UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

ANÁLISIS DEL SISTEMA CIUDADANO DE MONITOREO DE ENFERMEDADES RESPIRATORIAS –REPORTA CON MINERÍA DE DATOS .

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

P R E S E N T A :

NOMBRE DEL ALUMNO:
ROCÍO RODRÍGUEZ RAMÍREZ

TUTORES:
DRA. NATALIA B. MANTILLA BENIERS
DR. CRISTOPHER R. STEPHENS STEVENS

2012

Hoja de Datos del Jurado

1.Datos del alumno	1.Datos del alumno
Apellido paterno	Rodríguez
Apellido materno	Ramírez
Nombre(s)	Rocío
Teléfono	5658 7205
Universidad Nacional Autónoma de México	Universidad Nacional Autónoma de México
Facultad de Ciencias	Facultad de Ciencias
Carrera	Actuaría
Número de cuenta	406015981
2. Datos del tutor	2. Datos del tutor
Grado	Dra.
Nombre(s)	Natalia Bárbara
Apellido paterno	Mantilla
Apellido materno	Beniers
3. Datos del co-tutor	3. Datos del co-tutor
Grado	Dr.
Nombre(s)	Chistopher Rhodes
Apellido paterno	Stephens
Apellido materno	Stevens
4. Datos del sinodal 1	4. Datos del sinodal 1
Grado	Dra.
Nombre(s)	María del Pilar
Apellido paterno	Alonso
Apellido materno	Reyes
5. Datos del sinodal 2	5. Datos del sinodal 2
Grado	M. en C.
Nombre(s)	Raúl
Apellido paterno	Sierra
Apellido materno	Alcocer
6. Datos del sinodal 3	6. Datos del sinodal 3
Grado	M. en C.
Nombre(s)	Elio Atenógenes
Apellido paterno	Villaseñor
Apellido materno	García
7. Datos del sinodal 4	7. Datos del sinodal 4
Grado	Act.
Nombre(s)	Harim
Apellido paterno	García
Apellido materno	Lamont
8. Datos del trabajo escrito	8. Datos del trabajo escrito
Título	Análisis del Sistema Ciudadano de Monitoreo de Enfermedades Respiratorias Reporta con minería de datos.
Número de páginas	80 p.
Año	2012

Agradecimientos

Gracias a mi familia por todo su amor y apoyo incondicional de principio a fin.

Gracias a mis amigos por todos los inolvidables momentos que me han regalado, por la fuerza y alegría que me han inyectado siempre.

Gracias a mis tutores: Natalia y Chris, por todo el apoyo y dedicación en la realización de mi tesis y por darme la oportunidad de pertenecer al proyecto Reporta. Ha sido un honor para mi haber conocido y trabajado con personas de gran calidad humana y profesional como la de ustedes.

Gracias al equipo de Reporta por todas las atenciones brindadas a lo largo de mi estancia en el Centro de Ciencias de la Complejidad UNAM - C3 y por todas las sugerencias positivas para mejora de la tesis. Vic, Cheko y Adolfo, gracias por su amistad y por todas las enseñas computacionales compartidas.

Gracias al ICyT por su apoyo y patrocinio para la realización de este proyecto de tesis.

Índice general

Índice de figuras

Índice de cuadros

Introducción

De acuerdo a la Organización Mundial de la Salud (OMS), la influenza es una enfermedad respiratoria viral, contagiosa y aguda, cuyas manifestaciones características son fiebre, cefálea, dolores musculares, postración, inflamación de la mucosa nasal, dolor de garganta y tos. Los síntomas y los signos difieren en función de la edad de las personas infectadas; los infantes, ancianos y enfermos crónicos corren mayores riesgos (mueren con una fracción mayor de quienes contraen influenza que en otros grupos) por lo que son clasificados como de alto riesgo.

La OMS define a una pandemia como la expansión de una enfermedad infecciosa a lo largo de un área geográficamente muy extensa, a menudo por todo el mundo. Dicha enfermedad debe tener un alto grado de infectabilidad, cierta tasa de mortalidad y fácil contagio de una zona geográfica a otra. Para que pueda aparecer una pandemia es necesario que aparezca un nuevo virus o una nueva mutación de uno ya existente, que no haya circulado anteriormente y que la población no sea inmune a él, que el virus sea capaz de producir casos graves de la enfermedad con una mortalidad significativa y que el virus tenga la capacidad de transmitirse de persona a persona de forma eficaz provocando un rápido contagio entre la población.

La pandemia ocasionada por la influenza se refiere a la ocurrencia masiva de casos, con una elevada tasa de infección y mortalidad, ocasionada por la aparición de un nuevo subtipo de virus A, contra el cual la población no tiene inmunidad natural. Hay que recordar que durante el último siglo la humanidad se vió envuelta por tres pandemias de influenza tipo AH1N1, la primera en 1918, que dejó un saldo de 40 millones de personas a nivel mundial, siendo los más vulnerables los adultos jóvenes. Posteriormente se presentaron otras dos pandemias: en 1957, de influenza tipo AH2N2 y durante 1968, influenza tipo AH3N2, lo que trajo consigo incrementos en los índices de mortalidad. Sin duda alguna éstas fueron las tres pandemias que causaron mayores estragos a nivel mundial, especialmente en la población joven [24].

En marzo de 2009, México se convierte en el epicentro de lo que posteriormente se conocería como la pandemia por influenza AH1N1. La falta de datos de incidencia, hizo evidente la necesidad de mejorar la vigilancia epidemiológica, para así detectar la ocurrencia de comportamientos epidemiológicos anómalos y recabar la información pertinente con rapidez, razón por la que un grupo interdisciplinario con cede en el Centro de Ciencias de la Complejidad (C3) en la UNAM, abrió el portal de "Reporta", un sistema de monitoreo de enfermedades respiratorias cuyo objetivo principal es el de monitorear enfermedades respiratorias a nivel nacional, así como procesar y analizar la información capturada desde internet.

El objetivo de esta tesis es analizar, mediante algoritmos de minería de datos, a la población de usuarios de Reporta para identificar los factores de riesgo con base en sus características sociodemográficas y sintomatológicas a fin de predecir el nivel de riesgo de participantes de nuevo ingreso al sistema. Se usaron cuatro algoritmos para estudiar la muestra: conglomerados, para realizar un recorrido exploratorio de la base; Epsilon, para identificar las principales variables que describen a una clase; clasificador bayesiano ingenuo, para cuantificar la probabilidad de que dichos participantes pertenezcan a distintos grupos de riesgo de acuerdo con criterios epidemiológicos y de frecuencia de participación; y análisis de series de tiempo para estudiar los patrones temporales que caracterizan a las series de tiempo de participación, influenza y gripe.

La tesis está estructurada de la siguiente manera, en el capítulo 1 se describe al proyecto Reporta de manera general, se da a conocer el método de colecta de los datos en el sistema, el diseño de la base de datos y arquitectura del proyecto a nivel computacional; también se realiza un escaneo estadístico de las características de los usuarios con el fin de conocer a la población muestra.

El capítulo 2 presenta el marco teórico de la tesis. Inicialmente se da una breve reseña de lo que es la minería de datos, para después describir y definir los algoritmos que se usarán para el análisis de la población muestra: conglomerados, epsilon, clasificador bayesiano ingenuo, modelo de riesgo y análisis de series de tiempo. Se realiza el análisis exploratorio de datos y quedan definidas las clases de usuarios que se analizarán a lo largo del trabajo, estás son: sospechosos de influenza, usuarios que padecieron gripe y usuarios más participativos.

En el capítulo 3 se implementa computacionalmente los cuatro algoritmos descritos en el capítulo anterior para las clases de usuarios definidas. Se usaron los paquetes de: R y Octave

para optimizar el análisis y manejo de los datos; y Gnuplot para graficar los resultados.

En el último apartado se discuten y concluyen los resultados obtenidos a lo largo de la tesis, también se sugieren posibles acciones para mejora del desarrollo de l proyecto con base en los resultados analizados.

Todas las figuras y cuadros presentados en esta tesis son de elaboración propia, salvo la de entidad relacion del proyecto Reporta (figura 1.3) que fue elaborada por el M. en C. Victor Mireles, anotación que se hace al pie de dicha gráfica.

Capítulo 1

El Proyecto Reporta

En este capítulo se describe al proyecto Reporta: Sistema Ciudadano de Monitoreo de Enfermedades Respiratorias, se da a conocer el método de colecta de los datos en el sistema, el diseño de la base de datos y arquitectura del proyecto a nivel computacional; también se realiza un escaneo estadístico de las características de los usuarios con el fin de conocer a la población muestra.

1.1. Antecedentes

A raíz de la pandemia por influenza AH1N1 que ocurrió en 2009 se hizo evidente la necesidad de mejorar la vigilancia epidemiológica, monitoreando dicho fenómeno, con el fin de detectar la ocurrencia de comportamientos epidemiológicos anómalos y recabar la información pertinente con rapidez. Esta vigilancia epidemiológica debe complementarse con un procesamiento de la información que permita aprovecharla para planear el control eficiente de un brote.

En mayo de 2009, un grupo interdisciplinario con sede en el Centro de Ciencias de la Complejidad (C3) en la UNAM abrió Reporta: Sistema Ciudadano de Monitoreo de Enfermedades Respiratoria, cuyo objetivo principal es el monitorear enfermedades respiratorias a nivel nacional, así como procesar y analizar la información capturada desde su portal en internet http://reporta.c3.org.mx. En este portal se invita a las personas residentes en México a registrarse y llenar dos cuestionarios de opción múltiple. El primer cuestionario sólo se completa una vez y recaba información sociodemográfica del usuario. El segundo, es de llenado semanal y registra la presencia o ausencia de síntomas respiratorios en el participante, así como si éste ha acudido al médico y si tiene un diagnóstico de su enfermedad.

1.2. Método de colecta de datos

En el portal de Internet de Reporta: http: //reporta.c3.org.mx se orienta a quien visita la página en el proceso de su registro. Dicho registro involucra proporcionar un correo electrónico y elegir una contraseña. A continuación, el usuario proporciona información demográfica una única vez: cuestionario inicial (tabla 1.1) y llena semanalmente un breve cuestionario que permite valorar su estado de salud en relación con enfermedades respiratorias: cuestionario rutinario (tabla 1.2). Cuando el usuario termina de contestar el cuestionario de ingreso, se despliega en la pantalla una valoración de riesgo que el sistema calcula con base en los datos sociodemógraficos proporcionados minutos antes por el mismo usuario. Cada vez que completa el cuestionario semanal con su sintomatología (o reporta la ausencia de síntomas) la página emite una gráfica de las series de tiempo del número de participantes con síntomas sospechosos de influenza en el país y el estado de residencia del participante, así como la serie de tiempo que muestra si el propio participante ha tenido síntomas sospechosos de influenza desde que se registró en Reporta. Después de la inscripción, la participación continuada se motiva por medio del envío de correos electrónicos semanales con la liga directa al cuestionario de síntomas. Los datos generados integran una curva de incidencia que aparece en la página y se busca que sirvan para generar un sistema de alerta temprana. Parte de la información sociodemográfica recabada se visualiza en cinco gráficos y su totalidad se utilizará para diseñar modelos adecuados, obtener mapas de riesgo y, eventualmente, para parametrizar modelos predictivos que permitan evaluar las consecuencias de distintas intervenciones.

El proceso de almacenamiento de datos del proyecto Reporta se encuentra estructurada como lo muestra la figura 1.1. Inicialmente, los datos ingresados por los usuarios son registrados en tiempo real en los cuestionarios Inicial y Rutinario, posteriormente son almacenados en la base de datos (BD) de donde se extraen, transforman y descargan (ETL) a través de una aplicación programada en php que devuelve a la misma base dos tablas con la información reestrucutrada: reporteUsuarios y reporteRutinarios. Finalmente, estas tablas constituyen la fuente de procesamiento y análisis de datos.

La base de datos de Reporta está integrada por seis tablas (figura 1.2), en donde se realizarán las consultas para fines de esta tesis. La descripción del contenido de cada una de ellas se detalla en el cuadro 1.3.

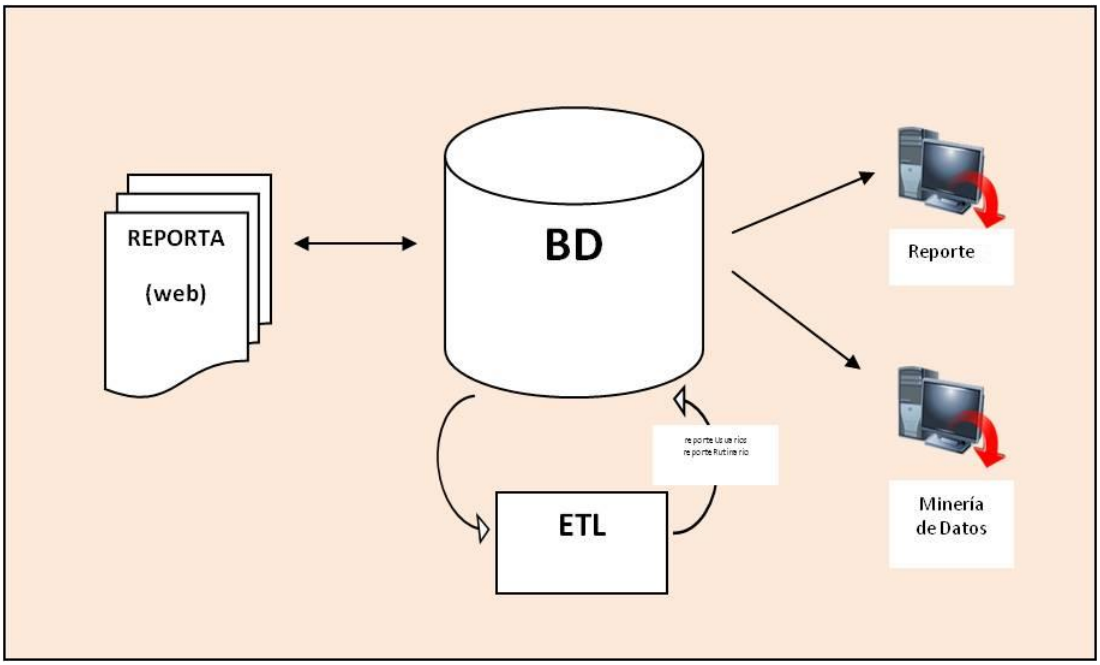

Figura 1.1: Estructura del proyecto Reporta

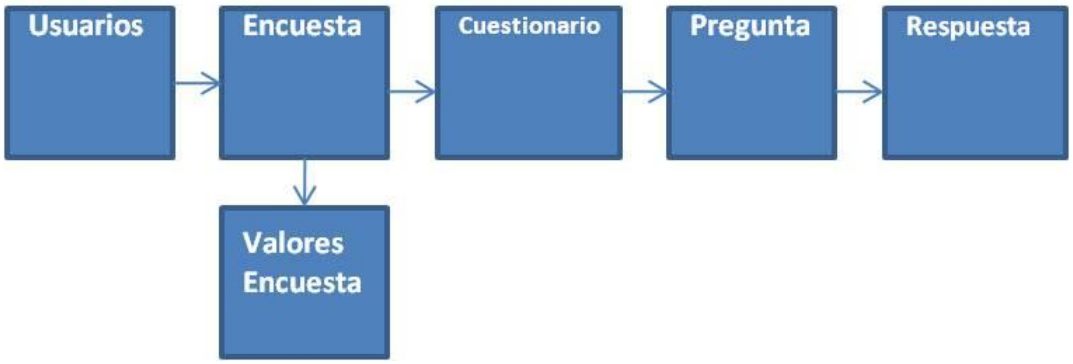

Figura 1.2: Relaciones entre tablas de la base de datos Reporta

Por último, en la figura 1.3 se muestra el diagrama Entidad-Relación (ER) de la base de datos enfocada a las tablas de captura que dan como resultado dos tablas planas de consulta: reporteUsuarios y reporteRutinario que cambian deacuerdo a los cambios actualizados en los cuestionarios.

	Pregunta	Respuesta
1	Sexo	Masculino
		Femenino
2	Año de nacimiento	A cuatro dígitos
3	Mes de Nacimiento	Enero - Diciembre
4	Principal Ocupación	No trabajo, estoy jubilado o pensionado
		Trabajo en casa (incluyendo amas de casa)
		Incapacitado para el trabajo
		Estudiante
		Maestro, investigador o académico
		Trabajador agropecuario (agricultor, ganadero, etc.)
		Trabajador industrial o de la construcción
		Conductor de medios de transporte o maquinaria móvil
		Vendedor ambulante
		Servicio al público (vendedor, mesero, cocinero, cajero, limpieza, etc)
		Trabajador de las fuerzas de seguridad o fuerzas armadas)
		Trabajador del sector salud
		Otro
		Trabajador administrativo que no da servicio directo al público
5	¿Cuál es tu máximo grado de estudios?	Ninguno
		Primaria
		Secundaria
		Preparatoria
		Universidad
6	¿Cuál es tu estado civil?	Casado / Unión Libre
		Soltero
		Viudo
		Divorciado
7	Medio de transporte de uso frecuente	A pie
8		Motocicleta o bicicleta
9		Automóvil o taxi
10		Transporte colectivo (autobús, metro, etc.)
11	¿Con cuántas personas compartes casa?	Ninguna
		1
		2
		3
		4
		5
		más de 5
12	Animales con lo que estas en contacto frecuentemente	Gatos
13		Pájaros
14		Perros
15		Cerdos
16		Aves de corral (gallinas, guajolotes, patos, etc.)
17		Otros
18	¿A qué tipo de médico puedes ir más fácilmente?	Ninguno
		IMSS
		ISSSTE
		Pemex
		Secretara de salud
		Sedena
		Marina
		Privado
		Otro
19	Típicamente, Cuántas veces te resfrías al año?	Entre dos y cinco
		Más de cinco
		Menos de dos
20	¿Recibiste la vacuna para la influenza (2008-2009)?	No
		Sí, en una campaña de vacunación en mi trabajo.
		Sí, por recomendación de mi médico.
		Sí, porque pertenezco a un grupo de riesgo.
		Sí, por alguna otra razón.
21	¿Con qué frecuencia haces ejercicio físico?	Menos de una hora por semana
		Una a cuatro horas por semana
		Más de cuatro horas por semana
22	Enfermedades crónicas que padece	Respiratorias (asma, enfisema, bronquitis, etc)
23		Cardiovasculares (hipertensión, arritmia, etc)
24		Diabetes
25		Otra
26	¿Eres fumador?	Sí
		Ocasional
		No
27	Acciones tomadas frecuentemente cuando enfermas	Me automedico
28		Tomo lo que la persona de la farmacia me indica
29		Uso homeopatía
30		Uso herbolaria o remedios caseros
31		Uso acupuntura
32		Voy con un médico
33	¿Perteneces a alguna de estas instituciones?	UNAM
		IPN
		otra
		UAM
		UACM

Cuadro 1.1: Preguntas que conforman el cuestionario inicial. Para cada pregunta, el usuario puede seleccionar una sola respuesta , salvo en los casos de: medio de transporte usado con mayor frecuencia, animales con los que tiene contacto frecuente, enfermedades crónicas padecidas y acciones tomadas cuando enferma, en donde puede seleccionar todas las respuestas que describan su caso.

	Pregunta	Respuesta
1	En los últimos siete días, ¿cuáles de estos síntomas se te presentaron?	Congestión o escurrimiento nasal
	(el usuario puede seleccionar más de una opción)	Dolor de cabeza
		Tos
		Dolor muscular
		Dolor de articulaciones (coyunturas)
		Vómito
		Diarrea
		Debilidad
		Dificultad para respirar
		Dolor al tragar
2	¿Tuviste fiebre?	No
		Entre 38y 38.5
		Entre 38.5 y 39
		Más de 39
		Si tuve fiebre pero no sé que temperatura tuve
3	¿Hace cuánto empezaron estos síntomas?	Hoy
		Ayer
		Antier
		Hace 3 días
		Hace 4 días
		Hace 5 días
		Hace 6 días
4	¿Estuviste recientemente con alguien que tuviera síntomas muy parecidos a los tuyos?	Sí
		No recuerdo
5	Si alteraste tu rutina, ¿por cuántos días?	Ninguno
		Uno
		Dos
		Tres
		Cuatro o cinco
		Seis o siete
		Más de una semana
6	¿Faltaste al trabajo o la escuela a causa de tu malestar?	No
		Sí, un día
		Sí, dos días
		Sí, tres días
		Sí, cuatro o más días
7	¿Fuiste al médico?	Sí, dijo que tenía gripe común
		Sí dijo que tenía influenza estacional
		Sí, dijo que tenía influenza AH1N1
		Sí, dijo que tenía otra cosa
		No, no fui
		Sí, pero no me atendieron
8	¿Cuáles de estos medicamentos te recetaron?	Amoxicilina (Amoxil, Amoxifur, etc)
		Claritomicina (Klaricid, Neo-clarosip, etc)
		Eritromicina (Eritropharma-S, Trofarma, etc)
		Gentamicina (Garamicina, Garamsa, etc)
		Doxiciclina (Vibramicina, Genobiotic-doxi, etc)
		Cefotaxima (Fotexina, Biosint, etc)
		Ninguno de los anteriores
		Otro antibiótico
9	¿Y de éstos?	Amantadina
		Rimantadina
		Oseltamivir (Tamiflí)
		Zanamivir (Relenza)
		Otro antiviral
		Ninguno de los anteriores
10	¿Te hicieron análisis de laboratorio? y con ellos te dijeron que tenías:	Influenza tipo A estacional
		Influenza tipo B
		Influenza tipo AH1N1
		Otra cosa
		No me hicieron análisis de laboratorio

Cuadro 1.2: Preguntas que conforman el cuestionario rutinario. Si el usuario no presentó síntoma ni estuvo enfermo, puede dar click en la opción: No presenté ningún síntoma esta semana. Este reporte de no sintomatología ni enfermedad queda registrado en la base de datos.

Tabla	Descripción
Usuario	Datos geográficos del usuario y datos para la identificación y localización del usuario a través del correo electrónico (La identificación de cada usuario es anónima)
Encuesta	Relación de los cuestionarios contestados por los usuarios indicados por fecha.
Cuestionario	Codificación de los tipos de cuestionarios existentes.
Pregunta	Codificación de las preguntas en la que se especifica su texto.
Respuesta	Codificación de la lista de respuestas a una pregunta específica junto con el texto a desplegar.
ValoresEncuesta	Relación que muestra las respuestas introducidas en cada una de las encuestas.
reporteUsuarios	Datos en detalle recabados por el cuestionario inicial a los usuarios que se inscriben en el proyecto.
reporteRutinario	Datos recabados por el cuestionario semanal de síntomas.

Cuadro 1.3: Descripción de las distintas tablas que conforman la base de datos de Reporta.

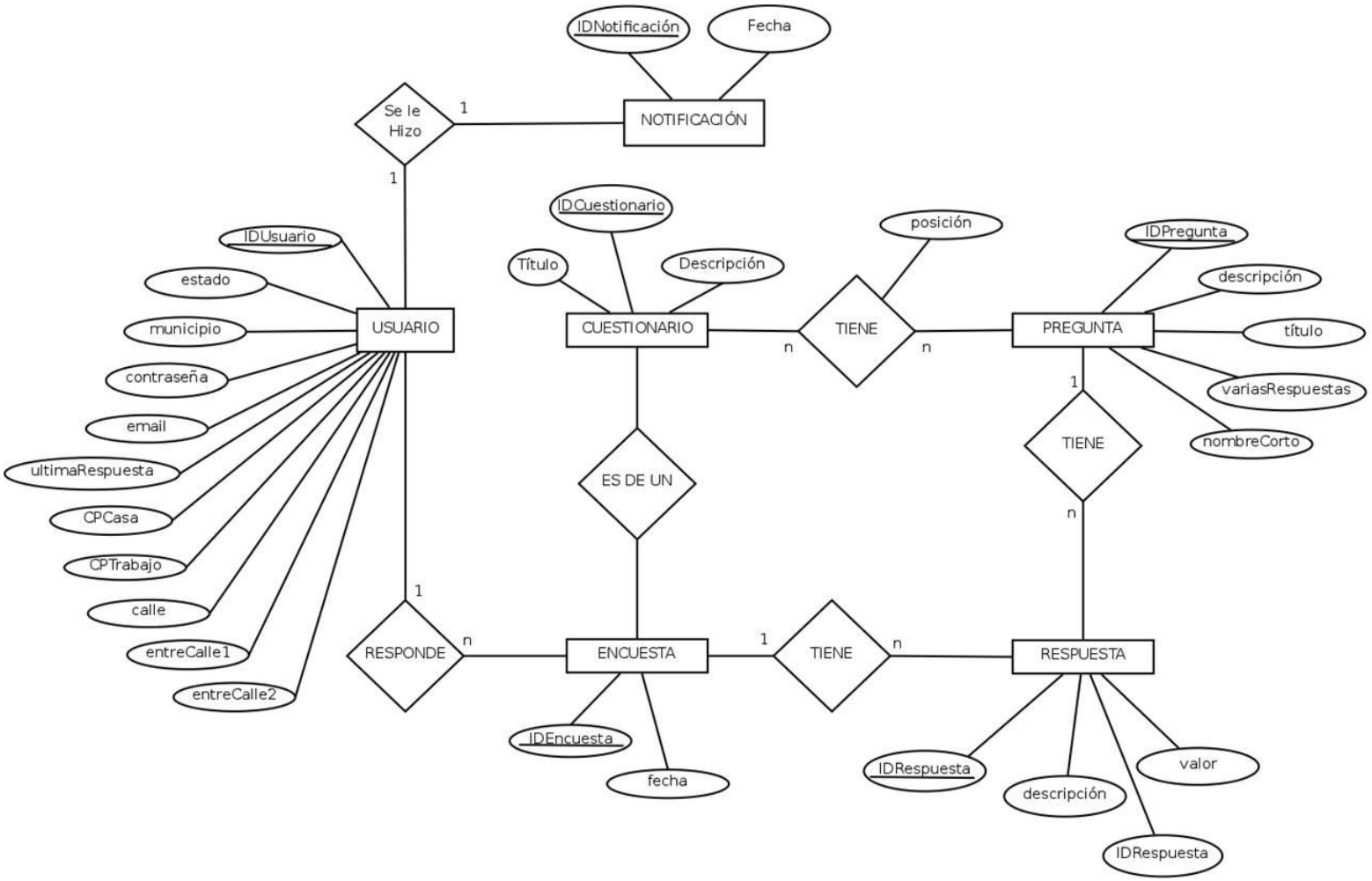

Figura 1.3: Diagrama ER de la Base de Datos REPORTA. Fuente: Elaborada por el M. en C. Victor Mirelles, tomada de un documento descriptivo del proyecto Reporta. Octubre 2010.

1.3. Objetivo del análisis

El objetivo de esta tesis es analizar, mediante algoritmos de minería de datos, a la población de usuarios de Reporta para identificar los factores de riesgo con base en sus características sociodemográficas y sintomatológicas a fin de predecir el nivel de riesgo de participantes de nuevo ingreso al sistema. Se usaron cuatro algoritmos para estudiar la muestra:

- Conglomerados: para realizar un recorrido exploratorio de la matriz de datos con la intención de conocer la existencia de asociaciones entre las variables de estudio. Estas relaciones encontradas serán el punto de partida para la generación de hipótesis del presente trabajo de investigación.
- Epsilon: para identificar las principales variables que describen a una clase.
- Clasificador bayesiano ingenuo: para cuantificar la probabilidad de que los participantes pertenezcan a distintos grupos de riesgo de acuerdo con criterios epidemiológicos y de frecuencia de participación.
- Análisis de series de tiempo: para estudiar los patrones temporales que caracterizan a las series de tiempo de participación, influenza y gripe.

1.4. Definición de la población objetivo

La población objetivo sobre la que se trabajó está conformada por los participantes del Sistema Ciudadano de Monitoreo de Enfermedades Respiratorias-Reporta registrados en la tabla reporteUsuarios desde mayo 2009 a septiembre 2011. Cabe señalar que de los 5515 usuarios registrados durante este periodo, son 4873 los que cuentan con la suficiente información para ser analizados. El motivo por el que se descartaron 642 usuarios de la base de datos para los subsecuentes análisis fue porque no completaron de manera correcta el cuestionario inicial o porque nunca enviaron registro alguno del cuestionario rutinario semanal.

1.5. Diseño de la matriz de datos

Para fines de este análisis se exportó la tabla reporteUsuarios a un archivo .csv, que contiene la información de cada uno de los usuarios que se ha registrado desde mayo de 2009 a septiembre de 2011 y se conviertieron los valores de respuesta numérica a texto para mejor comprensión y rápida lectura al momento del análisis.

La definición de las variables de estudio, así como el diseño de las matrices de datos toman forma y sentido a partir de las respuestas del cuestionario inicial 1.1 y del cuestionario rutinario 1.2 . Las variables sociodemográficas de cada usuario quedaron representadas en una matriz de datos con 4873 registros y 34 campos (variables de estudio). Cuando se necesitaba conocer la sintomatología de alguna clase de usuarios se realizaba una búsqueda directa en la base de datos de Reporta y mediante un join con la tabla reporteRutinario se recababa la información requerida. En la figura 1.4 se muestra el diseño de la matriz de datos a partir de la cual se generaron todos los análisis de esta tesis.

IDUsuario	Sexo	Edad	EstadoCivil	Escolaridad	Ocupacion	Apie	MotoBici	AutoTaxi	Tcolectivo	Cohabitantes	Mascotas	...	HerboCaseros	Acupun	Medico	Institucion
1	Masculino	27	Soltero	Universidad	Estudiante	Si	No	No	Si	1	Si	...	No	No	Si	UNAM
2	Masculino	57	Casado/Union Libre	Universidad	Docente	No	No	Si	No	5	Si	...	No	No	No	UNAM
3	Masculino	29	Soltero	Universidad	Estudiante	No	Si	No	No	2	No	...	No	No	No	UNAM
4	Femenino	28	Soltero	Universidad	Estudiante	Si	No	No	No	3	Si	...	No	No	Si	UNAM
5	Femenino	33	Soltero	Universidad	Docente	No	No	Si	No	1	No	...	No	No	No	UNAM
6	Femenino	28	Soltero	Universidad	Otro	No	No	No	Si	2	No	...	Si	No	No	Otra
8	Masculino	57	Soltero	Universidad	Docente	No	No	Si	No	2	No	...	No	No	No	UNAM
9	Masculino	70	Casado/Union Libre	Universidad	En casa	No	No	Si	No	1	Si	...	No	No	Si	Otra
10	Femenino	36	Soltero	Universidad	Docente	No	No	No	Si	Mas de 5	No	...	No	No	Si	UNAM
11	Masculino	32	Casado/Union Libre	Universidad	Docente	Si	No	No	Si	Ninguno	No	...	Si	No	Si	UNAM
12	Masculino	68	Casado/Union Libre	Universidad	Docente	No	No	Si	No	2	Si	...	No	No	Si	UNAM
13	Femenino	27	Soltero	Universidad	Otro	Si	No	Si	Si	5	Si	...	No	Si	Si	UNAM
14	Femenino	35	Soltero	Universidad	Estudiante	Si	No	Si	Si	4	Si	...	No	No	Si	NA
15	Femenino	24	Soltero	Universidad	Estudiante	No	No	No	Si	1	Si	...	No	No	Si	UNAM
16	Masculino	58	Viudo	Universidad	Docente	No	No	Si	Si	2	Si	...	No	Si	Si	UNAM
17	Masculino	35	Soltero	Universidad	Docente	No	No	Si	No	3	No	...	No	No	No	UNAM
18	Femenino	37	Casado/Union Libre	Universidad	Estudiante	No	No	Si	No	2	Si	...	Si	No	No	Otra
19	Femenino	25	Soltero	Universidad	Estudiante	No	No	Si	No	3	No	...	No	No	Si	Otra
20	Masculino	42	Casado/Union Libre	Universidad	En casa	No	Si	Si	No	2	Si	...	No	No	Si	Otra
21	Femenino	56	Soltero	Universidad	Docente	No	No	Si	Si	Ninguno	Si	...	No	No	Si	UNAM
22	Femenino	33	Soltero	Universidad	Docente	No	No	No	Si	Ninguno	No	...	No	No	Si	Otra
23	Masculino	41	Casado/Union Libre	Universidad	Docente	Si	No	Si	Si	1	No	...	No	No	Si	Otra
.	.	.	.	.	.	.	.	.	.	.	.	...	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	...	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	...	.	.	.	.
5507	Masculino	33	Soltero	Universidad	Docente	Si	Si	Si	Si	1	Si	...	No	No	No	NA
5508	Masculino	19	Soltero	Preparatoria	Estudiante	Si	No	Si	No	3	Si	...	Si	No	No	UNAM
5509	Femenino	36	Soltero	Universidad	Otro	No	No	Si	No	2	Si	...	No	No	No	Otra
5510	Masculino	27	Soltero	Universidad	Otro	No	No	No	Si	3	Si	...	No	No	Si	Otra
5511	Masculino	28	NA	Universidad	Estudiante	Si	No	Si	Si	2	Si	...	Si	No	Si	UAM
5512	Masculino	24	Soltero	Universidad	Estudiante	No	No	No	Si	4	Si	...	No	No	Si	Otra
5513	Masculino	20	Soltero	Preparatoria	Estudiante	Si	No	No	Si	Ninguna	Si	...	Si	No	Si	UAM
5514	Femenino	19	Casado/Union Libre	Universidad	Estudiante	No	No	No	Si	3	Si	...	No	No	Si	UAM
5515	Femenino	29	Soltero	Universidad	Estudiante	No	No	No	Si	NA	Si	...	Si	No	Si	Otra

Figura 1.4: Matriz de datos Tesis-Reporta

1.6. Descripción general de la población

La población participante en Reporta está conformada por un 56.41 por ciento de mujeres, un 42.46 por ciento de hombres y 1.13 por ciento no revela ese dato; reflejando muy cercanamente la proporción de género que existe a nivel nacional (51 por ciento mujeres, 49 por ciento hombres). La diferencia se acentúa aún más en las edades que corresponden a los usuarios del sistema; de acuerdo con la información emitida por el Instituto Nacional de Geografía y Estadística (INEGI) correspondiente al último Censo Nacional 2010 y comparándola con la población por edades de Reporta (figura 1.5), se tiene que que los usuarios que tienden a registrarse más que otros grupos de edad se encuentran entre los 20 y 60 años, también se puede observar que la muestra de estudio difiere mucho en la proporción por edades de la población mexicana. Tenemos una proporción de usuarios entre 20 y 60 años que supera la proporción nacional entre esas edades, caso contrario con las proporciones de infante y púberes; la diferencia menos acentuada la tiene el grupo de personas mayores de 60 años.

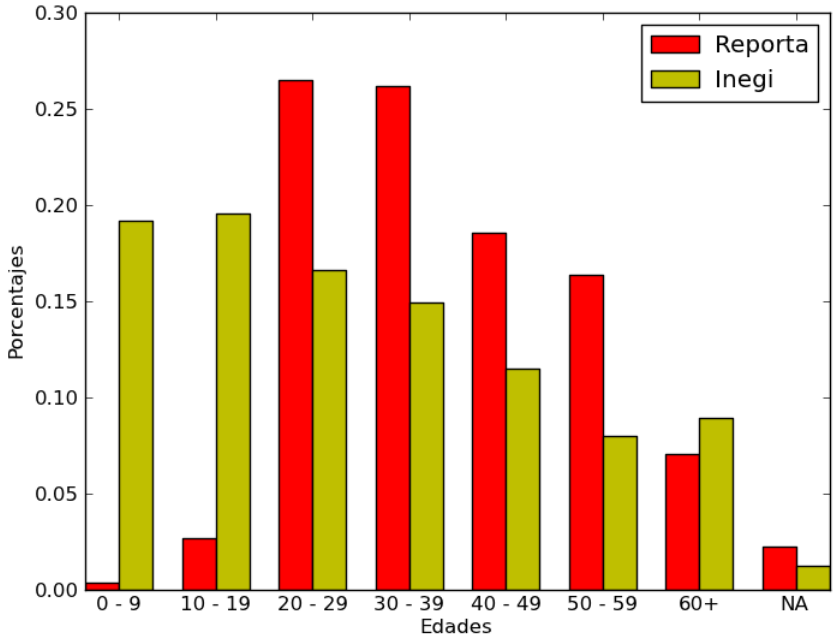

Figura 1.5: Relación por edades Población Reporta / Población Nacional.

En las gráficas 1.6 y 1.7 se puede observar la distribución de la escolaridad y ocupación de los usuarios respectivamente. Un alto porcentaje (85 por ciento) cuenta con formación universitaria y poco más del 50 por ciento pertenece al sector educativo (estudiantes, profesores, académicos, investigadores).

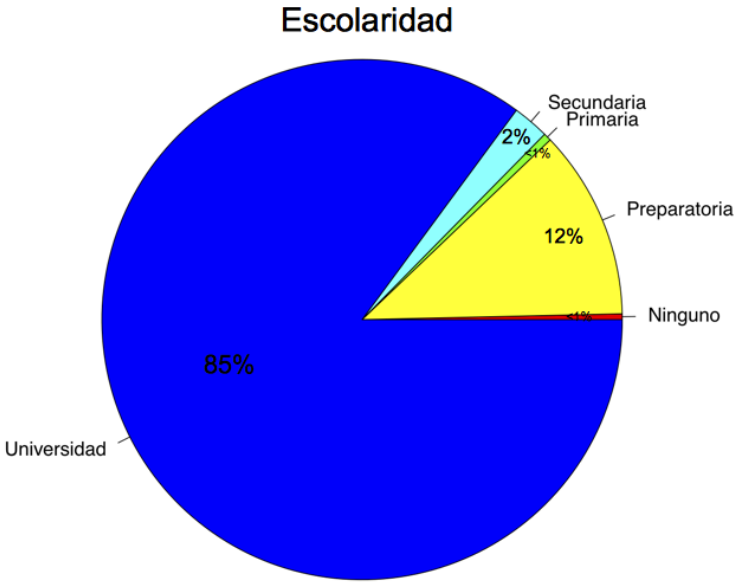

Figura 1.6: Escolaridad Reporta

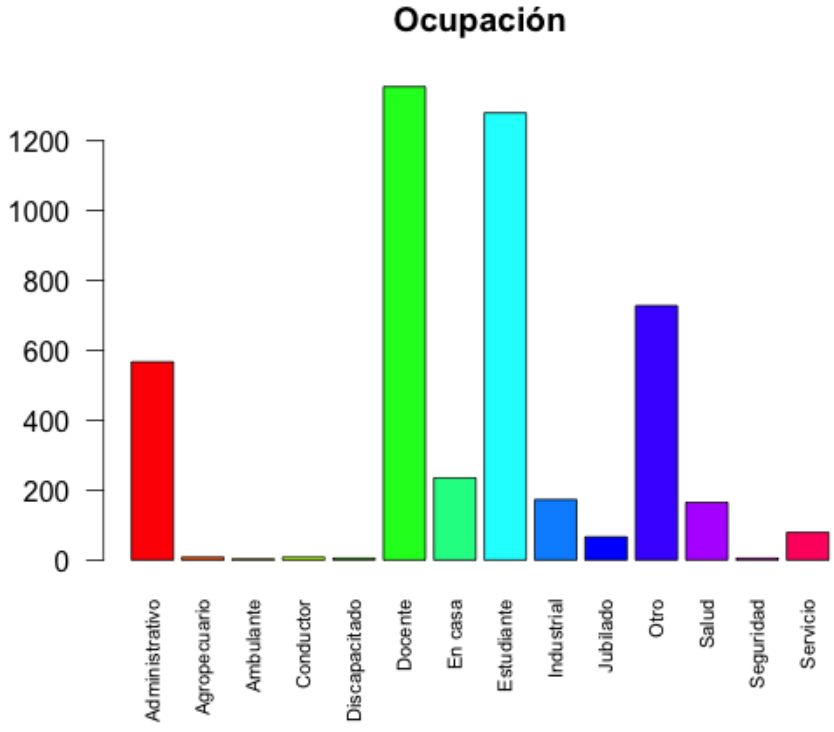

Figura 1.7: Ocupación Reporta

Los estados que tienen mayor participación son: Distrito Federal, estado de México, Morelos y Puebla, probablemente esto se deba a la cercanía que tienen con el Distrito Federal: punto principal de difusión.

Hasta se ha analizado de manera general las variables más representativas de Reporte Usuarios. En el siguiente capítulo se conjuntarán estas variables con las de Reporte Rutinario en una sola base y así comenzaremos la búsqueda de perfiles, grupos y tendencias que describan con mayor exactitud la población de Reporta.

Capítulo 2

Minería de Datos

Este capítulo presenta el marco teórico de la tesis, inicialmente se da una breve reseña de lo que es la minería de datos, para después describir y definir los algoritmos que se usarán para el análisis de la población muestra: conglomerados, epsilon, clasificador bayesiano ingenuo y análisis de series de tiempo. Se realiza el análisis exploratorio de datos y quedan definidas las clases de usuarios que se analizarán a lo largo del trabajo, estás son: sospechosos de influenza, usuarios que padecieron gripe, usuarios que padecieron influenza (según la confirmación en pruebas de laboratorio) y usuarios más participativos.

2.1. KDD y Minería de Datos

Por sus siglas en inglés el KDD es el Descubrimiento del Conocimiento en Bases de Datos, que se define como el proceso que corresponde al acceso, exploración, preparación, modelado y monitoreo de modelado de las bases de datos. Dentro de este proceso se encuentra la minería de datos, que se define como el uso de algoritmos de conocimiento automatizado para encontrar patrones de relación entre los elementos de una base de datos [1].

Según Dunham [2], el proceso KDD consiste en los siguiente cinco pasos:

1. Selección: es la obtención de la información a la que se le hará minería de datos. Dicha información está contenida en una base de datos.

2. Preprocesamiento: es el arreglo que se realiza a la base de datos con la finalidad de corregir errores, datos faltantes u anomalías.

3. Transformación: es el cambio de formato para el procesamiento de los datos.

4. Minería de datos: es la aplicación de algoritmos a la base de datos transformada para analizar los patrones que se presentan.

5. Interpretación: Lectura de los resultados obtenidos en términos de las variables de la base y del análisis que se realizó.

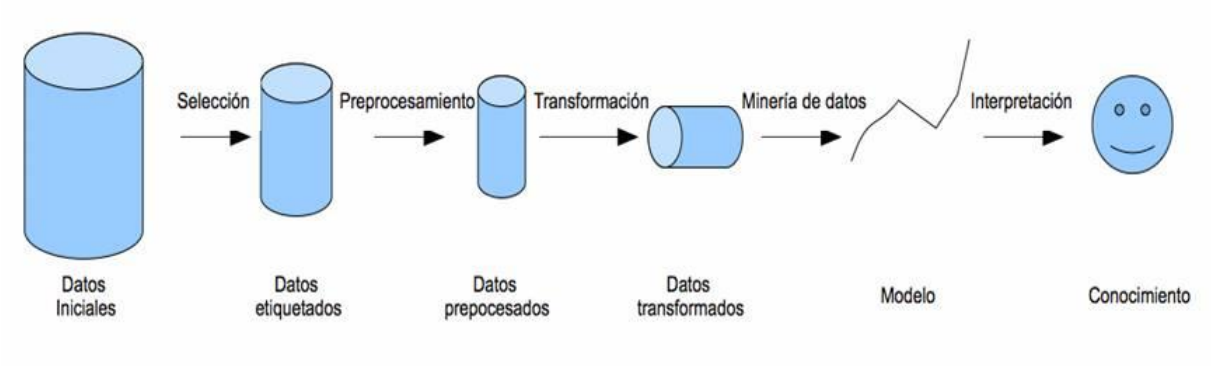

Figura 2.1: Proceso KDD

Para este proyecto de tesis, el proceso KDD se estableció de la siguiente manera:

Selección: La base de datos a analizar se extrajo mediante SQL directamente de las tablas existentes en la base de datos de Reporta. Nos interesa caracterizar a la población participante (en contraste con quienes no participan), conocer si hay patrones temporales distinguibles en su participación, así como analizar su sintomatología y vincularla a factores sociodemográficos que permitan detectar grupos de riesgo. También es importante saber qué regularidades temporales existen en la presentación de síntomas y enfermedades respiratorias, por lo que se trabaja con dos tablas planas extraídas de la base de datos del proyecto Reporta: la tabla de reporteUsuarios y la de reporteRutinario. La tabla reporteUsuarios contiene las variables sociodemográficas contestadas por el usuario en el cuestionario que llena al momento de resgistrarse como participante del proyecto Reporta mostrado en la tabla 1.1. La tabla reporteRutinario está integrada por el cuestionario que el usuario llena semanalmente, mostrado en la tabla 1.2, en el cuál reporta la síntomas que presento o no a lo largo de la semana anterior.

Preprocesamiento: El análisis exploratorio de datos (AED) se realiza con el software estadístico R. Inicialmente se convierten las categorías representadas mediante números a texto y posteriormente se realiza el análisis univariante.

Transformación: Los formatos usados fueron .csv separado por ";"para el análsis de datos en R, .dat para graficas en gnuplot y .txt para el cálculo de periodicidad en Octave.

Minería de datos: Los algoritmos que se aplican para el análisis de los datos son: conglomerados, épsilon, clasificación bayesiana ingenua y técnicas de análisis de series de tiempo.

Mediante la elaboración de conglomerados se pretende segmentar a la población en distintos grupos permitiendo que se revelen nexos entre rasgos de la población muestra y su frecuencia de participación, su deserción o la presencia de ciertos cuadros de síntomas. Mediante el análisis bayesiano se busca obtener predicciones de riesgos ligados a distintos factores sociodemográficos. Con el análisis de series de tiempo se caracterizan los patrones temporales de síntomas respiratorios en distintas localidades a fin de determinar si existe estacionalidad y si hay correlaciones entre el momento de presentación de distintos síntomas, también se analizará sobre el total de sospechosos, con la intensión de detectar tendencias y ciclos a nivel global.

2.2. Análisis exploratorio de datos (AED)

El análisis exploratorio de datos es el tratamiento (estadístico y gráfico) que permite explorar grandes cantidades de información en bases de datos con el fin de evaluar la calidad y consistencia de la información, resumir la información mediante diferentes estadísticas y gráficos, evaluar la necesidad de realizar transformaciones en las variables de interés asi como conocer la distribución de las mismas, detectar y abordar los valores faltantes y puntos extremos [3].

El tratamiento inicial estadístico de la matriz de datos se complementará con el uso del algoritmo de conglomerados para conocer la existencia de asociaciones entre las variables de estudio.

Cabe señar, que el análisis exploratorio de datos es el punto de partida para la generación de hipótesis de investigación [4].

2.2.1. Escalas de medición

Antes de comenzar el AED es fundamental conocer la escala de medición de las variables de estudio, pues de ello depende el tipo de técnica estadísitica que se usará para el análisis univariado, bivariado o multivariado [5].

Las variables pueden ser clasificadas como cuantitativas o cualitativas, dependiendo de si los valores presentados tienen un orden de magnitud natural o simplemente un atributo. En función de esta clasificación existen distintos tipos de escalas [6].

En el diagrama de la figura 2.2 se muestra claramente la clasificación de las variables según Hair [7], y en el texto e describen sus rasgos distintivos en detalle.

Figura 2.2: Clasificación de Variables

Variables Cualitativas o Categóricas

Son aquellas cuyos valores no pueden cuantificarse de forma significativa. Este tipo de variables se puede clasificar por su escala de medición en nominales y ordinales. La **escala nominal** asigna números u otro símbolo a las variables a fin de clasificarlas sin darle un significado cuantitativo. En este tipo de escala la asignación de números permite distinguir los objetos. La **escala ordinal**, que además de clasificar las variables establece una relación de orden o jerarquía entre los elementos. En este tipo de escala las calificaciones se ordenan con el empleo de la relación mayor que, menor que, igual que.

Variables Cuantitativas

Son aquéllas donde las características poseen un carácter numérico. Se dividen en intervalares y de razón; éstas se clasifican en discretas y continuas. Las **variables discretas** son aquéllas que toman valores enteros y son enumerables. Las **variables continuas** son aquéllas que pueden tomar infinidad de valores dentro de un intervalo dado. La **escala por intervalos**

corresponde a variables cuya cuantificación tiene significado, y en las que se establece una distancia, a partir de las cual las medidas reflejan grados de cercanía entre categorías. El cero pertenece a esta escala. La **escala de razones** tiene todas las propiedades de la escala de intervalo, excepto que aqí el cero indica la ausencia de la característica que se esté midiendo.

En el caso de la base de datos de Reporta, la matriz de datos está conformada por variable categóricas, algunas de ellas toman valores binarios, otras siguen algún orden. La única que al inicio era numérica, es la variable edad que para una mejor compresión de resultados, se deicidió categorizarla, cada valor fue representado por su rango quinquenal correspondiente.

2.2.2. Tratamiento de datos

La matriz de datos original que se extrajo directamente de la base de datos de Reporta, categorizaba a las variables con valores numéricos, los cuales se reemplazaron por su equivalente a texto para una rápida lectura y comprensión.

En la generación de conglomerados, factores de riesgo y perfilamiento de riesgo, de los usuarios se trabajó con los valores faltantes y puntos extremos del total de registros, ya que ninguno de los algoritmos utilizados para tales fines era sensible a estos casos, tampoco se quiso generar sesgo alguno en los resultados finales con el hecho de rellenar espacios en blanco utilizando técnicas de edición o imputación de datos. En cambio, para las series de tiempo, los valores extremos (ocasionados por caídas en el sistema) se editaron siguiendo el método de promedios móviles para evitar un sesgo en tendencias y en el análisis de estacionalidad.

De los 5,515 usuarios registrados durante el periodo mayo 2009 a septiembre 2011, son 4,873 los que cuentan con suficiente información para ser analizados. El motivo por el que se descartaron 642 usuarios de la base de datos para los subsecuentes análisis fue porque no completaron de manera correcta el cuestionario inicial o porque nunca enviaron registro alguno del cuestionario rutinario semanal.

El ajuste inicial que se realizó a la matriz de datos fue la transformación del valor numérico de las categorías de cada variable por su equivalente en texto y la categorización de los valores numéricos de la variable edad por su rango quinquenal equivalente. Se añadieron las siguientes variables: mascota, grupo de riesgo (derivadas de algunas variables sociodemográficas ya existentes y definidas en la tablas de reporteUsuario) y se crearon las siguientes clases de usuario: participa, sospechosos, gripe, influenza (derivadas de los registros sintomáticos y

pruebas de laboratorio reportados por los usuarios en el cuestionario rutinario). Estas nuevas variables y clases son categóricas y tomas valores binarios: sí (pertenencia del usuario a la clase que define la variable) y no (pertenencia a la clase complementaria).

Variables Añadidas	**Descripción**
Grupo de riesgo	Clase conformada por los usuarios que pertenecen a un grupo de riesgo según la definición de la OMS: menores de 2 años o mayores de 65 años o personas con que padezcan enfermedades crónicas.
Mascotas	Clase conformada por usuarios que tienen contacto frecuente con alguno de estos animales: cerdos, aves de corral, gatos, pájaros, perros, gatos y otros.

2.2.3. Tablas y gráficos

En el cuadro 2.2 se muestra el resumen estadístico de las variables sociodemográficas y de las clases de usuarios en Reporta: *participa, sospechosos, gripe e influenza*. Fueron 4,873 usuarios y con sus respectivos registros los que se consideraron para la descripción de este cuadro.

Para la variable edad, se tiene que la población de usuarios es relativamente joven, hay más usuarios con edades entre 20 a 40 años, y como se mostró en el primer capítulo, esta distribución no es similar con la presentada en la población general a nivel nacional. Hay muy poco usuarios infantes y púberes, así como pocas personas en edad avanzada.

Con base al resumen estadístico de la matriz de datos, se tiene que son mayoría los universitarios estudiantes y dedicados a la docencia, que participan casi en partes iguales personas solteras y casadas y que la mayoría acude al servicio médico privado. Cohabitan en casa con dos o tres personas en primer lugar, esto habla de núcleos familiares conformados por tres o cuatro integrantes; en segundo lugar están los usuarios que cohabitan con otra persona más; son minoría los usuarios que viven solos. El 75 por ciento de los participantes de Reporta convive con alguna mascota en casa, siendo perros y gatos los de su preferencia.

También se puede decir que la mayoría de los usuarios no padecen enfermedades crónicas y dicen presentar menos de 2 resfriados al año, así como no tener el hábito de fumar. Sólo una quinta parte es sospechoso de influenza y menos de la décima parte ha presentado gripe

Clase	Descripción
Participa	Clase conformada por los usuarios que tienen mayor participación en Reporta, criterio que incluye a los usuarios que han participado al menos un 60 por ciento de las veces a lo largo de todo el prooyecto Reporta.
Sospechosos	Clase conformada por los usuarios sospechosos de influenza de acuerdo con la definición de la Secretaría de Salud: fiebre y dolor de garganta o fiebre y tos.
Gripe	Clase conformada por los usuarios que tuvieron gripe de acuerdo con la definición de la Organización Mundial de la Salud (OMS: http://www.who.int/topics/influenza/es/): aparición súbita de fiebre alta, dolores musculares, cefalea, malestar general importante, tos seca, dolor de garganta y rinitis. Para definición de esta clase, se tomaron los síntomas de la tabla reporteRutinario más cercanos a las definición de la OMS: dolor muscular, dolor de cabeza, tos, dolor al tragar y temperatura mayor a 38 grados.
Influenza	Clase conformada por los usuarios que dieron positivo a Influenza tipo A estacional, infuenza tipo B e influenzas tipo AH1N1 en sus pruebas de laboratorio.

Cuadro 2.1: Cuatro clases de usuarios en las que se enfocará el análisis de minería de datos. De aquí en adelante aparecerán letra cursiva para identificarlas fácilmente a lo largo de la tesis.

(definida por la OMS). Hubo muy pocos casos de influenza, 41 usuarios enfermaron de ella, la confirmación del padecimiento fue bajo positivos en pruebas de laboratorio.

Por otro lado, la mayoría de los usuarios se ejercita menos de 60 minutos al día, sólo el 17 por ciento le dedica más de 4 horas a la semana y más de la mitad usa taxi, coche o transporte colectivo para transportarse. Un cuarto de los usuarios pertenece al grupo de riesgo definido por la OMS.

Para el tema de la frecuencias de participación por parte de los usuarios en el sistema Reporta tenemos que sólo la décima parte del total de usuarios, mantiene una participación

Variable	**Resumen estadístico**
Sexo	Femenino (2749, 56.41 %), Masculino (2069, 42.46 %, NA (56,1.13 %)
Estado Civil	Casado/Union Libre (2094, 44 %), Divorciado (319, 7 %), Soltero (2241, 48 %), Viudo (54, 1 %)
Escolaridad	Ninguno (17, < 1 %), Primaria (23, < 1 %), Secundaria(107,2 %) Preparatoria(556,12 %)Universidad(4017,85 %)
Ocupación	Administrativo(567,12 %), Agropecuario(8, < 1 %), Ambulante(3,< 1 %), Conductor(8,< 1 %); Discapacitado(5,< 1 %), Docente(1355,29 %), En casa(235,5 %), Estudiante(1280,27 %), Industrial(173,4 %), Jubilado(66,1 %), Salud(165,4 %), Seguridad(5,< 1 %), Servicio(79,2 %), Otro(728,16 %)
Medio de transporte habitual	A pie: Si (877, 18 %),No (3996, 82 %), Moto o bici: Si (176, 4 %),No (4697, 96 %), Auto o taxi: Si (3295, 68 %), No (1578, 32 %), Transporte Colectivo: No (2814, 58 %), Si (2059, 42 %)
Cohabitantes	Ninguno(349,7 %), Uno (894,19 %), Dos (994,21 %), Tres (1206,25 %), Cuatro (665,14 %), Cinco (334,7 %), Más de cinco(302,6 %)
Ejercicio	Menos de 1h/sem (2086, 44 %), 1-4 hrs/sem (1890, 40 %), Más de 4 hrs/sem (802, 17 %)
Servicio médico al que tiene acceso	Ninguno(250,5 %), IMSS(743,16 %), ISSSTE(528,11 %), Marina(7,< 1 %), Pemex(31,1 %), Privado(2699,57 %), Secretaria de Salud(172,4 %), Sedena(10,< 1 %), Otro(276,6 %)
Aplicación de la vacuna contra influenza en 2008	No (3770, 79 %), Sí, campaña de vacunación trabajo (422, 9 %) Sí, pertenezco a un grupo de riesgo (114, 2 %), Sí,por recomendación médica (152, 3 %) Sí, por alguna otra razón (288, 6 %)
Resfriados por Año estimados por el participante	Menos de 2 (2725, 57 %), Entre 2 y 5 (1778, 37 %), Más de 5 (241, 5 %)
Grupo de riesgo	No (3587, 74 %), Si (1286, 26 %)
Enfermedades Crónicas	Enf. crónicas respiratorias: Sí(404, 8 %), No (4469, 92 %), Enf. crónicas cardiovasculares: Sí (357, 7 %), No (4516, 93 %), Diabetes: Sí (107, 2 %), No (4766, 98 %), Otra: Sí(389, 8 %), No (4484, 92 %)
Fumador	No (3201, 68 %), Ocasional (802, 17 %), Si (697, 15 %)
Acciones que realizan los usuarios cuando enferman	Automedico: Sí(2330, 48 %), No (2543, 52 %), Farmacia: Sí (310, 6 %), No (4563, 94 %), Homeopatia: Sí(572, 12 %),No (4301, 88 %), Herbolaria/remedios caseros: Sí (1158, 24 %) No (3715, 76 %), Acupuntura:Sí (168, 3 %), No (4705, 97 %), Procura tención médica: Sí(3297, 68 %),No (1576, 32 %)
Mascotas	Sí (3658, 75 %), No (1215, 25 %)
Institución	UNAM (1762, 39 %),UACM (106, 2 %), UAM (96, 2 %),Otra (2576, 57 %)
Clase	**Resumen estadístico**
Participa	No (4412, 91 %), Sí(461, 9 %)
Sospechosos	No (3894, 80 %), Sí (979, 20 %)
Gripe	No (4507, 92 %), Sí (366, 8 %)
Influenza	No (4832, 99 %), Sí(41, 1 %)

Cuadro 2.2: Proporción de las variables y clases de estudio respecto a la población total

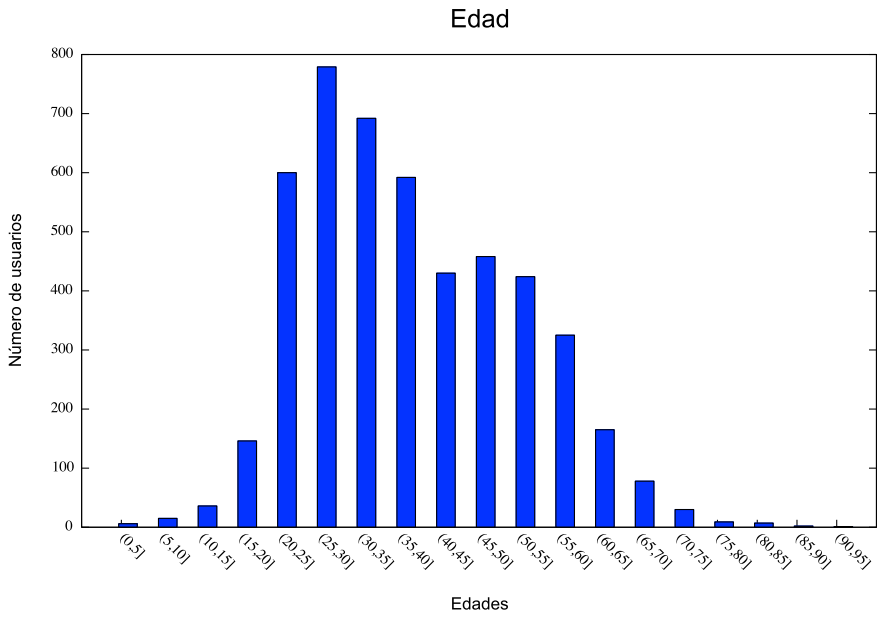

Figura 2.3: Gráfica de frecuencias de las edades de los usuarios de Reporta, dividida por quinquenios

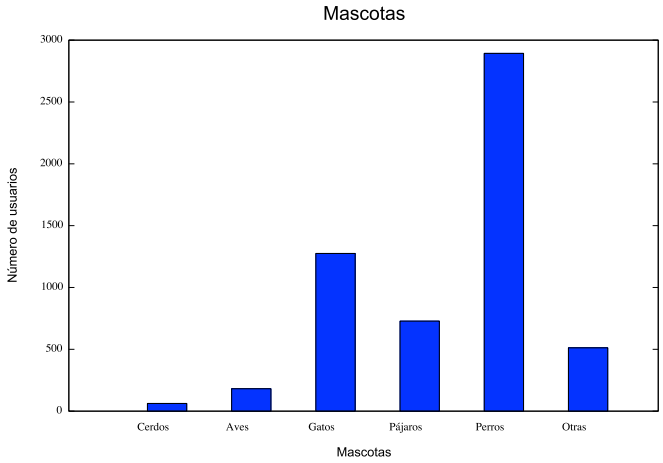

Figura 2.4: Gráfica de frecuencias de los usuarios que conviven con mascotas

activa al contestar semanalmente el cuestionario rutinario.

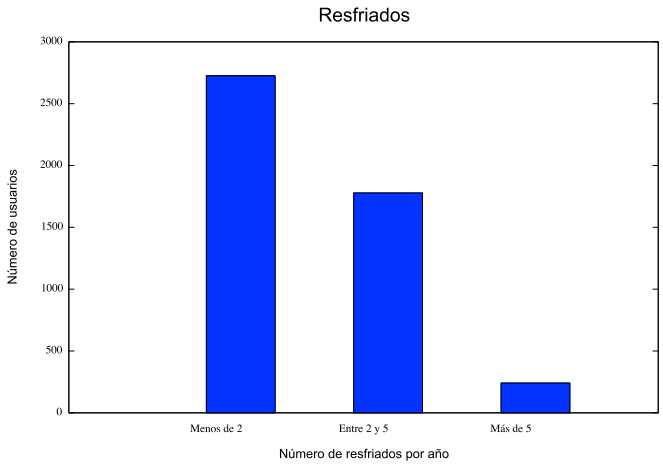

Figura 2.5: Gráfica de frecuencias del número de resfriados por año que padecen los usuario de Reporta

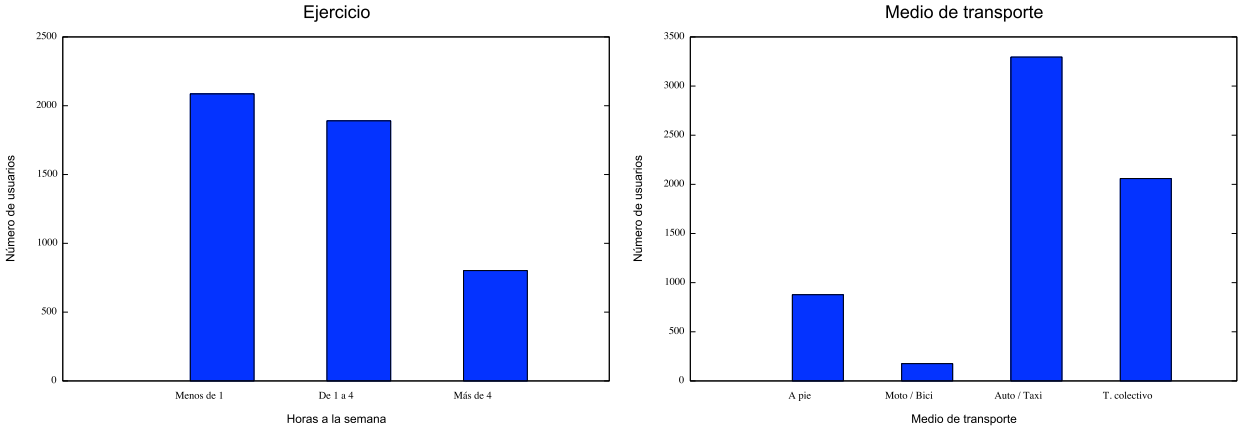

Figura 2.6: Gráfica de frecuencias del número de horas que los usuarios de Reporta hacen ejercicio y del medio de transporte que usan.

2.2.4. Análisis de conglomerados

El análisis de conglomerados es un método de clasificación no supervisado (sin conocimiento a priori de las clases de datos) cuyo principal propósito es agrupar objetos basándose en las características que poseen, de tal forma que cada objeto es muy parecido a los que hay en el conglomerado al que pertenece con respecto a algún criterio de selección. Los conglomerados de objetos resultantes deben mostrar un alto grado de homogeneidad interna (dentro del conglomerado) y alto grado de heterogeneidad externa (entre conglomerados) [10].

La matriz de datos se estructura de manera bidimensional con los n individuos u objetos

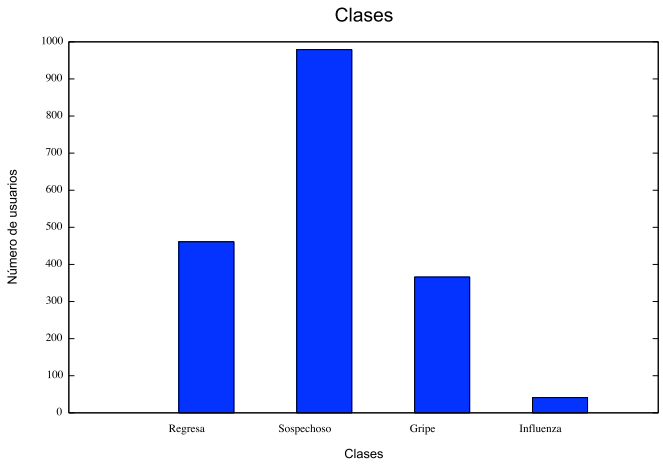

Figura 2.7: Gráfica de frecuencias las distintas clases de usuarios

en la muestra y las p variables observadas, es decir, X_{np} [11].

El proceso según Hair [7] es el siguiente:

1. Se parte de una matriz de datos de n individuos por p variables.
2. Se establece un criterio de similaridad para poder determinar la matriz de similaridad que permitirá relacionar los individuos entre sí (matriz de n individuos por n individuos).
3. Se establece un método de clasificación para agrupar a los individuos.
4. Se crea la estructura de clasificación mediante diagrama (dendogramas).

2.2.4.1. Criterios de Similaridad

Según Kaufman en [13] las medidas de similaridad miden el grado de semejanza entre dos objetos de forma que, cuanto mayor es su valor, mayor es el grado de similaridad existente entre ellos y con más probabilidad los métodos de clasificación tenderán a ponerlos en el mismo grupo. Se define la matriz de similaridad como S = [s(i,j)].

Las medidas de similaridad se rigen por los siguientes axiomas:

- $0 \leq s(i,j) \leq 1$
- $s(i,i) = 1$

- s(i,j) = s(j,i)

Dos objetos son más parecidos si el valor de s que se les atribuye, s(i,j), es cercano a 1.

Por otro lado, siendo D la matriz de disimilaridad, D= [d(i,j)], la disimilaridad cumple con los siguientes axiomas:

- d(i, j) $\geq$ 0
- d(i, i) = 0
- d(i, j) = d(j, i)

Se define la relación entre Similaridad y Disimilaridad de la siguiente manera: d(i,j)= 1 - s(i,j)

2.2.4.2. Medida de distancia para datos mixtos: variables numéricas, categóricas y binarias

La medida de distancia se selecciona de acuerdo al tipo de variables a analizar, en este caso, la matriz de datos de Reporta presenta datos mixtos, es decir, variables numéricas, categóricas y binarias. En presencia de tales variables, los artículos de Gonąlves [14] y Chávez [15] proponen calcular la matriz de distancia entre individuos a partir del coeficiente de similaridad de Gower, que se define de la siguiente manera:

$$S_{ij} = \frac{\sum_{k=1}^{p} W_{ijk} S_{ijk}}{\sum_{k=1}^{p} w_{ijk}} \tag{2.1}$$

Donde: W_{ijk} toma el valor de 0 o 1, cero si la variable k es desconocida para uno o ambos individuos, y en binarias para dobles ausencias; 1 en caso contrario.

S_{ij} : toma el valor de 0 o 1, será 1 cuando los dos individuos tienen el mismo valor y 0 en otro caso.

En variables cuantitativas se segue la siguiente fórmula:

$$S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k} \tag{2.2}$$

Donde R_k representa el rango de la k-ésima observación y se calcula restando a la observación mayor la menor, es decir $R = x_{max} - x_{min}$.

La ventaja de usar este coeficiente de similitud, es que es posible trabajar con bases de datos en las que faltan observaciones de algunas variables, sin necesidad de usar ningún método de imputación o edición.

2.2.4.3. Métodos de agrupamiento

Según Zheng [16], los métodos de agrupamiento o conglomerados se dividen en particionales y jerárquicos. La técnica particional obtiene una sola partición de los datos, donde previamente ya se ha elegido el número de conglomerados deseados. Dicha partición se realiza mediante la optimización de una función criterio ya sea de manera local o global.

La técnica jerárquica no requiere de ningún tipo de especificación inicial en cuanto al número de conglomerados a formar y se usa para fines exploratorios; consiste en una serie anidada de particiones creada a partir de una matriz de disimilaridad que da como resultado una estructura visual llamada dendograma. El dendograma es el que sugiere gráficamente el número de conglomerados a formar.

En la figura 2.8 se muestran los distintintos tipos de técnicas para los métodos de agrupamiento jerárquico y particional.

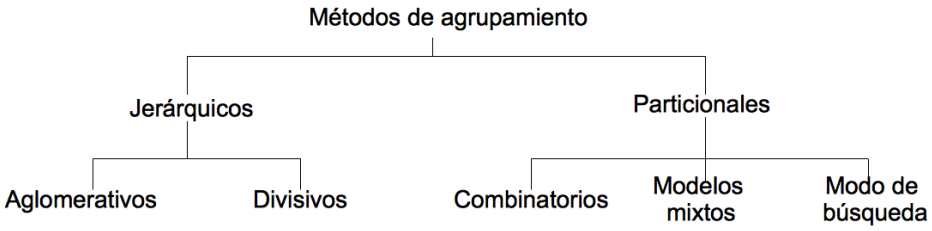

Figura 2.8: Métodos de Agrupamiento

Siguiendo la sugerencia de Gonąlves [14], se usará la técnica de agrupamiento jerárquica aglomerativa Ward, cuyo objetivo es minimizar la pérdida de información dentro de cada conglomerado y cuantificar dicha pérdida para que pueda ser interpretable. En cada iteración del algoritmo, se considera la posible fusión de todos los pares de grupos posibles y se escogen aquellos elementos cuyo incremento de pérdida de información en fusión sea mínimo. Esta pérdida se define con base a la suma de cuadrados mínimos (SSE, Sum of Squared Errors) dentro de cada conglomerado. Inicialmente, cuando todos los elementos son considerados individualmente, SSE=0. La distancia d(i,j) entre los datos univariados i y j es:

$$d(i,j) = \left(x - \frac{x+y}{2}\right)^2 + (y - \frac{x+y}{2})^2 = \frac{(x-y)^2}{2} \tag{2.3}$$

De forma similar, la distancia entre dos conglomerados G_i y G_j con n_i y n_j elementos respectivamente es:

$$d(G_i, G_j) = \frac{n_i n_j}{n_i + n_j} (\tilde{x}_i + \tilde{x_j}) \tag{2.4}$$

Donde x_i es la media de los elementos de G_i y x_j es la media de los elementos de G_j. En cada iteración del algoritmo, los grupos elegidos para fusionarse son aquellos que tienen mínima distancia entre ellos. Cuando dos grupos se fusionan, las distancias del grupo resultante con respecto al resto de grupos, se deben de recalcular.

Los datos serán analizados con el software estadístico R (http://www.r-project.org), usando el paquete çlusterz el algoritmo daisy para generar la matriz de disimilaridad y poder aplicar el método de Ward que finalmente nos estructurará el dendograma de agrupamiento.

2.3. Épsilon

A fin de obtener un perfil o una predicción de la pertenencia de individuos con rasgos descritos en un vector X a una clase dada por un vector C se define una función que se llamará Epsilon. Según Stephens en [17] la idea es identificar para la clase de interés: C, los factores X_i que están más correlaciones con ella, considerando la probabilidad condicional $P(C|X)$ y midiéndola con el punto de referencia P(C) que representa la hipótesis nula; de esta manera, al calcular $P(C|X) - P(C)$, se estará midiendo la incidencia de clase en la población general.

Como se está considerando la pertenencia de clase, cada individuo representa un ensayo Bernoulli (1 = pertence a la clase, 0 = no pertenece a la clase) y la distribución de probabilidad asociada es una distribución binomial, de esta manera, la significancia estadística para $P(C|X) - P(C)$ se puede determinar utilizando la prueba binomial Epsilon:

$$\varepsilon(C|X;C) = \frac{N_X \left[P(C|X) - P(C)\right]}{\sqrt{N_X P(C)(1 - P(C))}} \tag{2.5}$$

Donde:

N_X es el número de observaciones asociadas con el vector de características X.

$P(C|X)$ es la probabilidad de que una observación asociada con el vector de características X

pertenezca a la clase C. Se calcula como el número de observaciones asociadas con el vector de caracterísitcas X y que pertenece a las clase C, entre el número de observaciones asociadas con el vector de características X.
P(C) es la probablidad de que una observación pertenezca a las clase C. Se calcula como el número de muestras que pertenecen a la clase C entre el número total de muestras.

Así, el valor resultante indica cuantas desviaciones estándar se aleja el valor de lo que se observa $N_X P(C|X)$ del valor de lo que se espera observar $N_X P(C)$. Para este caso, como N_X es suficientemente grande y P(C) es constante, el comportamiento de la distribución binomial $B(N_X, P(C))$ se aproxima a una distribución normal de media $\mu = N_X P(C)$ y desviación estándar $\sigma = \sqrt{N_X P(C)(1 - P(C))}$ y como se sabe que en el intervalo $[\mu - 2\sigma, \mu + 2\sigma]$ de una distribución normal se encuentra aproximadamente el 95 por ciento de la distribución, interesaría conocer los casos con $\varepsilon > 2$ ya la diferencia $N_X P(C|X) - N_X P(C)$ no sería resultado del azar y la relación que exista entre X y C resultaría ser significativa.

En términos de un análisis bivariado vamos a considerar las probabilidades condicionales $P(C|X_i X_j)$ y $P(X_i X_j|C)$ en relación con diferentes hipótesis de nulidad que pueden proporcionar información complementaria. Las distribuciones de referencia serán: $P(C)$, $P(C|X_i)$, $P(C|X_j)$ y $P(X_i|C)P(X_j|C)$.

En primer lugar se tiene que $P(C|X_i X_j) - P(C)$ determinará la importancia de la presencia conjunta de las variables X_i y X_j en la pertenencia a la clase en relación con la población general. Por otro lado $P(C|X_i X_j) - P(C|X_i)$ será una medida del efecto de Xj en presencia de Xi, y análogamente para $P(C|X_i X_j) - P(C|X_j)$. Por último $P(X_i X_j|C) - P(X_i|C)P(X_j|C)$ refleja que tan correlacionas están X_i y X_j respecto a la clase.

Las pruebas binomiales asociadas a estas cuatro distribuciones de referencia y que complementan el análisis bivariante son: $\varepsilon(C|X_i, X_j; C)$, $\varepsilon(C|X_i X_j; C|X_i)$, $\varepsilon(C|X_i X_j; C|X_j)$ y $\varepsilon(X_i X_j|C; X_i|C X_j|C)$.

2.4. Clasificación bayesiana ingenua y modelo de riesgo

Según Jiawe en [18], los clasificadores bayesianos son clasificadores estadísticos que pueden predecir tanto las probabilidades del número de miembros de clase, como la probabilidad de que una muestra dada pertenezca a una clase particular. Los clasificadores bayesianos han

demostrado ser muy exactos y rápidos en la construcción de modelos cuando se han aplicado a grandes bases de datos. Estos clasificadores se basan en el teorema de Bayes (2.6)

$$P(C_j|X) = \frac{P(X|C_j)P(C_j)}{P(X)} \tag{2.6}$$

Donde:
$P(X|C_j)$ es la probabilidad de que ocurra X dado que es un patrón perteneciente a la categoría C_j. Generalmente a esta probabilidad se le conoce como verosimilidad de C_j con respecto a X.
$P(C_j)$ es la probabilidad a priori de ocurrencia de la categoría C_j.
$P(X)$ es la probabilidad de que X ocurra independientemente de su categoría y está definido por:

$$P(X) = \sum_{j=1}^{M} P(X|C_j)P(C_j) \tag{2.7}$$

La clasificación se distingue del análisis de conglomerados en que el proceso de conglomerados no dispone de información a priori respecto a la estructura de los datos o la agrupación deseada, mientras que la clasificación sí. La información a priori de que un proceso de clasificación suele disponer respecto a los datos será una definición de las clases dentro de las cuales se desea categorizar en cada caso a los datos [19].

La clasificación bayesiana ingenua (NB, naive bayes, por sus siglas en inglés) es uno de los modelos más simples y más utilizados en bases de datos. Su fundamento principal es la suposición de que todos los atributos son independientes del valor de la variable clase. La hipótesis de independencia que esta técnica toma como cierta supone que en el modelo de una red bayesiana existe un único nodo raíz (clase), y en la que todos los atributos son nodos hoja que tienen como único padre a la variable clase y cada arco representa una dependencia probabilística dada por la probabilidad condicional de cada variable dado su clase [20]. En la figura 2.9 se ilustra de manera gráfica de este modelo.

El clasificador bayesiano ingenuo sigue el teorema de Bayes y asume que las variables que componen el vector X son independientes así, la probabilidad $P(X|C)$ de la ecuación (2.6), bajo el supuesto de independencia, se comportaría de la siguiente manera:

$$P(X|C) = P(x_1|C)P(x_2|C)...P(x_N|C) \tag{2.8}$$

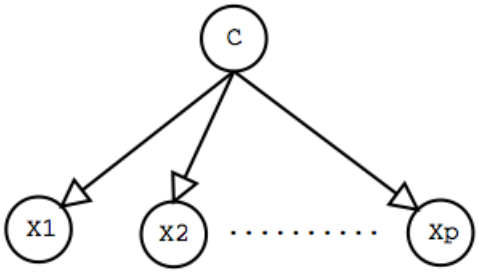

Figura 2.9: Estructura gráfica de un modelo bayesiano ingenuo.

es decir,

$$P(X|C) = \prod_{i=1}^{N} P(x_i|C) \tag{2.9}$$

Donde, N es el número de variables que componen al vector X.
$P(x_i|C)$ es la probablidad de tener la característica definida por la variable i, dada la clase C. Se calcula como el número de muestras que cumplen con la condición x_i y que además pertenecen a la clase C, entre el número de muestras que pertenecen a la clase C.

Conjuntando todo lo anterior, llegamos a que:

$$P_{NB}(C|X) = \prod_{i=1}^{N} P(x_i|C)\frac{P(C)}{P(X)} \tag{2.10}$$

Con la finalidad de predecir el perfil de riesgo de los participantes de nuevo ingreso, se aplica la teoría del clasificador bayesiano ingenuo (NB) antes descrita en la siguiente función:

$$f(X) = \log\left\{\frac{P_{NB}(C|X)}{P_{NB}(\neg C|X)}\right\}, \tag{2.11}$$

Aplicando (2.10) tenemos que:

$$f(X) = \log\left\{\frac{\prod_{i=1}^{N} P(x_i|C)\frac{P(C)}{P(X)}}{\prod_{i=1}^{N} P(x_i|\neg C)\frac{P(\neg C)}{P(X)}}\right\} = \sum_{i=1}^{N} \log\left\{\frac{P(x_i|C)}{P(x_i|\neg C))}\right\} + \log\left\{\frac{P(C)}{P(\neg C))}\right\} \tag{2.12}$$

Dado que $f(X)$ es una función creada para la observación y análisis de las categorías de X, el término $\log\left\{\frac{P(C)}{P(\neg C)}\right\}$ es una constante que se puede descartar para cada vector X, quedando la nueva función de la siguiente manera:

$$S(X) = \log\left\{\frac{P(X|C)}{P(X|\neg C)}\right\} \tag{2.13}$$

para cada $x_i \in X$, nuestra función es:

$$S(x_i) = \log \left\{ \frac{N_{C \cap x_i}/N_C}{N_{\neg C \cap x_i}/N_{\neg C}} \right\}, \tag{2.14}$$

y como cada $x_i \in X$ es independiente entre si, se tiene que:

$$S(x_i, ..., x_M) = \sum_{i=1}^{N} S(x_i) = Score \tag{2.15}$$

Para construir el modelo de riesgo se parte la base de datos en dos, seleccionando al azar 70 por ciento de los usuarios para fines de entrenamiento; el 30 por ciento restante se utiliza para probar los resultados obtenidos en el entrenamiento. Con base en los valores obtenidos $S(x_i)$ de la base de entrenamiento, se calcula el score de cada usuario en la base de prueba, posteriormente se ordenan de manera ascendente y esta lista se parte en 10 subconjuntos equivalente $A_1, ...A_{10}$. Sobre cada A_i se calcula el promedio de los scores, así como la probabilidad de pertenecia a la clase $P(C)$. Finalmente se grafican ambos valores (abscisas $P(C)$ y ordenas Scores promedio) y a partir del comportamiento que tomen estos puntos se trazan las fronteras de riesgo (bajo, medio y alto). Así para predecir el perfil de riesgo de un futuro participante bastará con calcular su score y ubicarlo en la región de riesgo correspondiente.

2.5. Series de tiempo

Una serie de tiempo es una colección de observaciones realizadas secuencialmente en el tiempo [21], su análisis constituye un importante área en estadística. A menudo los datos de las series de tiempo se examinan con la esperanza de descubrir un patrón que se pueda aprovechar para preparar un pronóstico.

Para llegar a tal objetivo, la series de tiempo se analizan desde sus diferentes componentes: tendencia, ciclo, variaciones estacionales y fluctuaciones irregulares.

La **tendencia** se refiere a un cambio suave que indica la orientación de la variable a largo plazo. El **ciclo** es un cambio oscilatorio ocasionado por la dinámica del sistema en estudio. La **estacionalidad** se refiere a un patrón o cambio regular de la serie dentro de un periodo de tiempo. Finalmente, la **irregularidad** corresponde a cambios que carecen de un patrón sistemático o regular, debido a lo cual no es posible pronosticar su comportamiento. Los movimientos irregulares se clasifican como erráticos cuando su causa es identificable (huelgas, terremotos, eventos especiales, etc.) o aleatorios cuando no son atribuibles a algún fenómeno

observable sino al azar [22].

Desde el punto de vista clásico, una serie de tiempo $Y = \{Y_t\}_{t=0}^{T}$ puede ser expresada como:

$$Y_t = f(T_t, C_t, E_t, I_t) \tag{2.16}$$

Donde T_t, C_t, E_t, I_t simbolizan, respectivamente los componentes tendencia, ciclo, estacionalidad e irregularidad en el tiempo t. Conviene considerar a f como una función sencilla y operativa, por lo cual generalmente se trabaja con alguno de los siguientes esquemas:

- Esquema aditivo: $Y_t = T_t + C_t + E_t + I_t$
- Esquema multiplicativo: $Y_t = T_t C_t E_t I_t$

Para conocer el modelo que mejor represente a la serie de tiempo se aplicará el criterio de diferencias y cocientes estacionales. A continuación se definirá cada uno de estos criterios. Sea $Y_{t,i}$ la observación de la variable en cuestión en el año y mes (o estación) i. El cálculo de las diferencias y cocientes estacionales $d_{t,i}$ y $c_{t,i}$ respectivamente se realiza de la siguiente manera:

$$d_{t,i} = Y_{t,i} - Y_{t-1,i} \tag{2.17}$$

$$c_{t,i} = \frac{Y_{t,i}}{Y_{t-1,i}} \tag{2.18}$$

Para hallar los coeficientes de variación se tiene que:

$$CV(d) = \frac{\text{Desviación estándar}(d)}{\text{Media}(d)} \tag{2.19}$$

$$CV(c) = \frac{\text{Desviación estándar}(c)}{\text{Media}(c)} \tag{2.20}$$

Si CV(c) > CV(d), se sugiere seguir el esquema aditivo; si CV(c) > CV(d), uno multiplicativo.

2.5.1. Análisis de la tendencia

La tendencia es el cambio suave que indica la orientación de la variable a largo plazo. En esta tesis, se estimará dicho componente mediante el ajuste de una función que dependa del tiempo. Se calculará un ajuste lineal y polinomial mediante el método de mínimos

cuadrádos, para lo cual, se usará la función de métodos númericos de Octave llamada *polyfit()* que devuelve el polinomio de orden n que mejor se ajusta a los puntos formados por (x,y).

El ajuste lineal estará representado por la ecuación:

$$Y_t = \alpha + \beta t + \epsilon_t. \tag{2.21}$$

Donde, α es la intersección con el eje Y, β es la pendiente estimada o razón de cambio en la frecuencia de la clase a evaluar por unidad de cambio (semanal o mensual) y ϵ_t es el error asociado al ajuste en el valor t.

El ajuste polinomial se calcula para restarlo a la serie de tiempo y sobre este nuevo suavizado se estimará la estacionalidad, método que mostró generar señales más claras en el análisis de estacionalidad que si sólo se le restara la tendencia lineal.

2.5.2. Análisis de la estacionalidad

Según el matemático francés Joseph Fourier (1768-1830) "Toda señal periódica, sin importar cuan complicada parezca, puede ser reconstruida a partir de sinusoides cuyas frecuencias son múltiplos enteros de una frecuencia fundamental, eligiendo las amplitudes y fases adecuadas."

Retomando la idea anterior, se usará la transformada discreta de Fourier para calcular y estudiar la estacionalidad de las series de tiempo. Este método permite el análisis de una cierta señal con un comportamiento periódico (frecuencias, periodos, amplitud, etc.), es decir, brinda información de la señal respecto a la frecuencia de repeticiones de eventos que la describen. Aplicada a una serie de tiempo, la transformada de Fourier muestra directamente su frecuencia característica para poder predecir sus comportamientos periódicos posteriores. Además, brinda una buena técnica de filtrado eliminando frecuencias no deseadas (altas frecuencias, ruido de la señal, interferencias, etc)[23].

En la práctica, la Transformada de Fourier se calcula con un algoritmo llamado Fast Fourier Transform o FFT. Por ejemplo en Matlab o Octave la función para calcular la transformada de Fourier se llama FFF y es la que usaremos en esta tesis para facilitar los cálculos de la misma.

Las frecuencias obtenidas mediante la FFT se pueden representar gráficamente en un periodograma, herramienta que nos permitirá detectar oscilaciones periódicas en el movimiento temporal de la variable en una serie de tiempo y determinar la duración de los periodos oscilatorios.

Capítulo 3

Implementación computacional

El análisis exploratorio de datos realizado en el capítulo anterior permitió tener un panorama más claro de los datos y el marco teórico desarrollado que marca el camino a seguir en el análisis de los mismos. Debido a que la base de datos contiene más de cinco mil registros fue necesaria la implementación de herramientas computacionales especializadas para permitir un óptimo manejo y análisis de la información.
Las herramientas computacionales de apoyamo para el desarrollo del proyecto fueron:

- R para el análisis exploratorio de datos, la obtención del dendograma de conglomerados y la implementación del algoritmo épsilon.
- Octave para el análisis de series de tiempo (tendencia y estacionalidad).
- Matplotlib y GNUPlot para graficar los resultados.

3.1. Clasificación de usuarios

Las 4 clases de usuarios en las que se enfocó el análsis fueron las descritas en el capítulo anterior y que se vuelven a mostrar en el cuadro 3.1.

3.2. AED: Conglomerados

Como se mencionó en el capítulo anterior, abordaremos la matriz de datos de manera exploratoria aplicando un método jerárquico de agrupamiento. Dentro del programa R (http://www.r-project.org), usaremos el paquete çlusterz el algoritmo daisy para generar la

Clase	Descripción
Participa	Clase conformada por los usuarios que tienen mayor participación en Reporta, criterio que incluye a los usuarios que han participado al menos un 60 por ciento de las veces a lo largo de todo el prooyecto Reporta.
Sospechosos	Clase conformada por los usuarios sospechosos de influenza de acuerdo con la definición de la Secretaría de Salud: fiebre y dolor de garganta o fiebre y tos.
Gripe	Clase conformada por los usuarios que tuvieron gripe de acuerdo con la definición de la Organización Mundial de la Salud (OMS: http://www.who.int/topics/influenza/es/): aparición súbita de fiebre alta, dolores musculares, cefalea, malestar general importante, tos seca, dolor de garganta y rinitis. Para definición de esta clase, se tomaron los síntomas de la tabla reporteRutinario más cercanos a las definición de la OMS: dolor muscular, dolor de cabeza, tos, dolor al tragar y temperatura mayor a 38 grados.
Influenza	Clase conformada por los usuarios que dieron positivo a Influenza tipo A estacional, infuenza tipo B e influenzas tipo AH1N1 en sus pruebas de laboratorio.

Cuadro 3.1: Cuatro clases de usuarios en las que se enfocará el análisis de minería de datos.

matriz de disimilaridad y así poder aplicar el método de Ward que finalmente estructura el dendograma de agrupamiento. El resultado se muestra en la figura 3.1. De acuerdo con este dendrograma, los usuarios de Reporta quedan perfilados en ocho conglomerados descritos de la siguiente manera:

- **Conglomerado 1** agrupa a 332 usuarios, que que usan el transporte colectivo con mucha frecuencia para transportarse, no tienen mascotas. No padecen ninguna enfermedad crónica ni pertenecen al grupo de riesgo. Ninguno de ellos usa acupuntura y no se enferman de gripe.

- **Conglomerado 2** agrupa a 211 usuarios con universidad como máximo nivel de estudios, que usan más el auto o taxi para transportarse que el colectivo, sin mascotas,

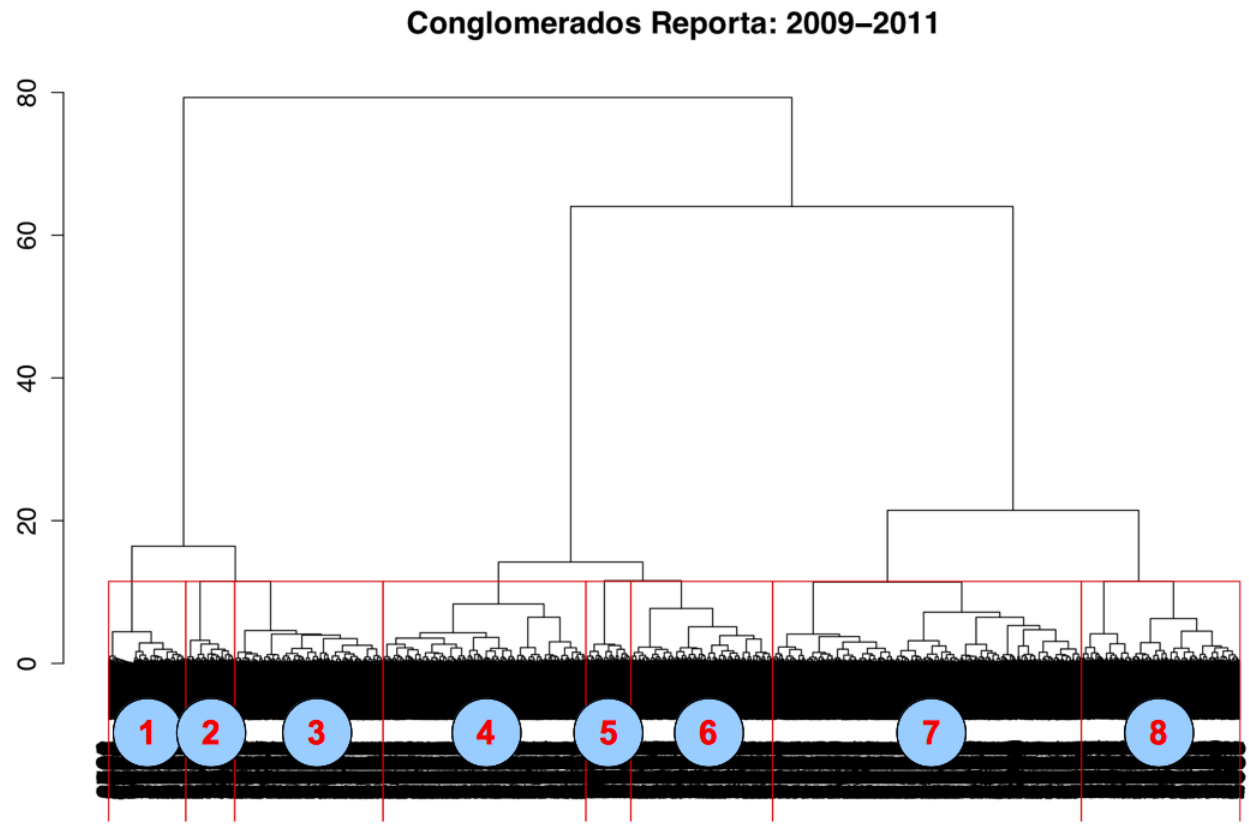

Figura 3.1: Dendograma que sugiere el número de conglomerados para la matriz de datos Reporta.

sin diabetes. Estos usuarios prefieren acudir al médico en vez de usar cualquier tipo de remedio para curarse cuando enferman. Pertenecen a un grupo de riesgo y presentan menos de 2 resfriados al año. (según la apreciación de ellos) no enferman de gripe ni fueron sospechos de influenza.

- **Conglomerado 3** agrupa a 639 usuarios con universidad como máximo nivel de estudios, que se transportan en auto o taxi, no tienen mascotas. No padecen ninguna enfermedad crónica y no se aplicaron la vacuna contra influenza en el 2008. Estos usuarios prefieren acudir al médico en vez de usar cualquier tipo de remedio para curarse cuando enferman. No pertenecen a un grupo de riesgo.

- **Conglomerado 4** agrupa a 874 usuarios estudiantes, que se transportan tanto a pie como en transporte colectivo y tienen mascotas. No padecen alguna enfermedad crónica

y acuden al médico cuando enferman. No pertenecen a un grupo de riesgo y no enferman de gripe (según la apreciación de ellos).

- **Conglomerado 5** agrupa a 194 usuarios que usan transporte colectivo tienen mascotas, no son fumadores, pertenecen a un grupo de riesgo, enferman de gripe y alguna vez han sido catalogados como sospechosos de influenza.

- **Conglomerado 6** agrupa a 610 usuarios universitarios que se transportan en colectivo, tienen mascotas no padecen de alguna enfermedad crónica ni pertenezen a un grupo de riesgo.

- **Conglomerado 7** agrupa a 1330 usuarios universitarios, que se transportan en auto o taxi, tienen mascotas. No padecen enfermedades crónicas ni pertenecen a un grupo de riego, son los que no se enferman de gripe.

- **Conglomerado 8** agrupa a 683 usuarios que se encuentran casados o viven en unión libre, que usan el transporte colectivo, auto o taxi para transportarse, tienen a un perro como mascota, usan servicio médico privado, pertenecen a un grupo de riesgo, y fueron catalogados en alguna ocasión como sospechos de influeza.

Al observar las relaciones existentes entre las variables y las clases de estudio en cada uno de los conglomerados propuestos, nos dimos cuenta de que los factores: usar el transporte público o transladarse a pie, tener mascotas y pertenecer a un grupo de riesgo; parecen estar muy relacionadas con la presencia de afecciones respiratorias (gripe o ser sospechosos de influenza), en contraparte los factores que parecieran ser protectores de estas afecciones son: no automediarse, acudir al médico al enfermar en vez de usar cualquier remedio casero, reportar haber tenido menos de 2 resfriados al año, no tener mascota, transportarse en auto o taxi.

Con el algoritmo de epsilon cuantificaremos la probabilidad de pertenencia a las distintas clases dado cada una de las variables de estudio con el fin de conocer las variables que determinan la participación de los usuarios en el proyecto Reporta y los factores de riesgo para las clases *sospechos*, *gripe* e *influenza*; de esta manera comprobaremos la veracidad de las hipótesis formuladas con base en el análisis de conglomerados.

3.3. Identificación de los factores de riesgo

Para la clase *participa*, la ponderación de variables sociodemográficas mediante la función *Épsilon* se muestra en la tabla 3.2.

Se recuerda que si la variable tiene un valor de $\varepsilon > 2$ indica que ésta es representativa de la clase en cuestión.

Los resultados de la tabla 3.2 revelan las caracterísiticas de los usuarios de Reporta que con mayor certeza estadística tienden a participar activamente en el sistema (60 por ciento o más de las veces a lo largo del proyecto). Las variables más importantes para determinar su participación activa en Reporta son: ocupación docente, institución UNAM, servicio médico privado, estado civil divorciado, cohabitantes 1, mascota al menos gatos y el rango de edad (55,65].

Por el contrario, las variables más importantes que determinan la poca participación por parte de los usuario en Reporta son : Institución UACM, no uso del auto o taxi para transportarse, cohabitantes 4 y más de 5, ocupación industrial, ejercicio menor de 1 hora a la semana, secundaria como escolaridad máxima, fumador, servicio médico público: IMSS y rango de edad (15,25].

También se analizó a las variables sociodemográficas de par en par, para conocer el conjunto de variables que describen a la clase de más participativos. Este análsis bivariante está complementado con: $\varepsilon(C|X_iX_j; X_i)$ y $\varepsilon(C|X_iX_j; X_j)$ que mide el efecto de una variable en relación con la otra y $\varepsilon(X_iX_j|C; X_iX_j)$ para conocer la correlación que tienen ambas variables respecto a la clase.

Así, el conjunto de variables tomadas por pares más importantes que determinan una participación activa en Reporta se muestra en la tabla 3.3.

Factor	*Épsilon*	N_x	N_{Cx}	$P(C\|x)$
Ocupación = Docente	5.55	1355	188	0.1387
Institución = UNAM	4.01	1762	216	0.1226
Rango de edad = (55,60]	3.65	325	50	0.1538
Cohabitantes = 1	3.48	894	115	0.1286
Servicio médico = Privado	3.07	2699	302	0.1119
Gatos = Sí	3.00	1275	152	0.1192
Rango de edad = (65,70]	2.95	78	15	0.1923
Otra enf. crónica = Sí	2.63	389	52	0.1337
Rango de edad = (60,65]	2.50	165	25	0.1515
Estado civil = Divorciado	2.26	319	42	0.1317
Institución = UACM	-2.00	106	4	0.0377
Auto o taxi = No	-2.00	1578	126	0.0798
Rango de edad = (20,25]	-2.06	600	42	0.0700
Cohabietantes = Más de 5	-2.08	302	18	0.0596
Ocupación = Industrial	-2.17	173	8	0.0462
Farmacia = Sí	-2.20	310	18	0.0581
Ocupación = Otro	-2.26	728	51	0.0701
Ejercicio = Menos de 1h/sem	-2.34	2086	166	0.0796
Escolaridad = Secundaria	-2.35	107	3	0.0280
Fumador = Sí	-2.45	697	47	0.0674
Rango de edad = (15,20]	-2.49	146	5	0.0342
Cohabitantes = 4	-2.51	665	44	0.0662
Servicio médico = IMSS	-3.04	743	46	0.0619

Cuadro 3.2: Factores que describen la clase de los usuarios con mayor participación en el Proyecto Reporta (60 por ciento o más de las veces a lo largo del proyecto).El número de usuarios que pertenecen a esta clase es $N(C) = 461$ lo que representa casi un 10 por ciento del total de usuarios en Reporta ($P(C) = 0.0946$).

Variables: X_i **y** X_j	$\varepsilon(C\|X_i, X_j; C)$	$\varepsilon(C\|X_i X_j; X_i)$	$\varepsilon(C\|X_i X_j; X_j)$	$\varepsilon(X_i X_j\|C; X_i X_j)$	$\varepsilon(X_i X_j\|\neg C; X_i X_j)$	$P(C\|X_i X_j)$	$P(X_i X_j\|C)$	$N_{X_i X_j}$	$N_{C\|X_i X_j}$
Ocupación=Docente y Servicio médico=Privado	7.07	2.19	4.93	2.3	4.46	0.1642	0.31	883	145
Ocupación=Docente y Institución=UNAM	6.2	2.14	3.46	1.41	4.26	0.1692	0.22	591	100
Rango de edad=(55,60] y Institución=UNAM	5.9	3.18	4.43	0.55	-2.72	0.2708	0.06	96	26
Ocupación=Docente y Farmacia=No	5.73	0.26	5.36	0.22	0.81	0.1412	0.4	1296	183
Ocupación=Docente y Aves=No	5.71	0.23	5.61	0.07	-0.27	0.1410	0.4	1298	183
Ocupación=Docente y Respiratorias=No	5.68	0.28	5.54	0.37	0.38	0.1415	0.39	1258	178
Ocupación=Docente y Fumador=No	5.65	0.73	4.48	1.29	4.07	0.1467	0.32	1009	148
Ocupación=Docente y Auto o taxi=Sí	5.6	0.49	4.64	2.28	6.43	0.1439	0.34	1105	159
Ocupación=Docente y Diabetes=No	5.55	0.06	5.58	0.05	-0.16	0.1393	0.4	1321	184
Ocupación=Docente y Sexo=Femenino	4.9	0.58	4.44	0.38	0.49	0.1460	0.25	781	114
Rango de edad=(55,60] y Otra enf. crónica=Sí	5.06	2.95	3.55	3.97	2.84	0.3061	0.0325	49	15
Rango de edad=(55,60] y Institución=UNAM	5.9	3.18	4.43	0.55	-2.72	0.2708	0.0564	96	26
Rango de edad=(70,75] y Gatos=Sí	4.48	4.33	3.89	2.02	-2.24	0.7500	0.0065	4	3
Rango de edad=(50,55] y Cohabitantes=Ninguna	4.38	3.68	3.71	2.31	-1.81	0.3462	0.0195	26	9
Cohabitantes=1 y Vacuna 2008=Sí, por alguna otra razón	4.19	2.88	3.12	2.14	-0.1	0.2542	0.0325	59	15
Rango de edad=(65,70] y Cohabitantes=1	4.1	1.73	3.05	2.73	2.36	0.3214	0.0195	28	9
Estado civil=Divorciado y Otra enf. crónica=Sí	3.65	2.49	2.44	2.43	1.27	0.2703	0.0217	37	10
Rango de edad=(55,60] y Ocupación=Servicio	3.39	2.35	3.24	2.29	-0.68	0.5000	0.0065	6	3
Otros animales=Sí y Rango de edad=(70,75]	3.39	2.99	3.27	2.61	-1.06	0.6667	0.0043	3	2
Resfriados por año=Más de 5 y Rango de edad=(50,55]	3.16	3.81	2.64	2.24	-2.06	0.3333	0.0108	15	5
A pie=Sí y Cohabitantes=2	2.71	2.17	2.07	2.08	0.86	0.1508	0.0651	199	30
Rango de edad=(45,50] y Otros animales=Sí	2.41	2.44	1.71	2.36	0.83	0.1864	0.0239	59	11
Cohabitantes=Más de 5 y Transporte colectivo=Sí	-2.08	-0.66	-2.15	0.08	3.73	0.0476	0.0174	168	8
Ocupación=Industrial y Rango de edad=(35,40]	-2.12	-1.44	-2.18	-0.99	5.55	0.0000	0.0000	43	0
Servicio médico=IMSS y Ocupación=Administrativo	-2.13	-1.15	-1.55	-0.09	2.78	0.0357	0.0087	112	4
Ejercicio=Menos de 1h/sem y Rango de edad=(30,35]	-2.15	-1.32	-0.95	0.39	2.11	0.0601	0.0434	333	20
Escolaridad=Preparatoria y Rango de edad=(15,20]	-2.19	-1.55	-0.22	3.96	19.59	0.0303	0.0065	99	3
Servicio médico=IMSS y Estado civil=Soltero	-2.23	-0.03	-1.56	1.26	2.39	0.0615	0.0521	390	24
Automedico=Sí y Rango de edad=(30,35]	-2.24	-1.99	-0.96	-0.05	2.78	0.0608	0.0499	378	23
Servicio médico=IMSS y Cohabitantes=5	-2.25	-1.62	-1.94	-0.99	2.64	0.0147	0.0022	68	1
Ocupación=Industrial y Sexo=Masculino	-2.29	-0.51	-2.19	0.95	7.18	0.0370	0.0108	135	5
Ocupación=Estudiante y Rango de edad=(15,20]	-2.34	-2.16	0.13	3.41	15.74	0.0362	0.0108	138	5
Servicio médico=IMSS y Perros=Sí	-2.38	0.11	-2.48	0.64	2.42	0.0631	0.0672	491	31
Cohabitantes=Más de 5 y Perros=Sí	-2.44	-0.84	-2.5	-0.26	3.04	0.0461	0.0217	217	10
Automedico=Sí y Rango de edad=(20,25]	-2.49	-2.27	-1.14	-0.5	2.01	0.0536	0.0369	317	17
Perros=Sí y Rango de edad=(20,25]	-2.61	-2.7	-1.02	-0.27	3.67	0.0573	0.0521	419	24
Perros=Sí y Cohabitantes=4	-2.75	-2.83	-0.82	-0.31	2.72	0.0564	0.0542	443	25
Cohabitantes=Más de 5 y Auto o taxi=No	-3.06	-1.99	-2.64	-0.87	5.18	0.0205	0.0065	146	3
Servicio médico=IMSS y Ejercicio=Menos de 1h/sem	-3.4	-1.51	-2.61	-0.14	3.13	0.0430	0.0347	372	16

Cuadro 3.3: Principales factores de par en par que describen a la clase de los usuarios con mayor participación en el proyecto Reporta.

Los resultados mostrados en la tabla 3.3, revelan que los usuarios docentes cuidadosos de su salud, sin diabetes ni enfermedades respiratorias crónicas, sin mascotas y que se transportan en auto o taxi y que cuentan con servicio médico privado y que son mujeres son fieles seguidores del proyecto, lo mismo se puede decir de quienes tienen una edad entre los 45 y 75 años de edad y comparten casa con ninguna, 1 o 2 personas. En contraparte, los usuarios cuyo servicio médico es público IMSS y: realizan poco ejercicio a la semana, tienen perros o cohabitan con más de 5 personas en casa o son solteros o su ocupación es administrativo o industrial son los que menos participan en el proyecto, lo mismo pasa con aquellos usuarios que cohabitan con más de 4 personas en su casa y no usan coche ni taxi pero si usan el transporte colectivo o tienen como mascota un perro, o con quienes tienen una edad entre 15 y 25 años tienen de mascota al menos un perro, se automedican y son estudiantes.

Para la clase *sospechosos* los factores de riesgo se muestran en la tabla 3.4.

Estos valores de *Épsilon* indican las características de los usuarios de Reporta que con mayor certeza estadística son catalogados como factores de riesgo de ser sospechosos de influenza según la definición de la SSA (fiebre y dolor de garganta o fiebre y tos). Dichos factores son: tener más de 5 resfriados por año, pertenecer al grupo de riesgo (tener menos de 2 años o más de 65 años o padecer alguna enfermedad crónica), tener alguna enfermedad crónica respiratoria, transportarse en colectivo o a pie, usar herbolarios caseros, acupuntura y automedicarse cuando enferman, convivir con pájaros y gatos, no tener servicio médico y ser mujer.

Las variables más importantes que determinan la ausencia de este diagnóstico clínico son: tener menos de 2 resfriados al año, no convivir con animales, no pertenecer al grupo de riesgo (tener menos de 2 años o más de 65 años o padecer alguna enfermedad crónica), tener una ocupación industrial, ser hombre y estar en el rango de edad (60,65].

El análisis de Épsilon para variables por pares para la clase *sospechosos* se muestra en la tabla 3.5.

Factor	*Épsilon*	N_x	N_{Cx}	$P(C\|x)$
Resfriados por año=Más de 5	5.08	241	80	0.3320
Grupo de riesgo=Sí	4.85	1286	328	0.2551
Resfriados por año=Entre 2 y 5	4.66	1778	436	0.2452
Respiratorias=Sí	4.57	404	118	0.2921
Otra enf. crónica=Sí	4.41	389	113	0.2905
Herbolaria o remedios caseros=Sí	3.69	1158	283	0.2444
Transporte colectivo=Sí	3.43	2059	476	0.2312
A pie=Sí	3.35	877	216	0.2463
Otros animales=Sí	3.21	512	132	0.2578
Institución=UNAM	3.21	1762	408	0.2316
Rango de edad=(25,30]	3.17	779	192	0.2465
Pájaros=Sí	3.12	728	180	0.2473
Homeopatía=Sí	3.04	572	144	0.2517
Servicio médico=Ninguno	2.96	250	69	0.2760
Sexo=Femenino	2.8	2749	611	0.2223
Estado civil=Soltero	2.2	2241	492	0.2195
Gatos=Sí	2.09	1275	286	0.2243
Automedico=Sí	2.06	2330	508	0.2180
Médico=No	-2.05	1576	284	0.1802
Herbolaria o remedios caseros=No	-2.06	3715	696	0.1873
Perros=No	-2.23	1980	358	0.1808
Ocupación=Industrial	-2.42	173	22	0.1272
Rango de edad=(60,65]	-2.55	165	20	0.1212
Sexo=Másculino	-2.78	2069	365	0.1764
Grupo de riesgo=No	-2.9	3587	651	0.1815
Trasporte colectivo=No	-2.93	2814	503	0.1787
Mascotas=No	-3.09	1215	201	0.1654
Resfriados por año=Menos de 2	-4.71	2725	449	0.1648

Cuadro 3.4: Factores de riesgo de ser sospechoso de influenza. El número de usuarios que pertenecen a esta clase es $N(C) = 979$, lo que representa un 20.09 por ciento del total de usuarios en Reporta ($P(C) = 0.2009$).

Variables: X_i **y** X_j	$\varepsilon(C\|X_i, X_j; C)$	$\varepsilon(C\|X_iX_j; X_i)$	$\varepsilon(C\|X_iX_j; X_j)$	$\varepsilon(X_iX_j\|C; X_iX_j)$	$\varepsilon(X_iX_j\|\neg C; X_iX_j)$	$P(C\|X_iX_j)$	$P(X_iX_j\|C)$	$N_{X_iX_j}$	$N_{C\|X_iX_j}$
Sexo=Femenino y Grupo de riesgo=Sí	6.31	4.69	2.44	0.89	-0.37	0.29	0.22	734	216
Sexo=Femenino y Respiratorias=Sí	5.53	4.52	1.71	1.38	0.57	0.34	0.09	249	85
Aves=Sí y Respiratorias=Sí	6.11	5.29	4.53	3.79	-1.91	0.78	0.01	18	14
Resfriados por año=Entre 2 y 5 y Sexo=Femenino	5.9	2.15	4.02	1.2	1.23	0.27	0.3	1056	289
Pájaros=Sí y Grupo de riesgo=Sí	5.5	3.5	3.2	2.35	0.98	0.35	0.08	224	78
Otros animales=Sí y Otra enf. crónica=Sí	4.69	3.36	2.72	2.26	0.21	0.46	0.02	52	24
Respiratorias=Sí y Pájaros=Sí	4.66	2.27	3.35	2.67	1.55	0.4	0.03	84	34
Ejercicio=Menos de 1h/sem y Respiratorias=Sí	4.47	4.37	1.23	1.42	-0.01	0.33	0.06	183	61
Gatos=Sí y Resfriados por año=Más de 5	4.39	3.73	1.34	1.39	0.49	0.41	0.03	74	30
Rango de edad=(25,30] y Pájaros=Sí	4.25	2.76	2.74	1.66	0.04	0.35	0.05	128	45
Pájaros=Sí y Otra enf. crónica=Sí	4.22	2.98	2	2.05	1.15	0.39	0.03	76	30
Rango de edad=(25,30] y Respiratorias=Sí	4.07	2.89	1.9	1.02	-0.02	0.39	0.03	71	28
Rango de edad=(20,25] y Resfriados por año=Más de 5	3.45	3.43	1.09	2.59	1.4	0.41	0.02	44	18
Escolaridad=Preparatoria y Resfriados por año=Más de 5	3.1	3.29	0.95	2.13	0.8	0.41	0.02	37	15
Otros animales=Sí y Cohabitantes=Más de 5	3.03	1.94	2.4	2.15	0.5	0.39	0.02	41	16
Sexo=Masculino y Gatos=No	-3.16	-0.75	-2.37	0.92	2.01	0.17	0.28	1601	271
Automedico=No y Mascotas=No	-3.41	-2.45	-1.15	0.78	2.48	0.15	0.11	697	104
Herbolaria o remedios caseros=No y Sexo=Masculino	-3.52	-2.18	-1.05	1.7	3.18	0.17	0.29	1698	283
Mascotas=No y Transporte colectivo=No	-3.53	-1.15	-2.08	1.32	2.45	0.15	0.12	773	116
Gatos=No y Transporte colectivo=No	-3.73	-2.82	-1.22	0.33	1.81	0.17	0.37	2141	361
Resfriados por año=Menos de 2 y Auto o taxi=Sí	-4.45	-0.51	-3.86	1.26	2.85	0.16	0.32	1950	313
Herbolaria o remedios caseros=No y Transporte colectivo=No	-4.6	-3.08	-2.07	0.42	2.94	0.16	0.37	2247	364
Herbolaria o remedios caseros=No y Trasporte=Auto o taxi	-4.64	-3.26	-1.79	0.6	2.84	0.16	0.3	1868	295
Resfriados por año=Menos de 2 y Estado civil=Casado/Unión Libre	-4.81	-1.76	-3.66	0.22	2.45	0.15	0.19	1244	182
Automedico=No y Resfriados por año=Menos de 2	-5.08	-3.66	-1.67	1	3.36	0.15	0.23	1537	229
Respiratorias=No y Resfriados por año=Menos de 2	-5.16	-4.18	-0.62	1.25	1.97	0.16	0.42	2584	414
Grupo de riesgo=No y Resfriados por año=Menos de 2	-5.97	-3.92	-2.02	0.65	1.68	0.15	0.31	2076	308
Resfriados por año=Menos de 2 y Transporte colectivo=No	-5.98	-2.47	-3.89	0.55	2.8	0.14	0.24	1672	238
Resfriados por año=Menos de 2 y Trasporte=Auto o taxi	-6.34	-3.19	-4	0.34	3.79	0.13	0.19	1417	189

Cuadro 3.5: Principales factores de riesgo de ser sospechoso de influenza.

Al observar los valores de la tabla 3.5 los usuarios con mayor certeza estadística de pertener a esta clase son usuarios de sexo femenino: pertenecientes al grupo de riesgo (tener menos de 2 años o más de 65 años o padecer alguna enfermedad crónica), que estiman tener más de 2 resfriados al año y que padecen enfermedades respiratorias crónicas. Del mismo modo ocurre para los usuarios que tienen una edad entre 20 y 25 años y: hayan estimado tener más de 5 resfriados al año, padezcan de enfermedades respiratorias crónicas y tengan contacto con pájaros, así como aquellos usuarios que cohabitan en casa con más de 5 personas y tienen contacto con algún animalito. Aquellos que padecen enfermedades respiratorias crónicas y: que sean mujeres o tengan contacto con pájaros o aves de corral o realicen ejercicio menos de 1 hora a la semana o tengan entre 20 y 25 años serán más propensos de ser sospechosos de influenza.

De acuerdo con los resultados de este análisis no son sopechososos de influenza aquellos usuarios que estimaron tener menos de 2 resfriados al año y: utilizan como principal medio de transporte el auto o taxi o no usan el transporten colectivo o no pertenezcan a un grupo de riesgo, no se automedican ni usan remedios caseros, acupuntura u homeopatía cuando enferman o que son casados o viven en unión libre. También no forman parte de esta clase quienes son hombres y: no tienen contacto con gatos ni usan herbolaria o remedios caseros cuando enferman.

Para la clase *gripe* la ponderación de variables sociodemográficas mediante la función *Épsilon* se muestra en la tabla 3.6.

Estos resultados revelan las características de los usuarios de Reporta que con mayor certeza estadísitica padecen gripe. Los rasgos asociados con la presentación de gripe son: haber estimado más de 5 resfriados al año, ser mujer, pertenecer al grupo de riesgo, tener gatos y pájaros como mascota, usar autobuses colectivos como medio de transporte, usar remedios caseros y tomar lo que le aconsejen en la farmacia cuando enferman y tener una edad entre 25 y 30 años.

En contraparte, aquellos usuarios de sexo masculino que estimaron tener menos de 2 resfriados al año, sin mascotas, que no pertenecen al grupo de riesgo (tener menos de 2 años o más de 65 años o padecer alguna enfermedad crónica), que no se transportan en autobuses colectivos, que son casados o viven unión libre y que tienen entre 65 y 70 años, son quienes casi no padecen gripe.

Factor	*Épsilon*	N_x	N_{Cx}	$P(C\|x)$
Resfriados por año=Más de 5	6.33	241	44	0.1826
Sexo=Femenino	3.73	2749	258	0.0939
Grupo de riesgo=Sí	3.64	1286	131	0.1019
Respiratorias=Sí	3.33	404	48	0.1188
Herbolaria o remedios caseros=Sí	3.24	1158	116	0.1002
Rango de edad=(25,30]	3.06	779	81	0.1040
Otra enf. crónica=Sí	3.04	389	45	0.1157
Otros animales=Sí	2.77	512	55	0.1074
Vacuna 2008=Sí, pertenezco a un grupo de riesgo	2.64	114	16	0.1404
Gatos=Sí	2.58	1275	120	0.0941
Resfriados por año= Entre 2 y 5	2.56	1778	162	0.0911
Transporte colectivo=Sí	2.54	2059	185	0.0898
Farmacia=Sí	2.31	310	34	0.1097
Institución=UNAM	2.23	1762	157	0.0891
Pájaros=Sí	2.15	728	70	0.0962
Rango de edad=(65,70]	-2.09	78	1	0.0128
Estado civil=Casado/Union Libre	-2.1	2094	132	0.0630
Transporte colectivo=No	-2.17	2814	181	0.0643
Grupo de riesgo=No	-2.18	3587	235	0.0655
Mascotas=No	-2.2	1215	71	0.0584
Resfriados por año=Menos de 2	-3.76	2725	153	0.0561
Sexo=Másculino	-4.12	2069	106	0.0512

Cuadro 3.6: Factores de riesgo de tener gripe definida por la OMS. El número de usuarios que pertenecen a esta clase es $N(C) = 366$, lo que representa un 7.51 por ciento del total de usuarios en Reporta ($P(C) = 0.0751$).

Al evaluar las variables sociodemográficas por pares con la función *Épsilon* se obtuvieron los resultados mostrados en la tabla 3.7 para la clase *gripe*.

Variables: X_i **y** X_j	$\varepsilon(C\|X_i, X_j; C)$	$\varepsilon(C\|X_iX_j; X_i)$	$\varepsilon(C\|X_iX_j; X_j)$	$\varepsilon(X_iX_j\|C; X_iX_j)$	$\varepsilon(X_iX_j\|\neg C; X_iX_j)$	$P(C\|X_iX_j)$	$P(X_iX_j\|C)$	$N_{X_iX_j}$	$N_{C\|X_iX_j}$
Sexo=Femenino y Resfriados por año=Más de 5	7.26	5.80	1.67	0.37	-0.28	0.2374	0.0902	139	33
Resfriados por año=Más de 5 y Mascotas=Sí	6.63	0.57	6.13	0.80	1.26	0.1980	0.1093	202	40
Aves=Sí y Grupo de riesgo=Sí	5.55	4.34	4.25	2.61	-1.67	0.2955	0.0355	44	13
Sexo=Femenino y Grupo de riesgo=Sí	5.44	3.18	2.35	0.2	0.07	0.1281	0.2568	734	94
Respiratorias=Sí y Herbolaria o remedios caseros=Sí	5.42	3.14	3.97	1.25	-1.59	0.2273	0.0546	88	20
Resfriados por año=Más de 5 y Grupo de riesgo=Sí	5.34	0.92	3.79	1.35	3.47	0.2188	0.0574	96	21
Resfriados por año=Más de 5 y Ejercicio=Menos de 1h/sem	5.33	0.81	5.25	0.68	-0.36	0.2136	0.0601	103	22
Respiratorias=Sí y Aves=Sí	5.05	3.54	4.1	3.03	-0.52	0.3889	0.0191	18	7
Resfriados por año=Más de 5 y Rango de edad=(20,25]	4.97	1.55	4.47	2.39	1.64	0.2727	0.0328	44	12
Ocupación=Docente y Aves=Sí	3.38	3.4	2.36	2.73	0.1	0.1930	0.0301	57	11
Cohabitantes=Ninguna y Perros=No	-2.76	-2.03	-2.36	-0.78	5.67	0.0244	0.0137	205	5
AutoMedico=No y Mascotas=No	-2.78	-2.03	-1.25	0.04	2.69	0.0473	0.0902	697	33
Resfriados por año=Menos de 2 y Auto o taxi=Sí	-3.22	-0.05	-2.97	0.92	3.02	0.0559	0.2978	1950	109
Sexo=Másculino y Estado civil=Casado/Unión Libre	-3.84	-1.08	-2.56	1.33	5.64	0.0439	0.1257	1049	46
Resfriados por año=Menos de 2 y Respiratorias=No	-3.89	-0.26	-3.2	0.99	2.18	0.0550	0.3880	2584	142
AutoMedico=No y Resfriados por año=Menos de 2	-4.11	-2.98	-1.47	0.31	3.55	0.0475	0.1995	1537	73
Estado civil=Casado/Unión Libre y Resfriados por año=Menos de 2	-4.13	-2.73	-1.83	-0.03	2.39	0.0442	0.1503	1244	55
Herbolaria o remedios caseros=No y Sexo=Masculino	-4.28	-3.22	-0.66	1.13	3.4	0.0477	0.2213	1698	81
Sexo=Masculino y Gatos=No	-4.67	-1.25	-3.81	-0.03	2.23	0.0443	0.1940	1601	71
Transporte colectivo=No y Resfriados por año=Menos de 2	-4.69	-3.24	-2.01	-0.09	2.98	0.0449	0.2049	1672	75
Sexo=Masculino y Resfriados por año=Menos de 2	-4.82	-1.98	-2.64	0.43	1.65	0.0387	0.1284	1215	47
Sexo=Masculino y Transporte colectivo=No	-4.82	-1.96	-3.63	-0.66	1.31	0.0389	0.1311	1233	48

Cuadro 3.7: Principales factores de riesgo de par en par de tener gripe definida por la OMS.

De esta lista de resultados (cuadro 3.7) se concluye que los usuarios que enferman de gripe con mayor frecuencia son aquellos que estimaron tener más de 5 resfriados al año y: son de sexo femenino o conviven con algún animal en casa o pertenecen al grupo de riesgo (tener menos de 2 años o más de 65 años o padecer alguna enfermedad crónica) o se ejercitan menos de una hora a la semana; lo mismo se concluye de aquellos que conviven con aves de corral y pertenecen al grupo de riesgo, usan herbolaria o remedios caseros cuando enferman y padecen enfermedades crónicas respiratorias.

Por otro lado, los usuarios de sexo masculino y: que estiman tener menos de 2 resfriados al año, que no usan transporte colectivo pero sí utilizan auto o taxi, que no pertenecen al grupo de riesgo, que cuentan con servicio médico privado, tienen menor probabilidad de enfermar de gripe. Del mismo modo ocurre para los usuarios que estiman tener menos de 2 resfriado al año y: de sexo masculino, que no usan transporte colectivo, que no pertenecen al grupo de riesgo y que están casado o viven en unión libre.

En la tabla 3.8 se muestra la ponderación de variables sociodemográficas mediante la función *Épsilon* para la clase *influenza*.

Factor	*Épsilon*	N_x	N_{Cx}	$P(C\|x)$
Cerdos=Sí	4.84	62	4	0.0645
Rango de edad=(80,85]	3.89	7	1	0.1429
Servicio médico=Marina	3.89	7	1	0.1429
Ocupación=Conductor	3.61	8	1	0.1250
Otra enf. crónica=Sí	3.18	389	9	0.0231
Acupuntura=Sí	3.03	168	5	0.0298
Diabetes=Sí	2.22	107	3	0.0280
Grupo de riesgo=Sí	2.19	1286	18	0.0140
Vacuna 2008=Sí, pertenezco a un grupo de riesgo	2.09	114	3	0.0263
Fumador=Ocasional	2.03	802	12	0.0150
Aves=Sí	2.02	181	4	0.0221

Cuadro 3.8: Factores de riesgo de padecer influenza determinada por estudios de laboratorio. El número de usuarios que pertenecen a esta clase es $N(C) = 41$, lo que representa un 0.84 % del total de usuarios en Reporta ($P(C) = 0.0084$).

Cabe señalar que entre quienes reportaron síntomas sospechosos de influenza y se hicieron análisis de laboratorio, sólo 41 personas tuvieron confirmación clínica de que la enfermedad que tuvieron estuvo causada por el virus de influenza y que los factores de riesgo que más llaman la atención en este grupo reducido de 41 usuarios fue que 18 de ellos pertenecían a un grupo de riesgo (tenían menos de 2 años de edad o más de 65, y padecían alguna enfermedad crónica) y 12 fumaban ocasionalmente.

En conclusión: la clase *Participa* está descrita por aquellas personas de clase media para arriba, tienen una edad edad entre 50 y 75 años y gozan de buena salud al no padecer enfermedades crónicas, no automedicarse ni usar remedios caseros o de medicina alternativa, se ejercitan más de 4 horas a la semana. En general estas características describen a personas bien informadas en temas de salud y conscientes de la importancia de mantener una vida sana. Para la clase de *sospechosos* y *gripe*, muchos de los factores de riesgo que en ambas clases coincidían están relacionados con la susceptibilidad de las personas de contraer fácilmente una enfermedad respiratoria, tal es el caso del número de resfriado padecidos al año (datos estimado por el usuario), la pertenencia a un grupo de riesgo o el padecer alguna enfermedad respiratoria crónica. Otros factores en común fueron los relacionados con el hecho de ser mujer y con el medio de transporte, enferman más quienes se transportan en colectivo, que los que lo hacen en auto o taxi y por último, una cosa curiosa que notamos es que el contacto con animales constituye un factor muy importante para la presencia de estas enfermedades respiratorias.

Por lo tanto con la prueba binomial Épsilon no solamente comprobamos las hipótesis formuladas con base en el análisis de conglomerados sino que también conocimos el amplio conjunto de factores de riesgo para las distintas clases de afecciones respiratorias, sus probabilidades condicionales asociadas y grado de riesgo.

3.4. Modelo de riesgo

Para conocer el perfil de riesgo de un futuro participante, se comenzó por calcular la función $S(X) = \log\left\{\frac{P(X|C)}{P(X|\neg C)}\right\}$ en los datos de entrenamiento. Así las puntuaciones de las distintas variables para la clase *sospechosos* son las siguientes:

Cuadro 3.9: Valores de $S(X)$ de las distintas variables para la clase *sospechosos*

Variable	$S(x_i)$	$P(C)$	N_C	N_{x_i}	$N_{C\cap x_i}$	$N_{\neg C\cap x_i}$
Rango de edad=(0,5]	1.3742	0.2019	689	4	2	2
Resfriados por año=Más de 5	0.6160	0.2019	689	163	52	111
Escolaridad=Ninguno	0.5270	0.2019	689	10	3	7
Respiratorias=Sí	0.5189	0.2019	689	295	88	207
Institución=UAM	0.4067	0.2019	689	69	19	50
Rango de edad=(5,10]	0.3934	0.2019	689	11	3	8
Otra enf. crónica=Sí	0.3750	0.2019	689	275	74	201
Vacuna 2008=Sí, pertenezco a un grupo de riesgo	0.3269	0.2019	689	77	20	57
Pájaros=Sí	0.3122	0.2019	689	506	130	376
Ocupación=Servicio	0.2997	0.2019	689	55	14	41
Grupo de riesgo=Sí	0.2771	0.2019	689	911	228	683
Rango de edad=(75,80]	0.2756	0.2019	689	8	2	6
Ocupación=Seguridad	0.2756	0.2019	689	4	1	3
Homeopatía=Sí	0.2623	0.2019	689	400	99	301
Herbolaria o remedios caseros=Sí	0.2591	0.2019	689	810	200	610
Rango de edad=(25,30]	0.2582	0.2019	689	539	133	406
Resfriados por año=Entre 2 y 5	0.2470	0.2019	689	1267	310	957
Servicio médico=Ninguno	0.2428	0.2019	689	164	40	124
Diabetes=Sí	0.2372	0.2019	689	70	17	53
A pie=Sí	0.2343	0.2019	689	619	150	469
Cerdos=Sí	0.2042	0.2019	689	38	9	29
Otros animales=Sí	0.2021	0.2019	689	351	83	268
Transporte colectivo=Sí	0.1857	0.2019	689	1426	333	1093
Fumador=Ocasional	0.1732	0.2019	689	588	136	452
Moto o bici=Sí	0.1668	0.2019	689	126	29	97
Institución=UNAM	0.1524	0.2019	689	1239	282	957
Acupuntura=Sí	0.1421	0.2019	689	124	28	96
Cohabitantes=1	0.1275	0.2019	689	636	142	494
Enfer. cardiovasculares=Sí	0.1266	0.2019	689	251	56	195
Servicio médico=Secretaria de Salud	0.1215	0.2019	689	126	28	98
Sexo=Femenino	0.1178	0.2019	689	1936	429	1507
Aves=Sí	0.1162	0.2019	689	122	27	95
Rango de edad=(55,60]	0.1158	0.2019	689	226	50	176
Ocupación=Estudiante	0.1157	0.2019	689	886	196	690
Estado civil=Soltero	0.1034	0.2019	689	1570	344	1226
Automedico=Sí	0.1023	0.2019	689	1681	368	1313
Auto o taxi=No	0.0947	0.2019	689	1089	237	852
Farmacia=Sí	0.0909	0.2019	689	212	46	166
Gatos=Sí	0.0884	0.2019	689	882	191	691
Ocupación=Docente	0.0872	0.2019	689	966	209	757
Estado civil=Viudo	0.0750	0.2019	689	42	9	33
Perros=Sí	0.0693	0.2019	689	2011	429	1582
Mascotas=Sí	0.0678	0.2019	689	2553	544	2009
Médico=Sí	0.0628	0.2019	689	2304	489	1815
Rango de edad=(30,35]	0.0621	0.2019	689	495	105	390
Cohabitantes=3	0.0453	0.2019	689	836	175	661
Cohabitantes=2	0.0435	0.2019	689	708	148	560
Rango de edad=(40,45]	0.0424	0.2019	689	316	66	250
Ejercicio=Menos de 1h/sem	0.0368	0.2019	689	1462	304	1158
Ejercicio=1-4 hrs/sem	0.0366	0.2019	689	1342	279	1063

Continúa en la siguiente página

Continuación de la tabla

Variable	$S(x_i)$	$P(C)$	N_C	N_{x_i}	$N_{C \cap x_i}$	$N_{\neg C \cap x_i}$
Vacuna 2008=No	0.0363	0.2019	689	2651	551	2100
Escolaridad=Universidad	0.0342	0.2019	689	2824	586	2238
Servicio médico=Privado	0.0338	0.2019	689	1938	402	1536
Rango de edad=(35,40]	0.0289	0.2019	689	392	81	311
Institución=UACM	0.0219	0.2019	689	73	15	58
Ocupación=Otro	0.0148	0.2019	689	509	104	405
Servicio médico=IMSS	0.0103	0.2019	689	501	102	399
Cohabitantes=Ninguna	0.0078	0.2019	689	251	51	200
Cohabitantes=Más de 5	0.0069	0.2019	689	197	40	157
Cerdos=No	-0.0024	0.2019	689	3374	680	2694
Aves=No	-0.0045	0.2019	689	3290	662	2628
Diabetes=No	-0.0053	0.2019	689	3342	672	2670
Acupuntura=No	-0.0056	0.2019	689	3288	661	2627
Farmacia=No	-0.0062	0.2019	689	3200	643	2557
Moto o bici=No	-0.0067	0.2019	689	3286	660	2626
Enfermedades cardiovasculares=No	-0.0105	0.2019	689	3161	633	2528
Servicio médico=Sedena	-0.0120	0.2019	689	5	1	4
Fumador=No	-0.0143	0.2019	689	2224	444	1780
Otros animales=No	-0.0248	0.2019	689	3061	606	2455
Fumador=Sí	-0.0279	0.2019	689	476	94	382
Gatos=No	-0.0319	0.2019	689	2530	498	2032
Ocupación=Salud	-0.0345	0.2019	689	112	22	90
Otra enf. crónica=No	-0.0369	0.2019	689	3137	615	2522
Homeopatía=No	-0.0380	0.2019	689	3012	590	2422
Vacuna 2008=Sí, por alguna otra razón	-0.0436	0.2019	689	200	39	161
Auto o taxi=Sí	-0.0463	0.2019	689	2323	452	1871
Rango de edad=(20,25]	-0.0484	0.2019	689	417	81	336
A pie=No	-0.0565	0.2019	689	2793	539	2254
Respiratorias=No	-0.0576	0.2019	689	3117	601	2516
Pájaros=No	-0.0605	0.2019	689	2906	559	2347
Herbolaria o remedios caseros=No	-0.0893	0.2019	689	2602	489	2113
Institución=Otra	-0.0916	0.2019	689	1802	338	1464
Estado civil=Casado/Union Libre	-0.0932	0.2019	689	1468	275	1193
Perros=No	-0.1047	0.2019	689	1401	260	1141
Rango de edad=(50,55]	-0.1048	0.2019	689	291	54	237
Automedico=No	-0.1057	0.2019	689	1731	321	1410
Escolaridad=Preparatoria	-0.1106	0.2019	689	379	70	309
Cohabitantes=5	-0.1130	0.2019	689	217	40	177
Grupo de riesgo=No	-0.1131	0.2019	689	2501	461	2040
Servicio médico=ISSSTE	-0.1197	0.2019	689	360	66	294
Servicio médico=Otro	-0.1233	0.2019	689	186	34	152
Ejercicio=Más de 4 hrs/sem	-0.1276	0.2019	689	549	100	449
Médico=No	-0.1387	0.2019	689	1108	200	908
Cohabitantes=4	-0.1401	0.2019	689	477	86	391
Vacuna 2008=Sí,recomendacion medica	-0.1409	0.2019	689	111	20	91
Vacuna 2008=Sí, campaña de vacunación trabajo	-0.1468	0.2019	689	290	52	238
Transporte colectivo=No	-0.1472	0.2019	689	1986	356	1630
Sexo=Másculino	-0.1542	0.2019	689	1442	257	1185
Ocupación=Administrativo	-0.1549	0.2019	689	393	70	323
Rango de edad=(15,20]	-0.1620	0.2019	689	96	17	79
Rango de edad=(45,50]	-0.1662	0.2019	689	323	57	266

Continúa en la siguiente página

Continuación de la tabla

Variable	$S(x_i)$	$P(C)$	N_C	N_{x_i}	$N_{C\cap x_i}$	$N_{\neg C\cap x_i}$
Ocupación=En casa	-0.1711	0.2019	689	165	29	136
Ocupación=Jubilado	-0.2099	0.2019	689	47	8	39
Mascotas=No	-0.2199	0.2019	689	859	145	714
Estado civil=Divorciado	-0.2298	0.2019	689	221	37	184
Resfriados por año=Menos de 2	-0.2333	0.2019	689	1887	315	1572
Rango de edad=(10,15]	-0.4175	0.2019	689	21	3	18
Ocupación=Agropecuario	-0.4175	0.2019	689	7	1	6
Servicio médico=Marina	-0.4175	0.2019	689	7	1	6
Rango de edad=(60,65]	-0.4583	0.2019	689	116	16	100
Servicio médico=Pemex	-0.4716	0.2019	689	22	3	19
Ocupación=Industrial	-0.5166	0.2019	689	122	16	106
Rango de edad=(70,75]	-0.5717	0.2019	689	24	3	21
Escolaridad=Secundaria	-0.6770	0.2019	689	79	9	70
Escolaridad=Primaria	-1.1907	0.2019	689	14	1	13
Rango de edad=(65,70]	-1.8244	0.2019	689	51	2	49
Rango de edad=(80,85]	-1.8244	0.2019	689	4	0	4
Rango de edad=(85,90]	-1.8244	0.2019	689	2	0	2
Rango de edad=(90,95]	-1.8244	0.2019	689	1	0	1
Ocupación=Ambulante	-1.8244	0.2019	689	2	0	2
Ocupación=Conductor	-1.8244	0.2019	689	8	0	8
Ocupación=Discapacitado	-1.8244	0.2019	689	4	0	4

Con base en los valores de la tabla 3.9 se calcula el $Score = \sum_{i=1}^{N} S(x_i)$ para cada usuario en la base de entrenamiento, posteriormente se ordenan de manera ascendente y esta lista se parte en 10 subconjuntos equivalentes $A_1, ...A_{10}$ y sobre cada A_i se calcula el promedio de los scores, así como la probabilidad de pertenecia a la clase $P(C)$. Finalmente se grafican ambos valores (abscisas $P(C)$ y ordenas Score promedio) para dar origen al modelo de riesgo. Las fronteras de riesgo (alto, medio, bajo) se construyen a partir de lo observado en esta gráfica y en la distribución de las frecuencias del score (número de casos que presentaron dicho score).

Para la clase de *sospechosos* se obtuvo la gráfica mostrada en la figura 3.2, en donde la distribución de frecuencias del score se concentra en su gran mayoría alrededor de la media (-0.0264), en el extremo derecho se encuentran los casos cuyos scores son los más altos y que presentan un alto riesgo de ser sospechos de influenza; a la extrema izquierda los que son poco probables de pertenecer a dicha clase.

Partiendo de que la probabilidad de ser sospechoso de influenza en la base de entrenamiento fue de P(C)= 20.19 por ciento se calculó la estimación de las fronteras de riesgo; inicialmente se fija el punto .2019 en el eje de las probabilidades y se establece que la distancia a tomar es el doble de P(C) tanto arriba como abajo, quedando los intervalos de riesgos

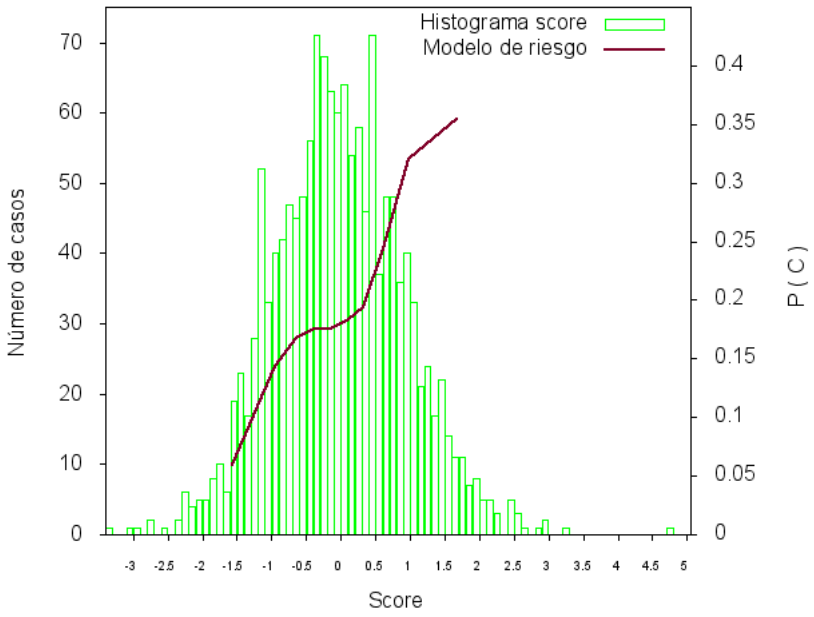

Figura 3.2: Modelo de riesgo y frecuencias score: Sospechosos de influenza

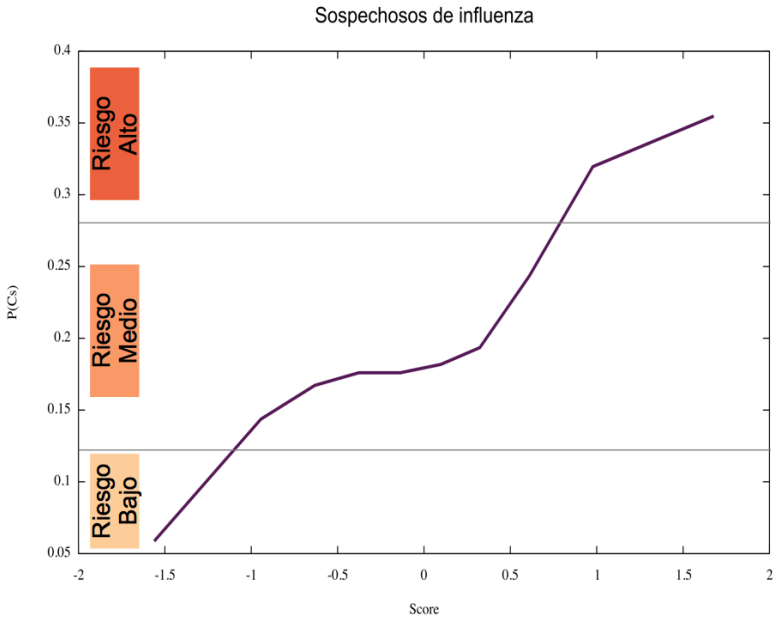

Figura 3.3: Niveles de riesgo: Sospechosos de influenza

delimitados de la siguiente manera:

- Riesgo alto: $P(C|X) > P(C) + 0.4P(C)$
- Riesgo medio: $0.6P(C) \leq P(C|X) \leq 1.4P(C)$
- Riesgo bajo: $P(C|X) < P(C) - 0.4P(C)$.

Es decir, aquellos usuarios cuyo score sea mayor que 1 tienen un riesgo alto de pertenecer a la clase *sospechosos*, el riesgo bajo se pronostica para puntuaciones menores de -1. El riesgo intermedio se sitúa entre estos dos valores. Al utilizar este modelo de riesgo ilustrado en la figura 3.5 para predecir el perfil de riesgo en la base de prueba, se tiene que los "verdaderos positivos"(aquellos casos pedichos como *sospechosos* y lo fueron) para riesgo alto fueron: 68 de 195 (34.87 por ciento) para riesgo medio: 199 de 1042 (19.10 por ciento) y para riesgo bajo: 23 de 224 (10.27 por ciento). Para tener una mayor precisión del modelo, se dicidió reajustar estas fronteras tomando en cuenta obtener una mayor proporción de verdaderos positivos en los perfiles de riesgo alto y medio; para el perfil bajo se consideró que el criterio de frontera inicialmente propuesto ($Score < 1$) es suficiente para pronosticar futuros casos.

Nivel de riesgo	**Rango**	**Verdaderos positivos**
Alto	$Score > 1$	34.87 %
	$Score > 1.5$	38.46 %
	$Score > 1.7$	43.40 %
Medio	$-1 \leq Score \leq 1$	19.10 %
	$-1 \leq Score \leq 1.5$	20.45 %
	$-1 \leq Score \leq 1.7$	20.61 %
Bajo	$Score < -1$	10.27 %

Cuadro 3.10: Proporciones de verdaderos positivos para cada criterio nivel de riesgo propuesto.

Analizando las proporciones de "verdaderos positivos"para los nuevos criterios de nivel de riesgo ilustrados en la tabla 3.10, se decidió que el criterio para pronosticar el perfil de riesgo de pertenecer a la clase *sospechoso* de un futuro participante es el siguiente:

- Riesgo alto: $Score > 1.7$
- Riesgo medio: $-1 \leq Score \leq 1.7$
- Riesgo bajo: $Score < -1$

Para la clase *gripe*, la tabla de resultados fue la siguiente:

Cuadro 3.11: Valores de $S(X)$ de las distintas variables para la clase *gripe*

Variable	$S(x_i)$	N_{x_i}	$N_{C \cap x_i}$	$N_{\neg C \cap x_i}$
Resfriados por año=Más de 5	0.91	163	28	135

Continúa en la siguiente página

Continuación de la tabla

Variable	$S(x_i)$	N_{x_i}	$N_{C \cap x_i}$	$N_{\neg C \cap x_i}$
Vacuna 2008=Sí, pertenezco a un grupo de riesgo	0.7	77	11	66
Servicio médico=Marina	0.7	7	1	6
Institución=UAM	0.59	69	9	60
Ocupación=Servicio	0.56	55	7	48
Rango de edad=(75,80]	0.54	8	1	7
Otra enf. crónica=Sí	0.53	275	34	241
Respiratorias=Sí	0.48	295	35	260
Farmacia=Sí	0.47	212	25	187
Rango de edad=(25,30]	0.43	539	61	478
Servicio médico=Ninguno	0.39	164	18	146
Grupo de riesgo=Sí	0.32	911	94	817
Otros animales=Sí	0.32	351	36	315
A pie=Sí	0.27	619	61	558
Acupuntura=Sí	0.25	124	12	112
Herbolaria o remedios caseros=Sí	0.25	810	78	732
Sexo=Femenino	0.24	1936	185	1751
Homeopatía=Sí	0.23	400	38	362
Ocupación=Estudiante	0.22	886	83	803
Gatos=Sí	0.21	882	82	800
Estado civil=Soltero	0.2	1570	145	1425
Transporte colectivo=Sí	0.2	1426	131	1295
Resfriados por año=Entre 2 y 5	0.19	1267	116	1151
Institución=UNAM	0.19	1239	113	1126
Servicio médico=Pemex	0.18	22	2	20
Aves=Sí	0.18	122	11	111
Automedico=Sí	0.16	1681	150	1531
Pájaros=Sí	0.16	506	45	461
Cohabitantes=1	0.13	636	55	581
Rango de edad=(20,25]	0.13	417	36	381
Servicio médico=Otro	0.12	186	16	170
Ocupación=En casa	0.11	165	14	151
Medico=Sí	0.11	2304	195	2109
Fumador=Sí	0.1	476	40	436
Cohabitantes=4	0.1	477	40	437
Auto o taxi=No	0.09	1089	91	998
Perros=Sí	0.09	2011	168	1843
Mascotas=Sí	0.09	2553	212	2341
Ejercicio=1-4 hrs/sem	0.08	1342	111	1231
Cohabitantes=Más de 5	0.06	197	16	181
Ejercicio=Menos de 1h/sem	0.05	1462	118	1344
Ocupación=Otro	0.05	509	41	468
Rango de edad=(55,60]	0.04	226	18	208
Rango de edad=(40,45]	0.03	316	25	291
Cohabitantes=2	0.03	708	56	652
Escolaridad=Universidad	0.03	2824	223	2601
Cerdos=Sí	0.03	38	3	35
Vacuna2008=No	0.02	2651	207	2444
Servicio médico=ISSSTE	0.01	360	28	332
Moto o bici=No	0.01	3286	254	3032
Cardio=No	0.01	3161	244	2917
Diabetes=No	0	3342	257	3085

Continúa en la siguiente página

Continuación de la tabla

Variable	$S(x_i)$	N_{x_i}	$N_{C \cap x_i}$	$N_{\neg C \cap x_i}$
Cerdos=No	0	3374	259	3115
Fumador=Ocasional	0	588	45	543
Rango de edad=(35,40]	0	392	30	362
Ocupación=Administrativo	-0.01	393	30	363
Aves=No	-0.01	3290	251	3039
Acupuntura=No	-0.01	3288	250	3038
Fumador=No	-0.02	2224	168	2056
Cohabitantes=3	-0.02	836	63	773
Vacuna2008=Sí, por alguna otra razón	-0.03	200	15	185
Servicio médico=Privado	-0.03	1938	145	1793
Rango de edad=(30,35]	-0.03	495	37	458
Pájaros=No	-0.03	2906	217	2689
HomeopatíaNo	-0.03	3012	224	2788
Farmacia=No	-0.04	3200	237	2963
Servicio médico=IMSS	-0.04	501	37	464
Otros animales=No	-0.04	3061	226	2835
Cohabitantes=5	-0.04	217	16	201
Auto o taxi=Sí	-0.05	2323	171	2152
Respiratorias=No	-0.06	3117	227	2890
Otra enf. crónica=No	-0.06	3137	228	2909
A pie=No	-0.07	2793	201	2592
Cardio=Sí	-0.07	251	18	233
Diabetes=Sí	-0.08	70	5	65
Estado civil=Viudo	-0.08	42	3	39
Servicio médico=Secretaría de Salud	-0.08	126	9	117
Escolaridad=Primaria	-0.08	14	1	13
Gatos=No	-0.08	2530	180	2350
Herbolaria o remedios caseros=No	-0.09	2602	184	2418
Ocupación=Docente	-0.09	966	68	898
Escolaridad=Preparatoria	-0.12	379	26	353
Grupo de riesgo=No	-0.14	2501	168	2333
Perros=No	-0.15	1401	94	1307
Institución=Otra	-0.15	1802	120	1682
Transporte colectivo=No	-0.16	1986	131	1855
Vacuna2008=Sí, campaña de vacunación trabajo	-0.17	290	19	271
Rango de edad=(45,50]	-0.18	323	21	302
Automedico=No	-0.18	1731	112	1619
Moto o bici=Sí	-0.2	126	8	118
Vacuna2008=Sí,recomendacion médica	-0.21	111	7	104
Rango de edad=(15,20]	-0.22	96	6	90
Médico=No	-0.26	1108	67	1041
Resfriados por año=Menos de 2	-0.28	1887	112	1775
Estado civil=Casado/Union Libre	-0.28	1468	87	1381
Rango de edad=(50,55]	-0.29	291	17	274
Mascotas=No	-0.3	859	50	809
Institución=UACM	-0.36	73	4	69
Ejercicio=Más de 4 hrs/sem	-0.36	549	30	519
Estado civil=Divorciado	-0.37	221	12	209
Sexo=Masculino	-0.42	1442	75	1367
Cohabitantes=Ninguno	-0.5	251	12	239
Rango de edad=(10,15]	-0.51	21	1	20

Continúa en la siguiente página

Continuación de la tabla				
Variable	$S(x_i)$	N_{x_i}	$N_{C \cap x_i}$	$N_{\neg C \cap x_i}$
Ocupación=Salud	-0.58	112	5	107
Ocupación=Industrial	-0.67	122	5	117
Escolaridad=Secundaria	-0.75	79	3	76
Ocupación=Jubilado	-1.34	47	1	46
Rango de edad=(60,65]	-1.56	116	2	114

En la figura 3.2 se muestra el modelo de riesgo para la clase *gripe* y la distribución de frecuencias del score.

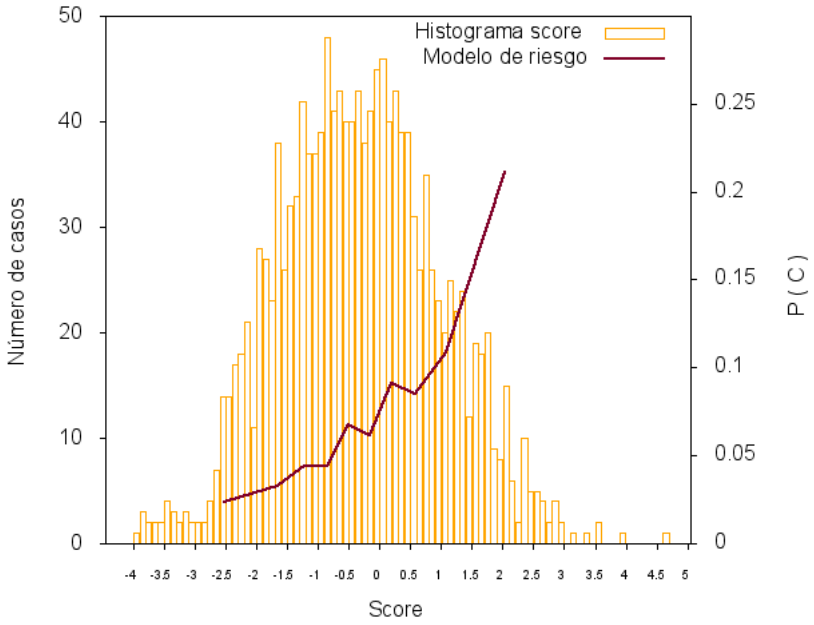

Figura 3.4: Modelo de riesgo y frecuencias score: Gripe

Sabiendo que $P(C) = .075$ y retomando el mismo criterio para estimar las fronteras de riesgo (pero esta vez cuadriplicando la proporción de $P(C)$), se tiene que los intervalos de riesgos quedan delimitados de la siguiente manera:

- Riesgo alto: $P(C|X) > 1.30P(C)$
- Riesgo medio: $0.70P(C) \leq P(C|X) \leq 1.30P(C)$
- Riesgo bajo: $P(C|X) < 0.70P(C)$.

Así, aquellos usuarios cuyo score sea mayor que 1 tienen un riesgo alto de pertenecer a la clase *gripe*, el riesgo bajo se pronostica para puntuaciones menores de -1. El riesgo intermedio se sitúa entre estos dos valores. Al utilizar este modelo para los casos de la base de prueba, los

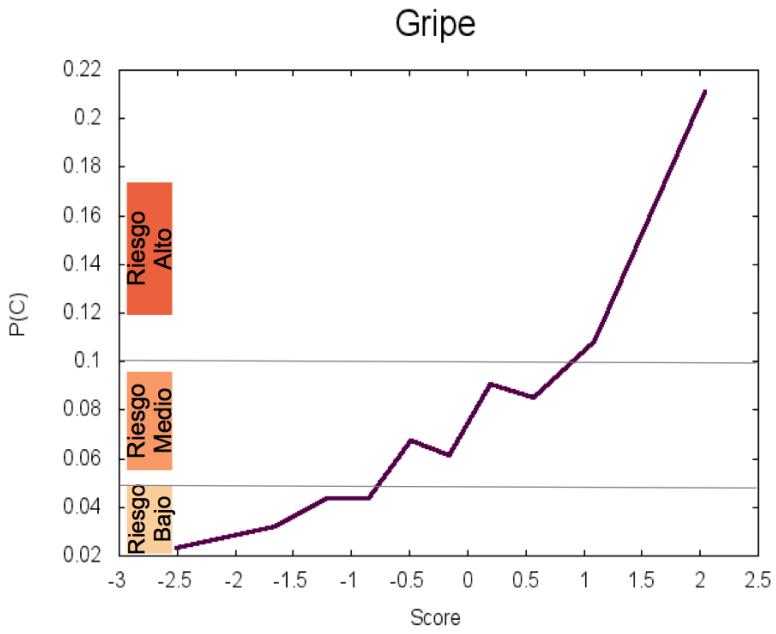

Figura 3.5: Niveles de riesgo: Gripe

"verdaderos positivos"para riesgo alto fueron 13 de 239 (5.44 por ciento), para riesgo medio 53 de 768 (6.90 por ciento) y para riesgo bajo: 38 de 454 (8.26 por ciento).

3.5. Análisis de series de tiempo

El objetivo de este análisis es conocer el comportamiento que presentan las series de tiempo de las clases: *sospechosos* y *gripe*, así como para cada uno de los 11 síntomas (enlistados en el cuestionario rutinario) relacionados con enfermedades respiratorias; se analizará la tendencia y estacionalidad así como las correlaciones existentes entre ellas. El periodo de estudios comprende desde el 1 de mayo del 2009 (inicio del proyecto) al 30 de septiembre del 2011.

Se sigue un modelo aditivo para la descomposición de las series de tiempo y se usará la trasformada rápida de Fourier para conocer el periodo de las mismas. La unidad de tiempo utilizada fueron semanas para las clases y meses para los síntomas individuales.

Cabe mencionar que el análisis de series de tiempo se realizará bajo el esquema aditivo de tiempo porque los coeficientes de variación diferencial fueron mayores que los coeficientes de variación estacional.

Serie de tiempo	$CV(d)$	$CV(c)$
Participación	-0.57	0.18
Sospechosos de influenza	-1.77	0.67
Gripe	-3.89	1.16

En primer lugar se analizó el número acumulado de personas registradas por mes, que se ilustra en la figura 3.6 y con mayor detalle en la tabla de eventos en donde se describen los sucesos que influyeron de forma espontánea o acciones que se efectuaron con la intención de promover la participación en Reporta.

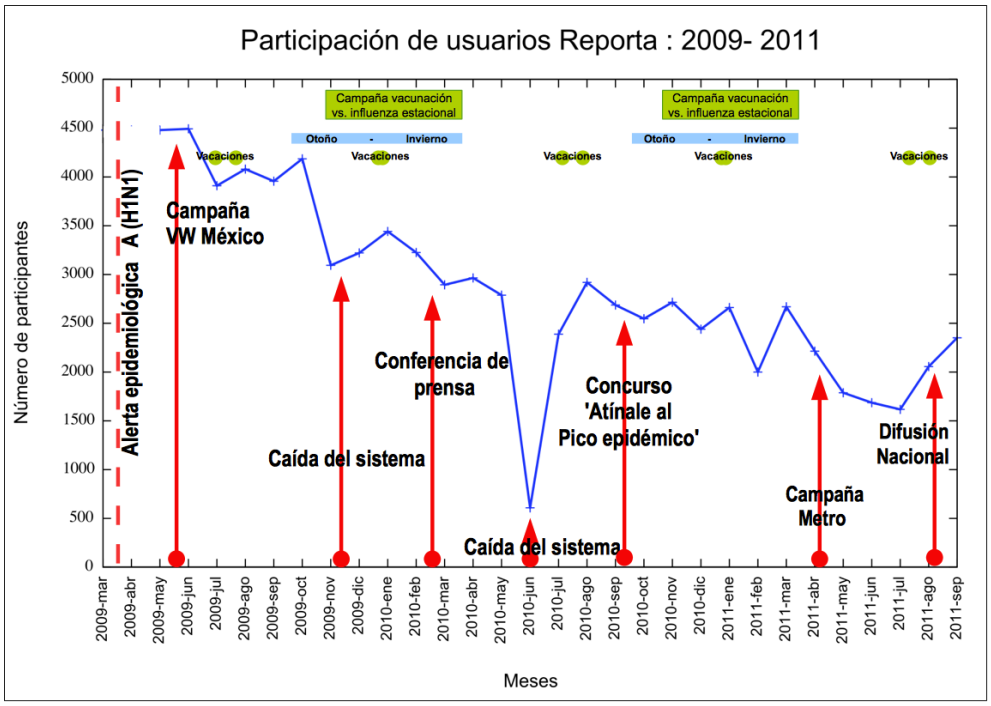

Figura 3.6: Serie mensual de la participación de usuarios en Reporta y contexto temporal

A continuación se describen cada uno de los eventos:

Evento	Descripción
Alerta epidemiológica A(H1N1)(17 de marzo 2009)	Se decreta la alerta epidemiológica en México para lo que se creía en un inicio influenza estacional y neumonía atípica grave; una semana más tarde, el 23 de abril, la Secretaría de Salud da a conocer la circulación de una nueva cepa de virus de influenza A (H1N1). Se tomaron medidas de promoción y prevención de esta enfermedad que consistían en la mejora e higiene del entorno, enfásis en nuevos hábitos de higiene personal y distanciamento social que con frecuencia la Secretaría de Salud junto con el Instituto Nacional de Enfermedades Respiratorias (INER) emitía en los distintos medios de comunicación. $http://portal.iner.gob.mx/archivos/criterios_param_amb.pdf$
Campaña de vacunación contra influenza estacional (octubre-marzo)	Campaña de vacunación anual gratuita realizada por la SSA. Inicia a finales del mes de octubre y continúa durante los meses de invierno.
Otoño-Invierno (23 sep- 21 mar)	Estaciones del año en los que se desarrolla la influenza estacional para el hemisferio norte.
Vacaciones verano (julio-agosto)	Receso escolar en verano.
Vacaciones invierno (diciembre -enero)	Receso escolar en invierno.

Evento	Descripción
Campaña VW México (12 de mayo 2009)	Volkswagen México, a través del área del comunicación corporativa, envió a sus colaboradores un mensaje adjunto vía e-mail que les informaba acerca del proyecto Reporta y les hacía la invitación a participar en el mismo.
Conferencia de prensa (16 de febrero 2010)	Presentación El Sistema Ciudadano de Monitoreo de Enfermedades Respiratorias Reporta ante medios de prensa. La conferencia tuvo lugar en Ciudad Universitaria.
Concurso 'Atínale al pico epidémico' (septiembre 2010)	Lanzamiento de la convocatoria al primer concurso del proyecto Reporta. Dirigido a todos los usuarios del sistema y consistió en calcular la semana con más casos sospechosos de influenza en el portal de Reporta.
Campaña en el metro de la Cd. de México (1 de abril 2011)	Se pegaron carteles informativos del Proyecto Reporta en distintas estaciones del Metro de la Cd. de México con el fin de dar a conocer el proyecto y así promover la participación del mismo.
Difusión nacional	Se envían carteles y trípticos a investigadores y encargados de difusión en distintas universidades del país con el fin de promover el proyecto.
Caídas del Sistema (2 al 8 nov 2009 y 31 may al 27 jun 2009)	Por fallas en el servidor de Reporta, el envío de correos rutinarios semanales no se completó satisfactoriamente.

La participación en Reporta, medida con el número de cuestionarios respondidos semanalmente, tiene una tendencia decreciente. Los primeros 5 meses del proyecto fueron los de mayor participación, hecho que fue motivado por la pandemia de influenza que vivía el país en aquellas fechas. También se observa que la participación incrementa en otoño e invierno y los descensos de participación coinciden con los periodos vacacionales. Las acciones que motivaron el registro de nuevos participantes al proyecto Reporta fueron: la conferencia de prensa (durante la semana posterior a dicho evento se unieron 202 usuarios a nivel nacinal), la difusión nacional (después de no tener nuevos participantes semanas atrás, comenzaron a registrarse 22 universitarios a nivel nacional) y el concurso 'Atínale al pico epidémico' (pasaron de haber 8 nuevos registros 2 meses, a 42 en los siguientes 2 meses).

Tomando una menor granularidad en la unidad de tiempo, se representan los datos de manera semanal para identificar de manera más clara los puntos de desplome por error en el

sistema que para fines de análisis, serán las semanas en las series de tiempo que se editarán usando promedios móviles.

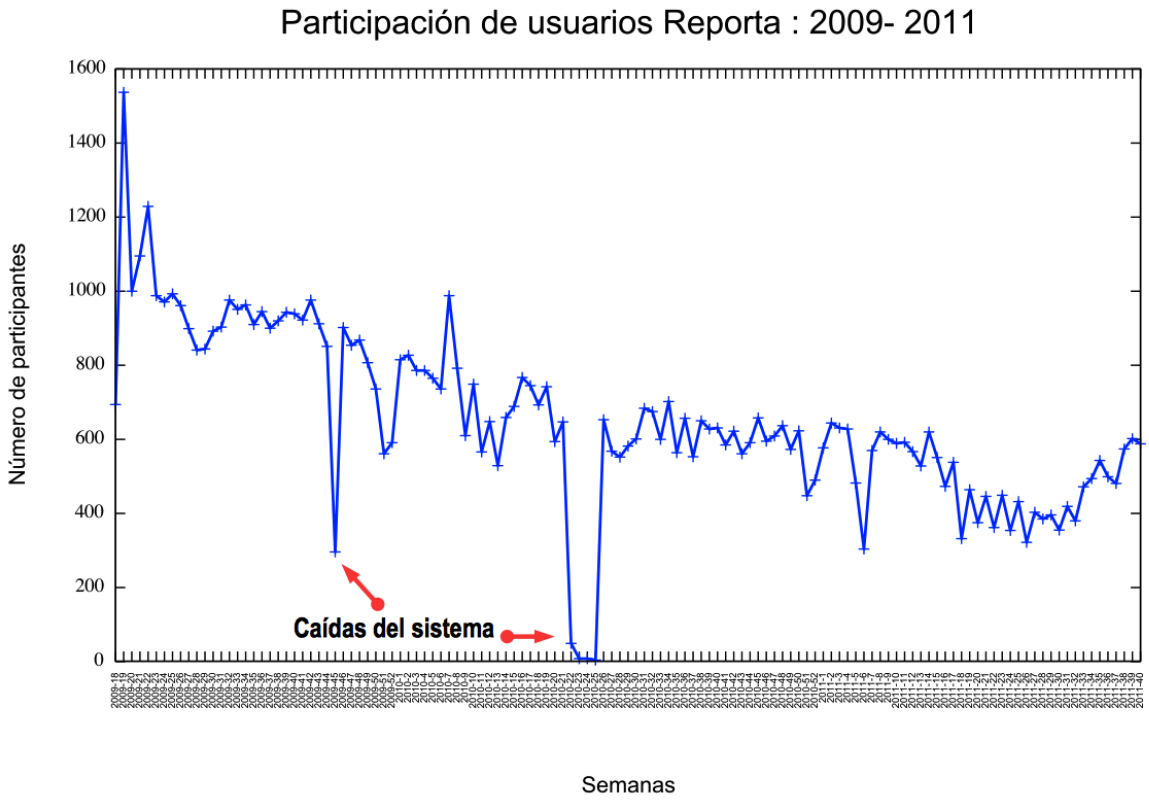

Figura 3.7: Serie semanal de la participación de usuarios en Reporta

Una vez editada la series de tiempo, se procedió a analizarla. La tendencia lineal está dada por la recta: $f(x) = -4.73x + 977.95$ cuya pendiente negativa indica una tendencia descendente con un cambio o razón promedio de casi 5 usuarios por semana. y la tendencia polinomial por la ecuación: $f(x) = 0.000009x^4 - 0.0029x^3 + 0.3493x^2 - 22.463x + 1283.7$.

Para conocer la periodicidad de la serie de tiempo se resta a la serie de tiempo la tendencia de mejor ajuste (polinomial) y sobre este nuevo refinamiento se calculó la transformada

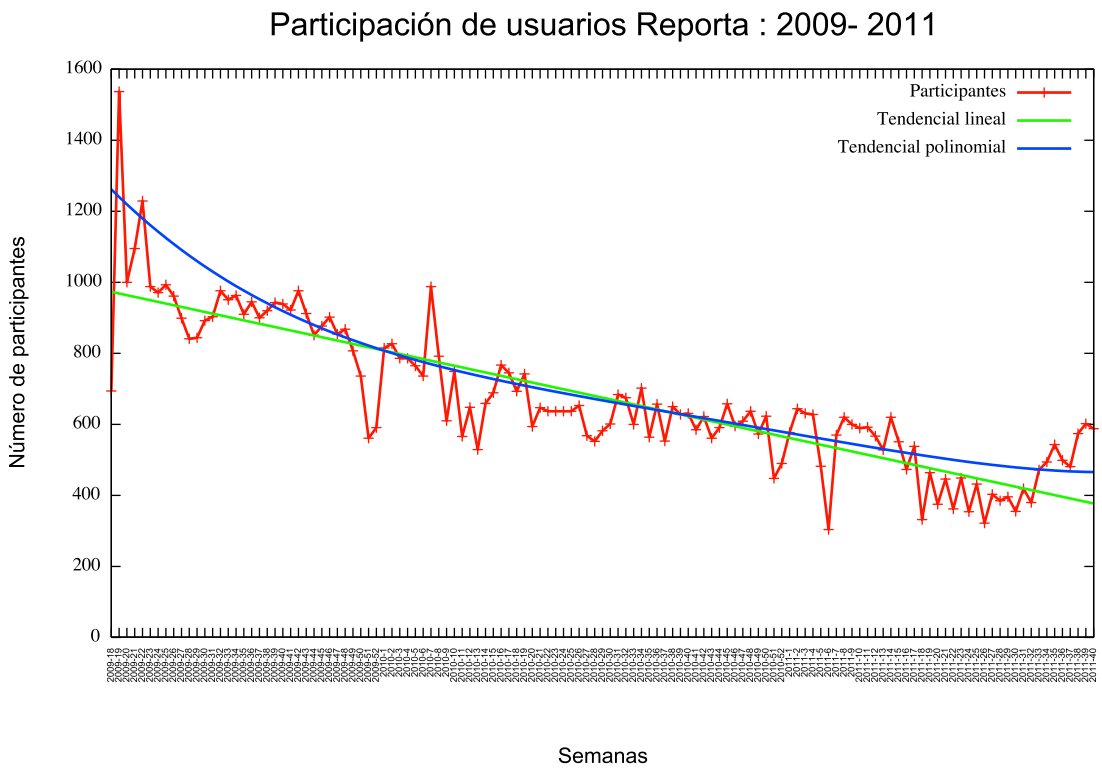

Figura 3.8: Serie semanal de registros en Reporta: Análisis de tendencia.

discreta de Fourier de la serie de tiempo semanal mediante el algoritmo de la transformada rápida de Fourier, análogamente se calculó y graficó el espectro de potencias de dicha serie. El periodograma resultante revela dos señales importantes: la señal de mayor magnitud representa un periodo de año y dos meses (equivalente a 63 semanas aproximadamente) y la segunda señal otro de 5.8154 meses equivalente a 25 semanas aproximadamente.

Para la clase *sospechosos*, la serie de tiempo se comporta como lo ilustra la figura 3.12.

La tendencia lineal está dada por la recta: $f(x) = -0.0827x + 18.309$ cuya pendiente negativa indica una tendencia ligeramente descendente con un cambio o razón promedio de casi .08 usuarios sospechosos de influenza por semana. La tendencia polinómica por la ecuación: $f(x) = 0.0000000015669x^6 - 0.00000063593x^5 + 0.000097526x^4 - 0.0069972x^3 + 0.23476x^2 - 3.2602x + 29.426$.

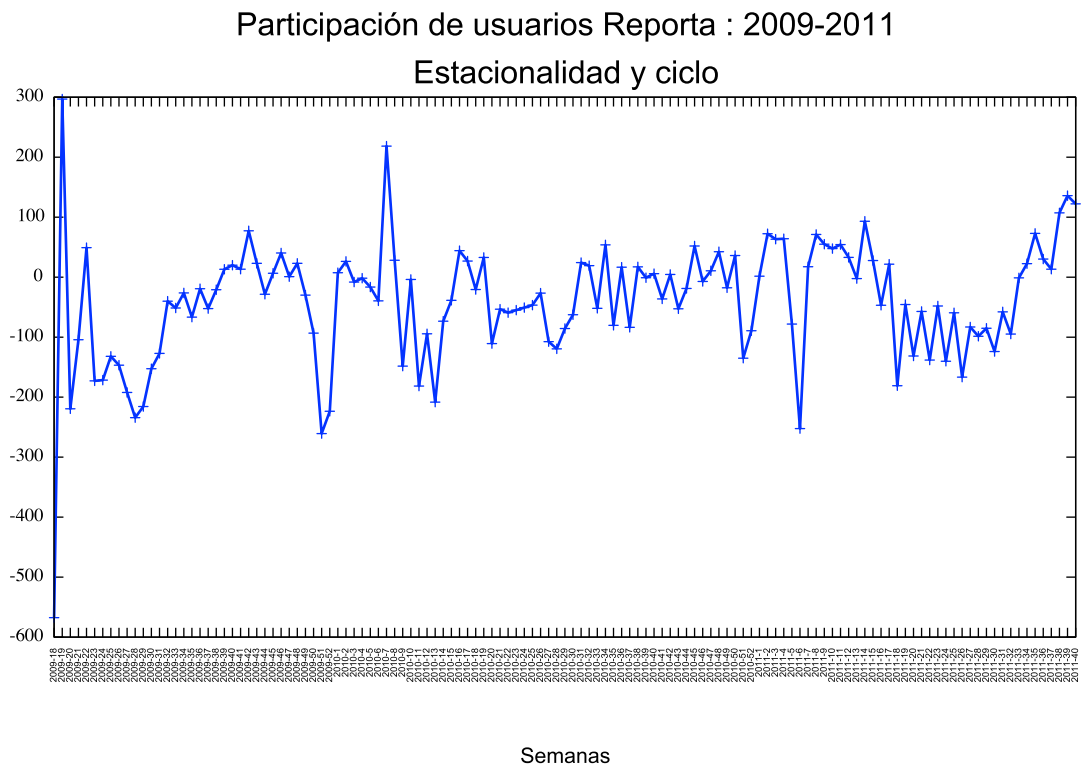

Figura 3.9: Serie de tiempo sin tendencial polinomial de la participación en Reporta.

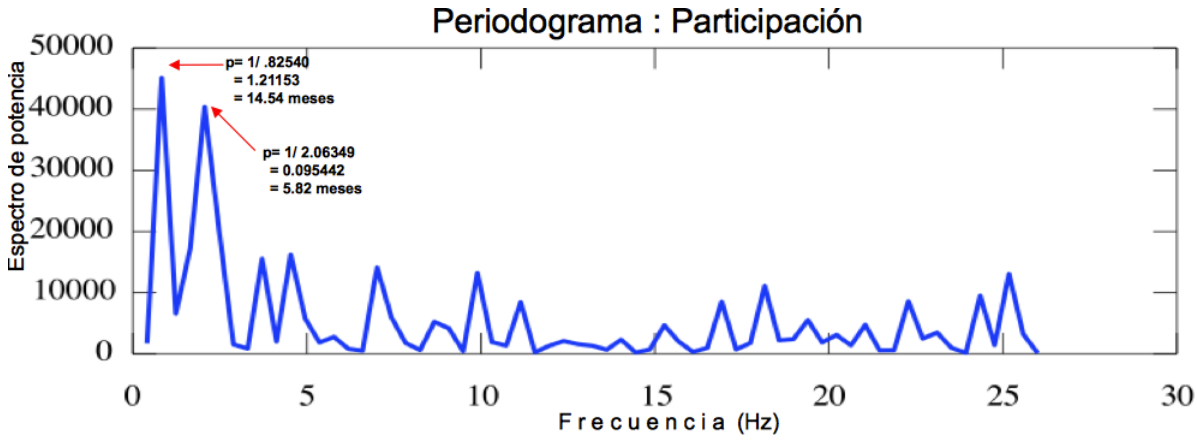

Figura 3.10: Periodograma para la serie de tiempo de participación

El periodograma revela dos señales importantes para la serie de usuarios sospechosos de influenza: la señal de mayor magnitud representa un periodo de 10.30 meses equivalente a 45 semanas aproximadamente y la segunda señal otro de 4.42 meses equivalente a 19 semanas aproximadamente.

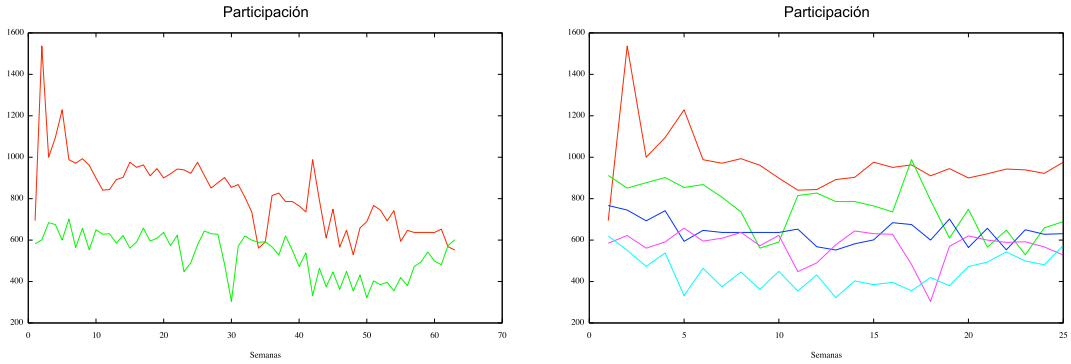

Figura 3.11: Gráficas que muestran a la serie de tiempo de participación dividida en pedazos de 63 y 25 semanas respectivamente.

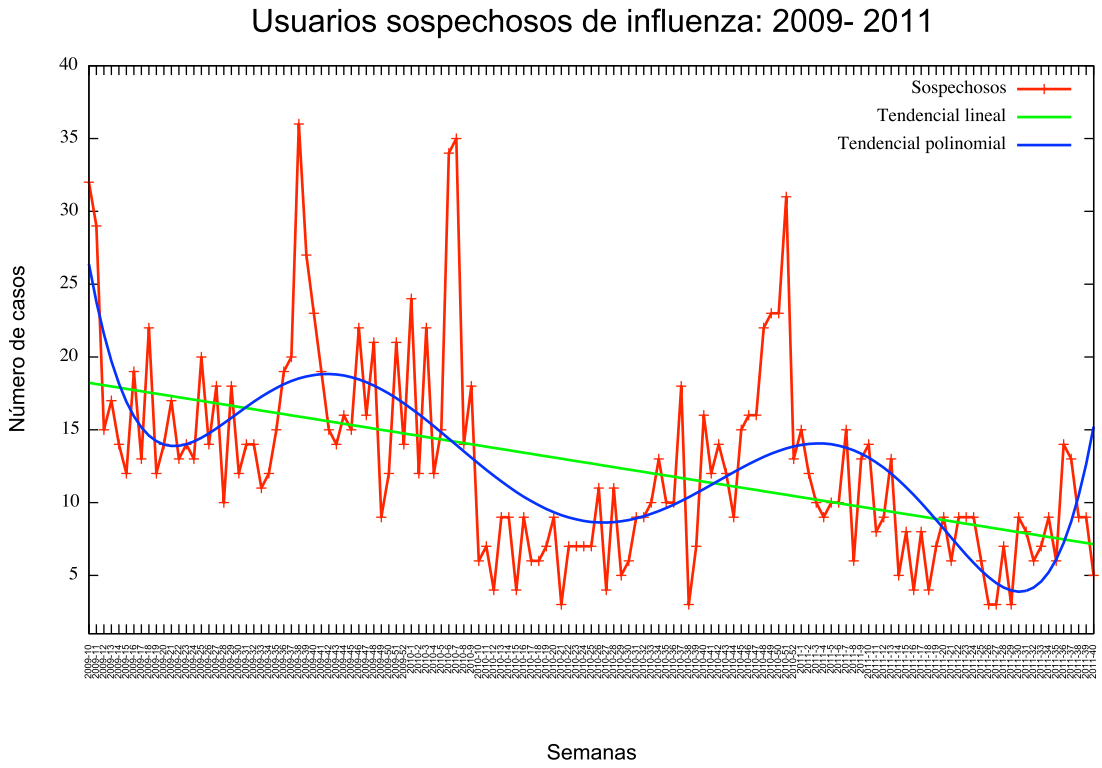

Figura 3.12: Análisis de tendencia

Para la clase *gripe*, la serie de tiempo se comporta como lo ilustra la figura 3.16. La tendencia lineal está dada por la recta: $f(x) = -0.0827x + 18.309$. La tendencia polinómica por la ecuación: $f(x) = 0.0000000015669x^6 - 0.00000063593x^5 + 0.000097526x^4 - 0.0069972x^3 + 0.23476x^2 - 3.2602x + 29.426$.

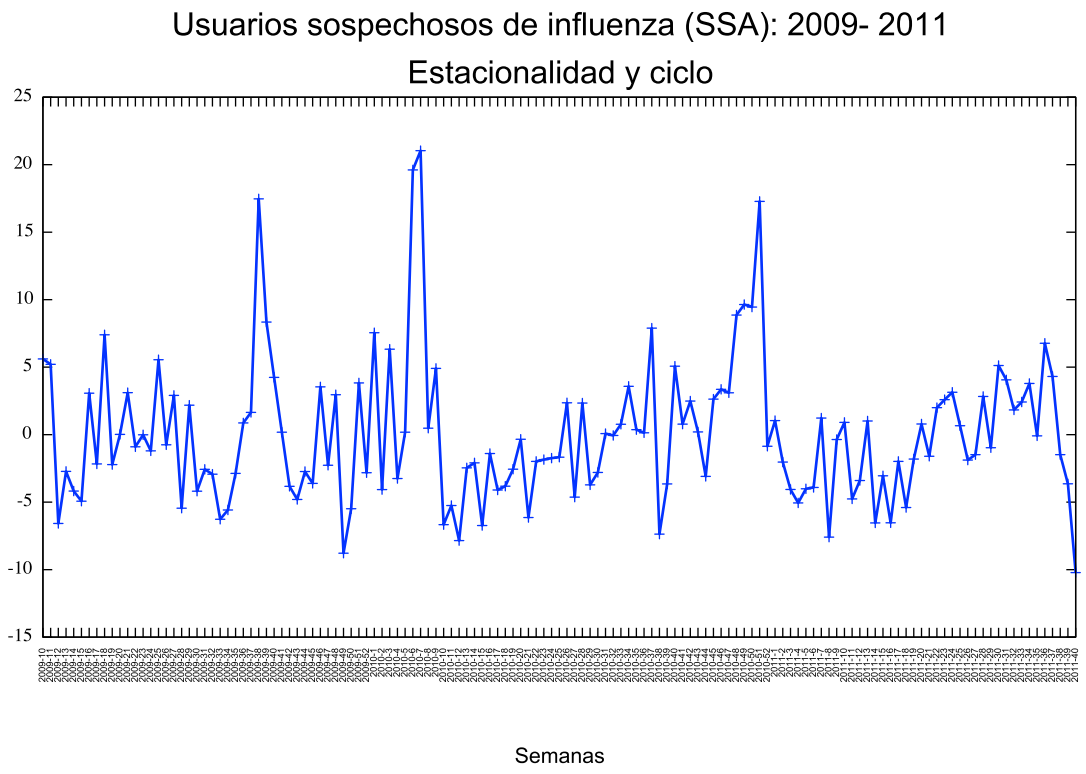

Figura 3.13: Serie de tiempo sin tendencial polinomial de los usuarios sospechosos de influenza según la definición de la SSA.

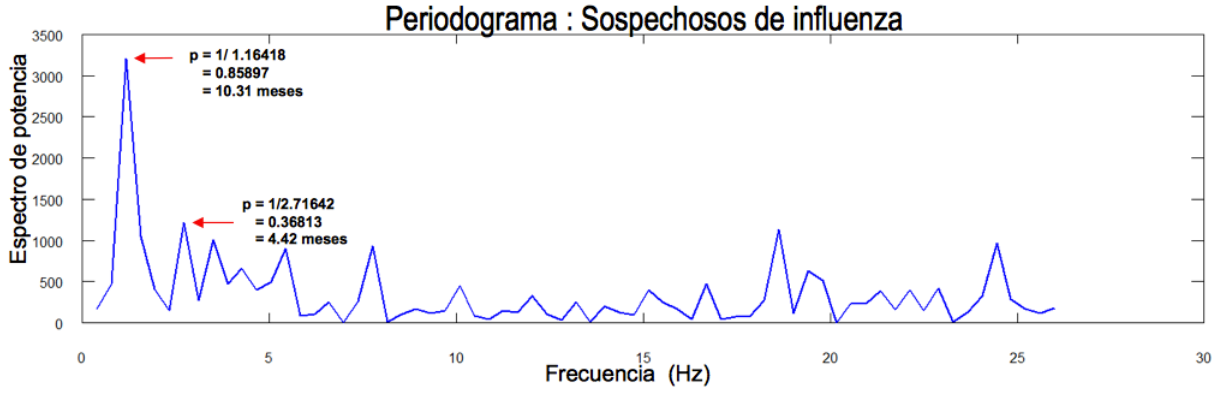

Figura 3.14: Periodograma para la serie de tiempo de los usuarios sospechosos de influenza según la definición de la SSA.

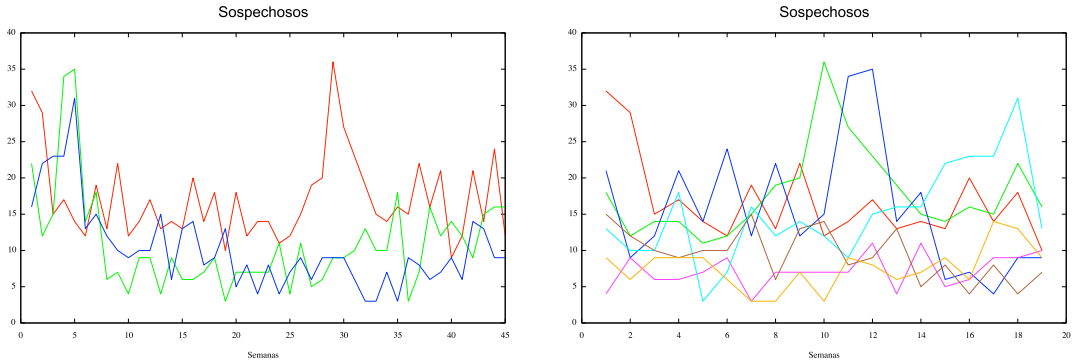

Figura 3.15: Gráficas que muestran a la serie de tiempo sospechosos de influenza dividida en pedazos de 45 y 19 semanas respectivamente.

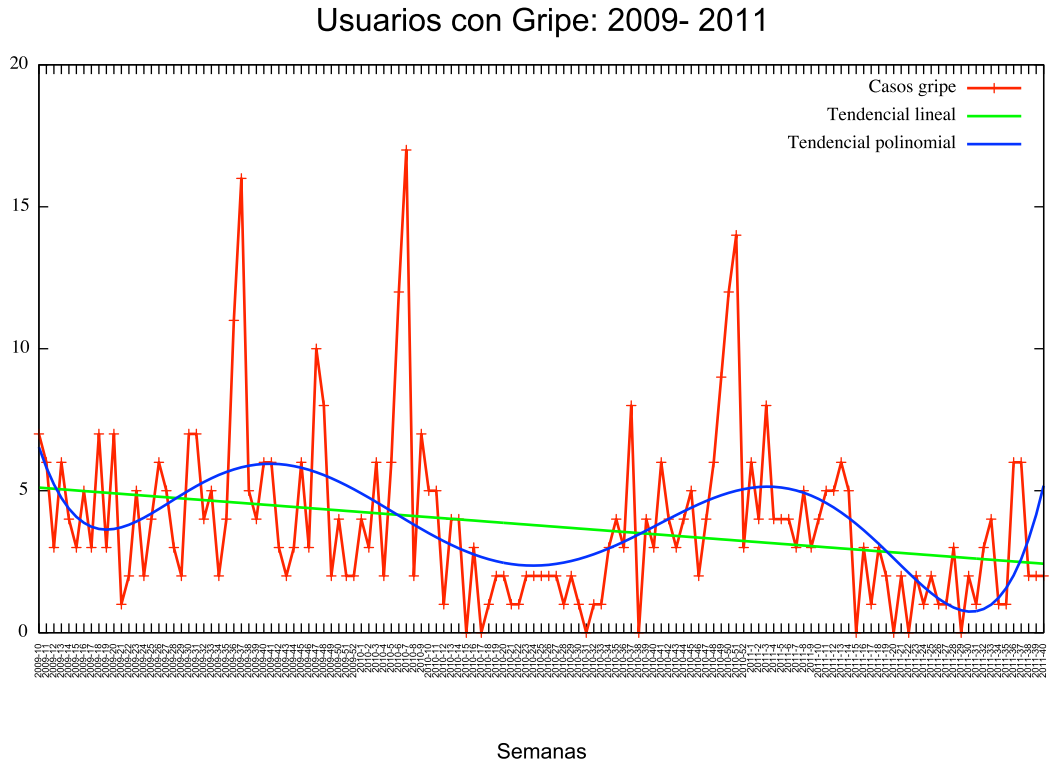

Figura 3.16: Análisis de tendencia

En este caso, el periodograma nos revela que el patrón temporal de los usuarios que padecieron gripe, se repite cada 5 semanas.

En la figura 3.20 se presenta el análisis individual de series de tiempo de los síntomas presentados por los usuarios de Reporta. De manera general se muestra el panorama de los

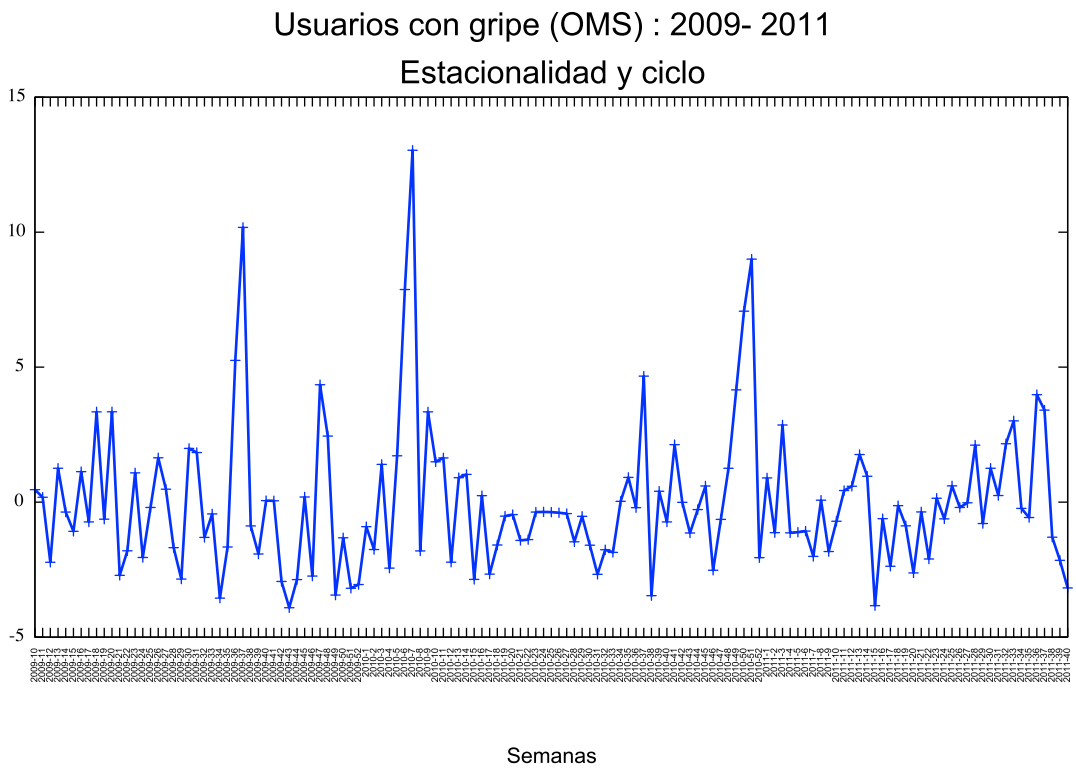

Figura 3.17: Serie de tiempo sin tendencial polinomial de los usuarios que tuvieron gripe según la definición de la OMS.

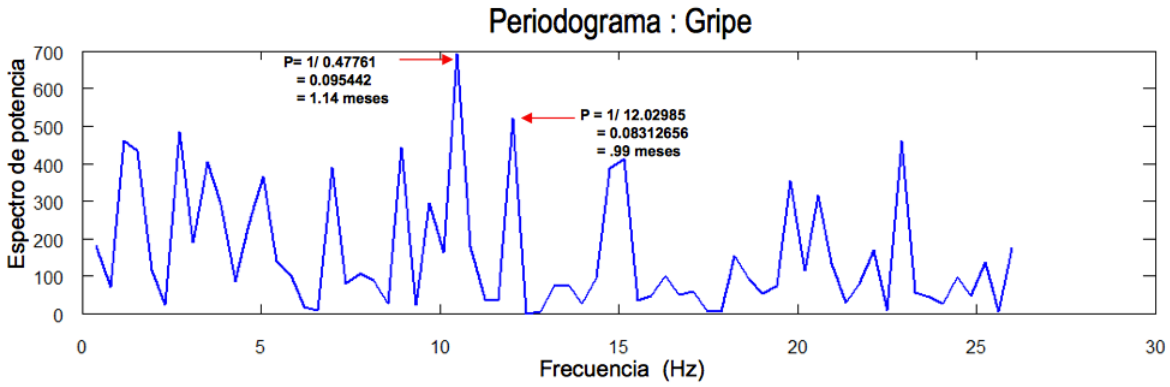

Figura 3.18: Periodograma para la serie de tiempo de gripe

11 síntomas.

La tabla 3.21 muestra cual es el comportamiento de cada uno de los síntomas a lo largo del tiempo de manera mensual, en ella también se incluye la tendencia de cada uno de los casos.

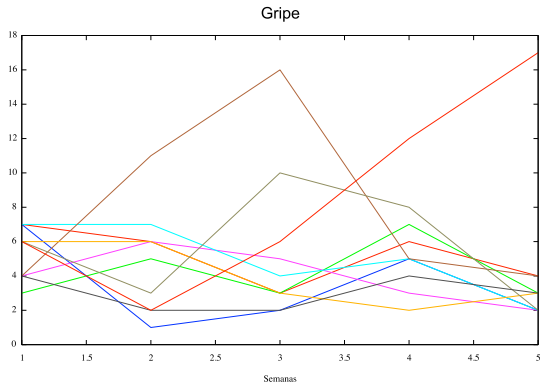

Figura 3.19: Serie de tiempo de gripe dividida en pedazos de 5 semanas.

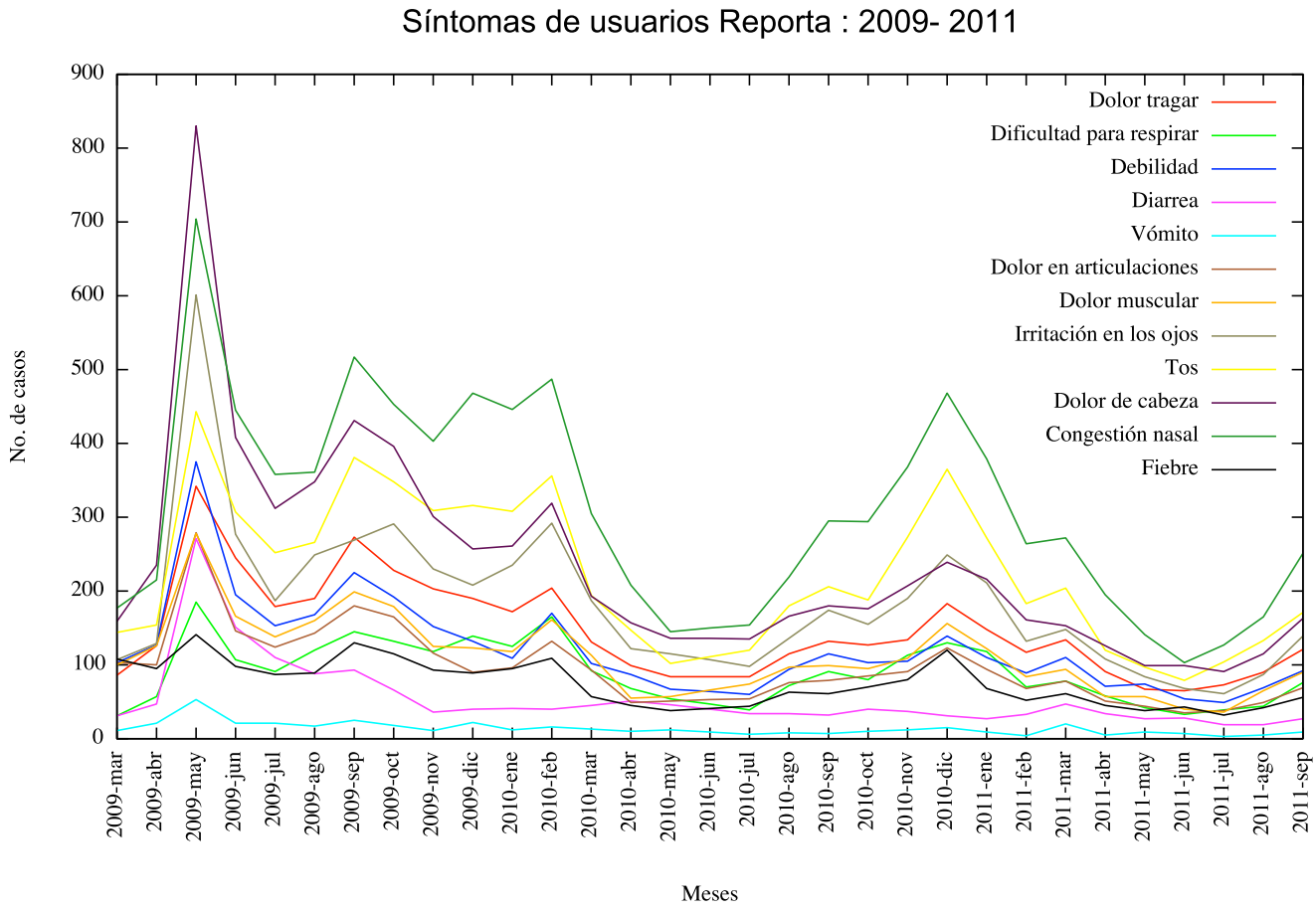

Figura 3.20: 11 síntomas de usuarios Reporta

Después de quitar la tendencial polinomial y calculado las frecuencias, los resultados revelan que los 11 síntomas siguen un patrón temporal de diez meses, que refleja la duración de ciclos en el número de participantes que reportan un síntoma dado. Los periodos secundarios oscilan entre los 3 y 5 meses, distancia que concuerda con el cambio de las estaciones de año.

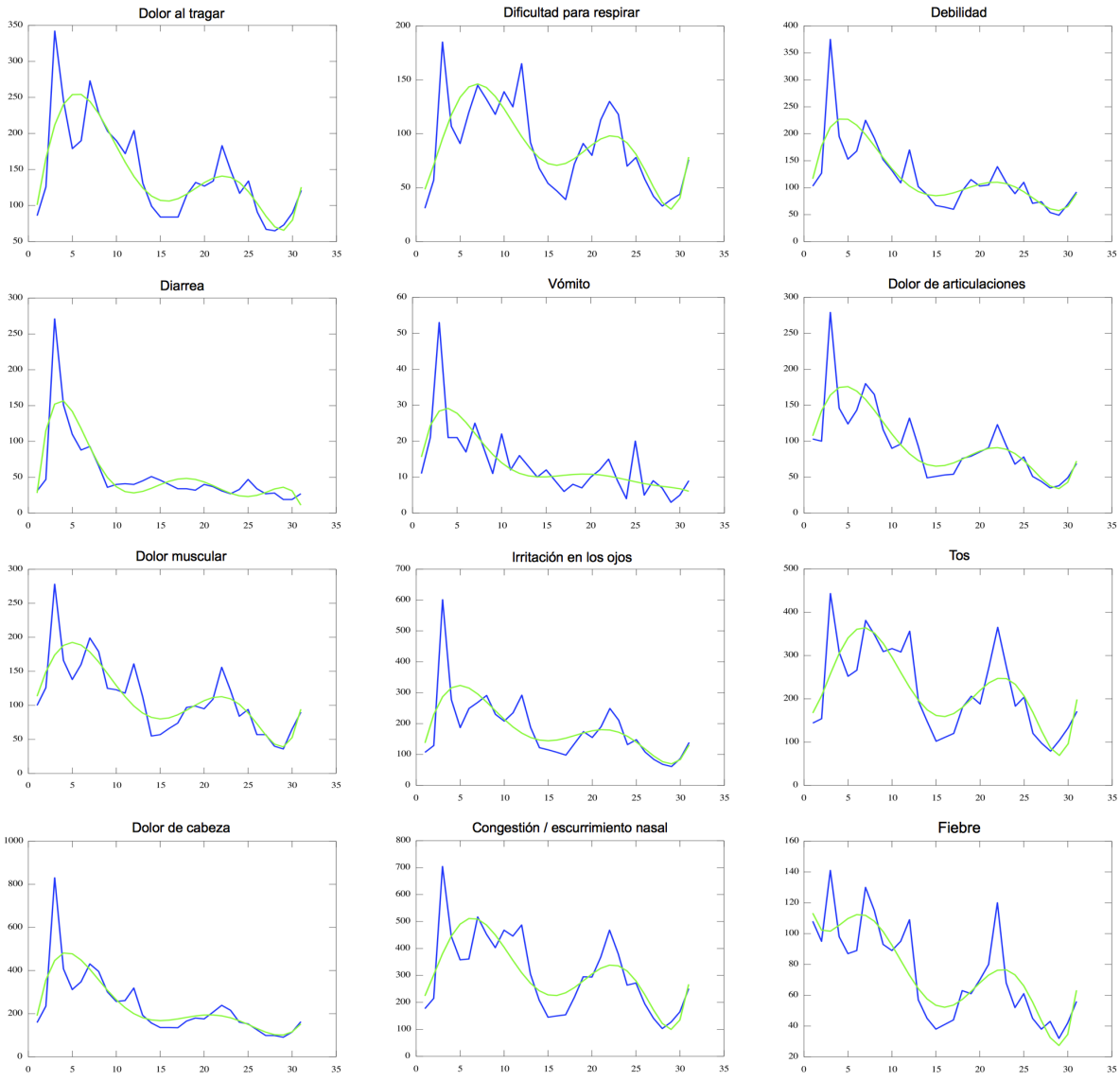

Figura 3.21: Series de tiempo con tendencia polinomial para cada uno de los síntomas.

Síntoma	**Periodo anual 1**	**Periodo mensual 1**	**Periodo anual 2**	**Periodo mensual 2**
Dolor al tragar	0.8333	10	0.3125	3.75
Dificultad para respirar	0.8333	10	0.3125	3.75
Debilidad	0.8333	10	0.3125	3.75
Diarrea	0.8333	10	0.2778	3.33
Vómito	0.2500	3	0.8333	10
Dolor en articulaciones	0.8333	10	0.3571	4.3
Dolor muscular	0.8333	10	0.4167	5
Irritación en los ojos	0.8333	10	0.3125	3.75
Tos	0.8333	10	0.3125	3.75
Dolor de cabeza	0.8333	10	0.3571	4.3
Congestión o escurrimiento nasal	0.8333	10	0.3125	3.75
Fiebre	0.8333	10	0.3571	4.3

Cuadro 3.12: Periodos más importantes de las series de tiempo de síntomas.

Cabe destacar las similitudes de periodo que existen entre las series de tiempo para los síntomas: dolor al tragar, dificultad para respirar, debilidad, irritación en los ojos, tos y congestión o escurrimiento nasal. Varios de estos síntomas describen afecciones respiratorias como resfriado o gripe, que se presentan con mayor frecuencia durante la época de otoño-invierno. Otros síntomas que tiene periodos similares son: dolor en articulaciones, dolor de cabeza y fiebre.

Para conocer de manera más precisa la relación que existe entre estas series de tiempo se calcuó sus coeficientes de correlación. Después de haber sustraido su tendencia polinomial, las correlaciones obtenidas se muestran en la figura 3.22.

	Dolor al tragar	Dificultad respirar	Debilidad	Diarrea	Vómito	Dolor articulaciones	Dolor muscular	Irritación ojos	Tos	Dolor cabeza	Congestión nasal	Fiebre	Sospechosos	Gripe
Dolor al tragar	1.0000													
Dificultad respirar	0.9062	1.0000												
Debilidad	0.9482	0.8706	1.0000											
Diarrea	0.7668	0.6506	0.8399	1.0000										
Vómito	0.7800	0.7479	0.8387	0.8065	1.0000									
Dolor articulaciones	0.9241	0.8542	0.9630	0.7881	0.7746	1.0000								
Dolor muscular	0.9297	0.9026	0.9297	0.6920	0.7742	0.9650	1.0000							
Irritación ojos	0.9297	0.9139	0.9529	0.8493	0.8033	0.9448	0.9238	1.0000						
Tos	0.9339	0.9427	0.8460	0.6076	0.6892	0.8572	0.9056	0.8767	1.0000					
Dolor de cabeza	0.9291	0.8637	0.9718	0.8782	0.8184	0.9550	0.9094	0.9782	0.8365	1.0000				
Congestión nasal	0.9486	0.9635	0.8822	0.7218	0.7646	0.8670	0.9029	0.9263	0.9673	0.8881	1.0000			
Fiebre	0.8333	0.8164	0.7564	0.4399	0.5996	0.8074	0.8562	0.7398	0.9097	0.7212	0.8352	1.0000		
Sospechosos	0.6148	0.6101	0.4764	0.1423	0.3283	0.5478	0.6232	0.4501	0.7567	0.4187	0.6368	0.9110	1.0000	
Gripe	0.5734	0.6138	0.4833	0.1619	0.3834	0.5493	0.6401	0.4388	0.6565	0.3849	0.5871	0.7329	0.7708	1.0000

Figura 3.22: Correlación de las 11 series de tiempo de síntomas y 2 clases: *gripe* y *sospechosos*.

A partir de la matriz de correlaciones se concluye lo siguiente:

- Existe una fuerte correlación positiva en el comportamiento de las series de tiempo de los síntomas: dolor al tragar, dificultad para respirar, dolor en las articulaciones, dolor muscular, irritación en los ojos, tos, dolor de cabeza, congestión nasal y fiebre; cuadro sintomático que sugiere la presencia de resfriados, gripes, infecciones de vías respiratorias o afecciones virales.

- Por otro lado, las series de tiempo de: vómito, diarrea, dolor de cabeza, debilidad e irritación en los ojos se correlacionan fuertemente.

- La serie de tiempo de fiebre correlaciona fuertemente con las series de tiempo: dolor en las articulaciones, dolor muscular, tos, congestión nasal, dificultas para respirar y dolor al tragar. Síntomas que representan la presencia de alguna infección en vías respiratoria o de alguna afeccion viral grave.

- La serie de sospechosos de influenza tiene una fuerte correlación con las series de tiempo de los síntomas: dolor al tragar, dificultad para respirar, dolor muscular, tos, congestión nasal y fiebre. La serie de tiempo de gripe tiene una fuerte correlación con las series de tiempo de los síntomas: dificultad para respirar, dolor muscular, tos, congestión nasal y fiebre.

- Por último, se observa que la serie de tiempo de gripe está muy altamente correlacionada con la serie de tiempo de sospechosos de influenza.

Capítulo 4

Discusión y conclusiones

El hecho de haberme permitido formar parte del proyecto Reporta del Centro de Ciencias de la Complejidad C-UNAM fue una verdadera fortuna para mi y para la realización de esta tesis, ya que además de todo el apoyo y orientación de mis tutores, coordinadores del proyecto, y de la asesoría computacional brindada por los talentosos compañeros de proyecto, desde un principio se me permitió acceder y disponer de su base de datos unificada, bien estructurada y con las variables sociodemográficas definidas, lo que permitió realizar el análisis y minería de datos sin contratiempos. La única modificación que se realizó fue la transformación de la variables: de su valor numérico a texto, misma que podría implementarse de manera permanente en la base fuente para que así las consultas futuras brindaran una óptima lectura de los datos.

Después de haber analizado los datos, encontrado los factores de riesgo para cada una de las clases, calculado la probabilidad de pertenencia a distintos grupos de riesgo de acuerdo con criterios epidemiológicos y analizado los patrones temporales que caracterizan a las series de tiempo de participación, influenza y gripe, se destacó destacaron detalles interesantes y hasta cierto punto curiosos. Los usuarios de Reporta, por sus características sociodemográficas, son en su mayoría de un nivel socioeconómico medio, se transportan en auto o taxi, conviven con frecuencia con algún tipo de animal o mascota, cuentan con universidad como nivel máximo de estudio, estudiantes y docentes conforman poco más del 50 por ciento de la población, un 72 por ciento viven en casa con no más de 3 personas y 57 por ciento acuden a un servicio médico privado cuando enferman. También se tiene que el 84 por ciento de la población realiza menos de una hora al día de ejercicio fśico, y que una cuarta parte de los usuarios pertenece al grupo de riesgo definido por la OMS (menor de 2 años, mayor de 65 y que padece alguna nefermedad crónica).

El algoritmo de conglomerados Ward, utilizado para complementar el análisis exploratorio de datos, sugirió buenas hipótesis (la relación entre medio de transporte, mascotas, la pertenencia a un grupo de riesgo, la estimación del número de resfriados al año por parte del participante y las medidas de salud tomadas al enfermar con la presencia de afecciones respiratorias) mismas que posteriormente fueron comprobadas y complementadas al realizar el análsis de los factores de riesgo para las distintas afecciones respiratorias (sospechosos de influenzas, gripe e influenza) mediante la prueba binomial Épsilon; algoritmo que reveló a su vez el amplio conjunto de factores de riesgo para las distintas clases de afecciones respiratorias, sus probabilidades condicionales asociadas y grado de riesgo.

Las personas de uno y otro sexo entre 55 y 65 años de edad son muy participativas de igual modo los docentes y usuarios de la UNAM son quienes colaboran en el proyecto de manera constante (47 por ciento y 41 por ciento respectivamente de la clase participativa). Aquellos que comparten casa con más de 4 personas, tienen una edad entre 15 y 25 años y se ejercitan menos de 1 horas a la semana son los que menos participan. La tendencia de la participación en Reporta va a la baja, para reactivarla y generar nuevos participantes convendría realizar más campaas de difusión enfocadas a personas con los perfiles participativos, invitándolos a ser "reporteros semanales"de un grupo de personas cercanas a ellos como familia, amigos, pacientes, etcétera, o pegar los carteles informativos en puntos estratégicos como estaciones de autobuses, aeropuertos, unidades habitacionales, en donde hay un mayor número de personas en movimiento de todas las edades y clases económicas y socioculturales. Difundir el proyecto e incentivar la participación mediante redes sociales también podría ser una buena opción.

En cuestión de afecciones respiratorias, se tiene que la convivencia con animales repercute de manera negativa sobre la salud de los usuarios y que las mujeres y jóvenes entre 25 y 30 años son más propensos a enfermar, así como todos aquellos usuarios con factores de salud que los hagan más suceptibles de pescar enfermedades respiratorias: padecer enfermedades respiratorias crónicas, la alta frecuencias de resfriados al año y la pertenencia al grupo de riesgo. En contraparte, resaltan los factores de aquellos usuarios que están alejados de pertenecer a estas clases: no usan con frecuencia el trasporte público, se transladan en auto o taxi, no tienen mascotas o contacto frecuente con animales, viven en unión libre o son casados y fueron los usuarios de sexo masculino quienes reportaron enfermar menos de estas afecciones respiratorias.

Se logró construir un modelo de riesgo basado en la clasificación bayesiana ingenua que puede predecir el perfil de riesgo de los participantes de nuevo ingreso de ser sospechosos de influenza o padecer gripe.

En series de tiempo, a pesar de que contabamos con una historia de dos años y medio de datos medidos de manera mensual y semanal, nos dimos cuenta de los siguientes patrones temporales: la participación aumenta en temporada de otoño invierno y los síntomas asociados a enfermedades respiratorias como los sospechosos de influenza tiene una periodicidad de 10 meses.

Finalmente se concluye que el trabajo de esta tesis podría representar el inicio de diferentes líneas de investigación, como lo son: la relación existente entre el tipo de mascotas o animales con los que se tiene contacto frecuente y las enfermedades respiratorias; los factores fisiológicos que hacen al género masculino ser menos vulnerable de contraer enfermedades respiratorias y también a aquellas personas que se encuentran casadas o viviendo en unión libre, así como la investigación de patrones sintomáticos en distintas zonas climáticas del país. Actualmente, participo con el equipo de Reporta en el desarrollando de un artículo de investigación científica titulado *Factors associated to risk of Influenza like illness* que aborda distintas hipótesis de investigación siendo una de ellas, el tema de la relación existente entre el contacto frecuente con animales y el ser sospechosos de influenza.

Bibliografía

[1] R. Nisbet, J. Elder y G. Miner. Handbook of Statistical Analysis and Data Mining Applications. Elsevier 2009.

[2] M. H. Dunham. Data Mining Introductory and Advanced Topics. Prentice Hall. 2003.

[3] M. Hernández-Ávila. Curso de Análisis de Datos para Epidemiología ambiental. Centro de Investigación en salud Poblacional. Instituo Nacional de Salud Pública. Bello Horizonte, Brasil.1997.

[4] J. De MAst y B.P. Kemper. Principles of Exploratory Data Analysis in Problem Solving: What Can We Learn from a Well-Known Case? Quality Engineering, 21, 366-375, 2009.

[5] M. Hardy y A. Bryman. Handbook of Data Analysis. Sage. 2009.

[6] J. Moncada. Estadística para ciencias del movimiento humano. EUCR. 2005.

[7] J.F. Hair, W.C. Black, B.J. Babin y R.E. Anderson. Multivariate data analysis, Pearson Prentice Hall. 2006.

[8] K-M Colimón. Fundamentos de epidemiología. Díaz de Santos,S.A. 1990.

[9] Pannekoek, Jeroen. Handbook of statistical data editing and imputation,Wiley. 2011.

[10] R. Chen. Intelligent Computing and Information Sciencie, V1. Springer. 2011.

[11] B Mirkin. Mathematical Classification and Clustering. Kluwer Academic Publishers.1996.

[12] J. Ortiz y A. Montenegro. Modelamiento estadístico, Universidad Nacional de Colombia. 2005.

[13] Kaufman, Leonard.Finding Groups in Data, Wiley.1990.

[14] L.S. Gonąlves, R. Rodrigues, A.T. Amaral, M. Karasawa y C.P. Sundré. Comparison of multivariate statistical algorithms to cluster tomato heirloom accesions, Brasil. 2008.

[15] D. Chávez, I. Miranda, M. Varela y L Fernández. Utilización del análisis de cluster con variables mixtas en la selección de genotipos de maíz. Revista de investigación Operacional, Vol. 30, No. 3, 209-216. 2010.

[16] N. Zheng y J. Zue.Statistical Learning and Pattern Analysis for Image and Video Processing.Springer. 2009.

[17] Artículo en desarrollo: Factors associated to risk of ILI. 2012.

[18] J. Han y M. Kamber. Data mining: concepts and techniques, Morgan Kaufmann publications. 2011.

[19] D. F. Nettleton. Técnicas para el análisis de datos clínicos. Ediciones Díaz de Santos. 2005.

[20] P. Domingos and M. Pazzani. Beyond independence: conditions for the opti- mality of the simple Bayesian classifier. Beyond independence: conditions for the optimality of the simple Bayesian classifier. University of California. 2003.

[21] C. Chatfield. The analysis of time series, An Introduction. Chapman and Hall/CRC. 1991.

[22] C. Rodríguez. Análisis de Series Temporales. Cuadernos de Estadística. Madrid: Editorial La Muralla, S.A. 2000.

[23] P. Bloomfield. Fourier Analysis of Time Series: An Introduction. Wiley. 2000.

[24] Cox NJ, Subbarao K. Influenza. Lancet.1999.

[25] Secretaría de Salud. Comunicación Social. Boletín 2009-127a. 17-04-2009.