UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

ESTIMACIÓN DE VENTANILLAS DE SUCURSALES CON TÉCNICA DE MINERÍA DE DATOS

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I A

P R E S E N T A :

LUZ ANGELICA SERRALDE VEGA

DIRECTOR DE TESIS:
M. EN C. JOSÉ SALVADOR ZAMORA MUÑOZ
Ciudad Universitaria, Cd. Mx., 2016

Hoja de datos del jurado

1. Datos del Alumno
Serralde
Vega
Luz Angélica
47 51 23 82
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
085313068

2. Datos del tutor
M en C
José Salvador
Zamora
Muñoz

3. Datos del sinodal 1
Dra Lizbeth
Naranjo
Albarrán

4. Datos del sinodal 2
Dra
Amparo
López
Gaona

5. Datos del sinodal 3
Act
Ángel Manuel
Godoy
Aguilar

6. Datos del Sinodal 4
Dr
Ricardo
Ramírez
Aldana

7. Datos del trabajo escrito
ESTIMACION DE VENTANILLAS DE SUCURSALES
CON TECNICA DE MINERIA DE DATOS
117 p
2016

Dedico este trabajo a mis padres.

Agradezco al equipo de Bancomer, que me apoyaron en el entendimiento del proyecto. A Azucena por su ayuda y apoyo constante.

Tabla de contenido

Presentación

El objetivo del presente trabajo constituye un ejemplo para mostrar la relevancia de la técnica de minería de datos aplicada a una red de sucursales de un banco, para proponer de manera óptima el número de personas necesarias en ventanilla y lograr un nivel de servicio establecido por el banco.

Se describe la importancia de los datos para la extracción del conocimiento de patrones y determinar, bajo cierto nivel de certidumbre, la certeza de la decisión del número ventanillas y de empleados necesarios para atender a los clientes que acuden a la sucursal.

Dentro del capítulo primero, se especifica, de manera general, la definición del término Minería de Datos; las definiciones de cada una de las fases que constituye el desarrollo de un modelo de minería, así como la trascendencia de aplicar una correcta limpieza y depuración de los datos para extraer el conocimiento de los mismos.

También se describen las principales técnicas a utilizar en la búsqueda de patrones y conocimiento, que expliquen y den solución a los diferentes problemas que tiene cualquier institución. Describe el funcionamiento y conveniencia de cada técnica, sus características comunes, así como la tipología de problema que enfrentan. Se estudiarán dos aspectos fundamentales, la expresividad y comprensibilidad de los modelos, que, junto a otras características, permitirán realizar una comparativa de las técnicas más comunes así como una herramienta que ayude a una toma de decisiones más asertiva.

El capítulo 2, muestra el ejemplo de minería de datos aplicado al problema de definir el número óptimo de ventanillas dentro de la red comercial de un banco. Se determina el costo monetario del problema y se cuantifica la posible solución. Se define la importancia de seleccionar y clasificar las variables relevantes para resolver el problema, así como una correcta la definición del objetivo, además, explica las técnicas de depuración de basura, ruido y valores ausentes dentro de la información.

Se explican los modelos que se adaptan al objetivo y resultado del problema, así como el desarrollo de las soluciones bajo diferentes técnicas de modelado. Se definen las herramientas para evaluar el mejor modelo.

El capítulo 2 finaliza con las conclusiones que se obtuvieron al resolver el problema de la red comercial, así como los siguientes pasos que se consideran dentro del proceso de minería de datos y que son esenciales para un proceso de aprendizaje continuo y resultados positivos en la toma de decisiones.

Introducción

Hoy en día es muy frecuente, sobre todo en las grandes empresas, la disponibilidad de grandes cantidades de datos, así como el uso generalizado de herramientas informáticas para el almacenaje y extracción de datos. Gran parte de esta información es histórica, es decir, representa las situaciones que se han producido en el pasado. Aparte de su función de "memoria de la organización", la información histórica es útil para explicar el pasado, entender el presente y predecir el futuro.

La mayoría de las decisiones de las empresas, organizaciones e instituciones se basan en información de decisiones pasadas extraídas de fuentes diversas.

Muchas veces, el método tradicional de convertir datos en conocimiento consiste en un análisis e interpretación de forma manual. El especialista en la materia, por ejemplo un grupo de médicos, puede analizar la evolución de enfermedades infecto-contagiosas entre la población, para determinar el rango de edad más frecuente de los infectados. Esta forma de actuar es lenta, cara y altamente subjetiva; de hecho, el análisis manual, es prácticamente imposible en situación donde el volumen de datos crece exponencialmente. La abundancia de datos desborda la capacidad humana de comprenderlos sin ayuda de herramientas potentes.

Consecuentemente, muchas decisiones importantes se realizan, no sobre la base de la gran cantidad de datos disponibles, sino siguiendo la propia intuición del usuario al no disponer de las herramientas necesarias, con lo que las decisiones tienen falta de credibilidad, falta de productividad e ineficiencia.

Este es el principal cometido de la minería de datos*: **Extracción implícita de conocimiento previamente desconocido y que se encuentra en datos presentes y disponibles de las bases de información.***

Lo que de verdad es interesante es el conocimiento que se puede inferir a partir de los datos, y más aún, la capacidad de poder usar ese conocimiento.

Existen algunas herramientas analíticas que han sido empleadas para analizar datos y que tiene su origen en la estadística, aunque algunos paquetes estadísticos son capaces de inferir patrones a través de los datos. El problema es que no funcionan bien para el tamaño de las bases de datos (cientos de tablas, millones de registros y un número considerable de dimensionalidad, atributos nominales, etc.) y que no se integran bien a los sistemas de información.

Todos los problemas y limitaciones de las aproximaciones clásicas, han hecho surgir la necesidad de nuevas herramientas y técnicas para soportar la extracción de conocimiento útil desde la información disponible, que se engloban bajo el nombre de minería de datos.

Todo este explosivo crecimiento de datos generó a finales de los 80's la aparición de un nuevo campo de investigación que se denominó KDD (Knowledge Discovery in Databases). El proceso KDD ha servido para unir a investigadores de áreas, al principio dispersas, como la Inteligencia Artificial, la Estadística, la Matemática, Técnicas de visualización, Aprendizaje automático o Bases de datos, en búsqueda de técnicas eficientes y eficaces que ayuden a encontrar el potencial de conocimiento que se encuentra inmerso en los grandes volúmenes de datos almacenados.

De modo resumido, puede considerarse la minería de datos como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos. La minería de datos se distingue de las aproximaciones anteriores porque no obtiene información extensional (datos), sino intencional (conocimiento), además de que el conocimiento no es, generalmente, una parametrización de ningún modelo preestablecido, sino que es un modelo novedoso y original extraído de la herramienta.

El resultado de la minería es un conjunto de reglas, ecuaciones, árboles de decisión, redes neuronales, etc. que utiliza técnicas de reconocimiento de patrones y otras técnicas de datos multivariados, las cuales pueden usarse para resolver preguntas como: ¿existe un número de clientes que se comportan de manera diferenciada?, ¿existen asociaciones entre los factores de riesgo para otorgar un seguro de automóvil?

1. Definición de Minería de Datos

EL KDD, es el proceso global de descubrir conocimiento útil desde las bases de datos, mientras que la minería de datos, se refiere a la aplicación de los métodos de aprendizaje y estadísticos para la obtencion de patrones y modelos. Al ser la fase de generación de modelos, comúnmente se asimila el KDD con minería de datos. Además, las connotaciones de aventura y dinero fácil del término de *minería de datos* han hecho que este se use como identificador de área, especialmente en el ámbito empresarial.

De las múltiples definiciones que existen del término **minería de datos**, la definición que hace el instituto SAS describe de mejor manera la idea de este concepto:
Es el proceso de seleccionar, explorar, modificar, modelar y valorar grandes cantidades de datos, con el objetivo de descubrir patrones de comportamiento que puedan ser utilizados como ventaja comparativa respecto a los competidores.

El proceso de minería de datos es aplicable en una gran gama de industrias y proporciona distintas técnicas de análisis según el problema que se quiere resolver.

La minería de datos consta de un conjunto de etapas que deben de realizarse, entre las que figuran las siguientes:

- Identificar un problema relevante.
- Integración y recopilación de datos
    - Selección de Datos
    - Preproceso de datos
    - Transformación de datos
- Aplicación de técnicas de modelado (herramientas para descubrimiento de patrones de comportamiento)
- Interpretación de los modelos obtenidos
- Difusión, uso y monitoreo

Algunas aplicaciones reales que se obtienen con la minería de datos, son:

- Predicción automática de tendencias y comportamientos
- Descubrimiento automático de patrones previamente desconocidos

A continuación se describe más detalladamente cada fase del proceso de minería de datos:

1.1 Identificar un problema relevante

En esta etapa se estudia el problema y se define cuál es la meta del proyecto, para ello se debe ser capaz de conocer:

- Cuál es el objetivo a cubrir
- Qué tipo de resultados se espera obtener
- Formular expectativas de éxito o fracaso del proyecto.

Si el planteamiento del problema se realiza de esta manera, se descubren fácilmente las fuentes de datos, así como los algoritmos de modelado que se aplicarán.

En esta etapa se incluyen costos y beneficios económicos de la realización del proyecto, así como una estimación del tiempo de duración.

1.2 Integración y recopilación de datos

En esta fase de integración y recopilación de datos, se determinan las fuentes de información que pueden ser útiles y cómo conseguirlas, es la que mayor esfuerzo requiere. Esta es una fase de enorme dificultad y a la cual se le da menor importancia porque no se obtienen datos definitivos, sin embargo, en conjunto con una correcta definición del problema, esta fase puede ser la diferencia entre el éxito y fracaso de proyecto de minería de datos; generalmente consta de 3 pasos:

Selección de datos: Se identifican las fuentes de datos y se selecciona el subconjunto de datos necesarios, ya sean tablas de una base de datos o ficheros de texto. El conocimiento que se tenga de la problemática a resolver es esencial, pues la experiencia y conocimiento ayudan a discernir qué características pueden ser interesantes para ser estudiadas.

El correcto entendimiento de la problemática es un factor principal para definir la profundidad histórica que se requiere en los datos. Es importante diferenciar en las fuentes de información, qué valor representa en el tiempo y si este valor seguirá siendo válido a través del tiempo para determinar si la variable ha dejado de ser significativa en el análisis. Generalmente, las variables demográficas y de estilo de vida, tienen un tiempo de vida corto. Por ejemplo, si se está analizando clientes y su nivel de pagos, la variable “Ingreso”, podrá o no ser significativa, dependiendo de la fecha en la cual el valor fue ingresado a la base de datos y la fecha del estudio.

El proceso de selección de datos muchas veces se engloba dentro del concepto más amplio denominado reducción de datos, aunque este término también puede incluir la agregación (si se pasa de instancias de cada día a instancias agregadas mensualmente), la generalización (por ejemplo, si se reemplaza el atributo ciudad por región) o incluso la compresión de datos, eliminando datos redundantes.

En general, cuando se trata de datos estilo-atributo-valor (es decir, una tabla), hay dos tipos de selección aplicables: Selección horizontal (muestreo), donde se eliminan algunas filas (individuos u observaciones) y selección vertical (reducción de dimensionalidad), donde se

eliminan características de todos los individuos. *La manera más directa de reducir el tamaño de una población o conjunto de individuos es realizar muestreo.*
En el caso de la minería de datos se puede plantear dos situaciones, dependiendo de la disponibilidad de la información:
Si se dispone de la población, en todo caso se ha de determinar qué cantidad de datos es necesario y cómo hacer la muestra. En algunos casos, una muestra aleatoria no es aconsejable. Por ejemplo, en el caso de una encuesta, se intenta escoger al menos un mínimo de individuos de cada tipología (edad, género, nivel económico, etc.). No obstante, ésta es la situación ideal, ya que se dispone (con mayor o menor facilidad) de la población total.
Los datos son ya una muestra de la realidad, por ejemplo, las reclamaciones registradas por el centro de atención a clientes solo son una muestra de la realidad, el resto de las reclamaciones no se registran.

Dentro de las razones para realizar un muestreo están, principalmente, la reducción del tamaño (con el objetivo de agilizar y facilitar los algoritmos de minería de datos), pero existen otras como balanceo de clases, cobertura equilibrada del espacio, entre otras.

Existen varios tipos de muestreo: Aleatorio simple, estratificado, por grupos o exhaustivo:

- Muestreo Aleatorio simple: La premisa de este muestreo es que cualquier observación (registro) tiene la misma probabilidad de ser extraída de la muestra. Puede ser con reemplazo o sin reemplazo.
- Muestreo aleatorio estratificado: El objetivo de este muestreo es obtener una muestra balanceada con suficientes elementos de todos los estratos o grupos.
- Muestreo de grupos: En cierto modo es el inverso al anterior, consiste en elegir sólo elementos de unos grupos y descartar aquellos que, por diversas razones, pueden impedir la obtención del comportamiento que nos interesa estudiar.
- Muestreo exhaustivo: Se trata de una generación del muestreo estratificado, con motivaciones similares. El objetivo es cubrir completamente el espacio de instancias y evitar poner muchas observaciones en zonas muy densas.

En algunos casos, se pueden realizar sobremuestras, es decir, “duplicación de registros”. Este concepto, en el que una misma instancia puede repetirse, admite muchísimas aplicaciones, pues además de nivelar las distribuciones de clases (balanceo), se pueden realizar copias que no sean exactamente iguales, como puede ser introduciendo un poco de ruido (error aleatorio) para facilitar la generalización o directamente generar ejemplos aleatorios utilizando la distribución de los atributos observados en los datos reales. En el caso de problemas de clasificación, se puede optar por generarlos etiquetados o no etiquetados. Este tipo de técnicas se utilizan en el caso de que haya pocos datos. Otro uso más popular del muestreo es el realizar muestras dispares para obtener modelos diferentes y después combinarlos, el objetivo es diferente, obtener muestras que varíen significativamente entre si.

Finalmente queda la pregunta, ¿Cuántos datos son necesarios? En el caso de validación de hipótesis (cuando el muestreo se utiliza para obtener el conjunto de prueba) existen técnicas estadísticas que ayudan a decidir cuántos son necesarios para obtener un nivel de significancia preestablecido (algunas basadas en la distribución binomial o en la aproximación normal). No obstante, el objetivo es determinar cuántos datos son necesarios para el entrenamiento del modelo. Esto depende, en general, del número de grados de libertad, que a que a su vez depende del número de atributos (y de los valores posibles de cada uno). Además depende del método de aprendizaje y su expresividad (por ejemplo, una regresión lineal requiere mucho menos datos que una red neuronal). En general, la tendencia en estos casos es utilizar un muestreo progresivo e incremental. En el que se crea la muestra cada vez mas grande (y diferente de ser posible) hasta que se vea que los resultados no varían significativamente entre un modelo y otro.
Con respecto a las variables o datos con las que la información se compone generalmente se dividen en:

Variables cuantitativas

- Discretas
  - Número de empleados
  - Edad
- Continuas
  - Salario
  - Monto del crédito

Variables cualitativas

- Nominales (no admiten un criterio de orden)
  - Género
  - Idioma
  - Estado de la república
- Ordinales (Aquellos donde se establece un orden)
  - Riesgo (Alto, medio, bajo)
  - Grado (Cualquier cualidad del sustantivo, ejemplo: Grado académico, militar)

Preproceso de datos: Tiene como objetivo la eliminación delmayor número posible de datos erróneos o inconsistentes e irrelevantes, y trata de presentar los datos de manera mas apropiada para la minería de datos.

Una vez que se indentificaron los datos a utilizar, se deben de estudiar para, por un lado, entender el significado de los atributos , y por otro lado, detectar errores de integración.

Dado que los datos provienen de diferentes fuentes, pueden contener valores erróneos o faltantes, el entendimiento del significado de los atributos o detectar errores de integración, como pueden ser datos repetidos, o con formato diferente, también se realizan en esta etapa. La inconsistencia, los valores nulos, los valores extremos y el ruido son propiedades de todas las bases de datos, para corregir estas anomalías, se deberán definir diferentes rutinas de limpieza de datos que ayudarán a rellenar datos faltantes, identificación de outliers, etc.
Para las variables categóricas, las distribuciones de frecuencia o gráficas de barras ayudan a la comprensión y significado de la variable. También es útil identificar valores nulos o algunos valores fuera de rango.

Para las variables cuantitativas, la estadística descriptiva con medidas de tendencia central y de dispersión y algunos gráficos nos ayudan a entender su distribución y ver si tienen valores no válidos o fuera de rango. También es útil entender la relación con otras variables.

Una vez analizado, se determina el tratamiento de las inconsistencias los valores con ruido y los valores nulos.

Valores perdidos: Al analizar los datos, existen variables que no tiene registros, y la decisión que se debe de tomar es, qué hacer con ellos. Es importante señalar que todas estas metodologías de eliminación de valores nulos, pueden causar más confusión si se tiene un gran volumen de datos sin valor:

- Ignorar el registro: Generalmente se utiliza cuando se trata de un problema de clasificación, este método es efectivo sólo cuando los valores nulos representan un porcentaje muy bajo.
- Eliminar toda la columna: Es una solución extrema, pero algunas ocasiones se debe de tomar, en casos de que la proporción de datos faltantes sea muy alta.
- Filtrar la fila: Sesga los datos, porque este faltante, puede estar relacionado con casos especiales.
- Rellenar el registro manualmente: En general esta opción es muy poco eficiente y lleva mucho tiempo, en particular si se trata de grandes volúmenes de datos.
- Utilizar una constante global para rellenar el valor nulo.
- Utilizar el valor de la Media o moda.
- Reemplazar el valor existente usando una técnica predictiva de aprendizaje automático. También existen algoritmos que se usan en estos casos, como el algoritmo EM (*Expectation Maximization*).

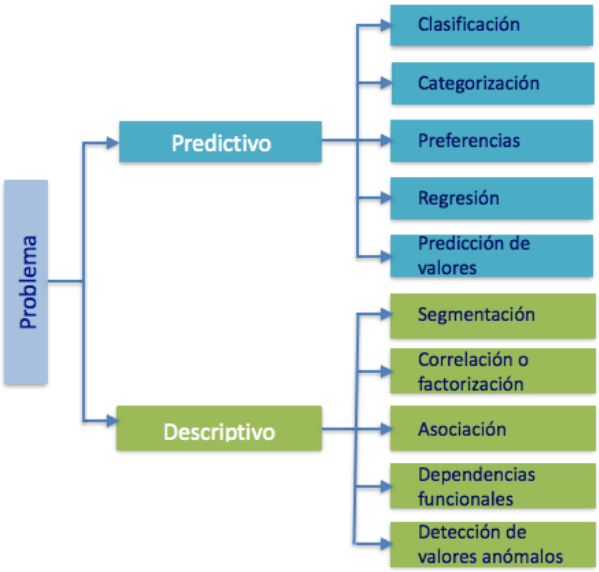
Valores con ruido: Uno de los problemas que afectan la calidad de los datos, es la presencia de valores que no se ajustan al comportamiento general de la variable, es decir, muestran valores fuera de los datos esperados. Estos valores se consideran ruido o excepciones,

pudiendo ser errores en la captura ó valores válidos. Un ejemplo claro de ruido por error de captura de datos es el sueldo negativo. Sin embargo, en el caso de modelos de detección de fraudes, las compras de un tarjetahabiente fuera del comportamiento normal suelen ser los más relevantes.

Transformación de los datos. Para la realización de esta tarea, es de gran ayuda tener una herramienta de visualización y métodos estadísticos, que incluye una fusión de datos tanto horizontal (columnas) como vertical (renglones). Se decide si se agrupan, clasifican, etc., y se elige el método o algoritmo que se va a aplicar. En esta fase se decide si hay que hacer alguna transformación de datos y bajo qué metodología.
En esta fase también se incluye la codificación de los datos para los algoritmos que el programa de minería de datos utiliza. Adicionalmente, en esta fase es cuando se crean nuevas variables.
Cuando se tienen muchas variables sobre una muestra, es presumible que una parte de la información sea redundante, o en el caso extremo, la cantidad de dimensiones (o atributos) sean muchos respecto a las observaciones o registros, para ello se aplica una reducción de dimensionalidad, que significa reducir el número de variables originales, creando nuevas que las representen, pero con la menor pérdida de información posible.

1.3 Aplicación y técnicas de modelado

Para que el proceso de modelado sea correcto es indispensable que se disponga ya de los datos preprocesados y toda la información de entendimiento derivada de los pasos anteriores.
En general, es complicado establecer un procedimiento para determinar la técnica de modelado idóneo a cada problemática, sin embargo, se explicará, a manera de guía, los mejores algoritmos a aplicar dependiendo del problema a resolver y el tipo de datos que se está utilizando (Cuadro 1.3.1).



Cuadro 1.3.1

A continuación se presenta una descripción más completa de cada tipo de problema. Antes de definir los tipos de problemas, se define el conjunto de objetos con los que se va a trabajar:

Sea E = Conjunto de todos los elementos posibles de entrada
Las instancias dentro de E = Conjunto de valores para una serie de atributos:

$$E = A_1 * A_2 * \ldots * A_n$$

Se define un elemento ***e*** como una n-tupla donde,

$$e = (a_1, a_2, \ldots, a_n) \quad \text{tal que } a_i \in A_i$$

1.3.1 Problema Descriptivo

Un problema descriptivo es aquel cuya meta es encontrar un entendimiento de los datos de estudio, por ejemplo, conocer cuáles son los clientes de una organización, o bien, encontrar los productos que se venden juntos de manera más frecuente. En ambos casos, lo que se pretende es encontrar grupos homogéneos pero con diferente enfoque. En el caso de los clientes de una organización, la meta es extraer características, en el segundo ejemplo, es asociar comportamientos.

Los elementos se presentan como un conjunto de datos sin etiquetar ni ordenar de ninguna manera.

$$\delta = \{e: e \in E\}$$

Donde:
δ = Conjunto de datos

El objetivo, por tanto, no es predecir nuevos datos sino describir los existentes. Los problemas descriptivos más comunes, son:

Análisis de segmentación, también llamados de aprendizaje no supervisado o clustering. El objetivo es encontrar grupos homogéneos entre los elementos de δ, de tal manera que los elementos asignados al grupo sean *similares*. Lo importante del agrupamiento respecto a la clasificación[1], es que son precisamente los grupos y la pertenencia a los grupos lo que se quiere determinar y a priori, no se sabe ni cómo son los grupos que se desea obtener. Otras veces se puede proporcionar el número de grupos que se desea obtener y, en otras ocasiones, este número se determina por el algoritmo de agrupamiento según las características de los datos.

El objetivo es obtener una función:

[1] Clasificación: Se describe mas adelante, en la seccón 1.3.2

$$\lambda: E \rightarrow S,$$

Donde,
λ es una función denominada clasificadora, que represente la correspondencia existente en los elementos.
E = Conjunto de los datos de entrada
S = Conjunto de datos de salida

Correlaciones y factorizaciones: Detectan atributos redundantes o dependencias entre atributos, o seleccionar un subconjunto, también se considera como problema de minería de datos. Los estudios factoriales y correlacionales se centran exclusivamente en los atributos numéricos (ya sean exclusivamente numéricos o después de la numerización).

El objetivo es, dados los datos del conjunto: $E=A_1 \times A_2 \times \ldots \times A_n$, determinar si dos o más atributos numéricos A_i y Aj A_j, están correlacionados linealmente o relacionados de algún otro modo. Este tipo de relaciones son bidireccionales o no orientadas.

Análisis de asociaciones: Es una de las técnicas más significativas en la minería de datos y que ha evolucionado con la misma minería de datos. El objetivo es similar a los estudios correlacionales y factoriales, pero para atributos nominales, que son frecuentes en las bases de datos. También se conoce con el nombre de análisis de vínculos.
Dados los datos del conjunto $E=A_1 \times A_2 \times \ldots \times A_n$, una regla de asociación se define generalmente de la siguiente forma:

$$\text{“Si } A_i = a \wedge A_j = b \wedge \ldots \wedge A_k = h \rightarrow A_r = u \wedge A_s = v \wedge A_z = w\text{”}$$

Donde todos los atributos son nominales y las igualdades se definen utilizando algún valor de los posibles para cada atributo.

La regla anterior está orientada, es decir, es una regla de asociación direccional. Por este motivo las reglas de asociación orientadas también se llaman dependencias de valor. También existen reglas de asociación bidireccionales en las que en lugar de tener un implicación se tiene una co-implicación, por ejemplo, si se tiene una regla de “Si, Compra_*Aguacate = sí* Λ *Compra*_Cebolla = Sí Λ Compra_Limones = sí”, se tiene una regla de asociación direccional u orientada y no puede ser entendida en orden inverso, es decir, si compra limones, no se puede concluir nada. En cambio si se tiene una regla “Si, Compra_Hamburguesa, sucede conjuntamente con Compra_Catsup”, esta regla se entiende en ambos sentidos, es decir, si compra cátsup, suele comprar hamburguesa y si se compra hamburguesa se suele comprar cátsup.

Dependencias funcionales: Consideran todos los valores posibles (a diferencia de las asociaciones o dependencias de valor). Se definen: “Dados los valores de A_i, A_j, A_k,

puedo determinar el valor de A_r". Es decir, el valor de A_r depende o es función de los valores de los atributos A_i, A_j, A_k. Por ejemplo, una dependencia funcional sería "Dada la edad, el nivel de ingresos, el código postal, si está casado o no, puedo determinar con bastante fiabilidad si el cliente tiene auto". Las dependencias funcionales, en particular cuando hay un atributo a cada lado, pueden ser orientadas y no orientadas, al igual que las reglas de asociación.

Detección de valores e instancias anómalas: Se utiliza generalmente para dar limpieza a los datos, no obstante, la detección de valores anómalos pueden sugerir fraudes, fallos, intrusos o comportamientos diferenciados. Este método es más general pues considera todos los atributos, y el objetivo es encontrar aquellas instancias que no son similares a ninguna (o muy poco) de las otras instancias. La manera de abordar el problema es generalmente la de agrupar los datos y ver aquellas instancias que se quedan desplazadas de los grupos mayoritarios. Para ello, son especialmente útiles los agrupadores suaves o los estimadores de probabilidad de agrupamiento, pues si un dato tiene baja probabilidad de agrupamiento en todos los grupos, se puede considerar un caso aislado y por tanto anómalo. Existen otros métodos como la medición de distancias (aquellas instancias cuyo vecino más próximo esté muy lejos puede considerarse una instancia anómala).

1.3.2 Problema Predictivo

El objetivo, es obtener un modelo que en un futuro pueda ser aplicado para predecir el comportamiento. Se denominan también problemas de aprendizaje supervisado, ya que el analista proporciona la variable respuesta deseada.
La variable a predecir puede ser una clase, una variable categórica o una variable numérica ó un órden entre ellos.

Problemas de clasificación: Hacen referencia a problemas en los que la variable a predecir tiene un número finito de valores, es decir, se trata de una variable categórica. Por ejemplo, clasificación de clientes (Bueno, Regular, Malo).
Se presentan como un conjunto de pares de elementos de dos conjuntos,

$$\delta = \{(e,s): e \in E, s \in S\}$$

Donde,
E es el conjunto de valores de entrada
S es el conjunto de valores de salida
Los elementos ***e***, al ir acompañados de un valor de ***S***, se denominan elementos etiquetados (*e,s*) y δ se denomina conjunto de datos etiquetado.

El objetivo es encontrar una función λ: E → S denominada clasificadora que represente la correspondencia existente en los elementos, es decir, para cada valor en *E*, tenemos un único valor en *S*. Además *S* es nominal, puede tomar un conjunto de valores c_1, c_2,..., c_m

denominados clases (cuando el número de clases es 2, se tiene una clasificación binaria). La función encontrada será capaz de determinar la clase para cada nuevo elemento sin etiquetar, es decir, dará un valor *S* para cada valor *e*.

Problemas de clasificación suave: La presentación del problema es el mismo que el de clasificación, pares de elementos de dos conjuntos, $\delta = \{(e, s): e \in E, s \in S\}$.

Además de la función λ: E → S, se busca otra $\theta: E \to \mathbb{R}$, función que representa el grado de certeza en la predicción hecha por la función λ. Lógicamente, siempre es preferible tener un clasificador suave que acompañe las predicciones (aunque sea sólo una estimación), pues permite realizar muchas otras aplicaciones como es la selección de los mejores elementos.

Problemas de estimación de probabilidad de clasificación: Se trata en realidad de una generalización de la clasificación suave. La presentación del problema es la misma que la de clasificación normal y suave, pares de elementos de dos conjuntos: $\delta = \{(e, s): e \in E, s \in S\}$.
Sin embargo, la función a encontrar es diferente al de la clasificación y la de clasificación suave. Se trata de encontrar ***m*** funciones:

$$\theta_i: E \to \Re,$$

Donde **m** es el número de clases, es decir, cada función a encontrar, retorna para cada elemento m un valor real p_i. Cada uno de estos valores p_i se denomina probabilidad de la clase ***i*** y representa el grado de certeza de que un elemento sea de la clase *i*. Si además se cumple que:

$$\forall p_i: 0 \leq p_i \leq 1 \; y \; \sum p_i = 1$$

Entonces, estas p_i, representan la probabilidad de que un elemento sea de la clase *i*. El conjunto de funciones encontradas se llama estimador de probabilidad.

Problemas de categorización. No se trata de encontrar una función, sino una correspondencia, es decir, cada elemento de

$$\delta = \{(e, s): e \in E, s \in S\}$$

así como la correspondencia λ: E → *S*, pueden asignar como varias categorías en S a un mismo ***e***, a diferencia de la clasificación, que sólo se asigna una y solo una.

Por ejemplo, dados un grupo de clientes, qué tipos de productos puede comprar. La categorización también puede ir asociada a una categorización suave (cada categoría asignada va a asociada de su certeza) o en forma de un estimador de probabilidades (se estima una probabilidad para todas las categorías). En este caso, la suma de probabilidades puede ser mayor a 1. Si bien en δ se puede tener varias etiquetas para el mismo elemento, la función aprendida por un estimador de probabilidades de clasificación y un estimador de probabilidades de categorización, viene a ser prácticamente lo mismo y usarlo como

clasificador consiste en seleccionar la clase con mayor probabilidad y crear un categorizador que seleccione las probabilidades más altas de las k mejores categorías.

Preferencias o Priorización. El aprendizaje de preferencias consiste en determinar, a partir de dos o más elementos, un orden de preferencia.
Cada elemento $(e_1, e_2, e_3, \dots, e_n), e_k \in E, k \geq 1,$ es una secuencia donde el orden de la secuencia representa la predicción. Un conjunto de datos para este problema es, por lo tanto, un conjunto de secuencias:

$$\delta = \{(e_1, e_2, e_3, \dots, e_n): e_k \in E\}$$

,

Otra manera alternativa de representar los datos es mediante un orden parcial, que se puede presentar como un caso particular de la definición anterior, donde las secuencias tienen sólo dos elementos, i.e., $k = 2$. Una simplificación del problema sería obtener sólo preferencias entre dos elementos, es decir, el modelo sólo podría decidir sobre la preferencia sobre dos objetos, pero no ordenar más de dos objetos. Lo más característico de esta técnica, es la presentación de los datos, ya que con un estimador de probabilidades también presenta una priorización, aunque a lo que aquí se da preferencia, es a la clase. Por ejemplo, dados una serie de candidatos para un trabajo, dar un orden priorizado de los candidatos para cubrir el puesto.

Regresión: Es quizá el problema más sencillo de definir. El conjunto de evidencias son correspondencias entre dos conjuntos

$$\delta: E \rightarrow S,$$

Donde ***S*** es el conjunto de valores de salida.

Al igual que con la clasificación, los registros al ir acompañados de un valor de ***S***, se denominan comúnmente registros etiquetados, y δ es un conjunto de datos etiquetado.
El objetivo es obtener una función λ: E→S que represente la correspondencia existente entre los datos, es decir, para cada valor de E se tiene un único valor para S. La diferencia con la clasificación es que S es un valor entero o real.
Un modelo de regresión se puede convertir en un modelo de clasificación binario suave, si se establece un umbral ***u*** para la salida ***y*** de la función λ, es decir, si $y < u$ asignaremos una clase, y si $y \geq u$ tendremos otra clase. Este es el caso del método denominado regresión logística que en realidad se utiliza para la clasificación.

Problemas de predicción de valores: Se refiere a problemas donde la variable a predecir es numérica. Por ejemplo, definir la probabilidad de que un cliente pague determinado préstamo o no.

Como se puede observar, algunas metodologías están relacionadas. Un caso muy particular ocurre cuando se tiene un problema de clasificación entre más de dos clases, y se quiere resolver mediante clasificadores que sólo consigan clasificar entre dos clases. El problema de

adaptar más de dos clases a una clasificación binaria se denomina **binarización** y existen varios métodos para realizarla:

One vs All: Se construye un clasificador binario utilizando todos los datos de una clase y agrupando en una misma clase el resto de los datos. Esto se realiza para todas las clases. Si se tienen n clases, el resultado será n clasificadores binarios que posteriormente se han de combinar. Esta misma aproximación se utiliza para convertir problemas de categorización en problemas de clasificación, con la diferencia de que, en la categorización, no hay que fusionar después los resultados.
All-pairs: Se construye un clasificador binario utilizando todos los elementos de dos clases e ignorando el resto. Esto se realiza para todos los pares de clases. Si se tienen n clases, el resultado serán n(n-1)/2 clasificadores binarios que posteriormente se habrán de combinar.

All-halves: Se construye un clasificador binario, por un lado, utilizando los datos de la mitad de las clases y el resto por el otro.

A continuación se resume en el cuadro 1.3.2, de manera descriptiva más no limitativa las técnicas que pueden aplicar para resolver problemas de minería de datos:

Cuadro 1.3.2

Problema	Modelo	Metodología	Descripción
Clasificación	Predictivo	Árbol de decisión Análisis discriminante Regresión logística Algoritmo genético evolutivo Red neuronal Vecinos próximos	Estas técnicas, utilizan un conjunto de datos para entrenamiento para crear el modelo, con el cual posteriormente clasifican individuos desconocidos.
Segmentación de Bases de datos	Clustering no jerárquico	Clustering Análisis factorial Análisis de clases latentes Componentes principales	Estos modelos utilizan un conjunto de datos de entrenamiento que posteriormente utilizan para clasificar a los individuos desconocidos.
Estimación de valores	De aprendizaje supervisado	Árboles de decisión Regresión Logística Red Neuronal Regresión Lineal y logarítmica Regresión no lineal Algoritmo genétivo evolutivo	

Segmentación de Bases de datos	Clustering jerárquico	Clustering Análisis de conglomerados Red Neuronal	Técnica apropiada cuando no se conocen o no se tiene información suficiente de los grupos
Análisis de relaciones	Asociaciones	Árboles de decisión Algoritmos genéticos evolutivos Regresión logística A priori CN2 rules (cobertura)	Cuenta las ocurrencias de los elementos y crea un vector de valores de atributos que suceden más frecuentemente

En la siguiente sección se describe con mayor detalle las técnicas de árboles de decisión y regresión lineal, pues son las técnicas que se usarán en el ejemplo del capítulo 2.

1.3.2.1 Arbol de decisión

Es una técnica para el aprendizaje *de modelos comprensibles de decisión*, elaborados a partir de una muestra de datos disponibles.
De todos los métodos de aprendizaje, los sistemas basados en árboles de decisión son quizás los más fáciles de utilizar y entender. Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.

Una de las grandes ventajas de los árboles de decisión es que, en su forma más general, las opciones posibles a partir de una determinada condición son mutuamente excluyentes. Esto permite analizar una situación y, siguiendo el árbol de decisión apropiadamente, llegar a una sola acción o decisión a tomar. Otra ventaja de los árboles de decisión, es que, permiten tratar a los datos perdidos como categorías independientes dentro de cada variable.

1.3.2.1.1 Metodología de construcción de reglas del algoritmo del árbol

La primera parte del algoritmo se llama la búsqueda de división. La búsqueda de división comienza con la selección de una entrada para la partición de los datos de disponibles. Si la escala de medición de la entrada seleccionada es un intervalo, cada valor único sirve como un posible punto de división para los datos.
Si la entrada es categórica, el valor promedio o el objetivo se toman dentro de cada nivel de esta entrada categórica. Los promedios tienen el mismo papel que los valores de entrada de intervalo único.

Para una entrada seleccionada y un punto de división fija, se generan dos grupos, los casos con valores de entrada menor que el punto de división, se registran en la rama derecha. Los casos con valores de entrada mayor que el punto de división, se registran en la rama izquierda. Los grupos, junto con los resultados del objetivo, forman una tabla de contingencia 2X2, con columnas que especifican la rama (izquierda o derecha), y filas especificando el valor objetivo. (0 ó 1). El estadístico Pearson de la ji-cuadrada, se utiliza para cuantificar la independencia de los recuentos en las columnas de las tablas.

Los valores grandes del estadístico de la ji-cuadrada, sugieren que la proporción de ceros y unos en la rama izquierda es diferente que la proporción de la rama derecha. Una diferencia grande en la proporción de las ramas izquierda y derecha, significa una buena división. Debido a que el estadístico de la ji-cuadrada de Pearson se puede aplicar al caso de divisiones múltiples, y objetivos múltiples, la estadística se convierte en un valor de probabilidad o p-value. El indicador de p-value indica la probabilidad de obtener el valor observado del estadístico, suponiendo proporciones del objetivo idénticos en cada dirección de la rama. Para grandes conjuntos de datos, este p-value pueden estar muy cerca de cero, por esta razón, la calidad de la división de una rama es calculado como un logaritmo =-log (ji-cuadrado del p-value). Por defecto, esta función, debería al menos una vez superar el umbral para una rama que se produzca con esta entrada.

El algoritmo busca la partición que maximice el valor *–log (p-valor)* sujeto al límite del número de ramas y al límite del número de observaciones mínimas que pueden ser asignados a una rama. El valor depende del criterio de división, que se define mediante el criterio de medida de la reducción de la entropía, gini y varianza, y que puede ser expresado genéricamente de la siguiente manera:

I (Nodo) – Suma a lo largo de las b ramas de P (b) I (b)

Donde,
I (Nodo) = la medida de entropía, Gini o varianza en el nodo
P (b) = Proporción de observaciones en el nodo asignadas a la rama b
Si se especifican probabilidades a priori, I (b), entonces, las proporciones P (b), serán modificadas por este valor
Dependiendo de la definición tenemos:
Entropía:

$$I(nodo) = -\sum_{\substack{Todas \\ Clases}} P_{Clase} log 2 P_{clase}$$

Gini:

$$I(nodo) = \left[1 - \sum^{Número\_de\_casos\_con\_clase\_i}_{Número\_total\_casoso\_en\_nodo}\right]^2$$

Varianza:

$$I(Nodo) = \sum_{Obs\_i} (Y_i - \bar{Y})^2 \text{ siendo } \bar{Y} \text{ la media de } Y \text{ en el nodo}$$

Los test de Chi-cuadrado y F usan la medida del –log (p-valor). Para estos criterios, la mejor partición es la que maximiza el estadístico (presenta el p-valor más reducido). Estos p-valores son ajustados para evitar los sesgos por calcular múltiples pruebas. La búsqueda de la partición óptima requiere el cálculo de distintas tablas de contingencia.
Si usamos la tabla original sin cambios en las categorías, el test ji-cuadrado puede ser usado. Este test asume que solo existe una población de la que solo extraemos una única muestra y calculamos un único test. Sin embargo realizar el test de forma repetida, viola este supuesto. Ello aumenta la posibilidad de encontrar alguna relación simplemente por el hecho de incrementar el número de veces en la búsqueda lo que puede llevar a encontrar relaciones ficticias o a magnificar las relaciones encontradas.

El problema surge, por lo tanto, cuando la tabla de contingencia original es reducida. Ahora para comparar los p-valores (niveles α de significancia) de los distintos predictores, estos deberán ser modificados por un factor multiplicador de Bonferroni. Este proceso dará como resultado un nuevo error de tipo I (α) que resuelve el problema anterior.
Para ello, se tiene en cuenta el número en que las ***c*** categorías originales de cada tipo de variable explicativa pueden ser reducidas a ***r*** grupos ($I \leq \boldsymbol{r} \leq c$) y se usa este resultado para obtener el nuevo nivel de significancia. Los multiplicadores pueden calcularse de dos formas, asumiendo a que la variable explicativa sea ordinal o nominal.

En el caso de variables ordinales, solo pueden ser definidas categorías continuas. Así la corrección de Bonferroni se deriva fácilmente al ser un coeficiente binomial:

$$B_{ordinal} = \begin{bmatrix} c-1 \\ r-1 \end{bmatrix}$$

En el caso de variables ordinales que presentan valores perdidos, éstos quedarían comprendidos en una nueva categoría *perdida.* Excepto para esta categoría perdida de nuevo, sólo podrán ser agrupadas categorías continuas como en el caso anterior. La categoría perdida puede quedar separada de las demás o ser combinada con cualquier otra categoría o grupo de categorías. El multiplicador de Bonferroni sería por lo tanto una simple extensión del caso anterior:

$$B_{ord\_Perdidos} = \begin{bmatrix} c-1 \\ r-1 \end{bmatrix} + r \begin{bmatrix} c-2 \\ r-1 \end{bmatrix} = \frac{r-q+r(c-r)}{c-1} B_{odinal}$$

Para las variables categóricas, cualquier agrupación de categorías es permisible, en este caso el multiplicador propuesto sería:

$$B_{categórica} = \sum_{i=0}^{r-1} (-1)^i \frac{(r-1)^c}{i!\,(r-i)!}$$

1.3.2.2 Modelo de regresión

La regresión ofrece un enfoque diferente para la predicción en comparación con los árboles de decisión. La regresión, como modelo paramétrico asume una estructura de asociación específica entre las entradas y objetivo. Por el contrario, los árboles, como los algoritmos predictivos, no asumen ninguna estructura de asociación; simplemente buscan aislar las concentraciones de los casos al igual que con valores de evaluación de los objetivos.
La modelización estadística consiste en explicar el comportamiento de una variable a partir del conocimiento de otras. Subyacente al concepto de modelización, está la idea de que una variable tiene una cierta variabilidad y que esta variabilidad está relacionada con el comportamiento de otras variables.
La modelización estadística es seguramente la técnica estadística más empleada. En minería de datos una parte importante de los problemas es de predicción, de forma que, basta que las variables explicativas estén asociadas a la variable de respuesta, para que, sabiendo el valor que toman aquellas, podamos hacer predicciones sobre el valor que tomará la variable objetivo.
Por otro lado, en regresión, es normal suponer a las variables explicativas fijadas por un diseño experimental, es decir, no aleatorias. Este no es el caso de la minería de datos, donde las variables explicativas se obtienen en contextos observacionales, por ejemplo una base de datos. Esto no es un problema, dado que en regresión se estudia el comportamiento de la respuesta condicional a las variables explicativas.
La modelización estadística consiste pues en descomponer los valores que toma la variable *y* para cada individuo en dos componentes, uno, función de las variables explicativas y otro que es específico del individuo en cuestión:

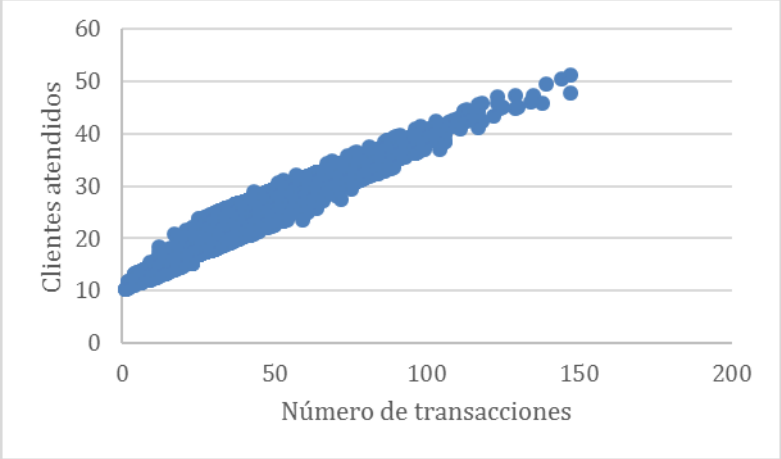
$$y_i = r(X_{i1}, \dots, x_{p1}) + \varepsilon_i,$$

Donde ***r*** es la función que relaciona los valores de la variable de respuesta con las explicativas, mientras que $\boldsymbol{\varepsilon_i}$ es la parte específica del individuo ***i*** que no se explica por ninguna variable.
La función ***r*** representa la parte determinista, estructural del modelo, que explica el comportamiento de la variable de respuesta y hacer predicciones sobre ella, mientras que la parte específica representa la parte impredecible, aleatoria ***y*** se denomina término de error. Este término se caracteriza a partir de una distribución de probabilidad generadora de los distintos valores para cada individuo, que sin pérdida de generalidad, se espera que su valor sea 0:

$$E[\varepsilon_i] = E[y_i - r(x_{i1}, \dots, x_{1p})] = 0$$

Los modelos estadísticos se diferencian respecto al tipo de la variable de respuesta, que puede ser numérica, binaria o categórica. El tipo de variables explicativas que pueden ser numéricas o categóricas, respecto a la función *r* y respecto a la distribución de probabilidad del término del error.

Si solo disponemos de una variable explicativa hablamos de regresión simple, mientras que si disponemos de varias variables explicativas se trata de un problema de regresión múltiple.
Para visualizar la relación entre la variable de respuesta y una variable explicativa, obtendremos el diagrama bivariante entre ambas variables. La forma de dicho diagrama aporta información sobre el tipo de relación que guardan estas variables, esto es sobre la función ***r***. Por ejemplo en la gráfica 1.3.2.2.1, se observa una relación lineal.



Gráfica 1.3.2.2.1

Para estimar dicha función ***r***, buscaremos una función tal que, en promedio, las desviaciones al cuadrado respecto de los puntos sean mínimas. Esto es, minimizaremos el error cuadrático medio:

$$\min_r E\left[(y_i - r(x_{i1}, \dots, x_{ip})^2\right]$$

El mínimo de esta función corresponde a la media condicional de y_i respecto de los valores de las variables explicativas.

$$r(x_{i1}, \dots, x_{ip}) = E(y_i | x_{i1}, \dots, x_{ip})$$

A esta función se le llama función de regresión.
La función de regresión más simple es la lineal, esto es, cada variable explicativa participa de forma aditiva y constante para todo el dominio observado en la formación de la variable respuesta:

$$E(y_i | x_{i1}, \dots, x_{ip}) = \beta_0 + \beta_1 x_{i1} +, \dots, + \beta_p x_{ip}$$

Por tanto, el modelo de regresión lineal se escribe:

$$y_i = \beta_0 + \beta_1 x_{i1} +, \dots, + \beta_p x_{ip} + \varepsilon_i$$

La suposición de linealidad puede parecer muy restrictiva pero en realidad no lo es tanto. Si el número de individuos con los que se cuenta es pequeño, los modelos lineales muestran una capacidad de generalización mayor que otros modelos más sofisticados y flexibles, debido a la excesiva dependencia de estos últimos respecto de los datos utilizados. Por otro lado, las variables explicativas utilizadas pueden ser transformaciones de las originales, lo que amplía la versatilidad del modelo lineal. Por ejemplo, pueden ser transformaciones logarítmicas o de raíz cuadrada para normalizar las variables originales, o pueden ser transformaciones polinómicas (x, x^2, $x^{3)}$, para ajustar una función ***r*** no lineal.

También, es posible tomar como variables explicativas, las variables binarias fruto de la recodificación en intervalos de las variables originales (binning) y aproximar relaciones no lineales cualesquiera, mediante funciones en escalera, o bien, podemos incluir en el modelo términos de interacción entre variables explicativas, mediante el respectivo producto (X_2*X_3), ampliando la flexibilidad del método a relaciones no aditivas La linealidad del modelo se refiere a la forma de la función de regresión, independientemente de si las variables explicativas son una transformación no lineal del espacio original.

Por otro lado, existen modelos no lineales de amplio uso que se convierten en modelos lineales con simples transformaciones. Por ejemplo, el modelo econométrico de elasticidad constante o doble logarítmico (representado en la parte izquierda de la gráfica 1.3.2.2.2) y = αx^{β}, siendo β el coeficiente de elasticidad, se convierte en un modelo lineal tomando logaritmos de ambas variables: $\dot{y} = ln(y)$, $\dot{x} = \ln(x)$, obteniendo $\dot{y} = \ln(\alpha) + \beta\dot{x}$. Este modelo expresa que una variación porcentual de la variable explicativa produce una variación porcentual constante en la variable de respuesta.

El modelo de crecimiento exponencial y=α*exp {βx} (Representado en la parte central de la gráfica 1.3.2.2.2) se vuelve lineal tomando logaritmos de la variable de respuesta ẏ= In (y), resultando ẏ=In (α)+βx. En este modelo, variaciones en términos absolutos de la variable explicativa producen una variación porcentual constante en la variable de respuesta.
El modelo logístico, que indica la tasa de variación entre cero y uno, de una respuesta (representado en la parte derecha de la gráfica 1.3.2.2.2), definida por:

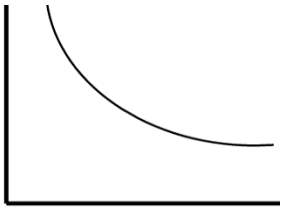
$$y = \frac{exp\{\alpha + \beta x\}}{1 + exp\{\alpha + \beta x\}}$$

Se linealiza mediante la transformación:

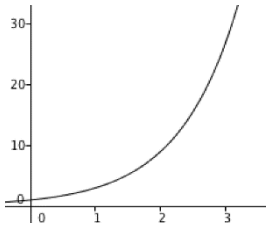
$$\dot{y} = \ln \frac{y}{1 - y}$$

Obteniendo entonces $\dot{y} = \alpha + \beta x$.

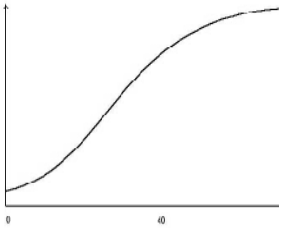
El modelo logístico se aplica en situaciones en que el modelo de crecimiento exponencial indefinido no es adecuado por llegarse a situaciones de saturación en el crecimiento.



Elasticidad constante



Crecimiento exponencial

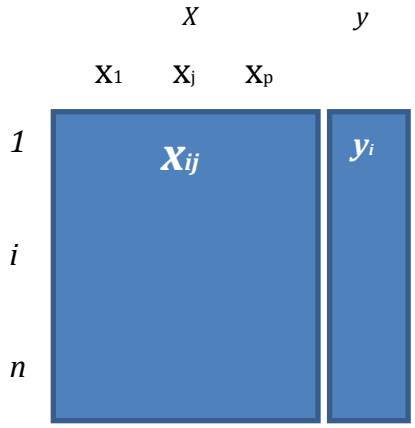


Crecimiento logístico

Cuadro 1.3.2.2.2

1.3.2.2.1 Estimación de la función de regresión lineal

Estimar la función de regresión lineal, significa estimar los coeficientes β_j. Para ello, dispondremos de una muestra de aprendizaje de valores, de la variable de respuesta con sus correspondientes variables explicativas X (Cuadro 1.3.2.2.1)



Cuadro 1.3.2.2.1.1

Para estimar dichos coeficientes, usamos el mismo criterio utilizado para definir la función de regresión, aplicado ahora a los datos muestrales:

$$\min_{b_0,\dots,b_p} \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} \right)^2 = \sum_{i=1}^{n} e_i^2$$

Llamamos residuos a los errores calculados con los datos muestrales. Por lo tanto el criterio de estimación es la minimización de la suma del cuadrado de los residuos (SCR). Esta minimización tiene una interpretación geométrica simple. Llamemos $\hat{y}_i$, al valor ajustado:

$$\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}$$

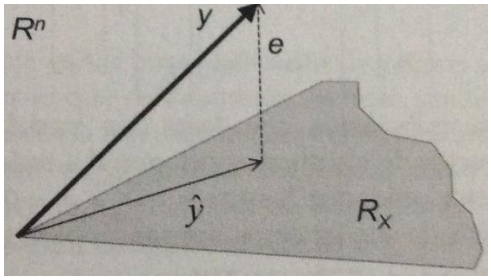
Por lo tanto, el ajuste mínimo cuadrático, se escribe:

$$y_i = \hat{y}_i + e_i \quad i = 1, \ldots, n$$

El cual expresado en notación matricial, queda:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

Donde ***y*** es el vector cuyos términos son los valores observados de la variable respuesta, $\hat{y}$ es el vector de valores ajustados y ***e***, el vector de residuos.
Esto es, el vector ***y*** se descompone en suma de dos términos, uno que procuramos sea un vector lo más parecido posible a ***y***, y que necesariamente debe estar contenido en el espacio generado por las variables explicativas (por ser combinación lineal de éstas) y otro vector de residuos, diferencia entre los dos anteriores. Claramente, hacer mínima la cantidad SCR, es hacer mínima la longitud del vector residual. Y la forma de obtener un vector de residuos mínimo es mediante la proyección ortogonal de la variable de respuesta sobre el espacio generado por las variables explicativas (Gráfica 1.3.2.2.1.2).



Gráfica 1.3.2.2.1.2

Por lo tanto, el vector ajustado debe pertenecer al espacio ***Rx*** generado por las variables independientes $\hat{y} = Xb$, y debe ser ortogonal al vector de residuos $\langle \hat{y}, y - \hat{y} \rangle = 0$.

Desarrollando esta expresión, se llega a que $b'X'y = b'X'Xb$, y por lo tanto suponiendo **X** de rango completo:

$$b=(X'X)^{-1}X'y$$

Así, el vector ajustado es:

$$\hat{y} = Xb = X(X'X)^{-1}X'y = Hy$$

Donde H es la matriz de proyección ortogonal de ***y*** sobre el espacio de $\boldsymbol{R_x}$ generado por las variables $\boldsymbol{x_j}$.
Análogamente, podemos ver que el vector de residuos ***e***, también se obtiene por proyección ortogonal de la variable de respuesta ***y***:

$$e = y - \hat{y} = (1 - H)y,$$

Siendo ***I*** la matriz de identidad, e ***I-H*** el operador de proyección sobre el espacio $R_x^{\perp}$, ortogonal al generado por los vectores x_j.
Puede verse que la estimación mínima cuadrática es un estimador insesgado de la función de regresión: $E[b] = (X'X)^{-1}X'E[y] = (X'X)^{-1}X'X\beta = \beta$ y por lo tanto,

$$E[\hat{y}] = X\beta = E\left(y|x_1, \dots, x_p\right)$$

Respecto de la varianza de los estimadores se tiene el siguiente resultado:

$$b = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon,$$

$$var(b) = E[(b-\beta)(b-\beta)'] = E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}] = \sigma^2(X'X)^{-1}$$

Los coeficientes $\boldsymbol{b_j}$, cumplen la propiedad de ser óptimos (no sesgados y de mínima varianza) dentro de los estimadores lineales de los valores observados $\boldsymbol{y_i}$.

1.4 Interpretación del modelo

Medir la calidad de los patrones encontrados por un algoritmo de minería de datos no es un problema trivial, ya que esta medida puede tener varios criterios, algunos bastante subjetivos. Idealmente los patrones descubiertos deben tener tres cualidades:

- Ser precisos
- Comprensibles
- E interesantes (es decir novedosos)

Según las aplicaciones puede interesar mejorar algún criterio y sacrificar ligeramente otro.

Para entrenar y probar un modelo se parten los datos en dos conjuntos: El conjunto de entrenamiento y el conjunto de prueba o test. Esta separación es necesaria para garantizar que la validación de la precisión del modelo es una medida independiente, de lo contrario, la precisión será sobreestimada teniendo estimaciones muy optimistas.

La precisión es una buena estimación de cómo se comportará el modelo para datos futuros similares a los de prueba, esta técnica no garantiza que el modelo sea correcto, sino que simplemente indica que si se usa la misma técnica con una base de datos similar a la de prueba, la precisión media será bastante parecida a la obtenida con éstos.

La validación simple reserva un porcentaje de la base como conjunto de prueba, este porcentaje varía desde el 5% al 50% y no debe usarse durante el entrenamiento, y la selección debe de ser aleatoria.

Medidas de evaluación de modelos: Dependiendo de la tarea de minería de datos, existen diferentes medidas de evaluación de los modelos, como se describe en el siguiente cuadro (Cuadro 1.4.1):

Cuadro 1.4.1	
Modelo	**Descripción**
Clasificación	Se evalúa la calidad de los patrones encontrados con respecto a su precisión predictiva. La cual se calcula como el número de instancias del conjunto de prueba dividido por el número de instancias totales en el conjunto de prueba.
Reglas de asociación	Se suele evaluar de forma separada cada una de las reglas con objeto de restringirnos a aquellas que puedan aplicarse a un mayor número de instancias y que tienen una precisión relativamente alta. Esto se hace en base a dos conceptos: Cobertura: Número de instancias a las que la regla predice correctamente Confianza: Porción de instancias que la regla predice correctamente, es decir la cobertura dividida por el número de instancias a las que puede aplicar la regla.
Regresión	La manera más habitual de evaluar este tipo de modelos es mediante el error cuadrático medio del valor predicho respecto al valor que se utiliza como validación. Esto promedia los errores y tiene más en cuenta aquellos errores que se desvían más del valor predicho (ponderación cuadrática). Aunque pueden usarse otras medidas de error, ésta es la más utilizada.
Agrupamiento	Las medidas de evaluación dependen del modelo utilizado, aunque suelen ser función de la cohesión de cada grupo y de la separación entre grupos, éstos se pueden formalizar utilizando la distancia media del centro al grupo de los miembros del grupo y entre grupos.

1.5 Interpretación y contextualización

En muchos casos, es necesario evaluar el contexto donde el modelo se va a utilizar. Por ejemplo, en el caso de la clasificación y las reglas de asociación, utilizar la precisión como medida de calidad tiene sus desventajas, primero porque no se tiene en cuenta el problema de tener distribuciones de clases no balanceadas, es decir, tiene muchas instancias de ciertas clases y muy pocas de otras.

En algunas ocasiones los errores no valen por igual, o los costos de error son difíciles de estimar o incluso desconocidos. En ese caso se usan estrategias como la curva ROC (Receiver Operating Characteristic).

1.6 Difusión, uso y monitoreo

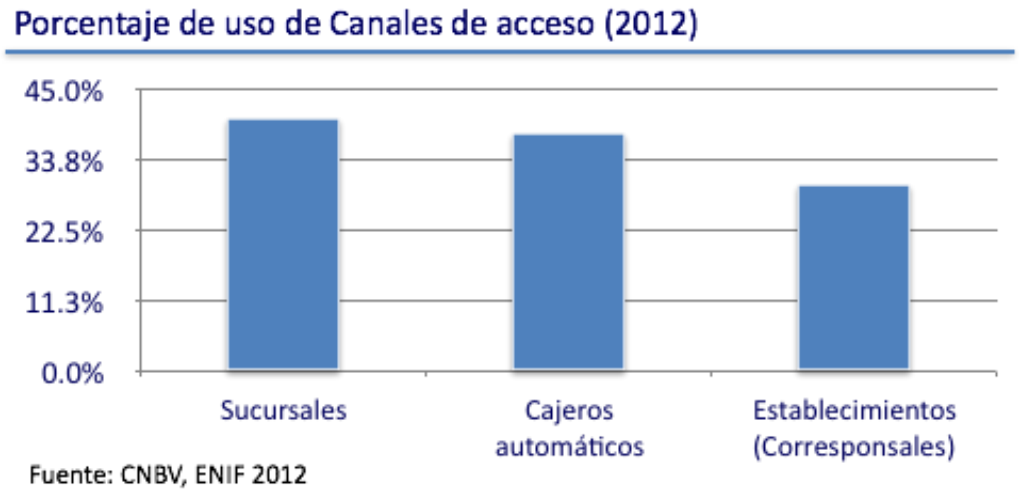
Una vez construido y evaluado el modelo, se suele usar principalmente con dos finalidades: Para que el analista recomiende acciones basadas en el modelo o bien para aplicar el modelo a diferentes conjuntos de datos. También puede incorporarse a otras aplicaciones, como por ejemplo a un sistema de análisis de créditos bancarios que asista al empleado al momento de evaluar a los solicitantes.

Es importante la difusión del modelo y que se distribuya y se comunique a los usuarios. El nuevo conocimiento obtenido debe de integrar el "know-how" de la organización.
También es fundamental medir lo bien que el modelo evoluciona y, aun cuando funcione de manera correcta, se debe de comprobar de manera continua el nivel de predicción del mismo, esto es principalmente porque los patrones de comportamiento suelen cambiar. Por lo tanto, el modelo debe de ser monitoreado, evaluado, re-entrenado y, tal vez, reconstruido nuevamente.

2. Ejemplo de aplicación a la red de sucursales Bancaria

2.1 Antecedentes

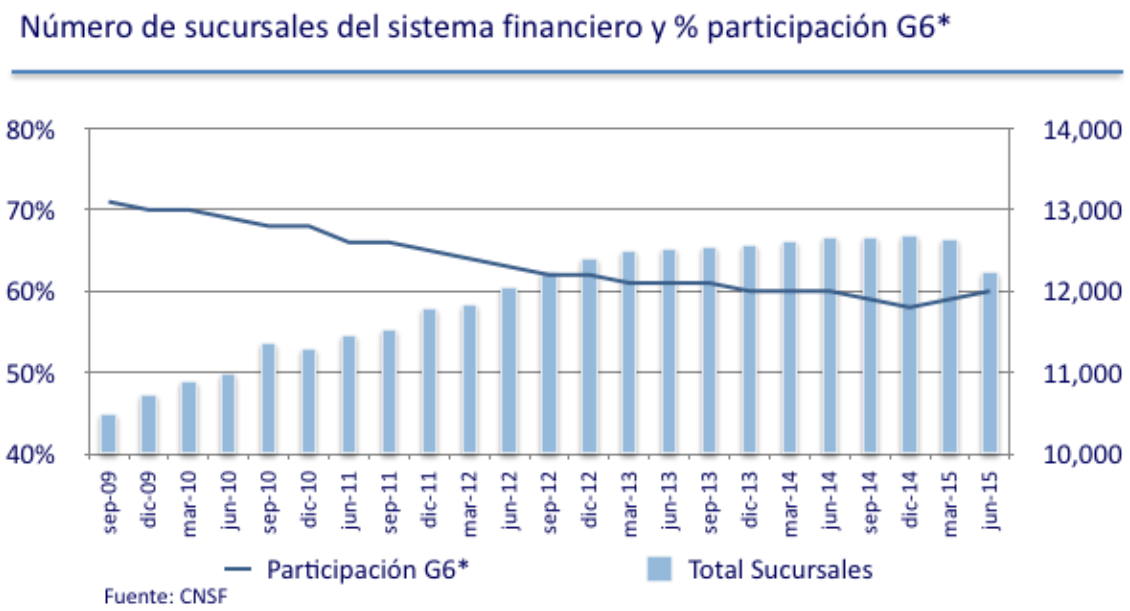
De acuerdo con la CNBV y Banco de México, cada vez existe mayor reconocimiento a nivel internacional sobre los beneficios sociales y económicos de contar con mayor acceso a servicios financieros.
Las sucursales bancarias, son uno de los principales puntos de acceso para ofrecer servicios y productos financieros. De acuerdo con los resultados de la ENIF[2] 2012, se identificó a las sucursales como el canal, que la mayoría de la población utiliza con el 40.5%. Seguido por los cajeros automáticos con el 38% (gráfica 2.1.1).



Gráfica 2.1.1

Al observar el comportamiento del número de sucursales, se observa un crecimiento continuo durante los últimos 9 años (gráfico 2.1.2). Sin embargo, las instituciones de banca múltiple han perdido participación en este canal de acceso, pues, resulta costoso establecer nuevos puntos de acceso para proveer sus servicios en todos los municipios. Solo dos instituciones han apostado al crecimiento de sucursales: Santander y Bancomer. En donde su estrategia es incrementar su participación de mercado, acercando a la población a tener contacto con un canal de acceso sólido.

[2] ENIF: Encuesta nacional de inclusión financiera, elaborada por el Consejo Nacional de Inclusión Financiera



Gráfica 2.1.2

[3]

Por otro lado es importante otorgar un nivel de servicio aceptable a los clientes que asisten a las sucursales, pues a pesar de los nuevos canales de acceso, las sucursales siguen siendo el canal principal para captación de nuevos clientes y colocación de servicios y productos.

La institución, donde se está realizando el presente trabajo, desarrolló un proyecto de expansión y actualización de la infraestructura de las sucursales, con el objetivo de captar más mercado y ampliar la ventaja competitiva existente entre sus competidores más cercanos. Alineado a la infraestructura, existe el proyecto de homogeneizar el modelo de servicio y perfiles de los empleados que trabajan en las sucursales.

Enseguida, se enlistan los cambios principales que afectan el servicio y personal que labora en la sucursal:

- Proyecto Ulises: Proyecto de modernización de las sucursales
  - Actualización de infraestructura y modernización de las sucursales (incluye cambios en el número de ventanillas).
  - Incorporación del equipo de asignación de fichas para dar servicio en la sucursal (Podios)
  - Actualización de autoservicio
  - Modernización y/o incremento de cajeros automáticos (ATM´s)
  - Incorporación y/o incremento de equipo para recibir depósitos (Practicajas)

[3] * G6: Los seis principales bancos del sistema bancario: Banamex, BBVA Bancomer, Banorte-Ixe, HSBC, Santander y Scotiabank. 85% de las sucursales.

- Proyecto de modelo de servicio: Diseño de protocolo de servicio, estableciendo un empleado que da la bienvenida y orienta al cliente, de acuerdo al trámite o servicio que desea.
- Cajero universal: Consolidación de los puestos de cajero en sucursal, dejando solo 3 perfiles:
    - Cajero universal A: Es el cajero con mayor experiencia ó cajero Sr. Tiene actividades principalmente administrativas, de control de flujo de efectivo y reportes centrales, así como dotar y mantener el autoservicio en funcionamiento, además de apoyar en ventanilla, dando servicio al cliente.
    - Cajero universal B: Cajero Jr. Proporciona servicio al cliente en ventanilla. También tiene la tarea de orientar al cliente al ingresar a la sucursal y otorgar fichas para asignar los turnos de atención, en caso de existir equipo.
    - Cajero part time: Tiene las mismas actividades que el cajero universal B, pero solo labora la mitad de los días del mes.
- Gerente Administrativo: Se encarga de coordinar las actividades de todos los cajeros de manera que se cumplan todas las actividades administrativas. Coordina el número de empleados que dan servicio en ventanilla.

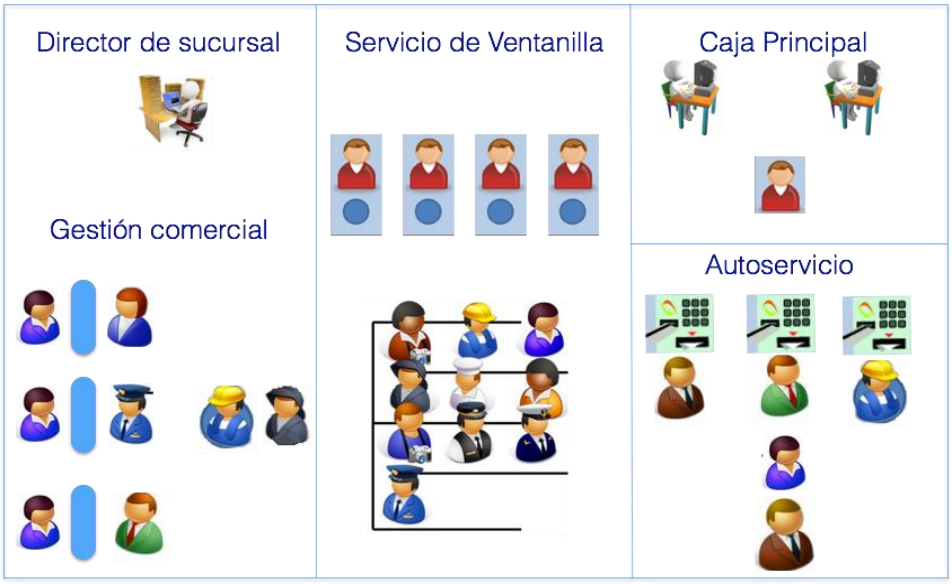
2.2 Alcance del proyecto

La organización de una sucursal tradicional, se divide en 3 grandes actividades (cuadro 2.2.1):

- **Gestión comercial:** Donde se atiende a clientes nuevos, asesoría acerca de productos y servicios del banco, es atendido por los ejecutivos.
- **Servicio de Caja principal ó actividades Back:** Desarrolla actividades para apoyar al servicio de ventanilla en flujo de efectivo, conteo de valores, reportes a áreas centrales, mantenimiento del equipo de autoservicio, entre otras actividades.
- **Servicio de ventanilla ó actividades Front:** Relación directa con el cliente, otorgando depósitos, retiros, pagos de servicios, cambio de cheques, etc.
- **Autoservicio:** No existe una relación directa con el cliente, el servicio es a través del cajero automático o las llamadas practicajas.
    - Cajero Automático (ATM), se utiliza para retiros de efectivo, pagos de servicio.
    - Practicajas: Depósitos, retiros y pagos de servicio.

Las primeras tres áreas se evalúan bajo un nivel de servicio, sin embargo, la gestión comercial no se considera una actividad operativa, pues la diversidad de operaciones y necesidades, es prácticamente a nivel cliente, siendo su labor básica de venta y captación.
El servicio de ventanilla y autoservicio se consideran actividades operativas, pues se limitan a cierto tipo de transacciones[4] y siempre se tienen que ejecutar de la misma manera.

[4] Cada solicitud que el cliente solicita o realiza: Retiro de efectivo, pago de cheques, pago de servicios, consulta de saldo, etc.



Cuadro 2.2.1

El servicio de ventanilla en una sucursal no sólo está sujeto al número de ventanillas disponibles, el número de dispositivos impacta directamente en la calidad del servicio de la sucursal (por ejemplo, si no tiene o no son suficientes los ATM's, en una sucursal), el cliente optará por usar el servicio de ventanilla. Por otro lado, el cliente valora que el autoservicio también otorgue cierto nivel de servicio. Sin embargo, el alcance del presente proyecto se acotará a dar un nivel de servicio óptimo en ventanillas de las sucursales (Cuadro 2.2.2) y mantener la disponibilidad del autoservicio las 24 horas del día los 365 días del año.

Cuadro 2.2.2

2.3 Objetivo

El objetivo del presente trabajo, es calcular la plantilla que necesitan las sucursales para otorgar un nivel de servicio óptimo en ventanilla y mantener la disponibilidad del autoservicio las 24 horas del día los 365 días del año, este cálculo es conocido como balanceo de sucursales o plantilla ideal.

El proceso de Balanceo en la red de tiene como objetivo mantener, de manera equilibrada, el nivel de servicio en la ventanilla y el personal necesario para otorgar dicho nivel de servicio. Para ello, se diseñó un proceso que mide las cargas de trabajo en el Back y las cargas de trabajo en el Front y de esta manera, determinar la plantilla óptima para hacer frente a esas cargas de trabajo. También se incluye el nivel de servicio ideal de manera que equilibre los costos de oportunidad y costos de servicio.

Se estudiarán la información y eventos que determinan el nivel de servicio que es necesario durante periodos de media hora y durante 5 meses de historia, se incluirá el dinamismo del nuevo perfil de Cajero Universal[5] A y B para realizar actividades en ventanilla y direccionamiento[6].

El modelo y/o proceso final arrojará una propuesta de plantilla de la sucursal que incluye:

- Número de ventanillas en sucursal.
- Número de personas y su perfil para el mantenimiento óptimo del autoservicio.
- Número de personas y su perfil para el mantenimiento óptimo en la caja principal.
- Número de personas y su perfil para el adecuado servicio en ventanilla.
- Número de personas necesariasy su perfil para cubrir todas las actividades en sucursal de direccionamiento.

2.4 Planteamiento del objetivo

El modelo de plantilla ideal de las sucursales dado un nivel de servicio, es sumamente complejo dadas las variaciones que existen en los días y horarios llamados pico y no pico, pues no se tienen los niveles de servicio en esos horarios y días para todas las sucursales, esto sin contar que cada sucursal es diferente en tamaño, cantidad y tipología de clientes. También el tamaño del autoservicio es diferente en cada sucursal y ayuda a liberar cargas de trabajo en ventanilla. A lo anterior se debe de agregar las cargas de trabajo administrativas y

[5]Cajero universal: Nombre del puesto que llevan las personas que se hacen cargo de las actividades administrativas (Cajero A), servicio a clientes en ventanilla y asesoría a de clientes (direccionamiento). Los Cajeros B, sólo realizan actividades en ventanilla y direccionamiento.

[6] Direccionamiento: Actividad que realiza el personal al orientar al cliente una vez entrado a la sucursal. Y otorga ficha para la gestión de filas.

operativas que no otorgan servicio al cliente (llamado back), como son las actividades para mantener el equipo del autoservicio y caja principal disponible.

Para poder librar estos inconvenientes, se decidió separar el proceso en varias etapas, (Cuadro 2.4.1). En donde se dividirán las actividades Back, de la actividad de ventanilla (front), aplicando una metodología diferente para cada estas dos grandes actividades.

Cuadro: 2.4.1

- **Actividades Back[7]:**
  - **A**ctividades de autoservicio: Total de tareas que se deben de realizar y tiempo promedio de cada actividad.
  - Actividades en caja principal: total de tareas que se deben de realizar y tiempo promedio de cada actividad.

Se calcula el FTE[6] sumando los tiempos promedio de las actividades.

- **Actividades Front:**
  - Actividades de ventanilla para otorgar un nivel de servicio óptimo:

Modelo de minería de datos.

- **Cálculo de plantilla ideal:**
  - Consolidando los dos puntos anteriores y considerando los perfiles que deberán de cubrirse.

[7] Las actividades y los tiempos promedio se realizan mediante un levantamiento en sitio, durante 15 días consecutivos. Las sucursales a medir.

[6] FTE: Full time equivalent, equivalente a jornada completa de trabajo o personas con jornadas completas de trabajo.

2.5 Beneficio Esperado

El modelo de Balanceo permitirá a las sucursales tener el personal necesario para cubrir todas sus actividades, a su vez, también optimiza el costo del personal, pues maximiza las capacidades del Cajero Universal, debido al intercambio de actividades en el back y front. De la misma manera, al mantener un nivel de servicio óptimo, se mejorará la relación cliente-sucursal.
Asimismo, se podrán detectar las sucursales cuyo nivel de servicio estén por debajo del deseable y/o estén en deterioro y proveer una propuesta de mejora de manera proactiva.

Las consecuencias de no realizar este proyecto se verán reflejadas en el nivel de servicio, costos por remodelación de sucursales y/o plantilla insuficiente para los casos en los que las sucursales manejan una gran transaccionalidad y mucho flujo de clientes. También, en el caso contrario, donde la sucursal mantiene una plantilla sobrada, la institución incurrirá en costos innecesarios de personal.

2.6 Actividades Back

2.6.1. Clasificación de actividades:

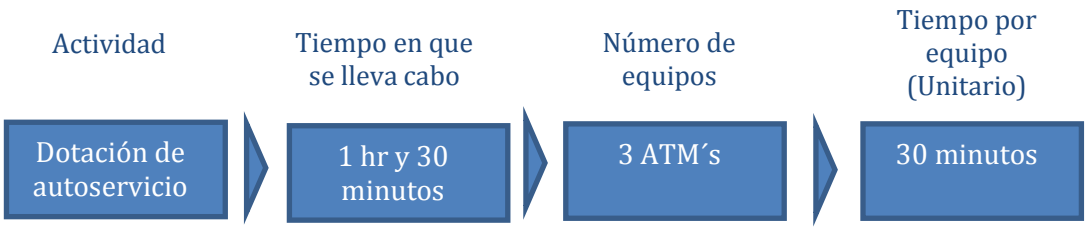
Como se observa en el cuadro 2.4.1, las actividades del back de la sucursal se divide en dos áreas principales, las que se derivan del autoservicio y las que se derivan de la Caja Principal.

Ambas áreas tienen actividades administrativas y operativas y se atacarán con el mismo procedimiento, sin embargo, se considerarán por separado pues, si bien son actividades que se pueden ejecutar de manera secuencial, hay actividades en ambos grupos que son de suma importancia y el ejecutarse fuera de tiempo podría implicar impactos en el nivel de servicio.

Actualmente la institución cuenta con información operativa y administrativa, para algunas sucursales, de las actividades del autoservicio y caja principal, así como tiempos de ejecución de estas actividades. Como anteriormente se comentó, las actividades reflejadas y tiempos de servicio están relacionados con el número de dispositivos y/o tamaño de la sucursal.

Para ello se clasificaron las actividades en unitarias y actividades relacionadas con equipamiento de la sucursal. A estas mismas actividades se asigna la frecuencia con que se deben de ejecutar a la semana.

Para evitar sesgos en las actividades que se derivan del equipamiento de las sucursales, se obtuvo el equipamiento de cada sucursal y se calcularon tiempos unitarios.
Por ejemplo, en la actividad de “Dotación de autoservicio” (Cuadro 2.6.1.1):



Cuadro 2.6.1.1

2.6.2. Definición del agenda y cálculo del personal necesario

Una vez teniendo el catálogo de actividades y la frecuencia, se elabora una distribución de actividades, donde el objetivo es equilibrar actividades durante la semana que cumplan con la frecuencia y los tiempos de jornadas diarios de cada cajero (Cuadro 2.6.2.1):

	Tiempos Autoservicio					Tiempos Caja Principal				
Autoservicio	L	M	M	J	V	L	M	M	J	V
Impulso Comercial	1.0	1.0	1.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0
Admin. Oficina	18.0	9.0	18.0	9.0	18.0	20.0	22.0	20.0	19.0	22.0
Actividad Gestión	0.0	0.0	0.0	0.0	0.0	2.0	1.6	2.0	1.6	1.6
Servicing	1.0	0.6	0.6	0.6	0.6	2.0	2.0	2.0	2.0	2.0
Total	20.0	10.6	19.6	10.6	19.6	26.0	27.6	26.0	24.6	27.6

* Tiempos en horas

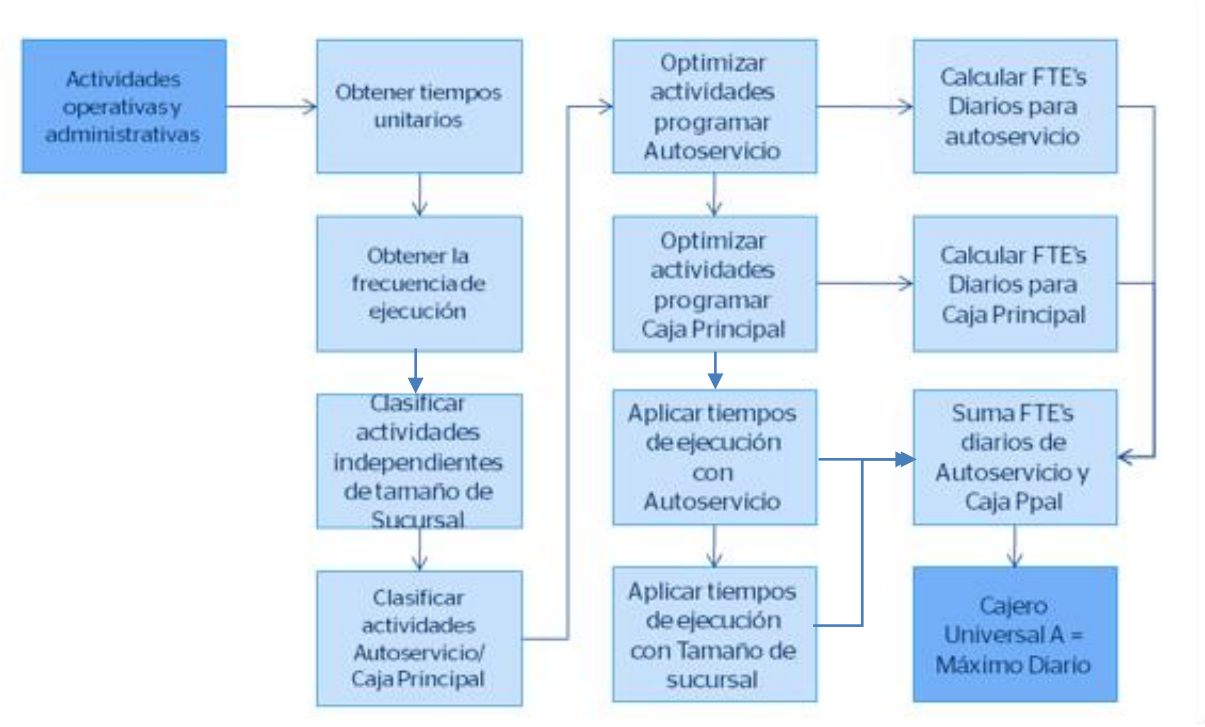
Cuadro 2.6.2.1

Una vez teniendo los tiempos unitarios, se asigna a cada sucursal y se individualiza de acuerdo a las características de cada sucursal.
El resultado del proceso anterior, nos indica cuántos minutos son necesarios diarios y a la semana, para hacer frente a las actividades en ambas áreas de la sucursal (Autoservicio y Caja principal) y se dividen entre las horas trabajadas de manera diaria de cada cajero Universal (sin incluir el descanso).

Posteriormente, se suman los requerimientos de ambas áreas y se obtienen las personas (FTE[8] de 40 horas de actividad diaria).

[8]FTE: Full time equivalent, equivalente a jornada completa de trabajo o personas con jornadas completas de trabajo.

Una vez teniendo la actividad diaria, se obtiene el máximo de la semana, siendo el resultado de número de cajeros Universales A. En el cuadro 2.6.2.2, se ilustra este procedimiento:



Cuadro 2.6.2.2

2.6.3. Resultados

Los resultados de la propuesta del cajero Universal A son los que se muestran en el cuadro 2.6.3.1, donde las divisiones Metropolitanas, son los que deben de ajustar a la baja, y la división Norte requiere solo de 4 cajeros más.

Cuadro 2.6.3.1

División	Sucursales	Cajero A Actual	Cajero A propuesto	Diferencial
Bajío	210	444	461	-17
Metro I	223	471	526	-55
Metro II	220	490	530	-40
Noreste	182	376	372	4
Noroeste	221	442	450	-8
Occ. I	111	237	249	-12
Occ II	192	384	407	-23
Sur	207	442	461	-19
Sureste	154	328	342	-14
Total	1,720	3,614	3,798	-184

2.7 Actividades front

Muchos clientes al visitar una sucursal bancaria, se han encontrado con grandes filas, con dos o tres personas atendiendo y muchas ventanillas vacías, lo que provoca gran descontento y una mala experiencia al cliente.
Los bancos que se consideren exitosos, tienen que aprovechar las visitas de los clientes a la sucursal para que esta, sea una experiencia satisfactoria. Siendo uno de los principales retos el otorgar un nivel de servicio satisfactorio, mejorando en lo más posibles los tiempos de espera del cliente y manteniendo los costos de personal más bajos posibles.

2.7.1 Identificar un problema relevante

En el año 2014, se realizó una actualización de los perfiles y actividades que realizan el personal que laboran en las sucursales, impactando en las capacidades y gestión de las mismas. Por otro lado, se está realizando una actualización en la infraestructura e imagen de todas las sucursales (Proyecto ULISES[9]), por ello es importante asegurar que este cambio no solo sea de imagen, también asegurar la mejora en los índices de calidad del servicio.

Para las actividades en ventanilla, el nivel de servicio será una variable decisoria para obtener comportamientos diferenciadores y características en la operativa y sucursales que determinen este nivel de servicio que se desea.

El objetivo es el número de ventanillas y con un análisis de minería de datos, se determinará las características y operativa que se necesita para establecer las ventanillas y el nivel de servicio que se ofrece de acuerdo al nivel de la transaccionalidad. Determinar el comportamiento de las sucursales, dado el nivel de servicio, es vital para un buen resultado del modelo.

La metodología para abordar el modelo de minería de datos, se constituye con las actividades que proponen las mejores prácticas en esta materia.

- **Identificar un problema relevante**
  - Análisis de la situación
  - Descripción del requerimiento del negocio y objetivos
  - Alcance del proyecto
  - Puntos clave que serán base de los modelos a realizar
- **Integración y recopilación de datos**
  - Software
  - Análisis del requerimiento inicial de información
  - Selección de datos
  - Preproceso de datos

[9] Este proyecto incluye fusiones y cierres de sucursales, así como nuevas sucursales, además de la actualización de los ATM`s e instalación de practicajas.

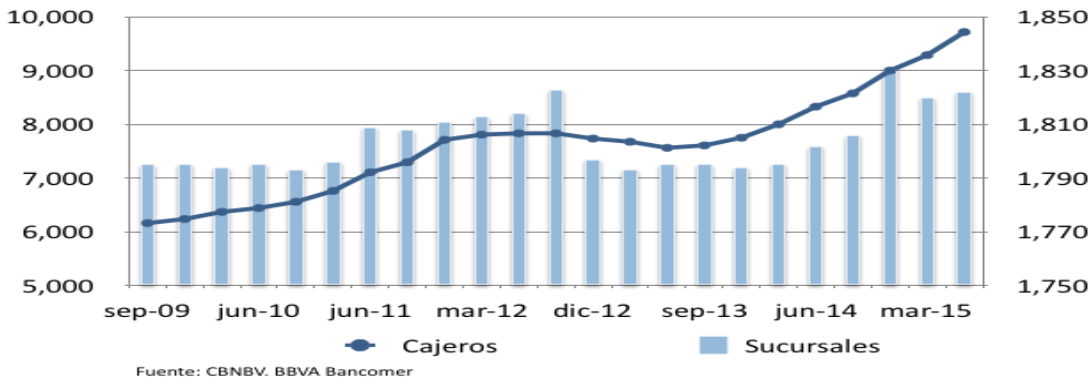
- **Aplicación y técnicas de modelado**
  - Selección y conjunto de datos para modelar
  - Diseño del diagrama de minería
  - Interpretación del modelo
- **Interpretación del modelo**
  - Elección del mejor modelo
- **Interpretación y contextualización**
- **Difusión, uso y monitoreo**

2.7.1.1 Análisis de la situación

El reto del proyecto es determinar una metodología que se adecue a las necesidades de cada sucursal. El método elegido, debe de tener las características para adaptarse al cambio de las sucursales mismas, por ejemplo, su evolución en la transaccionalidad, su equipamiento del autoservicio, etc.

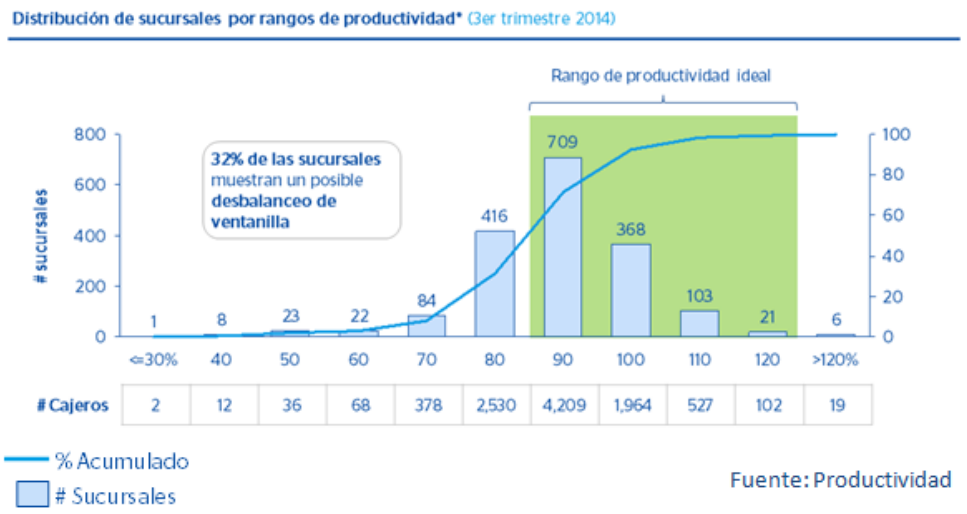
Además, se tiene como actual política reforzar la presencia de la institución bancaria con la modernización de las sucursales y la expansión del autoservicio. (Cuadro 2.7.1.1.1).

Volumen histórico de cajeros y sucursales (Sep. 09 a Mar. 15)



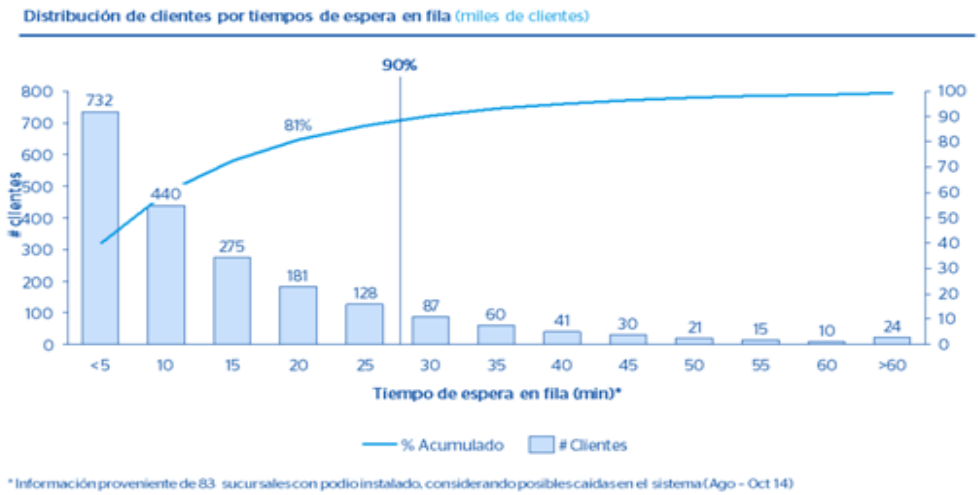
Cuadro 2.7.1.1.1

Por otro lado, si se analiza la productividad de las sucursales en la operativa de ventanilla, basados en un estándar de 14 transacciones por cada media hora, se obtiene que el 32% de las mismas están fuera de esta medida estándar y la media de las transacciones por cada media hora es de 10 (Cuadro 2.7.1.1.2)
Este resultado no necesariamente implica una disminución de plantilla, sino una propuesta de mejora en la asignación del personal por rangos medios horarios.



Cuadro 2.7.1.1.2

Si se analiza el nivel de servicio de las sucursales que cuentan con información de tiempos de servicio y clientes atendidos (sucursales con podio[11], Cuadro 2.7.1.1.3) se puede observar que el 80% de los clientes atendidos por media hora otorgan un nivel de servicio menor a 20 minutos y el 10% es atendido en más de 25 minutos.



Cuadro 2.7.1.1.3

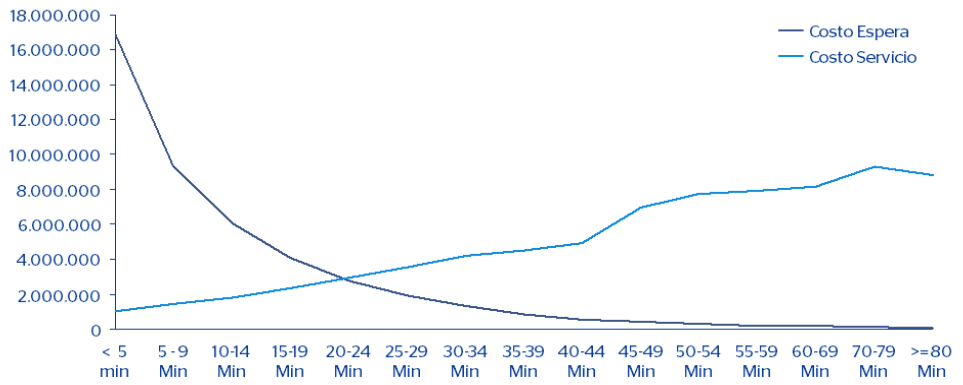
[11]Podio: Equipo electrónico en la sucursal que clasifica el tipo de cliente, el tiempo de espera y la caja en la cual fue atendido.

Ahora bien, para establecer un punto de equilibrio entre el nivel de servicio y el costo de otorgar ese nivel de servicio, se calcula el ingreso que genera un cliente y el supuesto de que cancele la relación con la institución. Por otro lado, se tienen los clientes que abandonan la sucursal asociado al nivel de servicio. La información anterior se asocia a las transacciones atendidas acorde al nivel de servicio otorgado y el costo de cada transacción.

Por lo tanto, se puede saber cuánto nos cuesta proporcionar un nivel de servicio (mientras menor sea el tiempo de atención, mayor costo de atender, pues se necesitan más personal) y cuánto nos cuesta que el cliente abandone la sucursal, dado mayor tiempo de servicio en sucursal.
El punto de equilibrio, corresponde al punto donde el costo de que el cliente abandone la sucursal sea igual al costo de que sea atendido es entre 20 y 24 minutos, como se observa en el cuadro 2.7.1.1.4

Dado lo anterior, se determina como nivel óptimo de tiempo de espera de 20 a 24 minutos.

Costos de espera y servicio, por tiempo promedio de espera (Ago a Oct 2014)



El costo de Espera estimado con clientes que abandonan la Sucursal, Estimando una pérdida por Ingreso $140

El costo de servicio, es el costo de $14.00 por transacción

Cuadro 2.7.1.1.4

2.7.1.2 Descripción del requerimiento del negocio y objetivos:

Como se comentó en la sección anterior (2.7.2.1), el objetivo es determinar la plantilla y tiempo de servicio óptimo para la red de sucursales de la Banca Comercial. El requerimiento de plantilla depende de los espacios físicos y el nivel de servicio que se desea otorgar.

El tiempo de servicio óptimo se obtiene mediante dos criterios, que son el costo de oportunidad vs. el costo del nivel de servicio. El otorgar un mayor tiempo de servicio costaría más a la institución por clientes que abandonan o perciben mal servicio, mismos que a su vez son detractores de la marca. Por el contrario, si se quiere otorgar un excelente nivel de servicio, el costo de personal, inmuebles, etc. puede ser muy alto y superar el beneficio de mantener al cliente, por lo tanto, el punto de equilibrio de estos dos costos se encuentra en el rango de 20 a 24 minutos. Tratando de ser menos conservador en el objetivo y más proactivos en el nivel de servicio, se establecerá el primer objetivo de nivel de servicio en el límite inferior del rango óptimo, que es de 20 minutos.

2.7.1.3 Alcance del proyecto.

La agrupación y separación de la información por nivel de servicio se determina conforme la información de podio[12], que sirve como base para establecer los 20 minutos de nivel de servicio (Actualmente las sucursales con podio (83 sucursales) otorgan un nivel de servicio de 20 minutos o menos al 80% de los clientes).
Adicional, evaluando costo de servicio vs. costo por tiempo de espera, el rango óptimo donde se equilibran ambos costos es en el rango de 20 a 24 minutos. Por lo tanto, se determina 20 minutos como el ideal para elaborar el balanceo y dar un servicio óptimo.

El objetivo de determinar la plantilla, se centra en un concepto más de espacio y capacidad física de la sucursal, ya que uno de los principales obstáculos para mejorar la sucursal es el espacio físico, sobre todo si es incrementar la capacidad de la sucursal, pues esto deriva en costo del personal, más costo de remodelación y adaptación, por lo tanto, el objetivo de plantilla óptima se centrará en obtener el número de espacios físicos, es decir, ventanillas que son necesarias para otorgar ese nivel de servicio e incurrir en el menor de los gastos.

2.7.1.4 Puntos clave que serán la base de los modelos a realizar

Los podios están programados para asignar un máximo tiempo de espera y prioridad de acuerdo al tipo de cliente (Empresa, Preferente, Cliente bancario y no cliente o usuario). Por lo que el llamado de cada tipo de cliente, depende de la clasificación y prioridad.

[12] Podio: Equipo en la sucursal que contabiliza el acceso de clientes y dirige al cliente a la ventanilla disponible.

Se propone modelar las mejores prácticas en sucursal que determinan los mejores niveles de servicio dado un flujo de clientes. Ahora bien, como no se tiene flujo de clientes para todas las sucursales, se analizará si las transacciones reflejan o se correlacionan con el flujo de clientes.

2.7.2 Integración y recopilación de datos

El objetivo es el número de ventanillas y con un análisis de minería de datos, se determinará las características y operativa que se necesita para establecer las ventanillas y el nivel de servicio que se ofrece de acuerdo al nivel de la transaccionalidad. Determinar el comportamiento de las sucursales, dado el nivel de servicio, es vital para un buen resultado del modelo.

Una vez definido el objetivo, el siguiente paso es definir la información que se necesitará y que se encuentra disponible, para resolver la problemática del número de ventanillas

2.7.2.1 Software

Se eligió a SAS-Enterprise-guide y SAS MINER para aplicar las técnicas de minería de datos, principalmente porque es el software institucional que utiliza para los procesos de análisis y minería. El SAS-Enterprise-guide tiene mucha facilidad para el manejo y manipulación de grandes cantidades de datos. Por otro lado el SAS MINER, es un módulo exclusivo de aplicación de técnicas de minería cuyo uso es muy sencillo y simple. Las técnicas estadísticas de minería que contiene, son muy estables y están validadas y supervisadas por SAS Institute.

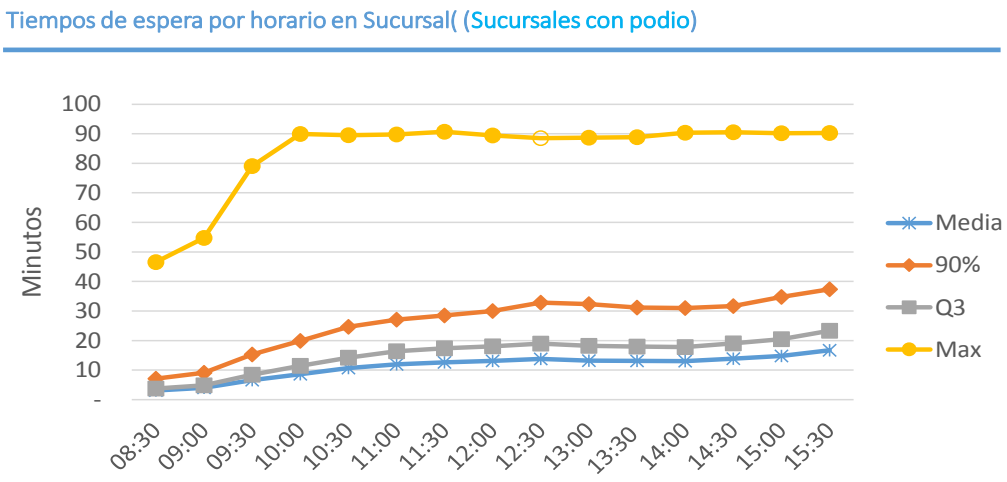
2.7.2.2 Análisis de requerimiento inicial de información

Un dato indispensable para el análisis, es el tiempo de espera de los clientes que asisten a las sucursales. Para ello se determinó utilizar la información de las 83 sucursales de podios que se tiene registrados con información de clientes y máximos tiempos de espera acorde a cada tipo de cliente.

Debido a que la información de tiempos de espera no se tiene a nivel cliente, sino que está agrupada por tipo de cliente y promedio por medias horas, se analizará la información por medias horas. Para ser conservadores en los tiempos de servicio para el análisis, se define como tiempo de servicio el máximo observado en esa media hora, sin considerar la tipología del cliente [14]. Por otro lado la gestión de personal en ventanilla, por parte del gerente administrativo, la verifica cada media hora.

[14] Nota: Los podios dan la versatilidad de establecer prioridades de atención de acuerdo al tipo de cliente que se registre en la sucursal. Por ejemplo los clientes preferentes y/o empresariales tendrán preferencia en la atención en ventanilla que una persona que no es cliente del Banco, es decir que no tiene cuenta en la institución.

La distribución de las observaciones de acuerdo a los tiempos de espera y el horario en el que son atendidos, se grafican en el cuadro 2.7.2.2.1.



Cuadro 2.7.2.2.1

La información que se utilizará para determinar comportamientos en las ventanillas es la información transaccional, características de la sucursal.
Debido a que durante 2014 se realizó un cambio en los perfiles del personal que atiende la ventanilla, y esto impacta en la gestión de la misma, se tomará la información desde el momento en que las sucursales operan bajo este nuevo esquema, el cual concluyó en el mes de junio. Por lo tanto, la información histórica utilizada para el modelo se realizará a partir del mes de junio y hasta el mes de octubre.

En seguida se detalla la información utilizada para realizar en análisis:

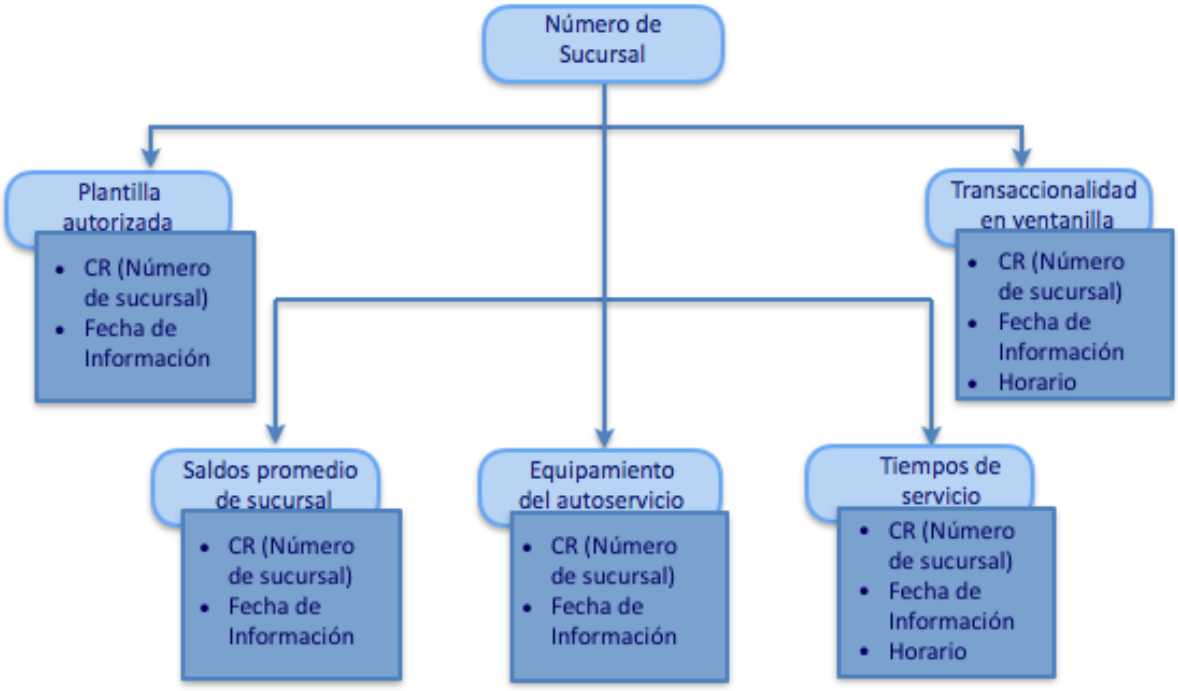
Características de la Sucursal	Clase de sucursal (Singular, A, B, etc.) Número de ejecutivos en sucursal Número de ATM´s en sucursal Número de practicajas en sucursal Número de cajeros B Número de cajeros A Número de cajeros part Saldos promedio en bóveda
Información transaccional	Fecha Horario de sucursal Número de ventanillas Número de transacciones

Información de Tiempos de Servicio	Fecha Horario Tiempos máximos de espera por cliente Tiempos máximos de espera por cliente preferente Tiempos máximos de espera por empresas Tiempos máximos de espera por usuario Número de clientes registrados Número de clientes atendidos Número de clientes que abandonan la sucursal

2.7.2.3 Selección de datos

En este apartado se hace referencia a las fuentes de información y la profundidad histórica que es necesaria y deseable. Así también se mapea la relación de las entidades y se muestra los orígenes y cálculo de las variables sintéticas.

En esta fase, es importante recordar que los datos suelen tener diferentes fuentes, como consecuencia se necesita integrar los datos antes de proceder al análisis.
Con la definición del objetivo, se necesita tener información operativa de la sucursal y su asociación con el servicio otorgado. El diagrama de información (Cuadro 2.7.2.3.1) que se muestra a continuación, menciona el nombre de las tablas y el campo llave con que se relacionan.



Cuadro 2.7.2.3.1

Una de las principales dimensiones con las que se trabaja en la minería de datos, es con el tiempo, los datos utilizados para obtener patrones comportamentales, tienen que estar alineados a la misma historia de datos.
En este caso, a pesar de tener datos transaccionales de más de 24 meses, se deberán acotar a la información de podios (que es donde se almacena el dato del tiempo de servicio y el número de clientes que arriban a la sucursal), que es del mes de Agosto al mes más reciente que, en este caso, es el mes de octubre.

2.7.2.3.1 Fechas de referencia

El análisis y resultado del modelo determina que deberá de cubrir los periodos de alta y baja afluencia de clientes. Como información disponible de los niveles de servicio se tiene historia de 3 meses, del mes de agosto al mes de octubre. Se determina este periodo suficiente para avanzar en el análisis, pues incluye días de alta transaccionalidad, como el regreso a clases en agosto, la semana de 15 de septiembre (día de mayor transaccionalidad en el año) y octubre que es el preámbulo al fin de año e inicio de ofertas de fin de año.

2.7.2.3.2 Definición del público objetivo

El modelo de balanceo, propone a las sucursales un número de ventanillas suficiente para otorgar un nivel de servicio deseado, en este caso de hasta 20 minutos. Deberá de contemplar las variaciones que existen en los días pico y horas pico, de manera que no quede corto en su estimación y disminuya la certidumbre del nivel de servicio.
Dado que los podios limitan la información de tiempo de espera en sucursal y funcionalmente no tienen ninguna excepción en cuanto a tipo de sucursal, no se pretende excluir ningún filtro en los datos.

2.7.2.4. Preproceso

En esta etapa, se analiza la calidad y suficiencia de los datos. Debido a las características propias de las técnicas de minería de datos, es necesario realizar una transformación de los datos para obtener una materia prima adecuada para el propósito particular.

2.7.2.4.1 Base para Análisis

Para poder utilizar la metodología de minería de datos, es necesario crear una tabla de información que contengan la profundidad histórica y de sucursales necesaria y suficiente para que el resultado sea exitoso.

La información que se tiene disponible para poder alinear esta información, es la información de podios, esta información representa 83 sucursales. Para decidir si es la adecuada para explotar la información de niveles de servicio y predecir un comportamiento, se analizará si la muestra es suficiente.

La institución tiene actualmente un total de 1746 sucursales, que incluyen módulos aduanales y sites[18] (Cuadro 2.7.2.4.1).

Podio /Tipo Sucursal	A	B	C	D	NUEVA	SINGULAR	Total	Núm. Suc.
Con Podio	21%	28%	35%	0%	9%	6%	5%	85
Sin Podio	9%	21%	52%	6%	10%	3%	95%	1,657
Total	9%	21%	51%	6%	10%	3%	100%	1,742

Cuadro 2.7.2.4.1

Por lo que se puede observar, la muestra tiene información de los tipos de sucursal más relevantes. Por otro lado, al evaluar el tamaño de la muestra se tiene:

$$n = \frac{Z_{\alpha}^2 * p * q * N}{NE^2 + Z_{\alpha}^2 * P * q}$$

Donde,
Z_{α} = Nivel de confianza deseado, en este caso del 95%
N = Tamaño de la población = 1742 sucursales
p = proporción esperada, que en este caso es del 5%.
q = 1 – p.
E = Precisión ó el error, en este caso se desea que fuese un 3%.

Con esto, nuestra formula se traduce en:

$$\mathbf{181 = \frac{\left(1.96^2 *.05 * (1-.05)\right)1742}{1742 *.03^2 + 1.96^2 *.05 * (1-.05)}}$$

Ahora bien, como ya se tiene el número de sucursales con Podio y si se cambia el error o precisión a 4.6% o la precisión, se obtiene:

$$\mathbf{84 = \frac{\left(1.96^2 *.05 * (1-.05)\right)1742}{1742 *.046^2 + 1.96^2 *.05 * (1-.05)}}$$

Siendo errores ó precisiones aceptables de 4 a 6%, se considera que el tamaño de la muestra de 85 sucursales con podios es suficiente para utilizar la información.

[18] Site: Servicio otorgado al cliente donde no hay personal de oficina, pues solo tiene dispositivos de autoservicio (ATM's) y las Cajas Automáticas que aceptan depósitos, llamadas comúnmente practicajas

2.7.2.4.2 Definición de variables sintéticas

Con el análisis de información se decidieron las siguientes variables sintéticas a construir para un mejor ajuste y discriminación de la variable objetivo.
Las variables sintéticas que se utilizarán en el análisis y/o selección de información:
Meses de maduración de Ulises[19] *(Meses_apl_ulises)*: Meses transcurridos desde que se modernizó el equipo y actualizaron las ventanillas y la fecha que se está evaluando.
Tiempo Máximo de Espera (MAX_TME): Max (Tiempo Medio de espera Clientes, Tiempo Medio de espera Usuarios, Tiempo Medio de espera Empresas, Tiempo medio de espera Preferentes).
Tiempo Máximo de Atención (MAX_TMA): Max (Tiempo Medio espera Clientes, Tiempo Medio de espera Usuarios, Tiempo Medio de espera Empresas, Tiempo medio de espera Preferentes)
Media Tiempo Medio de Espera (MEAN_TME): Max (Tiempo Medio espera Clientes, Tiempo Medio de espera Usuarios, Tiempo Medio de espera Empresas, Tiempo medio de espera Preferentes)
Clientes totales registrados (TOT_R): Suma (Clientes registrados, Usuarios registrados, Empresas registradas, Clientes Preferentes registrados)
Clientes totales atendidos (TOT_A): Suma (Clientes Atendidos, Usuarios Atendidos, Empresas Atendidas, Clientes Preferentes Atendidos)
Clientes Totales que abandonaron la fila (TOT_AB): Suma (Clientes que abandonaron la fila, Usuarios que abandonaron la fila, Empresas que abandonaron la fila, Clientes Preferentes que abandonaron la fila)
Clientes Promedio atendidos (CTES_PROM_ATEN): Clientes Total Atendidos entre número de ventanillas (TOT_A / NVentanillas).
Transacciones promedio X Cliente (TXN_PROM_CTE): Total transacciones entre total clientes atendidos (SUM_TXN / TOT_A).
Número de personas en la media hora observada (Fila): Fila de la media hora anterior más Total de clientes registrados menos Clientes Abandonados menos Clientes Atendidos. FILA_A + TOT_R - TOT_AB - TOT_A
Transacciones Promedio x Ventanilla (TXN_PROM_VT): Número de transacciones entre número de ventanillas. SUM_TXS/NVENTANILLAS.
Total Autoservicio (total_at): Total Atm's más total Practicajas: (Atm´s + practicajas).
Rango de Tipo de sucursal (R_Tipo_Suc), transformación de variable de grupo a variable numérica (Cuadro 2.7.2.4.2.1):

Cuadro 2.7.2.4.2.1

Tipo sucursal	R_Tipo_Suc
A	1
B	2
C	3
D	4

[19] Sucursal que actualizó la infraestructura inmobiliaria y de equipo para autoservicio (Proyecto ULISES)

NUEVA	5
SINGULAR	6

Bandera Ulises: (R_Ban_ulises): 0 si Ban_Ulises = "NO"; 1 si Ban_Ulises = "SI".
Logaritmo Máximo Tiempo Medio de Espera (LN_MAX_TME): Logaritmo del Máximo tiempo de espera: LOG (MAX_TME).
Logaritmo Máximo Tiempo Medio de Atención (LN_MAX_TMA): Logaritmo del Máximo tiempo de atención: LOG (MAX_TMA).
Logaritmo Media Tiempo Medio de Espera (LN_MEAN_TME): Logaritmo de la Media del tiempo de espera LOG (MEAN_TME).
*Target_Tiempo Media de la Media de Tiempo de Espera (*TRG_MEX_TME): Si MAX_TME >30, 1; en caso contrario , 0
*Target_Tiempo Máximo de la Media del Tiempo de Espera (*TRG_MAX_TME) Si MAX_TME > 25 => 1; en caso contrario 0.

2.7.2.4.3 Validaciones de auditoría

Para auditar la construcción del modelo, se realizarán diferentes validaciones a lo largo de su construcción:

Análisis de la información y resumen de sus características, identificadores utilizados

- Número de registros y atributos de cada tabla utilizada
- Garantizar existencia y consistencia de la información para cada identificador en las varias fuentes de información
- Otras características relevantes para la descripción de los mismos

Obtención de un conocimiento más detallado sobre los datos

- Técnicas utilizadas: consultas de datos y visualización gráfica
- Verificación de la distribución de los valores de los atributos y de la relación entre los atributos (validación con know-how de negocio)

Verificación de la completitud de los datos

- Análisis de errores: cuál es la frecuencia de los errores o de valores ausentes
- Descriptivo de los problemas encontrados en los datos, las correcciones realizadas en los mismos y los errores que aún persisten
- Revisión y optimización de los datos para que cumplan con los estándares de calidad necesarios para el análisis con técnicas de minería de datos.
- Corrección de anomalías
- Tratamiento de los valores ausentes

En el anexo se especifica más detalladamente los análisis que se realizaron de cada variable.

En los cuadros siguientes se resumen las medidas de tendencia central y de dispersión para excluir y redefinir variables en el análisis:

Cuadro 2.7.2.4.3.1					
Variable	N	Suma	Valores ausentes	Desviación estándar	Media
nfecha	83,942	1,677,309,809	-	26.83711	19,981.77
CTE_T_M_E_	83,942	845,703	-	11.61771	10.07485
CTE_T_M_A_	83,942	342,193	-	2.50889	4.07654
CTE_C__R_	83,942	1,306,908	-	10.91407	15.56918
CTE_C__A_	83,942	1,199,969	-	9.49798	14.29522
CTE_C__Ab_	83,942	110,149	-	2.65643	1.31220
EMP_T_M_E_	83,942	144,600	-	5.17491	1.72262
EMP_T_M_A_	83,942	142,197	-	5.20398	1.69399
EMP_C__R_	83,942	27,718	-	0.80356	0.33020
EMP_C__A_	83,942	22,849	-	0.71489	0.27220
EMP_C__Ab_	83,942	1,960	-	0.17802	0.02335
PRE_T_M_E_	82,526	341,416	1,416	5.14925	4.13707
PRE_T_M_A_	82,526	331,523	1,416	4.41367	4.01720
PRE_C__R_	83,942	243,905	-	3.29399	2.90564
PRE_C__A_	83,942	219,384	-	3.00182	2.61352
PRE_C__Ab_	83,942	24,913	-	0.70630	0.29679
USU_T_M_E_	82,526	609,550	1,416	12.23287	7.38615
USU_T_M_A_	82,526	175,343	1,416	2.88511	2.12470
USU_C__R_	83,942	466,340	-	7.63272	5.55550
USU_C__A_	83,942	425,894	-	6.94271	5.07367
USU_C__Ab_	83,942	41,897	-	1.48584	0.49912
SUM_TXS	83,942	3,295,419	-	20.74087	39.25829
NVENTANILLAS	83,942	350,542	-	1.54449	4.17600
TOT_R	83,942	2,044,871	-	12.44447	24.36052
TOT_A	83,942	1,868,096	-	10.12422	22.25460
TOT_AB	83,942	178,919	-	3.51493	2.13146
MAX_TME	83,942	1,005,319	-	13.43000	11.97635
MAX_TMA	83,942	575,243	-	5.74275	6.85286
CTES_PROM_ATEN	83,942	468,646	-	2.43440	5.58298
TXN_PROM_CTE	83,636	185,348	306	3.54590	2.21613
FILA_A	83,942	781,316	-	31.39709	9.30781
FILA	83,942	781,316	-	31.39709	9.30781
TXN_PROM_VT	83,942	795,550	-	3.69822	9.47738
MEAN_TME	80,015	738,919	3,927	10.11242	9.23476
cu_a	83,942	183,434	-	0.53991	2.18525

Cuadro 2.7.2.4.3.1					
Variable	N	Suma	Valores ausentes	Desviación estándar	Media
cu_b	83,942	314,244	-	1.82433	3.74358
cu_pt	83,942	29,079	-	0.52671	0.34642
ejecutivos	83,942	308,078	-	2.05399	3.67013
total	83,942	622,322	-	3.60505	7.41371
atms	83,942	252,695	-	1.02856	3.01035
practicajas	83,942	158,360	-	0.68729	1.88654
total_at	83,942	411,055	-	1.52366	4.89689
atm	83,942	252,695	-	1.02856	3.01035
practicaja	83,942	157,336	-	0.67621	1.87434
total_atm	83,942	410,031	-	1.50263	4.88469
saldo_promedio	83,942	132,957,162,788	-	816,112.610	1,583,917
maximo	83,942	286,169,373,596	-	1,938,857.30	3,409,132
aut_prom_diario	83,942	3,047,513	-	14.27499	36.30498
am_ulises	83,942	1,645,238,705	-	194.15652	19,599.71
R_TIPO_SUC	83,942	219,930	-	1.37146	2.62002
R_BAN_ULISES	83,942	83,942	-	-	1.00000
LN_MAX_TME	83,942	156,424	-	1.21658	1.86348
LN_MAX_TMA	83,942	144,741	-	0.58758	1.72429
LN_MEAN_TME	83,942	129,675	-	1.22439	1.54481
TRG_MEX_TME	83,942	72,507	-	0.34303	0.86378
TRG_MAX_TME	83,942	72,507	-	0.34303	0.86378
MES_AP_UL	83,942	1,013,886	-	6.41568	12.07841

Otra información para análisis es la obtención de percentiles, con la cual podemos tener una idea de la distribución y si tiene valores extremos.

Cuadro 2.7.2.4.3.2										
Variable	Media	Mínimo	Máximo	1st Pctl	10th Pctl	25th Pctl	75th Pctl	90th Pctl	95th Pctl	99th Pctl
nfecha	19,983	19,936	20,027	19,936	19,946	19,957	20,005	20,019	20,024	20,026
CTE_T_M_E_	5.85	-	90.48	-	0.78	2.07	13.95	25.28	33.82	53.87
CTE_T_M_A_	3.68	-	62.20	-	2.30	2.93	4.72	6.18	7.53	12.40
CTE_C__R_	14.00	-	122.00	-	2.00	7.00	22.00	30.00	35.00	46.00
CTE_C__A_	13.00	-	87.00	-	3.00	7.00	20.00	27.00	32.00	41.00
CTE_C__Ab_	1.00	-	169.00	-	-	-	2.00	3.00	5.00	10.00

Cuadro 2.7.2.4.3.2										
Variable	Media	Mínimo	Máximo	1st Pctl	10th Pctl	25th Pctl	75th Pctl	90th Pctl	95th Pctl	99th Pctl
EMP_T_M_E_	-	-	90.62	-	-	-	-	6.18	11.38	25.12
EMP_T_M_A_	-	-	62.38	-	-	-	-	6.08	11.63	26.05
EMP_C__R_	-	-	13.00	-	-	-	-	1.00	2.00	4.00
EMP_C__A_	-	-	15.00	-	-	-	-	1.00	2.00	3.00
EMP_C__Ab_	-	-	8.00	-	-	-	-	-	-	1.00
PRE_T_M_E_	3.20	-	90.70	-	-	0.42	5.85	8.95	11.98	23.32
PRE_T_M_A_	3.33	-	62.92	-	-	1.18	5.38	8.27	11.13	20.70
PRE_C__R_	2.00	-	51.00	-	-	1.00	4.00	7.00	9.00	16.00
PRE_C__A_	2.00	-	45.00	-	-	1.00	4.00	6.00	8.00	15.00
PRE_C__Ab_	-	-	38.00	-	-	-	-	1.00	2.00	3.00
USU_T_M_E_	1.87	-	90.68	-	-	-	9.58	22.88	33.50	56.97
USU_T_M_A_	1.98	-	62.35	-	-	-	3.22	4.63	6.08	11.93
USU_C__R_	2.00	-	89.00	-	-	-	9.00	16.00	21.00	31.00
USU_C__A_	1.00	-	64.00	-	-	-	8.00	15.00	19.00	28.00
USU_C__Ab_	-	-	186.00	-	-	-	1.00	2.00	2.00	5.00
SUM_TXS	36	1	281	5	18	26	48	64	78	111
NVENTANILLAS	4	1	12	2	3	3	5	7	7	9
TOT_R	24	-	125	-	7	17	32	40	45	57
TOT_A	22	-	98	1	9	16	28	35	40	49
TOT_AB	1	-	210	-	-	-	3	5	7	13
MAX_TME	7.07	0.02	90.70	0.33	1.25	2.78	16.25	29.58	40.08	63.42
MAX_TMA	5.13	-	62.92	1.55	3.02	3.82	7.58	12.27	17.08	32.27
CTES_PROM_ATEN	5.50	-	33.00	0.25	2.60	4.00	7.00	8.67	9.67	12.00
TXN_PROM_CTE	1.63	0.03	125.00	0.84	1.16	1.35	2.06	2.95	4.20	14.00
FILA	4.00	-514.00	492.00	-80.0	-18.00	-4.00	22.00	45.00	62.00	103.0
TXN_PROM_VT	9.25	1.00	54.50	2.00	5.14	7.00	11.50	14.00	15.75	20.00

Cuadro 2.7.2.4.3.2										
Variable	Media	Mínimo	Máximo	1st Pctl	10th Pctl	25th Pctl	75th Pctl	90th Pctl	95th Pctl	99th Pctl
MEAN_TME	5.66	-	89.22	0.27	0.96	2.17	12.82	22.28	29.52	47.24
cu_a	2.00	2.00	5.00	2.00	2.00	2.00	2.00	3.00	3.00	5.00
cu_b	3.00	1.00	12.00	1.00	2.00	3.00	4.00	6.00	7.00	12.00
cu_pt	-	-	2.00	-	-	-	1.00	1.00	1.00	2.00
ejecutivos	3.00	1.00	11.00	1.00	2.00	2.00	5.00	7.00	8.00	11.00
total	6.00	3.00	22.00	3.00	4.00	5.00	9.00	13.00	15.00	22.00
atms	3.00	1.00	5.00	1.00	2.00	2.00	4.00	4.00	5.00	5.00
practicajas	2.00	1.00	4.00	1.00	1.00	1.00	2.00	3.00	3.00	4.00
total_at	5.00	2.00	9.00	2.00	3.00	4.00	6.00	7.00	8.00	9.00
atm	3.00	1.00	5.00	1.00	2.00	2.00	4.00	4.00	5.00	5.00
practicaja	2.00	1.00	4.00	1.00	1.00	1.00	2.00	3.00	3.00	4.00
total_atm	5.00	2.00	9.00	2.00	3.00	4.00	6.00	7.00	8.00	9.00
saldo_promedio	1,362,643	548,312	4,535,346	548,312	769,722	1,043,815	1,834,085	2,602,674	3,614,821	4,535,346
maximo	2,899,537	886,716	9,694,628	886,716	1,272,989	1,945,792	4,334,873	6,159,520	7,726,617	9,694,628
aut_prom_diario	32.01	13.50	85.54	13.50	22.55	25.67	42.91	63.37	66.39	73.99
am_ulises										
R_TIPO_SUC	2.00	1.00	6.00	1.00	1.00	2.00	3.00	5.00	6.00	6.00
R_BAN_ULISES	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LN_MAX_TME	1.955	-4.094	4.508	-1.099	0.223	1.024	2.788	3.387	3.691	4.150
LN_MAX_TMA	1.636	-0.629	4.142	0.438	1.104	1.339	2.026	2.507	2.838	3.474
LN_MEAN_TME	1.646	-5.347	4.491	-1.261	-	0.607	2.511	3.081	3.369	3.845
TRG_MEX_TME	1.0	-	1.0	-	-	1.0	1.0	1.0	1.0	1.0
TRG_MAX_TME	1.0	-	1.0	-	-	1.0	1.0	1.0	1.0	1.0
MES_AP_UL	9.0	4.0	36.0	4.0	6.0	7.0	17.0	19.0	23.0	35.0

2.7.2.4.4 Depuración de información

Como depuración de información, se eliminarán observaciones donde el tiempo de espera de un cliente sea cero.
Si el horario no está definido, se eliminará la observación.
Si el número de ventanillas no está informado, se borra la observación
Si el máximo tiempo de espera no está informado, se borra la información.

2.7.2.4.5 Definición del target

Derivado de que el nivel de atención y las ventanillas atienden a todos los clientes independientemente de a qué segmento pertenecen y tratando de ser conservadores, el nivel de atención estará definido como el máximo tiempo observado en cualquier segmento que asiste a la sucursal, de esta manera se eliminará la información por segmento que se tiene en la información de Podios.

Derivado del objetivo, que es determinar la operación que se observa en ventanilla y que está dando el tiempo de servicio deseado, el target será el número de ventanillas.

Ahora bien, para definir si es posible combinar la información de ventanillas y tiempos del nivel de servicio, se deberá de analizar si la información está correlacionada.
Para ello se realizó un análisis de correlación entre las variables finales (23 variables), ver Anexo I. Donde se puede observar que sí hay correlación entre las variables de podio, (Clientes atendidos y recibidos) vs Transacciones y ventanillas.

Por lo tanto, se puede utilizar la información de tiempo de servicio, para determinar las características de las sucursales que operan con el nivel de servicio definido.
Por otro lado, se analizaron distribuciones por diferentes niveles de servicio, y se observa una separación de los datos

2.7.2.4.6 Definición de base de datos y target

Base de datos: Información de transaccionalidad donde el máximo tiempo de espera es menor o igual a 20 minutos. La nueva base de datos, con 77,580 con registros (Cuadro 2.7.2.4.6.1).

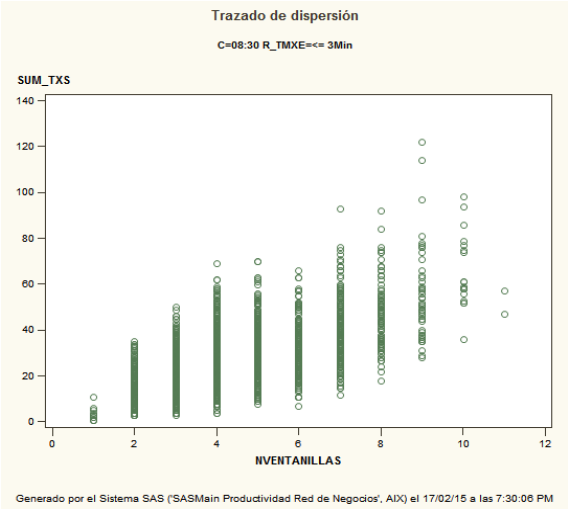
Cuadro 2.7.2.4.6.1

Rango Tiempo Max Espera	Núm. Registros	% Acumulado
<= 3Min	21,378	28%
3.0 - 7.0	18,513	51%
7.0 - 14.0	16,408	73%
14.0 - 21.0	8,638	84%
21.0 - 28.0	5,184	90%
28.0 - 35.0	2,903	94%
35.0 - 42.0	1,745	96%

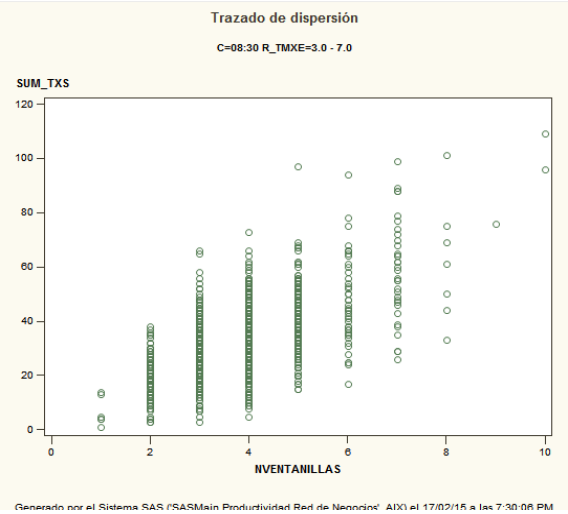
Cuadro 2.7.2.4.6.1		
Rango Tiempo Max Espera	Núm. Registros	% Acumulado
42.0 - 50.0	1,194	98%
>50	1,617	100%
Total	77,580	

En cuanto a análisis de información, se realizaron análisis de distribuciones, gráficas, que se especifican en el anexo. A manera de resumen se presentan algunos de los análisis. Se analizaron posibles distribuciones por clases de niveles de servicio, lo cual corrobora, de manera gráfica, la correlación con el nivel de servicio y número de ventanillas.

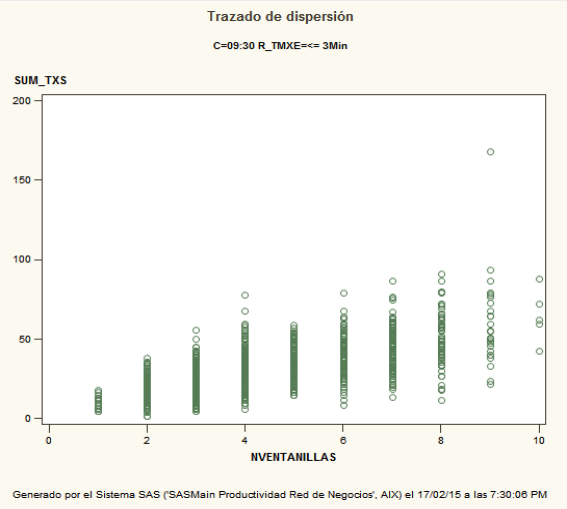
Como se puede observar (Gráficas 2.7.2.4.6.2 a 2.7.2.4.6.5), hay una correlación entre las ventanillas y las transacciones, también dependiendo de los niveles de servicio, se observan diferentes acumulados de observaciones. También se puede observar que a mayor nivel de servicio menor el número de ventanillas que están dando servicio.



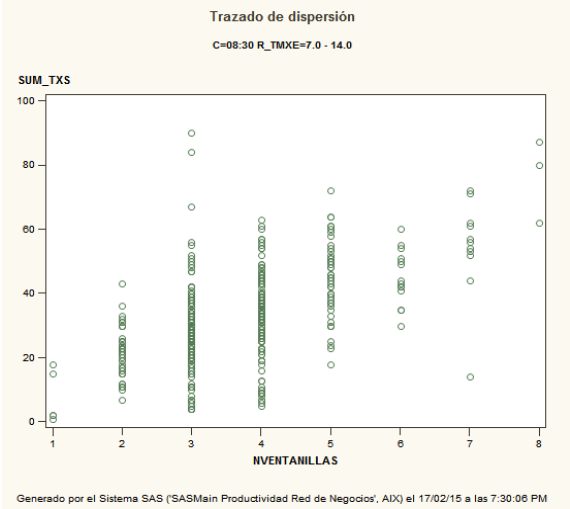
Gráfica 2.7.2.4.6.2



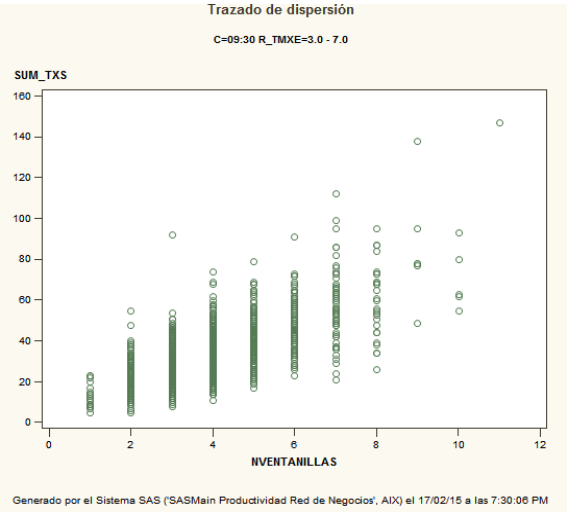
Gráfica 2.7.2.4.6.3



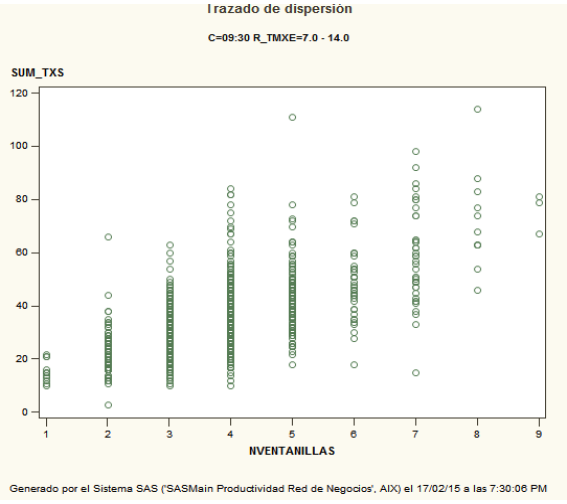
Gráfica 2.7.2.4.6.4



Gráfica 2.7.2.4.6.5



Gráfica 2.7.2.4.6.6



Gráfica 2.7.2.4.6.7

2.7.3 Aplicación y técnicas de modelado

En esta sección se define el tipo de problema y la técnica a utilizarse. Lo que se pretende es obtener una estimación de las ventanillas que dan servicio dentro del tiempo menor a 20 minutos, y encontrar variables independientes que determinen la operativa de la sucursal.

En este caso se puede descartar que sea un problema descriptivo, pues no requerimos clasificar a las sucursales, sino estimar las ventanillas bajo ciertas características o patrones de operación. Por lo tanto, es un problema predictivo.

2.7.3.1 Algoritmo del modelado

El objetivo del modelo (Número de ventanillas), la definiremos como una variable continua, dado que puede tomar cualquier valor entero. En la práctica, se puede acotar el número de ventanillas en el caso de que se obtengan valores muy altos[20].

Derivado del problema de tipo predictivo y que el objetivo es una variable continua, la técnica de modelización puede abordarse con regresión lineal múltiple y árboles de decisión. Aun cuando las redes neuronales son otra opción de técnica de estimación, no será utilizada por su complejidad en su interpretación de cara al negocio.

De las metodologías de árboles de decisión y regresión, se utilizará la que tenga mejor ajuste.

2.7.3.2 Selección de conjuntos de datos para modelar

El conjunto de datos de modelación (ModelSet) se divide típicamente en 3 bloques, con distintas finalidades en el proceso de desarrollo del modelo.

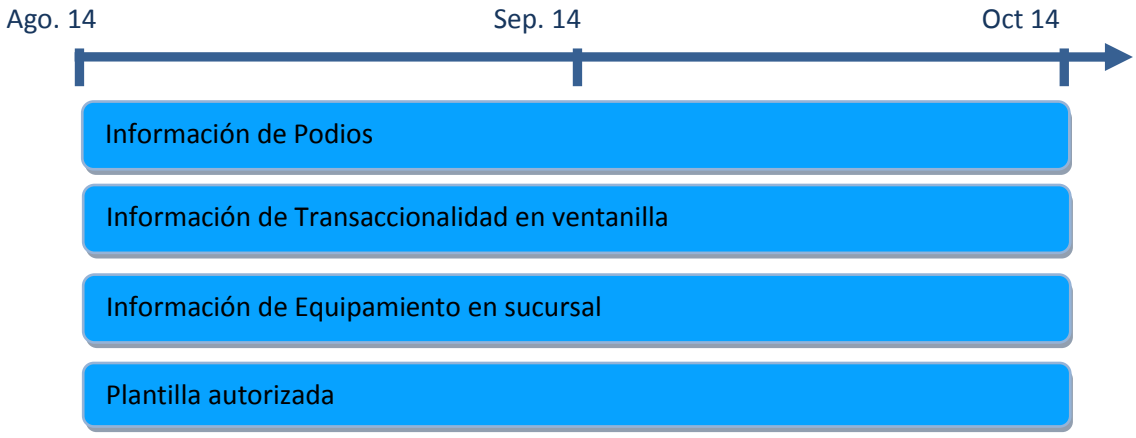
Datos de entrenamiento: se detectan patrones de comportamiento que predigan el evento a partir de las variables independientes. Es el conjunto de información que permitirá el desarrollo de varios modelos, generando las reglas y parámetros funcionales (función de puntuación). Aproximadamente 50% de las observaciones del ModelSet serán asignadas a entrenamiento. Este conjunto puede ser ligeramente incrementando si el número de observaciones totales fuera escaso, para así sobre ponderar el conjunto de entrenamiento.
Datos de validación: conjunto de datos que interviene en el modelado, aunque de forma indirecta. Se usa para seleccionar el mejor modelo de la fase de entrenamiento. Aproximadamente 30% de las observaciones serán dedicadas a este conjunto.
Datos de prueba: un % de clientes del ModelSet (20%) que no se usa en la construcción del modelo. Se aplicará el modelo en este conjunto de clientes como forma de identificar posibles problemas de sobre-ajuste. Es sobre este subconjunto del ModelSet donde se analizará las principales medidas de desempeño del modelo desarrollado.

[20] En la práctica, no existen sucursales con más de 15 ventanillas operando.

Adicionalmente, se recomienda evaluar el desempeño del modelo en otros periodos de tiempo. Esta es la mejor forma de validar la predictibilidad del modelo ya que presentará los resultados más cercanos a la implementación.

No obstante, y debido a la escasa profundidad histórica que se ha podido utilizar para modelar, se ha optado por no utilizar base de test y que la información se divida en entrenamiento y validación con un 50% en ambas muestras.

El periodo de evaluación:



Cuadro 2.7.3.2.1

El éxito en la modelación se revisa en dos vías:

- La primera, donde se mide la eficacia del modelo, mediante indicadores estadísticos
- Vía Negocio, donde se miden indicadores de productividad y se contrastan con resultados del modelo evaluando que la clasificación sea coherente y tenga sentido del negocio.

2.7.3.3 Tabla destino y variables que entran al modelo

La base de datos general se separará de acuerdo al nivel de servicio que se tiene para evaluar varios escenarios y en el impacto de las ventanillas requeridas, así también para determinar el costo de mejorar el nivel de servicio.

De esta manera, se separará la información y se incluirán en el modelo las diferentes muestras y diferentes escenarios, incluyendo aquellos cuyo nivel de servicio no es el óptimo (Cuadro 2.7.3.3.1).
Con esta separación obtendremos, en donde se tiene tiempo de servicios óptimos y dentro del rango deseado y niveles de servicio poco deseables.

Cuadro 2.7.3.3.1				
Nombre de la tabla	Nivel de servicio	Media de TME	Observaciones	Comentarios
PODIO_1	Hasta 5 min	2.00	33,703	El volumen de información es suficiente, sin embargo, el nivel de servicio es demasiado ambicioso, pues requeriría un gran número de cajeros, impactando en el costo, con beneficio marginal. Por lo tanto queda fuera de las expectativas del negocio.
PODIO_2	5 a 10 Min	7.20	17,026	El volumen de observaciones es suficiente aunque el objetivo del nivel de servicio aún está fuera de las expectativas del negocio.
PODIO_3	10 - 15 Min	12.30	10,311	El volumen de observaciones es suficiente y el objetivo del nivel de servicio razonable.
PODIO_4	15-20 min	17.30	6,760	El volumen de observaciones es suficiente y el objetivo del nivel de servicio razonable.
PODIO_5	20-25 Min	22.30	4,707	El volumen de observaciones es suficiente y el objetivo del nivel de servicio es el óptimo en cuanto a costo-beneficio.
PODIO_6	25 - 30 Min	27.30	3,261	El volumen de observaciones es menor a los anteriores escenarios, el resultado puede no ser el óptimo en cuanto a estimación y expansión a todas las sucursales.
PODIO_7	30-35 min	32.30	2,337	El volumen de observaciones es menor a los anteriores escenarios , el resultado puede no ser el óptimo en cuanto a estimación y expansión a todas las sucursales.
PODIO_8	35-40 min	37.30	1,619	El volumen de observaciones es muy bajo, se recomienda unificar ambos niveles de servicio para obtener resultados más consistentes.
PODIO_9	40-45 min	42.30	1,227	
NE	45-50 min	47.20	869	Muy pocas observaciones, adicional a que el nivel de servicio no es lo que la institución busca.
NE	50-55 min	52.30	638	Muy pocas observaciones, adicional a que el nivel de servicio no es lo que la institución busca.

Cuadro 2.7.3.3.1				
Nombre de la tabla	Nivel de servicio	Media de TME	Observaciones	Comentarios
NE	55-60 min	57.20	402	Muy pocas observaciones, adicional a que el nivel de servicio no es lo que la institución busca.
NE	> 60 min	71.50	1,084	Muy pocas observaciones, adicional a que el nivel de servicio no es lo que la institución busca.
Podio_20M	18-20 min	19.90	4,458	El volumen de observaciones es suficiente y el objetivo del nivel de servicio es mejor al óptimo en cuanto a costo-beneficio.
Podio_H20M	<= 20 min	6.60	67,800	El volumen de observaciones es suficiente, las observaciones están dentro del nivel de servicio.
Podio_H30M	<= 30 min	8.40	75,768	El volumen de observaciones es suficiente, aun cuando las observaciones están fuera del nivel de servicio, se evaluará para determinar el costo-beneficio de otorgar este nivel de servicio.

Una vez ya definido el objetivo se deben elegir las variables que entrarían al modelo. A continuación se detallan las variables que entran y el porqué, de la exclusión del resto (Cuadro 2.7.3.3.2).

Cuadro 2.7.3.3.2				
Base de datos de Entrada para cálculo de ventanillas dado un nivel de servicio				
#	**Variable**	**Descripción**	**En Modelo (Si/No)**	**Comentarios**
1	CR	Identificación de la sucursal	Sí	Como ID
2	nfecha	Fecha de la observación	Sí	Como Fecha
3	HORARIO	Horario	Sí	Como variable de clase
4	CTE_T_M_E_	Tiempo medio de espera del cliente	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
5	CTE_T_M_A_	Tiempo medio de atención del cliente	No	Si sale significativa, no se podría contar con esta información para todos los clientes.

Cuadro 2.7.3.3.2				
Base de datos de Entrada para cálculo de ventanillas dado un nivel de servicio				
#	Variable	Descripción	En Modelo (Si/No)	Comentarios
6	CTE_C__R_	Total de clientes recibidos en la sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
7	CTE_C__A_	Total de clientes atendidos en la sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
8	CTE_C__Ab_	Total de clientes que abandonan la sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
9	EMP_T_M_E_	Tiempo medio de espera de la empresa	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
10	EMP_T_M_A_	Tiempo medio de atención de la Empresa	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
11	EMP_C__R_	Total de clientes empresariales recibidos en sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
12	EMP_C__A_	Total clientes empresariales atendidos en sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
13	EMP_C__Ab_	Total de clientes empresariales que abandonan la sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
14	PRE_T_M_E_	Tiempo medio de espera del cliente preferente	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
15	PRE_T_M_A_	Tiempo medio de atención del cliente preferente	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
16	PRE_C__R_	Total de clientes preferentes recibidos en la sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
17	PRE_C__A_	Total de clientes preferentes atendidos en sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.

Cuadro 2.7.3.3.2				
Base de datos de Entrada para cálculo de ventanillas dado un nivel de servicio				
#	Variable	Descripción	En Modelo (Si/No)	Comentarios
18	PRE_C__Ab_	Total clientes preferentes que abandonan la sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
19	USU_T_M_E_	Tiempo medio de espera del usuario	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
20	USU_T_M_A_	Tiempo medio de atención del usuario	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
21	USU_C__R_	Total de usuarios recibidos en la sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
22	USU_C__A_	Total de usuarios atendidos en la sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
23	USU_C__Ab_	Total de Usuarios que abandonan la sucursal.	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
24	SUM_TXS	Suma de transacciones hechas en el tiempo y día de observación.	Sí	
25	NVENTANILLAS	Número de ventanillas en sucursal en el horario y día de observación.	Target	Lo que se desea estimar
26	TOT_R	Total de personas que se registraron en la sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
27	TOT_A	Total de personas atendidas en sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
28	TOT_AB	Tota de personas que abandonaron la sucursal	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
29	MAX_TME	Máximo tiempo de espera del promedio por segmento (cliente,	No	Si sale significativa, no se podría contar con esta información para todos los clientes.

Cuadro 2.7.3.3.2				
Base de datos de Entrada para cálculo de ventanillas dado un nivel de servicio				
#	Variable	Descripción	En Modelo (Si/No)	Comentarios
		usuario, Empresa, preferente)		
30	MAX_TMA	Máximo tiempo de atención del promedio por segmento (cliente, usuario, empresa, preferente)	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
31	CTES_PROM_ATEN	Clientes promedio atendidos	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
32	TXN_PROM_CTE	Transacciones promedio por cliente	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
33	FILA_A	Número de personas en fila en la media hora anterior	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
34	FILA	Número de personas en fila en el momento de la observación	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
35	TXN_PROM_VT	Transacciones promedio por ventanilla	No	Se deriva del target, por lo cual puede sesgar los datos.
36	MEAN_TME	Media del tiempo medio de espera de los segmentos	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
37	R_TMXE	Rango del máximo tiempo de espera	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
38	R_TME	Rango de tiempo medio de espera	No	Si sale significativa, no se podría contar con esta información para todos los clientes.
39	tipo_suc	Tipo de sucursal (A, B, Singular, etc.)	Sí	
40	cu_a	Número de cajeros universales A	Sí	
41	cu_b	Número de cajeros universales B	Sí	
42	cu_pt	Número de cajeros universales part	Sí	

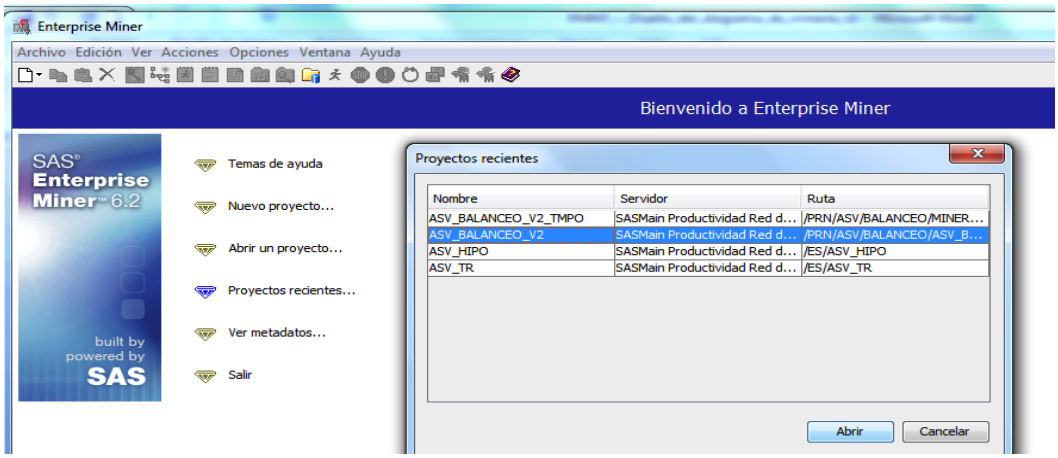
Cuadro 2.7.3.3.2				
Base de datos de Entrada para cálculo de ventanillas dado un nivel de servicio				
#	**Variable**	**Descripción**	**En Modelo (Si/No)**	**Comentarios**
43	ejecutivos	Número de personas que dan servicio en la sucursal	Sí	
44	total	Total de personas en la sucursal	Sí	
45	atms	Número de ATM's	Sí	
46	practicajas	Número de Practicajas	Si	
47	total_at	Total de equipamiento (ATM's + practicajas)	Sí	
51	saldo_promedio	Saldo promedio de la sucursal que maneja en bóveda	Sí	
52	maximo	Máximo promedio observado en la sucursal	No	No se cuenta con los datos, y tener la información puede atrasar el proceso.
53	aut_prom_diario	Autorizaciones promedio diarias	No	No se cuenta con los datos, y tener la información puede atrasar el proceso.
54	ban_ulises	Marca de si es Ulises o no (SI/NO)	No	No se considera porque todas las que tienen podio son Ulises
55	am_ulises	Año-mes de fecha de entrega Ulises	Sí	
56	R_TIPO_SUC	Transformación a valor numérico del tipo de sucursal	Sí	
57	R_BAN_ULISES	Transformación a valor numérico del campo tipo de ban - ulises	Sí	
58	LN_MAX_TME	Logaritmo de máximo tiempo de espera	No	Datos sólo de podio, al expandir no se puede contar con la información para todas las sucursales.
59	LN_MAX_TMA	Logaritmo de máximo tiempo de atención	No	Datos sólo de podio, al expandir no se puede contar con la información para todas las sucursales.

Cuadro 2.7.3.3.2				
Base de datos de Entrada para cálculo de ventanillas dado un nivel de servicio				
#	Variable	Descripción	En Modelo (Si/No)	Comentarios
60	LN_MEAN_TME	Logaritmo de la media del tiempo de espera	No	Datos sólo de podio, al expandir no se puede contar con la información para todas las sucursales.
61	TRG_MEX_TME	Dicotómico que marca si cumple con Hasta 25 minutos de atención a sucursal o no	No	No es el objetivo del Modelo.

2.7.3.4 Diagrama de flujo de los modelos desarrollados

SAS Enterprise Miner (SAS EM) es una solución de minería de datos que proporciona gran cantidad de modelos y de alternativas y además compara los resultados de las distintas técnicas de modelaje en términos estadísticos.

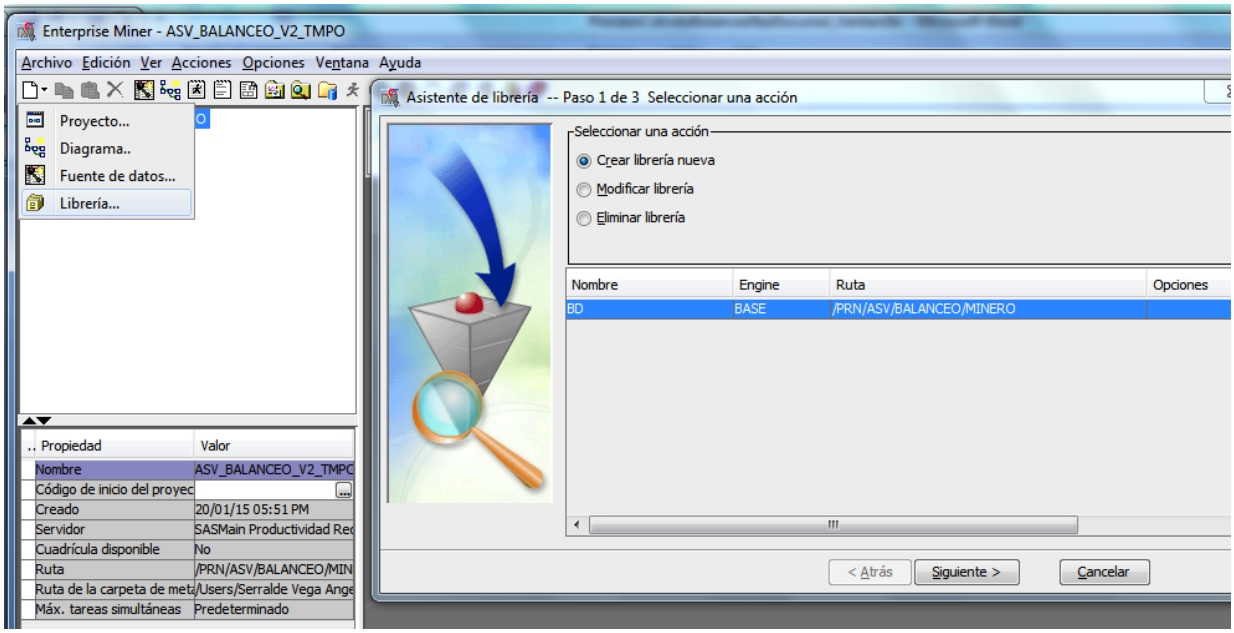
Los diagramas generados en SAS EM realizados contienen todas las tareas de modelaje, desde la carga de la información hasta la comparación de los modelos resultantes. Por lo tanto, el modelo de estimación de número de ventanillas, dado un nivel de servicio bajo diferentes escenarios, se desarrollará en la herramienta SAS Miner:



Cuadro 2.7.3.4.1

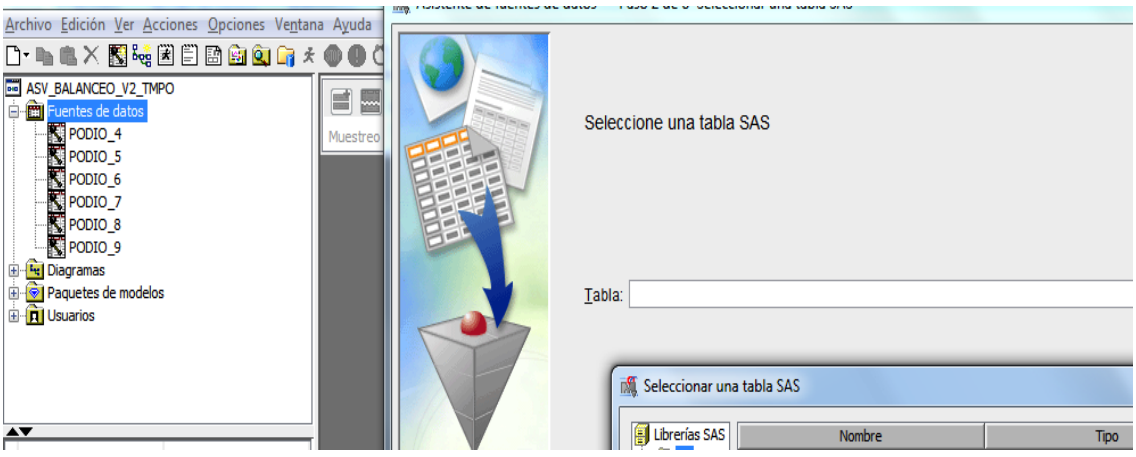
Al ser muchos escenarios, se decidió separar en dos proyectos de SAS, que son los dos primeros que se encuentran en el cuadro anterior, llamados ASV_BALANCEO_V2 Y ASV_BALANCEO_V2_TMPO.

La ruta donde residen los proyectos se especifica en la misma pantalla.
Por otro lado, en la biblioteca donde residen los metadatos, se define la opción de Menú: Archivo, biblioteca y se especifica la ruta (Cuadro 2.7.3.4.2)



Cuadro 2.7.3.4.2

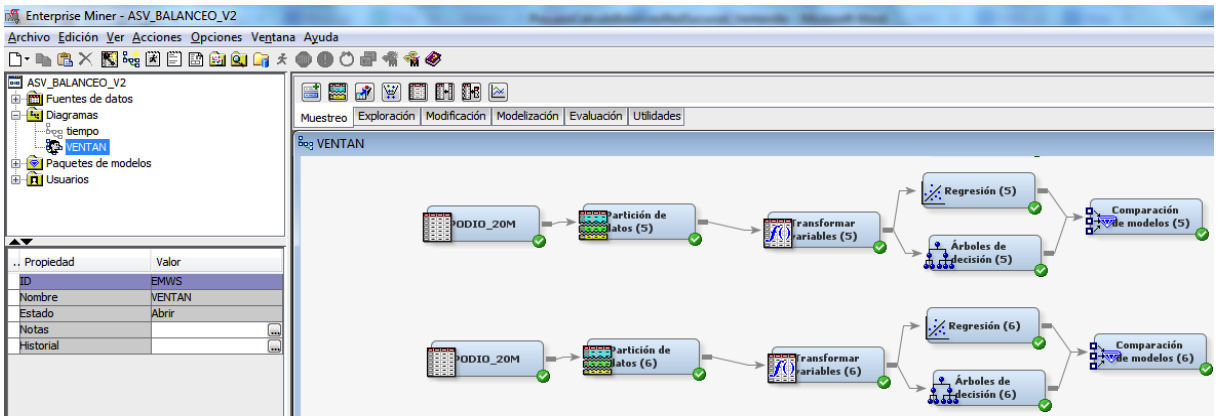
Para agregar la base para iniciar el modelaje(Cuadro 2.7.3.4.3), se deberá de crear una fuente de datos en la opción de fuente de datos, con clic derecho y crear fuente de datos, examinar y se selecciona la librería previamente generada, posteriormente se mostrarán las tablas de esa librería para elegir la base final.



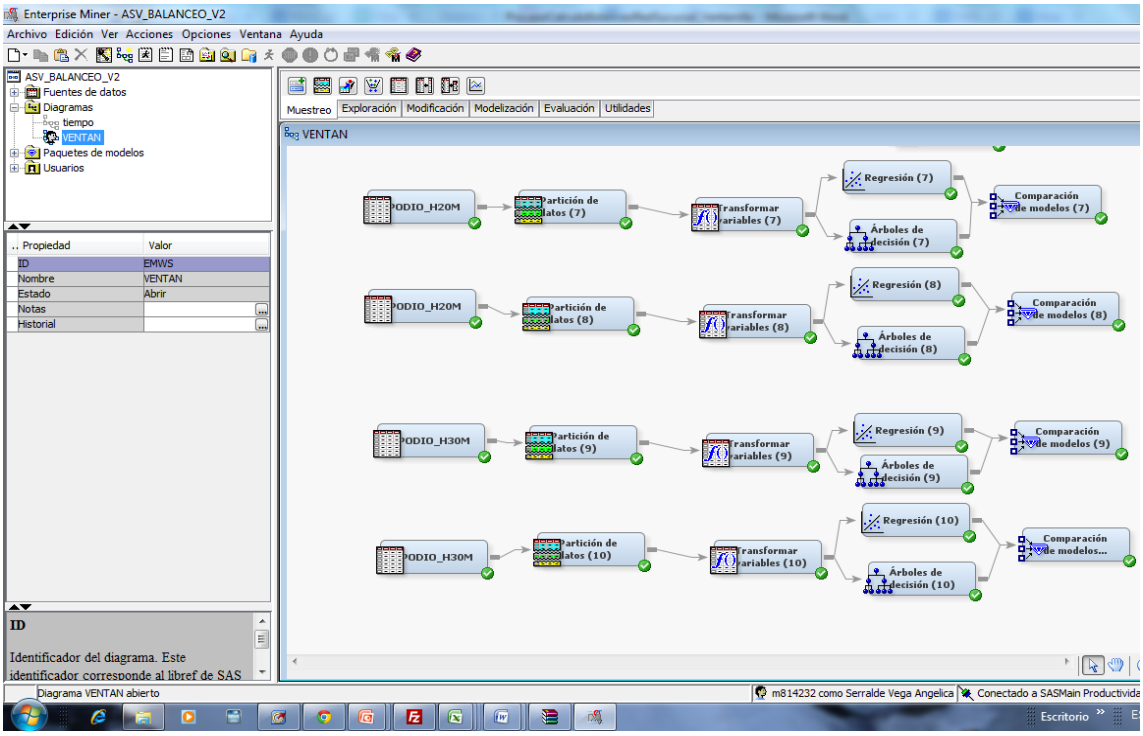
Cuadro 2.7.3.4.3

Posteriormente se agrega un diagrama nuevo.

Para iniciar en el diagrama se van agregando los nodos del proceso, sólo con arrastrar el icono deseado dentro del diagrama. El diagrama de flujo del proyecto se presenta en las siguientes láminas, donde se evalúan los diferentes escenarios y la información que se va a utilizar en cada escenario. El árbol del modelo de minería completo se presenta en los cuadros 2.7.3.4.4 al 2.7.3.4.7.

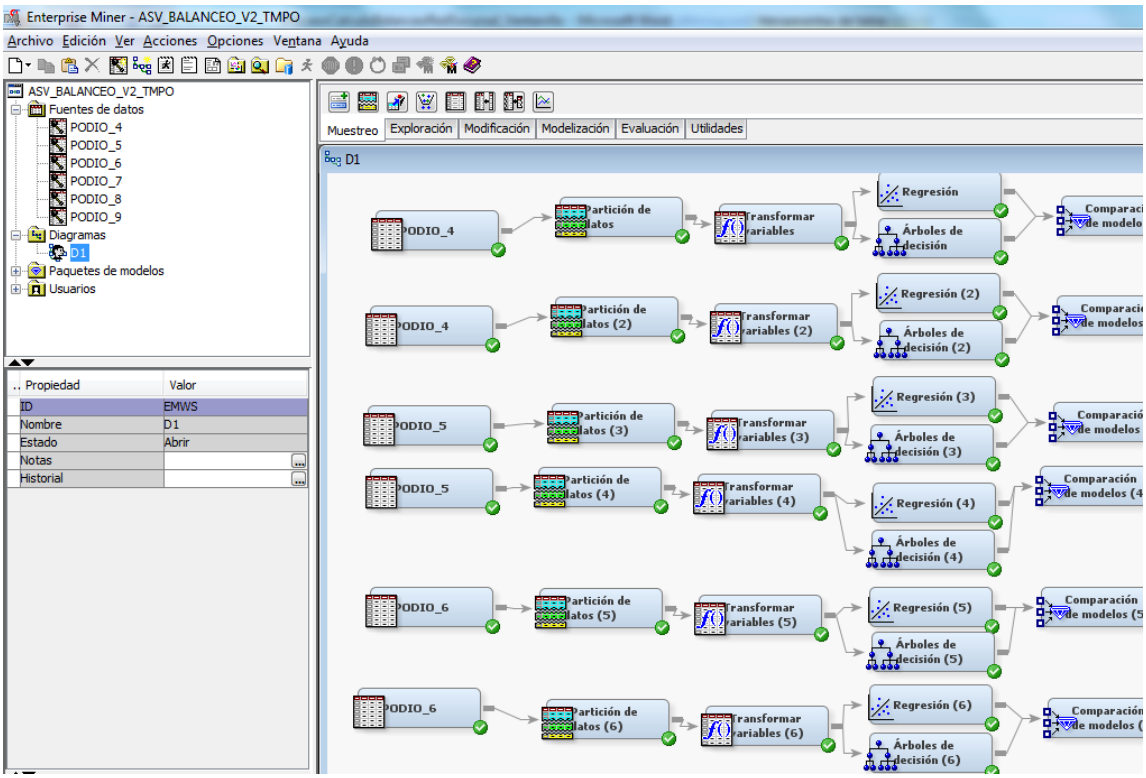


Cuadro 2.7.3.4.4

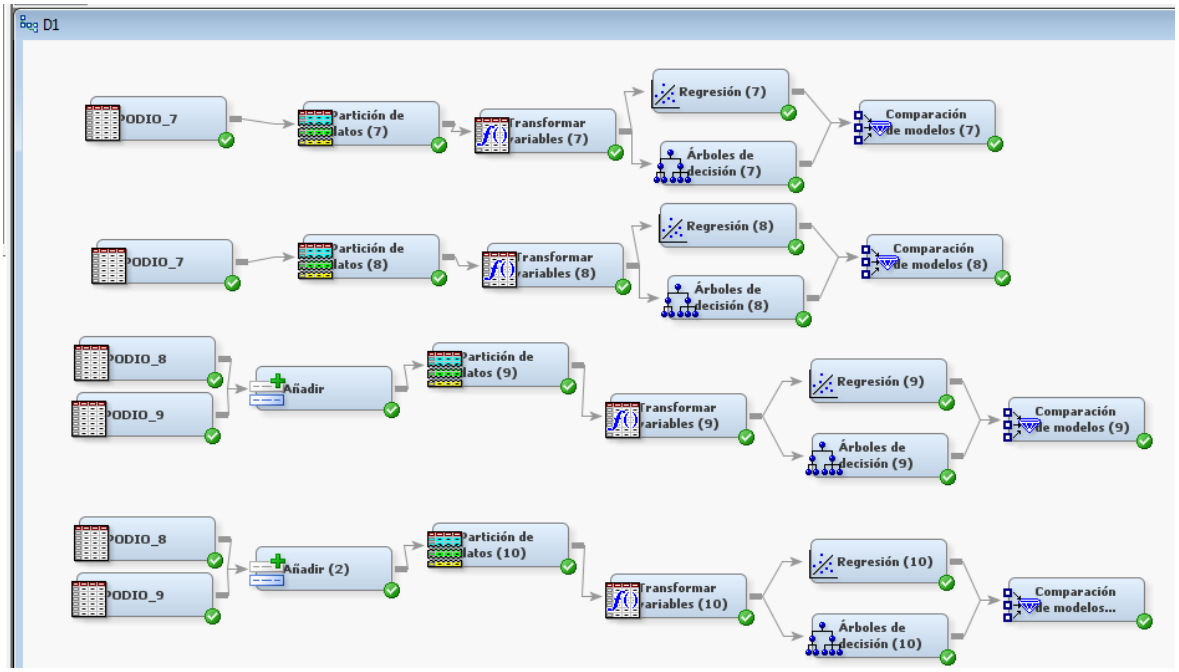


Cuadro 2.7.3.4.5

Escenarios de tiempo (Cuadro 2.7.3.4.6 y 2.7.3.4.7)



Cuadro 2.7.3.4.6



Cuadro 2.7.3.4.7

2.7.3.5 Parámetros de cada nodo

A continuación se presentan los nodos utilizados en los diagramas de los diferentes escenarios, en el entendido de que no cambian para ningún escenario, así también sólo se muestran los valores que se cambiaron, el resto permanece con los valores default de la herramienta.

Cuadro 2.7.3.5.1

Nodo	Descripción	Parámetros	Valor
	Criterios de participación de las muestras	% De muestra para entrenamiento % De muestra para validación % De muestra para prueba	50% 50% 0%
	Transformación de datos	Saldo Promedio	Método aplicado: Log
	Modelo de estimación	Tipo de regresión: Modelo de regresión:	Lineal Progresivo[22]
	Modelo de árbol de decisión	Sin cambio en ningún parámetro: Criterio de intervalo: Criterio Nominal Criterio ordinal	 Prob. F ProbChiSq Entropía
	Comparativa de los modelos	Sin cambio en ningún parámetro: Criterio seleccionado	Escenario de datos de validación: Average Squared Error
	Une bases de datos	Sin cambios en ningún parámetro	

2.7.4 Interpretación del modelo

A continuación se presenta un resumen de los resultados de los modelos de todos los escenarios con base a las salidas que genera el proceso de comparación de modelos.

[22] El modelo lineal progresivo (Forward) en SAS, parte de un modelo sin ninguna variable y de forma sistemática añade o excluye variables hasta que se cumplan los criterios de entrada y salida especificados (nivel de significancia, número de pasos y /o forzar que exista un número de variables).

Cuadro 2.7.4.1					
PROYECTO	NODO	DATOS[23]	ESCENARIO	Adj R-Sq	ERROR
BALANCEO V2	Árbol 5	Con Plantilla	Tiempo de servicio de 18 a 20 Min		0.44806
	Regresión 5	Con Plantilla		0.7504	0.43563
BALANCEO V2	Árbol 6	Sin plantilla			0.53541
	Regresión 6	Sin Plantilla		0.6647	0.61516
BALANCEO V2	Árbol 7	Con Plantilla	Tiempo de servicio de hasta 20 minutos		0.57154
	Regresión 7	Con Plantilla		0.738	0.65035
BALANCEO V2	Árbol 8	Sin plantilla			0.6698
	Regresión 8	Sin Plantilla		0.6589	0.82754
BALANCEO V2	Árbol 9	Con Plantilla	Tiempo de servicio de hasta 30 minutos		0.55174
	Regresión 9	Con Plantilla		0.7361	0.63247
BALANCEO V2	Árbol 10	Sin plantilla			0.64808
	Regresión 10	Sin Plantilla		0.656	0.82089
BALANCEO V2 TMPO	Árbol	Con Plantilla	Tiempo de servicio de 15 a 20 minutos		0.49484
	Regresión	Con Plantilla		0.7627	0.50283
BALANCEO V2 TMPO	Árbol 2	Sin plantilla			0.56294
	Regresión 2	Sin Plantilla		0.7627	0.68875
BALANCEO V2 TMPO	Árbol 3	Con Plantilla			0.45121
	Regresión 3	Con Plantilla		0.7725	0.46240
BALANCEO V2 TMPO	Árbol 4	Sin plantilla			0.4815
	Regresión 4	Sin Plantilla		0.6883	0.65093
BALANCEO V2 TMPO	Árbol 5	Con Plantilla	Tiempo de servicio de 25 a 30 minutos		0.46205
	Regresión 5	Con Plantilla		0.7924	0.47079

[23] Esta columna refiere a incluir o no en el análisis, a la plantilla que trabaja en la sucursal: Cajeros Back, front y ejecutivos

Cuadro 2.7.4.1					
PROYECTO	NODO	DATOS[23]	ESCENARIO	Adj R-Sq	ERROR
BALANCEO V2 TMPO	Árbol 6	Sin plantilla			0.55729
	Regresión 6	Sin Plantilla		0.7144	0.66531
BALANCEO V2 TMPO	Árbol 7	Con Plantilla	Tiempo de servicio de 30 a 35 minutos		0.49068
	Regresión 7	Con Plantilla		0.7879	0.46388
BALANCEO V2 TMPO	Árbol 8	Sin plantilla			0.49068
	Regresión 8	Sin Plantilla			0.46388
BALANCEO V2 TMPO	Árbol 9	Con Plantilla	Tiempo de servicio de 35 a 45 minutos		0.51106
	Regresión 9	Con Plantilla		0.7707	0.46752
BALANCEO V2 TMPO	Árbol 10	Sin plantilla			0.51221
	Regresión 10	Sin Plantilla			0.57006

Como se muestra en el cuadro 2.7.4.1, el objetivo del proyecto con escenario de, hasta 20 minutos, tiene un coeficiente de determinación $r^2 = 0.73$ en la regresión.
El árbol de decisión bajo este mismo escenario tiene menor error que la regresión (0.57154 contra 0.65035). Si se elige considerando este indicador, el modelo sería el árbol de decisión.

A continuación, se analizarán las salidas por deciles por rangos de predicción. En el caso del árbol de decisión, se observa una limitación al calcular como máximo número de ventanillas (8.7). Por otro lado, evaluando los rangos de la predicción vs lo observado, se ve que en la regresión donde existe un mayor diferencial de los valores predichos contra los valores observados es en los valores extremos. Esto significa que existe menor impacto por volumen, a diferencia del árbol.

Finalmente, como el diferencial de error cuadrático medio entre ambos modelos no es considerable, se elige a la regresión como el mejor modelo.

Cuadro 2.7.4.1.2					
SALIDA DEL ARBOL DE DECISION					
Rango predicho	Media de la variable objetivo	Media predicha	Número de observaciones	Puntuación del modelo	Diferencial
8.544 - 8.914	8.914	8.914	313	8.729	-0.1851

Cuadro 2.7.4.1.2					
SALIDA DEL ARBOL DE DECISION					
Rango predicho	Media de la variable objetivo	Media predicha	Número de observaciones	Puntuación del modelo	Diferencial
8.173 - 8.544	.	.	0	8.359	
7.803 - 8.173	8.169	8.169	498	7.988	-0.1803
7.433 - 7.803	7.797	7.797	182	7.618	-0.1784
7.063 - 7.433	7.277	7.277	505	7.248	-0.0291
6.693 - 7.063	6.748	6.748	1,925	6.878	0.1300
6.323 - 6.693	6.591	6.591	308	6.508	-0.0830
5.953 - 6.323	6.028	6.028	938	6.138	0.1100
5.583 - 5.953	5.736	5.736	948	5.768	0.0313
5.212 - 5.583	5.466	5.466	352	5.397	-0.0684
4.842 - 5.212	4.892	4.892	259	5.027	0.1355
4.472 - 4.842	4.813	4.813	2,072	4.657	-0.1555
4.102 - 4.472	4.372	4.372	3,064	4.287	-0.0853
3.732 - 4.102	3.966	3.966	6,141	3.917	-0.0490
3.362 - 3.732	3.621	3.621	5,489	3.547	-0.0742
2.992 - 3.362	3.138	3.138	7,272	3.177	0.0392
2.622 - 2.992	2.765	2.765	2,596	2.807	0.0416
2.251 - 2.622	2.295	2.295	78	2.436	0.1416
1.881 - 2.251	2.014	2.014	793	2.066	0.0524
1.511 - 1.881	1.659	1.659	167	1.696	0.0375

Cuadro 2.7.4.1.3					
SALIDA DE LA REGRESION					
Rango Predicho	Media de la variable objetivo	Media predicha	Número de observaciones	Puntuación del modelo	Diferencial
12.202 - 12.786	8.000	12.786	1	12.494	-0.2920
11.618 - 12.202	.	.	0	11.910	
11.033 - 11.618	7.500	11.411	2	11.326	-0.0859
10.449 - 11.033	8.833	10.656	6	10.741	0.0859
9.865 - 10.449	7.750	10.062	8	10.157	0.0950
9.281 - 9.865	8.200	9.492	35	9.573	0.0814
8.697 - 9.281	8.167	8.934	144	8.989	0.0549
8.113 - 8.697	7.924	8.385	329	8.405	0.0202
7.529 - 8.113	7.753	7.813	481	7.821	0.0077
6.945 - 7.529	7.422	7.224	650	7.237	0.0133
6.361 - 6.945	7.033	6.638	1,003	6.653	0.0146
5.776 - 6.361	6.424	6.052	1,447	6.069	0.0162
5.192 - 5.776	5.704	5.468	1,851	5.484	0.0166
4.608 - 5.192	4.756	4.862	3,434	4.900	0.0383
4.024 - 4.608	4.156	4.295	5,965	4.316	0.0209
3.440 - 4.024	3.647	3.728	7,028	3.732	0.0046
2.856 - 3.440	3.212	3.156	7,359	3.148	-0.0076
2.272 - 2.856	2.747	2.624	3,355	2.564	-0.0599
1.688 - 2.272	2.051	2.089	760	1.980	-0.1087
1.104 - 1.688	1.476	1.519	42	1.396	-0.1234

El modelo elegido para estimar el número de ventanillas dado un nivel de servicio máximo de 20 minutos, es la regresión.
Una de formas de medir la calidad del modelo es a partir de la descomposición de la suma de cuadrados de la variable respuesta en dos componentes ortogonales, una debida al ajuste del modelo y otra residual. Donde la longitud residual es mínima y el ajuste del modelo de longitud máxima.

Esta descomposición se interpreta como la descomposición de la variabilidad total de la variable de respuesta en variabilidad explicada por el modelo y variabilidad residual.

Esta descomposición se presenta en forma de tabla con el nombre de análisis de varianza (cuadro 2.7.4.1.4)
En este caso, la suma de cuadrados total de la variable de respuesta, es de 83129, de las cuales 61374 son explicadas por el modelo y quedan sin explicar 21755.
La suma de cuadrados medios, es resultado de la división de la suma de cuadrados medios entre los grados de libertad. Donde el valor obtenido (0.642386) es la estimación de la varianza del error real.

El valor F, es el cociente del cuadrado medio del modelo y el cuadrado medio del error. Se prueba la bondad del modelo en su conjunto (ajustado por la media) que representa el comportamiento de la variable dependiente. Esta prueba es una prueba de la hipótesis nula de que todos los parámetros excepto la intersección son cero. La prueba en general es significativa al obtener un p-valor menor que 0.0001.
El coeficiente de determinación ajustado tiene en cuenta la complejidad del modelo, al tener en cuenta el número de variables que lo componen. Por lo tanto es el más adecuado para medir la bondad del modelo, en este caso con un valor de 0.738.

Cuadro 2.7.4.1.4
Análisis de varianza

Fuente	DF	Suma de cuadrados	Cuadrados medios	F Value	Pr > F
Model	33	61374	1859.80915	2895.16	<.0001
Error	33866	21755	0.642386		
Corrected Total	33899	83129			

Con una r^2 de 0.783
Y una r^2 Ajustada = 0.738

2.7.5. Interpretación y contextualización

2.7.5.1 Las variables significativas y su coeficiente

Los valores de los coeficientes de la regresión del modelo elegido se muestran en el siguiente cuadro (Cuadro 2.7.5.1.1):

Cuadro 2.7.5.1.1
Analysis of Maximum Likelihood Estimates

Parámetro		DF	Estimador	Error estándar	t Value	Pr > |t|
Intercept		**1**	**-3.8894**	**0.1681**	**-23.14**	**<.0001**
HORARIO	**08:00**	**1**	**0.9742**	**0.0643**	**15.14**	**<.0001**

Cuadro 2.7.5.1.1						
Analysis of Maximum Likelihood Estimates						
Parámetro		DF	Estimador	Error estándar	t Value	Pr > |t|
HORARIO	08:30	1	0.6104	0.0539	11.32	<.0001
HORARIO	09:00	1	0.344	0.054	6.37	<.0001
HORARIO	09:30	1	0.3452	0.054	6.39	<.0001
HORARIO	10:00	1	0.3095	0.0541	5.72	<.0001
HORARIO	10:30	1	0.3126	0.0542	5.77	<.0001
HORARIO	11:00	1	0.2858	0.0544	5.25	<.0001
HORARIO	11:30	1	0.3116	0.0545	5.72	<.0001
HORARIO	12:00	1	0.3519	0.0546	6.45	<.0001
HORARIO	12:30	1	0.3746	0.0546	6.87	<.0001
HORARIO	13:00	1	0.4081	0.0545	7.48	<.0001
HORARIO	13:30	1	0.3996	0.0546	7.33	<.0001
HORARIO	14:00	1	0.4196	0.0546	7.69	<.0001
HORARIO	14:30	1	0.3941	0.0547	7.21	<.0001
HORARIO	15:00	1	0.4175	0.0547	7.63	<.0001
HORARIO	15:30	1	0.4083	0.055	7.42	<.0001
HORARIO	16:00	1	0.1324	0.0556	2.38	0.0172
HORARIO	16:30	1	-0.7077	0.1645	-4.3	<.0001
HORARIO	17:00	1	-0.8631	0.2752	-3.14	0.0017
HORARIO	17:30	1	-1.5982	0.4444	-3.6	0.0003
HORARIO	18:00	1	-2.3716	0.7662	-3.1	0.002
LOG_saldo_promedio		1	0.2668	0.0115	23.18	<.0001
MES_AP_UL		1	0.015	0.000765	19.63	<.0001
SUM_TXS		1	0.0283	0.000302	93.79	<.0001
cu_a		1	0.2556	0.0121	21.18	<.0001
cu_b		1	0.3494	0.00382	91.57	<.0001
cu_pt		1	0.3504	0.0101	34.54	<.0001
ejecutivos		1	0.049	0.00474	10.33	<.0001
tipo_suc	A	1	0.165	0.011	15.04	<.0001
tipo_suc	B	1	0.1549	0.00972	15.93	<.0001
tipo_suc	C	1	0.00281	0.0116	0.24	0.8093
tipo_suc	NUEVA	1	0.1764	0.0162	10.87	<.0001
total_at		1	0.086	0.00377	22.84	<.0001

Si bien es cierto, que el horario de 16:00 hrs. en adelante se encuentra en la base de datos, el horario de operación al público es hasta las 16:00 horas[24]. Por lo tanto se puede omitir los resultados posteriores al horario de servicio general.

Una de las variables más relevantes que se obtuvieron en el modelo, es el horario, esto se confirma con la operativa misma de la sucursal, pues actualmente se tienen definidos días y horarios pico y valle.

Otra variable relevante son los saldos promedio que se manejan en la bóveda. Esta variable explica si la sucursal maneja grandes cantidades de efectivo y por ende más transacciones que deben ser atendidas en ventanilla.

2.7.5.2. La salida de estimación

El resultado de la estimación, es el número de ventanillas que se necesita cada media hora durante el periodo de tiempo observado, donde se puede tener para cada sucursal 1,760 observaciones y propuestas de ventanilla.

Para efectos prácticos, el valor estimado de la ventanilla se redondea con el siguiente valor entero a partir de 0.5. Un ejemplo de una estimación para la sucursal 004, se presenta en el cuadro 2.7.5.2.1, donde se presentan los resultados de todos los escenarios de tiempo obtenidos.

Cuadro 2.7.5.2.1

CONCEPTO	VALORES						
FECHA	**1/10/14**	**1/10/14**	**1/10/14**	**1/10/14**	**2/10/14**	**2/10/14**	**2/10/14**
CR (Sucursal)	**0004**	**0004**	**0004**	**0004**	**0004**	**0004**	**0004**
TIPO_SUC	**C**	**C**	**C**	**C**	**C**	**C**	**C**
RANGO HORARIO	**14:00 a 14:29**	**14:30 a 14:59**	**15:00 a 15:29**	**15:30 a 15:59**	**09:00 a 09:29**	**09:30 a 09:59**	**10:00 a 10:29**
V_VENT_15_20M_REG_CP	**3**	**3**	**3**	**3**	**2**	**2**	**3**
V_VENT_15_20M_REG_SP	**3**	**4**	**4**	**3**	**3**	**3**	**3**
V_VENT_20M_REG_SP	**3**	**4**	**4**	**3**	**3**	**3**	**3**
V_VENT_20M_REG_CP	**3**	**3**	**3**	**3**	**2**	**3**	**3**
V_VENT_20_25M_R	**4**	**4**	**4**	**3**	**3**	**3**	**3**

[24] Existen sucursales que por temas de servicio a Corporativos y apoyo a cobranza, tienen horarios posteriores a las 16:00 horas. Los cajeros con este horario extendido es a lo más dos por sucursal.

Cuadro 2.7.5.2.1							
CONCEPTO	VALORES						
EG_SP							
V_VENT_20_25M_REG_CP	**3**	**3**	**3**	**3**	**2**	**2**	**3**
V_VENT_25_30M_REG_SP	**4**	**4**	**4**	**4**	**3**	**3**	**4**
V_VENT_25_30M_REG_CP	**3**	**3**	**3**	**3**	**3**	**2**	**3**
V_VENT_30_35M_REG_SP	**3**	**3**	**3**	**3**	**2**	**3**	**3**
V_VENT_30_35M_REG_CP	**3**	**3**	**3**	**3**	**2**	**3**	**3**
V_VENT_35_45M_REG_SP	**4**	**4**	**4**	**4**	**3**	**3**	**3**
V_VENT_35_45M_REG_CP	**3**	**3**	**4**	**3**	**3**	**3**	**3**
V_VENT_H20Mn_ARB_CP	**4**	**4**	**4**	**4**	**3**	**3**	**3**
V_VENT_H20Mn_REG_CP	**3**	**3**	**4**	**3**	**3**	**3**	**3**
V_VENT_H20Mn_ARB_SP	**4**	**4**	**4**	**4**	**3**	**3**	**4**
V_VENT_H20Mn_REG_SP	**4**	**4**	**4**	**4**	**3**	**3**	**3**
V_VENT_H30M_REG_CP	**3**	**3**	**4**	**3**	**2**	**3**	**3**
V_VENT_H30M_REG_SP	**3**	**4**	**4**	**4**	**3**	**3**	**3**
V_VENT_15_20M_ARB_CP	**4**	**4**	**4**	**4**	**3**	**3**	**4**
V_VENT_15_20M_ARB_SP	**4**	**4**	**4**	**4**	**4**	**4**	**4**
V_VENT_20M_ARB_SP	**3**	**4**	**4**	**3**	**3**	**3**	**3**
V_VENT_20M_ARB_CP	**4**	**4**	**4**	**4**	**4**	**4**	**4**
V_VENT_20_25M_ARB_SP	**4**	**3**	**3**	**4**	**4**	**4**	**4**
V_VENT_20_25M_ARB_CP	**3**	**3**	**4**	**3**	**3**	**3**	**3**
V_VENT_25_30M_ARB_SP	**4**	**4**	**4**	**4**	**3**	**3**	**4**
V_VENT_25_30M_ARB_CP	**4**	**4**	**4**	**4**	**3**	**3**	**3**
V_VENT_30_35M_A	**3**	**3**	**3**	**3**	**3**	**3**	**3**

Cuadro 2.7.5.2.1							
CONCEPTO	VALORES						
RB_SP							
V_VENT_30_35M_ARB_CP	**3**	**3**	**3**	**3**	**3**	**3**	**3**
V_VENT_35_45M_ARB_SP	**4**	**4**	**4**	**4**	**3**	**3**	**3**
V_VENT_35_45M_ARB_CP	**3**	**3**	**3**	**3**	**3**	**3**	**3**
V_VENT_H30M_ARB_CP	**4**	**4**	**4**	**4**	**4**	**4**	**4**
V_VENT_H30M_ARB_SP	**4**	**4**	**3**	**4**	**3**	**3**	**4**

Una vez teniendo los resultados estimados de las ventanillas de acuerdo a la transaccionalidad observada de cada media hora, la metodología para definir las ventanillas se decidirá a nivel sucursal, considerando la media, mediana y el quintil 3 de las observaciones. La decisión de estimar a nivel sucursal, es porque hay mucha variabilidad entre sucursales y de englobar esta información, se perdería la heterogeneidad de cada una.

A continuación se presentarán dos ejemplos de sucursales con comportamiento diferente. Para presentar los resultados se utilizará la opción de "Análisis de capacidad" de la herramienta SAS[26].

2.7.5.3. Ejemplo de Sucursales y análisis de resultados

Análisis de capacidad de: V_VENT_H20Mn_REG_CP
The CAPABILITY Procedure

Variable: V_VENT_H20Mn_REG_CP
CR=0021

Moments			
N	3760	**Sum Weights**	3760
Mean	3.91914894	**Sum Observations**	14736
Std Deviation	0.55441378	**Variance**	0.30737464
Skewness	0.37835507	**Kurtosis**	2.24566945
Uncorrected SS	58908	**Corrected SS**	1155.42128
Coeff Variation	14.14628	**Std Error Mean**	0.00904149

Basic Statistical Measures			
Location		**Variability**	
Mean	3.919149	**Std Deviation**	0.55441
Median	4	**Variance**	0.30737
Mode	4	**Range**	4
		Interquartile Range	0

Cuadro 2.7.5.3.1

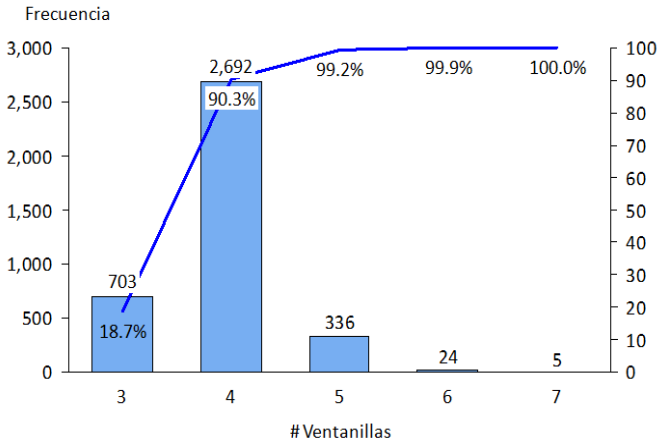
[26] La forma de uso de la herramienta, se presentan en el anexo III de este mismo documento.

Variable a analizar: Estimado de número de ventanillas modelo hasta 20 minutos de nivel de servicio regresión: V_VENT_H20Mn_REG_CP

Quantiles (Definition 5)								
Quantile	Estimate	95% Confidence Limits Assuming Normality		95% Confidence Limits Distribution Free		Order Statistics		
						LCL Rank	UCL Rank	Coverage
100% Max	7							
99%	5	5.1754618	5.2437262	5	6	3,711	3,735	95.14
95%	5	4.8043775	4.8587669	5	5	3,546	3,599	95.28
90%	4	4.6061238	4.6539743	4	5	3,348	3,421	95.28
75% Q3	4	4.2736665	4.3129475	4	4	2,768	2,873	95.20
50% Median	4	3.9014222	3.9368756	4	4	1,820	1,941	95.16
25% Q1	4	3.5253504	3.5646314	4	4	888	993	95.20
10%	3	3.1843236	3.2321740	3	3	340	413	95.28
5%	3	2.9795309	3.0339203	3	3	162	215	95.28
1%	3	2.5945717	2.6628361	3	3	26	50	95.14
0% Min	3							

Cuadro 2.7.5.3.2

Y su gráfica de distribución de ventanillas:



Gráfica 2.7.5.3.3

Como se puede observar en los estadísticos y gráficas, esta sucursal tiene poca variabilidad en los resultados. La media y moda son 4 ventanillas, con valor máximo de 7. El 90% de la distribución acumulada se obtiene en 4 ventanillas.

Si se toma otro ejemplo de sucursal, en este caso la 0061, y se obtienen mismos estadísticos, se observa mayor variabilidad y donde el 90% de los casos llega a 9 ventanilla, y siendo la media de 7 ventanillas.

Análisis de capacidad de: V_VENT_H20Mn_REG_CP
The CAPABILITY Procedure

Variable: V_VENT_H20Mn_REG_CP
CR=0061

Moments			
N	3765	Sum Weights	3765
Mean	7.40159363	Sum Observations	27867
Std Deviation	0.91782305	Variance	0.84239916
Skewness	0.5212129	Kurtosis	0.58858518
Uncorrected SS	209431	Corrected SS	3170.79044
Coeff Variation	12.4003438	Std Error Mean	0.0149581

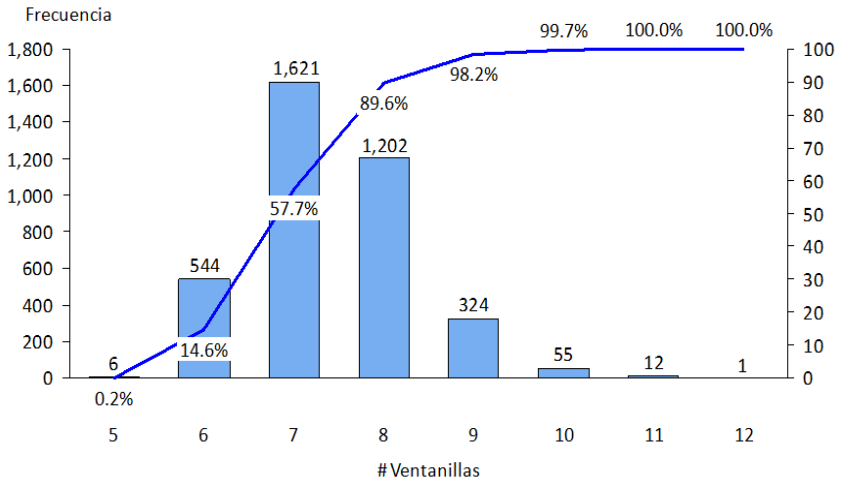
Basic Statistical Measures			
Location		Variability	
Mean	7.401594	Std Deviation	0.91782
Median	7	Variance	0.8424
Mode	7	Range	7
		Interquartile Range	1

Cuadro 2.7.5.3.4

Quantiles (Definition 5)								
Quantile	Estimate	95% Confidence Limits Assuming Normality		95% Confidence Limits Distribution Free		Order Statistics		
						LCL Rank	UCL Rank	Cover
100% Max	12							
99%	10	9.481	9.594	10	10	3,716	3,740	95.12
95%	9	8.867	8.957	9	9	3,551	3,604	95.25
90%	9	8.539	8.618	8	9	3,353	3,426	95.26
75% Q3	8	7.989	8.053	8	8	2,772	2,877	95.18
50% Median	7	7.372	7.431	7	7	1,823	1,944	95.14
25% Q1	7	6.750	6.815	7	7	889	994	95.18
10%	6	6.185	6.264	6	6	340	413	95.26
5%	6	5.846	5.936	6	6	162	215	95.25
1%	6	5.209	5.322	6	6	26	50	95.12
0% Min	5							

Cuadro 2.7.5.3.5

En la gráfica se observan los valores que toman los datos.



Gráfica 2.7.5.3.6

Es importante obtener la variación por horario, pues también los días y el horario pico pueden requerir mayor capacidad de atención en ventanilla, como se puede ver en los siguientes cuadros:

Cuadro 2.7.5.3.7. CR 9221						
The MEANS Procedure						
Variable de análisis: V_VENT_H20Mn_REG_CP						
HORARIO	N Obs	Media	Mínimo	Máximo	90th Pctl	95th Pctl
09:00	23	2.5217	2	3	3	3
09:30	23	2.6522	2	3	3	3
10:00	23	2.7391	2	3	3	3
10:30	23	2.7826	2	3	3	3
11:00	23	2.7391	2	3	3	3
11:30	23	2.6087	2	3	3	3
12:00	23	2.8261	2	3	3	3
12:30	23	2.8261	2	4	3	3
13:00	23	3.0435	3	4	3	3
13:30	23	2.8696	2	4	3	3
14:00	23	2.9565	2	4	3	3
14:30	23	3.0000	3	3	3	3
15:00	23	3.0870	3	4	3	4
15:30	23	3.1739	3	4	4	4
16:00	20	2.0500	2	3	2	2.5
17:00	1	1.0000	1	1	1	1

Cuadro 2.7.5.3.8.	
CR 9221	
Cuantil	Estimado
100% Max	4
99%	4
95%	3
90%	3
75% Q3	3
50% Mediana	3
25% Q1	3
10%	2
5%	2
1%	2
0% Min	1

Cuadro 2.7.5.3.9						
CR 9961 The MEANS Procedure						
Variable de análisis: V_VENT_H20Mn_REG_CP						
Horario	N Obs	Media	Mínimo	Máximo	90th Pctl	95th Pctl
09:00	27	6.7037	5	9	7	8
09:30	27	7.2963	6	10	8	9
10:00	27	6.9259	6	9	8	9
10:30	27	7.1852	6	9	8	8
11:00	27	7.2593	6	9	9	9
11:30	27	7.5185	6	9	9	9
12:00	27	7.2593	6	9	8	8
12:30	27	7.7407	6	10	9	9
13:00	27	7.7407	6	10	9	9
13:30	27	7.8148	6	10	9	9
14:00	27	7.6667	6	9	9	9
14:30	27	7.8889	6	10	9	9
15:00	27	7.6667	5	10	9	9
15:30	27	8.1481	6	10	9	10
16:00	27	5.5185	5	9	7	7
16:30	5	4.2000	4	5	5	5
17:00	1	4.0000	4	4	4	4
17:30	1	3.0000	3	3	3	3

Cuadro 2.7.5.3.10	
(CR 9961) Percentiles	
Cuantil	Estimado
100% Max	10
99%	10
95%	9
90%	9
75% Q3	8
50% Mediana	7
25% Q1	7
10%	6
5%	5
1%	4
0% Min	3

En el siguiente cuadro se muestran los resultados de la estimación de las ventanillas considerando los diferentes horarios de servicio de la sucursal. Se observa es que no hay gran diferencia de la media entre el alcance de ventanillas considerando la mediana, media y hasta el 75% de los horarios, también su desviación estándar es muy constante, lo que nos hace concluir que son pocos días donde la sucursal se satura y alcanza niveles máximos.

Por lo tanto, el alcance definido para determinar el número de ventanillas por horario, sería la mediana.

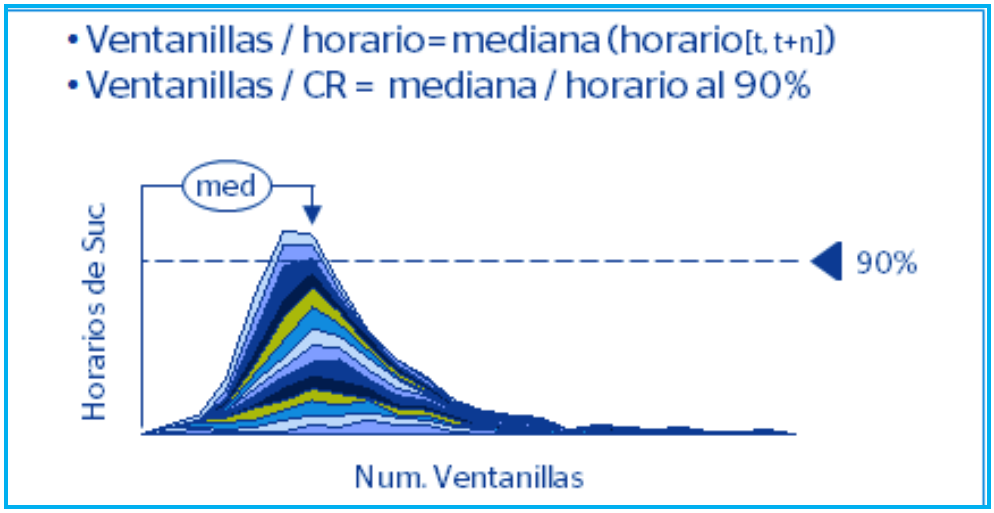
Cuadro 2.7.5.3.11							
Resultados de número de ventanillas con valores de la Mediana, Media, 75%, quintil 90% y máximo por horario							
Horario	N Obs	Variable	N	Media	Desv. Est.	Mín.	Máx.
08:30	1653	Mediana_H20Mn_REG_CP	1,653	3.3191	1.4241	1	10.00
		media_H20Mn_REG_CP	1,653	3.3621	1.4139	1	9.87
		Q3_H20Mn_REG_CP	1,653	3.4982	1.4935	1	
		Prct90_H20Mn_REG_CP	1,653	3.7178	1.5776	1	10.00
		MAX_H20Mn_REG_CP	1,653	4.1845	1.7482	1	12.00
09:00	1758	Mediana_H20Mn_REG_CP	1,758	3.0333	1.3853	1	10.00
		media_H20Mn_REG_CP	1,758	3.0624	1.3754	1	9.89
		Q3_H20Mn_REG_CP	1,758	3.2079	1.4429	1	10.00
		Prct90_H20Mn_REG_CP	1,758	3.3811	1.5161	1	11.00
		MAX_H20Mn_REG_CP	1,758	3.8544	1.7105	1	11.00
09:30	1762	Mediana_H20Mn_REG_CP	1,762	3.1512	1.4304	1	10.00

Cuadro 2.7.5.3.11							
Resultados de número de ventanillas con valores de la Mediana, Media, 75%, quintil 90% y máximo por horario							
Horario	N Obs	Variable	N	Media	Desv. Est.	Mín.	Máx.
		media_H20Mn_REG_CP	1,762	3.1811	1.4252	1	10.13
		Q3_H20Mn_REG_CP	1,762	3.3473	1.5023	1	10.00
		Prct90_H20Mn_REG_CP	1,762	3.5627	1.5839	1	11.00
		MAX_H20Mn_REG_CP	1,762	4.0919	1.7898	1	12.00
10:00	1763	Mediana_H20Mn_REG_CP	1,763	3.1554	1.4640	1	10.00
		media_H20Mn_REG_CP	1,763	3.1799	1.4510	1	10.19
		Q3_H20Mn_REG_CP	1,763	3.3503	1.5383	1	10.50
		Prct90_H20Mn_REG_CP	1,763	3.5377	1.6063	1	11.00
		MAX_H20Mn_REG_CP	1,763	4.1100	1.8166	1	12.00
10:30	1764	Mediana_H20Mn_REG_CP	1,764	3.2330	1.4937	1	10.00
		media_H20Mn_REG_CP	1,764	3.2575	1.4789	1	10.31
		Q3_H20Mn_REG_CP	1,764	3.4382	1.5663	1	11.00
		Prct90_H20Mn_REG_CP	1,764	3.6457	1.6344	1	11.00
		MAX_H20Mn_REG_CP	1,764	4.2205	1.8376	1	12.00
11:00	1763	Mediana_H20Mn_REG_CP	1,763	3.2535	1.5242	1	11.00
		media_H20Mn_REG_CP	1,763	3.2738	1.5017	1	10.70
		Q3_H20Mn_REG_CP	1,763	3.4586	1.5930	1	11.00
		Prct90_H20Mn_REG_CP	1,763	3.6727	1.6557	1	11.00
		MAX_H20Mn_REG_CP	1,763	4.2541	1.8759	1	12.00
11:30	1763	Mediana_H20Mn_REG_CP	1,763	3.3182	1.5441	1	11.00
		media_H20Mn_REG_CP	1,763	3.3479	1.5298	1	11.27
		Q3_H20Mn_REG_CP	1,763	3.5352	1.6056	1	12.00
		Prct90_H20Mn_REG_CP	1,763	3.7760	1.7117	1	12.00
		MAX_H20Mn_REG_CP	1,763	4.3364	1.9034	1	13.00
12:00	1764	Mediana_H20Mn_REG_CP	1,764	3.4022	1.5756	1	11.00
		media_H20Mn_REG_CP	1,764	3.4357	1.5590	1	11.42
		Q3_H20Mn_REG_CP	1,764	3.6287	1.6429	1	12.00
		Prct90_H20Mn_REG_CP	1,764	3.8747	1.7393	1	12.00
		MAX_H20Mn_REG_CP	1,764	4.4654	1.9359	1	13.00
12:30	1764	Mediana_H20Mn_REG_CP	1,764	3.4646	1.5912	1	11.00
		media_H20Mn_REG_CP	1,764	3.4982	1.5798	1	11.49
		Q3_H20Mn_REG_CP	1,764	3.6987	1.6670	1	12.00
		Prct90_H20Mn_REG_CP	1,764	3.9507	1.7577	1	12.00
		MAX_H20Mn_REG_CP	1,764	4.5527	1.9697	1	14.00

Cuadro 2.7.5.3.11							
Resultados de número de ventanillas con valores de la Mediana, Media, 75%, quintil 90% y máximo por horario							
Horario	N Obs	Variable	N	Media	Desv. Est.	Mín.	Máx.
13:00	1763	Mediana_H20Mn_REG_CP	1,763	3.5320	1.6181	1	11.00
		media_H20Mn_REG_CP	1,763	3.5667	1.5993	1	11.54
		Q3_H20Mn_REG_CP	1,763	3.7703	1.6968	1	12.00
		Prct90_H20Mn_REG_CP	1,763	4.0425	1.7862	1	13.00
		MAX_H20Mn_REG_CP	1,763	4.6557	2.0006	1	14.00
13:30	1762	Mediana_H20Mn_REG_CP	1,762	3.5528	1.6224	1	12.00
		media_H20Mn_REG_CP	1,762	3.5905	1.6116	1	11.64
		Q3_H20Mn_REG_CP	1,762	3.8121	1.7119	1	12.00
		Prct90_H20Mn_REG_CP	1,762	4.0766	1.7986	1	12.00
		MAX_H20Mn_REG_CP	1,762	4.7032	2.0396	1	15.00
14:00	1763	Mediana_H20Mn_REG_CP	1,763	3.5871	1.6295	1	12.00
		media_H20Mn_REG_CP	1,763	3.6345	1.6239	1	11.63
		Q3_H20Mn_REG_CP	1,763	3.8576	1.7221	1	12.00
		Prct90_H20Mn_REG_CP	1,763	4.1305	1.8255	1	12.00
		MAX_H20Mn_REG_CP	1,763	4.7623	2.0356	1	13.00
14:30	1762	Mediana_H20Mn_REG_CP	1,762	3.5951	1.6483	1	12.00
		media_H20Mn_REG_CP	1,762	3.6415	1.6331	1	11.65
		Q3_H20Mn_REG_CP	1,762	3.8746	1.7328	1	12.00
		Prct90_H20Mn_REG_CP	1,762	4.1612	1.8431	1	13.00
		MAX_H20Mn_REG_CP	1,762	4.8059	2.0540	1	14.00
15:00	1764	Mediana_H20Mn_REG_CP	1,764	3.6264	1.6456	1	12.00
		media_H20Mn_REG_CP	1,764	3.6798	1.6380	1	11.70
		Q3_H20Mn_REG_CP	1,764	3.9229	1.7368	1	12.00
		Prct90_H20Mn_REG_CP	1,764	4.2007	1.8508	1	13.00
		MAX_H20Mn_REG_CP	1,764	4.8475	2.0855	1	14.00
15:30	1763	Mediana_H20Mn_REG_CP	1,763	3.6875	1.6610	1	12.00
		media_H20Mn_REG_CP	1,763	3.7246	1.6463	1	11.73
		Q3_H20Mn_REG_CP	1,763	3.9663	1.7520	1	12.00
		Prct90_H20Mn_REG_CP	1,763	4.2377	1.8560	1	13.00
		MAX_H20Mn_REG_CP	1,763	4.8871	2.1011	1	14.00

Ahora bien, dado que el impacto en la actividad de la sucursal diaria no es significativa y la variabilidad por horario es mayor a la observada por día, así como su impacto en la sucursal,

entonces, se propone cubrir la variabilidad del horario en un 90% de lo observado en el día, lo que implica que sólo se estará dejando fuera del nivel de servicio en promedio 48 minutos diarios.



Gráfica2.7.5.3.12

2.7.6. Difusión, uso y monitoreo

La difusión del proyecto es inmediata, pues dados los proyectos de expansión, actualización de infraestructura, modelos de gestión y perfiles del personal, requiere una actualización de la plantilla alineada a todos estos cambios.

El uso a este modelo se dará de manera gradual, en el momento que se planee la actualización de la infraestructura. Una vez terminado la modernización de las sucursales, el balanceo de sucursales, se realiza cada año.

Por otro lado, las sucursales ya modernizadas, se evaluarán y se dará prioridad en donde la propuesta del modelo sea mayor.
El monitoreo que se dará a las sucursales, es siguiendo mediante indicadores de productividad que se construyen mes a mes:

- Productividad del cajero:
- Transacciones promedio diarias
- Saturación de la sucursal
- Indisponibilidad del autoservicio

2.7.6.1 Resultados de expansión

Como resultado de evaluar el total de sucursales, el número de ventanillas resultantes es de, 6,427 ventanillas, en 1,767 sucursales.

Cuadro 2.7.6.1.1		
División	**# Sucursales**	**# Ventanillas Propuestas**
Bajío	213	809
Metro Norte[27]	232	888
Metro Sur[28]	227	925
Noreste	192	622
Noroeste	220	754
Occidente	117	411
Occidente I	193	636
Sur	213	808
Sureste	160	574
TOTAL	1,767	6,427

2.7.6.2 Uso del conocimiento obtenido

Como se planteó en el punto 2.1, la institución está en un periodo de transformación, en donde se actualizarán las sucursales en infraestructura y actividades del personal que labora en ellas. Es importante recalcar la fuerte dependencia de una y otra, pues una infraestructura nueva, sin el suficiente personal, no sirve, y demasiado personal, conlleva a gastos extras.

El personal que apoya a la sucursal, son el cajero universal A (Perfil más completo y que está capacitado para dar servicio en el back, autoservicio, front y direccionamiento), cajero universal B (Perfil junior, que está capacitado para dar servicio en el front y direccionamiento), y el Cajero Part, con el mismo perfil que el cajero universal B, pero laborando la mitad de los días.

Para usar el modelo de estimación de ventanillas, es necesario completarlo con el número de personas que darán servicio a la sucursal. Lo que se espera como resultado es una propuesta que cubra bajo el nivel de servicio deseado todas las actividades derivadas de las cargas de trabajo en la sucursal.

2.8 Asignación de Perfiles

Si se retoma el diagrama de capítulo 2.6.1, donde el objetivo es determinar el número de personas óptimo para la plantilla de las sucursales, considerando un nivel de servicio de hasta 20 minutos, entonces ya tendremos los resultados de cada área de la sucursal

[27] Metro Norte: Corresponde al Estado de México y delegaciones al norte del DF. (Gustavo A. Madero, Azcapotzalco y Venustiano Carranza).

[28] Corresponde al DF, excepto GA. Madero, Azcapotzalco y Venustiano Carranza

- **Actividades Back[29]:**
  - **A**ctividades de autoservicio: Total de tareas que se deben de realizar y tiempo promedio de cada actividad.
  - Actividades en caja principal: total de tareas que se deben de realizar y tiempo promedio de cada actividad.

Se calcula el FTE[6] sumando los tiempos promedio de las actividades. ✓

- **Actividades Front:**
  - Actividades de ventanilla para otorgar un nivel de servicio óptimo:
- **Cálculo de plantilla ideal:**
  - Consolidando los dos puntos anteriores y considerando los perfiles que deberán de cubrirse.

Modelo de minería de datos. ✓

Una vez que se obtuvieron el número de cajeros para las actividades del back y el número de ventanillas, el siguiente paso es determinar las personas que son necesarias para hacer frente a las cargas de trabajo en sucursal.

No necesariamente el número de ventanillas es igual al número de cajeros, pues la afluencia de clientes es variable de acuerdo al horario. Por otro lado, dadas las actividades que pueden realizar los cajeros universales A, se espera que apoye en ventanilla, una vez que sus actividades administrativas o de autoservicio concluyan.

Por lo tanto, el siguiente paso es determinar el perfil que se necesita dadas las transacciones que se hacen por cada media hora, la necesidad de un direccionador[30] permanente en la sucursal, la gestión del autoservicio por parte del cajero A y el apoyo que puede dar el cajero Universal A en la ventanilla.

El apoyo que el Cajero A puede dar en ventanilla lo llamaremos CajeroA_Disp y será la diferencia de las actividades del back y el número de cajeros A total, por ejemplo:

[29] Las actividades y los tiempos promedio se realizan mediante un levantamiento en sitio, durante 15 días consecutivos. Las sucursales a medir.

[6] FTE: Full time equivalent, equivalente a jornada completa de trabajo o personas con jornadas completas de trabajo.

[30] Direccionador: Persona que está en la entrada de la sucursal y orienta y organiza a los clientes antes de formarse o tomar un ticket del podio.

Sucursal 1040
Número total de Cajeros A: 2

Cuadro 2.8.1

Actividad por media hora de los cajeros A en el back						Disponibilidad del Cajeros A (A)				
Horario	Lu	Ma	Mi	Ju	Vi	Lu	Ma	Mi	Ju	Vi
08:30	2	2	2	2	2	0	0	0	0	0
09:00	2	2	2	2	2	0	0	0	0	0
09:30	2	2	2	2	2	0	0	0	0	0
10:00	2	2	2	2	2	0	0	0	0	0
10:30	2	2	2	2	2	0	0	0	0	0
11:00	2	2	2	2	2	0	0	0	0	0
11:30	2	1	2	1	2	0	1	0	1	0
12:00	2	1	2	1	2	0	1	0	1	0
12:30	2	1	2	1	2	0	1	0	1	0
13:00	1	1	1	1	1	1	1	1	1	1
13:30	1	1	1	1	1	1	1	1	1	1
14:00	1	1	1	1	1	1	1	1	1	1
14:30	1	1	1	1	1	1	1	1	1	1
15:00	1	1	1	1	1	1	1	1	1	1
15:30	1	1	1	1	1	1	1	1	1	1

La definición de plantilla óptima, se aborda bajo dos premisas:

1. La plantilla debe de ser fija en el tiempo.
2. Debe de considerar la variabilidad de los horarios y los días, manteniendo en lo más posible el nivel de servicio

Bajo estas dos premisas, se define la plantilla óptima, como:

$$\mathbf{PlanOpt}_{(s,\,t,\,h)} = \mathbf{CajeroA}_{\mathbf{Back}(s,\,t,\,h)} + \mathbf{Ventanillas}_{(s,\,t,\,h)} + \mathbf{Direccionamiento}_{(s,\,t,\,h)} \qquad (1)$$

Donde,
PlantOpt = Plantilla optima
CajeroAback = Cajeros que llevan actividades administrativas
Ventanillas = Ventanillas que deben estar abiertas para dar servicio óptimo
Direccionamiento = Cajero universal B que lleva a cabo las actividades de asesoramiento
S = Sucursal
t = tiempo dentro del periodo de estudio
h = horario rangos de medias horas de servicio, de 8:30 – 14:00 hrs

Para obtener el número de empleados que cubrirán las actividades de ventanilla y partiendo de que los Cajeros Universales A, apoyan en estas actividades, se define:

$$Vent_{(s,\,t,\,h)} = CajeroA\_Disp_{(s,\,t,\,h)} + CajeroBVentanilla_{(s,\,t,\,h)} \quad (2)$$

Donde,
Vent = Número de ventanillas óptimas
t = día del periodo de estudio
h = Horario durante el día (medido por medias horas)
CajeroA_Disp = Cajero Universal A, sin actividad en el Back en el horario t
CajeroBVentanilla $_{(t,\,h)}$ = Cajeros universales B que se necesitan para cubrir el requerimiento de Ventanillas de la sucursal en el horario t

Dado que el número de ventillas ya se conoce, así como el número de Cajeros A y su tiempo disponible, lo que se necesita saber es el número de cajeros B.

$$CajeroBVentanilla_{(s,\,t,\,h)} = Vent_{(s,\,t,\,h)} - CajeroA\_Disp_{(s,\,t,\,h)} \quad (3)$$

Por lo tanto la ecuación queda de la forma:

$$PlanOpt_{(s,\,t,\,h)} = CajeroA_{Back(s,\,t,\,h)} + Vent_{(s,\,t,\,h)} - CajeroA\_Disp_{(s,\,t,\,h)} + Direccionamiento_{(s,\,t,\,h)} \quad (4)$$

Con este resultado obtenemos tantas propuestas de ventanilla, como días y horarios dentro del periodo de estudio.

De lo anterior,
Por otro lado, toda sucursal, debe de tener un direccionador o personal que asesore y oriente a cada cliente al momento de entrar a la sucursal, a este cajero o llamaremos $Cajero\_B_{Direccionador}$

De esta manera, la necesidad de la plantilla en sucursal, se define así:

$$Plantilla_{(s,\,t,\,h)} = Cajero\ A_{(s,\,t,\,h)} + CajeroA\_Disp_{(s,\,t,\,h)} + CajeroBVentanilla_{(s,\,t,\,h)} + Cajero\_B_{Direccionador(s,t,h)} \quad (5)$$

Donde,
s = Sucursal de estudio
t = día del periodo de estudio
h = Rango de horario de servicio de 30 minutos, que va desde las 8:30 a 16:00
$Cajero\_B_{Direccionador(s,\,h)} = 1$

Ahora bien, la información obtenida es por media hora y por el periodo de estudio, sin embargo, el personal en sucursal debe ser fijo, por lo que se tiene que dar una propuesta que

satisfaga, de la mejor manera, las variaciones de necesidades de personal en la sucursal y mantener una plantilla fija.

Como bien se sabe, hay dos tipos de Cajeros B en la sucursal, los de tiempo completo y los Cajeros part que cubren los días de mayor transaccionalidad de la sucursal y sólo trabajan 12 días hábiles del mes.

Como es sabido, no todos los días se atiende al mismo número de personas en las sucursales, pues existen días que tienen gran afluencia de clientes y otros días cuyo flujo de clientes baja considerablemente. Por lo tanto no es necesario que todos los días se tenga que cubrir a todas las ventanillas.
Ahora bien, para aquellas sucursales que tienen comportamientos más heterogéneos entre un día y otro, se propone que esta variabilidad sea cubierta por el Cajero B Part.

La propuesta es que el Cajero B cubra el promedio de días como tiempo completo y las plazas que se necesiten para cubrir las ventanillas, sea cubierto por el Cajero part, es decir:

$$CajeroBVentanilla_{(s)} = MedianaCajeroBVentanilla_{(s)} \quad (6)$$

y

$$CajeroPart_{(s)} = Ventanillas_{(s)} - MedianaCajeroBVentanilla_{(s)} \quad (7)$$

Resumiendo, tenemos los Cajeros A, cajeros y Cajeros Part, definidos de manera que la ecuación queda así:

$$Plantilla_{(S,)} = Cajero\ A_{(s)} + MedianaCajeroB(s) + CajeroPart_{(s)} + Cajero\_B_{Direccionador(s)} \quad (8)$$

Quedando por definir el Cajero B direccionador.

Finalmente, para completar la ecuación (1), La sucursal debe tener siempre al menos una persona en la entrada de la sucursal.

Este se definirá bajo el mismo argumento que el Cajero Universal B. Es decir partiendo de la necesidad de tener una empleado en esta actividad, la mayor parte del tiempo.

Lo anterior se puede escribir de la forma en que está la ecuación 9. entonces

$$Plantilla(s, t, h) - Cajero\ A_{(s)} + MedianaCajeroB(s) + CajeroPart_{(s)} > 0 \quad (9)$$

La definición de acotar la ecuación 9, será del mismo modo que se definió el Cajero B, es decir si la mediana de el requerimiento de direccionador es mayor o igual a 0.5, entonces es

necesario un Cajero Universal B para cubrir esta actividad, de lo contrario, se decidirá que no es necesario:

Si ***Mediana (Plantilla(s, t, h) - Cajero A$_{(s)}$ +MedianaCajeroB(s) + CajeroPart$_{(s)}$) > 0.5 ➔ 1, 0 en otro caso 1*** (10)

Retomando el ejemplo de la disponibilidad del Cajero A, se mostrarán los cálculos del resto de los perfiles en la sucursal 1040[31]:

Datos:
Sucursal: 1040
Número de ventanillas óptimas: 6
Personal para actividad back (Administrativo + Autoservicio): 2

Ventanillas óptimas para dar un nivel de servicio óptimo (de acuerdo a la transaccionalidad histórica) Cuadro 2.8.2:

Cuadro 2.8.2
Ventanillas en servicio necesarias (B)

RANGO	Lu	Ma	Mi	Ju	Vi	Lu	Ma	Mi	Ju	Vi
08.30 A 08.59	4	3	2	3	3	4	3	3	5	3
09.00 A 09.29	4	3	2	3	3	3	3	3	5	4
09.30 A 09.59	3	4	4	3	4	5	5	4	5	4
10.00 A 10.29	3	3	3	3	4	3	3	4	5	5
10.30 A 10.59	4	3	4	3	3	4	4	5	5	5
11.00 A 11.29	4	4	3	4	3	3	4	5	5	4
11.30 A 11.59	4	4	3	4	5	3	5	5	4	3
12.00 A 12.29	4	4	3	4	6	5	4	4	4	5
12.30 A 12.59	5	5	5	3	4	6	5	5	5	5
13.00 A 13.29	5	5	5	5	5	4	5	5	5	6
13.30 A 13.59	6	5	5	5	4	5	4	5	4	6
14.00 A 14.29	6	5	4	5	5	5	5	5	5	6
14.30 A 14.59	6	4	5	5	6	6	5	6	6	6
15.00 A 15.29	6	4	5	5	6	6	5	6	5	5
15.30 A 15.59	6	5	5	5	6	6	6	6	6	6

[31] El periodo de estudio real es de 3 meses, en este caso a modo ilustrativo se presentarán datos de solo dos semanas.

Aplicamos La fórmula (3), obtenemos los cajeros B, para ventanilla:

Cuadro 2.8.3										
Cajeros B Necesarios en ventanilla y direccionamiento (C)										
RANGO	Lu	Ma	Mi	Ju	Vi	Lu	Ma	Mi	Ju	Vi
08.30 A 08.59	4	3	2	3	3	4	3	3	5	3
09.00 A 09.29	4	3	2	3	3	3	3	3	5	4
09.30 A 09.59	3	4	4	3	4	5	5	4	5	4
10.00 A 10.29	3	3	3	3	4	3	3	4	5	5
10.30 A 10.59	4	3	4	3	3	4	4	5	5	5
11.00 A 11.29	4	3	3	3	3	3	4	5	5	4
11.30 A 11.59	4	3	3	3	5	3	5	5	4	3
12.00 A 12.29	4	3	3	3	6	5	4	4	4	5
12.30 A 12.59	5	4	5	2	4	6	5	5	5	5
13.00 A 13.29	5	4	5	4	5	4	5	5	5	6
13.30 A 13.59	5	4	4	4	3	5	4	5	4	6
14.00 A 14.29	5	4	3	4	4	5	5	5	5	6
14.30 A 14.59	5	3	4	4	5	6	5	6	6	6
15.00 A 15.29	5	3	4	4	5	6	5	6	5	5
15.30 A 15.59	5	4	4	4	5	6	6	6	6	6

Al calcular la mediana, obtenemos el valor de cajeros B para ventanilla. **Mediana = 4**
El siguiente paso es calcular el part, que es la mediana de los días (Ecuación 7):

$$\text{Cajero Part} = 6 - 4 = 2$$

Entonces tenemos 2 cajeros Part.
Ahora bien, sustituyendo los valores de la ecuación 9, tenemos:

Cuadro 2.8.4										
Cajeros en Sucursal (4 + (A))										
RANGO	Lu	Ma	Mi	Ju	Vi	Lu	Ma	Mi	Ju	Vi
08.30 A 08.59	4	4	4	4	4	4	4	4	4	4
09.00 A 09.29	4	4	4	4	4	4	4	4	4	4
09.30 A 09.59	4	4	4	4	4	4	4	4	4	4
10.00 A 10.29	4	4	4	4	4	4	4	4	4	4
10.30 A 10.59	4	4	4	4	4	4	4	4	4	4
11.00 A 11.29	4	5	4	5	4	4	4	4	4	4
11.30 A 11.59	4	5	4	5	4	4	4	4	4	4
12.00 A 12.29	4	5	4	5	4	4	4	4	4	4
12.30 A 12.59	4	5	4	5	4	4	4	4	4	4
13.00 A 13.29	4	5	4	5	4	4	4	4	4	4
13.30 A 13.59	5	5	5	5	5	4	4	4	4	4
14.00 A 14.29	5	5	5	5	5	4	4	4	4	4

14.30 A 14.59	5	5	5	5	5	4	4	4	4	4
15.00 A 15.29	5	5	5	5	5	4	4	4	4	4
15.30 A 15.59	5	5	5	5	5	4	4	4	4	4

Con la diferencia de los cuadros 2.8.4 – 2.8.3, tenemos el requerimiento del direccionador:

Cuadro 2.8.5
Cajeros en Sucursal

RANGO	Lu	Ma	Mi	Ju	Vi	Lu	Ma	Mi	Ju	Vi
08.30 A 08.59	0	0	0	0	0	0	0	0	1	0
09.00 A 09.29	0	0	0	0	0	0	0	0	1	0
09.30 A 09.59	0	0	0	0	0	1	1	0	1	0
10.00 A 10.29	0	0	0	0	0	0	0	0	1	1
10.30 A 10.59	0	0	0	0	0	0	0	1	1	1
11.00 A 11.29	0	0	0	0	0	0	0	1	1	0
11.30 A 11.59	0	0	0	0	1	0	1	1	0	0
12.00 A 12.29	0	0	0	0	2	1	0	0	0	1
12.30 A 12.59	1	0	1	0	0	2	1	1	1	1
13.00 A 13.29	1	0	1	0	1	0	1	1	1	2
13.30 A 13.59	0	0	0	0	0	1	0	1	0	2
14.00 A 14.29	0	0	0	0	0	1	1	1	1	2
14.30 A 14.59	0	0	0	0	0	2	1	2	2	2
15.00 A 15.29	0	0	0	0	0	2	1	2	1	1
15.30 A 15.59	0	0	0	0	0	2	2	2	2	2

La mediana de los datos anteriores es: Mediana = 0, entonces no es necesario el direccionador.

En resumen, el requerimiento de Plantilla para la sucursal 1040, es:

Cuadro 2.8.6

Perfil	Requerimiento de ventanilla	Plantilla actual	Diferencial
Cajero A:	2	2	0
Cajero B (Ventanilla + Direccionamiento)	4	4	0
Cajero Part	2	1	1
Total	**8**	**7**	**1**
FTE	**7**	**6.5**	**.5**

Finalmente, la propuesta final de plantilla de las sucursales, dado un nivel de servicio de hasta 20 minutos, arroja un diferencial por FTE[32] de 1,184 personas más para dar el nivel de servicio óptimo de 20 minutos. Donde mayor requerimiento de plantilla hay es en las divisiones Metro I y II.

	Plantilla propuesta				Plantilla actual				Diferencia FTE
División	Cajero A	Cajero B	Cajero Part	FTE	Cajero A	Cajero B	Cajero Part	FTE	
Bajío	461	734	204	1,297	444	682	115	1,184	113
Metro I	526	823	202	1,450	471	782	86	1,296	154
Metro II	530	857	188	1,481	490	790	84	1,322	159
Noreste	372	593	151	1,041	376	459	114	892	149
Noroeste	450	644	230	1,209	442	601	102	1,094	115
Occ. I	249	386	100	685	237	326	60	593	92
Occ II	407	578	183	1,077	384	496	116	938	139
Sur	461	742	206	1,306	442	692	102	1,185	121
Sureste	342	552	149	969	328	466	65	827	142
Total	3,798	5,909	1,613	10,514	3,614	5,294	844	9,330	1,184

2.8.1 Descripción del proceso del cálculo

Finalmente, el último paso es el proceso de automatización. El proceso de automatización tiene las siguientes características:

- Está programado en SAS.
- Se encuentra en el servidor del área de productividad (/PRN), con todos los permisos, de manera que cualquiera usuario puede correr el proceso y/o modificar cálculo y/o parametría.
- Tiene opción de elegir diferentes escenarios de nivel de servicio.
- Tendrá opción de determinar las tablas entradas de los datos para el cálculo de manera manual o automática (dadas las bases productivas transformadas en SAS), esto para cualquiera o todas las fuentes.
- Tendrá la opción de cambiar de modelo y parametría en el caso de cambio en actividades del cajero universal A.
- La salida será una tabla donde se especifica la sucursal y las variables , lista para su exportación a cualquier otra herramienta.

[32]FTE: Las siglas en inglés de equivalente en tiempo completo y son las horas de trabajo de los trabajadores de tiempo parcial y/o completo (=1) y se divide en la cantidad de horas de 1 día laboral completo.

La información de sistema y dirección física del proyecto:

Nombre del proyecto en SAS: Calculo_Balanceo_2015_V0_0

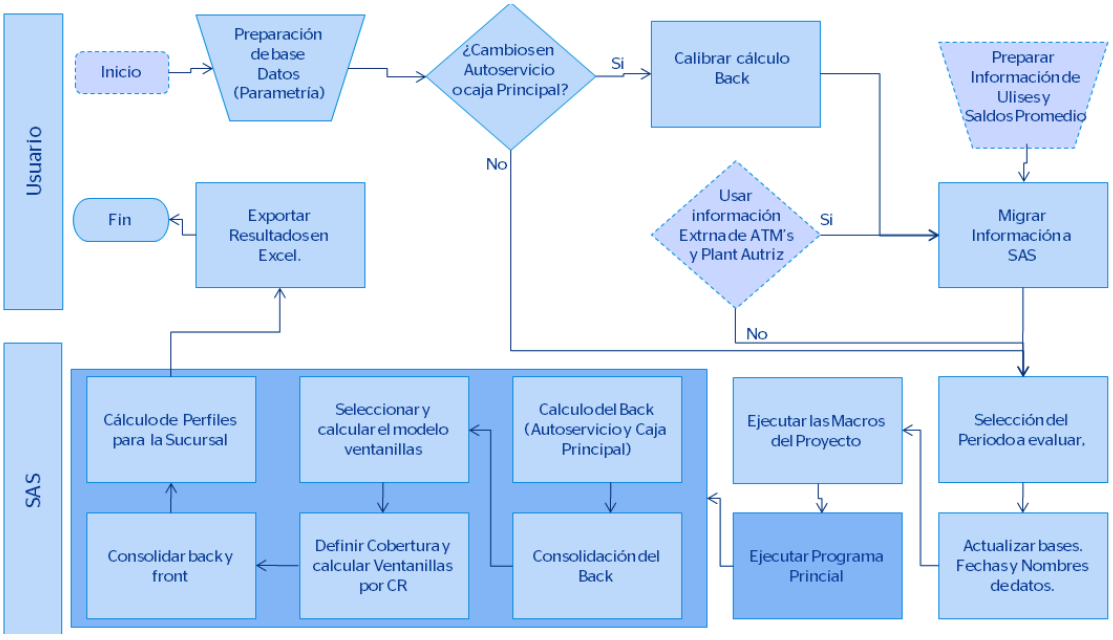
Dirección:
/PRN/Balanceo_2015

Programa principal: 20150206_PRINCIPAL_BALANCEO.sas

Programa	Descripción
Librerías_Balanceo	Procesos varios de cruces de información y cálculos simples.
Calcula_ACTIV_BACK_AUT_SEC	Proceso de cálculo para obtener Cajeros Universales A (Actividad secuencial)
Calculo_ACTIV_BACK_PARAL_SEC	Proceso de cálculo para obtener Cajeros Universales A (Actividad Paralela)
Calculo_ACTIV_BACK_CJ_PPAL	Proceso de cálculo para obtener Cajeros Universales en Caja principal
Consolida_back	Método de consolidación y cálculo de plantilla final total del Back Cajero Universal
CALC_H20MnREG_CP	Cálculo de estimación de ventanillas dado nivel de servicio deseado, NOTA: Este modelo se puede cambiar por cualquier otro escenario.
CALC_VENTANILLAS_C_ESCEN	Metodología de cálculo para obtener número de ventanillas.
CALC_PERFILES_C_ESCEN	Metodología de cálculo para obtener la plantilla con perfiles
20150206_PRINCIPAL_BALANCEO	Unifica y ejecuta las macros anteriores.

Cuadro 2.8.1.1

El proceso de ejecución del programa se esquematiza en el siguiente diagrama(2.8.1.2):



Gráfica 2.8.1.2

2.8.2 Nombres de las bases de Salida Principales y layouts

Los insumos de los dos tipos de Cajeros se listan a detalle en el Anexo II. Las tablas principales son los catálogos, las entradas y salidas. A continuación se menciona el nombre de cada una de ellas y una breve descripción.

Catálogos:

- PARAM_AUTOS_PARA Catálogo de actividades del autoservicio.
- PARAM_AUTOS_SEC Catálogo de actividades del autoservicio.
- PARAM_BACK_PPAL Catálogo de actividades de la Caja Principal

Input para el cálculo de plantilla optima en Ventanilla:

- PLANT_AUT_AAMM Plantilla autorizada del mes
- ULISES_TERM_201412: Información de la sucursales Ulises terminadas
- EQUIPAM_201410: Número de autoservicio a fin de mes del periodo deseado
- TXS_HORARIO: Transacciones por Mes y medias horas.
- SALDO_PROM_MMMAA: Saldos promedio mensuales que maneja la sucursal.
- BAL.SALDO_Periodo: Saldos promedio mensuales que maneja la sucursal consolidado periodo.

3. Conclusiones

El objetivo de proponer una plantilla con un nivel de servicio óptimo de hasta 20 minutos, en las sucursales que se planteó en el punto 2.4, ha sido cubierto con las diferentes metodologías propuestas:

- **Actividades Back[33]:**
  - Actividades de autoservicio: Número de personas para mantener el servicio contínuo.
  - Actividades en caja principal: Número de personas para mantenimiento optimo.

  } Se calcula el FTE[6] sumando los tiempos promedio de las actividades.

- **Actividades Front:**
  - Actividades de ventanilla para otorgar un nivel de servicio óptimo: Número de personas para un adecuado servicio.

  } Modelo de minería de datos.

- **Cálculo de plantilla ideal de la sucursal:**
  - Consolidando los dos puntos anteriores y considerando los perfiles que deberán de cubrirse.

La propuesta de plantilla de actividades del back ha sido cubierta con una metodología basada en la observación de tiempos de espera promedios de una muestra de sucursales, donde el resultado es incorporar 184 cajeros universales A. La ventaja de esta metodología es que se basa en cargas de trabajo y promedio de tiempos de cada actividad.
Si existen cambios en la operativa de las sucursales (simplificación de actividades, incorporación de otras, actualización de la infraestructura, etc.), estas actividades deberán actualizarse o agregarse con mediciones de tiempos de espera.

Por otro lado, el modelo de minería de datos, propone 9,525 ventanillas, como se mostró en el cuadro **2.7.6.1.1**. La minería de datos arrojó un modelo de Regresión lineal múltiple con un nivel de 0.73 de significancia, nos muestra a la minería de datos es una herramienta útil para estimar la plantilla en las ventanillas de las sucursales.

[33] Las actividades y los tiempos promedio se realizan mediante un levantamiento en sitio, durante 15 días consecutivos. Las sucursales a medir.

[6] FTE: Full time equivalent, equivalente a jornada completa de trabajo o personas con jornadas completas de trabajo.

Por otro lado las variables que explican el número de ventanillas, están relacionadas con la operatividad de las sucursales, y sobre todo el horario que determina el flujo de clientes (aunque esto se sabía de antemano, no se conocía con certeza la aportación de cada rango horario, pues anteriormente se clasificaba a los horarios picos iguales para todas las sucursales).

Finalmente las ventajas que otorga el modelo de minería de datos son:

- Estima una plantilla acorde a cada sucursal de la institución, ajustándose a la heterogeneidad de cada sucursal.
- Se ajusta a los cambios que las sucursales van teniendo en el tiempo, pues si cambian los valores de las variables, cambia el estimado de las ventanillas
- No requiere de grandes presupuestos, al relacionar los clientes y las transacciones, pues explota la información que ya existe en las bases de datos de la institución.
- Se puede complementar con información de tipo de clientes y determinar

En la consolidación para el cálculo de plantilla ideal de las sucursales, se requieren 1,184 personas adicionales, lo que implica un 12% de incremento en la plantilla actual para dar un nivel de servicio óptimo definido como 20 minutos de tiempo de espera. Con el método propuesto, se aprovecha al máximo las características del Cajero Universal A, pues considera el tiempo libre para apoyo en sucursal. Se determina, con cierto nivel de certidumbre, si es necesario para las actividades de la sucursal, los cajeros Part time y el apoyo de direccionamiento.

3.1 Consideraciones particulares

Para el cálculo del Cajero Universal A en el back, revisar anualmente las actividades de las dos áreas que lo conforman, o en el caso de implementar un cambio en las funciones que impacten de manera significativa los tiempos de espera de cada tarea en la operativa:

- Autoservicio: revisar anualmente las actividades y los tiempos promedio, la fuente de esta información es el MAR[34].
- Caja Principal: Revisar anualmente las actividades y los tiempos promedio, la fuente de información es el MAR. Incluye actividades cuyos tiempos depende del número de ejecutivos y del tamaño de la sucursal.

Para el cálculo de ventanillas, la técnica es válida para un periodo de 6 meses, por lo que se recomienda validar el nivel de significancia dos veces al año y evaluar la capacidad predictiva de las variables.

[34] MAR: Proceso de medición de tiempo de actividades administrativas. Se ejecuta cada año. Las sucursales son elegidas mediante un proceso aleatorio estratificado.

En cuanto al proceso de cálculo de la estimación de ventanillas, el periodo mínimo de historia es un trimestre, de lo contrario, la variabilidad de los meses podría impactar en el resultado del mismo.
Este proceso de cálculo puede usarse para simular cambios en los siguientes puntos:

- En flujo transaccional
- Infraestructura del autoservicio
- Nuevas transacciones o simplificación de transacciones[35] (Por ejemplo, implementar el cambio de cheques en la practicaja)

[35] Las transacciones, pueden ser, pago de cheques, pago de servicios, disposiciones de efectivo y depósitos.

Anexos

Anexo I

Análisis de Correlación de las 23 variables finales por horario:

The CORR Procedure
C=11:30

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	SUM_TXS	NVENTANILLAS	TOT_R	TOT_A	MAX_TME	MAX_TMA	CTES_PROM_ATEN	TXN_PROM_CTE	FILA	TXN_PROM_VT	cu_a
SUM_TXS	1	0.68495	0.59376	0.60174	0.1558	0.15796	-0.00702	0.15768	0.2081	0.53968	0.4699
		<.0001	<.0001	<.0001	<.0001	<.0001	0.6126	<.0001	<.0001	<.0001	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
NVENTANILLAS	0.68495	1	0.52664	0.53437	-0.00008	0.21001	-0.37867	0.06294	0.10653	-0.17776	0.5109
	<.0001		<.0001	<.0001	0.9953	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
TOT_R	0.59376	0.52664	1	0.7348	0.16196	0.08319	0.26652	-0.10474	0.52917	0.2147	0.3558
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
TOT_A	**0.60174**	**0.53437**	**0.7348**	**1**	**0.09486**	**-0.00423**	**0.51967**	**-0.30765**	**-0.04031**	**0.23169**	**0.3015**
	<.0001	**<.0001**	**<.0001**		**<.0001**	**0.7604**	**<.0001**	**<.0001**	**0.0036**	**<.0001**	**<.0001**
	5209	**5209**	**5209**	**5209**	**5209**	**5209**	**5209**	**5208**	**5209**	**5209**	**520**
MAX_TME	0.1558	-0.00008	0.16196	0.09486	1	-0.00418	0.08947	0.01281	0.55218	0.19389	0.0119
	<.0001	0.9953	<.0001	<.0001		0.763	<.0001	0.3552	<.0001	<.0001	0.389
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
MAX_TMA	0.15796	0.21001	0.08319	-0.00423	-0.00418	1	-0.2064	0.02931	0.09804	-0.03554	0.0905
	<.0001	<.0001	<.0001	0.7604	0.763		<.0001	0.0344	<.0001	0.0103	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
CTES_PROM_ATEN	-0.00702	-0.37867	0.26652	0.51967	0.08947	-0.2064	1	-0.35396	-0.13081	0.47909	-0.1549
	0.6126	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
TXN_PROM_CTE	0.15768	0.06294	-0.10474	-0.30765	0.01281	0.02931	-0.35396	1	0.22283	0.14438	0.032
	<.0001	<.0001	<.0001	<.0001	0.3552	0.0344	<.0001		<.0001	<.0001	0.0197
	5208	5208	5208	5208	5208	5208	5208	5208	5208	5208	520
FILA	0.2081	0.10653	0.52917	-0.04031	0.55218	0.09804	-0.13081	0.22283	1	0.13276	0.1348
	<.0001	<.0001	<.0001	0.0036	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
TXN_PROM_VT	0.53968	-0.17776	0.2147	0.23169	0.19389	-0.03554	0.47909	0.14438	0.13276	1	0.0245
	<.0001	<.0001	<.0001	<.0001	<.0001	0.0103	<.0001	<.0001	<.0001		0.0768
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
cu_a	0.46993	0.51093	0.35589	0.30152	0.01194	0.09059	-0.15493	0.0323	0.13487	0.02452	
CU_A	<.0001	<.0001	<.0001	<.0001	0.389	<.0001	<.0001	0.0197	<.0001	0.0768	
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
cu_b	0.55719	0.76138	0.43155	0.43239	-0.01212	0.22161	-0.23779	0.02999	0.08921	-0.0766	0.493
CU_B	<.0001	<.0001	<.0001	<.0001	0.3817	<.0001	<.0001	0.0305	<.0001	<.0001	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
cu_pt	0.28519	0.47466	0.22953	0.23345	-0.03446	0.13668	-0.18057	0.01313	0.02395	-0.11218	0.1557
CU_PT	<.0001	<.0001	<.0001	<.0001	0.0129	<.0001	<.0001	0.3435	0.0839	<.0001	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
ejecutivos	0.47132	0.6538	0.35831	0.31534	-0.04146	0.17958	-0.24651	0.03053	0.08116	-0.07015	0.5932
EJECUTIVOS	<.0001	<.0001	<.0001	<.0001	0.0028	<.0001	<.0001	0.0276	<.0001	<.0001	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
total	0.55045	0.75773	0.4225	0.39846	-0.02974	0.21445	-0.26075	0.03256	0.09138	-0.07872	0.5877
TOTAL	<.0001	<.0001	<.0001	<.0001	0.0318	<.0001	<.0001	0.0188	<.0001	<.0001	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
atms	0.30146	0.39467	0.30407	0.29114	0.06563	0.04214	-0.06731	0.01242	0.10271	-0.04172	0.3474
ATMS	<.0001	<.0001	<.0001	<.0001	<.0001	0.0023	<.0001	0.3701	<.0001	0.0026	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
practicajas	0.41635	0.52812	0.33276	0.34413	0.03356	0.09198	-0.13232	0.02925	0.09429	-0.03816	0.4727
PRACTICAJAS	<.0001	<.0001	<.0001	<.0001	0.0154	<.0001	<.0001	0.0348	<.0001	0.0059	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
total_at	0.39103	0.50428	0.3551	0.35151	0.05939	0.06989	-0.10505	0.02157	0.11179	-0.04534	0.4474
TOTAL_AT	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1197	<.0001	0.0011	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
saldo_promedio	0.40108	0.46279	0.38006	0.35872	0.06323	0.09695	-0.05374	0.01524	0.12488	0.02773	0.4043
Saldo Promedio	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0001	0.2714	<.0001	0.0454	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
maximo	0.27994	0.36471	0.30514	0.30136	0.08299	0.06963	-0.01215	0.00272	0.10828	-0.0009	0.2505
MAXIMO	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.3808	0.8447	<.0001	0.9483	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
aut_prom_diario	0.53492	0.72644	0.41975	0.39438	-0.0183	0.20284	-0.24701	0.04108	0.09975	-0.07308	0.6291
AUT_PROM_DIARIO	<.0001	<.0001	<.0001	<.0001	0.1867	<.0001	<.0001	0.003	<.0001	<.0001	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
R_TIPO_SUC	0.03564	-0.08107	0.06619	0.02415	0.069	-0.04895	0.10763	0.00834	0.09001	0.10351	0.1554
	0.0101	<.0001	<.0001	0.0813	<.0001	0.0004	<.0001	0.5475	<.0001	<.0001	<.0001
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520
R_BAN_ULISES	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.
	5209	5209	5209	5209	5209	5209	5209	5208	5209	5209	520

Cuadro AI.1

Análisis de Correlación de las 23 variables finales por horario

The CORR Procedure
C=11:30

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	cu_b	cu_pt	ejecutivos	total	atms	practicajas	total_at	saldo_promedio	maximo	aut_prom_diario	R_TIPO_SUC	R_BAN_ULISES
SUM_TXS	0.55719	0.28519	0.47132	0.55045	0.30146	0.41635	0.39103	0.40108	0.27994	0.53492	0.03564	.
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0101	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
NVENTANILLAS	0.76138	0.47466	0.6538	0.75773	0.39467	0.52812	0.50428	0.46279	0.36471	0.72644	-0.08107	.
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
TOT_R	0.43155	0.22953	0.35831	0.4225	0.30407	0.33276	0.3551	0.38006	0.30514	0.41975	0.06619	.
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
TOT_A	**0.43239**	**0.23345**	**0.31534**	**0.39846**	**0.29114**	**0.34413**	**0.35151**	**0.35872**	**0.30136**	**0.39438**	**0.02415**	.
	<.0001	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**0.0813**	**.**
	5209	**5209**	**5209**	**5209**	**5209**	**5209**	**5209**	**5209**	**5209**	**5209**	**5209**	**5209**
MAX_TME	-0.01212	-0.03446	-0.04146	-0.02974	0.06563	0.03356	0.05939	0.06323	0.08299	-0.0183	0.069	.
	0.3817	0.0129	0.0028	0.0318	<.0001	0.0154	<.0001	<.0001	<.0001	0.1867	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
MAX_TMA	0.22161	0.13668	0.17958	0.21445	0.04214	0.09198	0.06989	0.09695	0.06963	0.20284	-0.04895	.
	<.0001	<.0001	<.0001	<.0001	0.0023	<.0001	<.0001	<.0001	<.0001	<.0001	0.0004	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
CTES_PROM_ATEN	-0.23779	-0.18057	-0.24651	-0.26075	-0.06731	-0.13232	-0.10505	-0.05374	-0.01215	-0.24701	0.10763	.
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0001	0.3808	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
TXN_PROM_CTE	0.02999	0.01313	0.03053	0.03256	0.01242	0.02925	0.02157	0.01524	0.00272	0.04108	0.00834	.
	0.0305	0.3435	0.0276	0.0188	0.3701	0.0348	0.1197	0.2714	0.8447	0.003	0.5475	.
	5208	5208	5208	5208	5208	5208	5208	5208	5208	5208	5208	5208
FILA	0.08921	0.02395	0.08116	0.09138	0.10271	0.09429	0.11179	0.12488	0.10828	0.09975	0.09001	.
	<.0001	0.0839	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
TXN_PROM_VT	-0.0766	-0.11218	-0.07015	-0.07872	-0.04172	-0.03816	-0.04534	0.02773	-0.0009	-0.07308	0.10351	.
	<.0001	<.0001	<.0001	<.0001	0.0026	0.0059	0.0011	0.0454	0.9483	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
cu_a	0.4938	0.15572	0.59322	0.58777	0.34747	0.47271	0.44747	0.40432	0.25058	0.62914	0.15545	.
CU_A	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
cu_b	1	0.42713	0.7276	0.92055	0.32735	0.49853	0.44554	0.39568	0.29186	0.76284	-0.02621	.
CU_B		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0586	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
cu_pt	0.42713	1	0.47391	0.48609	0.30946	0.40478	0.39121	0.29211	0.24527	0.4394	-0.12361	.
CU_PT	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
ejecutivos	0.7276	0.47391	1	0.93776	0.28039	0.49823	0.41373	0.41	0.27304	0.85015	-0.12814	.
EJECUTIVOS	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
total	0.92055	0.48609	0.93776	1	0.32538	0.53608	0.46114	0.43377	0.30323	0.87027	-0.08624	.
TOTAL	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
atms	0.32735	0.30946	0.28039	0.32538	1	0.56207	0.92789	0.30213	0.29682	0.45034	-0.14933	.
ATMS	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
practicajas	0.49853	0.40478	0.49823	0.53608	0.56207	1	0.82992	0.31365	0.21394	0.61362	-0.08993	.
PRACTICAJAS	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
total_at	0.44554	0.39121	0.41373	0.46114	0.92789	0.82992	1	0.34518	0.29665	0.58038	-0.14127	.
TOTAL_AT	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
saldo_promedio	0.39568	0.29211	0.41	0.43377	0.30213	0.31365	0.34518	1	0.87596	0.51794	0.14901	.
Saldo Promedio	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
maximo	0.29186	0.24527	0.27304	0.30323	0.29682	0.21394	0.29665	0.87596	1	0.40526	0.03362	.
MAXIMO	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	0.0153	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
aut_prom_diario	0.76284	0.4394	0.85015	0.87027	0.45034	0.61362	0.58038	0.51794	0.40526	1	-0.18275	.
AUT_PROM_DIARIO	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
R_TIPO_SUC	-0.02621	-0.12361	-0.12814	-0.08624	-0.14933	-0.08993	-0.14127	0.14901	0.03362	-0.18275	1	.
	0.0586	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0153	<.0001		.
	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209	5209
R_BAN_ULISES	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.
	5209	5209	5209	!			5209	5209	5209	5209	5209	5209

Cuadro AI.2

Anexo II

Layout de las bases de datos de entrada

Cuadro AII.1

Nombre: LB_ENT.EQUIPAM_201410

Ruta: /PRN/ASV/BALANCEO/ENTRADAS/equipam_201410.sas7bdat

Layout de la información del equipamiento de las sucursales durante el periodo de estudio						
#	**Variable**	**Tipo**	**Longitud**	**Formato de entrada**	**Formato de salida**	**Descripción**
1	cr	Char	4	$CHAR4.	$CHAR4.	CR carácter de 4
2	FECHA	Num	8	DATE9.	DATE9.	Mensual en formato AAMMMDD
3	atm	Num	8	BEST12.	BEST12.	ATM
4	practicajas	Num	8	BEST12.	BEST12.	PRACTICAJAS
5	NFECHA	Char	7			Caracter con formato AAAAMM
6	TOTAL_AT	Num	8			Total Atm's + Practicajas

Cuadro AII.2

Nombre: LB_ENT.PLNT_AUTOR

Ruta: /PRN/ASV/BALANCEO/ENTRADAS/plnt_autor.sas7bdat

Layout Información de Plantilla Autorizada						
#	**Variable**	**Tipo**	**Longitud**	**Formato de entrada**	**Formato de salida**	**Descripción**
1	FECHA	Num	8	DATE9.	DATE9.	Fecha
2	CR	Char	12	$CHAR12.		Caracter de 4
3	CU_A	Num	8	BEST2.		Número de cajeros universales A
4	CU_B	Num	8	BEST2.		Número de cajeros universales B
5	EJEC	Num	8	BEST2.		Número de Personal dedicado a la gestión
6	REST	Num	8	BEST2.		Resto (Excepto pool , sedae)
7	PART	Num	8	BEST2.		Número de cajeros part

8	NFECHA	Char	7			Carácter AAAAMM
9	NOMBRE	Char	55	$CHAR55.	$CHAR55.	Nombre de la s
10	DD	Char	17	$CHAR17.	$CHAR17.	Nombre de la división
11	DZ	Char	23	$CHAR23.	$CHAR23.	Nombre de la zona
12	T_SUC	Char	14	$CHAR14.	$CHAR14.	Tipo suc en Plantilla Autorizada (Tradicional, Empresa, etc)
13	TIPO_SUC	Char	8	$CHAR8.	$CHAR8.	Clase_Suc en Plantilla autorizada (A, B, Singular, etc)

Cuadro AII.3

Nombre: LB_ENT.SALDO_PROM_T4

Ruta: /PRN/ASV/BALANCEO/ENTRADAS/saldo_prom_t4.sas7bdat

Layout de los Saldos promedio de bóveda manejados por la sucursal						
#	**Variable**	**Tipo**	**Longitud**	**Formato de entrada**	**Formato de salida**	**Descripción**
1	Cr	Char	4	$CHAR4.	$CHAR4.	CR carácter de 4
2	SALDO_PROMEDIO	Num	8			Saldo Promedio

Cuadro AII.4

Nombre: LB_ENT.ULISES_TERM_201412

Ruta: /PRN/ASV/BALANCEO/ENTRADAS/ulises_term_201412.sas7bdat

Layout de sucursales Ulises con fechas de Entrega						
#	**Variable**	**Tipo**	**Longitud**	**Formato de entrada**	**Formato de salida**	**Descripción**
1	Cr	Char	4	$CHAR4.	$CHAR4.	Caracter de 4
2	Ulises	Char	2	$CHAR2.	$CHAR2.	SI/NO
3	fecha_de_entrega	Num	8	DATE9.	DATE9.	Fecha de entrega Ulises

Cuadro AII.5

Nombre: LB_SAL.TXS_HORARIO

Ruta: /PRN/ASV/BALANCEO/SALIDAS/txs_horario.sas7bdat

Layout de las Transacciones por sucursal diaria y media horaria durante el periodo de estudio						
#	**Variable**	**Tipo**	**Longitud**	**Formato de entrada**	**Formato de salida**	**Descripción**
1	FECHA	Num	8	DATE9.		
2	RANGO	Char	13	13	13	RANGO
3	NVENTANILLAS	Num	8			
4	SUM_TXS	Num	8			
5	Cr	Char	12	$CHAR12.		
6	NFECHA	Char	7			

Layout de la base de salida

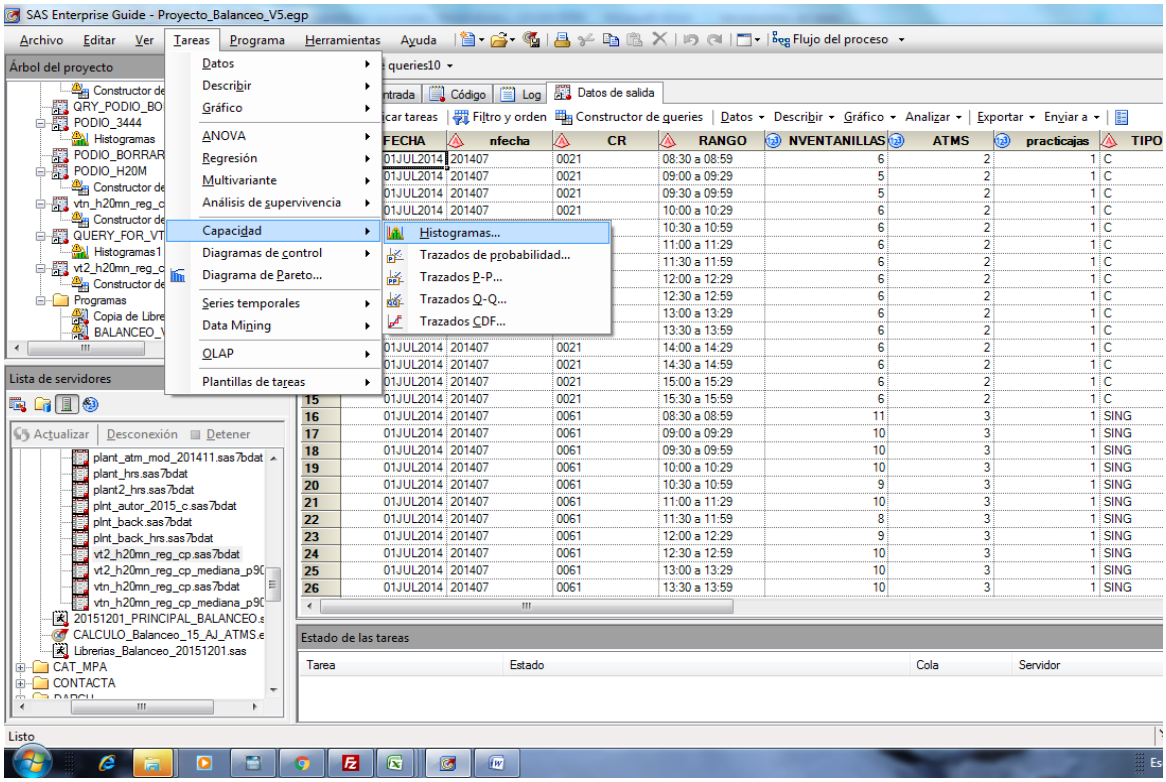
Cuadro AII.6				
Layout Plantilla Propuesta				
#	Variable	Tipo	Longitud	Descripción
1	CR	Alfanumérico	12	Número de la sucursal
2	CB_VF_Mediana_P90_H20Mn_REG_CP	Numérico	8	Número de Cajeros B propuestos bajo el escenario elegido
3	CB_PT_Mediana_P90_H20Mn_REG_CP	Numérico	8	Número de Cajeros Part Time Bajo el escenario elegido
4	CD_MED_Mediana_P90_H20Mn_REG_CP	Numérico	8	Número de Cajeros para direccionar del escenario elegido
5	NV_Mediana_P90_H20Mn_REG_CP	Numérico	8	Número de ventanillas propuestas bajo el escenario elegido
6	CU_A_PRP	Numérico	8	Número de cajeros A propuestos para el Back
7	NOMBRE	Alfanumérico	55	Nombre de la sucursal
8	DD	Alfanumérico	17	División a la que pertenece la sucursal
9	DZ	Alfanumérico	23	Zona a la que pertenece la sucursal
10	T_SUC	Alfanumérico	14	Clase de sucursal
11	TIPO_SUC	Alfanumérico	8	Tipo de sucursal
12	MES_ULISES	Numérico	8	Meses que tuvo actualización de equipamiento
13	BAN_ULISES	Alfanumérico	2	Si tiene equipamiento actualizado
14	CU_A	Numérico	8	Cajeros universales A actuales
15	CU_B	Numérico	8	Cajeros universales B actuales
16	PART	Numérico	8	Cajeros part time actuales

Anexo III

Análisis de variables con SAS

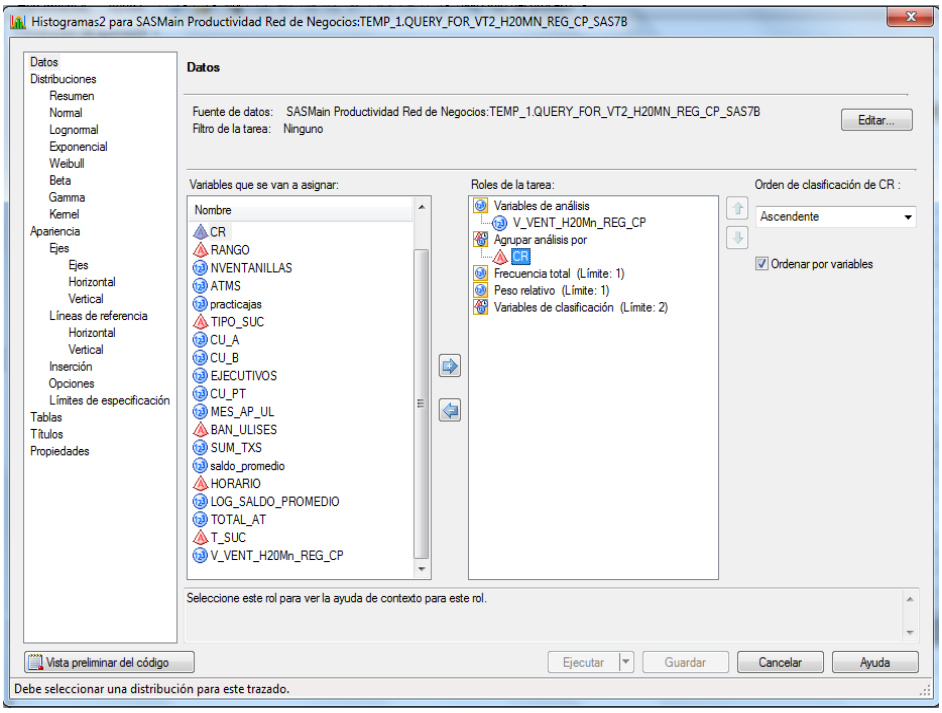
La herramienta SAS tiene varias opciones para analizar variables estadísticas, a continuación se presentará una opción para analizar variables y una validación de su distribución.

Se elige la opción de Histogramas en la opción de Tareas, Capacidad.



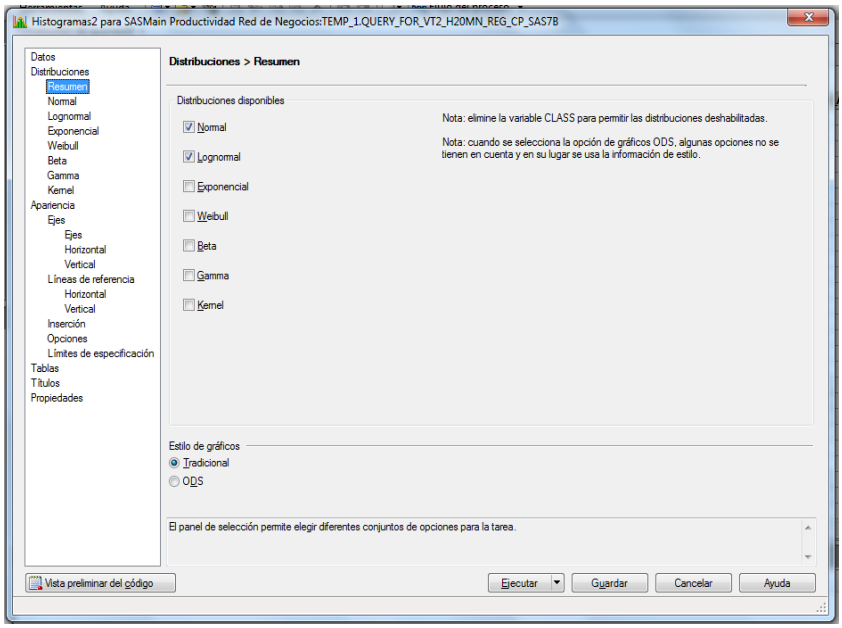
Cuadro AIII.1

Posteriormente abre un menú donde se definen los datos a analizar, (en nuestro caso es el campo de la estimación del número de ventanillas). Dado que el análisis es por sucursal, en la opción de **agrupar análisis por**, se asigna el campo de CR (Sucursal)



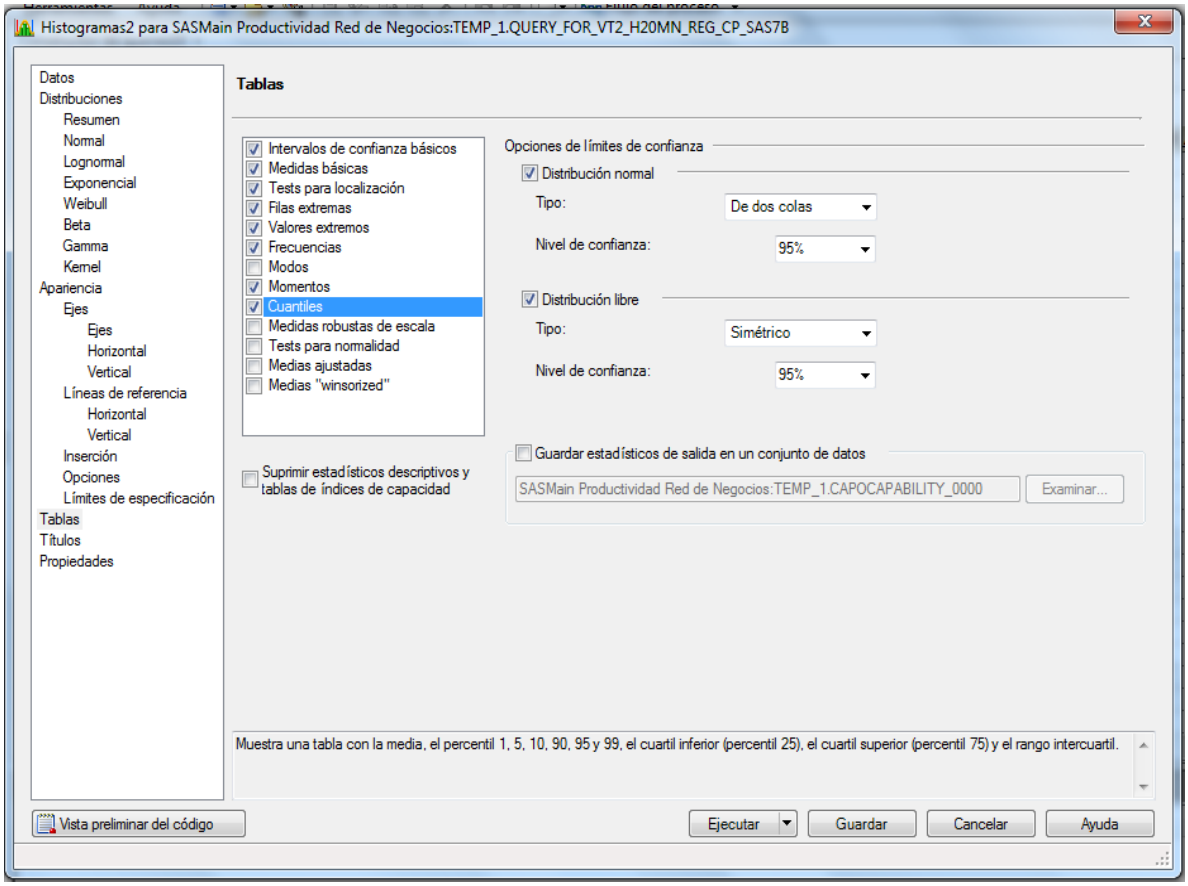
Cuadro AIII.2

La siguiente opción, es **Distribución**, en donde se realiza un análisis de la variable para determinar su posible distribución, en la opción de resumen, se marcan las distribuciones a analizar. Las siguientes opciones de las distribuciones, se dejarán con los valores predefinidos.



Cuadro AIII.3

En la opción de **Apariencia,** se definen los colores y la gráfica de la distribución. Se dejarán los valores que tiene predefinidos.
En la propuesta de **Tablas**, se especifican los análisis estadísticos que se requieren para analizar el comportamiento de la variable, en este caso se elegirán estadísticas básicas, valores extremos, frecuencias, cuartiles e intervalos de confianza básicos.



Cuadro AIII.4

La salida que muestra el paquete SAS, es la siguiente:

Análisis de capacidad de: V_VENT_H20Mn_REG_CP
The CAPABILITY Procedure

Variable: V_VENT_H20Mn_REG_CP
CR=0021

Cuadro AIII.5

Moments			
N	3760	Sum Weights	3760
Mean	3.91914894	Sum Observations	14736
Std Deviation	0.55441378	Variance	0.30737464
Skewness	0.37835507	Kurtosis	2.24566945
Uncorrected SS	58908	Corrected SS	1155.42128
Coeff Variation	14.14628	Std Error Mean	0.00904149

Cuadro AIII.6 Basic Statistical Measures			
Location		**Variability**	
Mean	3.919149	**Std Deviation**	0.55441
Median	4	**Variance**	0.30737
Mode	4	**Range**	4
		Interquartile Range	0

Basic Confidence Limits Assuming Normality			
Parameter	**Estimate**	**95% Confidence Limits**	
Mean	3.919149	3.901422	3.936876
Std Deviation	0.554414	0.542161	0.567237
Variance	0.307375	0.293938	0.321758

Cuadro AIII.7

Cuadro AIII.8 Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	433.4626	**Pr > |t|**	<.0001
Sign	**M**	1880	**Pr >= |M|**	<.0001
Signed Rank	**S**	3535340	**Pr >= |S|**	<.0001

Cuadro AIII.9 Quantiles (Definition 5)								
Quantile	**Estimate**	**95% Confidence Limits Assuming Normality**		**95% Confidence Limits Distribution Free**		**Order Statistics**		
						LCL Rank	**UCL Rank**	**Coverage**
100% Max	7							
99%	5	5.1754618	5.2437262	5	6	3,711	3,735	95.14
95%	5	4.8043775	4.8587669	5	5	3,546	3,599	95.28
90%	4	4.6061238	4.6539743	4	5	3,348	3,421	95.28
75% Q3	4	4.2736665	4.3129475	4	4	2,768	2,873	95.20
50% Median	4	3.9014222	3.9368756	4	4	1,820	1,941	95.16
25% Q1	4	3.5253504	3.5646314	4	4	888	993	95.20
10%	3	3.1843236	3.2321740	3	3	340	413	95.28
5%	3	2.9795309	3.0339203	3	3	162	215	95.28
1%	3	2.5945717	2.6628361	3	3	26	50	95.14
0% Min	3							

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
3	3735	7	3275
3	3732	7	3276
3	3718	7	3350
3	3717	7	3425
3	3712	7	3427

Cuadro AIII.10

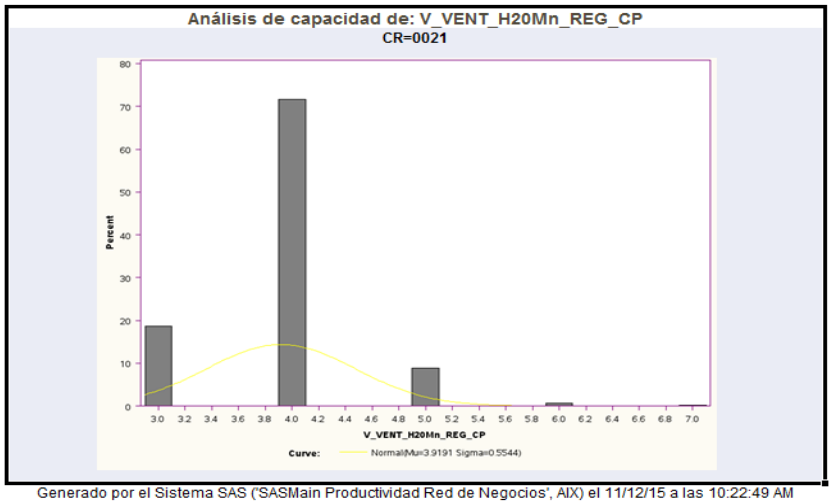
Cuadro AIII.11 Extreme Values

Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	3	703	1	3	703
2	4	2692	2	4	2692
3	5	336	3	5	336
4	6	24	4	6	24
5	7	5	5	7	5

Frequency Counts			
Value	Count	Percents	
		Cell	Cum
3	703	18.7	18.7
4	2,692	71.6	90.3
5	336	8.9	99.2
6	24	0.6	99.9
7	5	0.1	100

Cuadro AIII.12

Generado por el Sistema SAS ('SASMain Productividad Red de Negocios', AIX) el 11/12/15 a las 10:22:49 AM



Generado por el Sistema SAS ('SASMain Productividad Red de Negocios', AIX) el 11/12/15 a las 10:22:49 AM

Cuadro AIII.13

Salida del proceso “Capability” de la sucursal 0021.

Análisis de capacidad de: V_VENT_H20Mn_REG_CP
The CAPABILITY Procedure

Fitted Normal Distribution for V_VENT_H20Mn_REG_
CR=0021

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	3.919149
Std Dev	Sigma	0.554414

Cuadro AIII.14

Goodness-of-Fit Tests for Normal Distribution					
Test	**Statistic**		**DF**	**p Value**	
Kolmogorov-Smirnov	**D**	0.371		**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	119.374		**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	560.56		**Pr > A-Sq**	<0.005
Chi-Square	**Chi-Sq**	216398.06	18	**Pr > Chi-Sq**	<0.001

Cuadro AIII.15

Quantiles for Normal Distribution		
Percent	**Quantile**	
	Observed	**Estimated**
1	3	2.62939
5	3	3.00722
10	3	3.20864
25	4	3.5452
50	4	3.91915
75	4	4.2931
90	4	4.62966
95	5	4.83108
99	5	5.20891

Cuadro AIII.16

Bibliografía

[1] Aprendizaje automático, conceptos básicos y avanzados. Basilio Sierra Araujo, Editorial, Pearson Prentice Hall, Madrid 2006

[2] Data Mining, Soluciones con Enterprice Miner. César Pérez, Daniel Santin.: Alfaomega Ra-Ma Madrid, España 2007

[3] Introducción a la Minería de Datos. José Hernández Orallo, Ma José Ramírez Quintana, César Efraín Ramírez. Pearson Prentice Hall Madrid 2004

[4] Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications, Kattamuri S. Sarma Second Edition [Edición Kindle]

[5] http://support.sas.com/documentation/cdl/en/emgsj/67981/HTML/default/viewer.htm#titlepage.htm

[6] Probabilidad y Estadística, Morris H. DeGroot Addison-Wesley Iberoamericana México, 1988

[7] Cochran, W. G., Sampling Techniques, Third Edition, New York, USA, John Wiley and Sons, 1977.

[8] Comisión Nacional Bancaria y de Valores, http://portafoliodeinformacion.cnbv.gob.mx/bm1/Paginas/infoper.aspx

http://www.cnbv.gob.mx/Inclusi%C3%B3n/Documents/Reportes%20de%20IF/Reporte%20de%20Inclusion%20Financiera%205.pdf

[9] http://www.sas.com/content/dam/SAS/es_mx/doc/assets/27-gestionando-ciclo-vida-2.pdf

[10] http://www.banxico.org.mx/divulgacion/sistema-financiero/sistema-financiero.html

[11] México 2040, una nueva visión, futuro para todos. Claudio Loser, Harinder Kohli y José Fajgenbaum. Taurus México, 2012