versatilidad para superar las barreras de formato y sistema. En otras palabras, la web 3.0 busca asegurar que los usuarios puedan acceder a la información y herramientas de Internet sin importar la plataforma o el sistema operativo que utilicen.

#### 1.3.1.4 La web 4.0

La web 4.0, empezó en 2012, representa el siguiente gran avance en la evolución de la web y se enfoca en permitir un comportamiento más adecuado y predecible para que los usuarios puedan realizar acciones específicas al hacer una declaración o solicitud y obtener los resultados deseados (Nath & Iswary, 2015).

¿Cómo se llegó a la web 4.0? Los avances en la propia tecnología han llevado a empresas como Google, Microsoft y Facebook a desarrollar nuevos sistemas que pueden procesar información similar al cerebro humano gracias a los aprendizajes tanto automático como profundo (Nath & Iswary, 2015).

Por medio de la web 3.0 los motores de búsqueda desempeñan un papel crucial en nuestra vida diaria en la web. Cuando utilizamos plataformas como Google, Bing o Yahoo!, por ejemplo, introducimos una palabra clave y luego revisamos una serie de resultados para encontrar la información que buscamos. Es decir, estos motores de búsqueda nos muestran diversas opciones relacionadas con nuestra búsqueda, y nosotros debemos navegar a través de ellos hasta encontrar lo que realmente estamos buscando (Nath & Iswary, 2015).

La web 4.0 es una evolución de la web que ofrece una experiencia de usuario más personalizada y completa. En lugar de simplemente mostrar información, la web 4.0 tiene como objetivo actuar como un modelo inteligente que puede proporcionar soluciones específicas a las necesidades individuales de los usuarios. Esta nueva web utiliza tecnologías avanzadas y tiene una mejor comprensión de las intenciones y necesidades individuales de los usuarios para mejorar su experiencia (Nath & Iswary, 2015).

Las tecnologías incorporadas en la web 4.0 permiten a los usuarios interactuar con los dispositivos de una forma más natural y humana, incluyendo el reconocimiento de voz, la inteligencia artificial, la realidad aumentada y la realidad virtual. Además, estas tecnologías hacen posible una mayor personalización de la experiencia, adaptando la información y los servicios a las preferencias y necesidades de cada usuario de manera más precisa. Por ejemplo, se puede utilizar un dispositivo digital, como un teléfono inteligente o una computadora, para dar órdenes verbales, como "Comprar un boleto de autobús con ciertas características" o "Solicitar un Uber para un lugar y hora específicos", y el dispositivo realizará las acciones correspondientes

automáticamente. Esto implica que estamos avanzando de sitios web que nos proporcionan información a sitios web que nos brindan soluciones específicas para satisfacer nuestras necesidades.

La web 4.0 es capaz de aprovechar la capacidad de la computación cognitiva a través de computadoras potentes para almacenar y procesar una gran cantidad de datos y solicitudes. Esto no sólo permite el uso de dispositivos conectados a Internet, sino que también permite que las computadoras interactúen con los usuarios y recopilen datos para mejorar la experiencia y personalización del servicio. En otras palabras, la web 4.0 permite una interacción más profunda entre los usuarios y los sistemas informáticos, lo que lleva a una experiencia más personalizada y adaptada a las necesidades individuales (Nath & Iswary, 2015).

### 1.3.2 Localizador Universal de Recursos

URL (**Localizador Universal de Recursos**) es el “Sistema estandarizado de atribución de direcciones en Internet. Por extensión, URL designa igualmente la dirección de un sitio o de una página en la web” (Virga & Menning, 2000, p. 251).

Las URLs utilizan diferentes componentes de la dirección para dirigir las solicitudes de los observadores al servidor adecuado (Wessels, 2001).

La URL se compone de diferentes secciones, siendo la primera el **Protocolo de Transferencia de Hipertexto** (HTTP), que permite que el usuario realice solicitudes a los servidores web. El uso del protocolo HTTP en la cadena de inicio indica que los paquetes de datos se envían al servidor web. Sólo el servidor web puede entender la parte faltante de la cadena del localizador universal de recursos, que identifica el recurso específico que se está solicitando. La computadora del usuario envía una solicitud al servidor web para acceder a ese recurso utilizando el protocolo HTTP. Los recursos de la web suelen estar escritos en **Lenguaje de Marcado de Hipertexto** (HTML) o XML (Wessels, 2001).

La segunda sección de la URL es una doble barra inclinada //, lo que significa que el nombre de la computadora se encuentra a continuación.

La tercera sección de una URL identifica el tipo de computadora host, y comúnmente se utiliza la identificativa **WWW**: *World Wide Web* (telaraña

alrededor del mundo) para servidores web remotos. Si la URL incluye el protocolo "http", esto indica que la máquina es un servidor web, por lo tanto, las URL para acceder a archivos remotos en la web suelen comenzar con <http://www>.

La URL está compuesta por tres componentes principales que son:

1. El identificador de servicio de "http:".
2. Nombre de dominio del servidor.
3. La ruta que se debe seguir para ingresar al servidor.

**Protocolo de Transferencia de Hipertexto (HTTP)** (figura 4), "protocolo que sigue el modelo cliente/servidor usado por lo general entre un navegador de web y un servidor web. Realmente lo que establece HTTP es cómo recuperar hiper documentos distribuidos y enlazados a través de la web" (Prieto Espinosa, LLoris Ruiz, & Torres Cantero, 2004, p. 720). HTTP señala que el observador está enviando la solicitud mediante el Protocolo de transferencia de hipertexto. Este protocolo permite que se realicen solicitudes de red a un servidor web y está diseñado para responder a las solicitudes del navegador. Para usarlo, las estaciones de trabajo deben estar configuradas con el **Protocolo de Control de Transmisión TCP** (Severance, 2005). Se ejemplifica el funcionamiento del protocolo http en la Figura 3:

*Figura 3 Comunicaciones en el servidor / cliente utilizando protocolo de transferencia de hipertexto*



Fuente: <https://s3.amazonaws.com/s3.timetoast.com/public/uploads/photos/10088535/descarga.png>

Figura 4 Protocolo de Transferencia de Hipertexto



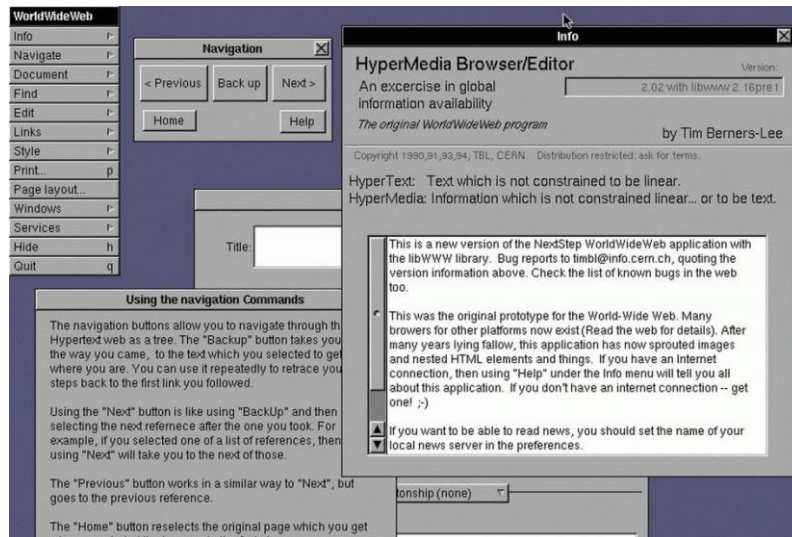
Fuente:[https://reader012.staticcloud.net/reader012/html5/20190414/547e8810b4af9fbe158b57ce/bg\\_3.png](https://reader012.staticcloud.net/reader012/html5/20190414/547e8810b4af9fbe158b57ce/bg_3.png)

### 1.3.3 Navegador web

Un Navegador web (o visor web) es un medio por el cual los clientes pueden navegar por la *World Wide Web*. Cuando un usuario ingresa una dirección de sitio web, el navegador web busca el servidor web que contiene esa página y solicita su acceso. Luego, el navegador espera a que el servidor envíe la información solicitada. Finalmente, la información es mostrada en la pantalla del usuario para su visualización (Comer, 2018).

Diferentes compañías, como *Netscape Communicator*, han creado navegadores web y los han comercializado. Tenemos los siguientes ejemplos: Microsoft con Internet Explorer, Google con Google Chrome y desarrollado mediante código abierto: Mozilla Firefox. Todos ellos ofrecen una variedad de funciones que cubren las necesidades diarias de los usuarios, por ejemplo: libretas de direcciones, comercio electrónico, correo electrónico, páginas de noticias, blogs, buscadores, redes sociales, entre otras. (Comer, 2018). A continuación, se visualiza en la figura 5. el primer navegador de la historia:

Figura 5 World Wide Web primer navegador de la historia



Fuente: <https://upload.wikimedia.org/wikipedia/commons/thumb/8/8c/WorldWideWeb.png/250px-WorldWideWeb.png>

La *World Wide Web* es una gran cantidad de información, como archivos, videos, imágenes, sonidos y otros, que se encuentran alojados en distintas computadoras alrededor del mundo. Se requiere un software especial llamado "navegador web" para acceder a Internet y disponer la computadora del usuario con todas las capacidades (de software y hardware) necesarias (Comer, 2018).

### 1.3.4 Internet

“Internet es la totalidad de todas las computadoras que están en red (utilizando diversas tecnologías de red) y emplean el conjunto de protocolos de Internet además de sus sistemas de red. El conjunto de protocolos de Internet implementa una red de conmutación de paquetes de área amplia que puede interconectar redes utilizando diferentes protocolos de red y características de conexión muy diferentes.” (Wilde, 2012, p. 18)

En el centro de Internet, hay una red principal de líneas de datos de alta velocidad que conectan los principales nodos de servidores de la red. Esta red incluye diversos tipos de sistemas informáticos, incluyendo instituciones educativas, empresas, agencias gubernamentales, organizaciones científicas y otros. Estos sistemas son responsables de enrutar diversos tipos de información, como datos, audio, vídeo, etc. La comunicación entre ellos se lleva a cabo mediante una variedad de medios, como antenas, satélites, líneas telefónicas, fibras

ópticas, medios digitales, inalámbricos, módems, líneas dedicadas y conexiones similares (Pollard, 2019). Se ejemplifica en la figura 6.

*Figura 6 Internet Global*



Fuente: <https://img.genial.ly/5f9c8e7f6a58bf21b87a9c8a/8d044b8d-0e1d-4831-9548-603a11659385.jpeg>

### 1.3.5 *World Wide Web*

En 1989, la **World Wide Web** (WWW) comenzó a surgir como una serie de páginas interactivas en Internet. La palabra WWW, que en español significa "telaraña alrededor del mundo" (figura 7), fue propuesta por el físico británico Tim Berners-Lee (Berners-Lee, Cailliau, Iuononen, Nielsen, & Secret, 1994). Él propuso un **sistema de comunicación** apoyado en el uso de redes informáticas, permitió a los investigadores que trabajaban en el mismo tema (a menudo apartados a miles de kilómetros de distancia) acceder instantáneamente a los datos generados por sus colegas. información y bases de datos, documentos, etcétera, sin verse en la necesidad de viajar por el mundo (Pollard, 2019).

Figura 7 www “Telaraña alrededor del mundo”



Fuente:<https://s3.amazonaws.com/s3.timetoast.com/public/uploads/photo/16264081/image/original-e0ea3c47a04ee924aa43853c73ed66f5.jpg>

En 1993, Internet empezó a hacerse presente en la cultura popular gracias al **primer navegador que representaba de manera gráfica las páginas de la World Wide Web.**

“Básicamente, la *World Wide Web* es un sistema hipermedia distribuido, con información almacenada en forma de páginas web, que se vinculan entre sí mediante enlaces web (más conocidos por sus nombres oficiales URI o URL). Esta propiedad de la web hace necesario disponer de un medio de acceso a información remota desde cualquier sistema recuperando información de la base de datos de la web (la cual está formada por todas las páginas web que están disponibles a nivel mundial)” (Wilde, 2012, p. 53).

Las redes conectadas no tienen que estar en la misma localización geográfica o edificio; pueden estar físicamente separadas, y conectarse mediante líneas de datos dedicadas como satélites, radios, enlaces infrarrojos, televisión por cable, módems, líneas telefónicas regulares, y fibras ópticas. Esto permite que la computadora remota se sienta como si estuviera en el mismo lugar físico y facilita la transferencia de archivos, correo electrónico, comercio electrónico, entre otros servicios. Además, también es posible compartir recursos como impresoras, carpetas, unidades de almacenamiento y otras funciones, incluyendo el acceso a Internet.

En el modelo de la web, un visualizador en la máquina del cliente se utiliza para acceder y visualizar páginas web alojadas en servidores. Los hipervínculos son palabras o frases destacadas que permiten a los usuarios saltar a otras páginas web en la misma máquina o en cualquier otra parte de la red. La capacidad de vincular información entre diferentes sistemas de computadoras de la web es una de las funciones más importantes de la red. El **Lenguaje de Marcas de**



**Hipertexto** (HTML) especifica la dirección del servidor web central que está asociada con la palabra o frase resaltada. Al seleccionar un hipervínculo, el usuario accederá a ese servidor y descargará el archivo o información especificada en la dirección, que se descargará en su computadora. (Wilde, 2012)

### 1.3.6 Servidor web

Un servidor web es una computadora que ofrece servicios a los clientes que solicitan información a través de un navegador web o un motor de búsqueda. Un servidor web no necesita ser una computadora exclusiva, es decir, un servidor de una manera física que sea un aparato que sólo se utilizaría para esa finalidad. En un contexto relacionado con la web, el término servidor se refiere a un proceso en una computadora que implementa funciones para responder a las peticiones de los usuarios. Técnicamente, cualquier computadora con una conexión de red puede realizar la función de un servidor web. Un ejemplo para entenderlo mejor es cuando dos programas se comunican entre sí a través de una red. Uno de los programas es el que inicia la comunicación y se llama cliente, mientras que el otro programa espera la conexión y se llama servidor. Es importante destacar que cualquier programa puede desempeñar el papel de servidor para un servicio específico y como cliente para otro servicio diferente. En otras palabras, un programa puede realizar ambas funciones según sea necesario (Wilde, 2012).

El servidor es una estructura informática que administra y controla el acceso a redes y sus recursos, como impresoras y archivos compartidos. Algunos servidores permiten el acceso a información alojada en bases de datos o publicada en sitios web, también es importante considerar que otros servidores se encargan de dirigir flujos de datos entre diferentes servidores y sistemas que se encargan de las copias de seguridad. El objetivo de utilizar un servidor es satisfacer las necesidades del cliente. (Wilde, 2012).

En la mayoría de los casos en Internet, la comunicación se establece entre un cliente y un servidor. Por ejemplo, cuando un usuario utiliza una aplicación de correo electrónico, el cliente (la aplicación) solicita información al servidor de correo, como los mensajes entrantes y la información del remitente. A su vez, el cliente puede enviar información al servidor, como el correo electrónico que se

está enviando. La información es un recurso muy valioso en Internet y permite a los usuarios acceder a una variedad de datos tanto públicos como privados (Severance, 2005).

El objetivo principal de un servidor web es almacenar un conjunto de páginas web, recibir y atender las peticiones de los navegadores web para mostrar dichas páginas, y facilitar la interacción entre el navegador y el servidor.

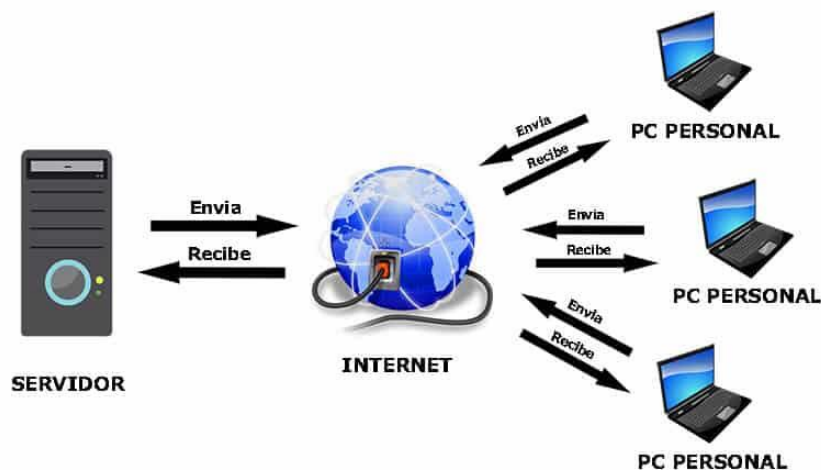
Un servidor web es como una biblioteca que contiene libros (páginas web y aplicaciones), pero en lugar de prestarlos físicamente, los proporciona a través de un navegador web. Este servidor puede ser una computadora con sistema operativo Linux, host, Solaris, Windows 10 o Windows Server 2022, que ha sido configurada con el hardware y software adecuados para atender las solicitudes de los usuarios. En términos generales, los servidores web descargan y presentan páginas web y aplicaciones a los usuarios.

Cuando un servidor web deja de funcionar, los usuarios no podrán acceder a las páginas web que están alojadas en ese servidor hasta que se solucione el problema y se restaure el servicio. Cuando un navegador web solicita una página, ésta es descargada y mostrada por el servidor web correspondiente. El servidor espera a recibir la solicitud del navegador antes de descargar y mostrar la página desde el sitio seleccionado. Una vez que se recibe la petición de descarga o visualización, el servidor web busca el documento o lugar solicitado y lo envía de vuelta al navegador para que el usuario pueda visualizarlo. En resumen, la función principal del servidor web es responder a las solicitudes del navegador para descargar y acceder a las páginas o sitios solicitados (Severance, 2005).

Un servidor web es fundamental en un sistema de intranet web, ya que permite publicar información para los clientes de una corporación o institución. Los documentos publicados pueden contener textos, gráficos, videos y audios para que los clientes puedan acceder a ellos mediante un navegador web desde sus propias computadoras. Además, los servidores web también pueden ejecutar programas que interactúen con bases de datos y otros dispositivos. A modo de ejemplo, un servidor web puede enviar notificaciones de correo electrónico automáticamente cuando ocurren procesos importantes en el servidor, y también proporcionar herramientas de administración para facilitar la gestión del servidor de forma más eficiente y efectiva (Severance, 2005).

Los servidores web tienen programas que permiten configurar distintos niveles de seguridad y administrarlos desde la computadora o dispositivo. Esto significa que se pueden establecer diferentes niveles de acceso para diferentes oficinas, departamentos y personas dentro y fuera de la organización. De esta manera, se puede controlar quiénes tienen acceso a qué información y garantizar que los datos confidenciales estén protegidos. Además, puede usar el administrador de archivos del programa para establecer permisos de lectura y escritura, lo que le permite establecer diferentes niveles de acceso para carpetas y archivos. Establecer niveles de seguridad ayuda a proteger la información privada mientras se permite que cierta información llegue al público en general. Esta opción también está disponible en configuraciones típicas de redes de área local (LAN) (Severance, 2005). La figura 8 representa los elementos comunes que suelen formar parte de un servidor web:

*Figura 8 Servidor web*



Fuente: <https://www.webebre.net/wp-content/uploads/2019/10/servidor-web.jpg>

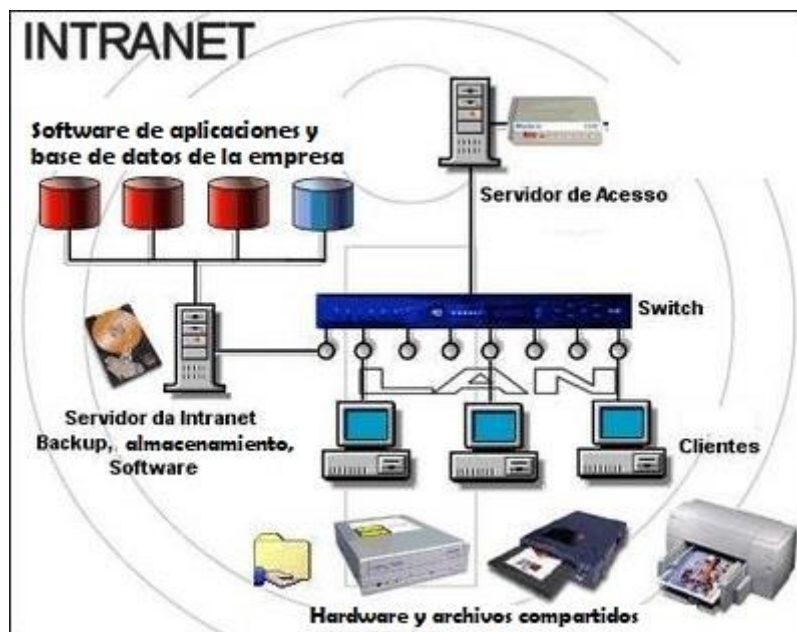
### 1.3.7 Sistema Intranet

Las intranets son una herramienta valiosa para las empresas ya que facilitan la comunicación y el intercambio de información entre los empleados. Sin embargo, no es suficiente con sólo implementar estas herramientas. Es importante tener en cuenta que el éxito de la comunicación y el intercambio de información depende de muchos factores, como la colaboración entre las personas, la eficacia de los procesos, la calidad del contenido y la tecnología utilizada. Para lograr una adopción y un intercambio de información efectivos, es necesario

adoptar una visión amplia y considerar todos estos factores. (Kennedy & Dysart, 2007).

**Definición de intranet:** “Red local que utiliza protocolos y aplicaciones de Internet para resolver los problemas de comunicación interna en una empresa. Se trata de un Internet en miniatura, sólo que limitada a la red de la empresa y, sobre todo, separada del mundo exterior por unas barreras de protección llamadas “firewalls”” (Virga & Menning, 2000, p. 139). Se refiere a una red privada que conecta varias computadoras y utiliza los protocolos de Internet estándar, pero que requiere una identificación de usuario para acceder a ella. Esta red se ha creado para organizar y distribuir información, así como para llevar a cabo transacciones digitales dentro de una organización o empresa. Se trata de una red privada que utiliza aplicaciones de Internet como páginas web, sistemas, navegadores, correo electrónico, directorios de correo y teléfono, grupos temáticos y comercio electrónico. Sin embargo, el acceso a estas aplicaciones está limitado sólo a los miembros de la organización que han sido autorizados para acceder a ellas (Kennedy & Dysart, 2007). La figura 9 representa el sistema de intranet que fue mencionado previamente.

*Figura 9 Estructura de intranet*



Fuente: [http://2.bp.blogspot.com/\\_5A6TC8quXS8/TE8n3yspz\\_I/AAAAAAAAAEE/vXq\\_LKO9bR4/s1600/intranet.jpg](http://2.bp.blogspot.com/_5A6TC8quXS8/TE8n3yspz_I/AAAAAAAAAEE/vXq_LKO9bR4/s1600/intranet.jpg)

La intranet es un sistema informático utilizado por empresas u organizaciones para administrar sus operaciones. Utiliza los protocolos TCP/IP o WAP para conectar los dispositivos en una red y permitir la interacción entre ellos.

Las intranets se basan principalmente en la tecnología de Internet creada en torno al protocolo TCP/IP. En la actualidad, los encargados de sistemas de información tienen la opción de elegir entre diferentes tecnologías que compiten con la intranet de la organización o empresa y que también utilizan Internet. Por lo tanto, deben decidir cuál de estas tecnologías es la más adecuada para sus necesidades y objetivos empresariales.

## 1.4 Minería web

“La minería web tiene como objetivo descubrir información útil o conocimiento de la estructura de hipervínculos web, el contenido de la página y los datos de uso. Aunque web *mining* utiliza muchas técnicas de minería de datos, como se mencionó anteriormente, no es puramente una aplicación de las técnicas tradicionales de minería de datos debido a la heterogeneidad y la naturaleza semiestructurada o no estructurada de los datos web. En la última década se inventaron muchas tareas y algoritmos de minería nuevos. Según los principales tipos de datos utilizados en el proceso de minería, las tareas de minería web se pueden clasificar en tres tipos: minería de estructura web, minería de contenido web y minería de uso web” (Liu, 2011, p. 7).

La investigación web se enfoca en cómo los sitios web se comparan con otros y cómo pueden atraer y retener a los usuarios convirtiéndolos en clientes potenciales o visitantes recurrentes. Para lograr esto, se utilizan herramientas que estudian el comportamiento de los usuarios y sus patrones de navegación para mejorar la experiencia del usuario en función de sus necesidades. Más tarde surgió la minería web, que busca recopilar información para enriquecer aún más la experiencia de navegación del usuario. Esto implica reconocer datos, documentos y elementos multimedia para analizarlos posteriormente y encontrar patrones útiles (SCIME, 2005).

La minería de datos es el análisis de diferentes conjuntos de datos para descubrir relaciones entre ellos. Cuando esta metodología se aplica a los datos web, se

llama **minería web**. La finalidad de la minería web es descubrir patrones en el contenido, la estructura y el uso de la web (SCIME, 2005)

En otras palabras: minería de datos + datos web = minería web.

La cantidad de datos que se encuentran disponibles actualmente para su análisis ha llevado a un importante crecimiento en la investigación en esta área (Kosala & Blockeel, 2000)

#### 1.4.1 Proceso de minería web para obtener conocimiento útil y significativo

La minería web es un proceso de exploración y análisis de datos web con el objetivo de obtener conocimientos útiles y significativos. Este proceso se lleva a cabo en varias fases, que se describen a continuación:

1. Recopilación de datos: en esta fase, se recopilan datos web relevantes que serán utilizados para el análisis. Esto puede incluir la extracción de datos de sitios web, la obtención de datos de registros de servidores web o la recopilación de datos de redes sociales (Liu, 2011).
2. Preprocesamiento de datos: los datos web recopilados pueden estar en diferentes formatos y pueden contener errores o información redundante. En esta fase, se realiza el preprocesamiento de datos para asegurarse de que los datos sean coherentes y estén en el formato correcto para el análisis posterior (Liu, 2011).
3. Análisis de datos: en esta fase, se utilizan técnicas de minería de datos para explorar y analizar los datos recopilados. Esto puede incluir técnicas de clasificación, agrupamiento, asociación y predicción (Liu, 2011).
4. Interpretación y evaluación de resultados: después de completar el análisis de datos, se interpretan y evalúan los resultados para determinar su relevancia y utilidad en el contexto de la minería web. Esto puede incluir la identificación de patrones, tendencias y relaciones entre los datos (Liu, 2011).
5. Aplicación de resultados: finalmente, se utilizan los resultados de la minería web para tomar decisiones y hacer recomendaciones. Esto puede

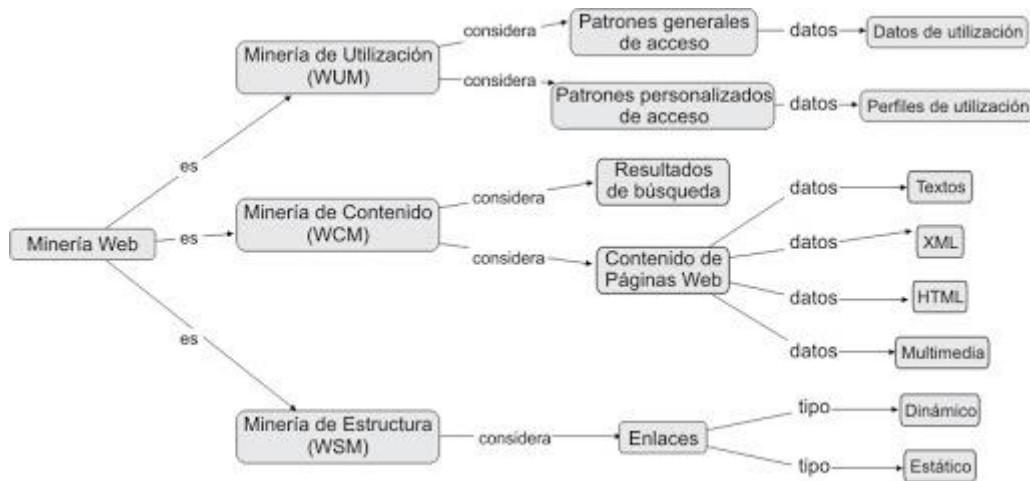
incluir la mejora de la experiencia del usuario en un sitio web, la optimización del motor de búsqueda o la identificación de oportunidades de mercado (Liu, 2011).

La minería web es un proceso que utiliza el contenido de las páginas web, la estructura de los enlaces y la estadística para ayudar a los usuarios a encontrar información. Hay tres tipos de datos web que son importantes en este proceso: contenido (textos, imágenes, sonidos, videos, etc.), estructura (datos que ayudan a determinar la organización del contenido, como HTML, XML, PHP) y uso (datos que indican las preferencias del usuario mientras navega por el sitio, como el tiempo de permanencia en la página o los artículos leídos, y en ocasiones datos más detallados como el nombre, intereses y correo electrónico del usuario). El análisis de estos datos es fundamental para identificar patrones y relaciones que resulten valiosos en la mejora de la experiencia del usuario y en la eficacia del sitio web en cumplir sus objetivos (Velásquez J. D., Yasuda, Aoki, & Weber, 2004). Los pasos clave en este proceso son:

- **Descubrir Recursos:** Este paso involucra la identificación de fuentes y datos web relevantes que serán analizados.
- **Selección y Preprocesamiento:** Aquí se eligen los datos que serán útiles y se someten a procesos de limpieza y organización para su análisis.
- **Reconocimiento de Patrones:** En esta etapa, se aplican técnicas de minería de datos para identificar patrones y relaciones en los datos recopilados.
- **Validación e Interpretación:** Finalmente, los patrones descubiertos son validados y se interpretan en función de los objetivos del análisis, lo que permite tomar decisiones informadas y mejorar la eficiencia y relevancia de un sitio web.

Como se ha mencionado anteriormente, se pueden identificar tres áreas principales en la taxonomía de la investigación en minería web, las cuales están representadas jerárquicamente en la figura 10.

Figura 10 Representación jerárquica de las áreas de la minería web



Fuente: <http://www.hipertexto.info/images/mapa-webmining.gif>

#### 1.4.1.1 Minería de utilización:

La minería de uso es un proceso en el que se analizan los registros de actividad de los usuarios en un sitio web para obtener información sobre su comportamiento y comprender su significado e interpretación. La personalización de los sitios web es un aspecto clave que se beneficia de la minería de uso, ya que ayuda a los propietarios del sitio a adaptar la experiencia del usuario a sus necesidades y preferencias (Liu, 2011).

#### 1.4.1.2 Minería de estructura:

La minería de estructura se refiere al proceso de identificar cómo está organizado un sitio web y las relaciones entre sus páginas, por ejemplo, la clasificación de las páginas, la divulgación de la página y las relaciones de la página. Esto se logra mediante el análisis de los hipervínculos y representando el sitio web como un grafo canalizado (Liu, 2011).

#### 1.4.1.3 Minería de contenido:

La minería de contenido implica analizar el contenido de una página web para encontrar la información más importante y relevante para los usuarios. Hay dos formas importantes de hacerlo: la primera es enfocarse en una página web en particular para extraer información de ella, y la segunda es mejorar los resultados de búsqueda al adaptar el contenido de la página a los patrones de búsqueda de los usuarios (Liu, 2011).



Los tres procesos de minería web (uso, estructura y contenido) siguen un procedimiento secuencial que se compone de los siguientes pasos: recopilación de datos, preparación de los datos, extracción de datos, exposición de resultados, valoración e interpretación de resultados, y toma de acción basada en los resultados obtenidos. Es decir, primero se recopilan los datos relevantes, luego se preparan para su análisis, después se extraen las informaciones relevantes, se exponen los resultados obtenidos, se valora e interpreta dichos resultados y finalmente se toman decisiones y acciones en función de los mismos.

#### 1.4.2 Técnicas de minería web

Las técnicas de minería web surgieron para aplicar la teoría de la minería de datos a los datos de la web y descubrir patrones similares entre ellos. Los primeros métodos utilizados para representar la minería de contenido web son:

**Clasificación:** es una técnica que se utiliza para asignar una o varias categorías a documentos como páginas web. Para hacer esto, se necesita analizar previamente un grupo de datos ya clasificados para luego aplicar el mismo criterio a otro grupo similar de datos. Se repite el proceso de preparación hasta obtener una clasificación correcta que coincide con la esperada. A este enfoque se le llama aprendizaje supervisado, ya que se basa en datos previamente clasificados para mejorar la precisión de la clasificación (Liu, 2011).

**Clustering:** se basa en agrupar documentos en función de su similitud o diferencia, sin la necesidad de una clasificación previa, es decir, es una preparación no supervisada. En este proceso, se busca que cada sección tenga elementos similares entre sí y que sean distintos de otras secciones. El objetivo es dividir la agrupación de datos total en grupos definidos por una medida de homogeneidad, con el fin de maximizar la diversidad entre las clases (Liu, 2011).

Las técnicas de clústeres son:

- El *clustering* particionado: es un método que consiste en dividir un grupo de elementos en subgrupos de tal manera que cada elemento se encuentre en al menos uno de estos subgrupos definidos.

- El *clustering* jerárquico es una técnica para agrupar elementos de un conjunto de datos en grupos jerárquicos. Para hacer esto, existen dos enfoques principales: la aglomeración y la división. En la técnica de **aglomeración** éste comienza considerando cada elemento como su propio grupo y luego los combina gradualmente en función de cuán parecidos o cercanos son entre sí. Este proceso continúa hasta que se cumpla una cierta condición. Mientras que en la técnica de la **división** inicialmente, se asume que todos los datos pertenecen a un solo grupo. Luego, se dividen en grupos más pequeños basándose en su similitud, nuevamente siguiendo una regla predefinida. En ambos casos, el proceso se repite varias veces hasta que se cumpla una condición de terminación. Esto permite crear una jerarquía de grupos que representan diferentes niveles de similitud o diferencia entre los elementos de datos.
- El *clustering* basado en la densidad: es un enfoque que se basa en el concepto físico de densidad y utiliza tres conceptos principales: densidad de umbral, cardinalidad y radio. En este método, un elemento pertenece a un grupo si su distancia al centroide de ese grupo es menor que el radio. La agrupación se detiene cuando la cantidad de elementos en cada grupo supera la densidad umbral. En caso contrario, se redefine el centroide y se repite el proceso de agrupación. Este método se utiliza para encontrar agrupaciones de alta densidad en un conjunto de datos.

### **Reglas de asociación:**

Las reglas de asociación son utilizadas para encontrar relaciones entre documentos basados en patrones comunes dentro del conjunto de datos completo. La regla de asociación se expresa como "si <A>, entonces <B>", donde A y B son subconjuntos de términos que no comparten ningún elemento en común. El objetivo es descubrir una relación implícita entre A y B, en función de las páginas web que los usuarios visitan. El soporte  $\alpha$  y la confianza  $\beta$  son medidas importantes de la regla de asociación. El soporte  $\alpha$  representa el porcentaje de veces que A y B aparecen juntos en un conjunto de visualizaciones V, mientras que la confianza  $\beta$  representa el porcentaje de veces que B aparece junto con A o cualquier otro elemento de B en V. Por lo tanto, el  $\beta\%$  de los

usuarios que ven la página A también ven la página B en un  $\alpha\%$  de las visualizaciones totales (Liu, 2011).

### **Descubrimiento de patrones secuenciales:**

La técnica de descubrimiento de patrones secuenciales se enfoca en encontrar patrones de eventos que suceden en un orden específico en el tiempo, como la secuencia de páginas web visitadas en un sitio. Esta técnica se utiliza para analizar el comportamiento de los usuarios y descubrir patrones en su comportamiento en el sitio web. Es una extensión de las reglas de asociación que se utilizan para encontrar patrones en conjuntos de datos (Liu, 2011).

#### 1.4.3 Desafíos y limitación de la minería web

El crecimiento constante de la web ha creado muchos desafíos para la investigación en este campo, como la falta de datos confiables, la presencia de información falsa, la naturaleza cambiante de los datos, la ambigüedad semántica, entre otros. Estos problemas hacen que sea difícil determinar qué información es relevante y cómo procesar diferentes tipos de datos, lo que lleva a limitaciones en la recuperación de información. Por esta razón, es importante utilizar técnicas de minería de datos para hacer frente a estas limitaciones y analizar los datos de manera efectiva (Markov & Larose, 2007).

Teniendo en cuenta el desarrollo en la web con los datos distribuidos, con grandes dimensiones y en constante cambio, se vislumbran desafíos para la minería web. Es necesario desarrollar técnicas más avanzadas y confiables para seleccionar los datos que son importantes y relevantes en un momento dado, y también para poder entender, interpretar y visualizar los patrones y tendencias que se detectan durante el proceso de análisis. En resumen, la minería web necesita adaptarse a las características cambiantes y dinámicas de la web y desarrollar herramientas más poderosas para procesar y analizar los datos de manera efectiva (Markov & Larose, 2007).

Dado los desafíos previos de la minería web, se está utilizando el *soft computing*, que es una subdisciplina de la Inteligencia Artificial que se utiliza para problemas complejos. *Soft computing* es un conjunto de técnicas diseñadas para aprovechar la tolerancia de la imprecisión y la incertidumbre para lograr una

solución controlable, robusta y de bajo costo. Sus principales componentes son la computación neuronal, la lógica, la neurocomputación y el razonamiento probabilístico (Chakraverty, Sahoo, & Mahato, 2019). *Soft computing* se basa en el modelo de la mente humana, por lo que ha suscitado mucho interés y actualmente tiene un gran potencial y progreso, especialmente en varias áreas y en las primeras etapas de las aplicaciones de minería web.

---

## Capítulo 2. Comportamiento del usuario en la búsqueda de contenidos en el sitio web

---

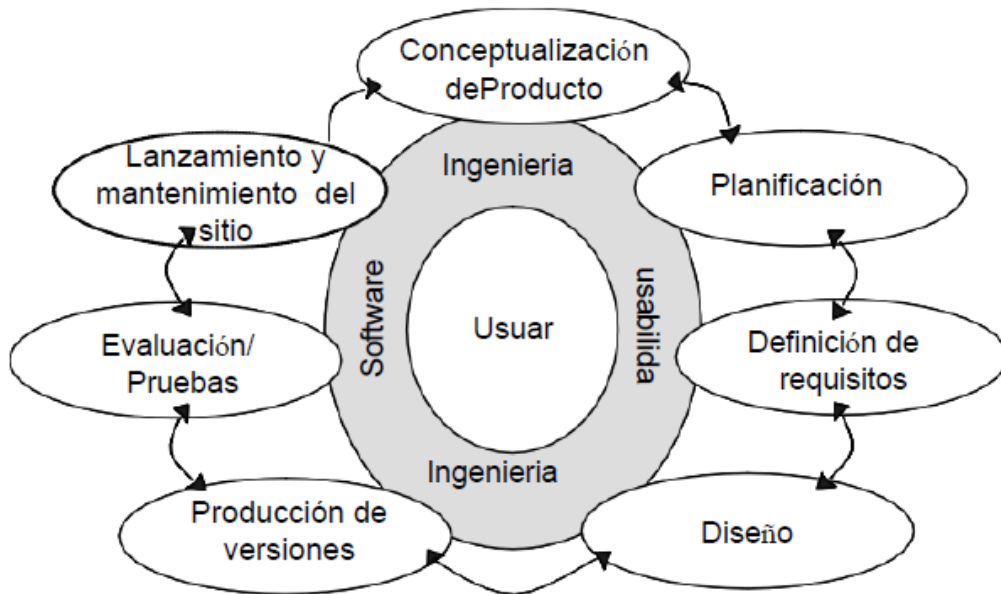
### Introducción

En este capítulo, se profundiza en el comportamiento del usuario y cómo éste influye en la búsqueda de contenidos en un sitio web. Se exploran los diferentes factores que afectan el comportamiento del usuario, como la accesibilidad, la funcionalidad, la encontrabilidad, la utilidad, la estética y la credibilidad. Se destaca la importancia de la usabilidad como factor esencial en la experiencia del usuario en un sitio web. Este capítulo establece la relación entre el comportamiento del usuario y la optimización en un sitio web, proporcionando una comprensión detallada de los aspectos que deben tenerse en cuenta para mejorar la experiencia del usuario.

### 2.1 El usuario y la web

Las personas usan la web de diversas maneras. Su interacción con la web puede ser auto motivada o externamente motivada; su habilidad puede ser novata o experta; sus necesidades y expectativas pueden ser simples o complejas. Para crear una experiencia de usuario exitosa y satisfactoria en un sitio web, necesitamos comprender cuestiones como por qué las personas van a un sitio web; lo que esperan e intentan lograr en el sitio; y todo lo que afecta su experiencia. Un sitio web es el resultado de un conjunto de procesos, generalmente iterativos, que comienzan con la conceptualización, la planificación y la definición de requisitos, luego pasan al diseño, la producción de versiones y la evaluación y prueba, antes de culminar en el lanzamiento del sitio. Por ejemplo, la figura 11 incorpora elementos que los ciclos de vida iterativos típicamente incluyen. En la práctica, la secuencia y la frecuencia de actividades pueden variar (Zaphiris & Kurniawan, 2007).

Figura 11 Un ciclo de vida genérico, variable e iterativo de desarrollo de sitios web ilustra los puntos en los que la ingeniería de usabilidad es más beneficiosa



Fuente: (elaborado con base en AdapZaphiris & Kurniawan, 2007, p. 2)

Los sitios web se diferencian de las aplicaciones de software convencionales en que, en lugar de ser productos, son servicios. No son bienes de consumo tangibles que los usuarios toquen/disfruten después de comprarlos, sino la manera virtual de cómo compran, se distraen, se instruyen y se relacionan. Si los usuarios no logran sus objetivos, o si el sitio web no satisface sus necesidades, los usuarios pueden simplemente abandonar el sitio y buscar alternativas, podrán checar e irse con la competencia de otros sitios web de interés o que bien capten su atención (Verbeek & Slob, 2006).

Es lógico que la satisfacción del usuario final sea el principal factor para determinar si una página web es exitosa o no. Es más probable que los usuarios satisfechos pasen más tiempo en el sitio, lo visiten nuevamente o lo recomienden a otros.

La satisfacción del usuario es un criterio complicado de definir y explicar, pero uno de los factores importantes que la influyen es el diseño del sitio web, que puede moldear la experiencia del usuario para que logre sus objetivos de la manera más efectiva posible. Podemos llamar satisfacción condicionada al diseño del sitio web como satisfacción del usuario. (Zaphiris & Kurniawan, 2007)

La primera tarea para establecer un marco teórico que respalde el diseño y la evaluación de sitios web centrados en el usuario es identificar los factores de diseño que afectan la satisfacción del usuario y cómo se relacionan entre sí.

## 2.2 Factores del comportamiento del usuario

Se van a considerar siete factores de diseño clave que influyen en la satisfacción-insatisfacción de uso del usuario. La Tabla 1 define los siguientes criterios para clasificar los factores como higiénicos o motivadores (Morbille, 2005).

Se puede distinguir entre dos tipos de factores en la evaluación de un sitio web: los factores de higiene, que se refieren a aspectos necesarios para que el sitio web sea funcional y útil, y los factores motivadores, que buscan estimular y evocar la intención de uso del sitio web, por lo anteriormente expuesto se forma una moneda con dos caras distintas, no separables y por necesidad complementarias (Herzberg, Mausner, & Snyderman, 2017).

Las diferencias en las características perceptibles de dichos elementos de diseño pueden derivarse de su impacto en la satisfacción del usuario. Los factores de higiene no están destinados a la frustración, por lo que los usuarios no los notarán. Los usuarios desconocen si un sitio web cumple con estas características, sólo si están ausentes o fallan. Por el contrario, los factores motivadores son percibidos directamente por los usuarios a través de su satisfacción resultante (Herzberg, Mausner, & Snyderman, 2017).

*Tabla 1 Criterios para la clasificación higiénico-motivadora de los factores de diseño*

| <b>Criterios/Tipo de factor</b>          | <b>Higiénicos</b>        | <b>Motivadores</b>        |
|--|--------------------------|---------------------------|
| <b>Carácter perceptible</b>              | Carácter desapercibido   | Carácter apercibido       |
| <b>Impacto en la intención de uso</b>    | Impiden la desmotivación | Provocan intención de uso |
| <b>Impacto en la satisfacción de uso</b> | Impiden la frustración   | Provocan satisfacción     |

Fuente: (Montero, 2006, p. 244)

Se identificaron tres factores higiénicos, que son accesibilidad, funcionalidad y *findability* (se tradujo del inglés a “encontrabilidad”), mientras que los factores motivadores son utilidad, calidad estética y credibilidad. El factor de usabilidad, el cual es el séptimo factor, tiene una doble influencia tanto higiénica como motivadora, debido a su naturaleza objetiva y subjetiva, y su rol que desempeña como elemento principal en relación con otros factores de diseño (Morille, 2005). Los factores expuestos anteriormente se muestran en la figura 12.

*Figura 12 Factores de diseño enfocados a la satisfacción-no frustración de uso*



Fuente: (Montero, 2006, p. 244)

Más adelante, se define cada factor, se examina su relación con la usabilidad y se expone las razones por las que se clasifican como factores higiénicos o motivadores.

## 2.3 Accesibilidad

Se trata de una característica de calidad que denota la capacidad de un sitio web para ser accesible y utilizado por la mayor cantidad posible de personas, sin importar las limitaciones que puedan tener en cuanto a su capacidad o al entorno en el que se encuentren. (Hassan & Martín Fernández, 2004)

La accesibilidad está tan estrechamente relacionada con la usabilidad que puede considerarse incluida en ella. Al analizar un sitio web desde el punto de vista de la comunicación, la accesibilidad estará más enfocada por el perfecto funcionamiento de los canales de comunicación, mientras que la usabilidad estará más enfocada porque la información sea percibida y entendida correctamente, sin considerar ambas carecería de sentido. Aunque ambas



tienen sus propios métodos de diseño y evaluación, puede observarse en muchas obras que son completamente compatibles. (Zaphiris & Kurniawan, 2007).

Este es el factor que podemos clasificar de forma más clara como higiénico. La accesibilidad se refiere a la capacidad de “**acceso**” a un sitio web, por lo que, si un usuario no puede acceder y, como resultado, no puede utilizar el sitio web, experimentará frustración. Ahora bien, si un sitio web es accesible y no causa frustración al usuario, es posible que este atributo no sea notado o tomado en cuenta por parte del usuario. Esto significa que la accesibilidad como una característica del diseño sólo será percibida cuando falte o en ausencia de la misma.

La importancia ética de la accesibilidad radica en que este factor garantiza derechos fundamentales, como el acceso a la información, la no discriminación y la inclusión digital. Además, la accesibilidad otorga una mayor relevancia ética a los aspectos higiénicos del diseño de un sitio web.

## 2.4 Funcionalidad

El término de funcionalidad hace referencia a la correcta funcionalidad técnica de un sitio web. Independientemente de si un sitio web (como una aplicación de correo web) es accesible y fácil de utilizar, en caso de que no pueda llevar a cabo su función (como enviar correos electrónicos a sus destinatarios sin errores), no tendrá ningún valor para los usuarios. (Verbeek & Slob, 2006).

Como hemos visto, la funcionalidad está estrechamente relacionada con la utilidad, y la funcionalidad se puede definir como **la utilidad objetiva** y ésta se refiere a la habilidad de una aplicación para cumplir con las funciones técnicas que los usuarios desean realizar (Verbeek & Slob, 2006).

La relación entre la funcionalidad y la usabilidad radica en que muchos problemas o carencias de la funcionalidad también son considerados como rasgos de la usabilidad, debido a que pueden causar frustración al usuario durante la interacción con la aplicación. Un ejemplo al respecto de los más reconocidos de usabilidad web es el de los «vínculos rotos».

Del mismo modo, los usuarios pueden percibir los defectos de usabilidad como defectos funcionales cuando no lo son. Este sería el caso del incumplimiento en el diseño del principio heurístico de “visibilidad del estado del sistema”, porque si el sistema no reporta lo que está sucediendo todo el tiempo, después de cierto tiempo es posible que el usuario crea que la aplicación ya no funciona (Zaphiris & Kurniawan, 2007).

La frustración del usuario suele ser causada principalmente por los errores en una aplicación. Por lo tanto, la funcionalidad es un factor higiénico importante, ya que su presencia o ausencia afecta directamente la satisfacción del usuario durante el uso de la aplicación. Esto es similar a la accesibilidad, donde la falta de acceso puede causar frustración y desmotivación en el uso de la aplicación.

## 2.5 Encontrabilidad

Encontrabilidad se refiere a la capacidad del usuario para encontrar la información que está buscando en un período de tiempo razonable. Es importante asegurarse de que la información esté organizada de manera adecuada, con una estructura, descripción y categorización claras, para facilitar su búsqueda y recuperación. En definitiva, se trata de medir la facilidad con la que un usuario puede encontrar la información deseada en una página web (Lautenbach, Schegget, Schoute, & Witteman, 1999).

La organización de la información en un sitio web está estrechamente relacionada con la capacidad de los usuarios para encontrarla. Por lo tanto, la encontrabilidad se refiere a la habilidad del diseño de un sitio web para permitir que los usuarios encuentren la información que están buscando de manera efectiva. Es un factor importante para la recuperación de la información en los procesos de interacción. Una organización adecuada del contenido favorecerá las actividades de recuperación de información en cuanto a eficiencia y efectividad (Morbille, 2005).

Si los contenidos de una aplicación de software no están bien organizados y no hay una guía clara para navegar por ellos, esto puede afectar negativamente la facilidad de uso objetiva. La categorización y la etiqueta clara de los contenidos, junto con la ayuda para la navegación, son importantes para mejorar la

usabilidad de la aplicación. Por lo tanto, se concluye que la fase de diseño de la estructura es la etapa más importante para lograr una buena usabilidad en la aplicación final (Morbille, 2005).

La estructura de la información en un sitio web puede pasar desapercibida para el usuario, ya que se enfoca principalmente en la interfaz visual de la aplicación. Sin embargo, cuando el usuario tiene dificultades para encontrar la información deseada, se hace evidente la importancia de una estructura adecuada. El usuario experimenta frustración al no poder lograr su objetivo y esto puede deberse a una estructura de información inadecuada en el sitio web. Es por eso que es importante prestar atención a la organización de la información y asegurarse de que sea fácilmente accesible y comprensible para los usuarios. (Zaphiris & Kurniawan, 2007).

Aunque la anchura y profundidad de la infraestructura del navegador son factores importantes para la eficacia en la recuperación de información, no parecen tener un impacto significativo en la percepción de usabilidad por parte de los usuarios (Velásquez & Palade, 2008).

Estos datos exponen el factor higiénico de la encontrabilidad. Si el usuario necesita invertir una mayor cantidad de tiempo y esfuerzo para encontrar lo que busca, aumentan las posibilidades de que se sienta frustrado.

## 2.6 Utilidad

La utilidad de un sitio web se refiere a la cantidad de valor que proporciona a sus usuarios, en términos de ingresos, beneficios y ventajas. Podemos definirla como la relación entre las actitudes emocionales de los usuarios hacia el sitio web. Es importante destacar que la utilidad de un sitio web está relacionada con la percepción subjetiva que los usuarios tienen sobre él, y no necesariamente con factores técnicos u objetivos (Verbeek & Slob, 2006).

Aunque la utilidad no es un mero factor de diseño, no se puede separar esa característica del proceso de diseño. Los agentes que intervienen en este proceso no sólo necesitan asegurarse de que el producto/servicio de su trabajo sea utilizable, accesible o estéticamente agradable; pero también contribuyen de

manera activa con su conocimiento y creatividad para hacerla útil (Morbille, 2005).

Se puede establecer una similitud entre la relación de la usabilidad con la utilidad y la relación de la usabilidad con la accesibilidad. Si un sitio web no es accesible o no se carga correctamente, no se puede utilizar y, por lo tanto, no se puede obtener ningún beneficio de su utilidad. Esto es similar a la relación entre la usabilidad y la utilidad: si un sitio web no es fácil de usar, los usuarios no podrán aprovechar su utilidad. La usabilidad representa la medida a la cual un usuario puede beneficiarse de la utilidad de un sitio web.

Por otra parte, el impacto positivo en la percepción de utilidad se atribuye a la usabilidad, puesto que se describe como un factor de percepción y motivación en la definición de utilidad. La percepción de utilidad produce satisfacción, que son todos los factores que conducen a utilizar la intención con el mayor peso.

## 2.7 Estética

Este factor hace referencia a un aspecto hermoso, placentero y atractivo. El diseño estético es un diseño que agrada a los sentidos, a la imaginación y a nuestro entendimiento. (Zaphiris & Kurniawan, 2007).

El diseño estéticamente agradable también se considera más fácil de usar por parte de los usuarios. Esta correlación no está presente en todos los casos, lo que sugiere que la disponibilidad percibida tiene una interrelación más compleja, como lo muestra el modelo propuesto en la figura 12 (Morbille, 2005).

La relación entre la estética y la usabilidad de un sitio web también funciona en sentido contrario; si se diseña un sitio web con una alta usabilidad, también se logra un diseño estético y atractivo para los usuarios.

La influencia del aspecto estético no se limita a la usabilidad percibida, ya que también puede afectar positivamente otros factores como la utilidad subjetiva. Mientras que otros factores, tales como la accesibilidad, afectan la estética, el efecto no es necesariamente negativo, en lugar de lo que podría suponerse erróneamente. Los diseños visualmente elaborados y estéticamente placenteros no son incompatibles con la accesibilidad de esos diseños (Verbeek & Slob, 2006).

La estética son factores del diseño que se pueden catalogar claramente como factor motivador en base a la lista de criterios que se muestra en la Tabla 1. La estética evoca emociones que modulan el comportamiento emocional de los usuarios, influyen en el propósito de utilizar y generar satisfacción (Greever, 2020).

Entre todas las cualidades del diseño, la estética es la cualidad más importante que perciben los usuarios, el aspecto visual de un sitio web influye desde el primer instante de su utilización.

Se consideró que otros factores como la utilidad y la credibilidad del sitio web son más difíciles de evaluar en comparación con los factores higiénicos, que son percibidos sin que el usuario sea consciente de ellos. Debido a que la estética es un elemento que puede influir en la elección de los usuarios, se puede inferir que es importante para la decisión de empezar a usar un sitio web. (Zaphiris & Kurniawan, 2007).

Si bien otros factores motivadores, como la utilidad, tienen una mayor influencia sobre el propósito de utilizar, esto no se percibe inmediatamente como estético, por lo que los usuarios que se enfrentan a un sitio poco atractivo pueden no tener la motivación para usarlo, independientemente de la utilidad o de otras características favorables del sitio (Morbille, 2005).

Debido a esto, la estética y la accesibilidad son los principales factores que afectan el uso inmediato de un sitio web, ya que la estética puede motivar al usuario a usarlo y la accesibilidad lo habilita para hacerlo

## 2.8 Credibilidad

La credibilidad se refiere a cuán creíble es un sitio web; es una cualidad percibida a juzgar por el visitante. Un objetivo común en el diseño de sitios web es hacer que el sitio sea más creíble. Esto ayuda a transmitir un mensaje. Cuanto más creíble sea su sitio, más eficazmente podrá llegar a su audiencia y lograr sus objetivos. Una buena navegación lo ayuda a persuadir y animar a los visitantes a hacer lo que usted quiere que hagan (Kalbach, 2007).

Aunque, del mismo modo que el factor de utilidad, no es únicamente un factor de diseño ya que también está sujeto de variables externas, el diseño sigue siendo un componente esencial para proyectar confianza (Kalbach, 2007).

La forma en que los usuarios perciben la facilidad de uso de un sitio web es un factor muy importante para aumentar su credibilidad. Si los usuarios encuentran fácil de usar un sitio web, lo considerarán más confiable y profesional. Por otra parte, si examinamos un sitio web desde un punto de vista comunicativo, no meramente instrumental, la percepción de usabilidad se convierte en la certeza del usuario en el sitio web. Entre los elementos de diseño más importantes para generar confianza se encuentra la usabilidad del sitio web y el diseño gráfico de la página web, que además están relacionados con el factor de la estética (Kalbach, 2007).

En consecuencia, podemos categorizar la credibilidad como un elemento que estimula la motivación, porque se trata de una característica percibida que incentiva la intención de utilizar.

## 2.9 Usabilidad como factor esencial

Se puede distinguir la usabilidad por ser un factor higiénico y motivador al mismo tiempo, ya que tiene dos aspectos diferentes: la usabilidad objetiva (medida en términos técnicos y objetivos) y la usabilidad subjetiva (medida por la experiencia y percepción del usuario).

Los factores de higiene, como la funcionalidad, accesibilidad y encontrabilidad, están relacionados con la dimensión objetiva o intrínseca de la usabilidad, mientras que los factores motivadores, como la utilidad, estética y credibilidad, están vinculados con la dimensión subjetiva o percibida de la usabilidad, como hemos comprobado en los anteriores puntos.

La Organización Internacional para la Estandarización (ISO) describe la usabilidad de un producto de software basándose en la interacción entre el usuario y el producto en un contexto específico de uso, evaluando tres atributos: eficiencia, efectividad y satisfacción del usuario. La usabilidad se compone de diferentes propiedades que deben ser evaluadas para medirla adecuadamente. (Zaphiris & Kurniawan, 2007). La ingeniería de usabilidad se enfoca en métodos

y mecanismos de evaluación para corroborar las decisiones de diseño desde la perspectiva del usuario, lo que es una de las funciones principales de la práctica de usabilidad. En resumen, la ISO establece un marco de referencia para la evaluación de la usabilidad de los productos de software y ayuda a los diseñadores y evaluadores a garantizar que los productos sean fáciles y agradables de usar para los usuarios (Verbeek & Slob, 2006).

El desempeño del usuario en actividades interactivas, tales como hacer más fácil el aprendizaje, la habilidad de la memoria, la efectividad, la productividad y la comprensión, garantizará experimentalmente que los usuarios no se sientan frustrados ni desanimados. En este sentido, que un diseño sea usable depende sobre todo de que funciona debidamente, de que es fácilmente accesible y de que cuenta con la estructura de la información correcta (Morville, 2005).

Por otra parte, la valoración de los factores percibidos en una actividad de interrelación (usabilidad percibida, atractividad, comodidad y disfrute de la utilización) determinará la habilidad del diseño para generar satisfacción e incentivar su uso. Dicha habilidad dependerá de la utilidad, la credibilidad y la estética del sitio web (Kalbach, 2007).

Los profesionales de usabilidad también son responsables de sugerir diseños y rediseños de interfaces en función de los resultados de los estudios de evaluación, heurística de diseño utilizable, guías de usabilidad o modelos de diseño de interrelación.

Podemos concluir que la usabilidad es fundamental en el diseño debido a su dualidad conceptual, su relación con otros factores de diseño y su capacidad para predecir y evaluar la calidad de dichos factores en la práctica. Esto la convierte en un elemento fundamental para el diseño.

---

## Capítulo 3. Ciencia de datos

---

### Introducción

En este capítulo, se introduce el campo de la ciencia de datos y su evolución en el análisis de datos. Se exploran conceptos como el análisis con ciencia de datos o predicciones, el aprendizaje automático y el uso de la ciencia de datos en diversas áreas, como el servicio al cliente, los coches sin conductor y las predicciones. Además, se discuten aspectos relacionados con los datos, su estructura, tipos y técnicas de análisis. Se presenta el proceso CRISP-DM como una metodología ampliamente utilizada en proyectos de ciencia de datos. También se detallan los algoritmos genéticos y su aplicación en la optimización. Este capítulo proporciona una base sólida en ciencia de datos, que será fundamental para la aplicación del modelo de algoritmo en los siguientes capítulos.

### 3.1 La ciencia de datos en la era de los datos masivos

En el mundo actual, las organizaciones y las personas se centran más en los grandes datos y la inteligencia artificial. Puede parecer sorprendente que las personas y las organizaciones creen y recopilen más de 2,5 exabytes (1 exabyte equivale a  $10^{18}$  bytes) cada día. Esto significa que el volumen de datos ha aumentado significativamente al paso de los años. La mayor parte de las empresas han cambiado su modelo de negocio y lo han centrado más en los datos. Algunas organizaciones también han agregado nuevos departamentos en la empresa para realizar análisis de datos. Los estadísticos necesitarían analizar los datos cuantitativamente en el pasado, pero esto no es suficiente ya que los resultados del análisis sólo podrían hablar sobre el presente. Cuando surgieron procesos informáticos sólidos, tecnología en la nube y herramientas analíticas; la gente comenzó a usarlos para realizar análisis. Comenzaron a desarrollar modelos para analizar datos (Campbell, 2021).



“IBM define *Data Science* o Ciencia de Datos — ya en fechas más recientes— como el proceso de describir (extraer) conocimiento (*insights*) oculto a partir de cantidades masivas de datos estructurados y no estructurados, utilizando métodos como estadística, aprendizaje automático, minería de datos y analítica predictiva. Es un área multidisciplinar que está cambiando el modo en que las organizaciones resuelven problemas y ganan ventaja competitiva, y que lo concentra en las tres grandes disciplinas: *Computer Science* (informática), matemáticas / estadística y dominio del conocimiento” (Aguilar, 2019, p. 421).

Figura 13 Disciplinas de Ciencia de Datos



Fuente: (Aguilar, 2019, p. 422)

En términos simples, la ciencia de datos es una rama de las matemáticas y la estadística para obtener información útil y significativa sobre el conjunto de datos y las tendencias a partir de los datos o la información sin procesar. Puede procesar y administrar el conjunto de datos usando habilidades analíticas, comerciales y de programación (Campbell, 2021)

El campo de la ciencia de datos se remonta a sus raíces en la estadística. Este campo es una combinación de programación, perspicacia empresarial y estadísticas. Es importante aprender más sobre cada tema, para que se tenga

una idea de cómo abordar el proceso de aprendizaje. El arte de encontrar ideas y tendencias ocultas en el conjunto de datos se remonta a mucho tiempo atrás. Los antiguos egipcios analizaban los datos del censo para ayudarlos a recaudar impuestos de manera eficiente. También usaron análisis de datos para pronosticar cuándo podría haber inundaciones en el Nilo. Es importante aprender de los datos anteriores para identificar una tendencia o una idea en el conjunto de datos. Esto ayuda a la empresa para tomar decisiones informadas (Campbell, 2021).

Los datos se han convertido en el nuevo petróleo y todas las empresas, independientemente de la industria, buscan formas de administrar y almacenar grandes volúmenes de datos, como es el caso de Python, que es una de las herramientas más ampliamente utilizada en el campo de la ciencia de datos.

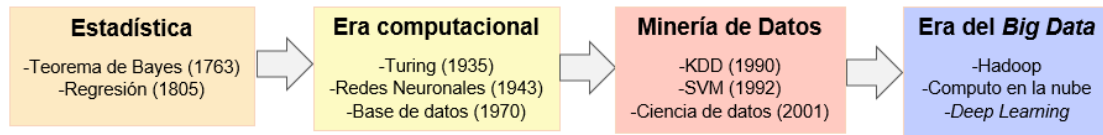
Éste fue un desafío para la mayoría de las empresas hasta 2010. El objetivo de cada empresa era definir un marco o solución que les permitiera almacenar grandes volúmenes de datos (Campbell, 2021).

La introducción de Hadoop y otras plataformas ha proporcionado a las organizaciones una forma más fácil de almacenar grandes volúmenes de datos, por lo que ahora se centran en métodos y soluciones para procesar la información. Esto sólo se puede hacer usando la ciencia de datos. Es importante tener en cuenta que la ciencia de datos constituye el futuro de la tecnología. Es importante saber qué es la ciencia de datos, especialmente si se desea agregar algún valor al negocio (Campbell, 2021).

## 3.2 Evolución del análisis de datos

Los datos han sido recopilados y analizados desde tiempos muy antiguos. La captura y difusión de información ha sido parte de la gestión de reinos, teniendo registros de tierras y ejército. El uso moderno de las estadísticas comenzó en el siglo XVIII con formas sistemáticas de capturar datos, y la evolución de la impresión del sistema ayudó a almacenar datos. (Probyto Data Science and Consulting Pvt. Ltd., 2020)

Figura 14 Evolución del análisis de datos



Fuente: (elaborado con base en Probyto *Data Science and Consulting* Pvt. Ltd., 2020, p. 3)

Las estadísticas han existido durante mucho tiempo con moderación en todo el mundo. El desarrollo sistemático de las estadísticas ocurrió en el siglo XVIII y se utilizó con fines administrativos en todo el mundo. Los grandes avances en la ciencia durante la era postindustrial sentaron las bases para la era de la informática. Comenzando con el trabajo innovador de Alan Turing en la Teoría de la computación y el avance en los semiconductores, las computadoras comenzaron a volverse más poderosas año tras año (Probyto Data Science and Consulting Pvt. Ltd., 2020).

A mitad del siglo XIX, se invirtió mucho trabajo de investigación en la comprensión de cómo aprende un cerebro humano y los avances en la comprensión de la estructura del cerebro. Surgieron los primeros artículos sobre redes neuronales. La metodología para aprender y repetir algunos eventos era completamente diferente de cómo explicaban los métodos estadísticos basados en distribución. Las bases de datos también comenzaron a ser de uso exclusivo en la década de 1980. La digitalización también comenzó a ser bien reconocida en la industria y el gobierno. Al mismo tiempo, estaban surgiendo los primeros experimentos con redes, correos electrónicos y una web interconectada (Probyto Data Science and Consulting Pvt. Ltd., 2020).

A finales de la década de 1990, las computadoras se volvieron comunes en los hogares de EE. UU. Microsoft, con su potente sistema operativo MS Windows 98, y la Macintosh de Apple también estaban disponibles en el mercado. Al mismo tiempo, la informática empresarial estaba en auge con gigantes como IBM convirtiéndose en los principales proveedores de potentes servidores e Internet siguiendo su ritmo. Esta vez, la minería de datos se volvió prominente para crear informes, analizar datos de clientes y tomar decisiones basadas en análisis de datos básicos (Excel estaba en Windows en ese momento). Las bases de datos de conocimiento, la máquina de vectores de soporte (SVM), las bases de datos SQL y el poder mayor de cómputo comercializaron las primeras etapas de la

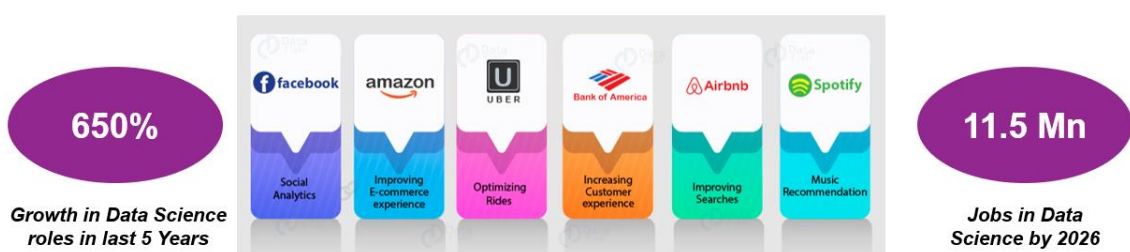
ciencia de datos antes de que explotara alrededor de 2008 (Probyto Data Science and Consulting Pvt. Ltd., 2020).

En 2006, el científico informático Doug Cutting lanzó el proyecto Hadoop, que revolucionó el almacenamiento y procesamiento de archivos distribuidos al anticipar el surgimiento del *Big Data*. A partir de marzo de 2009, Amazon ofreció el servicio de alojamiento de *MapReduce* conocido como *Elastic MapReduce*, marcando así el inicio de la era del Big Data. Al mismo tiempo, el uso de unidades de procesamiento gráfico (GPU) y la computación en la nube redujeron significativamente los costos de computación, lo que impulsó un rápido avance en el aprendizaje profundo y la infraestructura en la nube. En resumen, el proyecto Hadoop, junto con el surgimiento de servicios como *Elastic MapReduce* y la evolución de la GPU y la computación en la nube, han sido factores clave en el desarrollo del procesamiento de grandes volúmenes de datos y la implementación de soluciones de aprendizaje profundo en la nube (Probyto Data Science and Consulting Pvt. Ltd., 2020).

Hoy, tenemos una mejor comprensión de cómo la ciencia de datos crea valor para las empresas al hacerlas basadas en datos, así como de todo el ecosistema y los requisitos previos para hacer que el valor de los datos esté disponible ahora a un costo razonable (Probyto Data Science and Consulting Pvt. Ltd., 2020).

La ciencia de datos es una disciplina que ha experimentado un gran crecimiento y es muy atractiva para todas las organizaciones debido a sus beneficios. Las grandes empresas tecnológicas están liderando el desarrollo de la industria hacia organizaciones que utilizan datos como base para sus decisiones y estrategias (Probyto Data Science and Consulting Pvt. Ltd., 2020). En la figura 15, se puede ver que algunas de las empresas más grandes de la actualidad se basan en la última tecnología y en el proceso de toma de decisiones basado en datos:

Figura 15 Crecimiento de la ciencia de datos



Fuente: (Probyto Data Science and Consulting Pvt. Ltd., 2020, p. 4)

### 3.3 Análisis con ciencia de datos o predicciones

La ciencia de datos es una mezcla de numerosos algoritmos, herramientas, principios y lenguajes para identificar los patrones ocultos dentro de las variables en el conjunto de datos. Esto puede llevar a preguntarse en qué se diferencia esto de lo que se ha hecho con los datos durante años. Los datos se han vuelto en el nuevo petróleo y todas las empresas, independientemente de la industria, buscan formas de administrar y almacenar grandes volúmenes de datos. Éste fue un desafío para la mayoría de las empresas hasta 2010. El objetivo de cada empresa era definir un marco o solución que les permitiera almacenar grandes volúmenes de datos. La respuesta es que antes sólo podíamos usar herramientas y algoritmos para explicar las variables en el conjunto de datos, pero usando la ciencia de datos, se vuelve más fácil predecir los resultados. La ciencia de datos se utiliza para tomar decisiones informadas basadas en predicciones realizadas utilizando el conjunto de datos existente. Puede aplicar numerosos análisis al conjunto de datos para obtener esta información (Campbell, 2021).

#### 3.3.1 Analítica casual predictiva

Si se desea desarrollar un modelo que prediga las posibilidades o los resultados de un evento futurista, debe utilizarse la analítica causal predictiva. Supongamos que trabaja para una compañía de crédito y le presta dinero a la gente en función de su crédito. Le va a preocupar la capacidad de sus clientes para pagar la cantidad que les ha prestado. Puede desarrollar modelos para realizar un análisis predictivo de los datos utilizando el historial de pagos. Esto puede ayudarlo a determinar si el cliente le pagará a tiempo o no (Campbell, 2021).

#### 3.3.2 Análisis prescriptivo

Es posible que se necesite utilizar un modelo que pueda tomar las decisiones necesarias y modificar los parámetros en función del conjunto de datos o de la pregunta. Para hacer esto, es necesario usar análisis prescriptivos. Esta forma de análisis trata más de proporcionar la información correcta para que se pueda tomar una decisión informada. También se puede usar este tipo de análisis para predecir una variedad de resultados asociados y acciones prescritas. Un ejemplo

de este tipo de análisis es un automóvil autónomo. Siguiendo el ejemplo del automóvil autónomo puede ejecutar numerosos algoritmos en los datos recopilados de los automóviles y utilizar los resultados para hacer que el automóvil sea más inteligente. Esto facilita que el automóvil tome las decisiones correctas para girar, reducir la velocidad, acelerar o identificar la dirección a tomar (Campbell, 2021).

## 3.4 Aprendizaje automático

### 3.4.1 Hacer predicciones

Numerosos algoritmos de aprendizaje automático permiten hacer predicciones utilizando conjuntos de datos no estructurados, semiestructurados y estructurados. Supongamos que trabaja para una compañía financiera y tiene los datos transaccionales disponibles y necesita desarrollar un modelo para determinar la tendencia de transacciones futuras. Para realizar este análisis, debe utilizar un algoritmo de aprendizaje automático supervisado. Dichos algoritmos se utilizan para entrenar la máquina con un conjunto de datos existente. También puede usar algoritmos de aprendizaje automático supervisado para desarrollar y entrenar un modelo para detectar futuros fraudes en función de la información histórica (Campbell, 2021).

### 3.4.2 Descubrimiento de Patrones

No todos los conjuntos de datos tienen variables que pueden usarse para hacer las predicciones necesarias. Esto no es verdad. Hay un patrón oculto en cada conjunto de datos, y se necesita encontrar esos patrones para hacer las predicciones requeridas. Para hacer esto, se debe utilizar un modelo sin supervisión ya que carece de etiquetas predefinidas en el conjunto de datos con las que pueda agrupar las variables. Uno de los algoritmos más comunes utilizados para identificar patrones es el agrupamiento (Campbell, 2021).

Supongamos que trabaja para una compañía telefónica y tiene la tarea de identificar dónde instalar torres en un área para establecer una red. Entonces puede usar el algoritmo de agrupamiento para identificar dónde puede configurar torres para garantizar que todos los usuarios en el área reciban la potencia de señal óptima. Según los ejemplos anteriores, es importante comprender cómo los enfoques de ciencia de datos y análisis de los datos son diferentes. Este último incluye el uso de predicciones y análisis descriptivos sólo hasta cierto punto. Por otro lado, la ciencia de datos se trata más del uso del aprendizaje automático y el análisis casual predictivo (Campbell, 2021).

### 3.5 Uso de la ciencia de datos

Antes, las organizaciones trabajaban con cantidades pequeñas de datos y era sencillo analizarlos y comprender las relaciones utilizando herramientas de inteligencia empresarial. Sin embargo, con el tiempo, las organizaciones comenzaron a recopilar datos de una gran variedad de dispositivos, lo que resultó en volúmenes mucho más grandes de datos. Además, la mayoría de los datos recopilados en la actualidad son semiestructurados o no estructurados, lo que significa que no siguen un formato predefinido y no encajan fácilmente en las herramientas tradicionales de inteligencia empresarial diseñadas para datos estructurados. Es importante tener en cuenta este cambio en la naturaleza de los datos recopilados, ya que requiere enfoques y herramientas diferentes para su análisis y comprensión (Campbell, 2021).

Las herramientas simples de inteligencia comercial no pueden procesar este tipo de datos, especialmente porque se recopilan grandes volúmenes de datos de diferentes instrumentos. Es por esta razón que requerimos desarrollar herramientas de análisis complejas y avanzadas y algoritmos para procesar, analizar y extraer información de los datos. La ciencia de datos ha ganado popularidad no sólo por esta razón, sino también por su aplicación en diversos ámbitos. Ahora exploraremos cómo se utiliza la ciencia de datos en diferentes sectores. (Campbell, 2021).

### 3.5.1 Servicio al cliente

¿Qué beneficios podríamos obtener si pudiéramos saber exactamente lo que los clientes desean? ¿Cómo podemos aprovechar los datos existentes, como el historial de pedidos, la actividad de navegación web, los ingresos y la edad, para obtener más información sobre los clientes? Estos datos son generalmente accesibles a través de registros web. Gracias al uso de modelos matemáticos y estadísticos, podemos trabajar de manera eficiente con grandes volúmenes de datos y determinar los productos más adecuados para recomendar a los clientes existentes o potenciales. Esta estrategia puede resultar en un aumento de las ventas y la diversificación de productos para una empresa. (Campbell, 2021).

### 3.5.2 Coches sin conductor

¿Cómo te sentirías si tu auto pudiera llevarte a casa? Numerosas empresas están intentando desarrollar y mejorar el funcionamiento de un coche autónomo. Los autos recopilan información en vivo de varios sensores, como láseres, radares y cámaras. para crear un mapa del entorno circundante. El algoritmo del automóvil utiliza estos datos para decidir acelerar, reducir la velocidad, estacionar, detener, adelantar, etc. Estos algoritmos suelen ser algoritmos de aprendizaje automático (Campbell, 2021).

### 3.5.3 Predicciones

La ciencia de datos se puede utilizar para el análisis predictivo, como en la previsión meteorológica. Los algoritmos utilizados toman datos de aviones, satélites, radares, barcos y otras partes para recopilar y analizar datos. Esto sirve para ayudar a construir los modelos requeridos (Campbell, 2021).

## 3.6 Datos

### 3.6.1 Datos estructurados frente a no estructurados

Ciertos conjuntos de datos están bien estructurados, como las tablas de una base de datos o un software de hoja de cálculo. Otros registran datos sobre el estado del mundo, pero de forma más heterogénea. Tal vez sea un gran



compendio de texto con imágenes y enlaces como Wikipedia, o la complicada combinación de notas y resultados de pruebas que aparecen en los registros médicos personales. Los datos a menudo se representan mediante una matriz, donde las filas de la matriz representan elementos o registros distintos, y las columnas representan propiedades distintas de estos elementos. Por ejemplo, un conjunto de datos sobre ciudades de EE. UU. puede contener una fila para cada ciudad, con columnas que representan características como el estado, la población (Skiena, 2017).

Cuando nos enfrentamos a una fuente de datos no estructurada, como una colección de *tweets* de Twitter, nuestro primer paso generalmente es construir una matriz para estructurarla. Un modelo de bolsa de palabras construirá una matriz con una fila para cada *tweet* y una columna para cada palabra de vocabulario de uso frecuente. La entrada de matriz  $M[i, j]$  denota el número de veces que el *tweet*  $i$  contiene la palabra  $j$  (Skiena, 2017).

### 3.6.2 Datos cuantitativos frente a categóricos

Los datos cuantitativos consisten en valores numéricos, como la altura y el peso. Dichos datos pueden incorporarse directamente en fórmulas algebraicas y modelos matemáticos, o mostrarse en gráficos y cuadros convencionales. Por el contrario, los datos categóricos consisten en etiquetas que describen las propiedades de los objetos que se investigan, como el género, el color del cabello y la ocupación. Esta información descriptiva puede ser tan precisa y significativa como los datos numéricos, pero no se puede trabajar con las mismas técnicas (Skiena, 2017).

Los datos categóricos a menudo se pueden representar numéricamente. Por ejemplo, podemos asignar un valor numérico para representar el género, como 0 para masculino y 1 para femenino. Sin embargo, surge un desafío cuando hay más de dos categorías y no existe un orden implícito entre ellas. Tomemos el ejemplo de los colores de cabello, donde asignaríamos valores distintos a cada tono, como 0 para canas, 1 para cabello rojo y 2 para cabello rubio. Sin embargo, debemos tener en cuenta que estos valores numéricos no tienen un significado más allá de identificar las categorías. No tendría sentido hablar de un "color de cabello máximo" o "color de cabello mínimo". Tampoco podemos realizar operaciones aritméticas con estos valores, ya que no tienen una interpretación

significativa. En teoría, tener más datos siempre es mejor que tener menos, porque siempre se puede desechar algunos mediante muestreo para obtener un conjunto más pequeño si es necesario. *Big data* es un fenómeno emocionante. Pero en la práctica, existen dificultades para trabajar con grandes conjuntos de datos. En general, las cosas se vuelven más difíciles una vez que el volumen es demasiado grande. Los desafíos de los grandes datos incluyen (Skiena, 2017):

El tiempo del ciclo de análisis se ralentiza conforme crece el tamaño de los datos: las operaciones computacionales en conjuntos de datos toman más tiempo a medida que aumenta su volumen. Las hojas de cálculo pequeñas brindan una respuesta instantánea, lo que le permite experimentar y jugar; pero las hojas de cálculo grandes pueden ser lentas y complicadas para trabajar, y los conjuntos de datos lo suficientemente grandes pueden tardar horas o días. Los algoritmos inteligentes pueden permitir que se hagan cosas asombrosas con *big data*, pero mantener pequeña cantidad de datos generalmente conduce a un análisis y una exploración más rápidos (Skiena, 2017).

Los grandes conjuntos de datos son complejos de visualizar: los gráficos con millones de puntos son imposibles de mostrar en pantallas de computadora o imágenes impresas, y mucho menos entender conceptualmente. ¿Cómo podemos esperar entender realmente algo que no se puede ver? Los modelos simples no requieren datos masivos para ajustarse o evaluarse: una tarea típica de ciencia de datos podría ser tomar una decisión (por ejemplo, si debo ofrecer un seguro de vida a este compañero) en función de una pequeña cantidad de variables: digamos edad, género, altura, peso y la presencia o ausencia de condiciones médicas existentes (Skiena, 2017).

Si se tienen los anteriores datos sobre un millón de personas con sus resultados de vida asociados, debería de poderse construir un buen modelo general de riesgo de cobertura. Probablemente no sería de mucha ayuda para construir un modelo sustancialmente mejor si se tuvieran datos de cientos de millones de personas. Los criterios de decisión sobre sólo unas pocas variables (como la edad, género y el estado civil) no pueden ser demasiado complejos y deben ser sólidos para una gran cantidad de solicitantes. Cualquier observación que sea tan sutil que requiera datos masivos para desentrañar resultará irrelevante para una gran empresa que se basa en el volumen. Los grandes datos a veces se denominan datos incorrectos. A menudo se recopilan como el subproducto de un sistema o procedimiento determinado, en lugar de recopilarse a propósito

para responder a su pregunta en cuestión. El resultado es que tal vez tengamos que hacer esfuerzos heroicos para darle sentido a algo simplemente porque lo tenemos (Skiena, 2017).

Considere el problema de conocer las preferencias de los votantes entre los candidatos presidenciales. El enfoque de *big data* podría analizar *feeds* masivos de Twitter o Facebook, interpretando pistas de sus opiniones en el texto. El enfoque de datos pequeños podría ser realizar una encuesta, preguntar a cientos de personas una pregunta específica y tabular los resultados. De esta manera podemos decir que: el conjunto de datos correcto es el directamente más relevante para las tareas en cuestión, no necesariamente el más grande (Skiena, 2017).

### 3.6.3 Clasificación y regresión

Surgen dos tipos de problemas repetidamente en las aplicaciones tradicionales de ciencia de datos y reconocimiento de patrones, los desafíos de la clasificación y la regresión. a) Clasificación: a menudo buscamos asignar una etiqueta a un elemento de un conjunto discreto de posibilidades. Problemas tales como predecir el ganador de una determinada competencia deportiva (¿equipo A o equipo B?) o decidir el género de una película dada (¿comedia, drama o animación?) son problemas de clasificación, ya que cada uno implica seleccionar una etiqueta de las opciones posibles; b) Regresión: otra tarea común es pronosticar una cantidad numérica dada. Predecir el peso de una persona o cuánta nieve tendremos este año es un problema de regresión, donde pronosticamos el valor futuro de una función numérica en términos de valores anteriores y otras características relevantes. Quizás la mejor manera de ver la distinción prevista es mirar una serie de problemas de la ciencia de datos y etiquetarlos (clasificarlos) como regresión o clasificación (Skiena, 2017).

Se emplean diversos métodos algorítmicos para resolver estos dos tipos de problemas, aunque las mismas preguntas a menudo se pueden abordar de cualquier manera (Skiena, 2017):

¿El precio de una acción en particular será más alto o más bajo mañana? (clasificación)

¿Cuál será el precio de una acción en particular mañana? (regresión)

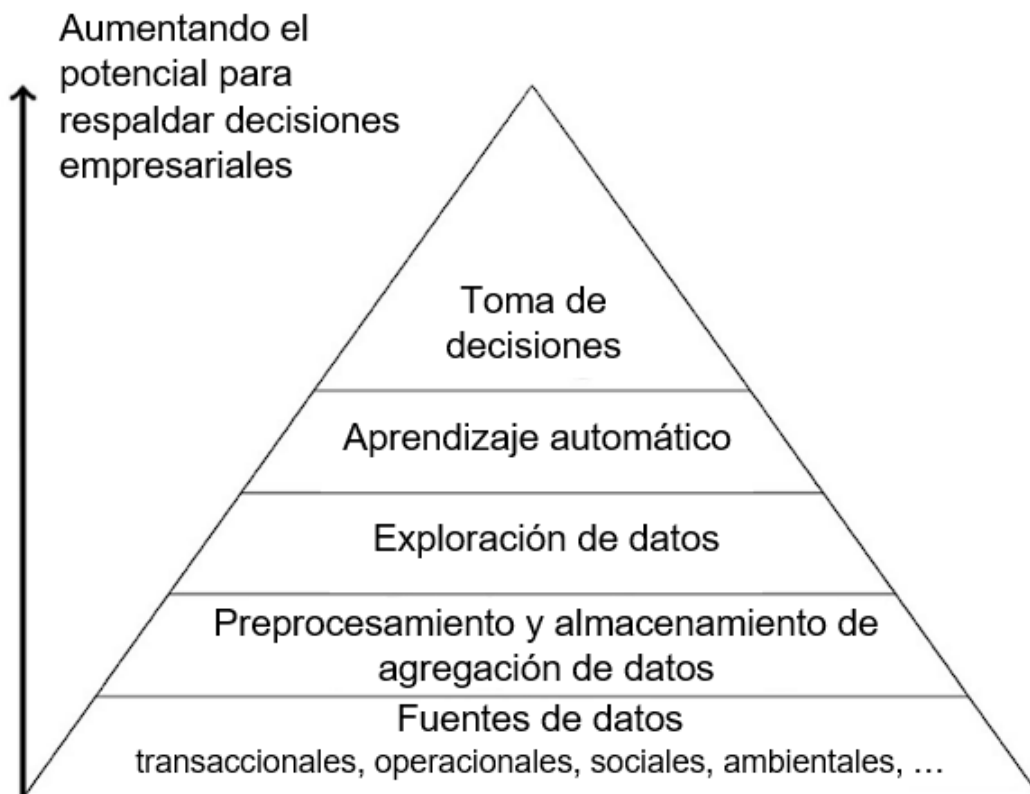
¿Considerando cierta persona, se tiene un alto o bajo riesgo para venderle una póliza de seguro? (clasificación)

¿Cuánto tiempo esperamos que viva cierta persona? (regresión)

### 3.7 Proceso CRISP-DM

Muchos individuos y compañías envían con regularidad propuestas sobre el procedimiento más adecuado que hay que seguir para ascender en la pirámide de ciencia de datos tal como se expone en la figura 16, el procedimiento CRISP-DM su acrónimo significa “*Cross Industry Standar Process for Data Mining*”. Una gran ventaja de CRISP-DM y el motivo más importante de su uso generalizado, es que está previsto para no depender de ningún software, distribuidor o método de análisis de datos (Kelleher & Tierney, 2018).

Figura 16 Pirámide de la ciencia de datos



Fuente: (elaborado con base en Kelleher & Tierney, 2018, p. 57)

CRISP-DM fue desarrollado inicialmente por una asociación de organizaciones formada por los principales proveedores de ciencia de datos, clientes finales,

firmas consultoras e investigadores. El CRISP-DM original fue financiado parcialmente por la Comisión Europea en el marco del programa ESPRIT, y el proceso fue presentado por primera ocasión en un seminario en 1999 (Kelleher & Tierney, 2018).

La metodología de CRISP-DM se compone de 6 fases, como se indica en la figura 1 (que se encuentra en la página 6): comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Los datos se encuentran en el punto central de todas las acciones de ciencia de datos, y razón por la cual la ilustración (figura 1) CRISP-DM tiene datos en su núcleo. Las flechas entre las fases muestran el sentido característico del proceso. El proceso es semiestructurado, lo que quiere decir que un investigador de datos no necesariamente pasa por estas seis fases linealmente. Según el resultado de una fase en concreto, un investigador de datos puede regresar a una de las fases anteriores, volver a hacer la fase en curso o avanzar al siguiente (Kelleher & Tierney, 2018).

En las dos primeras fases, “comprensión del negocio” y “comprensión de datos”, los objetivos del proyecto serán definidos entendiendo las necesidades del negocio y los datos disponibles para la organización. Las primeras fases de un proyecto, se deberá alternar entre centrarse en la organización y analizar cuáles son los datos disponibles. Esta relación mutua suele implicar la identificación un problema empresarial y, a continuación, examinar si los datos adecuados estén disponibles para elaborar una solución basada en datos para ese problema. Si se dispone de los datos, el proyecto puede continuar; de otra manera, habrá que identificar otro problema para resolverlo. A lo largo de esta fase de un proyecto, se dedicará mucho tiempo en reunirse con gente especializada en la empresa (como, por ejemplo: de compras/ventas, mercadotecnia, comercialización, operaciones) para entender sus problemas y con los *DBAs* (administradores de la base de datos) para lograr un entendimiento mutuo, de qué datos se encuentran en las fases previas, y por lo tanto repita la fase actual o seguir adelante con la siguiente (Kelleher & Tierney, 2018).

Una vez que se ha definido claramente un problema de negocios y con los datos disponibles, se continua con la próxima fase del CRISP-DM: “preparación de datos”. La fase de preparación de los datos se centrará en crear un conjunto de datos que puedan usarse para el análisis de los datos. Generalmente, para crear de este conjunto de datos es necesario integrar fuentes de datos de diferentes

bases de datos. Esta integración de datos es relativamente fácil si la empresa tiene un almacenamiento de datos. Después de crear el conjunto de datos, es necesario verificar y corregir su calidad. Los problemas habituales de la calidad de los datos implican valores atípicos y los valores que faltan. Tiene una gran importancia verificar la calidad de los datos debido a que los errores de datos pueden afectar gravemente al desempeño de los algoritmos de análisis de los datos (Kelleher & Tierney, 2018).

La próxima fase de CRISP-DM es la fase de “modelado”. Este es la fase en la cual los algoritmos automatizados son utilizados para obtener patrones beneficiosos de los datos y crear modelos que codifican estos patrones. El *machine learning* (aprendizaje automático) es el campo de la computación que se enfoca en el desarrollo de estos algoritmos. En la fase de modelado, los científicos de datos suelen entrenar muchos modelos distintos en un conjunto de datos utilizando diversos algoritmos de aprendizaje automático. Un modelo es entrenado en un conjunto de datos ejecutando un algoritmo de aprendizaje automático en el conjunto de datos para identificar pautas apropiadas en los datos y retornar un modelo que codifica estas pautas (Kelleher & Tierney, 2018).

En ciertos casos, un algoritmo de aprendizaje automático funciona mediante la adaptación de una estructura de modelo de patrones a un conjunto de datos mediante el establecimiento de parámetros de patrones en valores adecuados para ese conjunto de datos (por ejemplo: ajustar un modelo de red neuronal o regresión lineal a un conjunto de datos). En otras situaciones, un algoritmo de aprendizaje automático desarrolla un modelo fragmentario (por ejemplo, realizando un árbol de toma de decisiones de tal forma que se vaya creando un nodo a la vez que empieza en el nodo raíz del árbol). En la mayor parte de las investigaciones de ciencia de datos, se trata de un modelo que se genera mediante un algoritmo de aprendizaje automático que, en último término, es el programa que una empresa implementa para solucionar el problema que es abordado por la investigación de ciencia de datos (Kelleher & Tierney, 2018).

Cada modelo se entrena con un tipo distinto de algoritmo de aprendizaje automático, y cada algoritmo realiza una búsqueda de las distintas clases de pautas en los datos. En la presente fase del proyecto, por lo general no se sabe cuáles son los patrones más adecuados para la búsqueda de datos, entonces en este panorama es coherente experimentar con un repertorio de algoritmos

distintos y ver cuál algoritmo retorna los modelos más exactos al ejecutarse en el conjunto de datos (Kelleher & Tierney, 2018).

En la mayor parte de los proyectos de ciencia de datos, los resultados de las primeras pruebas del modelo pondrán al descubierto los problemas de datos. Estos errores de datos en ocasiones se descubren al investigar por qué el desempeño de un modelo es inferior al previsto o cuando se comprende que un modelo puede estar funcionando cuestionablemente bien. Alternativamente, un científico de datos puede examinar la estructura del modelo para determinar que el modelo se basa en cualidades inesperadas y examinar los datos para asegurarse de que esas cualidades sean correctas, de tal manera que se corrobora que se haya codificado correctamente los datos. En consecuencia, no es inusual que un proyecto atraviese por diversos recorridos de estas dos fases del proceso: modelado y preparación de datos; A modo de ejemplo, un científico de datos “X” y sus colaboradores expusieron que, a lo largo de un proyecto de ciencia de datos, restablecieron su conjunto de datos 12 veces en un período de siete semanas, y en la quinta semana, tras pasar por varias repeticiones de datos de limpieza y de preparar los datos, detectaron un gran error en los datos. El proyecto no habría tenido un resultado satisfactorio si este error no se hubiera reconocido y corregido (Kelleher & Tierney, 2018).

Las dos fases finales del proceso CRISP-DM, evaluación e implementación, se centran en la forma en que los modelos se adaptan en la organización y sus procesos intrínsecos del mismo. Las pruebas llevadas a cabo durante la fase de modelado se enfocan exclusivamente en la exactitud de los modelos para el conjunto de datos. La fase de evaluación consiste en la evaluación de modelos en el ámbito más extenso determinado por las necesidades de la organización. ¿El modelo se lleva a cabo con el objetivo o los objetivos del proceso? ¿Hay ciertas razones de negocio por el que un modelo no sea adecuado? Durante este proceso, también resulta útil realizar una inspección general del control de calidad de las operaciones del proyecto: ¿falta algo? ¿Se puede hacer una cosa mejor? Según la valoración de los modelos, la decisión más importante que se toma durante la fase de evaluación consiste en determinar si algún modelo debería ser implementado en la organización o si es necesario volver a realizar el proceso CRISP-DM para generar nuevos modelos que sean adecuados. Asumiendo que el procedimiento de evaluación apruebe uno o más modelos, entonces se avanza hacia la última fase del proceso: “implementación”. La fase

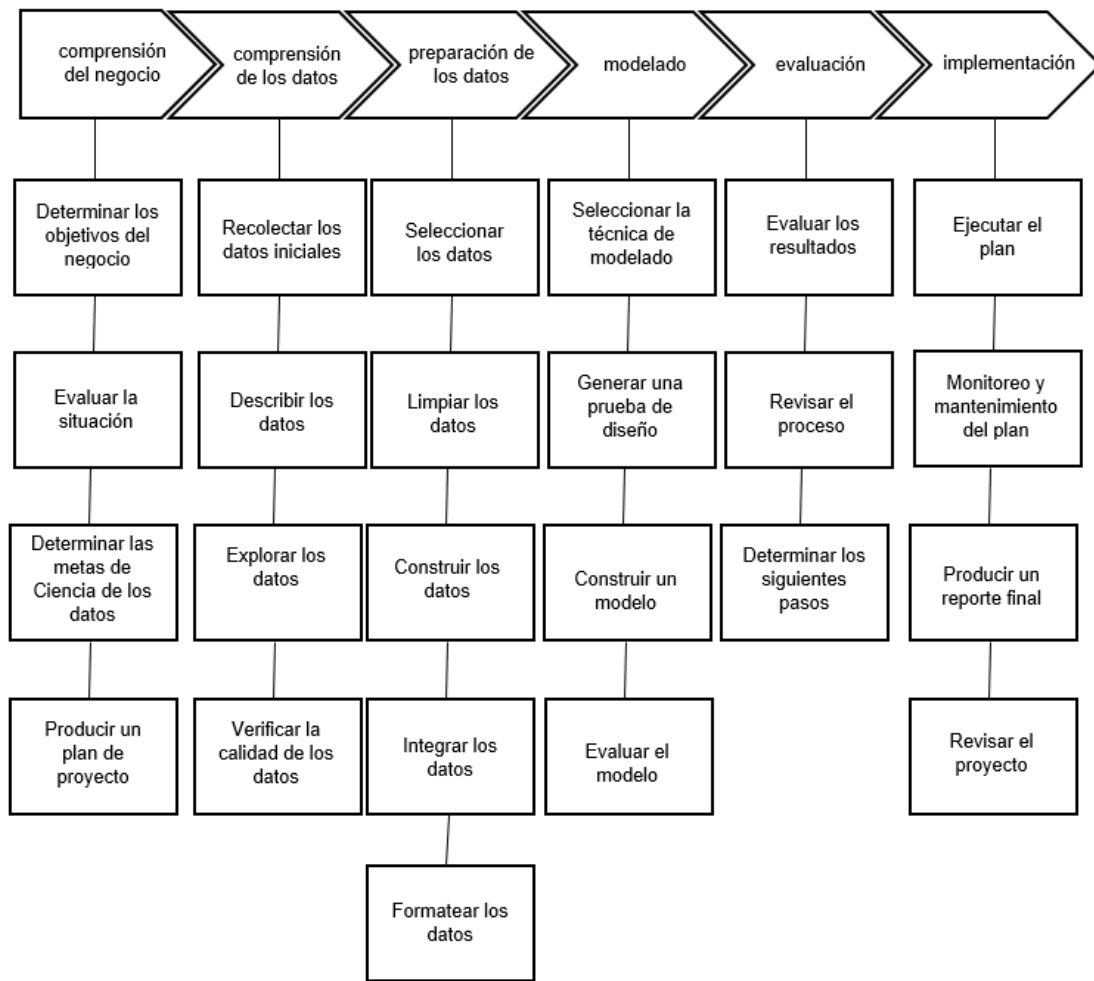
de implementación implica explorar cómo se puede implementar los modelos elegidos en un contexto de la empresa. Esto involucra la planificación de la integración de los modelos de la base técnica y los procesos de negocio de la organización. Los mejores modelos son aquellos se ajustan fácilmente a las prácticas recientes. Los modelos que se adaptan a las prácticas presentes tienen un conglomerado natural de usuarios que tienen una problemática bien definida que el modelo les apoya a solucionar. Otra cuestión de la implementación es poner en práctica un plan de revisión periódica del funcionamiento del modelo (Kelleher & Tierney, 2018).

El círculo externo en la figura 1 del proceso CRISP-DM, resalta la iteración del proceso en su conjunto. El carácter iterativo de los proyectos de ciencia de datos es tal vez el elemento de esos proyectos que la mayoría de las veces pasa desapercibida en los debates sobre la ciencia de datos. Una vez que se ha elaborado y puesto en práctica un modelo en un proyecto, debe revisarse periódicamente para asegurarse de que aún cumple con las necesidades comerciales y no ha quedado obsoleto. Los modelos basados en datos pueden volverse obsoletos por varias razones: los requisitos comerciales pueden haber cambiado; el proceso mediante el cual el modelo simula y ofrece información pudo haber cambiado, a modo de ejemplo: alteración en el comportamiento de los usuarios, modificaciones en los correos electrónicos no deseados, etc.; o también las fuentes de datos utilizados por el modelo pueden haberse modificado, como ejemplo: es posible que se haya actualizado un sensor que suministra información a un modelo, de tal forma que la nueva versión del sensor ofrece lecturas levemente distintas, lo que disminuye la precisión del modelo. La periodicidad de dicha comprobación dependerá de qué tan rápido evolucione el ambiente organizacional, así como los datos utilizados por el modelo. Es necesario monitorear constantemente para determinar el mejor tiempo para rehacer el proceso (Kelleher & Tierney, 2018).

Esto es lo que interpreta el círculo externo del proceso CRISP-DM mostrado en la figura 1. Por ejemplo, en función de los datos, la cuestión de negocio y el dominio, puede que deba pasar por este proceso iterativo cada año, cada mes, cada semana, o inclusive de manera diaria. En la figura 17 se presentan una visión general de las distintas fases del proceso del proyecto de ciencia de los datos y de las principales tareas que se llevan a cabo en cada fase (Kelleher & Tierney, 2018).



Figura 17 Las fases y tareas del proceso CRISP-DM



Fuente: (elaborado con base en Kelleher & Tierney, 2018,p 66)

## 3.8 Algoritmos Genéticos

### 3.8.1 Introducción

Se podría afirmar que la aparición de las computadoras electrónicas ha sido el avance más innovador en la historia de la ciencia y la tecnología. Esta revolución continua está aumentando significativamente nuestra habilidad para anticipar y manejar la naturaleza de maneras que apenas se imaginaban hace cincuenta años. La creación de software inteligente y posiblemente formas de vida a través de programas informáticos es lo que la mayoría de las personas considera como el logro más importante de esta revolución. Desde los inicios de la era de las computadoras, se han tenido como objetivos la creación de inteligencia artificial

y vida artificial. Los pioneros de la informática como Alan Turing, John Von Neumann y Norbert Wiener, entre otros, tenían una visión motivada por dotar a los programas informáticos de inteligencia, autorreplicación real, capacidad de adaptación para el aprendizaje y control de su entorno (Mitchell, 1998).

Los primeros pioneros de la informática se mostraron igualmente interesados en la biología y la psicología, además de la electrónica, y tomaron como inspiración los sistemas naturales para lograr sus objetivos. Por lo tanto, no es sorprendente que desde los primeros días de las computadoras se utilizaran no sólo para cálculos de misiles y códigos militares, sino también para modelar el cerebro, imitar el aprendizaje humano y simular la evolución biológica. A pesar de que estas actividades informáticas motivadas biológicamente han tenido altibajos a lo largo de los años, desde principios de la década de 1980 han experimentado un resurgimiento en la comunidad de investigación informática. El primer enfoque se ha convertido en el campo de las redes neuronales, el segundo en el aprendizaje automático y el tercero en lo que ahora se llama "computación evolutiva", donde los algoritmos genéticos son uno de sus ejemplos más destacados. En las décadas de los años 1950 y 1960, hubo varios científicos informáticos que investigaron de manera independiente los sistemas evolutivos, con la intención de usar la evolución como una herramienta para optimizar problemas de ingeniería. La idea en todos estos sistemas era crear una población de soluciones candidatas para un problema determinado, utilizando operadores que estuvieran basados en la variación genética natural y la selección natural (Mitchell, 1998).

En la década de 1960, Rechenberg introdujo las "estrategias de evolución" como método para optimizar parámetros de valor real para dispositivos como superficies aerodinámicas. Schwefel continuó desarrollando esta idea en los años de 1970, y desde entonces, las estrategias de evolución se han convertido en un área activa de investigación. Aunque la mayoría de dicha investigación se ha realizado independientemente del campo de los algoritmos genéticos, en los últimos tiempos ambas comunidades han comenzado a interactuar. Fogel, Owens y Walsh desarrollaron la "programación evolutiva" en 1966, una técnica que representa soluciones candidatas a tareas dadas como máquinas de estado finito que evolucionan mediante mutaciones aleatorias en sus diagramas de transición de estado y seleccionando la más adecuada. Esta técnica sigue siendo un área de investigación activa y se ha formulado de manera más amplia. En

conjunto, las estrategias de evolución, la programación evolutiva y los algoritmos genéticos son los pilares fundamentales del campo de la computación evolutiva. (Mitchell, 1998).

En las décadas de 1950 y 1960, varios investigadores, como Box, Friedman, Bledsoe, Bremermann, Toombs y Baricelli, desarrollaron algoritmos inspirados en la evolución para la optimización y el aprendizaje automático. Sin embargo, su trabajo ha recibido poca atención y no ha tenido el mismo éxito que las estrategias más modernas, como la programación evolutiva y los algoritmos genéticos. Además, algunos biólogos evolutivos también utilizaron la computación para simular la evolución con fines experimentales. En resumen, la computación evolutiva ya estaba presente en los primeros días de la computadora electrónica. (Mitchell, 1998).

En los años de 1960 y 1970, John Holland y su equipo de la Universidad de Michigan inventaron los algoritmos genéticos (AG). A diferencia de otras estrategias evolutivas y de programación, el objetivo original de Holland era estudiar la adaptación en la naturaleza y cómo se podía aplicar a los sistemas informáticos. El libro de Holland de 1975 "Adaptación en sistemas naturales y artificiales" presentó los AG como una forma de imitar la evolución biológica y estableció un marco teórico para la adaptación mediante el uso de estos algoritmos (Mitchell, 1998).

El método de Holland, conocido como Algoritmo Genético, implica la transformación de una población de "cromosomas" que se componen de "genes" o instancias de un "alelo" específico. Estos cromosomas se someten a un proceso de selección natural y se utilizan operadores de cruce basados en la genética, mutación e inversión. El operador de selección determina qué cromosomas tienen permitido reproducirse y los cromosomas más aptos tienen una mayor probabilidad de producir descendencia. El cruce, que imita la recombinación biológica, intercambia subpartes de dos cromosomas, mientras que la mutación cambia al azar los valores de los alelos en algunas ubicaciones del cromosoma. Por último, la inversión invierte el orden de una sección contigua del cromosoma, lo que reorganiza el orden en que se disponen los genes. En la literatura sobre Algoritmos Genéticos, "cruce" y "recombinación" se usan indistintamente (Mitchell, 1998).

Holland fue pionero en el desarrollo del algoritmo genético basado en poblaciones, que utiliza operadores de cruce, inversión y mutación para crear una nueva generación a partir de una población anterior. En comparación con otras estrategias evolutivas como las de Rechenberg y la programación evolutiva de Fogel, Owens y Walsh, el algoritmo de Holland incorporó la idea de una población grande y diversa, así como la selección natural y el cruce inspirado en la genética. Holland también hizo un esfuerzo para establecer una base teórica sólida para la evolución computacional, utilizando la noción de "esquemas" que se convirtió en la base teórica para la mayoría de los trabajos posteriores sobre algoritmos genéticos. Además, a diferencia de otros métodos de optimización, los algoritmos genéticos no requieren un conocimiento profundo del problema que se está tratando de resolver, sino que trabajan con códigos que representan los parámetros del problema (Mitchell, 1998).

### 3.8.2 Evolución Biológica

Para los investigadores en computación evolutiva, la evolución proporciona una inspiración valiosa para resolver algunos de los problemas más complejos en diversos campos. Muchos problemas en computación implican buscar a través de una gran cantidad de posibles soluciones. Por ejemplo, el problema de la ingeniería computacional de proteínas, que busca una secuencia de aminoácidos que produzca una proteína con propiedades específicas, y el problema de encontrar un conjunto de reglas o ecuaciones que puedan predecir los cambios en los mercados financieros, como el mercado de divisas. En estos casos, los mecanismos de la evolución parecen ser muy adecuados para ayudar a encontrar soluciones efectivas. Tales problemas de búsqueda pueden ser optimizadas mediante un uso adecuado del paralelismo, lo que permite explorar múltiples opciones de forma eficiente y simultánea. Por ejemplo, en la búsqueda de proteínas con propiedades específicas, en lugar de evaluar una secuencia de aminoácidos a la vez, sería mucho más rápido evaluar muchas simultáneamente. Para esto, se requiere tanto paralelismo computacional, que implica el uso de muchos procesadores para evaluar las secuencias simultáneamente, como una estrategia inteligente para seleccionar el siguiente conjunto de secuencias a evaluar (Mitchell, 1998).

La evolución biológica se presenta como una fuente inspiradora para resolver problemas. La evolución es un proceso que consiste en explorar un gran número de posibles soluciones, en biología estas posibilidades son las secuencias genéticas y las soluciones son organismos que pueden sobrevivir y reproducirse en su entorno. La evolución también puede ser vista como un método para generar soluciones innovadoras a problemas complejos, como lo hizo el sistema inmunológico de los mamíferos para luchar contra los gérmenes. Estos mecanismos pueden inspirar métodos de búsqueda computacional. La aptitud de un organismo biológico depende de muchos factores, como la resistencia a características físicas del entorno y la capacidad de competir y cooperar con otros organismos. Los criterios de aptitud cambian continuamente a medida que evolucionan las criaturas, por lo que la evolución busca un conjunto de posibilidades en constante cambio, lo que es útil para programas informáticos adaptativos. La evolución es un método de búsqueda masivamente paralelo, probando y cambiando millones de especies en paralelo, en lugar de trabajar en una especie a la vez. Finalmente, las "reglas" de la evolución son simples: las especies evolucionan a través de variaciones aleatorias seguidas de selección natural, donde los organismos más aptos sobreviven y se reproducen, propagando así su material genético a las generaciones futuras. Estas simples reglas son responsables de la gran complejidad y variedad que vemos en la biosfera (Mitchell, 1998).

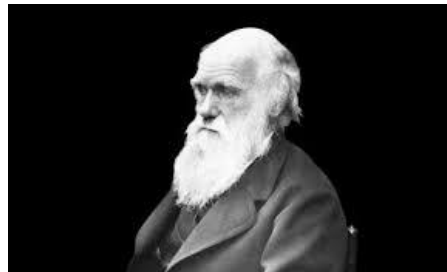
La totalidad de los seres vivos están compuestos por células, y en cada una de ellas se encuentra un conjunto de uno o más cromosomas, que actúan como un modelo para el organismo en cuestión. Cada cromosoma se puede dividir en secciones llamadas genes, que contienen información codificada para producir una proteína específica. De manera simplificada, se puede pensar en un gen que determina un rasgo particular, como el color de ojos, y los diferentes "ajustes" posibles para ese rasgo se llaman alelos. Cada gen tiene una ubicación específica, también conocida como locus, en el cromosoma correspondiente (Mitchell, 1998).

En los organismos, los cromosomas son estructuras que contienen material genético y muchos tienen varios cromosomas en cada célula. La colección completa de todos los cromosomas en un organismo se llama genoma. El término genotipo se utiliza para describir el conjunto específico de genes que se encuentran en el genoma de un organismo, y dos individuos con genomas

idénticos se dice que tienen el mismo genotipo. Durante el desarrollo fetal y posterior, el genotipo da lugar al fenotipo del organismo, que se refiere a sus características físicas y mentales, como el color de ojos, altura, tamaño del cerebro e inteligencia. (Mitchell, 1998).

Los seres vivos que presentan sus cromosomas en pares son conocidos como diploides, mientras que aquellos que no tienen pares cromosómicos se les denomina haploides. La evolución de las poblaciones en la naturaleza, a lo largo del tiempo, se rige por los principios de la selección natural y la supervivencia de los más aptos, teorías propuestas por Charles R. Darwin (figura 18) (Mitchell, 1998).

*Figura 18 Charles Darwin*



Fuente: <https://i2.wp.com/historia-biografia.com/wp-content/uploads/2020/02/Charles-Darwin.jpg?w=1000&ssl=1>

En la naturaleza, los organismos compiten por recursos como comida, agua y refugio. Incluso dentro de la misma especie, compiten por tener más descendencia. Sin embargo, los individuos menos capacitados tendrán menos descendencia, lo que significa que los genes de los individuos más aptos se transmitirán a más individuos en las siguientes generaciones. A veces, la combinación de características beneficiosas de diferentes antepasados puede dar lugar a descendientes "superiores" cuya adaptación al entorno es mucho mejor que la de cualquiera de sus ancestros. De esta manera, las especies evolucionan y adquieren características cada vez más adecuadas a su entorno (Sivanandam & Deepa, 1998).

En los algoritmos genéticos, los cromosomas son como un conjunto de instrucciones para resolver un problema, codificados a menudo como una cadena de caracteres. Los "genes" son partes individuales de esas instrucciones que codifican una acción específica para la solución propuesta. Por ejemplo, en el contexto de la creación de una imagen digital, los genes podrían ser las

instrucciones para definir el color y la ubicación de cada píxel. Cada "alelo" en una cadena de caracteres puede ser una letra o número diferente, y en alfabetos más grandes, puede haber aún más opciones para cada "locus". La "cruza" implica combinar partes de dos soluciones propuestas para crear una nueva, mientras que la "mutación" implica cambiar una parte de una solución existente al azar, para ver si mejora el resultado (Mitchell, 1998).

En general, la mayoría de las aplicaciones de los algoritmos genéticos utilizan individuos haploides, especialmente aquellos que son monos cromosómicos. En un algoritmo genético que utiliza cadenas de bits, el genotipo de un individuo se refiere a la configuración de bits en su cromosoma. Aunque en algunos casos no existe la noción de "fenotipo" en el contexto de los algoritmos genéticos, algunos investigadores han estado experimentando con algoritmos genéticos que incluyen tanto un nivel genotípico como fenotípico. Un ejemplo de esto es la codificación de una red neuronal como cadena de bits y la propia red neuronal como fenotipo (Mitchell, 1998).

### 3.8.3 Métodos de representación en algoritmos genéticos

El método utilizado para representar la optimización de parámetros en algoritmos genéticos tiene un gran impacto en su desempeño. Diferentes esquemas de representación pueden conducir a diferentes niveles de precisión y tiempo de cómputo. Es importante elegir cuidadosamente el método de representación adecuado para cada problema para obtener los mejores resultados posibles. Para la optimización numérica, se utilizan comúnmente dos métodos de representación (Affenzeller, Winkler, Wagner, & Beham, 2009):

- El enfoque preferido para la representación de datos en algoritmos genéticos es el uso de cadenas binarias. Este método es popular porque el alfabeto binario permite la mayor cantidad de combinaciones posibles en comparación con otras técnicas de codificación. Algunos esquemas de codificación binarios comunes incluyen el código uniforme y el código de escala de Gray.
- Otra opción de representación es mediante un vector de números enteros o reales, en donde cada valor representa un solo parámetro a optimizar. En el caso de utilizar la representación binaria, es crucial determinar la

cantidad adecuada de bits a utilizar en el código de cada parámetro para maximizar el rendimiento del sistema. Utilizar muy pocos o demasiados bits puede impactar negativamente en la precisión de las soluciones, por lo que es importante encontrar el número óptimo de bits para cada parámetro. (Affenzeller, Winkler, Wagner, & Beham, 2009).

#### 3.8.4 Creación de la población inicial

En los algoritmos genéticos, primero se genera una población inicial de posibles soluciones (también llamados "individuos" o "cromosomas") de manera aleatoria o mediante una heurística. En cada paso de iteración, también conocido como "generación", se evalúan los individuos de la población actual y se les asigna una puntuación de aptitud. Para crear una nueva población, se seleccionan primero los individuos (generalmente con una probabilidad proporcional a su puntuación de aptitud) y luego se generan posibles descendientes que formarán la siguiente generación de individuos. Esto asegura que los individuos más aptos tengan una mayor probabilidad de ser seleccionados y de transmitir sus características a las generaciones futuras. Para generar nuevos posibles candidatos de solución, los algoritmos genéticos utilizan dos operadores: cruce y mutación (Affenzeller, Winkler, Wagner, & Beham, 2009).

- Una de las principales operaciones genéticas es el cruce, que consiste en tomar dos individuos, conocidos como padres, y crear uno o dos nuevos individuos, conocidos como descendencia, combinando partes de ambos padres. De manera básica, este operador intercambia secciones de la cadena genética de dos padres seleccionando un punto de cruce aleatorio para separar las subcadenas antes y después del mismo (Affenzeller, Winkler, Wagner, & Beham, 2009).
- El segundo operador genético, conocido como mutación, es una modificación aleatoria que tiene como objetivo evitar que el algoritmo converja prematuramente y explorar nuevas soluciones en el espacio de búsqueda. En el caso de las cadenas binarias, la mutación se logra al intercambiar bits al azar dentro de una cadena, y se controla mediante una tasa de mutación que determina la probabilidad de que se produzca una mutación en cada bit (Affenzeller, Winkler, Wagner, & Beham, 2009).



Hay dos formas de formar esta población inicial. El primero es utilizar una solución aleatoria generada mediante la creación de un generador de números aleatorios. Este método es el preferido para problemas sin conocimiento previo o para evaluar el desempeño de un algoritmo. El segundo método se basa en la información previa sobre un problema de optimización específico. A partir de esta información, se identifican un conjunto de requisitos y se recopilan soluciones que cumplan con estos requisitos para formar un grupo inicial. En este enfoque, el algoritmo genético comienza a optimizar a partir de soluciones aproximadas previamente conocidas, lo que permite una convergencia más rápida hacia la solución óptima en comparación con los métodos anteriores (Sivanandam & Deepa, 1998).

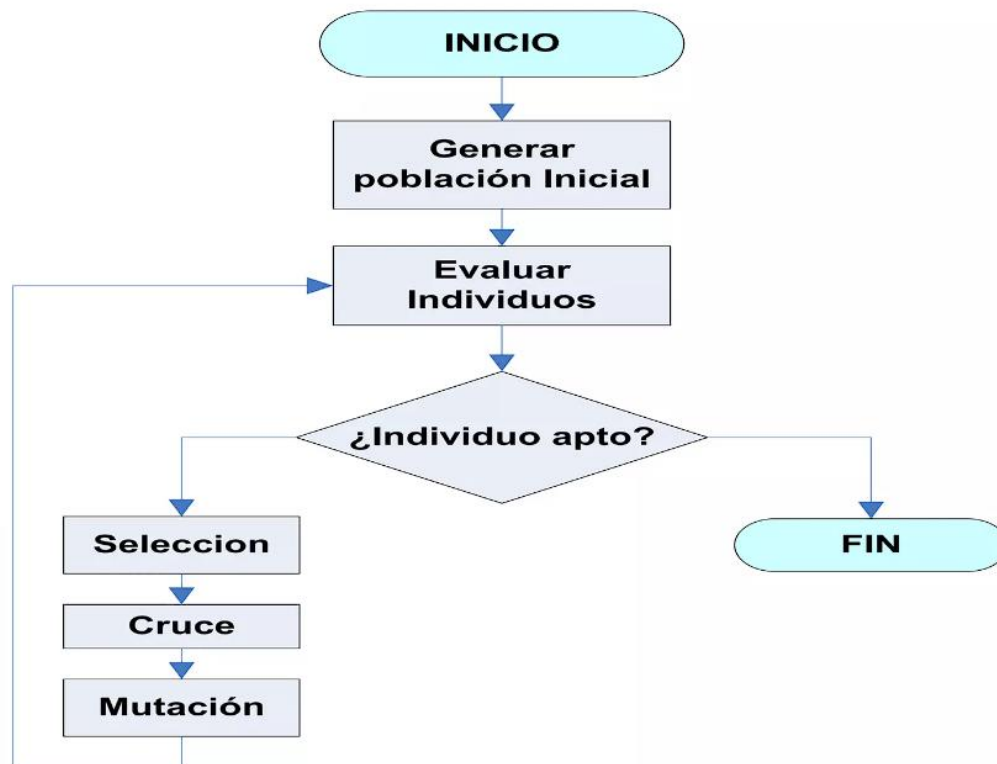
Por ejemplo, supongamos que estamos tratando de encontrar la mejor combinación de ingredientes para hacer una galleta perfecta. En una población inicial, podríamos tener varios candidatos de solución (cromosomas) que representan diferentes combinaciones de ingredientes y proporciones. Durante cada generación, evaluamos el sabor, textura y apariencia de cada galleta y le asignamos una puntuación de aptitud. Luego, seleccionamos los candidatos más aptos (con una mayor puntuación de aptitud) y producimos descendientes que combinan los ingredientes de los padres mediante el cruce. También introducimos pequeñas mutaciones aleatorias en algunos de los ingredientes para generar más variedad en la población de candidatos. Continuamos repitiendo este proceso de selección, cruce y mutación en cada generación hasta que encontremos la mejor combinación de ingredientes que resulte en la galleta perfecta.

### 3.8.5 Operadores Genéticos

La figura 19 presenta un esquema de un algoritmo genético básico en forma de diagrama de flujo. Este algoritmo utiliza tres operadores genéticos principales: selección, cruzamiento y mutación. También se puede utilizar un cuarto operador de reproducción, conocido como el operador de inversión. Además, a veces se aplican algunos operadores basados en la evolución natural, y se pueden encontrar muchas versiones de estos operadores en la literatura de algoritmos

genéticos. No es obligatorio emplear todos los operadores mencionados en un algoritmo para optimizar el problema, ya que cada función es autónoma y no depende de las demás. Las opciones o planes del transportista dependen de la pregunta y de la representación oficial del plan utilizado. Por ejemplo, los operadores diseñados para cadenas binarias no se pueden usar directamente en cadenas codificadas con números enteros o reales

Figura 19 Diagrama de flujo de para un algoritmo genético sencillo



Fuente: <https://image.slidesharecdn.com/presentaciontesis-120712182305-phpapp02/85/modelos-de-algoritmo-genetico-36-320.jpg?cb=1342117615>

#### A) La selección

Para permitir la convergencia hacia soluciones óptimas, se deben seleccionar las mejores soluciones de descendencia para que sean padres en la nueva población de padres. Se genera un excedente de soluciones descendientes y se seleccionan las mejores para lograr un progreso hacia el óptimo. Este proceso de selección se basa en los valores de aptitud de la población. En el caso de problemas de minimización se prefieren valores bajos de aptitud y viceversa en

el caso de problemas de maximización. Los problemas de minimización pueden transformarse fácilmente en problemas de maximización con negación. Por supuesto, esto también funciona para transformar problemas de maximización en problemas de minimización (Kramer, 2017).

Los operadores de selección elitistas seleccionan las mejores soluciones de las soluciones descendientes como padres. La selección Plus elige las mejores soluciones  $\mu$  de las soluciones descendientes de  $\lambda$ . Luego, la selección Plus selecciona también los  $\mu$  padres antiguos que llevaron a su creación (Kramer, 2017).

Muchos algoritmos de selección se basan en la aleatoriedad. La rueda de la ruleta, también conocida como selección proporcional de aptitud, selecciona soluciones parentales al azar con una distribución uniforme. La probabilidad de que una solución sea seleccionada está determinada por su aptitud. Por este motivo, la aptitud relativa de las soluciones se normaliza con la suma de todos los valores de aptitud en una población, generalmente por división. Esta fracción de aptitud puede entenderse como la probabilidad de que una solución sea seleccionada. La ventaja de los operadores de selección proporcionales a la aptitud es que cada solución tiene una probabilidad positiva de ser seleccionada (Kramer, 2017).

En el caso de la selección de padres, es importante destacar que pueden olvidarse buenos padres. Además, la aleatoriedad de la selección proporcional de aptitud permite olvidar las mejores soluciones. Aunque esto puede sonar contraproducente para el proceso de optimización al principio, el olvido puede ser una estrategia razonable para superar los óptimos locales. Otro operador de selección famoso es la selección de torneos, donde se selecciona aleatoriamente un conjunto de soluciones y dentro de este subconjunto de competencia, las mejores soluciones finalmente se seleccionan como nuevos padres. El segundo paso se puede implementar con la selección proporcional de *fitness* como ejemplo típico. La selección de torneos ofrece una probabilidad positiva para que cada solución sobreviva, incluso si tiene peores valores de aptitud que otras soluciones (Kramer, 2017).

Cuando se utiliza la selección como mecanismo para elegir a los padres de la nueva generación, se denomina selección de supervivencia. El operador de selección determina qué soluciones sobreviven y qué soluciones mueren. Esta

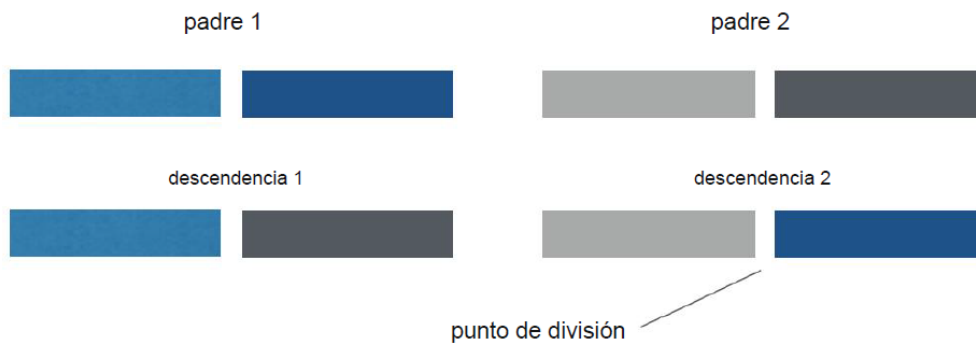
forma de implementación sigue directamente el principio de supervivencia del más apto de Darwin. Pero los operadores de selección introducidos también se pueden emplear para la selección de apareamiento que forma parte de los operadores de cruce. La selección de apareamiento es una estrategia para decidir qué padres participan en el proceso de cruzamiento. Tiene sentido considerar otros criterios para la selección de apareamiento que para la selección de supervivencia (Kramer, 2017).

## B) Cruzamiento

Cruzamiento es un operador que permite la combinación del material genético de dos o más soluciones. La mayoría de las especies en la naturaleza tienen la contribución genética de dos padres. Algunas excepciones no conocen sexos diferentes y por lo tanto sólo tienen uno de los padres. En los algoritmos genéticos podemos incluso extender los operadores de cruce a más de dos padres. El primer paso en la naturaleza es la selección de una pareja potencial. Muchas especies gastan muchos recursos en procesos de selección, pero también en la elección de un socio potencial y en estrategias para atraer socios. En particular, los machos gastan muchos recursos en impresionar a las hembras. Después de la selección de un compañero, el emparejamiento es el siguiente paso natural. Desde una perspectiva biológica, dos socios de la misma especie combinan su material genético y lo heredan a su descendencia (Kramer, 2017).

Los operadores de cruce en algoritmos genéticos implementan un mecanismo que mezcla el material genético de los padres. Uno famoso para la representación de cadenas de *bits* es el cruce de  $n$  puntos. Divide dos soluciones en  $n$  posiciones y las ensambla alternativamente en una nueva tal como se muestra en la figura 20. Por ejemplo, si “0010110010” es el primer padre y “1111010111” es el segundo, el cruce de un punto elegiría aleatoriamente una posición, supongamos 4, y generaría las dos soluciones candidatas descendientes “0010-010111” y “1111-110010”. La motivación para tal operador es que ambas cadenas pueden representar partes exitosas de soluciones que, cuando se combinan, incluso superan a sus padres. Este operador se puede extender fácilmente a más puntos, donde las soluciones se dividen y se vuelven a ensamblar alternativamente (Kramer, 2017).

*Figura 20 Ilustración de cruce de un punto que divide el genoma de dos soluciones en un punto arbitrario (aquí en el medio) y las vuelve a ensamblar para obtener dos soluciones nuevas*



Fuente: (Kramer, 2017)

Para representaciones continuas, los operadores de cruce están orientados a operaciones numéricas. El cruce aritmético, también conocido como cruce intermedio, calcula la media aritmética de todas las soluciones parentales por componentes. Por ejemplo, para los dos padres (1, 4, 2) y (3, 2, 3) la solución de descendencia es (2, 3, 2.5). Este operador de cruce se puede extender a más de dos padres. El cruce dominante elige sucesivamente cada componente de una de las soluciones parentales. El cruce uniforme utiliza una relación de mezcla fija como 0,5 para elegir aleatoriamente un *bit* de cualquiera de los padres. Surge la pregunta, cuáles de las soluciones parentales participan en la generación de nuevas soluciones. Muchos algoritmos genéticos simplifican este paso y eligen aleatoriamente los padres para la operación de cruce con distribución uniforme (Kramer, 2017).

### C) La mutación

Los operadores de mutación cambian una solución perturbándola. La mutación se basa en cambios aleatorios. La fuerza de esta perturbación se llama tasa de mutación. En espacios de solución continua, la tasa de mutación también se conoce como tamaño de paso (Kramer, 2017).

Hay tres requisitos principales para los operadores de mutación. La primera condición es la accesibilidad. Cada punto en el espacio de soluciones debe ser accesible desde un punto arbitrario en el espacio de soluciones. Un ejemplo que puede complicar el cumplimiento de esta condición es la existencia de

restricciones que reducen todo el espacio de soluciones a un subconjunto factible. Debe haber una posibilidad mínima de llegar a cada parte del espacio de la solución. De lo contrario, la probabilidad de que se pueda encontrar el óptimo no es positiva. No todos los operadores de mutación pueden garantizar esta condición, por ejemplo, los enfoques de decodificación tienen dificultades para cubrir todo el espacio de la solución (Kramer, 2017).

El segundo buen principio de diseño de los operadores de mutación es la imparcialidad. El operador de mutación no debe inducir una desviación de la búsqueda hacia una dirección particular, al menos en espacios de solución no restringidos sin mesetas. En el caso de espacios de solución restringidos, el sesgo puede ser ventajoso. También la idea de búsqueda de novedades que trata de buscar en partes del espacio de soluciones que aún no han sido exploradas, induce un sesgo en el operador de mutación (Kramer, 2017).

El tercer principio de diseño de los operadores de mutación es la escalabilidad. Cada operador de mutación debe ofrecer el grado de libertad que su fuerza sea adaptable. Esto suele ser posible para los operadores de mutación que se basan en una distribución de probabilidad (Kramer, 2017).

Por ejemplo, para la mutación gaussiana que se basa en la distribución gaussiana, la desviación estándar puede escalar las muestras extraídas aleatoriamente en todo el espacio de la solución. La implementación de los operadores de mutación depende de la representación empleada. Para cadenas de *bits*, generalmente se usa la mutación de cambio de *bits*. La mutación de cambio de *bit* cambia un *bit* cero a un *bit* y viceversa con una probabilidad definida, que desempeña el papel de la tasa de mutación. Suele elegirse en función de la duración de la representación (Kramer, 2017).

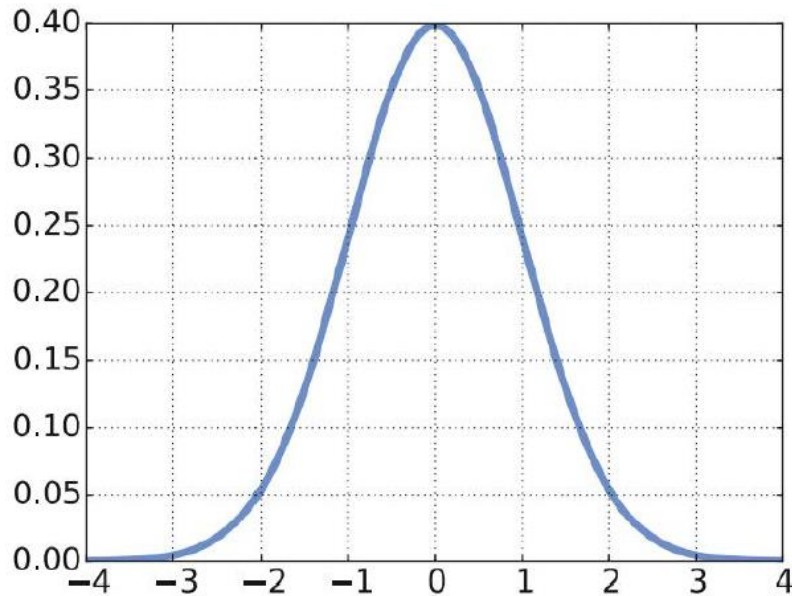
Si  $N$  es la longitud de la cadena de *bits*, cada *bit* se invierte con una tasa de mutación de  $1/N$ . Si la representación es una lista o cadena de elementos arbitrarios, la mutación elige aleatoriamente un reemplazo para cada elemento. Este operador de mutación se conoce como reinicio aleatorio (Kramer, 2017).

Sea  $[5, 7, -3, 2]$  el cromosoma con valores enteros que provienen del intervalo  $[-10, 10]$ , luego el reinicio aleatorio decide para cada componente, si se reemplaza. Si se reemplaza el componente, elige aleatoriamente un nuevo valor del intervalo. Por ejemplo, el resultado puede ser  $[8, -2, -5, 6]$  (Kramer, 2017).

Para representaciones continuas, la mutación gaussiana es el operador más popular. La mayoría de los procesos en la naturaleza siguen una distribución Gaussiana tal como se muestra en la figura 21. Esta es una suposición razonable para la distribución de soluciones exitosas (Kramer, 2017).

$$x' = x + \sigma \cdot N(0, 1)$$

*Figura 21 La distribución gaussiana es la base del operador de mutación gaussiana que añade ruido a cada componente del cromosoma*



Fuente: (Kramer, 2017)

Un vector de ruido gaussiano se agrega a un vector de solución continua. Si  $x$  es la solución descendiente generada mediante cruce, la mutación gaussiana utiliza  $N(0, 1)$  para representar un vector de ruido con distribución gaussiana como base. La variable  $\sigma$  se refiere a la tasa de mutación que escala la intensidad del ruido añadido. La distribución gaussiana tiene su punto máximo en el origen. Por lo tanto, con mayor probabilidad, la solución experimentará cambios mínimos o incluso nulos. La mutación gaussiana es un excelente ejemplo de un operador de mutación que satisface todos los criterios mencionados. Además,  $\sigma$  es adaptable según sea necesario. Además, con un valor escalable de  $\sigma$ , todas las regiones en espacios de solución continuos se vuelven accesibles. Debido a la simetría de la distribución gaussiana, no tiene preferencia por ninguna dirección y, por lo tanto, no induce sesgos en la búsqueda (Kramer, 2017).

### 3.8.6 Parámetros de control

La elección de los parámetros adecuados es fundamental para el éxito de los algoritmos genéticos. Los parámetros estáticos que se mantienen constantes durante la ejecución del algoritmo genético se pueden ajustar de antemano. Las estrategias de muestreo como el muestreo de hipercubo latino y la búsqueda en cuadrícula se aplican a menudo para ajustar los parámetros. Los métodos estadísticos pueden apoyar el proceso de ajuste. Algunos parámetros deben controlarse durante la ejecución para una mejora significativa de la búsqueda (Kramer, 2017).

Además de la estructura y los operadores utilizados en los algoritmos genéticos simples, existen ciertos parámetros de control que son críticos para su rendimiento. Entre éstos se incluyen el tamaño de la población, el número de individuos en la población, así como la tasa de cruce y la tasa de mutación proporcionada. Estos parámetros deben ser cuidadosamente ajustados para asegurar que el algoritmo converja a una solución óptima de manera eficiente (Kramer, 2017).

Un parámetro importante en los algoritmos genéticos es el tamaño de la población, que debe ser lo suficientemente grande para manejar múltiples soluciones y aumentar el tiempo de procesamiento. Sin embargo, se ha encontrado que el uso de una población más grande aumenta la probabilidad de encontrar la solución óptima en comparación con una población pequeña, aunque puede llevar más tiempo. La frecuencia de cruce es otra consideración importante, ya que puede ayudar a optimizar la salida al encontrar regiones prometedoras. Pero, si la frecuencia es demasiado baja, la convergencia a la solución puede ser más lenta, mientras que, si es demasiado alta, puede haber una saturación en torno a la solución (Kramer, 2017).

En cuanto a la operación de mutación, está controlada por la tasa de mutación, la cual puede introducir un alto grado de diversidad en la población si es alta y provocar inestabilidad. Por otro lado, una tasa de mutación baja puede dificultar que los algoritmos genéticos encuentren una solución globalmente óptima.



### 3.8.7 Función de evaluación de aptitud

En los algoritmos genéticos, la función de evaluación de aptitud es importante para medir la eficacia de una solución en la resolución de un problema. Se asigna un valor numérico a cada solución en la población, lo que influye en su probabilidad de ser seleccionado como padre para la próxima generación. Para que la función de evaluación sea útil, es esencial que se diseñe adecuadamente para reflejar el objetivo del problema y proporcionar una métrica clara para evaluar la calidad de las soluciones. Además, es importante que la evaluación de la función de aptitud para cada individuo se realice de manera rápida, de modo que el procesamiento sea eficiente. Si la evaluación es lenta, existen técnicas como la computación paralela y distribuida, la evaluación aproximada o la evaluación de solamente los elementos que han cambiado, que pueden ser utilizadas para mejorar el rendimiento (Affenzeller, Winkler, Wagner, & Beham, 2009).

La unidad de evaluación de aptitud es la conexión entre el algoritmo genético y el problema de optimización. En vez de utilizar información directa sobre la estructura del problema, los algoritmos genéticos mejoran la calidad de las soluciones evaluándolas con la información que produce la unidad de evaluación. El problema de optimización se diseña para cumplir con los requisitos funcionales especificados por el diseñador y lograr una estructura de desempeño que cumpla la función deseada dentro de las restricciones establecidas. La calidad de una solución propuesta generalmente depende del resultado de otra solución que haya cumplido con la función deseada y las restricciones establecidas. En el caso de un algoritmo genético, el cálculo de la calidad de la solución debe ser automático, y el desafío consiste en encontrar un procedimiento adecuado que pueda calcularla (Affenzeller, Winkler, Wagner, & Beham, 2009).

La evaluación de la aptitud puede ser realizada de forma manual o mediante un programa según la complejidad del problema de optimización. Si el problema no puede ser resuelto mediante ecuaciones matemáticas, se puede construir un programa basado en reglas o una combinación de ambos para evaluar la aptitud. En el caso de que existan pocas restricciones que sean muy importantes y no puedan ser violadas, es posible diseñar el esquema de representación de forma

adecuada para eliminar tempranamente las soluciones que no cumplan con estas restricciones (Affenzeller, Winkler, Wagner, & Beham, 2009).

### 3.8.8 Aplicaciones

Los algoritmos genéticos se emplean para solucionar problemas que resultan complicados de resolver con las técnicas de optimización convencionales, como aquellos que son difíciles de definir con precisión o que resultan complicados de modelar matemáticamente. Asimismo, se utilizan cuando la función objetivo presenta discontinuidades, no linealidades marcadas, aleatoriedad o características poco confiables o indefinidas.

La minería web presenta nuevos desafíos debido al constante aumento y cambio sin control de la información en la web. Para descubrir y analizar información útil, la inteligencia artificial debe integrarse en las herramientas web mediante la búsqueda, diseño y adquisición de algoritmos. Los algoritmos genéticos (AG) y sus características han mostrado resultados interesantes en varios campos de la minería web, como en la búsqueda y recuperación de información, donde se presentan como mecanismos de búsqueda y como motores de búsqueda complementarios para mejorar el rendimiento. En la optimización de consultas, se propone el uso de AG para construir perfiles de usuarios y monitorear el comportamiento de navegación. Además, en la representación de documentos basados en conjuntos de términos indexados y la minería distribuida, donde GEMGA (*Gene Expression Messy Genetic Algorithm*) es un algoritmo de búsqueda evolutiva paralela y como tal es una opción para abordar problemas de coincidencia de patrones en entornos con datos y recursos informáticos distribuidos (Velásquez & Palade, 2008).

Se destaca la investigación en relación a la estructura web y al comportamiento del consumidor. Se ha llevado a cabo la simulación de los efectos de las estrategias de marketing en un contexto de mercado competitivo mediante la implementación de un modelo que utiliza algoritmos genéticos y simulación multiagente para ajustar las características de los consumidores virtuales obtenidos a partir de un mercado real. Además, se ha utilizado una combinación de múltiples clasificadores en el modelo para predecir el comportamiento de compra de los consumidores en los sitios de comercio electrónico y encontrar las mejores combinaciones posibles. También se ha abordado la búsqueda web

como un problema de optimización mediante el uso de un AG para obtener las páginas más interesantes para el usuario. Por último, se ha propuesto construir un sitio web que proporcione a los usuarios la información deseada con la menor cantidad de hipervínculos posibles, lo que implica la construcción de un modelo que optimice la estructura web para una navegación más efectiva, utilizando un AG para precisar los hipervínculos que cumplan con este objetivo. (Velásquez & Palade, 2008).

Numerosas aplicaciones han demostrado el éxito de los Algoritmos Genéticos en aplicaciones prácticas. Un aspecto importante de esta historia de éxito es su amplia aplicabilidad. Una vez llegado al mundo de los Algoritmos Genéticos, suena atractivo modelar cada problema de optimización con funciones de *fitness* y adaptar y emplear Algoritmos Genéticos para resolverlos. También el diseño de operadores genéticos apropiados y la elección de parámetros suele ser una tarea conveniente. Además, los Algoritmos Genéticos de mayor éxito en aplicaciones incorporan conocimiento experto. Esto incluye procedimientos de inicialización, por ejemplo, el uso de soluciones que ya se conocen como padres en la primera generación y la modificación de partes de la solución con diseños expertos. En las aplicaciones, el profesional puede ser parte de un ciclo de optimización interactivo. La visualización de las ejecuciones del algoritmo genético puede respaldar la integración de las decisiones humanas en un proceso de optimización automatizado (Kramer, 2017).

### 3.8.9 Comparación con respecto a los métodos convencionales

Se pueden destacar las siguientes diferencias entre los AG y los métodos convencionales de optimización:

#### **1) Codificación**

La codificación de puntos en el espacio de búsqueda es utilizada por los algoritmos genéticos en lugar de los propios puntos. Los parámetros específicos no evolucionan, pero están completamente codificados. Esa es la importancia de esta representación (Sivanandam & Deepa, 1998).

## **2) Búsqueda paralela implícita**

En vez de llevar a cabo una búsqueda secuencial punto por punto, se realiza un conjunto de búsquedas simultáneamente, por lo que podemos decir que es una búsqueda implícitamente paralela de AG al mismo tiempo que cubre una amplia gama de posibles soluciones (Sivanandam & Deepa, 1998).

## **3) Uso directo de la función de adaptación**

Para la optimización, en general, el cálculo de la derivada es el primer método a seguir, sin embargo, no se puede aplicar a ninguna función, y tiene restricciones en el tipo de solución final, especialmente si son múltiples las soluciones posibles. Dado esto, los AG no sufren las limitaciones de los métodos de las derivadas, ya que no requieren derivadas u otras propiedades de la función objetivo, sino sólo de sí misma (Sivanandam & Deepa, 1998).

## **4) Transición probabilística**

La probabilidad de ocurrencia define las reglas de transición entre generaciones de operadores genéticos y de cumplimiento, en lugar de reglas deterministas (Sivanandam & Deepa, 1998).

## **5) Combinación de técnicas**

Los AG exploran regiones en el espacio de búsqueda de manera eficiente mientras aprovecha la evaluación de cada punto (Sivanandam & Deepa, 1998).

### **3.8.10 Factores relevantes para algoritmos genéticos**

En la aplicación de algoritmos genéticos en situaciones prácticas, suele suceder que la población de soluciones se vuelva muy similar, lo que indica que todos los individuos son casi iguales. Esta situación puede obstaculizar la capacidad del algoritmo para encontrar soluciones aún mejores, ya que puede quedar estancado en soluciones subóptimas que no son las mejores para el problema que se está tratando de resolver (Sivanandam & Deepa, 1998).

Se han desarrollado estrategias para combatir esta "deriva genética". Una técnica simple pero no muy efectiva consiste en introducir mutaciones después de la selección y el cruzamiento. Después de la selección y el cruzamiento, se escogen algunos bits de la población y se modifican aleatoriamente, cambiando unos pocos bits de 0 a 1 o de 1 a 0 (Sivanandam & Deepa, 1998).

El algoritmo genético es capaz de resolver problemas que no son demasiado complejos, además de ser eficiente y preciso. Esto lleva a una reducción en el costo computacional, con menos tiempo y recursos utilizados; debido a su simplicidad, se pueden apreciar grandes beneficios al usar un algoritmo genético.

La utilización de Algoritmos Genéticos en la Programación es un enfoque innovador que tiene la capacidad de abarcar diversas áreas de aplicación en las que no se cuenta con una solución clara para el problema. Sin embargo, es fundamental tener conciencia y la habilidad de distinguir cuáles soluciones son efectivas y cuáles no lo son.

---

## Capítulo 4. Elaboración de un modelo de algoritmo

---

### Introducción

En este capítulo, se presenta la construcción de un modelo de algoritmo para la optimización del comportamiento del usuario en un sitio web. Se describen los parámetros del algoritmo, la representación del individuo y el valor óptimo buscado. Se explica en detalle el modelo de caminos mínimos, incluyendo el razonamiento matemático y el pseudocódigo del modelo. Este capítulo es crucial para comprender cómo se llevará a cabo la optimización del comportamiento del usuario mediante el uso de un algoritmo específico.

### 4.1 Optimización Web y Usabilidad

Si una empresa quiere seguir siendo competidor en el ámbito digital, requiere de un sitio web para proveer de información específica de una manera simple y comprensible a los usuarios quienes la buscan. No obstante, las circunstancias suelen señalar que la disposición de la página web muchas veces impide a los usuarios hallar la información requerida, incluso si ésta se encuentra presente. Con esto en mente, para determinar la mejor estructura de red, a continuación, se describe esto a través de los estudios de los usuarios/consumidores (mediante el registro de la red), es posible determinar la mejor ruta o forma de moverse de manera más eficiente. (Maximizar la usabilidad de los hipervínculos existentes según los principios de usabilidad) en el sitio (Sivanandam & Deepa, 1998).

Teniendo en cuenta los datos de entrada para el algoritmo observados en la Tabla 2, se deben señalar de manera preliminar los supuestos y observaciones para su extracción de datos y operación.

Tabla 2 Datos entrantes del algoritmo

| Datos                                |
|--------------------------------------|
| Ruta                                 |
| Hipervínculo                         |
| Número de clics por página principal |
| Número de clics por usuario          |
| Número de clics por navegador        |
| Permanencia                          |
| Número total de paginas              |

Fuente: elaboración propia

### Supuestos y observaciones

1) Más visitas no significa necesariamente que sea mejor, **lo que buscamos es tener un alto acceso, pero de calidad**. Para saber si el tráfico es de calidad, se debe analizar las visitas considerando otras métricas como la duración de la visita y la tasa de rebote. (Fedesoft, Cluster IT de la cámara de comercio de Bogotá, 2021).

2) Se omite la consideración de la comparación de secuencias completas (conjuntos de sesiones) ya que el algoritmo genético opera iterativamente mediante la creación y eliminación de hipervínculos. Esto conduce a que la secuencia completa pierda su validez, es decir, la relación entre pares de páginas y el objetivo en términos de frecuencia de acceso (Bandyopadhyay & Pal, 2007).

3) No se tiene en cuenta la existencia de múltiples hipervínculos que lleven a diferentes  $x$  en una misma página  $y$ . Por razones de diseño, usabilidad y simplicidad de datos, se considera una ruta única entre estas páginas.

4) Los cálculos se realizan basados en los registros de visita del sitio, teniendo en cuenta cada clic realizado, por lo tanto, se basa implícitamente en los cálculos utilizados, ya que se tiene en cuenta el comportamiento del usuario para establecer una sesión.

5) Se considera que "adyacente" representa un estado estático en un momento dado. En este contexto, se sugiere llevar a cabo investigaciones en sitios con baja dinámica estructural, es decir, sitios que experimentan cambios mínimos. La recomendación se basa en la idea de que la estática de la adyacencia puede ser más efectiva de estudiar en entornos web con cambios limitados.

## 4.2 Parámetros del algoritmo

Los parámetros son valores entrantes que no se modifican durante la ejecución del algoritmo y son en gran parte dependientes de la etapa en la que se implementan. Para continuar con los algoritmos genéticos es necesario definir:

### **Tamaño de los elementos:**

Esto corresponde con una fila vectorial de longitud  $m \times m$ , donde  $m$  es la cantidad total de páginas en su sitio web. Su valor binario interno representa la presencia o ausencia de hipervínculos entre páginas (Sivanandam & Deepa, 1998).

### **Tamaño de la población:**

Muestra cuántas personas existirá en cada generación. En un sentido algorítmico, esto implica explorar el espacio de posibles soluciones a considerar (Bandyopadhyay & Pal, 2007).

### **Criterios de parada:**

Es el criterio utilizado para decidir cuándo termina el algoritmo. Pueden especificar la cantidad máxima de repeticiones que realizará el algoritmo, especificar un plazo de tiempo para ejecutar el algoritmo y especificando un límite del valor óptimo (el algoritmo se detiene cuando encuentra un valor menor o igual a dicha valor óptimo), especificando una tarea de tolerancia que tenga en cuenta la variación acumulativa en el valor óptimo pasando de generación en generación, o tenga en cuenta el intervalo de tiempo máximo. Por tanto, si la función objetivo no mejora, el algoritmo se detiene (Bandyopadhyay & Pal, 2007).

Los demás parámetros corresponden a los requeridos por el operador genético del AG, que dependerán del estudio.



- **Selección:** selección de la próxima generación de padres en base a sus valores de aptitud (Sivanandam & Deepa, 1998).
- **Reproducción:** método de producción de descendencia, indicando qué individuos tendrán garantía de supervivencia, el porcentaje resultante de cruces y el porcentaje resultante de mutaciones (Sivanandam & Deepa, 1998).
- **Mutación:** la tasa de variación aleatoria en los elementos de una población que proporciona diversidad genética y explora un espacio mayor (Sivanandam & Deepa, 1998).
- **Cruce:** una forma de unión de padres para producir un nuevo elemento (Sivanandam & Deepa, 1998).

### 4.3 Representación del Individuo

Para realizar una mejor organización estructural del sitio web, se crea una matriz de adyacencia, representación de una matriz binaria que contiene ligas como se muestra a continuación:

$$\bar{R} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & \dots & M \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ \dots \\ M \end{matrix} & \begin{bmatrix} 1 & 1 & 1 & 0 & \dots & 1 \\ 0 & 1 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & 1 & \dots & 1 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 1 \end{bmatrix} \end{matrix}$$

$R$  — Matriz de  $m \times m$  (donde  $m$  es la cantidad total de páginas del sitio web)

$R_{xy}$  significa la presencia (valor 1) o que no existe (valor 0) de un hipervínculo entre la página  $x$  y la página  $y$ , donde  $x, y \in \{1, 2, 3, \dots, m\}$

Para facilitar el análisis, la matriz se linealiza de manera más directa y lógica con los operadores genéticos necesarios para hacerlo. Por lo tanto, para crear un

elemento binario lineal para la aplicación del algoritmo genético, se procede de la siguiente manera:

**1.- Largo del elemento:** corresponde a todos los datos comprendidos en su matriz de origen. Es decir, la longitud  $M$  de un individuo lineal equivale a  $m \times m$ , de tal forma que  $m$  representa la cifra total de páginas de su sitio web.

**2.- Linealización:** la linealización se definirá como el procedimiento de formación de un arreglo lineal (cadena) de representaciones de individuos genéticos utilizando una matriz estructural, en la cual el primer *bit* corresponderá a la presencia o ausencia de un hipervínculo entre las páginas 1 y 2; el segundo *bit* corresponderá a la presencia de un hipervínculo entre las páginas 1 y 3, Para el tercer bit se continuará de la misma forma y así consecutivamente. La presencia se expresa en binario: 1 = hipervínculo presente, 0 = sin hipervínculo. Desde un punto de vista matemático, la relación entre los individuos lineales y sus componentes de matriz (hipervínculos) se puede resumir de la siguiente manera:

Sea:

$C$  = Elemento que se considerará (serie de representación)

$c_i$  = Cada *bit* de la serie, con  $x \in \{1,2,\dots,m\}$

De esta forma, con  $m$  igual a 3 páginas se obtendrían:

---

|   |            |                       |         |
|---|------------|-----------------------|---------|
| $c1 =$ Presencia de hipervínculo de 1 a 1 | $\implies$ | $c1: 1 \rightarrow 1$ |         |
| $c2 =$ Presencia de hipervínculo de 1 a 2 | $\implies$ | $c2: 1 \rightarrow 2$ |         |
| $c3 =$ Presencia de hipervínculo de 1 a 3 | $\implies$ | $c3: 1 \rightarrow 3$ | $m = 3$ |

---

|   |            |                       |          |
|---|------------|-----------------------|----------|
| $c4 =$ Presencia de hipervínculo de 2 a 1 | $\implies$ | $c4: 2 \rightarrow 1$ |          |
| $c5 =$ Presencia de hipervínculo de 2 a 2 | $\implies$ | $c5: 2 \rightarrow 2$ |          |
| $c6 =$ Presencia de hipervínculo de 2 a 3 | $\implies$ | $c6: 2 \rightarrow 3$ | $2m = 6$ |

---

|   |            |                       |          |
|---|------------|-----------------------|----------|
| $c7 =$ Presencia de hipervínculo de 3 a 1 | $\implies$ | $c7: 3 \rightarrow 1$ |          |
| $c8 =$ Presencia de hipervínculo de 3 a 2 | $\implies$ | $c8: 3 \rightarrow 2$ |          |
| $c9 =$ Presencia de hipervínculo de 3 a 3 | $\implies$ | $c9: 3 \rightarrow 3$ | $3m = 9$ |

---

En el ejemplo de la tabla 3 se muestra la presencia de hipervínculos a través de las páginas 1 y 3, 1 y 2, 3 y 1 y por último entre 2 y 3.

*Tabla 3 Ejemplificación de la presentación de los elementos genéticos con m igual a tres páginas*

| Bit =>       | C <sub>1</sub> | C <sub>2</sub> | C <sub>3</sub> | C <sub>4</sub> | C <sub>5</sub> | C <sub>6</sub> | C <sub>7</sub> | C <sub>8</sub> | C <sub>9</sub> |
|--------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Link =>      | 1 → 1          | 1 → 2          | 1 → 3          | 2 → 1          | 2 → 2          | 2 → 3          | 3 → 1          | 3 → 2          | 3 → 3          |
| Individuo => | 0              | 1              | 1              | 0              | 0              | 0              | 1              | 1              | 0              |

Fuente: elaboración propia

En la tabla 4 se hace una generalización.

*Tabla 4 Generalización de elementos*

|                |                |                |     |                |                  |                  |     |                  |                   |                   |     |                   |                   |     |                  |
|----------------|----------------|----------------|-----|----------------|------------------|------------------|-----|------------------|-------------------|-------------------|-----|-------------------|-------------------|-----|------------------|
| c <sub>1</sub> | c <sub>2</sub> | c <sub>3</sub> | ... | c <sub>m</sub> | c <sub>m+1</sub> | c <sub>m+2</sub> | ... | c <sub>m+m</sub> | c <sub>2m+1</sub> | c <sub>2m+2</sub> | ... | c <sub>2m+m</sub> | c <sub>3m+1</sub> | ... | c <sub>mxm</sub> |
|----------------|----------------|----------------|-----|----------------|------------------|------------------|-----|------------------|-------------------|-------------------|-----|-------------------|-------------------|-----|------------------|

Fuente: elaboración propia

Es decir, cada componente  $c_x$  del elemento será:

$$c_x \begin{cases} 0 & \text{si } Z_{kx}(y) = 0 \\ 1 & \text{si } Z_{kx}(y) = 1 \end{cases}$$

Donde:

$$Z_{kx}(y) \begin{cases} 0 & \text{ausencia de hipervínculo entre las páginas } k(y) \text{ y } j(y) \\ 1 & \text{presencia de hipervínculo entre las páginas } k(y) \text{ y } j(y) \end{cases}$$

De este modo, cada *bit* en la posición y el individuo, tendrá implícitamente un enlace entre las dos páginas según se indica en la siguiente ecuación 1.

*Ecuación 1 Enlace entre páginas y posición de los elementos lineales.*

$$\begin{array}{l} k = \left\lceil \frac{i}{m} \right\rceil \\ j = \begin{cases} m & \text{si } \text{res}(i, m) = 0 \\ \text{res}(i, m) & \text{si } \text{res}(i, m) \neq 0 \end{cases} \end{array}$$

Función cajón superior o función techo.  
Función *res* corresponde a la función residuo.

Fuente: elaboración propia con base en (Academia Lab, 2023)

La función de cajón, también llamada función de techo. Si coincide con el cajón superior, es definido como la función que se aplica a un número real  $x$  y retorna el número entero más pequeño  $k$  mayor que  $x$ , como puede verse en la ecuación 2 para el cajón inferior, encuentra el número entero más grande  $k$  menor que  $x$ .

*Ecuación 2 Función cajón inferior*

$$\lfloor x \rfloor = \min\{n \in \mathbb{Z} \mid n \geq x\}.$$

Fuente: (Academia Lab, 2023)

Por otro lado, la función  $\text{res}(i, m)$  siempre devuelve un valor entero, ya que sólo devuelve el resto de la división de  $i$  y  $m$ . La ecuación 3 representa la operación que permite utilizar la función de cajón inferior para la obtención del residuo de la función.

*Ecuación 3 Función residuo*


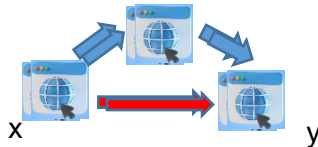



$$\text{res}(i, m) = i - m \cdot \left\lfloor \frac{i}{m} \right\rfloor$$

Fuente: elaboración propia con base en (Equipo editorial de IONOS, 2021)

## 4.4 Valor Óptimo

La función de valor óptimo es la clave para un correcto diseño, convergencia y regularización de datos reales (Mitchell, 1998). Antes de comenzar a trabajar, se realizan cálculos iniciales en base a los modelos a desarrollar para determinar sus funciones de aptitud para encontrar los mejores elementos. La Tabla 5 muestra el esquema que contiene los conceptos utilizados para describir los cálculos realizados, seguido de explicaciones y esquemas representativos para facilitar la comprensión de los cálculos (Bandyopadhyay & Pal, 2007).

Tabla 5 Características de Acceso y Navegación en el Sitio Web

| Característica         | Descripción  | Representación   |
|------------------------|--|--|
| <b>Acceso Directo</b>  | Se denomina acceso directo entre <b>x</b> y <b>y</b> al acto de avanzar (acceder) desde la página <b>x</b> a la <b>y</b> utilizando una liga directa que lo conecte.   |   |
| <b>Acceso Objetivo</b> | Es llamado acceso objetivo entre <b>x</b> y <b>y</b> al acto de avanzar (acceder) desde la página <b>x</b> a la <b>y</b> considerando que, en algún momento de la sesión paso por una o varias páginas intermedias, es decir, el número de ligas visitadas de una página a otra es superior a 1. |    |
| <b>Pasadas</b>         | Se denomina número de pasadas entre <b>x</b> a <b>y</b> al número de veces que se pasa de la posición <b>x</b> a la posición <b>y</b> .  |   |
| <b>Posición</b>        | Se le llama posición a la página <b>x</b> , cuando tienes acceso a esa página en una sesión específica (primera página consultada, segunda, tercera, etc.).  |   |
| <b>Longitud</b>        | El número de ligas que hay que pasar para llegar <b>x</b> a <b>y</b> se le denomina largo entre <b>x</b> y <b>y</b> .  |   |
| <b>Factor Tiempo</b>   | Factor tiempo es igual al tiempo neto (en segundos) considerando la página <b>j</b> de la sesión <b>S</b> , se divide por el total del tiempo de la sesión <b>S</b> (la suma de los tiempos de la totalidad de las páginas pertenecientes a la sesión).  | <p>Ecuación 4 Factor tiempo</p> $\text{Factor de Tiempo}_{j,S} = \frac{\text{Tiempo Neto}_{j,S}}{\sum_{i=1}^N \text{Tiempo}_{i,S}}$ <p>Fuente: Elaboración propia con base en (Velásquez J. , Yasuda, Aoki, &amp; Weber, 2004)</p> |

Fuente: elaboración propia con base en (Velásquez J. , Yasuda, Aoki, & Weber, 2004)

Considerando la Ecuación 4 se tiene lo siguiente:

- Factor de Tiempo $_{j,S}$  es el factor de tiempo para la página  $j$  en la sesión  $S$ .
- Tiempo Neto $_{j,S}$  es el tiempo neto (en segundos) en la página  $j$  durante la sesión  $S$ .
- Tiempo $_{i,S}$  es el tiempo (en segundos) en la página  $i$  durante la sesión  $S$ .
- La sumatoria se realiza sobre todas las páginas ( $N$ ) que pertenecen a la sesión  $S$ .

La ecuación 4 en la tabla 5 intenta capturar el verdadero significado del tiempo dedicado a una página versus el tiempo dedicado a visitar otras páginas en la misma sesión. La suposición subyacente es que el tiempo que se pasa en una página está relacionado con el interés del usuario en su contenido y estructura. En otras palabras, cuanto más tiempo pase un usuario en una página durante una sesión, mayor será el interés percibido en el contenido de esa página. Con esta idea en mente, se calcula un "factor de tiempo" que permite comparar la importancia relativa de las páginas entre diferentes sesiones. Por ejemplo, acceder a la página  $x$  durante 4 minutos en una sesión de 8 minutos informa que  $x$  tiene un factor de tiempo del 50 %, mientras que acceder a la página  $z$  durante 10 minutos en otra sesión de 0 minutos informa que  $z$  tiene un factor de tiempo del 33.33 %. En otras palabras, la página  $x$  es más importante que la página  $z$ , a pesar de que la página  $z$  tiene un tiempo de acceso total más largo (en minutos). (Velásquez J. D., Yasuda, Aoki, & Weber, 2004)

Al construir el valor óptimo se consideraron aspectos como: desarrollar una función que suponga una estructura del hipervínculo conveniente para la navegación, describir los beneficios de usar hipervínculos, medir el potencial de crear un hipervínculo que no existe y aparentemente nunca se utiliza. Este proceso produce modelos tentativos con una naturaleza probabilística común que se describen a continuación (EVANS & WALKER, 2004.) (P. Baldi & Smyth, 2003).

## 4.5 Modelo: caminos mínimos.

El siguiente modelo es el propuesto que usamos, se llama "caminos mínimos" porque utiliza el concepto de camino más corto de una liga a otra al realizar ciertos cálculos. La idea es generar un peso para cada hipervínculo existente en función del uso de cada hipervínculo existente y la importancia del tiempo de visita (en la misma sesión) de la página de destino del hipervínculo. Por otro lado, si no existe un hipervínculo entre **x** y **y**, su peso se construye a partir de multiplicar los pesos de los hipervínculos existentes pertenecientes a la ruta mínima requerida para que **x** y **y** se comuniquen (Fronita, Gernowo, & Gunawan, 2018).

Cálculos preliminares (fórmulas) Primero, se calcula la periodicidad de cada *hipervínculo*, de la siguiente manera:

### *Ecuación 5 Periodicidad de cada hipervínculo*

$$\lambda_{xy} = \frac{\text{Número de rutas de para pasar de } x \text{ a } y}{\text{Número total de usos del hipervínculo saliendo de } x}$$

Fuente: elaboración propia con base en (Ross, 2000)

Donde  $\lambda_{xy}$  representa la probabilidad de visitar la página **y** después de estar en la página **x**, y se calcula como la proporción entre el número de veces que se pasa de **x** a **y** y el número total de usos del hipervínculo que comienza en **x**. En otras palabras,  $\lambda_{xy}$  indica cuán probable es que un usuario cambie de la página **x** a la página **y**, considerando todos los casos en los que se utiliza ese hipervínculo desde **x**. Este concepto se basa en la probabilidad condicional de cambiar de una página a otra, también conocida como probabilidad de uso del hipervínculo (Ross, 2000).

Otra medida representativa de la importancia de cada hipervínculo correspondería al factor tiempo dedicado a la última página, se denota el tiempo del hipervínculo con  $t_{xy}$ :

$t_{xy}$  = Tiempo medio de estancia en la página cuando termina el acceso directo de la página inicial **y** comienza la página **x** destino.

Como se explicó,  $t_{xy}$  es el factor de tiempo promedio de la página de destino del hipervínculo  $y$  a partir de  $x$ . En concreto, se supone que el tiempo que se pasa en una página de objetivo/destino está relacionado con la importancia de los hipervínculos, lo que indica que cuanto más tiempo se pasa en una página de destino, más útiles son los hipervínculos que conducen a ella.

A continuación, se crearon los pesos de los hipervínculos existentes según lo definido en la ecuación 6.

*Ecuación 6 Peso del hipervínculo existente xy*

$$\pi_{xy} = \lambda_{xy} * t_{xy}$$

Fuente: elaboración propia con base en (Krug, 2000)

Esta ecuación refleja la importancia del hipervínculo  $\pi_{xy}$  al depender tanto de la probabilidad de uso  $\lambda_{xy}$  como del tiempo de permanencia en la página destino  $t_{xy}$ . Cuanto mayor sea el factor de probabilidad  $\lambda_{xy}$  y mayor el factor tiempo  $t_{xy}$ , mayor será el peso del hipervínculo.

La proporcionalidad directa entre  $\pi_{xy}$ ,  $\lambda_{xy}$ , y  $t_{xy}$  indica que el peso del hipervínculo se incrementa proporcionalmente a la probabilidad de uso y al tiempo de permanencia en la página destino.

Esta ecuación se aplica a los datos y solo se calcula para aquellos pares de páginas que tienen un hipervínculo entre ellas ( $x$  y  $y$ ). Los valores resultantes de  $\pi_{xy}$  se utilizan para tomar decisiones sobre mantener o borrar hipervínculos existentes y también pueden orientar la creación de nuevos enlaces, considerando el comportamiento del consumidor y calculando los caminos existentes para llegar de una página a otra en el caso de no poseer un enlace directo entre ellas.



### Ecuación 7 Camino

$$p = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), \dots, (x_{l_p}, y_{l_p})\}$$
$$\text{Tal que } y_k = x_{k+1}, \forall a \in [1, l_p - 1]$$

Fuente: elaboración propia con base en (Mitchell, 1998)

Vamos a considerar que  $P$  es el conjunto de todos los caminos posibles en el gráfico y un recorrido en específico de longitud  $l_p$  que recorre entre  $x_1$  y  $y_{l_p}$ , por lo tanto, se define un camino, como se muestra en la ecuación 7.

### Ecuación 8 Probabilidad mediante caminos posibles

$$\lambda'_{xy} = \sum_{p \in P(x,y)} \prod_{(k,h) \in p} \lambda_{kh}$$

Fuente: elaboración propia con base en (Mitchell, 1998)

La ecuación 8 se basa en el concepto de probabilidad (Ross, 2000) para determinar la probabilidad de transición  $\lambda'_{xy}$  al moverse de  $x$  a  $y$  sin utilizar hipervínculos. En este contexto, se exploran diversas opciones de rutas entre los nodos, cada una compuesta por una secuencia única de transiciones. Para ilustrar, podríamos imaginar diferentes escenarios, como avanzar de  $x$  a una ubicación ficticia ( $i$ ), luego de  $i$  a otra ubicación ficticia ( $j$ ), y finalmente de  $j$  a  $y$  en una primera ruta. O, por ejemplo, podríamos considerar una segunda ruta donde avanzamos de  $x$  a otra ubicación ficticia ( $z$ ) y de  $z$  a  $y$ .

La diversidad de elecciones para navegar a través del mapa está directamente influenciada por la cantidad de conexiones presentes en la estructura. Conforme aumenta el número de enlaces, se generan más alternativas para trazar rutas. Este enfoque nos brinda la capacidad de capturar la complejidad del grafo, abordando todas las posibles trayectorias que los usuarios podrían seguir al explorar el sitio web.

La Ecuación 6 ilustra que a medida que las rutas posibles se extienden (más páginas visitadas en la secuencia), el producto de las probabilidades asociadas disminuye, ya que estos valores son menores que 1. En lugar de considerar todas las rutas posibles, nos centramos en la ruta más corta, la que requiere el menor número de pasos para ir de  $x$  a  $y$ . La probabilidad de que un hipervínculo no exista refleja las conexiones potenciales en la estructura del sitio web. En otras palabras, tiene más sentido agregar un hipervínculo si su beneficio potencial supera la ruta más corta, ya que esto introduce una vía más directa y relevante, creando un acceso más eficiente (Fronita, Gernowo, & Gunawan, 2018).

Es importante señalar que a medida que disminuye el número de hipervínculos necesarios para acceder a la información deseada, la satisfacción del usuario tiende a aumentar. La ecuación definitiva para calcular la probabilidad potencial ( $\lambda'_{xy}$ ) se muestra a continuación:

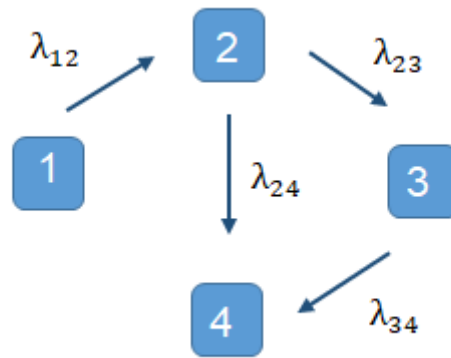
*Ecuación 9 Peso potencial camino mínimo*

$$\lambda'_{xy} = \prod_{(a,b) \in p^{min}} \lambda_{ab}$$

Fuente: elaboración propia con base en (Mitchell, 1998)

Por ejemplo, al analizar la figura 22, se concluye que faltan los pesos directos entre 1 y 4 y esto es debido a la falta de hipervínculos. En este contexto, el camino más corto es  $1 \Rightarrow 2 \Rightarrow 4$  en vez de  $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4$ , que es otro camino posible. Para este caso  $\lambda'_{14}$  sería como  $\lambda'_{14} = \lambda_{12}\lambda_{24}$

Figura 22 Ejemplo hipervínculos entre páginas



Fuente: elaboración propia

A medida que se calculan las probabilidades potenciales  $\lambda'$ , también se determinan los pesos potenciales para normalizar estas probabilidades en cada fila, siguiendo la regla de transición. Con base a la ley de los grandes números (Ross, 2000) se asume que el tiempo promedio necesario para llegar a la página de destino se mantiene constante.

### Consideraciones Finales

En este capítulo, se exploraron a fondo aspectos cruciales relacionados con la optimización web y la usabilidad de un sitio. Se destacó la importancia de proporcionar información de manera accesible y comprensible para los usuarios, subrayando la necesidad de estructuras de red eficientes. La utilización de algoritmos genéticos y modelos como "camino mínimo" se reveló como un enfoque prometedor para abordar la optimización de la estructura del sitio web.

Se establecieron supuestos y observaciones esenciales para la implementación del algoritmo, considerando factores como la calidad del tráfico y la simplicidad de datos. Además, se detallaron parámetros clave del algoritmo y la representación del individuo, proporcionando una base sólida para la aplicación práctica.

La sección dedicada al valor óptimo destacó la importancia de desarrollar funciones de aptitud para evaluar la idoneidad de los elementos generados. El modelo de "camino mínimo" se presentó como un enfoque efectivo que

incorpora la probabilidad de uso y el tiempo dedicado a las páginas, permitiendo decisiones informadas sobre la creación y eliminación de hipervínculos.

Estos hallazgos subrayan la necesidad continua de investigar y aplicar enfoques innovadores para mejorar la experiencia del usuario en entornos digitales. La combinación de algoritmos genéticos, modelos de optimización y consideraciones de usabilidad se posiciona como un campo crucial para el desarrollo web efectivo y la satisfacción del usuario.

---

## Capítulo 5. Validación del modelo

---

### Introducción

En este capítulo, se aborda la validación del modelo propuesto. Se analiza el comportamiento del usuario y se exploran diferentes técnicas de análisis, como el análisis de eventos de comportamiento, la retención de usuarios, el modelo de embudo, la ruta de comportamiento y el modelo de Fogg. Se presenta el modelo AISAS (*Attention, Interest, Search, Action, Share*) como un marco para comprender el comportamiento del usuario. Además, se detallan aspectos relacionados con la representación de los datos de entrada y se explica cómo se implementará el modelo de algoritmo en un estudio de caso. Este capítulo demuestra la aplicabilidad y relevancia del modelo propuesto.

### 5.1 Marketing digital: Un Estudio de Análisis de Comportamiento de Usuarios

Más recientemente, el *marketing* digital ha reemplazado a las estrategias de *marketing* tradicionales. Los negocios prefieren hacer publicidad de sus productos en páginas web y plataformas de redes sociales. Pero dirigir la atención al público adecuado sigue siendo un reto para el *marketing* en línea. Puede ser caro gastar millones en mostrar tu anuncio a un público que probablemente no comprará tu producto (Chaffey & Ellis-Chadwick, 2016).

Se trabajará con la información y ligas públicas de unos datos de publicidad, indicando si un usuario de Internet específico hizo clic o no en un anuncio en el sitio web de una empresa para crear un algoritmo de aprendizaje automático que prediga si un usuario concreto hace clic en un sitio. Los datos están formados de algunas variables, por ejemplo: "Tiempo diario empleado en el sitio", "Edad", "Ingresos del área", "Uso diario de Internet", "Línea de tema del sitio", "Ciudad", "Genero", "País", "Marca de tiempo" y "Haga clic en el sitio".

Se puede distinguir entre dos resultados posibles para la variable "Hacer clic en un sitio"(0 y 1): si el usuario no hace clic en el anuncio, se asigna un valor de 0; si el usuario hace clic en el anuncio, se asigna un valor de 1.

Veremos si se puede utilizar las otras variables para pronosticar de forma precisa el valor de la variable "Haga clic en el sitio". Asimismo, llevaremos a cabo un análisis exploratorio de los datos para saber cómo el "Tiempo Diario Pasado en el Sitio" en conjunto con la "Línea de Tema del Sitio" influye en la decisión del usuario para hacer clic sobre la liga (Chaffey & Ellis-Chadwick, 2016).

## 5.2 Análisis del comportamiento del usuario

A continuación, exploraremos en detalle el análisis del comportamiento del usuario en el contexto de la Facultad de Contaduría y Administración (FCA) de la UNAM durante el período 2019-2020. El análisis del comportamiento del usuario es esencial para comprender cómo los visitantes interactúan con los recursos en línea y las actividades proporcionadas por la facultad. A través de la tabla 6 que se observa a continuación, examinaremos las actividades clave, la frecuencia con la que se realizaron, los sitios web involucrados y las rutas de acceso necesarias para acceder a la información relevante. Esta información nos permitirá obtener una visión más profunda de cómo los usuarios interactuaron con los recursos en línea y qué áreas fueron de mayor interés durante el período especificado.

Tabla 6 Actividades Principales y nivel de profundidad de la FCA 2019-2020

| Actividades Principales 2019-2020                           | Número de casos | Sitio             | Ruta en clic  | URL   |
|---|-----------------|-------------------|---|---|
| <b>Examen general de conocimientos</b>                      | 58              | Titulación        | FCA/Alumnos/titulación/Examen   | <a href="http://titulacion.fca.unam.mx/egc_convocatoria.php">http://titulacion.fca.unam.mx/egc_convocatoria.php</a>                             |
| <b>Exámenes de maestrías, especialidades y doctorados</b>   | 1157            | Posgrado          | FCA/Posgrado/(Maestrías-Especializaciones-Doctorado)/Convocatoria                             | <a href="https://posgrado.fca.unam.mx/admision.php">https://posgrado.fca.unam.mx/admision.php</a>   |
| <b>Examen de conocimientos para entrar a maestría</b>       | 870             | Posgrado          | FCA/Posgrado/Admisión/Convocatoria  | <a href="https://posgrado.fca.unam.mx/admision.php">https://posgrado.fca.unam.mx/admision.php</a>   |
| <b>Diplomado en Línea</b>                                   | 2948            | Titulación        | FCA/Alumnos/titulación/Diplomado/Diplomado en línea   | <a href="http://titulacion.fca.unam.mx/dip_lin_consiste.php">http://titulacion.fca.unam.mx/dip_lin_consiste.php</a>                             |
| <b>Exámenes de colocación del idioma inglés</b>             | 3718            | Centro de Idiomas | FCA/Centro de Idiomas/Curso Inglés/(Presencial-A Distancia)                                   | <a href="http://idiomas.fca.unam.mx/curso_ingles_distancia.php">http://idiomas.fca.unam.mx/curso_ingles_distancia.php</a>                       |
| <b>Exámenes globales de conocimientos del idioma inglés</b> | 525             | Centro de Idiomas | FCA/Centro de Idiomas/Examen Global/(Presencial-A Distancia-Resultados) /Mes                  | <a href="http://idiomas.fca.unam.mx/examen_global_presencial_octubre.php">http://idiomas.fca.unam.mx/examen_global_presencial_octubre.php</a>   |
| <b>Examen de comprensión de lectura de textos en inglés</b> | 1114            | Centro de Idiomas | FCA/Centro de Idiomas/Examen Global/(Presencial-A Distancia-Resultados) /Mes                  | <a href="http://idiomas.fca.unam.mx/examen_lectura_presencial_octubre.php">http://idiomas.fca.unam.mx/examen_lectura_presencial_octubre.php</a> |
| <b>Exámenes Parciales del SUAYED</b>                        | 628             | SUAYED            | FCA/SUAYED/Inscripción Exámenes Parciales y Globales/Autenticarse /mandar datos de Formulario | <a href="https://suayedfca.unam.mx/ema">https://suayedfca.unam.mx/ema</a>   |
| <b>Exámenes Globales del SUAYED</b>                         | 104             | SUAYED            | FCA/SUAYED/Inscripción Exámenes Parciales y Globales/Autenticarse /mandar datos de Formulario | <a href="https://suayedfca.unam.mx/ema">https://suayedfca.unam.mx/ema</a>   |

Fuente: elaboración propia

La Tabla 6 ofrece una visión general de varias actividades principales realizadas en la FCA de la UNAM durante el período 2019-2020. Cada entrada en la tabla representa una actividad específica, proporcionando detalles importantes sobre

el número de casos, el sitio web relacionado, la ruta de clic necesaria para acceder a la información y la URL correspondiente.

**Número de Casos:** La columna "Número de casos" indica cuántas veces se realizó cada actividad en el período 2019-2020. Por ejemplo, se llevaron a cabo 58 exámenes generales de conocimientos, 1,157 exámenes de maestrías, especialidades y doctorados, 870 exámenes de conocimientos para entrar a maestría, 2,948 diplomados en línea, 3,718 exámenes de colocación del idioma inglés, 525 exámenes globales de conocimientos del idioma inglés, 1,114 exámenes de comprensión de lectura de textos en inglés, 628 exámenes parciales del Sistema de Universidad Abierta y Educación a Distancia (SUAYED) y 104 exámenes globales del SUAYED.

**Sitio:** La columna "Sitio" especifica la ubicación dentro del sitio web de la FCA donde se encuentra la información relacionada con cada actividad. Por ejemplo, las actividades relacionadas con exámenes generalmente se encuentran en las secciones de "Titulación," "Posgrado," "Centro de Idiomas," o "SUAYED," dependiendo de la actividad específica.

**Ruta en Clic:** La columna "Ruta en Clic" describe la secuencia de pasos o clics que los usuarios deben seguir para acceder a la información sobre cada actividad. Esto incluye la estructura de carpetas y enlaces necesarios para encontrar los detalles relevantes.

**URL:** La columna "URL" proporciona el enlace directo (dirección web) que conduce a la información detallada sobre cada actividad. Las URL son útiles para acceder directamente a la información sin necesidad de navegar manualmente a través del sitio web.

La Tabla 6 presenta una descripción detallada de diversas actividades realizadas en la FCA de la UNAM durante el período 2019-2020. Proporciona información valiosa sobre la frecuencia de estas actividades, su ubicación en el sitio web, cómo acceder a ellas y enlaces directos a la información relevante. Esto puede ser útil para la gestión y la planificación de actividades en el sitio web de la facultad.



Tabla 7 Actividades Principales y nivel de profundidad de la FCA 2020-2021

| Actividades Principales 2020-2021                           | Número de casos | Sitio             | Ruta en clic   | URL   |
|---|-----------------|-------------------|--|---|
| <b>Examen general de conocimientos</b>                      | 73              | Titulación        | FCA/Alumnos/titulación/Examen  | <a href="http://titulacion.fca.unam.mx/egc_convocatoria.php">http://titulacion.fca.unam.mx/egc_convocatoria.php</a>                             |
| <b>Exámenes de maestrías, especialidades y doctorados</b>   | 850             | Posgrado          | FCA/Posgrado/(Maestrías-Especializaciones-Doctorado)/Convocatoria                            | <a href="https://posgrado.fca.unam.mx/admision.php">https://posgrado.fca.unam.mx/admision.php</a>   |
| <b>Examen de conocimientos para entrar a maestría</b>       | 678             | Posgrado          | FCA/Posgrado/Admisión/Convocatoria   | <a href="https://posgrado.fca.unam.mx/admision.php">https://posgrado.fca.unam.mx/admision.php</a>   |
| <b>Diplomado en Línea</b>                                   | 3263            | Titulación        | FCA/Alumnos/titulación/Diplomado/Diplomado en línea  | <a href="http://titulacion.fca.unam.mx/dip_lin_consiste.php">http://titulacion.fca.unam.mx/dip_lin_consiste.php</a>                             |
| <b>Exámenes de colocación del idioma inglés</b>             | 2636            | Centro de Idiomas | FCA/Centro de Idiomas/Curso Inglés/(Presencial-A Distancia)                                  | <a href="http://idiomas.fca.unam.mx/curso_ingles_distancia.php">http://idiomas.fca.unam.mx/curso_ingles_distancia.php</a>                       |
| <b>Exámenes globales de conocimientos del idioma inglés</b> | 2718            | Centro de Idiomas | FCA/Centro de Idiomas/Examen Global/(Presencial-A Distancia-Resultados) /Mes                 | <a href="http://idiomas.fca.unam.mx/examen_global_presencial_octubre.php">http://idiomas.fca.unam.mx/examen_global_presencial_octubre.php</a>   |
| <b>Examen de comprensión de lectura de textos en inglés</b> | 2360            | Centro de Idiomas | FCA/Centro de Idiomas/Examen Global/(Presencial-A Distancia-Resultados) /Mes                 | <a href="http://idiomas.fca.unam.mx/examen_lectura_presencial_octubre.php">http://idiomas.fca.unam.mx/examen_lectura_presencial_octubre.php</a> |
| <b>Exámenes Parciales del SUAYED</b>                        | 626             | SUAYED            | FCA/SUAYED/Inscripción Exámenes Parciales y Globales/Autenticarse/mandar datos de Formulario | <a href="https://suayedfca.unam.mx/ema">https://suayedfca.unam.mx/ema</a>   |
| <b>Exámenes Globales del SUAYED</b>                         | 127             | SUAYED            | FCA/SUAYED/Inscripción Exámenes Parciales y Globales/Autenticarse/mandar datos de Formulario | <a href="https://suayedfca.unam.mx/ema">https://suayedfca.unam.mx/ema</a>   |

Fuente: elaboración propia

La Tabla 7 presenta una descripción detallada de diversas actividades realizadas en la FCA de la UNAM durante el período 2020-2021. Proporciona información valiosa sobre la frecuencia de estas actividades, su ubicación en el sitio web,

cómo acceder a ellas y enlaces directos a la información relevante. Esto puede ser útil para la gestión y la planificación de actividades en el sitio web de la facultad.

### 5.2.1 Comprensión del análisis del comportamiento del usuario

El análisis del comportamiento del usuario consiste en realizar un estudio de la conducta del consumidor en el producto y los datos detrás del comportamiento, mediante la construcción de modelos de comportamiento del usuario y retratos de usuarios. **Para cambiar las decisiones de productos, lograr operaciones refinadas y guiar el crecimiento empresarial** (Chaffey & Ellis-Chadwick, 2016)

En el proceso de operación del producto, al recopilar, almacenar, rastrear, analizar y aplicar datos de comportamiento del usuario, es posible encontrar factores virales, características de grupo y usuarios objetivo que se dan cuenta del crecimiento personal del usuario. Para restaurar profundamente los escenarios de uso del usuario, las reglas de operación, las rutas de acceso y las características de comportamiento (Rana, y otros, 2020).

### 5.2.2 El propósito del análisis del comportamiento del usuario

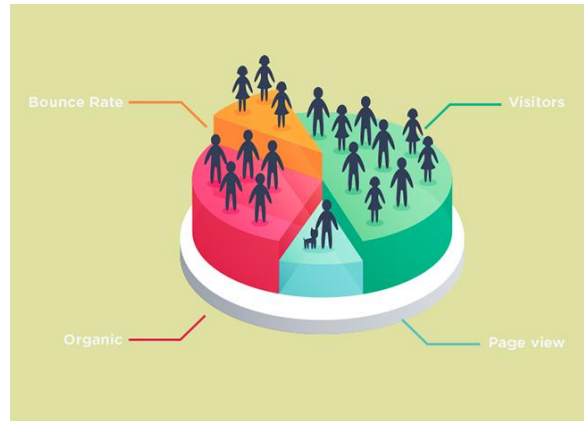
El análisis del comportamiento del usuario basado en datos es particularmente importante para productos en industrias como las finanzas por Internet, la nueva venta minorista, la cadena de suministro, la educación en línea, la banca y los valores. El propósito del análisis del comportamiento del usuario es: promover la iteración de productos, lograr un *marketing* de precisión, proporcionar servicios personalizados e impulsar las decisiones de productos (Verbeek & Slob, 2006).

Reflejado principalmente en los siguientes aspectos:

- Para los productos, ayuda a verificar la viabilidad del producto, estudiar las decisiones del producto, comprender claramente el comportamiento y los hábitos del usuario y descubrir los defectos del producto para facilitar la iteración y optimización de los requisitos (Chaffey & Ellis-Chadwick, 2016).

- Para el diseño, ayuda a aumentar la amabilidad de la experiencia, coincidir con las emociones del usuario, encajar delicadamente en el servicio personalizado del usuario y encontrar la falta de interacción para perfeccionar y mejorar el diseño (Chaffey & Ellis-Chadwick, 2016).

*Figura 23 Ejemplo del análisis del comportamiento del usuario*



Fuente: <https://blog.ida.cl/wp-content/uploads/sites/5/2016/08/image06.png>

### 5.2.3 Indicadores clave para analizar el comportamiento del usuario

La clave para analizar los datos del comportamiento del usuario es encontrar un indicador para medir los datos. Según el comportamiento del usuario, los indicadores múltiples se pueden subdividir en tres categorías: indicadores de rigidez, indicadores activos e indicadores de resultados (Chaffey & Ellis-Chadwick, 2016).

- **Índice de pegajosidad:** es un indicador que se utiliza para medir la frecuencia con la que los usuarios visitan un sitio web durante su ciclo de uso. Este indicador tiene varios componentes, que incluyen la cantidad y el porcentaje de nuevos usuarios, el número y porcentaje de usuarios activos, la tasa de permanencia, la tasa de abandono y la tasa de acceso (Chaffey & Ellis-Chadwick, 2016).
- **Indicadores activos:** examina principalmente la participación de visitas de usuarios, como usuarios activos, nuevos usuarios, usuarios recurrentes, usuarios perdidos, tiempo medio de estancia, frecuencia de uso, etc. (Chaffey & Ellis-Chadwick, 2016).

- **Indicadores de producto:** mide principalmente la producción de valor directo creado por los usuarios, como el número de páginas vistas, visitantes únicos, clics, frecuencia de consumo y cantidad de consumo (Chaffey & Ellis-Chadwick, 2016).

El propósito de desglosar estos indicadores es guiar la toma de decisiones operativas, lo que implica optimizar y ajustar las estrategias basándose en diferentes indicadores. En resumen, el objetivo fundamental de segmentar el análisis del comportamiento del usuario es triple: primero, aumentar la retención y conciencia del usuario; segundo, fomentar la participación y actividad del usuario; y tercero, incrementar el valor del usuario. Todo ello con el fin de fomentar la fidelidad y lealtad del usuario (Chaffey & Ellis-Chadwick, 2016).

#### 5.2.4 Implementar análisis de comportamiento del usuario

Después de determinar los indicadores de análisis del comportamiento del usuario, podemos utilizar algunos modelos para analizar los datos del comportamiento del usuario de forma cualitativa y cuantitativa (Missaoui, Abdessalem, & Latapy, 2017).

Los modelos de análisis más utilizados son:

- Análisis de eventos de comportamiento
- Análisis de retención de usuarios
- Análisis del modelo de embudo
- Análisis de la ruta de comportamiento
- Análisis del modelo de Fogg

##### 5.2.4.1 Análisis de eventos de comportamiento

El análisis de eventos de comportamiento consiste en analizar eventos específicos del usuario en función de indicadores operativos clave. Al rastrear o registrar los eventos de comportamiento del usuario, puede comprender rápidamente la tendencia de los eventos y la finalización de los usuarios (Verbeek & Slob, 2006).

Rol: principalmente para resolver quién es el usuario, de dónde viene, cuándo viene, qué hace, cómo hacerlo, y el resumen es la definición del evento a seguir. Principios: quién, cuándo, dónde, qué, cómo. Se utiliza principalmente para estudiar la influencia y el grado de ocurrencia de un determinado evento de comportamiento sobre el valor de la organización empresarial.

#### 5.2.4.2 Análisis de retención de usuarios

El análisis de retención de usuarios es un modelo utilizado para analizar la participación y actividad de los usuarios. A través de la retención y la tasa de retención, puede comprender la retención y pérdida de usuarios. Por ejemplo, utilice indicadores como la retención al día siguiente, la retención semanal y la retención mensual para medir la popularidad o la viscosidad del producto (Chaffey & Ellis-Chadwick, 2016).

La retención de usuarios generalmente se ajusta a la regla 40-20-10, es decir, la retención de nuevos usuarios al día siguiente debe ser superior al 40%, la retención semanal superior al 20% y la retención mensual superior al 10% para cumplir con los estándares comerciales. Realizamos análisis de retención de usuarios principalmente para verificar si se logran los objetivos de operación establecidos y luego afectar la siguiente decisión de producto (Chaffey & Ellis-Chadwick, 2016).

#### 5.2.4.3 Análisis del modelo de embudo

El análisis del modelo de embudo es describir la conversión del usuario y la tasa de abandono de los enlaces clave en cada etapa del uso del producto por parte del usuario. Por ejemplo, en las operaciones de actividad diaria, al determinar la tasa de abandono de cada enlace, analice cómo los usuarios abandonan, por qué y dónde lo hacen. Encuentre los enlaces que necesitan mejorar, céntrese en ellos y tome medidas efectivas para aumentar la tasa de conversión general (Doshi, Connally, Spiroff, Johnson, & Mashour, 2017).

Por lo que el análisis del modelo de embudo puede verificar si el diseño de todo el proceso es razonable. Al comparar la tasa de conversión de cada enlace, se puede encontrar que la tasa de conversión de ese enlace en la actividad de

operación no cumple con el índice esperado, para encontrar el problema y encontrar la dirección de optimización.

#### 5.2.4.4 Análisis de la ruta de comportamiento

El análisis de la ruta de comportamiento consiste en analizar la ruta de acceso del usuario durante el uso del producto. Mediante el análisis de datos de las rutas de comportamiento, los usuarios pueden encontrar las funciones y rutas de uso más utilizadas. Y desde el análisis multidimensional de la página, rastree la ruta de conversión del usuario, mejore la experiencia del usuario del producto. Ya sea que se trate de un inicio en frío de un producto o de un evento de *marketing* diario, el análisis de la ruta del comportamiento debe primero clasificar la trayectoria del comportamiento del usuario. Las trayectorias de comportamiento del usuario incluyen cognición, familiaridad, prueba, uso para la lealtad, etc. Detrás de la trayectoria están las características del usuario, que tienen un valor de referencia importante para las operaciones del producto (Chaffey & Ellis-Chadwick, 2016).

Al analizar la ruta de comportamiento del usuario, encontraremos que la ruta de comportamiento real del usuario tiene una cierta desviación de la ruta de comportamiento esperado. Esta desviación es un posible problema con el producto, y el producto debe optimizarse a tiempo para encontrar espacio para acortar el camino (Chaffey & Ellis-Chadwick, 2016).

#### 5.2.4.5 Análisis del modelo de Fogg

El modelo de Fogg ofrece un enfoque analítico para comprender por qué los usuarios se comportan de cierta manera. Se simplifica en una fórmula  $B = MAT$ , donde B representa el comportamiento, M representa la motivación, A representa la habilidad y T representa el disparador. Este enfoque sugiere que el comportamiento requiere simultáneamente motivación, habilidad y un estímulo desencadenante. En otras palabras, para que un usuario realice una acción específica, debe sentirse motivado, tener la capacidad de realizar la acción y estar expuesto a un estímulo que desencadene la acción. Este modelo se puede utilizar para evaluar la eficacia de un producto y su capacidad para lograr los objetivos previstos (Fogg, 2019).

## 5.2.5 Modelo AISAS

El modelo de análisis del comportamiento del usuario es en realidad un modelo AISAS: atención, interés, búsqueda, acción, compartir, también afectan las decisiones de comportamiento del usuario (Jun, y otros, 2021).

### 5.2.5.1 *Attention*

Atención significa que, si queremos obtener un determinado rendimiento, primero debemos llamar la atención de los usuarios. Si no hay usuarios, todas las actividades de *marketing* posteriores serán inútiles. Para atraer la atención de los usuarios, podemos partir de muchos aspectos, como a través del *marketing* interactivo para atraer el consumo del servicio (Jun, y otros, 2021).

### 5.2.5.2 *Interest*

Después de atraer usuarios, si queremos retener realmente a estos usuarios, debemos dejar que los usuarios tengan cierto interés en nuestros productos y hacer que quieran adquirir nuestros productos desde el fondo de su corazón. Esto requiere que llevemos a cabo cierta investigación de mercado sobre el grupo objetivo de antemano para comprender su interés (Jun, y otros, 2021).

### 5.2.5.3 *Search*

Cuando el grupo objetivo tiene cierto interés en nosotros, pueden recopilar información sobre nuestros productos a través de algunos canales en línea o fuera de línea. Esta etapa es la etapa de búsqueda. Si desea que los usuarios dejen una mejor impresión de nosotros, debe prestar atención al mejoramiento de los buscadores en línea, optimizar los servicios fuera de línea y mejorar la reputación (Jun, y otros, 2021).

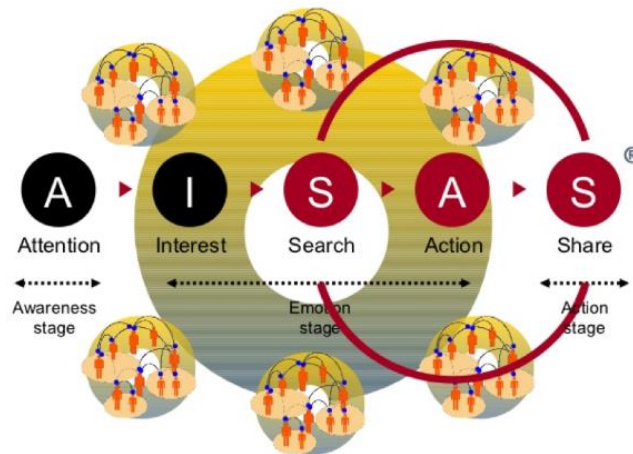
### 5.2.5.4 *Action*

Si el usuario queda satisfecho con los productos de la empresa después de una serie de encuestas, pasará directamente al consumo. En esta etapa, el vínculo más importante para promover las transacciones de pedidos es el vínculo de inscripción (Jun, y otros, 2021).

### 5.2.5.5 Share

Si un usuario usa los productos de la empresa para obtener una mejor experiencia, puede compartirlo con las personas que lo rodean y recomendar los productos de la empresa a quienes lo rodean. Esto también se denomina boca a boca. Debemos prestar atención al importante papel de la comunicación de boca en boca, y su poder persuasivo puede acabar con todas las actividades de *marketing* en segundos (Jun, y otros, 2021).

Figura 24 Modelo AISAS



Fuente: <https://www.marketing2.ca/wp-content/uploads/2018/10/AISAS-model.png>

## 5.3 Representación de los datos de entrada

En el diseño de algoritmos genéticos es de suma importancia representar los datos, considerando no solamente la representación de la solución o de los cromosomas, además de los datos de entrada. Estos datos a menudo se encuentran en bases de datos relacionales, pero se deben preparar estructuras matriciales antes de ejecutar el algoritmo genético para una recuperación y uso óptimos. Dichas estructuras corresponden con los datos de entrada esenciales del algoritmo genético (Wirsansky, 2020).



Tabla 8 Datos de entrada del algoritmo

| Variable                         | Descripción   |
|----------------------------------|---|
| Tiempo diario pasado en el sitio | Tiempo del consumidor en el sitio en minutos                                      |
| Edad                             | Edad del usuario en años  |
| Ingresos del área                | Promedio. Renta del área geográfica del consumidor                                |
| Uso diario de Internet           | Promedio. minutos al día que el consumidor está en Internet                       |
| Ciudad                           | Ciudad del consumidor   |
| Genero                           | Siendo 1 Masculino y 0 Femenino   |
| País                             | País del consumidor   |
| Marca de tiempo                  | Tiempo cuando el consumidor hizo clic sobre el anuncio o sobre la ventana cerrada |
| Clic en el anuncio               | Siendo 1 hizo clic en el anuncio y 0 no hizo clic en el anuncio                   |

Fuente: elaboración propia

Tabla 9 Rango de datos del algoritmo

|                 | Tiempo diario pasado en el sitio | Edad     | Ingresos del área | Uso diario de Internet | Genero  | Clic en el anuncio |
|-----------------|----------------------------------|----------|-------------------|------------------------|---------|--------------------|
| <b>Cantidad</b> | 1000                             | 1000     | 1000              | 1000                   | 1000    | 1000               |
| <b>Media</b>    | 65.0002                          | 36.009   | 55000.00008       | 180.0001               | .481    | .5                 |
| <b>Std</b>      | 15.853615                        | 8.785562 | 13414.634022      | 43.902339              | .499889 | .50025             |
| <b>Min</b>      | 32.6                             | 19       | 13996.50          | 104.78                 | 0       | 0                  |
| <b>25%</b>      | 51.36                            | 29       | 47031.8025        | 138.83                 | 0       | 0                  |
| <b>50%</b>      | 68.215                           | 35       | 57012.30          | 183.13                 | 0       | .5                 |
| <b>75%</b>      | 78.5475                          | 42       | 65470.6350        | 218.7925               | 1       | 1                  |
| <b>Max</b>      | 91.43                            | 61       | 79484.80          | 269.96                 | 1       | 1                  |

Fuente: elaboración propia

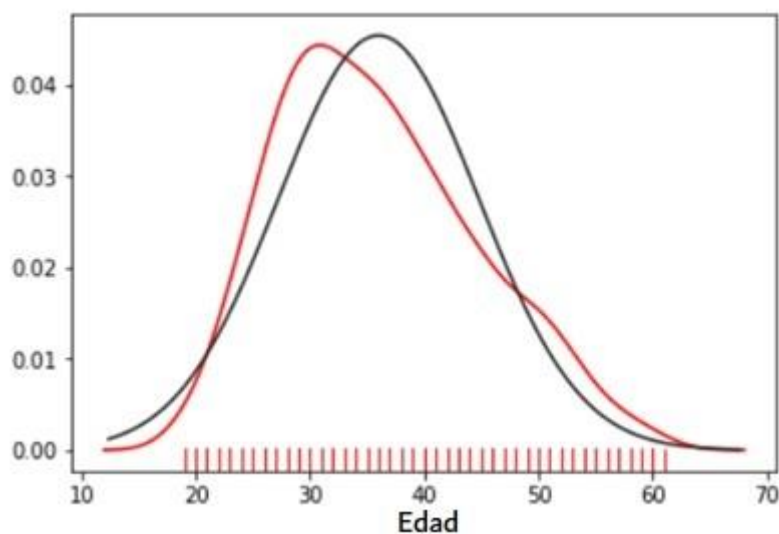
Un dato valioso de la tabla 9 es que el ingreso más pequeño es de \$ 13996.50 y un ingreso máximo del área de \$ 79484.80. Eso quiere decir que las personas que visitan el sitio son de distintas clases sociales. Además, los usuarios pasan de 32 a 91 minutos en el sitio web en una sesión, por lo que podemos concluir que es un sitio web popular.

Además, se puede observar que la edad promedio de los visitantes es de 36 años y que la edad más baja registrada es de 19 años y la más alta es de 61 años. Por lo tanto, se podría inferir que el sitio web está dirigido principalmente a una audiencia adulta. En cuanto a la distribución de género, se puede observar

que la proporción de mujeres y hombres es bastante similar, con un 52% de mujeres.

Para analizar mejor los datos, primero dibujemos un histograma con la estimación de la densidad del Kernel (“La Densidad de Estimación de Kernel, consiste en una técnica no paramétrica de la estadística multivariada usada para estimar la probabilidad de la función de densidad de una variable aleatoria. En muchos casos puede ser vista como una generalización de un histograma, pues, maneja datos concentrados en una región muy pequeña.” (Valbuena, 2018, p.317)) para la variable **Edad**, como puede verse en la Gráfica 1.

Gráfica 1 Edad

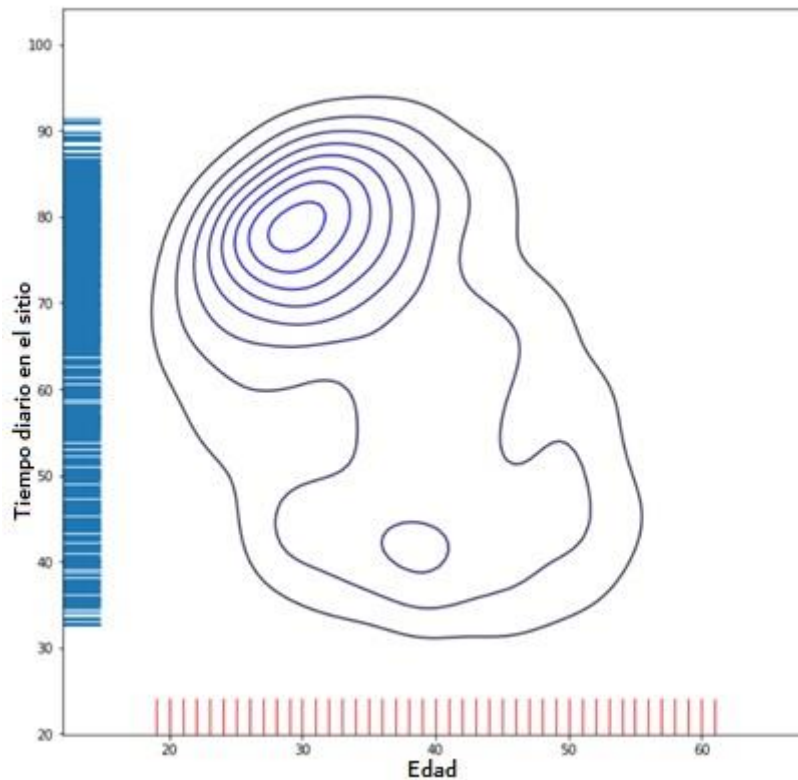


Fuente: elaboración propia

La gráfica 1 muestra que la variable **edad** presenta una distribución normal de los datos. Analizaremos por qué esto es adecuado para un tratamiento de datos eficaz.

A continuación, se muestra un diagrama de densidad de dos dimensiones para determinar la correlación entre las dos variables. En la gráfica 2 se observa cómo se relaciona las variables de la **edad** con el **tiempo pasado en el sitio**.

Gráfica 2 Edad y tiempo que pasan en el sitio



Fuente: elaboración propia

De la Gráfica 2, se puede deducir que los usuarios de menor edad dedican más tiempo al sitio web. Esto supone que el principal grupo objetivo al que podría ir dirigido las campañas de *marketing* serían los usuarios de 20 a 40 años de edad. En el caso de tener un producto pensado para gente de mediana edad, este es el lugar perfecto para promocionar. De manera contraria, es un error anunciar un producto que estuviera dirigido a adultos mayores de 60 años en este sitio.

A partir de la información se puede obtener una visión clara de las características de los usuarios que hacen clic en los anuncios, lo que permite realizar diversos estudios adicionales.

## 5.4 Modelo del Algoritmo

La codificación en el cromosoma es un aspecto crítico que influye en la eficacia de un algoritmo genético.

Se ha seleccionado utilizar un cromosoma no binario al diseñar este algoritmo, lo que implica la codificación directa de cada parámetro con valores enteros o reales. Este enfoque ha sido adoptado porque permite entender mejor comprensión el problema (Sivanandam & Deepa, 1998).

### 5.4.1 Población inicial o primera generación

Es posible crear la población inicial utilizando una heurística o basándose de una solución existente al problema, pero en general, la manera más común de generarla es al azar. (Sivanandam & Deepa, 1998).

En esta situación, se opta por una selección aleatoria de la población inicial o primera generación, lo que aumenta la probabilidad de que la calidad de la población inicial sea baja y obstaculice la convergencia rápida del algoritmo. Generar la población inicial mediante una heurística puede resultar en un rendimiento degradado del algoritmo y en la necesidad de invertir tiempo valioso en la búsqueda de soluciones iniciales óptimas. (Sivanandam & Deepa, 1998).

La asignación de la primera generación se basa en las siguientes consideraciones:

- Es necesario asignar a cada sección un sitio que contenga todas las páginas correspondientes.
- Los bloques adyacentes que pertenecen a una misma sección deben ser asignados al mismo sitio para que los usuarios no se cambien del sitio.
- Se asignan sólo ligas sin considerar archivos en PDF.
- Para una sección se considera como máximo un nivel de **profundidad de 6** niveles por sitio.

### 5.4.2 Conservación de las soluciones óptimas o de mayor calidad

Se utiliza una técnica de cruce destructiva en la que se aceptan nuevos individuos, aunque su función objetivo no sea mejor que la de sus progenitores. A pesar de ello, se asegura que los individuos con mejor función objetivo se transmitan a las siguientes generaciones mediante la selección de un porcentaje de los mejores individuos, los cuales son transferidos directamente a la siguiente generación, aunque con una probabilidad inicialmente establecida en cero o muy baja, para así introducir variabilidad en la población. (Sivanandam & Deepa, 1998).

#### 5.4.3 La elección de progenitores o padres

La estrategia de selección adoptada es el torneo determinístico, la cual implica elegir al azar dos individuos y comparar sus valores de función objetivo. El individuo con el mejor valor se selecciona como "padre", mientras que se hace lo mismo para seleccionar una "madre". De esta manera, se realiza la reproducción entre los dos individuos seleccionados. El algoritmo asegura que cada pareja de individuos se cruce sólo una vez. El único individuo excluido de la reproducción es el peor de su generación en términos de función objetivo, ya que siempre perdería en el torneo. A pesar de eso, esta técnica de selección no ofrece una garantía de que los individuos más destacados sean elegidos para crear la siguiente generación (Sivanandam & Deepa, 1998).

#### 5.4.4 Técnica de cruce destructiva con mutación aleatoria

La técnica de cruce elegida para la reproducción en el algoritmo se basa en la selección de un solo punto, y es considerada destructiva. Esto significa que un punto aleatorio es seleccionado en los padres, y sus dos secciones son combinadas para formar dos hijos que son añadidos a la población, independientemente de si son o no mejores que sus padres en términos de la función objetivo (Wirsansky, 2020).

Aunque el punto de cruce es seleccionado al azar, se tiene en cuenta evitar la división de la planificación de una sección, ya que esto podría generar ligas con una cantidad superior o inferior a la requerida (Wirsansky, 2020).

La mutación es una operación que se realiza seleccionando aleatoriamente un factor de mutación entre 1 y 100. Si el factor es menor o igual al valor mínimo de probabilidad de mutación definido, se modifica aleatoriamente el gen correspondiente a esa sección, incluyendo cambios en el sitio, la página y la liga. Es importante destacar que el gen puede quedar con los mismos parámetros que tenía inicialmente. Este proceso se repite para cada uno de los genes de un cromosoma, tanto para los cromosomas que provienen de una crucea como para aquellos que se preservan de la generación anterior (Wirsansky, 2020).

#### 5.4.5 Ajustes del algoritmo

Hay varios parámetros significativos en el algoritmo genético que necesitan ser ajustados en base a los resultados obtenidos durante la ejecución del algoritmo (Wirsansky, 2020).

Se establece como un parámetro el tamaño inicial de la población, que permanece constante en las generaciones siguientes. Inicialmente, se fija un valor de 8, lo que significa que se crearán ocho cromosomas en cada generación. Sin embargo, el algoritmo es adaptable a esa cantidad y permite ajustarla según las necesidades (Wirsansky, 2020).

Inicialmente, la probabilidad de mutación se establece en un 2%, lo que significa que solamente un 2% de los cromosomas mutará, cambiando de manera aleatoria una sección de sitio, página y liga. No obstante, el parámetro de probabilidad puede ser ajustado de acuerdo a los resultados que se obtengan durante la ejecución del algoritmo.

Se establece un parámetro en el algoritmo genético para la cantidad de generaciones a iterar, lo cual también es una condición de terminación del algoritmo. Esto significa que el algoritmo finaliza después de completar la cantidad de generaciones especificadas, sin importar la función objetivo obtenida. Se establece un límite máximo de iteración de 200 generaciones al inicio.

La segunda condición de finalización se define por un valor objetivo de la función, el cual se establece como un porcentaje de 100% al principio. Esto significa que no debe haber colisiones de enlaces entre páginas y sitios web. Sin embargo, en ejecuciones del algoritmo con muchas ligas o en condiciones muy específicas,

como un número limitado de sitios o páginas, este valor objetivo se puede disminuir a un 90% o incluso menos según sea necesario.

---

## Capítulo 6. Aplicación del modelo de algoritmo en un estudio de caso.

---

### Introducción

En este capítulo, se lleva a cabo la aplicación práctica del modelo de algoritmo, utilizando el proceso CRISP-DM como guía en la ciencia de datos. Se describen las acciones realizadas en el algoritmo, desde la selección de la población inicial hasta la evaluación de los resultados obtenidos. Este capítulo demuestra cómo el enfoque propuesto en los capítulos anteriores se materializa en un estudio de caso real, mostrando la relevancia y el impacto de la optimización del comportamiento del usuario en un sitio web mediante el uso de la ciencia de datos y algoritmos genéticos.

La comprensión del comportamiento del usuario en línea se ha convertido en una práctica esencial para empresas e instituciones que desean mejorar su presencia digital y mejorar la experiencia del usuario. En este capítulo, se presentará un estudio de caso enfocado en la página web de la Facultad de Contaduría y Administración de la UNAM ([www.fca.unam.mx](http://www.fca.unam.mx)) donde se aplicará el modelo de algoritmo genético y ciencia de datos en Python para analizar el comportamiento del usuario. Este análisis permitirá entender mejor cómo interactúan los usuarios con la página y qué elementos pueden mejorarse para optimizar la experiencia del usuario y aumentar el compromiso en la plataforma. Además, se demostrará cómo la combinación de técnicas de algoritmos genéticos y ciencia de datos en Python pueden utilizarse para abordar problemas complejos de manera efectiva y eficiente.

**Para dimensionar la magnitud de este análisis, consideremos el hecho de que estamos tratando con un total de 122,193 ligas distribuidas en 21 sitios web, si calculamos el total de combinaciones que podríamos obtener al mezclar estos sitios con sus diferentes enlaces, nos encontramos con un**



número asombroso de posibilidades  $21^{122193}$ , esto es equivalente a  $10^{10^{5.208}}$ , lo que equivale a un “1” seguido de 161,435 ceros. Esta cifra es extraordinariamente grande y subraya la importancia de emplear la ciencia de datos para gestionar y analizar esta abrumadora cantidad de información.

El reto que enfrentamos a lo largo del proyecto radica en tener que considerar las acciones del usuario: tener que considerar las interacciones con las ligas, el comportamiento del usuario para ver qué es lo que le interesa y a qué páginas realmente está entrando. Se tienen 122,193 ligas y también se cuentan con 21 sitios web, Es decir que por cada sitio web que tenemos podemos ocupar todas las ligas, para mostrar la combinación de todas estas ligas utilizaremos la fórmula de la combinatoria de variación con repetición, se aplica cuando se quiere contar el número de maneras diferentes en las que se pueden organizar elementos con repetición.

La fórmula de la variación con repetición es:

$$V(n, r) = n^r$$

Donde n es el número de elementos y r es el número de grupos que se forman. En este caso,  $n=21$  (sitios web) y  $r=122,193$  (ligas). Aplicando la fórmula:

$$V(21, 122193) = 21^{122193}$$

Este número es extremadamente grande y difícil de conceptualizar debido a su magnitud. La cantidad de subconjuntos posibles que se pueden formar usando la variación con repetición es igual a  $21^{122193}$ , lo cual es un número colosal.

Por lo tanto, usando la fórmula de la variación con repetición, hay  $21^{122193}$ , subconjuntos posibles de ligas que se pueden formar en los 21 sitios web, lo que subraya la complejidad de esta tarea.

## 6.1 Algoritmo

Se sigue una secuencia de pasos para la estructura del algoritmo genético utilizado (Mitchell, 1998):

1. Generar una población inicial de **P** individuos seleccionados aleatoriamente, cada uno representando un conjunto de valores variables.
2. Calcular el valor de aptitud (*fitness*) de cada uno de los individuos de la población, que se encuentra relacionado con el valor objetivo de la función. Si el valor de la función es mayor, el *fitness* del individuo también será mayor.
3. Se deben realizar los siguientes pasos de forma repetitiva hasta que se creen **P** nuevos individuos y genere una nueva población vacía.
  - 3.1 Elegir dos miembros de una población determinada, cuya probabilidad de elección esté en proporción directa a su aptitud (*fitness*).
  - 3.2 Combinar los rasgos de los dos individuos seleccionados mediante una técnica de cruzamiento para producir nueva descendencia. (*crossover*).
  - 3.3 Aplicar un cambio aleatorio en el material genético del individuo seleccionado, a fin de incrementar la diversidad de la población y prevenir la convergencia temprana hacia una solución subóptima.
  - 3.4 Agregar este nuevo individuo en la población creada.
4. Sustituir la población antigua con una nueva
5. En caso de que la condición de finalización no se cumpla, regresar al paso 2.

### 6.1.1 Población

Dentro del ámbito de los algoritmos genéticos, el concepto de **individuo** se utiliza para hacer referencia a cualquier solución potencial que se esté buscando para el problema en cuestión. Cada individuo representa una posible

combinación de valores de variables que puede llevar a la optimización de la función objetivo, ya sea maximizándola o minimizándola. Para expresar estas combinaciones, es posible emplear un vector (arreglo) que tenga una dimensión igual al número total de variables y un valor asignado en cada posición. Supongamos que la función objetivo  $J(x,y,z)$  tiene una dependencia en las variables  $x$ ,  $y$ , y  $z$ . La combinación de valores  $x=3.0$ ,  $y=9.5$ , y  $z=-0.5$  corresponde a uno de los posibles valores de la función objetivo (Mitchell, 1998).

### 6.1.2 Aptitud

En el contexto de los algoritmos genéticos, se requiere medir la habilidad de cada individuo en la población para resolver el problema que se está tratando de optimizar, lo que se conoce como **aptitud**. La relación entre la aptitud y la función objetivo  $f$  puede variar dependiendo de si el problema a resolver es de maximización o minimización (Sivanandam & Deepa, 1998):

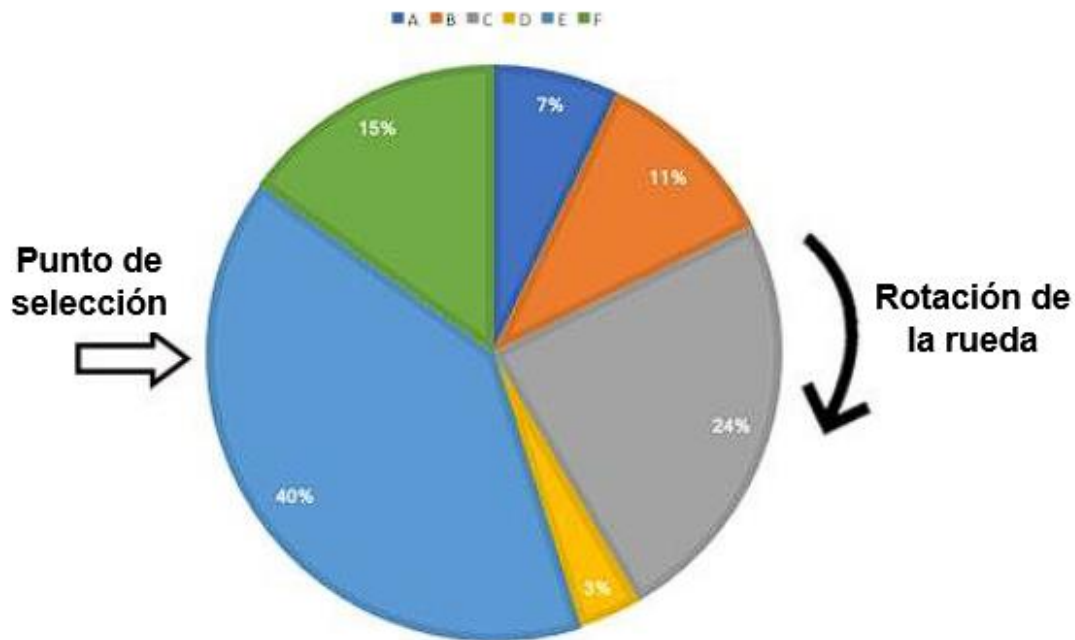
- Maximización: cuanto mayor sea la aptitud del individuo, mayor será el valor de la función objetivo  $f(\text{individuo})$ .
- Minimización: los individuos se ajustan mejor con valores más bajos de la función objetivo  $f(\text{individuo})$  o, de manera equivalente, se ajustan menos con valores más altos de la función objetivo. La aptitud del problema de minimización se puede calcular como  $-f(\text{individuo})$  o  $1/1+f(\text{individuo})$ , ya que el algoritmo genético selecciona a los elementos con la mayor aptitud.

### 6.1.3 Selección de individuos

La forma en que se eligen los individuos para participar en el proceso de cruzamiento varía en función de la implementación específica del algoritmo genético. En general, todo el mundo tiende a elegir a los elementos mejores (mayor aptitud). Algunas de las estrategias más comunes incluyen (Sivanandam & Deepa, 1998):

- Método de la ruleta: consiste en seleccionar a los individuos de la población con una probabilidad proporcional a su aptitud relativa. Esta aptitud relativa se calcula dividiendo la aptitud del individuo por la suma de las aptitudes de todos los individuos de la población. Si alguien es el doble de apto que otra persona, es más probable que sea elegido. Este enfoque puede ser problemático si algunos individuos tienen una aptitud física mucho más alta que otros (múltiples rangos), porque estos individuos serán reelegidos y la gran mayoría de los individuos de la siguiente generación serán "hijos" del mismo "padre". (No ha cambiado mucho) (Sivanandam & Deepa, 1998).

*Figura 25 Ruleta selección ejemplo*



Fuente: (elaborado con base en Wirsansky, 2020, p.29)

- Método de clasificación: después de clasificar a todos los individuos de mayor a menor aptitud, la selección de un individuo es menos probable a medida que su posición en la clasificación aumenta. Este enfoque es menos extremo que el método de la ruleta, ya que la diferencia entre la aptitud más alta y la más baja es mucho menor en comparación con otros métodos. (Sivanandam & Deepa, 1998).
- Selección competitiva (Torneo): dos parejas de individuos (ambos con igual probabilidad) se seleccionan aleatoriamente de la población. De

cada par, elige el par con la aptitud más alta. Por último, los dos finalistas se comparan y se escoge aquel que presente una aptitud superior. En comparación con los dos primeros métodos anteriores, podemos decir que este método tiene una tendencia a producir una distribución de probabilidades de selección más uniforme. (Sivanandam & Deepa, 1998).

*Figura 26 Ejemplo de selección por torneo con un tamaño de torneo de tres*

| Individual | Fitness |
|------------|---------|
| A          | 8       |
| B          | 12      |
| C          | 27      |
| D          | 4       |
| E          | 45      |
| F          | 17      |

Fuente: (Wirsansky, 2020, p.35)

Explicando el ejemplo de la figura 26, se observa que tenemos seis individuos y con los valores de aptitud que se muestran en la misma figura 26. Por lo que la figura ilustra la selección aleatoria de tres de ellos (A, B y F), luego se anuncia a F como el ganador, ya que tiene el valor de aptitud más alto (17) entre estos tres individuos.

- Selección truncada: se realiza una selección aleatoria de individuos y los n individuos con la menor aptitud de la población truncada son descartados en primer lugar.

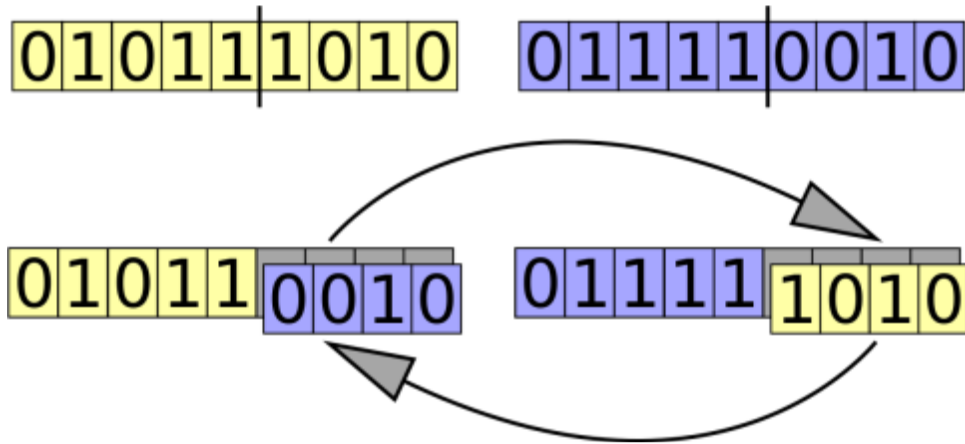
#### 6.1.4 Cruce de dos individuos (recombinación)

El propósito de esta fase es crear nuevos individuos (descendencia) a partir de individuos existentes (padres), con el objetivo de combinar las características de los individuos previos. Esta es otra etapa del algoritmo donde puedes seguir diferentes estrategias (Wirsansky, 2020). Los tres más utilizados son:

- Cruzamiento de un solo punto: implica seleccionar aleatoriamente una posición para ser utilizada como punto de intersección. Durante este proceso de cruce, se divide a cada padre en dos mitades y se

intercambian entre sí. Esta operación genera dos nuevos individuos en cada cruce realizado (Wirsansky, 2020).

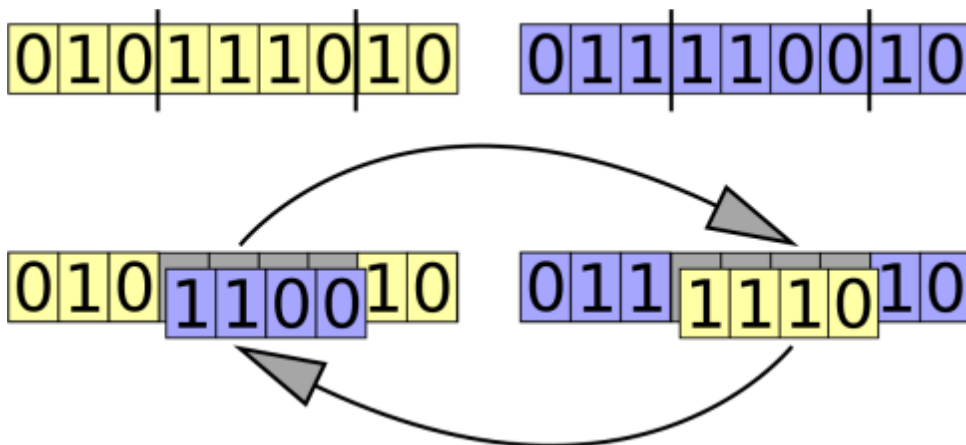
*Figura 27 Punto único Transversal ejemplo*



Fuente: (Wirsansky, 2020, p.37)

- Cruzamiento de varios puntos: se escogen ubicaciones de manera aleatoria como puntos de intersección. Cada padre se divide en secciones por medio de puntos de interrupción y estas secciones se intercambian para crear dos nuevos individuos por cada cruce. Este proceso da como resultado dos nuevos individuos para cada cruce (Wirsansky, 2020).

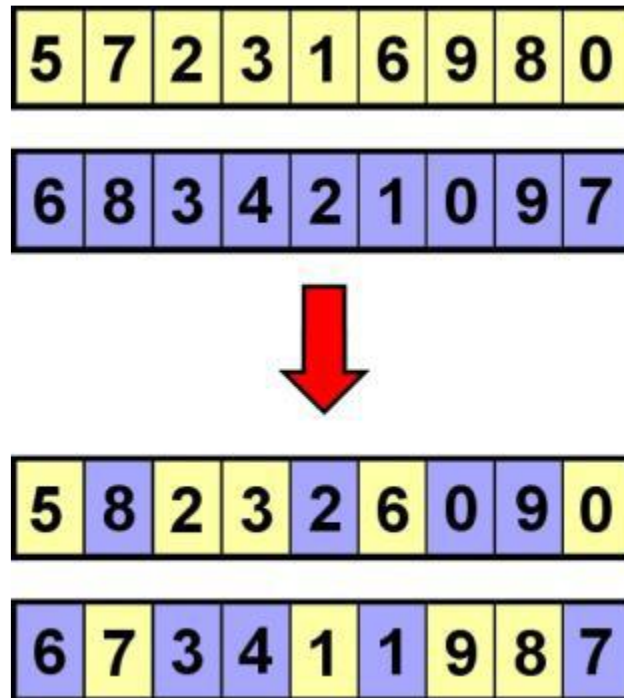
*Figura 28 Dos puntos Transversal ejemplo*



Fuente: (Wirsansky, 2020, p.38)

- Cruzamiento uniforme: se asigna un valor en cada posición del nuevo individuo tomando como fuente uno de los dos padres. En líneas generales, en esta estrategia de cruce, cada valor del nuevo individuo tiene igual probabilidad de ser heredado de cualquiera de los dos padres, aunque puede haber otros factores que afecten esta probabilidad, como la aptitud del individuo. A diferencia de la estrategia anterior, cada cruce con esta estrategia produce descendencia (Wirsansky, 2020).

*Figura 29 Uniforme Transversal ejemplo*



Fuente: (Wirsansky, 2020, p.38)

## 6.5 Mutar individuo

Después de que se genera cada nuevo individuo descendiente, se aplica un proceso de mutación que implica la posibilidad de que cada posición en el individuo mute con una probabilidad de  $pp$ . La mutación es una etapa importante del proceso ya que ayuda a aumentar la variabilidad y evita que el algoritmo se quede atrapado en mínimos locales. Esto se debe a que los individuos de una generación a otra son muy similares y la mutación introduce cambios en el material genético.

Se puede manejar la intensidad del cambio que puede causar una mutación mediante diferentes estrategias:

- Distribución uniforme: esta estrategia implica que la mutación de una posición determinada se logra sumando un valor aleatorio tomado de una distribución uniforme (como una distribución entre -1 y +1) al valor original de la posición (Kramer, 2017).
- Distribución normal: en este método, para mutar la posición  $i$  se añade al valor actual de dicha posición un número obtenido de una distribución normal, la cual tiene una media en 0 y una desviación estándar específica. La magnitud del cambio generado por la mutación dependerá de la desviación estándar de la distribución normal utilizada. Una mayor desviación estándar aumenta la probabilidad de que la mutación produzca cambios grandes (Wirsansky, 2020).
- Aleatorio: se realiza la mutación en la posición  $i$  mediante la sustitución del valor actual por un valor aleatorio que se encuentra dentro del rango permitido para esa variable. En general, esta técnica resulta en cambios más significativos que las dos estrategias anteriores (Wirsansky, 2020).

Es fundamental tener en cuenta que las mutaciones pueden hacer que un valor que estaba dentro del rango permitido salga de él. Para evitar esto, se puede utilizar una estrategia en la que, si el valor resultante de la mutación supera alguno de los límites establecidos, se reemplaza por el valor límite correspondiente. De esta forma, se otorga la posibilidad de que los valores se desvíen hasta un límite máximo establecido (Wirsansky, 2020).

## 6.3 Implementación del Algoritmo

### 6.3.1 Evaluar la aptitud del comportamiento del usuario

La aptitud de un individuo representa qué tan bueno es en la tarea que se está resolviendo. En el caso del ejemplo de modelar el comportamiento de un usuario en la página web de FCA UNAM, la aptitud podría medir qué tan rápido o eficiente es el comportamiento del usuario para encontrar la información que busca.



Para evaluar la aptitud de cada individuo en la población, se utiliza la función de aptitud. En general, esta función toma como entrada el cromosoma (es decir, la solución propuesta por el individuo) y devuelve un valor numérico que representa la aptitud del individuo.

La función de aptitud puede variar según la tarea que se esté resolviendo y cómo se defina la aptitud en ese contexto específico. Por ejemplo, en un problema de optimización de funciones, la aptitud podría medir la cercanía de la solución propuesta al valor óptimo de la función. (Wirsansky, 2020)

*Código 1 Captura de código de la función evaluar la aptitud del comportamiento del usuario en el editor de Sublime Text*

```
def evaluar_aptitud(cromosoma):
    """Evaluar la aptitud de un cromosoma."""
    # Definir una función que simule el comportamiento del usuario en la página web
    tiempo_en_pagina = 0
    paginas_visitadas = 0
    for accion in cromosoma:
        if accion == "ir_a_pagina":
            tiempo_en_pagina += random.randint(5, 30) # simular el tiempo que el usuario pasa en una página
            paginas_visitadas += 1
        elif accion == "ver_pagina":
            tiempo_en_pagina += random.randint(10, 60)
            if random.random() < 0.1:
                return tiempo_en_pagina * paginas_visitadas * 10
            # Aptitud es proporcional al tiempo, páginas visitadas y la tasa de conversión
        else:
            return tiempo_en_pagina * paginas_visitadas
    return tiempo_en_pagina * paginas_visitadas
```

Fuente: elaboración propia

Tal como se muestra en código 1 la función `evaluar_aptitud` toma como entrada un cromosoma (es decir, una lista de acciones) y simula el comportamiento del usuario en la página web de FCA UNAM según las acciones del cromosoma. En la función se utiliza un ciclo `for` para iterar sobre las acciones del cromosoma y se utilizan estructuras condicionales para determinar qué acción se debe realizar en cada iteración.

Para simular el comportamiento del usuario, la función utiliza la biblioteca `random` de Python para generar números aleatorios que representan el tiempo que el usuario pasa en cada página web. En cada iteración del ciclo `for`, se acumula el tiempo que el usuario pasa en la página web y se cuenta el número de páginas visitadas. Se devuelve un valor numérico que representa la aptitud del cromosoma basado únicamente en la cantidad de tiempo que el usuario pasó en la página web y el número de páginas visitadas.

La evaluación de la aptitud es importante en un algoritmo genético ya que es la medida de cuán bien un cromosoma se ajusta a la solución óptima del problema.

En este caso, la aptitud mide cuánto tiempo pasa el usuario en la página web de FCA UNAM y cuántas páginas visita, lo cual se considera un indicador del éxito del usuario en la navegación por la página web. La función evaluar\_aptitud es esencial en el proceso de selección de los cromosomas más aptos para la reproducción del algoritmo genético.

### 6.3.2 Función de selección por ruleta

En este paso, se utiliza la función de evaluación de aptitud definida en el anterior paso para clasificar todos los cromosomas en función de su aptitud. Los cromosomas con una aptitud más alta se consideran más aptos y tienen más probabilidades de ser seleccionados para la reproducción.

Existen diferentes técnicas de selección que se pueden utilizar en un algoritmo genético, pero una de las más comunes es la selección por ruleta o "*roulette wheel selection*". Esta técnica se basa en el hecho de que la probabilidad de ser seleccionado es proporcional a la aptitud del cromosoma. En resumen, la probabilidad de que un cromosoma sea seleccionado para la reproducción es mayor si tiene una mayor aptitud en comparación con otros cromosomas de la población.

Código 2 Captura de código de la función de selección por ruleta en el editor de Sublime Text

```
# Función de selección por ruleta
def seleccion(poblacion, aptitudes):
    total_aptitudes = sum(aptitudes)
    ruleta = [aptitud/total_aptitudes for aptitud in aptitudes]
    seleccionados = []
    for i in range(len(poblacion)):
        r = random.random()
        suma = 0
        for j in range(len(ruleta)):
            suma += ruleta[j]
            if r <= suma:
                seleccionados.append(poblacion[j])
                break
    return seleccionados

# Evaluar la aptitud de cada cromosoma en la población
aptitudes = [aptitud(cromosoma) for cromosoma in poblacion]

# Seleccionar los cromosomas para la reproducción
seleccionados = seleccion(poblacion, aptitudes)
```

Fuente: elaboración propia

En código 2 la función selección (población, aptitudes) toma como entrada una lista de cromosomas población y una lista de valores de aptitud correspondientes a cada cromosoma en la población aptitudes. La función primero calcula la suma de todas las aptitudes y luego construye una lista llamada ruleta en la que cada elemento es la aptitud del cromosoma dividida entre la suma total de aptitudes. Esto normaliza las aptitudes para que sumen 1 y las convierte en probabilidades para la selección por ruleta.

La función luego itera a través de cada cromosoma en la población y selecciona uno al azar utilizando la ruleta. La probabilidad de selección de un cromosoma es proporcional a su aptitud relativa en la población. La función agrega los cromosomas seleccionados a una lista seleccionados y los devuelve al final.

Después de definir la función de selección, se evalúa la aptitud de cada cromosoma en la población utilizando la función aptitud(cromosoma). Luego, se llama a la función de selección para seleccionar los cromosomas para la reproducción. La lista resultante de cromosomas seleccionados se guarda en la variable seleccionados

Resumiendo, la idea detrás de la selección es que los cromosomas más aptos deben tener una mayor probabilidad de ser seleccionados para la reproducción, ya que se espera que sus características se transmitan a las generaciones futuras y mejoren la población, es decir, la selección es el proceso de elegir a los mejores cromosomas para la reproducción y descartar a los peores.

### 6.3.3 Función de cruce y mutación

Este paso del algoritmo consiste en realizar el cruce y la mutación de los cromosomas seleccionados en el paso anterior para crear nuevos cromosomas. El cruce es la mezcla de dos cromosomas seleccionados para crear dos nuevos cromosomas hijos, mientras que la mutación es un proceso que se realiza en cada cromosoma con una probabilidad definida. La mutación consiste en cambiar una acción por otra acción aleatoria. Estos procesos se realizan para explorar nuevas soluciones y aumentar la diversidad genética en la población. Luego de realizar el cruce y la mutación, se crea una nueva población de cromosomas, que reemplaza a la población anterior para ser evaluada en la siguiente generación.

*Código 3 Captura de código de la función de cruce y mutación en el editor de Sublime Text*

```
# Función de cruce
def cruce(padre, madre):
    punto_corte = random.randint(1, len(padre)-1)
    hijo1 = padre[:punto_corte] + madre[punto_corte:]
    hijo2 = madre[:punto_corte] + padre[punto_corte:]
    return hijo1, hijo2

# Función de mutación
def mutacion(cromosoma):
    for i in range(len(cromosoma)):
        if random.random() < PROB_MUTACION:
            cromosoma[i] = random.choice(["ir_a_inicio", "ir_a_portal_profesor", "ir_a_investigación", "ir_a_licenciaturas"])
    return cromosoma
```

Fuente: elaboración propia

Al analizar código 3 se puede resaltar lo siguiente:

La función de cruce toma dos padres como argumentos y selecciona aleatoriamente un punto de corte en el cromosoma. Luego, mezcla los genes de ambos padres a partir del punto de corte para crear dos hijos. El primer hijo es creado tomando la primera parte del padre y la segunda parte de la madre, mientras que el segundo hijo es creado tomando la primera parte de la madre y la segunda parte del padre. Finalmente, la función devuelve los dos hijos.

La función de mutación toma un cromosoma como argumento y recorre cada uno de sus genes. Para cada gen, se comprueba si debe ser mutado comparando su probabilidad de mutación con un valor aleatorio. Si se cumple la condición, el gen se muta y se selecciona una acción aleatoria entre cuatro posibles acciones ("ir\_a\_inicio", "ir\_a\_portal\_profesor", "ir\_a\_investigación", "ir\_a\_licenciaturas"). La función devuelve el cromosoma mutado.

#### 6.3.4 Reemplazar la población anterior con la nueva población

Este paso del algoritmo consiste en reemplazar la población anterior con la nueva población generada a través de la reproducción y la mutación de los cromosomas seleccionados. Este proceso de reemplazo se puede realizar de varias formas.

Una de las formas más comunes de reemplazo es el reemplazo generacional. En este enfoque, la nueva población completa reemplaza a la población anterior. Es decir, todos los individuos de la población anterior son descartados y se reemplazan por los individuos generados en la nueva población. Este enfoque asegura que la nueva población tenga la misma cantidad de individuos que la población anterior.

En general, la forma en que se realiza el reemplazo de la población depende de los objetivos del problema que se está resolviendo y de los parámetros del algoritmo genético. Es importante tener en cuenta que el reemplazo debe permitir una diversidad adecuada en la población para evitar que se estanque en soluciones subóptimas.

#### Código 4 Captura de código de la función de nueva población en el editor de Sublime Text

```
# Generar la nueva población
nueva_poblacion = []
while len(nueva_poblacion) < TAM_POBLACION:
    # Seleccionar los padres
    padres = random.sample(seleccionados, 2)
    # Realizar el cruce
    hijos = cruce(padres[0], padres[1])
    # Mutar los hijos
    hijos_mutados = [mutacion(hijo) for hijo in hijos]
    # Agregar los hijos mutados a la nueva población
    nueva_poblacion.extend(hijos_mutados)

# Reemplazar la población anterior por la nueva
poblacion = nueva_poblacion
```

Fuente: elaboración propia

Al analizar el código 4 de la función se llega a lo siguiente:

En este código, se crea una lista vacía llamada "nueva\_poblacion" donde se almacenarán los nuevos cromosomas generados. Luego, se utiliza un bucle *while* para repetir el proceso de selección, cruce y mutación hasta que la nueva población tenga el tamaño especificado en el parámetro TAM\_POBLACION.

Dentro del bucle, se utilizan las funciones de selección, cruce y mutación previamente definidas para generar dos hijos a partir de dos padres seleccionados aleatoriamente de la lista de "seleccionados". Los hijos resultantes se mutan utilizando la función de mutación y se agregan a la lista de "nueva\_poblacion".

Finalmente, una vez que se ha generado la nueva población completa, se reemplaza la población anterior por la nueva población utilizando el operador de asignación "=". Es decir, la variable "poblacion" ahora hace referencia a la nueva lista de cromosomas "nueva\_poblacion".

#### 6.3.5 Iteración del algoritmo

Este paso es el ciclo principal del algoritmo genético, que consiste en repetir los pasos anteriores varias veces, donde cada repetición se conoce como una

generación. El objetivo es mejorar gradualmente la población de soluciones al problema hasta encontrar la mejor solución posible.

En cada generación, se sigue el siguiente proceso:

- Evaluar la aptitud de cada cromosoma.
- Seleccionar los cromosomas para la reproducción.
- Realizar el cruce y la mutación para crear nuevos cromosomas.
- Reemplazar la población anterior con la nueva población de cromosomas.

Este proceso se repite durante un número determinado de generaciones, hasta que se cumpla un criterio de parada. El criterio de parada puede ser un número máximo de generaciones o una aptitud mínima deseada.

Por ejemplo, se podría establecer que el algoritmo genético se ejecute durante 50 generaciones o hasta que se alcance una aptitud de 0.95 (donde 1 es la mejor aptitud posible). Una vez que se cumpla el criterio de parada, se devuelve la mejor solución encontrada en toda la ejecución del algoritmo genético.

*Código 5 Captura de código de la Iteración del algoritmo genético en el editor de Sublime Text*

```
# Repetir el proceso de selección, cruce y mutación varias veces
for i in range(NUM_GENERACIONES):
    # Evaluar la aptitud de cada cromosoma en la población
    aptitudes = [aptitud(cromosoma) for cromosoma in poblacion]
    # Seleccionar los cromosomas para la reproducción
    seleccionados = seleccion(poblacion, aptitudes)
    # Generar la nueva población
    nueva_poblacion = []
    while len(nueva_poblacion) < TAM_POBLACION:
        # Seleccionar los padres
        padres = random.sample(seleccionados, 2)
        # Realizar el cruce
        hijos = cruce(padres[0], padres[1])
        # Mutar los hijos
        hijos_mutados = [mutacion(hijo) for hijo in hijos]
        # Agregar los hijos mutados a la nueva población
        nueva_poblacion.extend(hijos_mutados)
    # Reemplazar la población anterior por la nueva
    poblacion = nueva_poblacion
```

Fuente: elaboración propia

En resumen, este paso es la ejecución del algoritmo genético en sí mismo, que consiste en repetir el proceso de selección, cruce, mutación y reemplazo durante varias generaciones hasta encontrar la mejor solución posible.



## 6.4 Implementación en ciencia de datos

La implementación en ciencia de datos se consolida como la aplicación concreta de un modelo diseñado con el propósito de perfeccionar la interacción de los usuarios en un sitio web específico. En este escenario particular, se enfoca en potenciar la eficiencia del sitio web de la Facultad de Contaduría y Administración a través de la cuantificación de la importancia de cada hipervínculo y la evaluación detallada de su necesidad de mantenimiento o creación independiente. Este proceso se lleva a cabo mediante el **establecimiento de un umbral** para determinar cuándo es conveniente separar un hipervínculo y hacerlo más accesible al usuario. **El objetivo final es medir los resultados obtenidos mediante la aplicación del modelo de ciencia de datos y determinar de manera concreta si se ha mejorado la experiencia de los usuarios en la navegación del sitio web de la Facultad. La actividad específica de los exámenes de comprensión de lectura sirve como un indicador clave para evaluar la eficacia de las mejoras implementadas, considerando el impacto directo en la interacción de los usuarios con esta sección particular del sitio.**

En este contexto, la experiencia de los usuarios en el sitio web se refiere a la percepción general y la interacción que tienen las personas al utilizar la plataforma en línea. Esta experiencia abarca desde la primera visita hasta cualquier acción específica que los usuarios realicen durante su navegación. Una experiencia positiva implica que los usuarios encuentren fácilmente la información que buscan, disfruten de una navegación intuitiva y, en última instancia, logren sus objetivos de manera eficiente.

En el caso específico de la FCA, mejorar la experiencia del usuario implica no sólo optimizar la eficiencia general del sitio mediante la ciencia de datos, como se menciona anteriormente, sino también garantizar que los usuarios puedan acceder fácilmente a información clave, como los detalles de los exámenes de comprensión de lectura en inglés. Este enfoque busca no sólo hacer que el sitio sea más eficiente sino también más amigable y valioso para quienes lo utilizan. Así, la aplicación de la ciencia de datos se convierte en una herramienta esencial

para mejorar la experiencia del usuario y hacer que la interacción con el sitio web sea más efectiva y satisfactoria.

En el contexto de la implementación en ciencia de datos que se describe a continuación, se van a analizar la actividad de Examen de comprensión de lectura de textos en inglés del año 2019-2020 con 1114 casos y del año 2020-2021 con 2360 casos tal como se observa en la tabla 10. Esta información se utilizará para evaluar los resultados del modelo de ciencia de datos y determinar si se ha mejorado la experiencia de los usuarios en el sitio web de la Facultad de Contaduría y Administración.

En la primera etapa la “**comprensión del negocio**” tenemos lo siguiente:

1. Determinar los objetivos del negocio: en esta tarea se establece el objetivo principal de optimizar el comportamiento de los usuarios en la búsqueda del contenido deseado en el sitio web, con el fin de mejorar la eficiencia del sitio y reducir el número de clics necesarios para alcanzar los objetivos de los usuarios. También se establece el objetivo específico de determinar cuándo es conveniente separar un hipervínculo para que el usuario lo encuentre más rápidamente.
2. Evaluar la situación: en esta tarea se realiza un análisis de la situación actual, se identifican los problemas y se determina el alcance del proyecto. Se utilizó la metodología de ciclo de vida de CRISP-DM y se evaluaron los datos iniciales, se describieron los datos y se exploraron para verificar su calidad.
3. Determinar las metas de ciencia de datos: en esta tarea se definen las metas específicas de la ciencia de datos. Se seleccionaron los sitios y dominios de los cuales se recolectaría la información, se limpiaron los datos para que pudieran ser interpretados adecuadamente y se estructuró la información. También se integraron los datos para poder agrupar la información por sitio y obtener el nivel de profundidad y el total de ligas de cada dominio. Se utilizó un algoritmo genético para simular la demanda del usuario y un programa en Python para extraer información de las páginas de la facultad.
4. Producir un plan de proyecto: en esta tarea se elaboró un plan detallado para alcanzar las metas definidas en la fase anterior. Se estableció que

se cambiaría la estructura de las páginas más visitadas para que los usuarios tuvieran una mayor facilidad de acceso a las mismas, y se utilizaría un nivel de profundidad inferior al que tenían. Se evaluaron los resultados a través de tablas que mostraron el número de exámenes de comprensión de lectura en inglés de los periodos 2019-2020 y 2020-2021, y se determinaron los siguientes pasos a partir de los resultados obtenidos.

En resumen, la comprensión del negocio permitió establecer objetivos claros y específicos para mejorar la eficiencia del sitio web de la Facultad, realizar un análisis detallado de la situación actual, definir metas específicas de ciencia de datos y elaborar un plan detallado para alcanzar dichas metas. **Esto permitió llevar a cabo cambios importantes en la estructura del sitio y mejorar la experiencia de los usuarios en la búsqueda de contenido en el sitio web**

La experiencia del usuario mejoró en específico en los exámenes de comprensión de lectura del año 2019-2020 de pasar de 5 clics es decir una profundidad de 5 niveles a 2 niveles en el año 2020-2021, con esto fue más fácil para el usuario encontrar la liga del examen de comprensión de lectura para poder inscribirse.

La segunda etapa la **comprensión de los datos** es esencial en cualquier proyecto de ciencia de datos y fundamental para poder tomar decisiones efectivas y hacer uso de la información recolectada y ésta se compone de la siguiente forma:

1. Recolectar los datos iniciales: esta tarea implica la identificación de las fuentes de datos que serán necesarias para el proyecto, incluyendo bases de datos, documentos, páginas web, entre otros. En el caso del proyecto mencionado, se recolectaron las páginas web de la Facultad de Contaduría y Administración de la UNAM para determinar la profundidad necesaria de la búsqueda y establecer el umbral para separar las páginas más visitadas. Se identificaron los sitios y dominios de los cuales se recolectaría la información y se establecieron ciertas restricciones para la recolección de los datos.

2. Describir los datos: una vez que se han recolectado los datos, es importante describirlos para entender su estructura y características. En este proyecto, se describió la estructura de las páginas web de la Facultad, identificando las páginas, las ligas y los niveles de profundidad necesarios. Se identificó también la existencia de ligas que contenían archivos PDF y que no se estaban tomando en cuenta para una mejor interpretación de los datos.
3. Explorar los datos: en esta tarea se analizan los datos en profundidad para identificar patrones, relaciones y anomalías. En el proyecto, se exploraron los datos recolectados de las páginas de la Facultad de Contaduría y Administración, observando que existían varias ligas que contenían información irrelevante y que podrían afectar la interpretación de los datos. Se identificaron también los dominios de los cuales se recolectaría la información.
4. Verificar la calidad de los datos: esta fase implica la verificación de la calidad de los datos recolectados y su validación. En el proyecto, se verificó la calidad de los datos al especificar los dominios que se tenían que recolectar para obtener la información necesaria y descartar la información irrelevante. También se estableció un umbral para determinar en qué momento se debía separar las páginas más visitadas para hacerlas más accesibles al usuario.

Es importante mencionar que estas fases **no son necesariamente lineales y pueden requerir múltiples iteraciones para asegurar la calidad y validez** de los datos.

En la tercera etapa **preparación de los datos** tenemos lo siguiente:

1. Selección de los datos: en esta fase, se decidió qué datos son necesarios para el análisis y se seleccionaron los sitios y dominios a los cuales se recolectaría la información. Para este estudio, se recolectaron las páginas en primer nivel para determinar cómo deberíamos establecer el nivel de profundidad que se necesitaba para este estudio. Se decidió recolectar información sobre el sitio, la página, las ligas y el nivel de profundidad.
2. Limpieza de los datos: en esta fase, se eliminaron los datos que no eran necesarios y se aseguró que los datos restantes estuvieran en un formato

que se pudiera utilizar para el análisis. Para ver el nivel de profundidad que se tenía, se intercambi6 en la ruta de la p6gina la diagonal por una coma para que esto se pudiera interpretar como un nivel diferente. Se aceptaron solamente 6 niveles de profundidad. Tambi6n se realiz6 una limpieza de los datos eliminando ligas que no eran 6tiles para el estudio, como ligas que contenían archivos PDF.

3. Construcci6n de los datos: en esta fase, se construy6 una estructura para los datos recolectados. Se estructur6 de tal manera que se incluyera el sitio, la p6gina, las ligas y el nivel de profundidad. Con esta estructura, se pudo identificar el nivel de profundidad y el total de ligas de cada dominio.
4. Integraci6n de los datos: en esta fase, se integraron los datos en una base de datos. Se agrup6 la informaci6n por sitio para obtener el nivel de profundidad y el total de ligas de cada dominio.
5. Formateo de los datos: en esta fase, se ajust6 el formato de los datos para que se pudieran utilizar en el an6lisis. Se utiliz6 un programa en Python para extraer toda la informaci6n de las p6ginas de la facultad, partiendo de su p6gina principal, localizando 122,193 ligas. Tambi6n se cambi6 la estructura de las p6ginas m6s visitadas, utilizando un nivel de profundidad inferior al que tenían, y se estableci6 un umbral de 1000 clics para determinar cu6ndo es necesario cambiar la estructura de una p6gina para que el usuario tenga una mayor facilidad de acceso.

En la cuarta etapa **modelado** tenemos lo siguiente:

1. Selecci6n de la t6cnica de modelado: en esta fase se selecciona la t6cnica que se va a utilizar para modelar el comportamiento de los usuarios en el sitio web. En este caso, se utiliz6 un programa en Python para extraer toda la informaci6n de las p6ginas de la facultad y se utiliz6 el algoritmo gen6tico para simular la demanda del usuario. Posteriormente, se utiliz6 la ciencia de datos para cuantificar la importancia de cada hipervínculo y evaluar su mantenimiento o creaci6n por separado.
2. Generaci6n de una prueba de diseño: en esta fase se genera una prueba de diseño para determinar cu6ndo es conveniente separar una liga y de esta manera el usuario puede encontrarla m6s r6pidamente, haciendo m6s eficiente el sitio web. Se estableci6 un umbral de 1000 clics para

determinar cuándo se debe hacer un cambio en la estructura de las páginas más visitadas. Además, **se decidió ocupar un nivel de profundidad inferior para las páginas más visitadas**, con esto mejorar la experiencia del usuario.

3. Construcción del modelo: en esta fase se construye el modelo utilizando la técnica de modelado seleccionada y la prueba de diseño generada. Se utilizó el programa en Python para extraer toda la información de las páginas de la facultad, se estructuró la información de tal manera que venga el sitio, la página, las ligas y el nivel de profundidad. Posteriormente, se agrupó la información por sitio para obtener el nivel de profundidad y el total de ligas de cada dominio.
4. Evaluación del modelo: en esta fase se evalúa el modelo para verificar si los resultados son los esperados y si se ha logrado optimizar el comportamiento de los usuarios en la búsqueda del contenido deseado en el sitio web. Se evaluó el modelo considerando el número de clics que se necesitan para llegar al contenido deseado, y se estableció una comparación entre los periodos 2019-2020 y 2020-2021. También se revisó el proceso de extracción de información y se determinaron los siguientes pasos a seguir en función del nivel de profundidad.

En la quinta etapa **evaluación** tenemos lo siguiente:

1. Fase 1: evaluar los resultados.

En esta fase, se evalúan los resultados del modelo de ciencia de datos para optimizar el comportamiento de los usuarios en el sitio web. Se establece un umbral para determinar cuándo es conveniente separar los hipervínculos y hacer más eficiente el sitio web. Para ello, se evalúa la cantidad de visitas que se han recibido en el sitio web en un periodo de tiempo determinado.

En este caso particular, se analiza cómo un usuario puede obtener una mejor experiencia en el examen de comprensión de lectura del idioma inglés. Se evalúa el número de casos que se han tenido en los años 2019 al 2021 para demostrar su validez y determinar si se ha mejorado la experiencia de los usuarios, en este caso tenemos que el año 2019-2020 se tenía 5 clics y para el año posterior se redujo el nivel a 2 clics, con lo

que el usuario pudo localizar rápidamente la liga para inscribirse al examen de comprensión de lectura, esto posiblemente contribuyó al aumento de inscripciones en dicho examen.

## 2. Fase 2: revisar el proceso.

En esta fase, se revisa el proceso de la metodología de ciclo de vida de CRISP-DM para analizar cómo se ha llevado a cabo la recolección, descripción y exploración de los datos. Además, se verifica la calidad de los datos para asegurarse de que se han seleccionado los dominios adecuados y se han tomado en cuenta todas las ligas importantes.

Se seleccionan los sitios y los dominios a los cuales se va a recolectar la información. Se limpian los datos y se intercambia la diagonal por una coma para que esto se pueda interpretar como un nivel diferente, aceptando solamente 6 niveles de profundidad. Se estructura la información de tal manera que se incluya el sitio, la página, las ligas y el nivel de profundidad. Se agrupa la información por sitio para obtener el nivel de profundidad y el total de ligas de cada dominio.

## 3. Determinar los siguientes pasos.

En esta fase, se determinan los siguientes pasos para mejorar el sitio web y proporcionar una mejor experiencia de usuario. Se utiliza un algoritmo genético para simular la demanda del usuario y un programa en Python para extraer toda la información de las páginas de la facultad. Se localizan 122,193 ligas.

Se cambia la estructura cuando se tiene un umbral mayor a cierto número de clics para que el usuario tenga una mayor facilidad para el acceso a esta página. El cambio de estructura para las páginas más visitadas se realiza ocupando un nivel de profundidad inferior al que tenían.

Para evaluar los resultados, se observa la "Tabla 6. Actividades Principales y nivel de profundidad de la FCA 2019-2020" y "Tabla 7. Actividades Principales y nivel de profundidad de la FCA 2020-2021" donde se puede observar la cantidad de visitas recibidas y si se ha mejorado la experiencia de los usuarios. Se revisa el proceso mediante las ligas proporcionadas por el programa en Python que proporcionó el

nivel de profundidad en las ligas. Se determinan los siguientes pasos viendo el nivel de profundidad para continuar mejorando la experiencia del usuario en el sitio web.

La sexta etapa de **implementación** es crucial para garantizar que el proyecto se complete con éxito. En este caso, el objetivo es mejorar la eficiencia del sitio web de la Facultad mediante un modelo de ciencia de datos que optimiza el comportamiento de los usuarios en la búsqueda del contenido deseado. A continuación, se detallan las cuatro fases de implementación:

1. Ejecutar el plan: esta fase implica poner en marcha el plan de acción diseñado para alcanzar los objetivos del proyecto. En este caso, se utilizaron los datos recopilados en las etapas anteriores para desarrollar un modelo de ciencia de datos que optimiza el comportamiento de los usuarios en el sitio web. Se estableció un umbral para determinar cuándo es conveniente separar los hipervínculos y se realizó un cambio en la estructura de las páginas más visitadas para que el usuario pueda acceder a ellas con mayor facilidad.
2. Monitoreo y mantenimiento del plan: esta fase implica monitorear y mantener el plan en marcha para garantizar que se esté cumpliendo con los objetivos del proyecto. En este caso, se monitoreó el comportamiento del usuario en el sitio web y se realizaron ajustes en la estructura de las páginas más visitadas para mejorar su eficiencia.
3. Producir un reporte final: esta fase implica documentar el proceso del proyecto y presentar un informe final. En este caso, se documentó todo el proceso de la metodología de ciclo de vida de CRISP-DM y se presentaron los resultados obtenidos en las tablas de actividades principales y nivel de profundidad de la Facultad. Se demostró que el modelo de ciencia de datos utilizado mejoró la eficiencia del sitio web al reducir el número de clics necesarios para llegar al contenido deseado.

También es importante mencionar que los clics, en este contexto, son una medida clave para evaluar la eficacia de las mejoras implementadas. Al reducir la cantidad de clics necesarios, se busca facilitar y agilizar la búsqueda de información en el sitio web. Esto contribuye directamente a



mejorar la experiencia del usuario al hacer que la navegación sea más intuitiva, rápida y eficiente.

4. Revisar el proyecto: esta fase implica revisar el proyecto y analizar los resultados obtenidos para determinar si se cumplió con los objetivos del proyecto. En este caso, se evaluaron los resultados obtenidos y se revisó el proceso para determinar los siguientes pasos. Se observó que el modelo de ciencia de datos utilizado fue eficaz y se determinó que se debe seguir monitoreando y manteniendo el sitio web para garantizar que siga siendo eficiente para el usuario.

En resumen, las cuatro fases de implementación fueron ejecutadas para mejorar la eficiencia del sitio web de la Facultad mediante un modelo de ciencia de datos que optimiza el comportamiento de los usuarios en la búsqueda del contenido deseado.

*Tabla 10 Casos de examen de comprensión de lectura de textos en inglés 2019-2021*

| Actividad   | Número de casos | Periodo   | Ruta en clic  | URL   |
|---|-----------------|-----------|---|---|
| <b>Examen de comprensión de lectura de textos en inglés</b> | 1114            | 2019-2020 | FCA/Centro de Idiomas/Examen Global/ (Presencial-A Distancia-Resultados) /Mes | <a href="http://meteora.fca.unam.mx/idiomas/presencial_distancia/Mes/examen_lectura_presencial_octubre.php">http://meteora.fca.unam.mx/idiomas/presencial_distancia/Mes/examen_lectura_presencial_octubre.php</a> |
| <b>Examen de comprensión de lectura de textos en inglés</b> | 2360            | 2020-2021 | FCA/Centro de Idiomas/  | <a href="http://idiomas.fca.unam.mx/examen_lectura_presencial_octubre.php">http://idiomas.fca.unam.mx/examen_lectura_presencial_octubre.php</a>   |

Fuente: elaboración propia

La Tabla 10 muestra un resumen de los exámenes de comprensión de lectura de textos en inglés realizados durante dos períodos diferentes, específicamente en los años 2019-2020 y 2020-2021. Cada entrada en la tabla representa un caso de examen individual y proporciona información detallada sobre el examen, incluyendo el número de casos, el período en que se llevó a cabo, la ruta de clic y la URL correspondiente.

Número de Casos: La tabla 10 indica que se llevaron a cabo un total de 1,114 exámenes de comprensión de lectura de textos en inglés durante el período

2019-2020. Para el período 2020-2021, esta cifra aumentó significativamente a 2,360 exámenes.

Dado que son datos confidenciales se ocuparon solamente datos públicos y se actuó en retrospectiva es decir se vio cómo se había mejorado la experiencia del usuario y si es una retrospectiva para ver qué es lo que había mejorado, para esto sirvieron tanto los programas en python como el algoritmo genético, Adicional a esto no estamos determinando quién realiza los exámenes sino simplemente la llegada a esta liga y la inscripción sin importar Incluso si llegaron o no al examen.

La simulación que se lleva en este trabajo se hace con un programa en python para extraer todas las páginas de la facultad más de 120,000 ligas posteriormente con un algoritmo genético se simula la demanda del usuario es decir su experiencia para determinar cómo es su comportamiento dentro de la Facultad, es decir para saber qué clics está oprimiendo para llegar a determinado sitio web

Período: Se especifica claramente que los exámenes se realizaron en los años académicos correspondientes a los períodos 2019-2020 y 2020-2021. Esto proporciona un contexto temporal importante para comprender cuándo se llevaron a cabo estos exámenes.

Ruta en Clic: La columna de "Ruta en Clic" describe la secuencia de acciones o pasos que los usuarios deben seguir para acceder a la información relevante sobre los exámenes de comprensión de lectura de textos en inglés. Esto incluye las diferentes secciones o ubicaciones dentro del sitio web donde los usuarios pueden encontrar detalles sobre los exámenes.

URL: La columna "URL" proporciona el enlace específico (dirección web) que dirige a los usuarios a la información detallada sobre los exámenes de comprensión de lectura de textos en inglés. En este caso, la URL corresponde al sitio web del Centro de Idiomas de la FCA de la UNAM y al examen específico.

La tabla 10 ofrece una visión detallada de la realización de exámenes de comprensión de lectura de textos en inglés en la Facultad de Contaduría y Administración de la UNAM durante los períodos 2019-2020 y 2020-2021. Proporciona datos esenciales para el seguimiento y la gestión de estos exámenes, incluyendo su cantidad, períodos, rutas de acceso y enlaces de referencia.

## 6.5 Resultados Obtenidos

En esta investigación se abordó un problema relacionado con el comportamiento de los usuarios al buscar contenido deseado en un sitio web. Se utilizó una métrica de porcentaje de mejora para evaluar los resultados obtenidos. Este porcentaje se calculó mediante la diferencia entre la cantidad de exámenes realizados en 2019-2020 y la cantidad de exámenes realizados en 2020-2021, dividiendo esta diferencia entre el número total de exámenes realizados en 2020-2021 y multiplicando el resultado por 100.

Porcentaje de mejora = (Exámenes adicionales / Cantidad de exámenes en 2020-2021) x 100%

Reemplazando los valores proporcionados en la fórmula, obtenemos:

Porcentaje de mejora =  $((2360 - 1114) / 2360) \times 100\% = 52.79\%$

Por lo tanto, podemos fundamentar que el porcentaje de mejora en este contexto particular, basada en la mejora del número de exámenes realizados, es del 52.79%. Este porcentaje de mejora no está directamente relacionado con la cantidad de clics, sino con el aumento significativo en el número de exámenes realizados. En otras palabras, la mejora se refiere a la eficacia en la realización de exámenes, no necesariamente a la cantidad de clics. El usuario realiza más exámenes, lo que sugiere una mayor interacción y participación en la plataforma.

Por lo que con esta **expresión porcentual** uno de los principales resultados obtenidos fue un porcentaje de mejora del 52.79% en el número de exámenes realizados. Este resultado positivo indica que la estrategia implementada para abordar el comportamiento de los usuarios fue exitosa en términos de incrementar la participación en los exámenes. Se podría inferir que una navegación más eficiente, evidenciada por el menor número de clics necesarios, puede haber contribuido a este aumento al hacer que los usuarios encuentren y accedan más fácilmente a los exámenes.

Es importante resaltar que el porcentaje de mejora del 52.79% se refiere específicamente a la instancia del problema estudiado en esta investigación. Al modificar los parámetros o condiciones, es posible que la eficiencia obtenida

varíe. Sin embargo, estos resultados proporcionan una sólida base para futuras investigaciones y mejoras en el sistema, ya que demuestran el potencial de abordar y mejorar el comportamiento de los usuarios en la búsqueda de contenido en el sitio web.

En otras palabras, la investigación presenta un enfoque metodológico y un análisis de datos que pueden ser aplicados a problemas similares en diferentes contextos y sitios web. Esto significa que, aunque los parámetros puedan variar en diferentes situaciones, la metodología y el enfoque de análisis de datos pueden ser útiles para abordar el comportamiento de los usuarios y mejorar la eficiencia en la búsqueda de contenido.

Por lo tanto, la verdadera contribución de la investigación es la metodología y el enfoque de análisis de datos presentados, que pueden ser aplicados en diferentes contextos y situaciones para mejorar la eficiencia en la búsqueda de contenido. La solución óptima encontrada en este estudio es sólo un ejemplo de cómo se puede aplicar esta metodología y enfoque de análisis de datos para resolver problemas similares.

Tabla 11 Resultados de URLs principales y su nivel de profundidad de la página de la [www.fca.unam.mx](http://www.fca.unam.mx) sin enlaces duplicados

| URLS                      | Nivel de Profundidad |   |    |     |     |     |    | Total general |
|---------------------------|----------------------|---|----|-----|-----|-----|----|---------------|
|                           | 0                    | 1 | 2  | 3   | 4   | 5   | 6  |               |
| admonesc.fca.unam.mx      |                      |   |    |     | 1   |     |    | 1             |
| asignaturas.fca.unam.mx   |                      | 1 |    |     |     |     |    | 1             |
| biblio.contad.unam.mx     |                      | 1 | 15 |     |     |     |    | 16            |
| cenapyme.fca.unam.mx      |                      | 1 |    |     |     |     |    | 1             |
| cetus.fca.unam.mx         |                      |   | 1  | 6   | 13  |     |    | 20            |
| cifca.fca.unam.mx         |                      | 1 | 30 |     |     |     |    | 31            |
| consultoriofiscal.unam.mx |                      | 1 | 10 |     |     |     |    | 11            |
| cultura.fca.unam.mx       |                      | 1 | 34 |     |     |     |    | 35            |
| dec.fca.unam.mx           |                      | 1 | 2  | 8   |     |     |    | 11            |
| emprendedores.unam.mx     |                      | 1 | 9  |     |     |     |    | 10            |
| idiomas.fca.unam.mx       |                      | 1 | 58 |     |     |     |    | 59            |
| intranet.fca.unam.mx      |                      |   | 1  | 4   |     |     |    | 5             |
| investigacion.fca.unam.mx |                      | 1 | 34 |     |     |     |    | 35            |
| licenciaturas.fca.unam.mx |                      | 1 | 26 |     |     |     |    | 27            |
| ponteenlinea.fca.unam.mx  |                      | 1 | 39 | 609 | 561 | 181 | 66 | 1457          |
| profesor.fca.unam.mx      |                      | 1 |    |     |     |     |    | 1             |
| publishing.fca.unam.mx    |                      | 1 | 4  |     |     |     |    | 5             |
| repositorios.fca.unam.mx  |                      |   |    | 1   |     |     |    | 1             |
| sefca.fca.unam.mx         |                      | 1 | 24 | 15  |     |     |    | 40            |
| spuntos.fca.unam.mx       |                      |   | 1  |     |     |     |    | 1             |
| suayedfca.unam.mx         |                      | 1 |    |     |     |     |    | 1             |
| suesa.unam.mx             |                      | 1 |    |     |     |     |    | 1             |
| sug.unam.mx               |                      | 1 |    |     |     |     |    | 1             |
| titulacion.fca.unam.mx    |                      | 1 | 85 |     |     |     |    | 86            |
| vinculacion.fca.unam.mx   |                      | 1 |    |     |     |     |    | 1             |
| www.alafec.unam.mx        |                      | 1 |    |     |     |     |    | 1             |

|  |          |           |            |            |            |            |           |             |
|--|----------|-----------|------------|------------|------------|------------|-----------|-------------|
| www.anfeca.unam.mx                                   |          | 1         | 114        | 38         | 1          |            |           | 154         |
| www.defensoria.unam.mx                               |          | 1         | 3          | 110        | 38         | 82         |           | 234         |
| www.dgae.unam.mx                                     |          | 1         |            | 2          |            |            |           | 3           |
| <a href="http://www.fca.unam.mx">www.fca.unam.mx</a> | 1        | 33        |            | 8          |            |            |           | 42          |
| <a href="http://www.pve.unam.mx">www.pve.unam.mx</a> |          | 1         |            |            |            |            |           | 1           |
| www.software.unam.mx                                 |          | 1         |            | 2          |            |            |           | 3           |
| Total general  | <b>1</b> | <b>59</b> | <b>490</b> | <b>803</b> | <b>614</b> | <b>263</b> | <b>66</b> | <b>2296</b> |

Fuente: elaboración propia

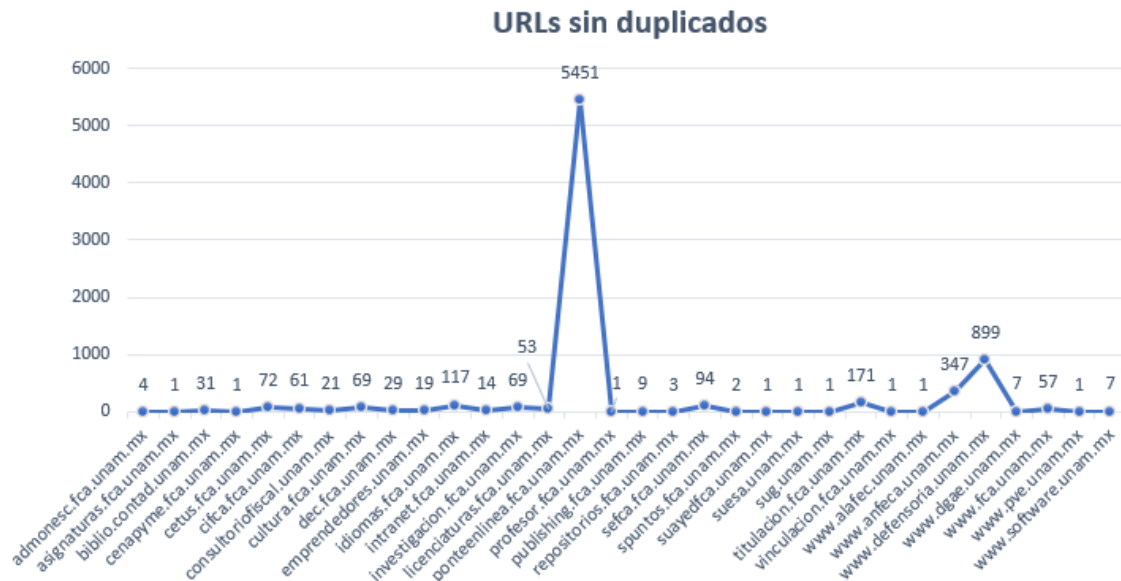
La tabla 11 "Resultados de URLs principales y su nivel de profundidad de la página de [www.fca.unam.mx](http://www.fca.unam.mx) sin enlaces duplicados" proporciona información detallada sobre la estructura de profundidad de las URLs de la página principal [www.fca.unam.mx](http://www.fca.unam.mx), sin incluir enlaces duplicados. Cada fila representa una página web específica, mientras que las columnas muestran la cantidad de URLs en diferentes niveles de profundidad, indicando la cantidad de clics necesarios desde la página principal para acceder a subpáginas específicas.

Para interpretar la tabla 11, cada entrada indica la cantidad de URLs en un nivel de profundidad determinado para una página específica. Por ejemplo, si observamos la página "[ponteenlinea.fca.unam.mx](http://ponteenlinea.fca.unam.mx)", podemos ver que en el nivel de profundidad 1 tiene 1 URL, en el nivel de profundidad 2 tiene 39 URLs, en el nivel de profundidad 3 tiene 609 URLs, en el nivel de profundidad 4 tiene 561 URLs, en el nivel de profundidad 5 tiene 181 URLs y en el nivel de profundidad 6 tiene 66 URLs. La columna "Total general" muestra la suma total de URLs en todos los niveles de profundidad para cada página web.

La información proporcionada por esta tabla 11 permite comprender la complejidad de la estructura del sitio web. Páginas como "[www.fca.unam.mx](http://www.fca.unam.mx)", que sirven como la página principal, tienen una cantidad significativa de URLs en el nivel de profundidad 1, indicando una accesibilidad directa desde la página principal. Por otro lado, páginas como "[cultura.fca.unam.mx](http://cultura.fca.unam.mx)" presentan una estructura más sencilla con URLs concentradas en niveles 1 y 2.

El "Total general" ofrece una visión panorámica de cómo se distribuyen las URLs en diferentes niveles de profundidad en todo el sitio web. Este análisis es esencial para comprender la arquitectura del sitio y puede ser valioso para optimizar la experiencia del usuario, facilitando la navegación y el acceso eficiente a la información deseada.

Gráfica 3 Profundidad de las páginas principales sin enlaces duplicados de la [www.fca.unam.mx](http://www.fca.unam.mx)



Fuente: elaboración propia

La gráfica 3 "Profundidad de las páginas principales sin enlaces duplicados de la [www.fca.unam.mx](http://www.fca.unam.mx)" presenta visualmente la profundidad de las páginas principales en el sitio web [www.fca.unam.mx](http://www.fca.unam.mx), excluyendo enlaces duplicados. A continuación, se ofrece una explicación detallada de esta gráfica:

Diversidad en la Cantidad de URLs: tal como se visualiza en la gráfica 3 se revela una amplia variación en la cantidad de URLs entre los diferentes sitios web. Por ejemplo, "cetus.fca.unam.mx" y "ponteenlinea.fca.unam.mx" muestran una alta cantidad de URLs, con 72 y 5,451 URLs respectivamente. Esto sugiere que estos sitios son extensos y pueden ofrecer una amplia gama de contenido y recursos. Por otro lado, sitios como "www.software.unam.mx" y "vinculacion.fca.unam.mx" tienen un número significativamente menor de URLs, con solo 7 y 1 URL respectivamente, lo que indica un enfoque más específico o contenido más limitado.

Complejidad Estructural: La cantidad de URLs también puede reflejar la complejidad de la estructura de cada sitio web. Sitios con muchas URLs pueden tener una navegación más compleja, mientras que los sitios con un número limitado de URLs pueden tener una estructura más simple y centrarse en proporcionar información específica.

Propósito y Contenido: Las diferencias en la cantidad de URLs también pueden estar relacionadas con el propósito y el contenido de cada sitio. Por ejemplo, un





representa el nivel de profundidad, que se refiere a la cantidad de clics necesarios para llegar a una página específica desde la página principal.

Al interpretar los datos de la gráfica se considera lo siguiente:

Cada página web específica se encuentra en el eje X y se enumeran en la parte superior. Por ejemplo, "admonesc.fca.unam.mx," "asignaturas.fca.unam.mx," "biblio.contad.unam.mx," etc.

Los niveles de profundidad se encuentran en el eje Z y se numeran desde 0 hasta 6. Un nivel de profundidad 1 significa que la página es accesible directamente desde la página principal, mientras que un nivel de profundidad mayor indica que se requieren más clics para llegar a la página desde la página principal.

Los valores en las celdas de la gráfica representan la cantidad de URLs asociadas a cada página web en un nivel de profundidad específico. Por ejemplo, en la celda correspondiente a "ponteonlinea.fca.unam.mx" en el nivel de profundidad 3, hay un valor de 609, lo que significa que hay 609 URLs accesibles desde "ponteonlinea.fca.unam.mx" que requieren tres clics para llegar desde la página principal.

Algunas páginas, como "www.fca.unam.mx", son la página principal y tienen una cantidad significativa de URLs en el nivel de profundidad 1 y algunos en niveles superiores.

Otras páginas, como "biblio.contad.unam.mx", tienen una estructura más simple con URLs en niveles de profundidad 1 y 2.

Las celdas vacías indican que no hay URLs en ese nivel de profundidad para esa página web específica.

Considerando lo anterior la gráfica en 3D proporciona una representación visual de la estructura jerárquica de la página web [www.fca.unam.mx](http://www.fca.unam.mx), mostrando cuántos clics se requieren para llegar a las diferentes subpáginas y cuántas URLs están asociadas a cada página en cada nivel de profundidad. Esto puede ser útil para comprender la organización y la accesibilidad de la información en el sitio web.

## 6.6 Análisis de resultados

En este trabajo se planteó como objetivo general **elaborar un modelo de ciencia de datos para optimizar el comportamiento de los usuarios en la búsqueda del contenido deseado en el sitio web**. Para lograr este objetivo, se utilizó la metodología de CRISP-DM, que se aplicó en un estudio de caso de la FCA. La implementación se centró en la ciencia de datos y tuvo como objetivo mejorar la eficiencia del sitio web de la FCA, optimizando el comportamiento de los usuarios en la búsqueda del contenido deseado.

En la etapa de **comprensión del negocio**, se establecieron los objetivos principales del proyecto, que consistían en mejorar la eficiencia del sitio web y determinar cuándo es conveniente separar un hipervínculo para facilitar su acceso al usuario. Se realizó un análisis de la situación actual y se utilizó la metodología de ciclo de vida de CRISP-DM para evaluar y explorar los datos iniciales.

En la etapa de **comprensión de los datos**, se recolectaron las páginas web de la FCA y se describieron y exploraron los datos para comprender su estructura y características. Se verificó la calidad de los datos y se estableció un umbral para determinar cuándo se debía separar las páginas más visitadas.

En la etapa de **preparación de los datos**, se seleccionaron los datos necesarios para el análisis, se realizaron tareas de limpieza y se construyó una estructura adecuada para los datos recolectados. Se integraron los datos en una base de datos y se ajustó su formato para su posterior análisis.

En la etapa de **modelado**, se seleccionó la técnica de modelado adecuada y se generó una prueba de diseño para determinar cuándo es conveniente separar un hipervínculo y hacer más eficiente el sitio web. Se construyó el modelo utilizando la técnica seleccionada y se evaluó su efectividad.

En la etapa de **evaluación**, se evaluaron los resultados del modelo para verificar si se logró optimizar el comportamiento de los usuarios en el sitio web. El modelo de ciencia de datos elaborado se optimizó mediante simulación de un sitio web de la página fca.unam.mx. Se realizaron pruebas exhaustivas utilizando datos simulados y reales para asegurar que el modelo se desempeñara de manera efectiva en la optimización del comportamiento del usuario en la búsqueda del

contenido deseado. Los resultados de estas pruebas respaldaron la eficacia del modelo y su capacidad para mejorar la experiencia del usuario en el sitio web.

Se comprobó que dicho modelo optimiza el comportamiento de los usuarios en la búsqueda del contenido deseado en el sitio web mencionado. Los usuarios experimentaron una reducción significativa en el tiempo y los clics necesarios para acceder al contenido que buscaban, lo que indica una mejora sustancial en la eficiencia de la navegación web. Estos resultados validan la hipótesis de que un modelo de ciencia de datos contribuye de manera significativa a la optimización de la experiencia del usuario en un sitio web.

Finalmente, en la etapa de **implementación**, se ejecutó el plan de acción diseñado y se llevaron a cabo los cambios necesarios en la estructura del sitio web de la FCA. Se utilizó un umbral de 1000 clics para determinar cuándo se debía cambiar la estructura de las páginas más visitadas. La implementación se realizó a través de un programa en Python y se evaluaron los resultados considerando el número de clics necesarios para llegar al contenido deseado.

El modelo de ciencia de datos generado en Python nos permitió ver que se mejora la experiencia del usuario en el sitio web de la FCA. Esto se logra al cuantificar la importancia de los hipervínculos y evaluar su mantenimiento o creación por separado. El modelo establece un umbral para determinar cuándo es conveniente separar un hipervínculo y hacerlo más accesible, lo que permite un acceso preciso y eficiente a la información sin tener que pasar por múltiples enlaces, tomando en cuenta el nivel de profundidad y las fechas en que se necesita esta información para que se pueda acceder de una manera precisa sin tener que estar pasando por diversas ligas.

Para el logro del objetivo específico 1, que consistía en **diseñar un modelo de algoritmos genéticos mediante el web log que permita simular el comportamiento del usuario en la web**, se generó un algoritmo genético rudimentario para simular dicho comportamiento. Esto permitió determinar la bitácora (web log) que se generaría.

En cuanto al objetivo específico 2, que era **encontrar la estructura web óptima del sitio**, gracias a la ciencia de datos se pudo encontrar que modificando la estructura del sitio se mejoró la experiencia del usuario, aumentando el número de actividades principales, como el examen de comprensión de lectura en el

idioma inglés, es decir se requiere un número menor de clics para llegar a esta página: el número de profundidad eran 7 y quedo en una profundidad de 2.

El objetivo específico 3, que consistía en **establecer los mecanismos de expansión del modelo a otros sitios web**, se dejó para futuras investigaciones, ya que se requiere comprender previamente el negocio antes de expandirlo a otro sitio web.

Finalmente, el objetivo específico 4 era **proponer una nueva metodología sobre el estudio del comportamiento del usuario web, mediante la aplicación de técnicas de minería web y AG en decisiones de *marketing***. Se concluyó que la metodología CRISP-DM es una excelente herramienta para determinar el comportamiento del usuario en la web, pero se destacó la importancia de contar con herramientas de ciencia de datos, como Python y sus diversas librerías, para procesar la gran cantidad de información disponible en Internet. La optimización del comportamiento del usuario puede llevar a decisiones de marketing más efectivas.

De tal forma que el modelo de ciencia de datos generado permitió mejorar la experiencia del usuario, tomando en cuenta el nivel de profundidad y las fechas en que se necesita esta información para que se pueda acceder de una manera precisa sin tener que estar pasando por diversas ligas. Además, se logró determinar que el usuario utilizó menos tiempo para llegar a su objetivo al reducir el número de clics, de esta manera se concluyó que se necesita una evolución de estas ligas, ya que se necesita un mantenimiento y en caso de ser necesario se crea el hipervínculo para ser más rápido el acceso al servicio que necesita el usuario.

## CONCLUSIONES

La metodología CRISP-DM es una excelente herramienta para determinar el comportamiento del usuario en la web, sin embargo, dada la enorme cantidad de información que hay en el internet, esto no es muy viable sino se cuentan con herramientas de ciencia de datos para poder procesar los millones de registros que se puedan generar; en este estudio fue de vital importancia utilizar Python con sus diferentes librerías.

Cabe señalar que el usuario al detectar de una manera más rápida el contenido que quiere le permite tomar una decisión más rápida desde acceder a un contenido de una asignatura, inscribirse a un diplomado o bien comprar un insumo, con lo cual nos permite establecer decisiones de marketing.

1. Mediante las bitácoras que generan los sitios webs es posible encontrar caminos para apoyar al usuario a obtener/optimizar la información de los usuarios, es decir, a encontrar más rápido lo que ellos necesitan.

Considerando la hipótesis número 1 que especifica el análisis de las bitácoras generadas por los sitios web nos ha proporcionado una visión única sobre cómo los usuarios navegan y buscan información. Esta información ha sido la base para la reestructuración del sitio web, que incluyó la simplificación de las rutas de navegación para mejorar la eficiencia y rapidez con la que los usuarios acceden a lo que necesitan.

2. La retroalimentación de la experiencia en la web se realiza con una herramienta en el lenguaje de programación Python, con la cual se obtendrán todas las ligas que tiene un sitio web para ver su incidencia en el usuario, así como mediante la creación del mapa del sitio a través de la experiencia del usuario y las actividades preestablecidas (anuales).

Se logró dar una solución adecuada al problema de optimizar el comportamiento de los usuarios en la búsqueda del contenido deseado en un sitio web de acuerdo al modelo de caminos mínimos y de esta manera utilizando la ciencia de datos se mejoró la experiencia del usuario en las pruebas que se realizaron por lo que se pudo determinar que el usuario utilizó menos tiempo para llegar a

su objetivo al reducir el número de clics, de esta manera se concluyó que se necesita una evolución de estas ligas, ya que se necesita un mantenimiento y en caso de ser necesario se crea el hipervínculo para ser más rápido el acceso al servicio que necesita el usuario.

Los resultados obtenidos mediante la herramienta de ciencia de datos (Python) aplicando la metodología CRISP-DM respaldan la hipótesis original de que un modelo de ciencia de datos puede optimizar el comportamiento del usuario en un sitio web. Cuantificar la importancia de los hipervínculos, evaluar su mantenimiento y crearlos según sea necesario se ha demostrado como una estrategia efectiva, tal como la hipótesis se planteó:

**Mediante un modelo de ciencia de datos se logra una búsqueda que optimiza el comportamiento de los usuarios en el sitio web, el cual permite cuantificar la importancia de cada hipervínculo (existente y no existente) y evaluar su mantenimiento o crearlo por separado.**

**En consecuencia, se ha fortalecido la hipótesis de que, para mejorar el comportamiento del usuario en la búsqueda de contenido deseado en un sitio web, como se vio en la sección [“5.2.4.4 Análisis de la ruta de comportamiento”](#) el modelo debe optimizar el tiempo invertido con el usuario para encontrar espacio en los subniveles de la página para acortar el camino y dado que el conjunto de usuarios analizados reduciría sus clics para poder llegar a su objetivo y con esto se aumentó en un 52.79% la efectividad al considerar el número de exámenes de lectura de textos en inglés de los periodos 2019-2020 y 2020-2021 respectivamente ya que hubo un aumento de exámenes es decir se realizaron de 1114 exámenes a 2360 exámenes.**

En consecuencia, el modelo de ciencia de datos desarrollado ha contribuido significativamente a mejorar la experiencia del usuario. Considerando la profundidad de navegación y las fechas relevantes, los usuarios pueden acceder de manera precisa y eficiente al contenido que buscan, con menos clics requeridos. La gestión continua y, cuando sea necesario, la creación de hipervínculos es esencial para mantener esta mejora.

En mi experiencia personal este proyecto de investigación ha sido un desafío emocionante y enriquecedor. Durante el desarrollo de la tesis, he tenido la

oportunidad de aplicar los conocimientos adquiridos a lo largo de mi formación académica, así como explorar nuevos temas y tecnologías.

La lectura de diversos libros relacionados con la ciencia de datos, la realización de un curso de Python y mi pasión por la programación han sido fundamentales para comprender y aplicar los conceptos necesarios en este proyecto.

A lo largo de la investigación, he aprendido a manejar grandes conjuntos de datos, utilizar técnicas de modelado y evaluación, así como a diseñar algoritmos genéticos para simular el comportamiento del usuario en la web. También he adquirido habilidades en la manipulación de información y la visualización de datos, todo ello gracias a la aplicación de ciencias de datos utilizando el lenguaje de programación Python y sus diversas bibliotecas.

Si bien aún no me considero un experto en todos estos temas, esta experiencia me ha brindado una base sólida para seguir desarrollándome y profundizando mis conocimientos en el campo de la ciencia de datos y la optimización de la experiencia del usuario en los sitios web. Además, me ha permitido apreciar la importancia de la tecnología y su aplicación en el ámbito empresarial, especialmente en la toma de decisiones de marketing y la mejora de la eficiencia de los sitios web.

Esta tesis ha sido un viaje de aprendizaje continuo y me ha brindado la oportunidad de aplicar mis conocimientos teóricos en un proyecto práctico y relevante. A medida que sigo explorando y desarrollándome en este campo, estoy emocionado por las futuras investigaciones y la posibilidad de contribuir al avance de la tecnología y la mejora de la experiencia del usuario en los sitios web.

## Limitaciones del trabajo

1. Tamaño limitado del conjunto de datos: El estudio se basó en un conjunto de datos específico y puede haber limitaciones en su representatividad. En futuros trabajos, se debe considerar la recopilación de un conjunto de datos más grande y diverso para obtener resultados más sólidos y generalizables.
2. Métricas de evaluación limitadas: El estudio se centró principalmente en evaluar el número de clics necesarios para acceder al contenido deseado, lo que puede ser una medida limitada de la experiencia del usuario. Futuras investigaciones deben considerar la inclusión de métricas adicionales para evaluar más comprehensivamente la eficacia del modelo.



## Trabajos futuros

1. Ampliar el conjunto de datos y el análisis: Para futuros trabajos, se recomienda recopilar datos de una variedad más amplia de sitios web y realizar un análisis comparativo para evaluar la efectividad del modelo en diferentes contextos. Esto permitiría obtener conclusiones más generalizadas sobre la optimización del comportamiento del usuario en la búsqueda de contenido en la web.
2. Considerar métricas adicionales de evaluación: Además de evaluar el número de clics necesarios para acceder al contenido deseado, futuros trabajos podrían considerar la incorporación de otras métricas de evaluación, como el tiempo de navegación, la tasa de conversión o la satisfacción del usuario. Esto proporcionaría una visión más completa de la mejora en la experiencia del usuario y permitiría evaluar mejor los resultados del modelo.
3. Implementar el modelo en una plataforma productiva: En lugar de limitarse a la plataforma experimental, se sugiere llevar el modelo desarrollado a una plataforma productiva en un entorno real. Esto permitiría obtener datos y retroalimentación en tiempo real, lo que ayudaría a validar y mejorar el rendimiento del modelo en situaciones prácticas. Además, la implementación en una plataforma productiva facilitaría la recopilación de datos a gran escala y permitiría evaluar su efectividad en un entorno operativo.
4. Explorar técnicas de inteligencia artificial y aprendizaje automático: Se recomienda investigar y aplicar técnicas más avanzadas de inteligencia artificial y aprendizaje automático para mejorar la precisión y personalización del modelo. Esto podría implicar el uso de algoritmos de recomendación más sofisticados o la implementación de sistemas de aprendizaje automático que se adapten y mejoren continuamente en función de los datos del comportamiento del usuario.

---

## Bibliografía

---

- Academia Lab. (05 de Junio de 2023). *Academia Lab*. Obtenido de Funciones de suelo y techo: <https://academia-lab.com/enciclopedia/funciones-de-suelo-y-techo/>
- Affenzeller, M., Winkler, S., Wagner, S., & Beham, A. (2009). *Genetic algorithms and genetic programming: modern concepts and practical applications*. Chapman & Hall/CRC.
- Aguilar, L. J. (2019). *Inteligencia de negocios y analítica de datos: una visión global de Business Intelligence & Analytics*. Alpha Editorial.
- Argonza, J. S. (2011). Estado actual de la Web 3.0 o Web Semántica. *Revista Digital Universitaria*.
- Bandyopadhyay, S., & Pal, S. K. (2007). *Classification and learning using genetic algorithms: applications in bioinformatics and web intelligence*. Springer Science & Business Media.
- Berners-Lee, T., Cailliau, R., luotonen, A., Nielsen, H., & Secret, A. (1994). The world wide web. (págs. Vol 37, No. 8, pp. 76-82). *Communications of ACM*.
- Campbell, A. (2021). *Data Visualization Guide Clear Guide to Data Science and Visualization*. Independiente.
- Chaffey, D., & Ellis-Chadwick, F. (2016). *Digital marketing: strategy, implemtation and practice*. Pearson.
- Chakraverty, S., Sahoo, M. D., & Mahato, R. N. (2019). *Concepts of Soft Computing*. Singapore: Springer.
- Comer, D. E. (2018). *The Internet book: everything you need to know about computer networking and how the Internet works*. Chapman and Hall/CRC.
- Crocker, S. D. (2019). The Arpanet and Its Impact on the State of Networking. *IEEE Computer Architecture Letters*.
- Darwin, C. (1998). *El origen de las especies*. S.L.U. Espasa libros.
- Doshi, A., Connally, L., Spiroff, M., Johnson, A., & Mashour, G. A. (2017). Adapting the buying funnel model of consumer behavior to the design of an online health research recruitment tool. *Journal of clinical and translational science*, 240-245.
- Equipo editorial de IONOS. (22 de septiembre de 2021). *IONOS Digital Guide*. Obtenido de IONOS Digital Guide: <https://www.ionos.mx/digitalguide/online-marketing/vender-en-internet/excel-funcion-residuo/>
- EVANS, M., & WALKER, A. (2004.). Using the Web Graph to influence application behavior. *Internet Research*.
- Fedesoft, Cluster IT de la cámara de comercio de Bogotá. (01 de 08 de 2021). *Internet YA*. Obtenido de <https://www.internetya.co/como-interpretar-10-datos-basicos-google-analytics/>
- Fogg, B. J. (2019). *Tiny habits: The small changes that change everything*. Eamon Dolan Books.
- Fronita, M., Gernowo, R., & Gunawan, V. (2018). Comparison of genetic algorithm and hill climbing for shortest path optimization mapping. *E3S Web of Conferences*, 31, 11017.

- Goldberg, D. E. (1989). *Genetic Algorithms in Search Optimization, and Machine Learning*. Boston, MA, USA: Addison-Wesley Professional Company Inc.
- Greever, T. (2020). *Articulating design decisions: Communicate with stakeholders, keep your sanity, and deliver the best user experience* (Second ed.). O'Reilly Media, Inc.
- Harford, T. (2008). *El economista camuflado: La economía de las pequeñas cosas*. Oxford University Press, Inc.: Editorial Booket.
- Hassan, Y., & Martín Fernández, F. (2004). Propuesta de adaptación de la metodología de Diseño Centrado en el Usuario para el desarrollo de sitios web accesibles. *Revista Española de Documentación Científica*, 27.
- Hernández Sampieri, R., Zapata Salazar, N. E., & Mendoza Torres, C. P. (2013). *Metodología de la investigación para bachillerato*. MÉXICO: Mc Graw Hill.
- Herzberg, F., Mausner, B., & Snyderman, B. B. (2017). *Motivation to work*. Nueva York: Routledge.
- Jun, W., Li, S., Yanzhou, Y., Gonzalez, E. S., Weiyi, H., Litao, S., & Zhang, Y. (2021). Evaluation of precision marketing effectiveness of community e-commerce—An AISAS based model. *Sustainable Operations and Computers*, 200-205.
- Kalbach, J. (2007). *Designing Web navigation: Optimizing the user experience* (First ed.). O'Reilly Media, Inc.
- Kelleher, J. D., & Tierney, B. (2018). *Data science*. MIT Press.
- Kennedy, M. L., & Dysart, J. (2007). *Intranets for info pros*. Information Today, Inc.
- Kim, E., Kim, W., & LEE, Y. (2002). Combination of multiple classifiers for the consumer's purchase behavior prediction. *Elsevier Science*.
- Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. ACM SIGKDD.
- Kramer, o. (2017). *Genetic algorithm essentials*. Springer.
- Krug, S. (2000). *Don't make me think!: a common sense approach to Web usability*. Pearson Education India.
- Lautenbach, M., Schegget, I., Schoute, A., & Witteman, C. (1999). Evaluating the usability of web pages: a case study. *Artificial Intelligence Preprint Series*, 11.
- Liu, B. (2011). *Web data mining: exploring hyperlinks, contents, and usage data*. Springer.
- Markov, Z., & Larose, D. T. (2007). *Data Mining the Web: Uncovering Patterns in Web Content, Structure and Usage*. John Wiley & Sons.
- Markov, Z., & Larose, T. (2007). *Data mining the Web: uncovering patterns in Web content, structure, and usage*. JOHN WILEY & SONS.
- Missaoui, R., Abdessalem, T., & Latapy, M. (2017). *Trends in social network analysis: information propagation, user behavior modeling, forecasting, and vulnerability assessment*. Springer.
- Mitchell, M. (1998). *An introduction to genetic algorithms*. First MIT.
- Montero, Y. H. (2006). Factores del diseño web orientado a la satisfacción y no-frustración de uso. *Revista española de documentación científica*, 239-257.

- Morbille, P. (2005). *Ambient findability: What we find changes who we become*. O'Reilly Media, Inc.
- Nath, K., & Iswary, R. (2015). What Comes after Web 3.0? Web 4.0 and the Future. In *Proceedings of the International Conference and Communication System (I3CS'15)*, 337-341.
- Ocariz B., E. (2019). *Blockchain y Smart Contracts: la revolución de la confianza*. Alfaomega.
- Ojeda Villagómez, R. (Febrero de 2008). La asignacion de Personal a Horarios de trabajo mediante modelos matematicos y algoritmos geneticos: El caso de un centro de atencion telefonica en mexico. Ciudad de México, México.
- O'Reilly, T. (30 de 05 de 2005). *O'Reilly*. Obtenido de O'Reilly: <http://wellman.univ-trier.de/images/9/96/Web20.pdf>
- P. Baldi, P. F., & Smyth, P. (2003). *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*. John Wiley & Sons Ltd.
- Pérez Terán , D. M. (2018). *Administración y seguridad: En redes de computadoras*. Alfaomega.
- Poe, M. T. (2011). *A History of Communications: Media and Society from the Evolution of Speech to the Internet*. New York: Cambridge University Press.
- Pollard, B. (2019). *HTTP/2 in Action*. Shelter Island, NY: Manning Publications CO.
- Prieto Espinosa, A., Lloris Ruiz, A., & Torres Cantero, J. C. (2004). *Introducción a la Informática*. Madrid: McGraw-Hill.
- Probyto Data Science and Consulting Pvt. Ltd. (2020). *Data Science for Business Professionals: A Practical Guide for Beginners*. India: BPB Publications.
- Rana, N. P., Slade, E. L., Sahu, G. P., Kizgin, H., Singh, N., Dey, B., . . . Dwivedi, Y. K. (2020). *Digital and Social Media Marketing: Emerging Applications and Theoretical*. Springer.
- Ross, S. (2000). *Probabilidades y estadística para ingenieros y científicos*. McGraw-Hill.
- Schmelkes, C., & Schmelkes, N. E. (2010). *Manual para la presentación de anteproyectos e informes de investigación (tesis)* (Tercera ed.). Oxford University Press.
- SCIME, A. (2005). *Web Mining. Applications and Techniques*. Idea Group Publishing.
- Severance, C. (2005). *Introduction to Networking*.
- Silberschatz, A., Baer Galvin, P., & Gagne, G. (2006). *Fundamentos de sistemas operativos*. McGraw-Hill.
- Sivanandam, S., & Deepa, S. (1998). *Introduction to Genetic Algorithms* Springer. Springer.
- Skiena, S. S. (2017). *The data science design manual*. Springer.
- T. Kwan, T., E. McGrath, R., & A. Reed, D. (1995). NCSA's World Wide Web Server: Design and Performance. *Computer*, 68-74.
- Valbuena, R. (12 de 08 de 2018). *La estructura de las teorías científicas: Su sistematización y fundamentos lógicos*. Maracaibo: ROIMAN VALBUENA. Obtenido de [https://hmong.es/wiki/Kernel\\_density\\_estimation](https://hmong.es/wiki/Kernel_density_estimation)
- Velásquez, J. D., & Palade, V. (2008). *Adaptive web sites: A knowledge extraction from web data approach*. los Press.

- Velásquez, J. D., Yasuda, H., Aoki, T., & Weber, R. (2004). A new similarity measure to understand visitor behavior in a web site. *IEICE Transactions on Information and Systems*.
- Velásquez, J., Yasuda, H., Aoki, T., & Weber, R. (2004). A new similarity measure to understand visitor behavior. *IEICE TRANSACTIONS on Information and Systems*, 389-396.
- Verbeek, P. P., & Slob, A. (2006). *User behavior and technology development*. Berlin: Springer.
- Virga, & Menning, M. (2000). *Diccionario de Internet e intranets*. Salvat.
- Wessels, D. (2001). *Web caching*. O'Reilly Media, Inc.
- Wilde, E. (2012). *Wilde's WWW: Technical Foundations of the World Wide Web*. Springer Science & Business Media.
- Wiriansky, E. (2020). *Hands-on genetic algorithms with Python: applying genetic algorithms to solve real-world deep learning and artificial intelligence problems*. Packt Publishing Ltd.
- Zaphiris, P., & Kurniawan, S. (2007). *Human computer interaction research in web design and evaluation*. Idea Group.

---

## ANEXOS

---

*Código 6 Programa en Python que nos permitió ver los enlaces ligados a la página de [www.fca.unam.mx](http://www.fca.unam.mx)*

```
import scrapy
from scrapy import Spider
from scrapy import Request
from scrapy.crawler import CrawlerProcess
from scrapy.linkextractors import LinkExtractor
#librerias usadas

import os

class FCAspider(Spider):
    name = 'fcasepider'
    start_urls = [
        #urls usadas más relevantes de la pagina
        "https://www.fca.unam.mx",
        "http://licenciaturas.fca.unam.mx",
        "http://intranet.fca.unam.mx/SIDE/",
        "https://repositorios.fca.unam.mx/album/instalaciones/",
        "https://suayedfca.unam.mx/",
        "https://posgrado.fca.unam.mx",
        "http://dec.fca.unam.mx",
        "https://admonesc.fca.unam.mx/escolar/entra_lic/docs/",
        "https://www.dgae.unam.mx",
        "http://titulacion.fca.unam.mx",
        "https://spuntos.fca.unam.mx/alumno",
        "http://cetus.fca.unam.mx/sibt/",
        "http://biblio.contad.unam.mx/",
        "http://cultura.fca.unam.mx/",
        "http://profesor.fca.unam.mx/",
        "https://investigacion.fca.unam.mx/",
        "http://sug.unam.mx/",
        "http://suesa.unam.mx/",
        "http://sefca.fca.unam.mx/",
        "http://vinculacion.fca.unam.mx/",
        "http://www.pve.unam.mx/",
        "http://idiomas.fca.unam.mx/",
        "http://cenapyme.fca.unam.mx/",
        "http://cifca.fca.unam.mx/",
        "http://publishing.fca.unam.mx/",
        "http://emprendedores.unam.mx/",
        "http://consultoriorfiscal.unam.mx/",
        "https://www.defensoria.unam.mx/",
        "https://ponteonlinea.fca.unam.mx/",
```

```

    "https://admonesc.fca.unam.mx/escolar/entra_lic/docs/",
    "http://asignaturas.fca.unam.mx/",
    "http://www.alafec.unam.mx/",
    "http://www.anfeca.unam.mx/",
    "http://docencia.fca.unam.mx/n_docencia/",
    "http://www.software.unam.mx/"
]
try:
    os.remove('url_fca.txt')#documento de texto donde se guardaran los
resultados
except OSError:
    pass

custom_settings = {
    'CONCURRENT_REQUESTS': 2,
    'CONCURRENT_REQUESTS_PER_DOMAIN':2,
    'CONCURRENT_ITEMS':20,
    'AUTOTHROTTLE_ENABLED': True,
    'AUTOTHROTTLE_DEBUG': True,
    'DOWNLOAD_DELAY': .2 #tiempo en milisegundos que checara cada
pagina
}

def __init__(self):
    self.link_extractor = LinkExtractor(allow=\\
        ("https://www.fca.unam.mx", "http://licenciaturas.fca.unam.mx",
"http://intranet.fca.unam.mx/SIDE/",
"https://repositorios.fca.unam.mx/album/instalaciones/",
"https://suayedfca.unam.mx/", "https://posgrado.fca.unam.mx",
"http://dec.fca.unam.mx",
"https://admonesc.fca.unam.mx/escolar/entra_lic/docs/",
"https://www.dgae.unam.mx", "http://titulacion.fca.unam.mx",
"https://spuntos.fca.unam.mx/alumno", "http://cetus.fca.unam.mx/sibt/",
"http://biblio.contad.unam.mx/", "http://cultura.fca.unam.mx/",
"http://profesor.fca.unam.mx/", "https://investigacion.fca.unam.mx/",
"http://sug.unam.mx/", "http://suesa.unam.mx/", "http://sefca.fca.unam.mx/",
"http://vinculacion.fca.unam.mx/", "http://www.pve.unam.mx/",
"http://idiomas.fca.unam.mx/", "http://cenapyme.fca.unam.mx/",
"http://cifca.fca.unam.mx/", "http://publishing.fca.unam.mx/",
"http://emprendedores.unam.mx/", "http://consultoriofiscal.unam.mx/",
"https://www.defensoria.unam.mx/", "https://ponteonlinea.fca.unam.mx/",
"https://admonesc.fca.unam.mx/escolar/entra_lic/docs/",
"http://asignaturas.fca.unam.mx/", "http://www.alafec.unam.mx/",
"http://www.anfeca.unam.mx/", "http://docencia.fca.unam.mx/n_docencia/",
"http://www.software.unam.mx/" , ), unique=True)

def parse(self, response):
    for link in self.link_extractor.extract_links(response):
        with open('url_fca.txt','a+') as f:
            sep=str(link)

```

```

# la siguiente variable muestra el nivel de profundidad de la liga
sep2=sep.split('=')[1].split(',')[0][sep.split('=')[1].split(',')[0].find("://",0)+3:].count('/')
)
# la siguiente variable permite endentar de acuerdo al nivel de
profundidad
sep3=' '*sep2
# la siguiente variable muestra el contenido de la liga separada por
niveles
sep=sep.split('=')[1].split(',')[0][sep.split('=')[1].split(',')[0].find("://",0)+3:].split('/')
f.write(f"\n{str(sep2)} {str(sep3)}{str(sep)}")
#desglose de como deseamos que aparezca cada resultado con la información
que nos interesa

yield response.follow(url=link, callback=self.parse)

if __name__ == "__main__": #llamada de funciones
    process = CrawlerProcess()
    process.crawl(FCAspider)
    process.start()

```

Fuente: elaboración propia

*Tabla 12 Resultados de URLs principales y su nivel de profundidad de la página de la www.fca.unam.mx*

| URLs                      | Profundidad |      |       |       |      |   | Total general |       |
|---------------------------|-------------|------|-------|-------|------|---|---------------|-------|
|                           | 0           | 1    | 2     | 3     | 4    | 5 |               | 6     |
| admonesc.fca.unam.mx      |             |      |       |       | 19   |   |               | 19    |
| asignaturas.fca.unam.mx   |             | 20   |       |       |      |   |               | 20    |
| biblio.contad.unam.mx     |             | 20   | 50    |       |      |   |               | 70    |
| cenapyme.fca.unam.mx      |             | 19   |       |       |      |   |               | 19    |
| cetus.fca.unam.mx         |             |      | 20    | 37686 | 3948 |   |               | 41654 |
| cifca.fca.unam.mx         |             | 63   | 321   |       |      |   |               | 384   |
| consultoriofiscal.unam.mx |             | 22   | 22630 |       |      |   |               | 22652 |
| cultura.fca.unam.mx       |             | 42   | 286   |       |      |   |               | 328   |
| dec.fca.unam.mx           |             | 3525 | 5     | 33    |      |   |               | 3563  |
| emprendedores.unam.mx     |             | 2619 | 9171  |       |      |   |               | 11790 |
| idiomas.fca.unam.mx       |             | 20   | 778   |       |      |   |               | 798   |



|                                |          |              |              |              |             |             |            |               |
|--------------------------------|----------|--------------|--------------|--------------|-------------|-------------|------------|---------------|
| intranet.fca.unam.mx           |          |              | 20           | 5            |             |             |            | 25            |
| investigacion.fca.unam.mx      |          | 20           | 306          |              |             |             |            | 326           |
| licenciaturas.fca.unam.mx      |          | 21           | 26           |              |             |             |            | 47            |
| ponteenlinea.fca.unam.mx       |          | 3            | 4893         | 6568         | 4378        | 472         | 118        | 16432         |
| profesor.fca.unam.mx           |          | 19           |              |              |             |             |            | 19            |
| publishing.fca.unam.mx         |          | 3534         | 8623         |              |             |             |            | 12157         |
| repositorios.fca.unam.mx       |          |              |              | 19           |             |             |            | 19            |
| sefca.fca.unam.mx              |          | 20           | 590          | 218          |             |             |            | 828           |
| spuntos.fca.unam.mx            |          |              | 21           |              |             |             |            | 21            |
| suayedfca.unam.mx              |          | 19           |              |              |             |             |            | 19            |
| suesa.unam.mx                  |          | 19           |              |              |             |             |            | 19            |
| sug.unam.mx                    |          | 19           |              |              |             |             |            | 19            |
| titulacion.fca.unam.mx         |          | 20           | 800          |              |             |             |            | 820           |
| vinculacion.fca.unam.mx        |          | 19           |              |              |             |             |            | 19            |
| Ste<br>anexowww.alafec.unam.mx |          | 45           |              |              |             |             |            | 45            |
| www.anfec.unam.mx              |          | 148          | 4259         | 38           | 2           |             |            | 4447          |
| www.defensoria.unam.mx         |          | 105          | 108          | 3117         | 911         | 882         |            | 5123          |
| www.dgae.unam.mx               |          | 20           |              | 2            |             |             |            | 22            |
| www.fca.unam.mx                | 1        | 453          |              | 10           |             |             |            | 464           |
| www.pve.unam.mx                |          | 19           |              |              |             |             |            | 19            |
| www.software.unam.mx           |          | 2            |              | 4            |             |             |            | 6             |
| <b>Total general</b>           | <b>1</b> | <b>10855</b> | <b>52907</b> | <b>47700</b> | <b>9258</b> | <b>1354</b> | <b>118</b> | <b>122193</b> |

Fuente: elaboración propia

La tabla 12 proporcionada muestra los resultados de las URL principales y su nivel de profundidad en la página del sitio web [www.fca.unam.mx](http://www.fca.unam.mx). Cada URL se presenta en la columna "URLs", mientras que el nivel de profundidad se muestra en las columnas numeradas del 0 al 6. El nivel de profundidad indica la distancia entre la página principal y la página correspondiente a la URL en cuestión.

Por ejemplo, si observamos la primera URL "[admonesc.fca.unam.mx](http://admonesc.fca.unam.mx)", podemos ver que tiene una profundidad de 0. Esto significa que esta URL se encuentra en la página principal del sitio web. Por otro lado, si nos fijamos en la URL "[cetus.fca.unam.mx](http://cetus.fca.unam.mx)", podemos ver que tiene una profundidad de 2. Esto indica

que esta URL se encuentra a dos niveles de profundidad desde la página principal.

La tabla también muestra el recuento de URLs en cada nivel de profundidad. Por ejemplo, en el nivel de profundidad 0, solo se encuentra la URL principal del sitio web ([www.fca.unam.mx](http://www.fca.unam.mx)), que aparece con un recuento de 1 en la columna correspondiente.

A medida que nos desplazamos hacia los niveles de profundidad más altos, podemos ver que el recuento de URLs aumenta. Esto sugiere que a medida que nos alejamos de la página principal, se generan más páginas adicionales o secciones dentro del sitio web.

En resumen, la tabla presenta una visión general de la estructura del sitio web [www.fca.unam.mx](http://www.fca.unam.mx), mostrando el número de URLs en cada nivel de profundidad. Esto puede ser útil para comprender la organización y la jerarquía de las páginas dentro del sitio web, así como para analizar la distribución y la complejidad de la estructura de navegación.

En las gráficas 5 y 6 se ilustra el contenido de la tabla 12



La Gráfica en 3D del nivel de profundidad con las URLs de la página de [www.fca.unam.mx](http://www.fca.unam.mx) ofrece una representación visual de la estructura de profundidad de las páginas web relacionadas con el sitio principal. En esta gráfica, el eje X representa las páginas web específicas, el eje Y muestra la cantidad total de URLs, y el eje Z indica el nivel de profundidad, que es la cantidad de clics necesarios para llegar a las subpáginas desde la página principal. A continuación, explicaré la gráfica:

En el eje X, encontramos una lista de páginas web específicas relacionadas con [www.fca.unam.mx](http://www.fca.unam.mx).

En el eje Z, los niveles de profundidad están representados desde 0 hasta 6. Un nivel de profundidad 0 indica que la página es la página principal, y los niveles posteriores representan la cantidad de clics necesarios para llegar a una subpágina específica desde la página principal.

Los valores en la tabla indican la cantidad de URLs que se encuentran en cada nivel de profundidad para cada página web específica. Por ejemplo, si observamos la página "[ponteenlinea.fca.unam.mx](http://ponteenlinea.fca.unam.mx)", podemos ver que en el nivel de profundidad 1 tiene 3 URLs, en el nivel de profundidad 2 tiene 4,893 URLs, en el nivel de profundidad 3 tiene 6,568 URLs, en el nivel de profundidad 4 tiene 4,378 URLs, en el nivel de profundidad 5 tiene 472 URLs y en el nivel de profundidad 6 tiene 118 URLs. Esto muestra que esta página tiene una estructura bastante profunda y compleja, con una gran cantidad de subpáginas a varios niveles de profundidad.

Algunas páginas, como "[www.fca.unam.mx](http://www.fca.unam.mx)", son la página principal y tienen una cantidad significativa de URLs en el nivel de profundidad 1 y algunos en niveles superiores.

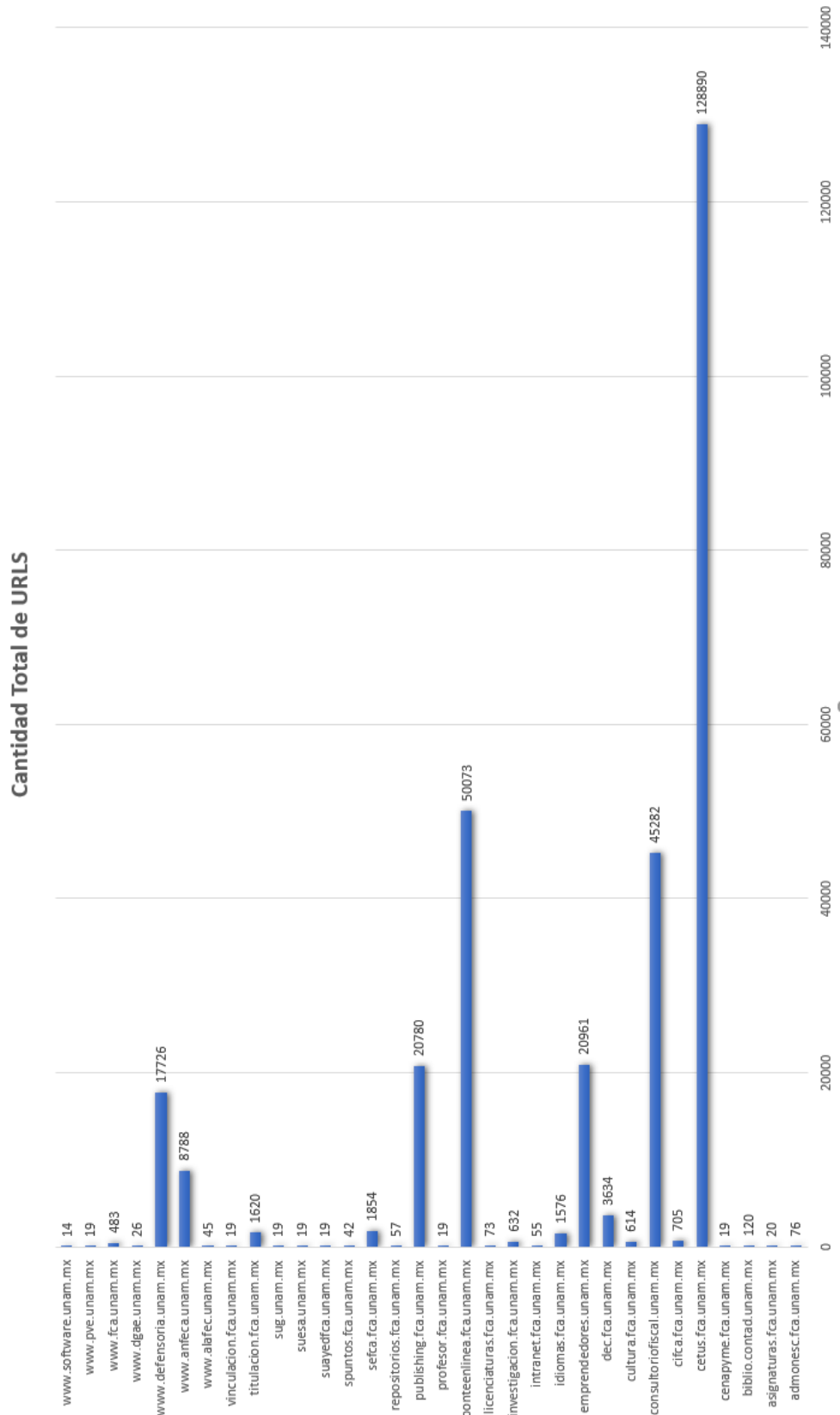
Otras páginas, como "[biblio.contad.unam.mx](http://biblio.contad.unam.mx)", tienen una estructura más simple con URLs en niveles de profundidad 1 y 2.

Las páginas "[dec.fca.unam.mx](http://dec.fca.unam.mx)" y "[publishing.fca.unam.mx](http://publishing.fca.unam.mx)" tienen una estructura interesante, ya que tienen un gran número de URLs en el nivel de profundidad 0 y algunas en niveles superiores, lo que sugiere que pueden contener enlaces directos a recursos importantes.

Considerando lo anterior esta gráfica en 3D proporciona una representación visual detallada de la estructura de profundidad de las páginas web relacionadas

con [www.fca.unam.mx](http://www.fca.unam.mx), lo que puede ayudar a comprender cómo se organiza y accede a la información en el sitio web, así como la cantidad de recursos disponibles en cada nivel de profundidad.

Gráfica 6 Cantidad de URLs totales de los sitios principales de la página [www.fca.unam.mx](http://www.fca.unam.mx)



Fuente: elaboración propia

La gráfica 6 ofrece una representación visual de la cantidad de URLs asociadas a varios sitios web que forman parte de la página principal de la Facultad de Contaduría y Administración (FCA). En esta gráfica, el eje Y muestra la cantidad total de URLs, mientras que el eje X enumera los nombres de los sitios web específicos.

Este tipo de representación gráfica es útil para comprender la estructura y la complejidad de un sitio web, así como para analizar la distribución de contenido en diferentes secciones o subsitios. A continuación, se presenta una explicación detallada de algunos aspectos clave de la gráfica:

**Diversidad en la cantidad de URLs:** La gráfica revela una amplia variación en la cantidad de URLs entre los diferentes sitios web. Algunos sitios, como "cetus.fca.unam.mx" y "ponteonlinea.fca.unam.mx", tienen una cantidad muy alta de URLs 128,890 y 50,073 URLs respectivamente, lo que sugiere que son sitios web extensos y complejos. Por otro lado, sitios como "www.software.unam.mx" y "vinculacion.fca.unam.mx" tienen un número significativamente menor de URLs 14 y 19 URLs respectivamente, indicando que pueden ser más específicos o menos extensos en contenido.

**Complejidad estructural:** La cantidad de URLs puede reflejar la complejidad de la estructura de cada sitio web. Sitios con una gran cantidad de URLs pueden tener una navegación más compleja y ofrecer una amplia gama de recursos y contenido. Por otro lado, los sitios con un número limitado de URLs pueden tener una estructura más simple y centrarse en proporcionar información específica.

**Diferencias en el propósito y contenido:** Las diferencias en la cantidad de URLs también pueden estar relacionadas con el propósito y el contenido de cada sitio. Por ejemplo, un sitio web como "consultoriofiscal.unam.mx" con 45,282 URLs probablemente esté destinado a proporcionar una amplia gama de información relacionada con asuntos fiscales, mientras que sitios como "www.software.unam.mx" y "vinculacion.fca.unam.mx" pueden tener un enfoque más específico.

**Gestión y mantenimiento:** La cantidad de URLs puede influir en la gestión y el mantenimiento de un sitio web. Sitios con un gran número de URLs pueden requerir una atención constante para garantizar que el contenido esté actualizado y sea accesible para los usuarios.

Por lo que, esta gráfica ofrece una visión general de la diversidad en la cantidad de URLs entre los sitios web principales de la FCA de la UNAM. Esto puede ser útil para la administración y la planificación de contenido, así como para comprender la complejidad de la presencia en línea de la facultad.