UNIVERSIDAD DE ANTIOQUIA

CAJA NEGRA BOT – AUTOMATIZACIÓN DE ORDENES DE SERVICIO

Autor

David Bedoya Llano

Universidad de Antioquia

Facultad de Ingeniería, Departamento de Ingeniería Electrónica y Telecomunicaciones

Medellín, Colombia

2020

CAJA NEGRA BOT – AUTOMATIZACIÓN DE ORDENES DE SERVICIO

David Bedoya Llano

Informe de Semestre de Industria como requisito para optar al título de:

Ingeniero Electrónico

Asesor

Juan Pablo Urrea Profesor Universitario

Universidad de Antioquia
Facultad de Ingeniería, Departamento de Ingeniería Electrónica y Telecomunicaciones
Medellín, Colombia

2020.

Agradecimiento

A mi asesor Juan Pablo Urrea, por su compromiso para guiarme durante el proceso.

A la empresa iData por permitirme mejorar mis conocimientos y poder participar del proyecto.

A la Universidad de Antioquia por ser mi segundo hogar, por permitirme formarme como profesional integral y por darme la oportunidad de conocer personas excepcionales.

A mi esposa por toda su paciencia y amor durante toda mi formación académica.

A mi familia, que siempre me ha apoyado.

Resumen

Durante el presente documento se menciona en ocasiones “La Empresa”, esta hace referencia al cliente con el cual se está desarrollando en conjunto una solución, La empresa contratante (iData) se reserva el derecho de revelar su nombre y/o tipo de vinculación.

El presente proyecto, en modalidad de semestre de industria, está compuesto por la implementación y puesta en productivo de dos modelos de Machine Learning, donde se presenta la solución a desarrollar, como se construyeron las sabanas de datos y definir la arquitectura en Microsoft Azure para su implementación.

La práctica fue llevada a cabo en iData, se contó con la asesoría del director técnico Víctor Manuel Hoyos Valencia, que estuvo presente para brindarme asesoría durante toda la ejecución del proyecto.

El propósito del proyecto es implementar una solución de analítica avanzada, que permita aprobar o rechazar de manera automática y óptima los insumos que componen las ordenes de servicios, con el fin de reducir la carga operativa actual en las ordenes de servicio que demanda a “La Empresa” y sus analistas. De esta manera se busca mantener un control sobre el gasto y propiciar un crecimiento sostenido en las operaciones de mantenimiento preventivo y correctivo sobre toda la flota, aumentando el tiempo de disponibilidad de la flota.

Esta solución contará con un BOT, que consiste en la unión de los 2 modelos de Machine Learning implementados en Microsoft Azure, se encargará de tomar la decisión final sobre el insumo, donde cada modelo tendrá un porcentaje diferente sobre la decisión.

Tabla de Contenido

Índice de Figuras

Índice de Tablas

Índice de ecuaciones

1. Introducción

"La Empresa" requiere implementar una solución que le permita automatizar la gestión de las ordenes de servicio de mantenimiento preventivo y correctivo sobre su flota de vehículos gestionada y no gestionada. A hoy esta gestión se realiza de forma manual basados en la experiencia y criterio de cada uno de los analistas que atiende la operación, analizando los insumos que componen las diferentes órdenes y validando si estos se encuentran dentro de los comportamientos normales con el fin de tomar una decisión.

Con el fin de automatizar dicha gestión se busca construir un histórico comportamental de los datos, permitiendo el análisis y procesamiento desatendido de los insumos que componen la orden de servicio, minimizando la carga operativa actual del proceso, teniendo control sobre el gasto y habilitando un crecimiento sostenido de la operación.

El proyecto Caja Negra BOT es parte de diversas iniciativas estratégicas de negocio que forman parte del marco de ejecución de su estrategia de transformación digital. El presente documento contiene el diseño y la solución analítica avanzada que se desarrolló en el proyecto, la descripción de la solución de analítica para la clasificación de insumos utilizando Machine Learning, así como los elementos asociados entre los datos y los componentes tecnológicos a utilizar.

Se presentan los resultados de las etapas llevadas a cabo, el preprocesamiento de la información, el análisis de los datos de la mano con el negocio y por último el modelo aplicado para obtener la respuesta automática de la orden de servicio generada por un vehículo, todo acompañado de un análisis visual, del comportamiento de los datos.

En el preprocesamiento de la información se preparó la sábana de datos con las variables de interés. Se crearon los respectivos KPI's (Costo, Cantidad, Recencia en tiempo, Recencia en Kilometraje), los 2 primeros KPI's son respecto al insumo que se requiere y los 2 últimos son respecto al comportamiento del vehículo. En el análisis de la información, se analizó el comportamiento de las distribuciones creadas para cada KPI, se hizo un análisis de curtosis y simetría como también la identificación de valores atípicos, para darle tratamiento correspondiente buscando normalizar las distribuciones de probabilidad de cada KPI.

2. Objetivo general

Reducir la carga operativa que demanda a “La Empresa” y sus analistas, por medio del diseño de un BOT. Estará compuesto por dos modelos de Machine Learning, un modelo estadístico No Supervisado y un modelo Supervisado, con el fin de Aprobar y Rechazar todos los insumos que componen una orden de servicio para los mantenimientos preventivos y correctivos de toda la flota.

3. Objetivos específicos

- Realizar unos análisis descriptivos de todas las variables que componen la sábana de datos que suministrará “La Empresa”, esto será el insumo para toda la fase de entrenamiento y tratamiento de datos.

- Identificar las variables relevantes al negocio con el fin de construir las llaves, con base en estas construir los diferentes históricos comportamentales.

- Construir un modelo estadístico No Supervisado a partir del comportamiento histórico de las llaves, creando distribuciones de probabilidad para definir umbrales de decisión.

- Diseñar un modelo Supervisado para intervenir insumos que no cuenten con un histórico comportamental, buscando un grado de similitud con otros para darle criterio al BOT sobre la decisión.

- Realizar pruebas de funcionamiento de los dos modelos de Machine Learning, en conjunto con “La Empresa” y poner en productivo los modelos para su consumo por parte de los analistas.

- Diseñar y crear visualizaciones de control del gasto, insumos aprobados y rechazados, ordenes completas aprobadas y otros indicadores, en herramientas como PowerBI.

4. Marco teórico

El aprendizaje automático también conocido como machine Learning, es un campo que lleva a otro nivel a la inteligencia artificial, es decir, hace que las computadoras aprendan a "Pensar". Son un conjunto de instrucciones que permiten aprender a las máquinas por su cuenta y responder a preguntas con bastante acierto. De esto se desprenden entonces dos modalidades: Aprendizaje Supervisado y No Supervisado.

Ahora como parte fundamental de la ejecución del proyecto es diseñar e implementar un BOT. Un BOT es un programa con un conjunto de instrucciones que imita el comportamiento humano en ciertas actividades con el fin de automatizarla (PWC and Microsoft, 2018), este BOT tendrá todo el conocimiento aprendido del conjunto de los dos modelos a implementar, reemplazando la actividad ejecutada por los analistas.

Será puesto en productivo en Azure Machine Learning Studio. Azure Machine Learning Studio es un portal web que contiene herramientas, y con opciones de poco código o sin código, para facilitar la construcción de modelos, con un esquema arrastrar y soltar facilita la puesta en productivo de proyectos como también su administración (Microsoft, 2019).

4.1. Microsoft Azure

Microsoft Azure brinda la posibilidad de configurar muchos de los servicios en infraestructura local en la nube, olvidándose de los gastos de ingeniería y renovación de servidores, brinda seguridad y comunicación entre servicios, puede lograr trabajar en ambientes híbridos mezclando las bondades de lo local y la nube, se crean las soluciones a necesidad y pagando por lo que usa y siempre está en constante innovación.

La plataforma Azure está compuesta por más de 200 productos y servicios en la nube diseñados para ayudarle a dar vida a nuevas soluciones que permitan resolver las dificultades actuales y crear el futuro (Microsoft Azure, 2021).

En Microsoft Azure existe 3 servicios de los cuales se estará haciendo uso:

- El Área de trabajo de Machine Learning Studio Classic, que permite arrastrar y soltar nodos para crear, probar e implementar soluciones de análisis predictivo, este estará dando solución al modelo No Supervisado (Estadístico).
- El Aprendizaje Automático, tiene las características similares y se presenta como la evolución de su predecesor Área de trabajo de Machine Learning Studio Classic, con nuevas características y servicios disponibles, pero manteniendo la filosofía de arrastrar y soltar nodos, en este se dio solución al Modelo Supervisado.
- El Servidor de SQL, es una base de datos administrada en la nube donde el acceso se proporciona como servicio, el servidor de SQL administrado se

encarga de la escalabilidad, respaldo y como también su disponibilidad, albergará todos los datos que se requieran para entrenamientos de ambos modelos.

4.2. Modelo estadístico (No Supervisado)

A diferencia de los algoritmos de aprendizaje supervisado, en el no supervisado solo se le otorgan las características sin proporcionar ninguna etiqueta (Rodriguez, 2018). es decir, no se suministra la etiqueta o variable de salida, la función es buscar una agrupación, por lo que el algoritmo buscar catalogar por similitud y poder crear así grupos de clasificación.

El modelo No Supervisado realiza la aceptación o rechazo de los insumos en función a estadísticos para determinados grupos de insumos, identificados como llaves.

Para cada una de las llaves se construirá una distribución de probabilidad, y toda variable aleatoria posee una distribución de probabilidad que describe su comportamiento, lo que se conoce entonces como función de distribución representa las probabilidades acumuladas ya se para variables continuas o discretas (Sergas, 2014).

4.3. Modelo XGBoost (Supervisado)

En el aprendizaje Supervisado se entrena el algoritmo otorgándole las variables o características y las etiquetas que serán las salidas o variables a pronosticar (Freidman J., 2008), todo con el fin de que el algoritmo combine e identifique las características para poder lograr las predicciones, y como resultado lograr relacionar las características con las etiquetas para obtener un resultado.

XGBoost pertenece a una familia de algoritmos de impulso y utiliza el marco de aumento de gradiente (GBM – Siglas en ingles) en su núcleo. Es una biblioteca optimizada de aumento de gradiente distribuido (DataCamp, 2019).

XGBoost tiene un conjunto de características que lo hace útil a la hora de su implementación.

- La velocidad y rendimiento, está escrito originalmente en C++.
- El algoritmo central es paralelizable, aprovechable en computadores de múltiples núcleos, en GPU o en grupos de computadores.
- Supera a otros métodos de algoritmos, y ha demostrado un mejor rendimiento respecto a otros algoritmos similares.
- Permite hacer validación cruzada por medio de un parámetro o permite el uso de funciones objetivas definidas de manera local.

4.4. API o Microservicio

Servicio API o Microservicio es la interfaz que un softwate utiliza para interactuar con otro software (Escobar, 2015).

El microservicio será utilizado para tener la comunicación entre el CI (Centro de información) y el BOT, se hará una petición POST con un Json en el cuerpo de la petición y el servicio obtendrá la respuesta en el mismo formato Json con la información procesada.

4.5. Visualización - PowerBI

Para la visualización de los resultados de los modelos implementados, se crearon diferentes indicadores de conocimiento del comportamiento, se hizo por medio de la herramienta Microsoft PowerBI.

PowerBI es un conjunto de herramientas de análisis empresarial, Power BI permite la conexión a cientos de orígenes de datos, como es el caso la conexión de una base de datos de Azure SQL DataBase para diseñar las visualizaciones de cada uno de los indicadores de interés.

4.6. Ecuaciones estadísticas

Se consideraron diferentes descriptores estadísticos para evaluar los diferentes KPI's, los cuales permitirán analizar e interpretar el comportamiento histórico de la flota administrada por "La Empresa".

4.6.1. Variables Cualitativas

- Mayor que el valor más probable: es una bandera que indica si una determinada variable es mayor (o igual) que el valor más probable (1) o si es menor (0).
- Probabilidad baja: es una medida cuyo valor es 1 si la probabilidad de un determinado KPI es mayor a un umbral (1) o menor (0).

4.6.2. Variables Cuantitativas

- Media: Es el valor promedio del KPI dado por la siguiente ecuación

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Ecuación 1. Media de una muestra

- Varianza: es una medida de dispersión de una variable definida como la esperanza del cuadrado de la desviación de dicha variable respecto a su media.

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$$

Ecuación 2. Varianza

- Curtosis: Es una característica de forma de su distribución de frecuencias/probabilidad. Según su concepción clásica, una mayor curtosis implica una mayor concentración de valores de la variable muy cerca de la distribución (pico) y muy lejos de la misma (colas), al tiempo que existe una relativamente menor frecuencia de valores intermedios (hombros).
- Asimetría Estadística: son indicadores que permiten establecer el grado de simetría (o asimetría) que presenta una distribución de probabilidad de una variable aleatoria sin tener que hacer su representación gráfica. La forma de medir la asimetría es a través del coeficiente de asimetría de Fisher, el cual tiene la siguiente forma:

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

Ecuación 3. Asimetría Estadística

Si el coeficiente es mayor a cero, la distribución es asimétrica positiva o a la derecha. Por el contrario, si es menor a cero, la distribución es asimétrica negativa o a la izquierda. (IBM, 2018).

- Falsos positivos: es la cantidad de insumos aceptados por el BOT que no debieron haber sido aceptados divido por el total de ítems aceptados.
- Falsos negativos: es la cantidad de insumos rechazados por el BOT que no debieron haber sido rechazados divido por el total de ítems rechazados.
- Cobertura: es la cantidad de insumos que no fueron enviados a APROBACIÓN MANUAL dividido la cantidad total de insumos.

5. Metodología

En este capítulo serán abordados los elementos utilizados durante el desarrollo del proyecto para el modelo no supervisado 'estadístico' y el modelo supervisado 'Scoring'. Se explicarán los procedimientos de diseño e implementación durante este proyecto para alcanzar la totalidad de los objetivos. El BOT incluye dos etapas entramiento y producción. A continuación, se explica el procedimiento:

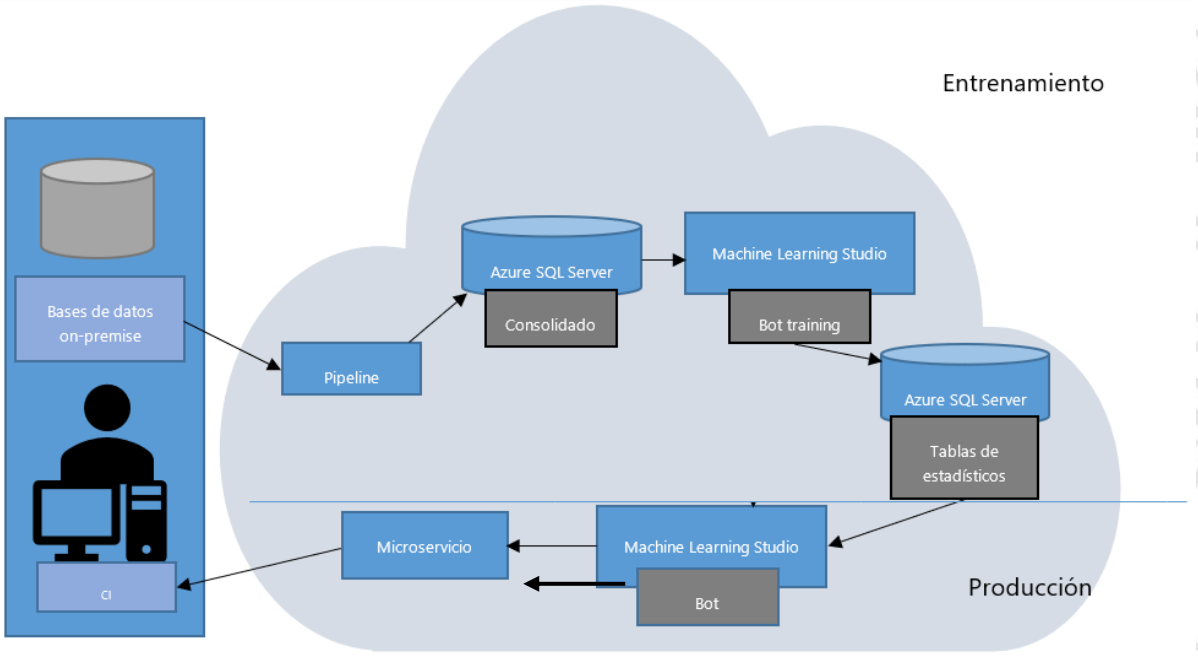


Figura 1. Procedimiento del BOT

Entrenamiento:

- Pipeline: Se leen las bases de datos on-premise para aplicarles transformaciones. Al final se crea una vista única o tabla de datos consolidado, la cual reúne las principales variables.
- Azure SQL Server: Se guarda la tabla *consolidado* en una base de datos Azure SQL. Esta tabla contiene toda la información y las columnas necesarias para entrenar el modelo.
- Machine Learning Studio Training: Se entrena el modelo supervisado y no supervisado en *Machine Learning Studio y Machine Learning classic respectivamente.*
- SQL Server: Se guardan los parámetros del modelo no supervisado en tablas de la misma base de datos SQL. Las tablas guardadas son las siguientes:
    - Tabla de estadísticos 1: *dbo.unsupervised_llave1*
    - Tabla de estadísticos 2 *dbo.unsupervised_llave2*
    - Tabla de estadísticos 3 *dbo.unsupervised_llave3*
    - Tabla de estadísticos 4 *dbo.unsupervised_llave4*

Cada tabla corresponde a los estadísticos de cada llave construida lo cual se aclara más adelante en la sección 5.3 (Variables relevantes, construcción de llaves e históricos comportamentales).

Producción:

- Machine Learning Studio BOT: Se llama la API del *BOT* la cual procesa la información del insumo para determinar si el insumo debe ser aprobado, rechazado o enviado a ajuste
- Microservicio: Se usa un Microservicio como interfaz entre el Bot y la aplicación del Centro de Inteligencia (CI) para garantizar redundancia y tolerancia a fallos.

A continuación, se presenta el esquema (Figura 2) del proceso para el funcionamiento del BOT y los elementos que intervienen.

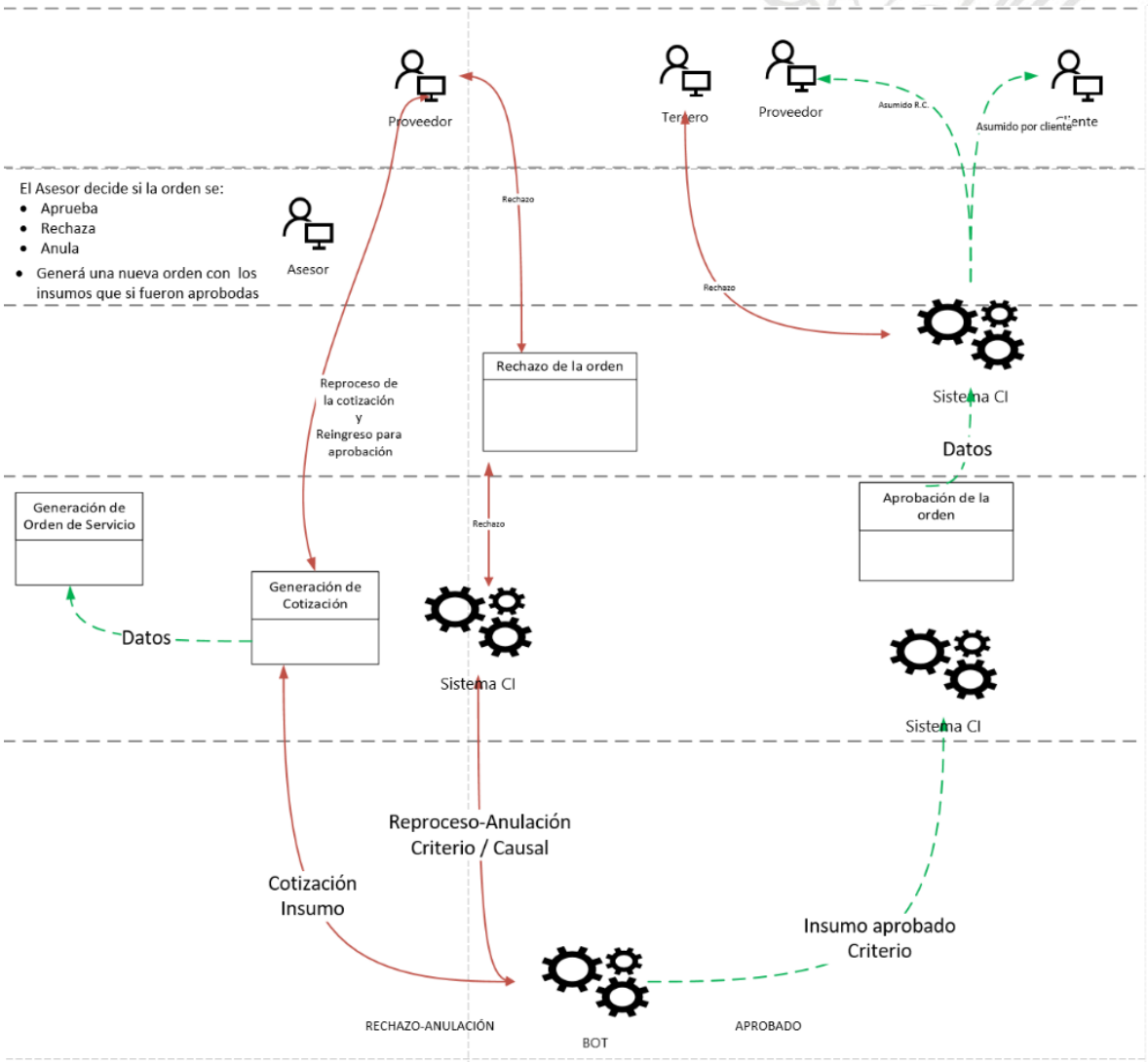


Figura 2. Diseño Conceptual

5.1. Construcción de KPI's

Para cada uno de los insumos ingresados se calcula un conjunto de KPI's. Los KPI's calculados corresponden a los siguientes:

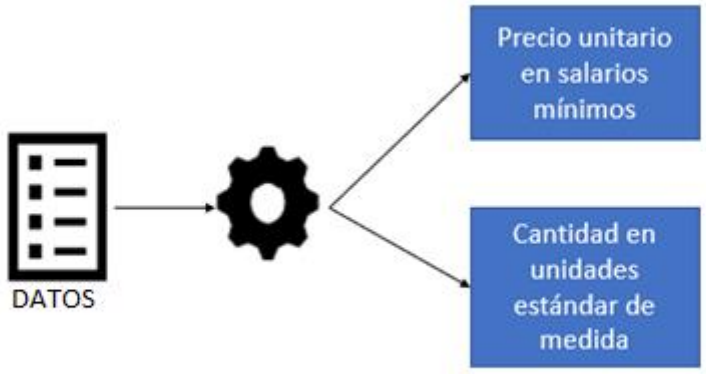


Figura 3 KPI's Calculados para cada insumo

5.1.1. Precio unitario del insumo en salarios mínimos:

Es el precio unitario del insumo, incluyendo IVA y sin incluir descuentos, divido el salario mínimo diario por el respectivo año. Se toma el precio total y se divide por la cantidad (normalizada). Por ejemplo, para el año 2020 el salario mínimo diario es 29260, es decir, que el respectivo valor del salario unitario seria:

$$precio.unitario = \frac{precio.total}{29260 * cantidad}$$

Ecuación 4. Precio Unitario

5.1.2. Cantidad normalizada de insumos:

Es la cantidad de insumos cotizados. Estas cantidades pueden estar en diferentes unidades (litros, cuartos, unidad, etc.). Por eso se lleva a cabo una normalización para convertir todas las unidades a una misma unidad de referencia dependiendo del tipo de medida usada. Este problema se presenta especialmente con las cantidades referentes a volúmenes. La Tabla 1 muestra el factor necesario para convertir una unidad a la unidad de referencia.

Tabla 1. Factores Conversión Unidades

Unidad	Referencia	Factor
Galón	Litro	3,7854
Hora	Hora	1
Cuarto	Litro	0,9463
Unidad	-	1
Pinta	Litro	0,473
Litro	Litro	1
Libra	Libra	1
Kilómetro	Kilómetro	1

La fórmula por utilizar:

$$Cantidad.referencia = cantidad * factor$$

Ecuación 5. Cantidad Referencia

5.1.3. Recencia de tiempo:

Es la cantidad de tiempo transcurrido desde la última vez que se hizo el cambio del respectivo insumo y el insumo cotizado, con la misma actividad y placa de vehículo. Los tiempos tomados como referencia para estos cálculos son la fecha de entrada de la orden asociada al insumo. Este KPI calculado solo aplica en el entrenamiento de la solución para encontrar los estadísticos asociados al insumo.

$$recencia.tiempo = fecha.actual - fecha.ultimo.cambio.insumo$$

Ecuación 6. Recencia en Tiempo

5.1.4. Recencia en Kilometraje:

Es la diferencia en kilómetros del contador entre el último cambio al mismo insumo (hecho para el mismo vehículo) y el kilometraje actual del vehículo cuyo insumo es cotizado. Este KPI calculado solo aplica en el entrenamiento de la solución para encontrar los estadísticos asociados al insumo.

$$recencia.kilometraje = kilometraje.actual - kilometraje.ultimo.cambio.insumo$$

Ecuación 7. Recencia en Kilometraje

El BOT consiste en la unión del resultado del modelo Estadístico y Scoring para tomar la decisión final sobre el insumo. Para ello el BOT consume una lista de insumos y de parámetros. La decisión final del BOT puede tener dos estados:

- APROBADO: Cuando el insumo cumple con los valores esperados por el negocio.

- RECHAZADO: Cuando el insumo tiene valores fuera del rango aceptable.

Aunque en la decisión final se tienen estas dos únicas respuestas como la unión de los dos modelos, quien primero toma una decisión sobre el insumo a procesar es el modelo Estadístico que a su vez tiene otros 2 estados, que son abordados por el modelo de Scoring, por tanto, los insumos que tengan un estado APROBADO y RECHAZADO no serán analizados por el modelo Supervisado.

- AJUSTE: Cuando el insumo tiene un valor fuera del rango aceptable y por lo tanto necesita un ajuste por parte del proveedor para ser aprobado.

- APROBACIÓN MANUAL: cuando la solución no cuenta con suficiente información para aprobar o rechazar el insumo.

5.2. Análisis de variables con sus respectivos descriptivos

Para el análisis de esa sección se usó una sábana de datos que está comprendida entre el 01/12/2019 hasta el 30/11/2020, suministrada por "La Empresa".

5.2.1. Estadístico

Al ser el modelo estadístico no supervisado "La Empresa" depositó mayor confianza en los resultados esperados sobre este modelo ya que los KPI's se calculan sobre el comportamiento histórico y se definen umbrales de decisión sobre las distribuciones de probabilidad. Por ende, se construyeron en conjunto las variables que tienen sentido al negocio para tomar una decisión, basados en el conocimiento experto.

Las variables suministradas fueron:

Tabla 2. Variables Estadístico

VARIABLE	DESCRIPCIÓN
ITEM_ID	ID UNICO DEL OBJETO
ACTIVITY_ID	ID DE LA ACTIVIDAD
USAGE_CONDITION_ID	ID DE LA CONDICION DE USO
CATEGORY_ID	ID DE LA CATEGORIA
BRAND_ID	ID DE LA MARCA
GAMUT_ID	ID DE LA GAMA
CAR_TYPE_ID	ID DEL TIPO DE CARRO
YEAR_CAR	AÑO DEL CARRO
SUPPLIER_ID	ID DEL PROVEEDOR DE SERVICIOS
SUPPLIER_CITY	CIUDAD DEL PROVEEDOR DE SERVICIOS
OP_CITY_ID	ID CIUDAD DE OPERACIÓN
OP_BRANCH_ID	ID DE LA RAMA DE OPERACIÓN

5.2.2. Scoring

El modelo no supervisado tiene la función de dar respuesta a los insumos que tuvieron un estado diferente a RECHAZADO y APROBADO, por lo tanto, las respuestas obtenidas por el modelo estadístico son el insumo para el modelo supervisado.

Scoring tiene como labor buscar similitud entre un sistema, un subsistema y una actividad que no tenga un comportamiento histórico el cual el modelo estadístico no puedo dar una respuesta de aprobado o rechazado ósea ajuste o aprobación manual.

Entonces, se determina un símil a dicho insumo el cual, si posea un comportamiento histórico con las variables mencionadas en la tabla 2, para así tener un criterio de decisión sobre el insumo sin histórico.

Variables escogidas de la sábana de datos resultante del análisis del modelo estadístico fueron:

Tabla 3. Variables Scoring

VARIABLE	DESCRIPCIÓN
CAR_TYPE	TIPO DE CARRO
GAMUT_ID	ID DE GAMA DE CARRO
SYSTEM_NAME	NOMBRE DE SISTEMA
SUBSYSTEM_NAME	NOMBRE DE SUBSISTEMA
ACTIVITY_NAME	NOMBRE DE ACTIVIDAD
QUANTITY	CANTIDAD (KPI)
MILEAGE_RECENCY	RENCENCIA EN KILOMETRAJE (KPI)
TIME_RECENCY	RECENCIA EN TIEMPO (KPI)
TOTAL_PRICE2_N	PRECIO TOTAL NORMALIZADO (KPI)
DECISION	DECISION (ESTADISTICO)

Para el análisis de las variables mencionadas en la Tabla 2., se separan en variables categóricas y variables numéricas, ya que para cada categoría se presenta un análisis diferente.

Para las variables numéricas se presenta un resumen estadístico, mostrado en la figura 4.

	count	mean	std	min	25%	50%	75%	max
GAMUT_ID	983414.0	2.680683	1.240157	1.0	2.0	2.0	4.0	5.0
TOTAL_PRICE	983414.0	121809.912268	692348.246592	0.0	20000.0	50000.0	119778.4	600000000.0
QUANTITY	983414.0	6.345060	994.415237	0.0	1.0	1.0	2.0	786000.0
MILEAGE_RECENCY	983414.0	38650.000000	0.000000	38650.0	38650.0	38650.0	38650.0	38650.0
TIME_RECENCY	983414.0	944.000000	0.000000	944.0	944.0	944.0	944.0	944.0

Figura 4. Estadísticos variables numéricas

Para el análisis de las variables numéricas se construyó un histograma para cada una, buscando identificar relevancias en los valores posibles.

GAMUT_ID: Existen 4 tipos de id de gama 1,2,4,5 cada una representa Económica, Media, Lujo y Maquinaria y Equipo respectivamente. En la gama 4 se encuentra la mayor concentración de vehículos, aunque intuitivamente se

pueda interpretar como vehículos de gama alta, también están incluidos camiones y camionetas de pequeñas prestaciones.

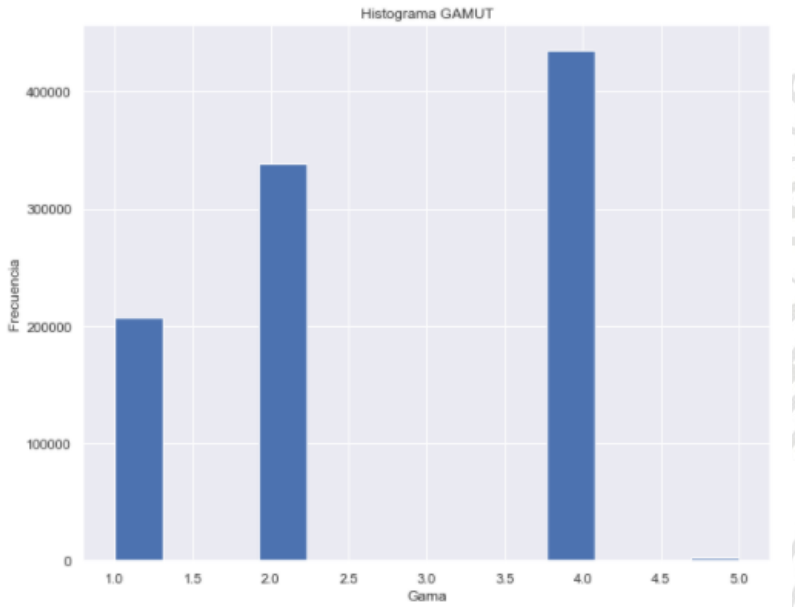


Figura 5. Histograma ID de Gama

Esta variable será de utilidad ya que, al buscar similitud entre vehículos, la gama presenta una gran cualidad.

Para el análisis de los 4 siguientes KPI's se aclara que son valores que ya se encuentran normalizados.

Se presentan en el siguiente orden, QUANTITY, MILEAGE_RECENCY, TIME_RECENCY, TOTAL_PRICE.

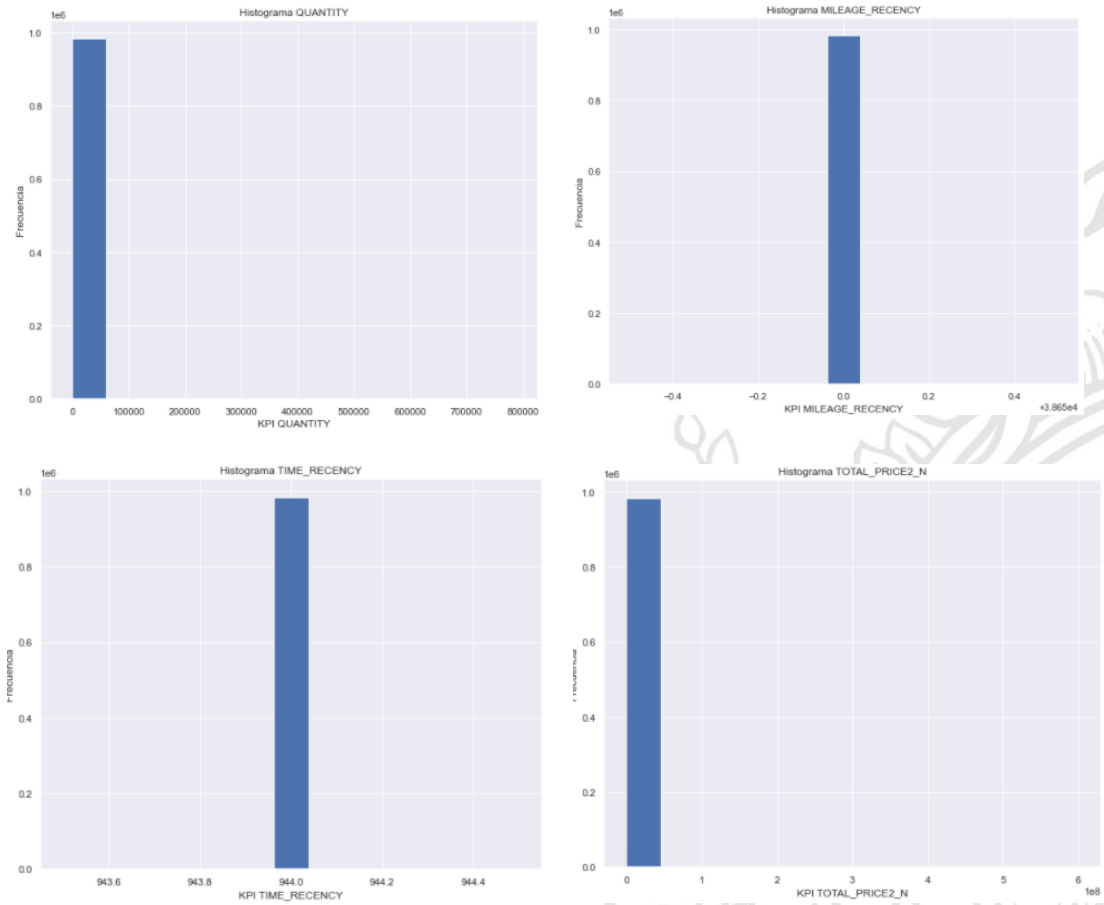


Figura 6. Histogramas KPI's

Los valores de cada gráfico se encuentran normalizados, al tener gran concentración de valores en QUANTITY y TOTAL_PRICE cercanos a cero, los

valores diferentes a estos parecen inexistentes, pero al presentar un diagrama de bigotes para el caso de TOTAL_PRICE se visualizarán los valores atípicos.

Diagrama Bigotes TOTAL_PRICE

Figura 7. Bigotes TOTAL_PRICE

Los KPI's se toman como recomendación de parte de "La Empresa" para tener más asertividad en el grado de similitud con otro insumo con características similares. Pero durante el análisis de los descriptivos se observó que realmente los valores a predecir son esencialmente esos cuatro KPI's con el fin de tomar una decisión, por lo tanto, no deben ser incluidos en la sábana de datos de Scoring.

Adicionalmente los insumos que presentan un estado de AJUSTE y APROBACIÓN_MANUAL, son precisamente insumos que no tienen un

histórico para construir su distribución de probabilidad, por ende, se tenían gran cantidad de nulos sobre las decisiones con estas categorías.

Algo que si dejo evidenciar la construcción de dichos histogramas es que los KPI's si presentan valores atípicos por analizar y que serán tratados en la sección 5.3.

Las variables categóricas en este caso fueron las variables con las cuales se construyó el modelo de Scoring. Se presenta el conteo para la variable CAR_TYPE y DECISION.

```
Camión Mediano     188562
Pickup             171699
Camión Liviano     154228
Dobletroque        121664
Automovil          110800
Tractocamión        60293
Camioneta - SUV     54614
Trailer             45929
Vans                35735
Campero             29670
Motocarro            5263
Otros                1592
Bus o Buseta         1506
Motos                1004
Compresores           511
Montacargas           240
Elevadores            104
Name: CAR_TYPE, dtype: int64
```

Figura 8. Conteo CAR_TYPE

```
APROBADO             703352
APROBACION_MANUAL    180573
AJUSTE                89255
RECHAZADO             10234
Name: DECISION, dtype: int64
```

Figura 9. Conteo DECISION

¿Qué porcentaje entonces de participación el modelo supervisado tendrá sobre la toma decisión del BOT en su respuesta final? Analicemos un poco los porcentajes en las categorías diferentes de Aprobado y Rechazado.

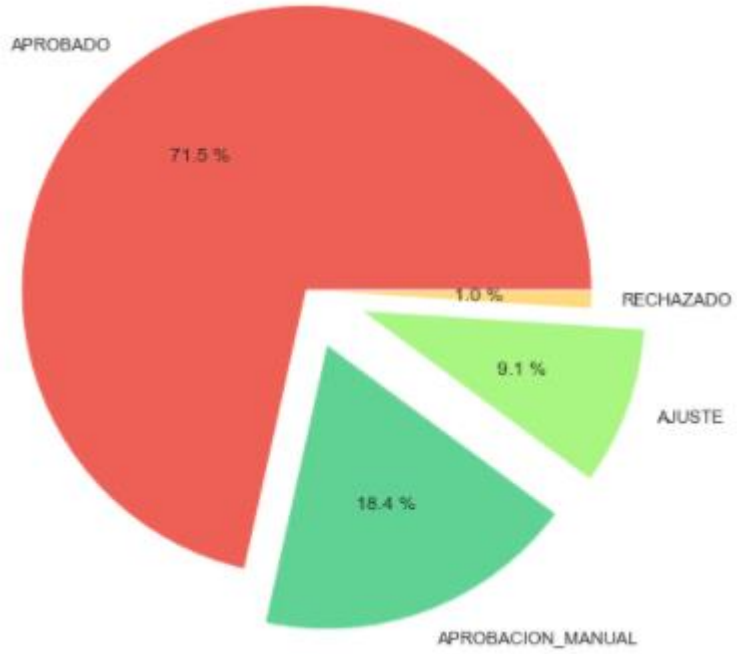


Figura 10. DECISION Modelo No Supervisado

Como el modelo Scoring solo aborda los insumos que tuvieron un estado de APROBACION_MANUAL y AJUSTE, será entonces un 27,5% la participación. Cabe aclarar que de este porcentaje no todo necesariamente terminará APROBADO, y es precisamente en la búsqueda de ese símil que se tendrá un criterio para decir si se APRUEBA o se RECHAZA el insumo sin un histórico comportamental.

Las variables aquí escogidas son la combinación de los análisis descriptivos de las variables mencionadas en la sección 5.2.2 y el conocimiento experto, ya que las variables como SYSTEM_NAME y SUBSYSTEM_NAME son etiquetas a las partes de los vehículos de la flota donde no es posible hacer ningún tipo de análisis.

5.3. Variables relevantes, construcción de llaves e históricos comportamentales

En la sección 5.2 se analizaron las variables para cada uno de los modelos. Para el estadístico se observó que no fue a partir de un análisis exploratorio de las

variables, sino a partir del conocimiento experto y variables que tengan sentido al negocio a la hora de los analistas tomar una decisión.

Por otra parte, en el modelo de Scoring no se logró utilizar los KPI's en busca de otro insumo con características similares. Lo que se hizo entonces fue buscar un insumo que tenga un histórico comportamental con valores semejantes en las variables seleccionadas para así usar los KPI's como referencia para el insumo con falta de histórico.

Lo que sigue a continuación es como se crearon entonces esas distribuciones de probabilidad, que finalidad tienen las variables suministradas por el equipo experto y que significado tienen las tablas de estadísticos presentadas en la sección de entrenamiento del procedimiento del BOT.

5.3.1. Estadístico

Se asigna una llave que identifica cada insumo. La llave del insumo permite especificar un insumo lo más concretamente posible usando variables adicionales. Cada llave corresponde, por tanto, a una combinación única de variables, para la cual se tiene unos KPI's de referencia hallados a través de métodos estadísticos.

Para todo insumo nuevo se encuentra su llave correspondiente en la cual todas las variables coinciden. Sin embargo, es posible que un insumo no tenga una llave correspondiente, ya que las combinaciones de sus variables son nuevas. Para reducir este tipo de problemas, se crean cuatro variables diferentes con cuatro niveles de especificidad diferentes. De tal manera que mientras menos específica sea una llave, más ítems agrupa y más ítems tiene asociados a ella.

Se construyeron 1027 llaves posibles. se presentan las primeras 19 combinaciones en la figura 11:

para consultar el archivo, revise los anexos al final del informe.

LLAVE	Porcentaje	Valor
Puntaje : (ITEM_ID,):	99%	0.9881578947368421
Puntaje : (ITEM_ID, ACTIVITY_ID):	95%	0.9519491312012087
Puntaje : (ITEM_ID, ACTIVITY_ID, CATEGORY_ID):	93%	0.9322047343238479
Puntaje : (ITEM_ID, ACTIVITY_ID, GAMUT_ID):	91%	0.9088567111558802
Puntaje : (ITEM_ID, ACTIVITY_ID, CATEGORY_ID, GAMUT_ID):	88%	0.8836313271216318
Puntaje : (ITEM_ID, ACTIVITY_ID, OP_BRANCH_ID):	88%	0.8771468144044321
Puntaje : (ITEM_ID, ACTIVITY_ID, USAGE_CONDITION_ID):	88%	0.8752367161923948
Puntaje : (ITEM_ID, ACTIVITY_ID, CAR_TYPE_ID):	86%	0.8589020397884664
Puntaje : (ITEM_ID, ACTIVITY_ID, CATEGORY_ID, CAR_TYPE_ID):	86%	0.8589020397884664
Puntaje : (ITEM_ID, ACTIVITY_ID, USAGE_CONDITION_ID, CATEGORY_ID):	86%	0.856008562075044
Puntaje : (ITEM_ID, ACTIVITY_ID, CATEGORY_ID, OP_BRANCH_ID):	84%	0.844534122387308
Puntaje : (ITEM_ID, ACTIVITY_ID, BRAND_ID):	83%	0.8297532107781416
Puntaje : (ITEM_ID, ACTIVITY_ID, USAGE_CONDITION_ID, GAMUT_ID):	83%	0.8252707126668346
Puntaje : (ITEM_ID, ACTIVITY_ID, CATEGORY_ID, BRAND_ID):	82%	0.818419793502896
Puntaje : (ITEM_ID, ACTIVITY_ID, GAMUT_ID, CAR_TYPE_ID):	81%	0.8086439184084614
Puntaje : (ITEM_ID, ACTIVITY_ID, CATEGORY_ID, GAMUT_ID, CAR_TYPE_ID):	81%	0.8086439184084614
Puntaje : (ITEM_ID, ACTIVITY_ID, GAMUT_ID, OP_BRANCH_ID):	81%	0.8052833039536641
Puntaje : (ITEM_ID, ACTIVITY_ID, USAGE_CONDITION_ID, CATEGORY_ID, GAMUT_ID):	80%	0.8027033492822967

Figura 11. Combinatoria Variables

Cada llave tiene un valor asignado en la tercera columna como se muestra en la figura 10. Este valor indica cual es el porcentaje de cada posible llave que tiene muestras mayores a 20 individuos, donde 20 es el valor que se tiene como muestra mínima para la construcción de la distribución de probabilidad. Esto con el fin de mostrar a "La Empresa" cuáles son las llaves que tendrán gran participación en la construcción de estadísticos.

Cada llave tiene una combinación única de variables y debe respetarse el orden, con lo anterior "La Empresa" decidió las siguientes 4 llaves:

Tabla 4. Conjunto de Llaves

LLAVE1	LLAVE2	LLAVE3	LLAVE4
ITEM_ID	ITEM_ID	ITEM_ID	ITEM_ID
ACTIVITY_ID	ACTIVITY_ID	ACTIVITY_ID	ACTIVITY_ID
USAGE_CONDITION_ID	USAGE_CONDITION_ID	USAGE_CONTIDITION_ID	GAMUT_ID
BRAND_ID	GAMUT_ID	BRAND_ID	CAR_TYPE_ID
GAMUT_ID	CAR_TYPE_ID	GAMUT_ID	SUPPLIER_ID
CAR_TYPE_ID	SUPPLIER_CITY	CAR_TYPE_ID	
SUPPLIER_ID			
SUPPLIER_CITY			

Para cada uno de los grupos de las llaves se calculan diferentes estadísticos de referencia que permitirán en un futuro tomar decisiones respecto a la aprobación o rechazo de insumos, como se muestra en la figura 11.

Para evaluar los estadísticos de una llave dada, debe encontrarse la llave a la que pertenece el insumo y buscar los estadísticos en las diferentes tablas.

Ahora para la toma de una decisión cuando un insumo no encuentre estadísticos sobre la llave 1 irá analizando cada una de las llaves y si por último en la llave 4 no encuentra resultado será enviado a aprobación manual de la siguiente manera

- Se calcula la llave 1 para el insumo. Se buscan los estadísticos para la llave 1 en la tabla *dbo.unsupervised_llave1*.

- Al no encontrar resultado en llave1 se calcula la llave 2 para el insumo y se calculan los estadísticos para la llave 2 en la tabla *dbo.unsupervised_llave2*.
- De la misma forma se comprobarán la llave 3 y 4, y si por último en la llave4 no existe un histórico para el insumo, el insumo no será aprobado ni rechazado, y tendrá un estado de Aprobación Manual.

Figura 12. Organización de Llaves

De esta manera para cada uno de los insumos tendremos 3 tres parámetros que corresponden a tres estadísticos de referencia caracterizados de la siguiente manera:

Tabla 5. Parámetros estadísticos

Parámetro	Estadístico correspondiente	Nombre en la tabla de estadísticos
Umbral Máximo	Percentil 0.95	NOMBRE_KPI_UP
Umbral Mínimo	Percentil 0.05	NOMBRE_KPI_DN
Valor Referencia	Mediana estadística (ej. Mediana de precio)	NOMBRE_KPI_REFERENCE

Se listan los diferentes campos que se almacenan en las tablas de estadísticos no supervisados (*dbo.unsupervised_llave1, dbo.unsupervised_llave2, dbo.unsupervised_llave3, dbo.unsupervised_llave4*)

- ID: Otorga un orden a la llave.
- LLAVE: llave que identifica el insumo.
- N: número de insumos del histórico asociados a la llave.
- TOTAL_PRICE_MEAN: Media del precio total normalizado de insumos (precio de los insumos multiplicado por la cantidad).
- QUANTITY_MEAN: Media de la cantidad cotizada del insumo.
- MIL_RECENCY_MEAN: media de la recencia en kilometraje.
- TIME_RECENCY_MEAN: media de la recencia en tiempo.
- TOTAL_PRICE_MEDIAN: mediana del precio total de insumos.
- QUANTITY_MEDIAN: mediana de la cantidad de insumos.
- MIL_RECENCY_MEDIAN: mediana de la recencia en kilometraje.
- TIME_RECENCY_MEDIAN: mediana de la recencia en tiempo.
- TOTAL_PRICE_KURT: curtosis del precio total normalizado.
- QUANTITY_KURT: curtosis de la cantidad.
- MIL_RECENCY_KURT: curtosis de la recencia en kilometraje.
- TIME_RECENCY_KURT: curtosis de la recencia en tiempo.
- TOTAL_PRICE_STD: desviación estándar del precio total.
- QUANTITY_STD: desviación estándar de la cantidad.
- MIL_RECENCY_STD: desviación estándar de la recencia en kilometraje.
- TIME_RECENCY_STD: desviación estándar del tiempo en kilometraje.
- TOTAL_PRICE_UP: umbral superior del KPI del precio.
- TOTAL_PRICE_DN: umbral inferior del precio total.
- QUANTITY_UP: umbral superior para cantidad.
- QUANTITY_DN: umbral inferior para cantidad.
- MIL_REC_UP: umbral superior para recencia de kilometraje.
- MIL_REC_DN: umbral inferior para recencia de kilometraje.
- DIFF_MM: Diferencia entre media y mediana de precio.
- T_PRICE: umbral de precio para la probabilidad del precio.
- T_QUANTITY: umbral de precio para la probabilidad de la cantidad.
- T_MIL_RECENCY: umbral de recencia en kilometraje.
- T_TIME_RECENCY: umbral de recencia en tiempo

Se muestra en la figura 13 la construcción de los estadísticos para cada una de las llaves con repetición mayor a 20 muestras, recordar que la llave es la concatenación de cada una de las variables.

	id	LLAVE	COUNT	RECOUNT	TOTAL_PRICE_MEAN	QUANTITY_MEAN	MIL_RECENCY_MEAN	TIME_RECENCY_MEAN	TOTAL_PRICE_MEDIAN	QU
1	1	3573-5-102-8-2-3-164-3	3567	17835	1,00044569160939	0,999953029436501	7,02875116344265	7,02875116344265	0,997809866710008	1
2	2	3302-1-102-8-2-3-164-3	3434	17170	2,50144342130234	1,83351450203841	7,3776262085032	7,3776262085032	2,50401796599658	1,7
3	3	3573-5-106-8-2-3-164-3	3299	16495	0,992803116303385	1,0005819945438	12,6150168414671	12,6150168414671	0,98082125995678	1
4	4	4243-5-102-8-2-3-164-3	3289	16445	0,58161152772366	7,42108069929901	7,27752260261473	7,27752260261473	0,578348073183232	7,5
5	5	427-4-106-3-2-15-6362-1	3283	16415	0,64047308410091	1	2,08053726469691	2,08053726469691	0,642672692588994	1
6	6	3302-1-106-8-2-3-164-3	3212	16060	2,49694858780002	2,0432516811955	12,5120885429639	12,5120885429639	2,50169394891095	2,0
7	7	3338-5-102-8-2-3-164-3	3166	15830	0,0670039962076709	1,00065808085913	7,92964030322169	7,92964030322169	0,0632750118625927	1
8	8	1907-5-102-8-2-3-164-3	3008	15040	1,771289805415	0,99956914893617	8,10405333776598	8,10405333776598	1,76007459915986	1
9	9	3338-5-106-8-2-3-164-3	2727	13635	0,0734773026968152	1	16,1849247084709	16,1849247084709	0,0715317506783533	1
10	1...	2981-5-115-37-4-8-35-1	2561	12805	0,365096261254418	2,46810910269427	13,3635393518157	13,3635393518157	0,361585593440165	2,4
11	1...	3302-1-106-8-1-1-164-3	2477	12385	2,49927872511472	1,64329608397257	7,60015460637865	7,60015460637865	2,50169394891095	1,6
12	1...	1907-5-106-8-2-3-164-3	2324	11620	1,74629823328822	1,00051635111876	16,3693063683305	16,3693063683305	1,72798840511113	1
13	1...	3573-5-106-8-1-1-164-3	2101	10505	1,00102347218717	0,999944792003808	8,76284047596383	8,76284047596383	0,995756203113382	1
14	1...	3018-1-102-8-2-3-164-3	2054	10270	2,50768404808468	0,515658617332038	9,54746025316453	9,54746025316453	2,50504018761848	0,5
15	1...	4242-5-106-8-1-1-164-3	2005	10025	0,516466605217679	3,80008479201928	8,88387413466332	8,88387413466332	0,510060358587559	3,7
16	1...	3302-1-115-37-4-8-35-1	1852	9260	1,93578498669386	5,71463282937374	39,8608500431966	39,8608500431966	1,93851731644619	5,8
17	1...	4243-5-106-1-2-6-116-2	1832	9160	0,534062589889345	12,1115962139738	7,97198251091703	7,97198251091703	0,534763545336385	12,
18	1...	3302-1-106-1-2-6-116-2	1820	9100	2,01451669867287	2,71617604395601	8,84202698901101	8,84202698901101	2,01524476582053	2,6
19	1...	3574-5-106-1-2-6-116-2	1807	9035	12,1115882476659	0,99997265965689	8,90800529053682	8,90800529053682	12,0848470695867	1
20	2...	3338-5-106-8-1-1-164-3	1803	9015	0,072499230303353	1	9,97042478092064	9,97042478092064	0,0692089778475866	1

Figura 13. Construcción de estadísticos

Las alertas por KPI's, se calculan usando los 3 estadísticos para cada uno de los insumos:

- Alerta de precio: Si el precio del insumo cotizado sobrepasa por encima el umbral máximo de precio o es menor al umbral mínimo de precio.
- Alerta de cantidad: Si la cantidad del insumo cotizado es mayor al umbral máximo de cantidad o menor al umbral mínimo de cantidad.
- Alerta de recencia en KM: Si la recencia en kilometraje es mayor al umbral máximo de recencia de kilometraje del insumo o es menor al umbral mínimo de kilometraje.
- Alerta de recencia en tiempo: Si la recencia en tiempo es mayor al umbral máximo de recencia de tiempo del insumo o es menor al umbral mínimo de tiempo.
- Alerta de precio alto: Si el precio del insumo cotizado está por encima del umbral superior.
- Alerta de cantidad alta: Si la cantidad del insumo cotizado es mayor al umbral superior.
- Alerta de recencia de kilometraje alta: Si la recencia de kilometraje que se envía desde el CI para el insumo cotizado es mayor a la recencia de kilometraje del umbral superior.
- Alerta de recencia de tiempo alta: Si la recencia de tiempo del insumo cotizado, que es enviada por el CI, es mayor al umbral superior de recencia de tiempo para ese insumo.

A continuación, en la Tabla 4 se muestra la lógica implementada para la generación de las alertas en los KPI's

Tabla 6. Alertas por KPI

Nombre de la alerta	Estado de la alerta
Si PRECIO > UMBRAL_MÁXIMO Ó PRECIO < UMBRAL_MÍNIMO	ALERTA PRECIO (O_PRICE==0)
Si CANTIDAD > UMBRAL_MÁXIMO Ó CANTIDAD < UMBRAL_MÍNIMO	ALERTA CANTIDAD (O_QUANTITY ==0)
Si RECENCIA_KM > UMBRAL_MÁXIMO Ó Si RECENCIA_KM < UMBRAL_MÍNIMO	ALERTA RECENCIA KM (O_MIL_RECENCY == 0)
Si RECENCIA_TIEMPO > UMBRAL_MÁXIMO Ó SI RECENCIA_TIEMPO<UMBRAL_MÍNIMO	Alerta Recencia Tiempo (O_Time_Recency ==0)
Si PRECIO > PRECIO_REFERENCIA	ALERTA PRECIO ALTO (HIGH_PRICE==1)
Si CANTIDAD > CANTIDAD REFERENCIA	ALERTA CANTIDAD ALTA (HIGH_QUANTITY==1)
Si RECENCIA_KM < RECENCIA_KM REFERENCIA	ALERTA RECENCIA KM BAJA (HIGH_MIL_RECENCY==0)
Si RECENCIA_TIEMPO < RECENCIA_TIEMPO REFERENCIA	ALERTA RECENCIA TIEMPO BAJA (HIGH_TIME_RECENCY==0)
RECENCIA_KM < UMBRAL_MÍNIMO Y RECENCIA_TIEMPO < UMBRAL MÍNIMO	ALERTA RECENCIA BAJA

Usando las diferentes alertas se decide en cuál de las 4 posibles categorías puede quedar el insumo, según la figura 14.

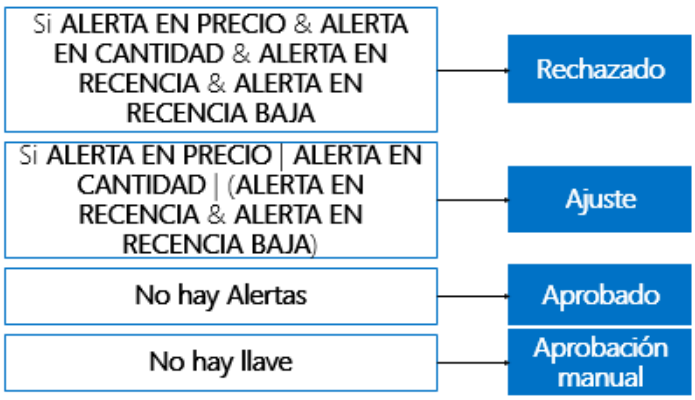


Figura 14. Reglas para la decisión final

Las anteriores reglas se interpretan de la siguiente manera:

- Un insumo es RECHAZADO si tiene alertas en precio, alerta en cantidad y alerta en recencia baja tanto en kilometraje como en tiempo.
- Un insumo es enviado a AJUSTE si tiene alerta en precio o en cantidad o por recencia baja.
- Si no se cumple ninguna de las condiciones anteriores, el insumo es APROBADO.
- Si no se encuentra llave, el insumo se envía a APROBACIÓN MANUAL.
- Si el insumo tiene un KPI no válido (su valor es negativo o igual a cero) se envía a APROBACIÓN MANUAL.
- Si el KPI de Precio o de Cantidad está en NA o tiene un valor nulo, se envía a APROBACIÓN MANUAL.

Se muestra la construcción en la herramienta de machine learning classic del entrenamiento del modelo estadístico en la figura 18, donde se calculan todos los estadísticos mencionados para cada una de las posibles llaves existentes.

Para el entrenamiento se acordó con “La Empresa” un histórico de dos años y medio, teniendo en cuenta que los registros del mes anterior se cargan los primeros 5 días de cada mes.

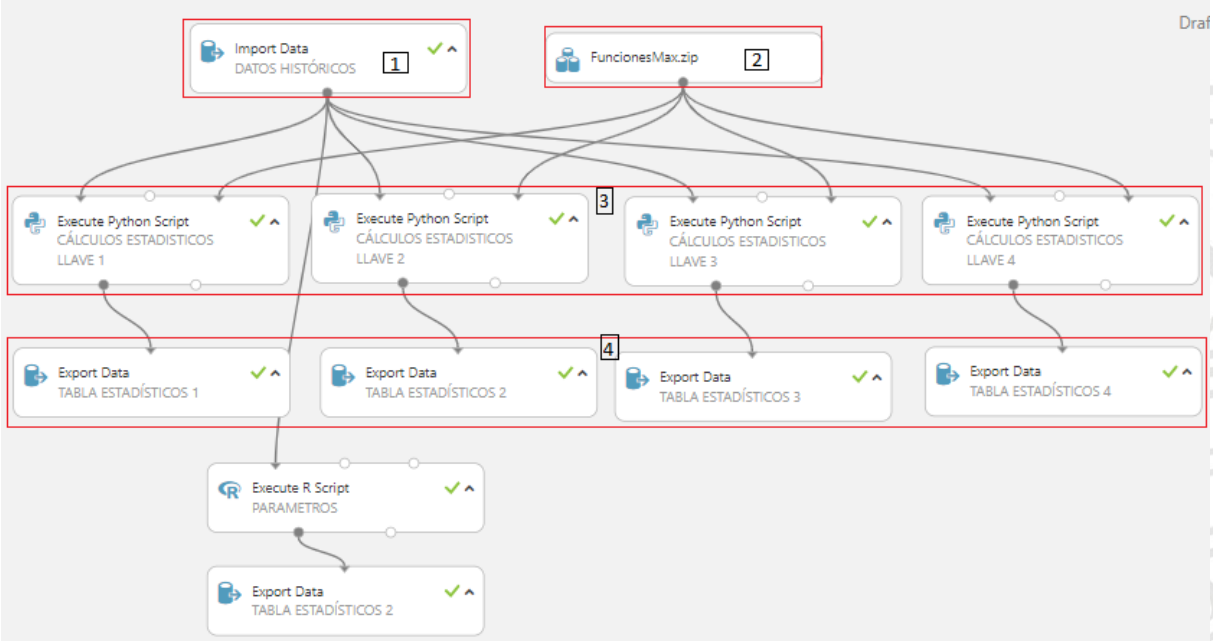


Figura 15. Entrenamiento Modelo Estadístico

- [1] Este nodo se encarga de hacer la extracción de los dos años y medio de datos para entrenamiento
- [2] Se importan funciones creadas para normalizar los datos, filtrar los datos, crear los KPI's para cada llave y la función de boostraping.
- [3] Se da uso a las funciones mencionadas en el item anterior
- [4] Se exportan cada una de las tablas de estadísticos como por ejemplo *dbo.unsupervised_llave1*

Las funciones mencionadas en el [2] todas se han explicado menos la función de bootstraping, a continuación, una explicación y la justificación de dicha función.

Revisando al detalle las distribuciones de cada KPI y los resultados en la tabla de estadísticos para alguna combinación de llave posible, se evidenció en la mayoría de los casos curtosis y asimetría, como lo muestra la figura 15.

Las decisiones del modelo estadístico están basadas en los umbrales en la Tabla 4, tomando como máximo de la cobertura el percentil 0.95 a partir del valor de la media de la muestra.

Veamos entonces un ejemplo en la figura 16 del comportamiento de la variable TOTAL_PRICE, para un valor de llave3 de muestra: "3573-5-3.0-2.0".

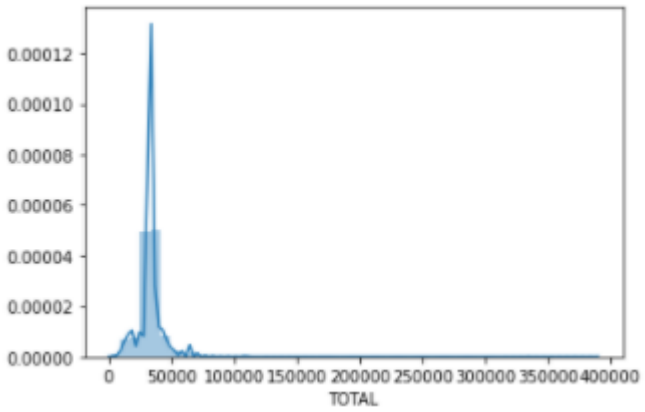


Figura 16. Distribución Precio para una combinación de llave

Si se hace un análisis de descriptivos vemos que la media está ubicada en 33.278 y el percentil 0.95 en 47.821 como lo muestra la figura 17.

```
count      7709.000000
mean      33278.314781
std       10785.871213
min        1889.000000
20%       30418.000000
40%       32683.000000
50%       32683.000000
90%       42000.000000
95%       47821.000000
max      388234.000000
Name: TOTAL, dtype: float64
```

Figura 17. Descriptivos Análisis de distribución

La figura 20 muestra que hay valores para esa llave que oscilan entre cero y valores cercanos a 400.000. Indagando sobre el conocimiento experto de los analistas y revisando al detalle el estado de esos insumos, se identificó la presencia de valores atípicos.

La función Boostraping es una técnica de remuestreo para cuando el tamaño de muestras es muy pequeño o cuando las distribuciones estan muy sesgadas, como la obtención de intervalos de confianza, de pruebas de significación estadística o de cualquier otro estadístico en el que estemos interesados (A., 2015).

Podemos decir que el remuestreo por bootstrap permite crear una aproximación de la distribución de los datos poblacionales partiendo de los datos observados (muestra). Tal distribución se construye partiendo de un remuestreo aleatorio de los datos, en donde los nuevos valores serán los más probables (comunes) con base en la distribución original (reduce datos atípicos, menos probables en la muestra).

Veamos otros 4 ejemplos, para otro conjunto de llaves sobre la misma variable TOTAL_PRICE, esta variable ya se encuentra normalizada.

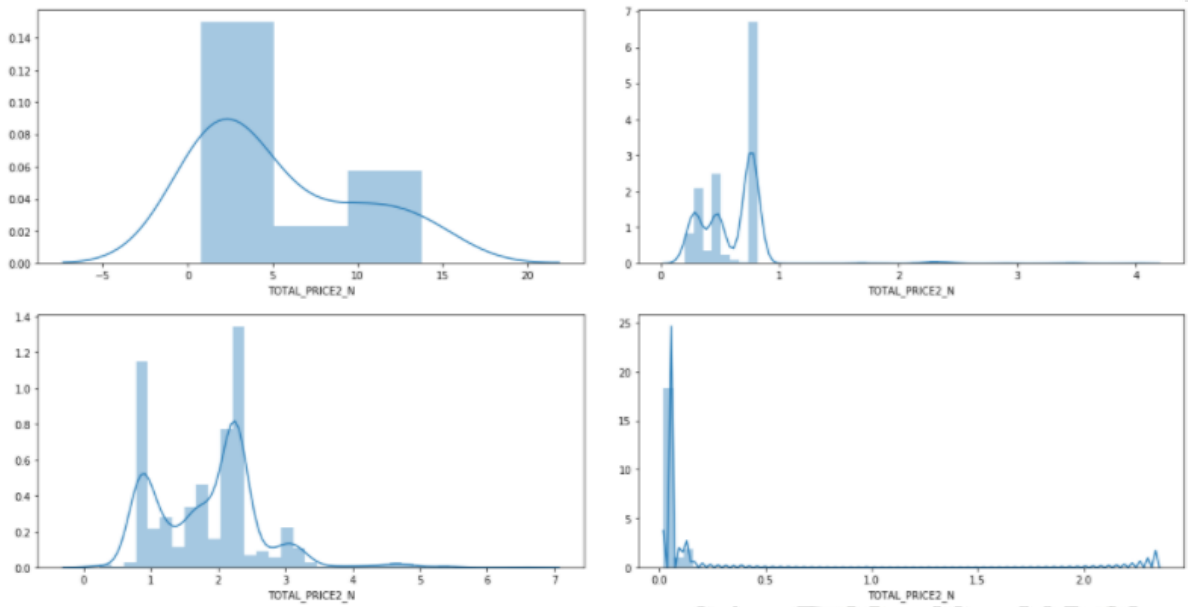


Figura 18. Distribuciones para cuatro llaves

Analizando las distribuciones de la figura 18, vemos que la más cercana a una distribución normal sería la posición [1,1] en la matriz de gráficos, las demás no tienen ninguna semejanza a alguna distribución conocida.

Cabe aclarar que son las 4 llaves con mayor repetición en el histórico comportamental.

```
count    4041.000000
mean        0.066728
std         0.063084
min         0.020335
25%         0.054875
50%         0.058075
75%         0.058168
95%         0.124972
max         2.354742
Name: TOTAL_PRICE2_N, dtype: float64
```

Figura 19. Descriptivos Ejemplo distribución

Los descriptivos de la distribución de la posición [2,2] de la matriz de distribuciones presentados en la figura 19, sucede lo mismo con el ejemplo anterior el percentil 0.95 está muy lejos de un valor máximo que tiene la muestra. Estos valores atípicos provocan la asimetría positiva muy pronunciada.

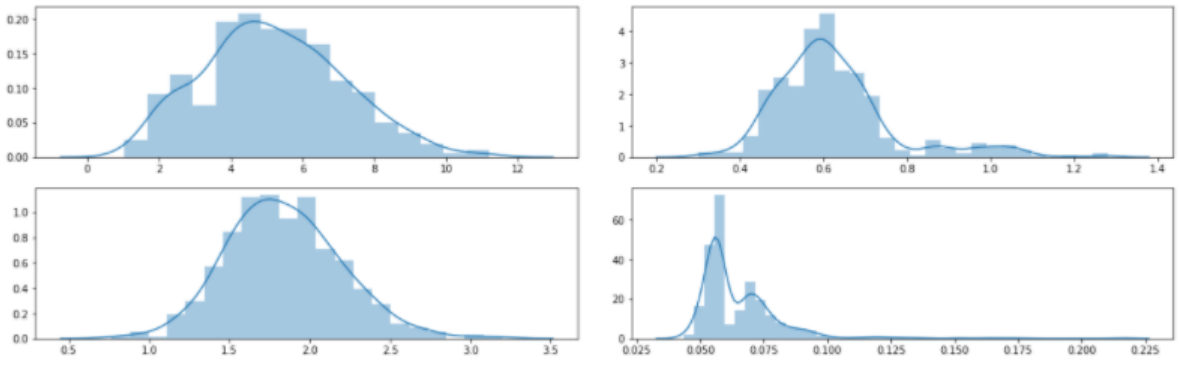


Figura 20. Resultados de aplicar boostrap a las distribuciones

Aplicando boostrap a los ejemplos notamos una diferencia contundente en el comportamiento de los datos observados, nuevamente analicemos los descriptivos para la posición [2,2] de la matriz de distribuciones presentados en la figura 20.

```
count    500.000000
mean       0.065285
std        0.016849
min        0.043665
50%        0.057989
99%        0.130200
max        0.215549
dtype: float64
```

Figura 21. Descriptivos boostrap aplicado a la muestra

El percentil 0.99 en la figura 21 se encuentra en 0.13 y aunque su valor máximo está en 0.21, los valores por fuera del percentil 0.95 serán omitidos o rechazados ya que no se encuentran dentro del comportamiento normal del histórico de la flota.

5.3.2. Scoring

El origen de los datos del modelo de Scoring es el resultado del procesamiento previo del modelo estadístico, aunque aquí no se tienen en cuenta los estadísticos construidos para el modelo, si toma 2 de los KPI's para buscar su símil (TOTAL_PRICE, QUANTITY).

En el preprocesamiento de los datos se parte de las variables anteriormente mencionadas en el origen de los datos para entrenar el modelo, y los análisis descriptivos de las variables nos permiten evidenciar que los KPI's no son necesarios para modelo de Scoring. Esto sucede ya que los insumos que el modelo va a clasificar no tienen un histórico para llave conformada, por lo tanto, las características que describirían la salida del modelo la conforman 5 variables categóricas y 2 variables numéricas. Las cuales son las siguientes:

- CAR_TYPE
- GAMUT_ID
- SYSTEM_NAME
- SUBSYSTEM_NAME
- ACTIVITY_NAME
- TOTAL_PRICE
- QUANTITY

La salida del modelo estadístico, representada por la variable DECISION, contiene 4 categorías como se explicó en la sección 5.2.2, al cual, según el conocimiento experto, se hizo la siguiente transformación:

Para todos los insumos que estaban etiquetadas como APROBACION_MANUAL se transformó en RECHAZADO. Esto sucede ya que los insumos que presentan esta etiqueta normalmente son rechazados por el negocio, contienen muchas inconsistencias dentro de los valores cotizados de los insumos. Por otro lado, los insumos representados con la etiqueta AJUSTE se transformaron en APROBADOS, ya que "La Empresa" normalmente aprueba dichos insumos con esta etiqueta.

Finalmente, la variable DECISION solo queda con dos categorías, APROBADO y RECHAZADO.

La sábana de datos que alimenta el modelo de Scoring al presentar en su mayoría variables categóricas, deben transformarse a numéricas mediante un método 'label encoding'. Este método convierte cada valor de una columna categórica en un número (Yadav, 2019). Finalmente, los datos no tendrán valores categóricos.

Para el entrenamiento del modelo se utilizó el 70% de la muestra y el 30% restante se utilizó para validación.

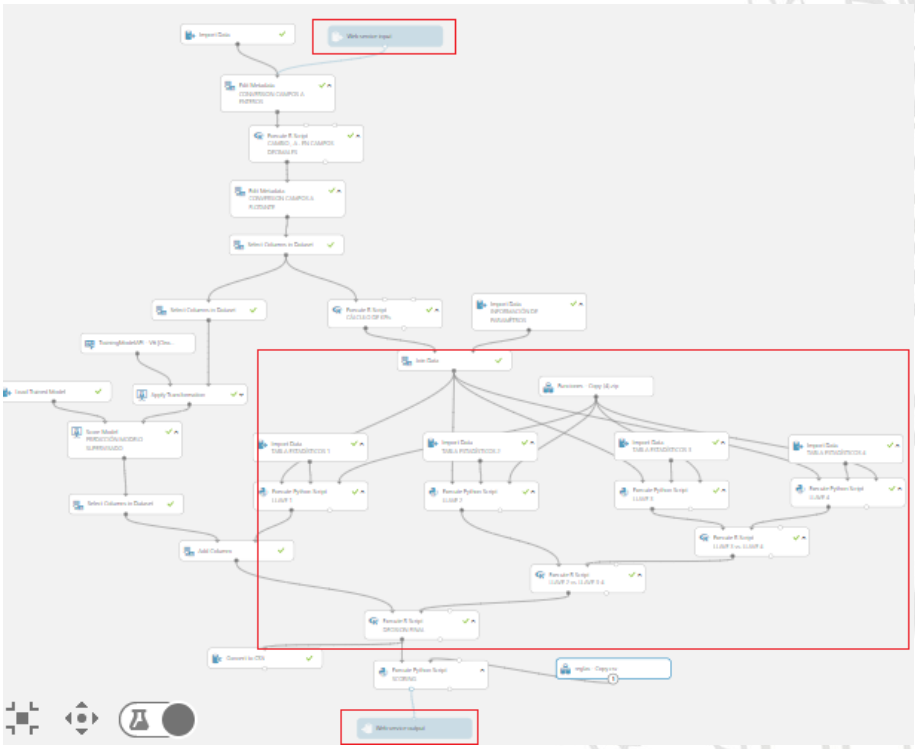
Ahora para la implementación del modelo de Scoring, se utiliza un clasificador XGBoost. El clasificador tiene los siguientes hiperparámetros escogidos:

- Learning_rate: 0.05
- Random_state: 42
- N_estimators: 100
- Objective: 'binary-logistic'

Lo demás hiperparámetros del modelo XGBoost se dejaron por defecto.

5.4. Construcción Modelo No Supervisado Estadístico

En Machine Learning Studio se tienen dos flujos: un flujo de entrenamiento donde se construyeron todos los estadísticos y uno para el flujo predictivo. En esta sección analizaremos el flujo predictivo.



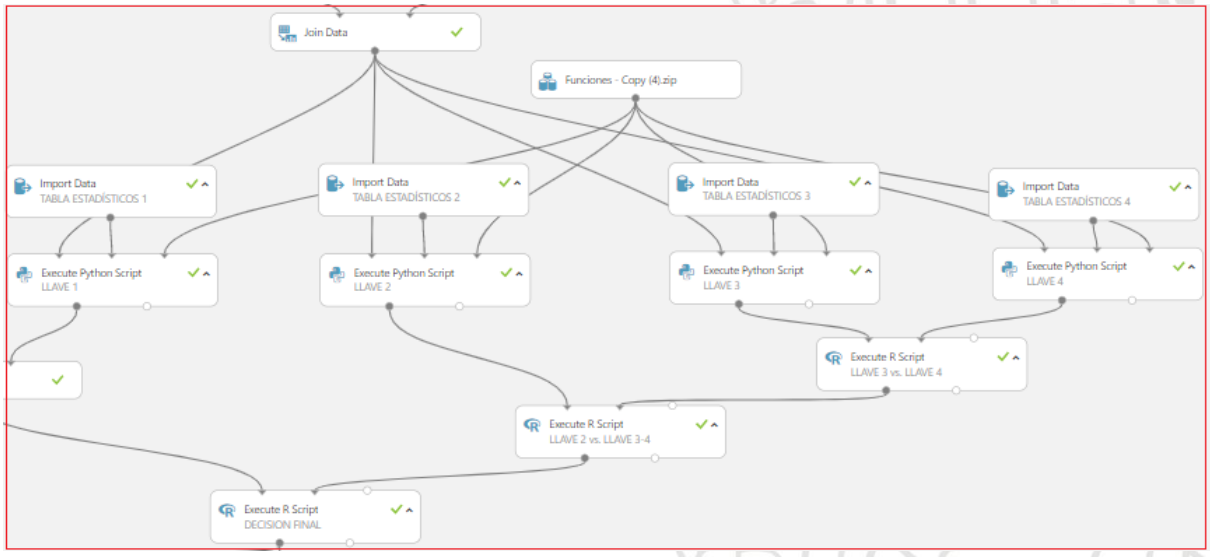


Figura 22. Modelo Estadístico Completo

En la figura 22 se muestra la implementación de modelo estadístico creado, en el recuadro rojo más grande, se muestra cuando se cargan las funciones necesarias y las 4 tablas de estadísticos creadas para cada llave.

Los cuadros pequeños en la parte superior e inferior corresponden a las entradas y salidas correspondientes a los datos que envía el CI (Centro de Información – "La Empresa") al servicio web respectivo expuesto desde Machine Learning Studio.

La parte fundamental de este modelo se encuentra en la figura 23, estos módulos de la herramienta implementan un código en Python encargados de realizar la asignación de las llaves, evaluación de estadísticos y generación de alertas en los KPI, para la orden que va a ser procesada.

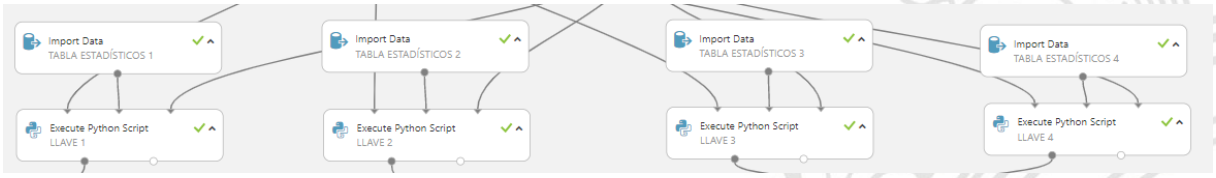


Figura 23. Módulos de asignación de llave, estadísticos y KPI's

Una vez determinada la respuesta del insumo procesada en el modelo encerrado por el recuadro amarillo en la figura 24, dicha respuesta tendrá los cuatro posibles estados en la decisión, APROBADO, RECHAZADO, AJUSTE o APROBACION_MANUAL.

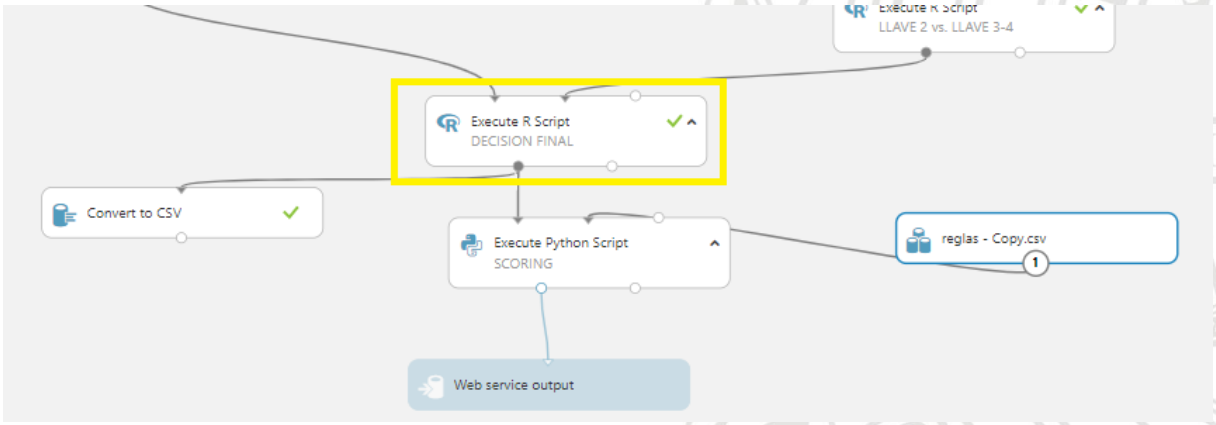


Figura 24. Modulo decisión final

Hasta este punto tiene participación el modelo estadístico, la respuesta puede enviarse por el microservicio web, pero quedaría incompleto ya que hace falta la participación del modelo Scoring, empleado en el siguiente módulo y analizado en la sección 5.5.

5.5. Construcción Modelo Supervisado Scoring

Para la construcción del modelo de Scoring se utilizó la herramienta Machine Lerning Studio, donde esta se cargó el modelo entrenado en local y lo que permite la herramienta es disponibilizar un microservicio utilizado posteriormente en el flujo de predictivo del BOT.

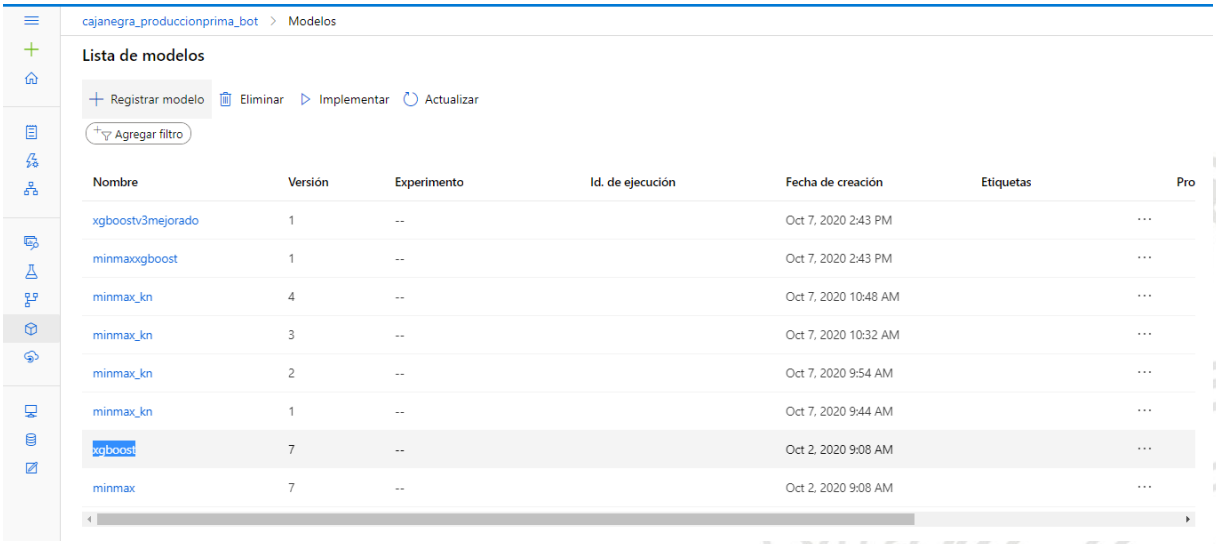


Figura 25. Carga modelo XGBoost entrenado

A diferencia del Machine Learning Classic que lleva un flujo de ejecución, el Machine Learning Studio se utilizó únicamente para la carga del modelo supervisado entrenado y crear un punto de conexión o microservicio de dicho modelo, que será consumido posteriormente dentro del flujo de Machine Learning Classic.

En la figura 26 observamos el módulo encerrado en el recuadro rojo que ejecuta un código en Python se envía el JSON al microservicio del modelo de Scoring, lo procesa y contesta otro JSON, y se une con la respuesta del modelo estadístico para dar una respuesta final.

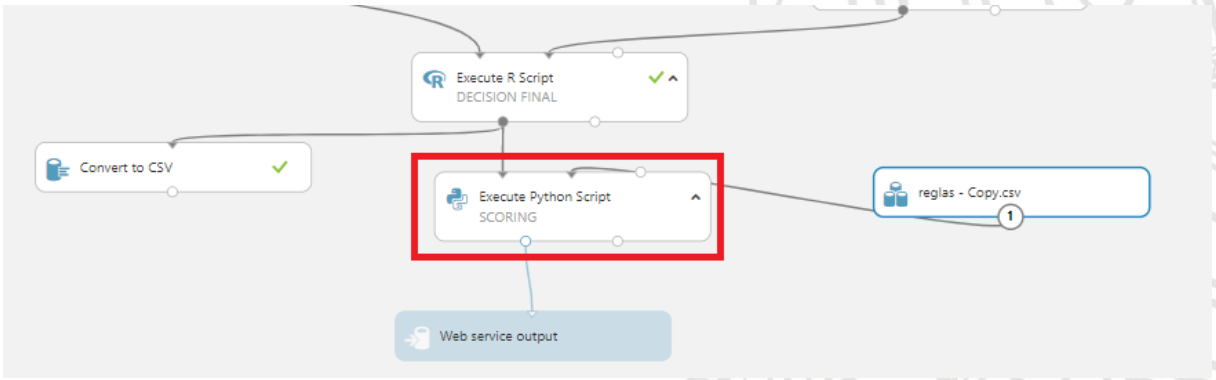


Figura 26. modulo para consumo de microservicio modelo Scoring

5.6. Pruebas de funcionamiento y puesta en productivo

5.6.1. Flujo de ejecución

Para el desarrollo del BOT en esta sección se detallan los distintos componentes para tener en cuenta:

El Flujo analítico define los diferentes procesos y transformaciones que se realizan sobre los datos para lograr la clasificación deseada.

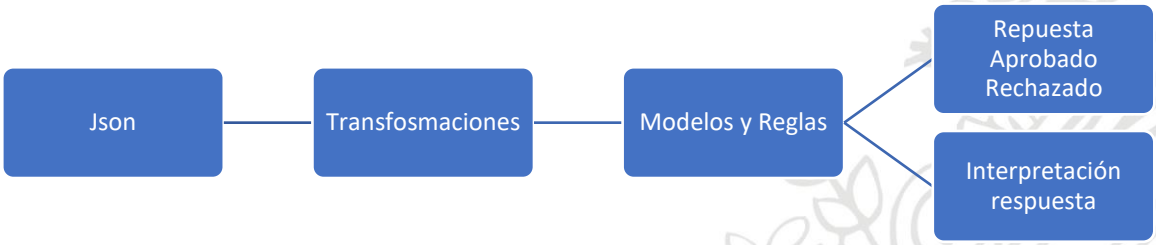


Figura 27. Flujo analítico de ejecución

A continuación, se describen los componentes de la figura 27:

- JSON: Tiene cada uno de los datos de la orden y los insumos que son necesarios para la clasificación de los insumos de la orden de servicio. Las variables que son importantes y que se tendrán en cuenta, se determinadan en la fase de entrenamiento.
- Transformaciones: Las transformaciones son las operaciones de generación de nuevas variables a partir de las que se entregan inicialmente. Estas variables nuevas enriquecen los modelos y permitirán decidir el accionamiento de las reglas.
  A continuación, se presentan algunas de las variables creadas y algunas variables que se consideran, a priori, como variables que pueden ser usadas en su formato original y variables que se pueden crear con transformaciones.
  Variables iniciales definidas por el negocio:
  - Nombre del insumo
  - Precio del insumo
  - Sistema del insumo
  - Subsistema del insumo
  - Placa
  - Marca del carro
  - Modelo
  - Región

- Cantidad
- Canon del insumo
- Fecha de la orden
- Contador
- Orden
- Revisión
- Última revisión
- Siniestro
- Abuso
- Aprobado/Rechazado
- Fotos adjuntas (¿)
- Proveedor
- Tipo de servicio
- Sucursal
- Descuento
- Tipo de contador
- Tipo de Vehículo
- Marca
- Cliente
- Condición de uso
- Gama

Variables creadas (indicadores de conocimiento):

- Último cambio a insumo en kilómetros
- Último cambio a insumo en tiempo
- Tipo de mantenimiento correctivo o preventivo
- Tiempo de vida útil estándar por insumo y por vehículo
- Cantidad estándar por insumo y por vehículo
- Precio estándar por insumo y por vehículo

- Modelos: Los modelos contienen parámetros encontrados durante el entrenamiento y permiten decidir, con base en las variables, la mejor clasificación para los datos de los insumos. Estos modelos se complementan con las reglas, ya que permiten encontrar patrones ocultos en los datos, los cuales, por su misma naturaleza, no pueden ser codificados en forma de reglas.
- Reglas: Las reglas del modelo consistirán en un conjunto de condicionales que utilizan variables originales y transformadas para decidir la salida del modelo (clasificación de los insumos en aprobada, ajuste o rechazo). Las reglas definidas se basarán en los criterios definidos por el negocio en combinación con reglas derivadas del análisis de los datos. Las reglas se aplican a nivel de insumo.

  Un ejemplo de regla sería: Si el *contador* tiene un valor cercano a la *revisión* y el mantenimiento es *preventivo* entonces el insumo es ACEPTADO.

- Salida: La salida consta de 2 partes: la respuesta del insumo y la interpretación de la decisión:
    - Respuesta del insumo: las posibles respuestas de la solución son:
        - APROBADO: Insumo que cumple con todos los KPI's
        - RECHAZADO: Se define cuando no se encuentra que el insumo cumple con al menos uno de los KPI's o el insumo con el que se asoció en Scoring tuvo un estado anterior de Rechazado.
    - Interpretación: Evidencian que variables influyeron en la clasificación o en la activación de las reglas.

5.6.2. Arquitectura

La arquitectura utilizada en el proyecto tiene 4 etapas. Cada una de ellas se utilizarán distintos componentes de Microsoft según se muestra en la figura 28.

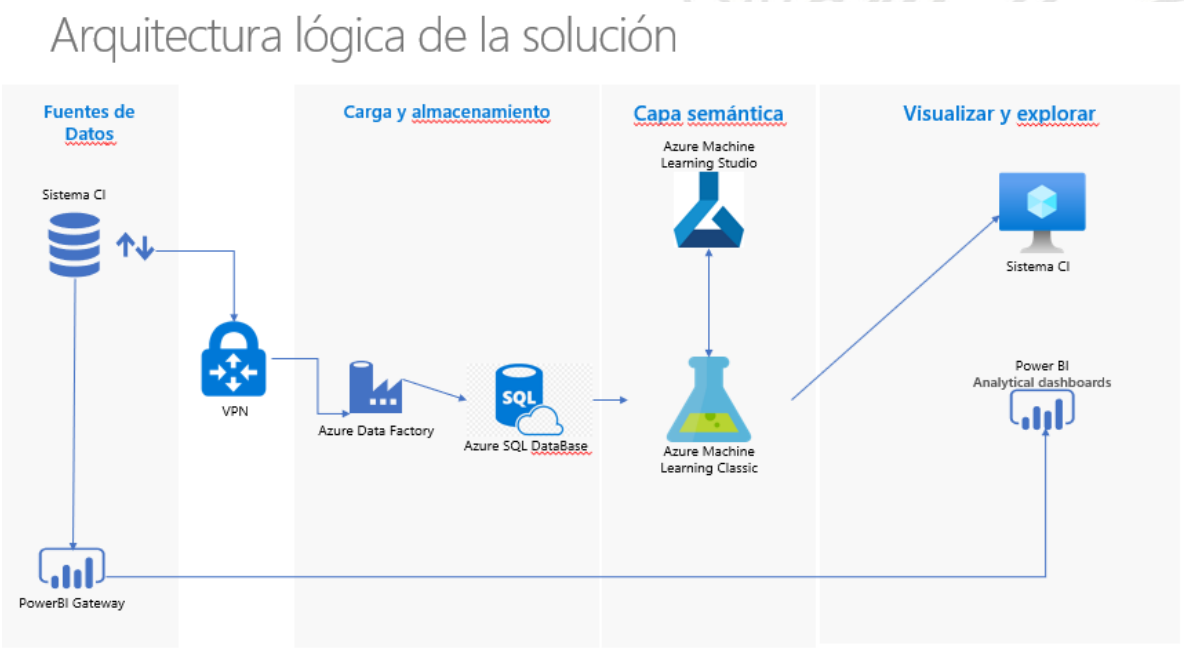


Figura 28. Arquitectura Base

Para el proyecto solo se trabajó en las 2 últimas etapas, creación de los modelos e informes de resultados.

5.6.3. Decisión final del BOT

La decisión final del BOT está compuesta por el conjunto de decisiones del modelo estadístico y el Scoring, en resumen, la orden de servicio es enviada por medio del microservicio, el modelo estadístico procesa de acuerdo con las llaves creadas y su histórico comportamental, en el campo DECISION agregará su etiqueta.

Antes del flujo del BOT dar una respuesta de salida en el microservicio, hace un llamado a otro ENDPOINT relacionado con el XGBoost como se explicó en la sección 5.5, y este modelo solo toma los insumos de las ordenes que tuvieron un estado diferente a RECHAZADO o APROBADO y actualiza la variable DECISION.

Por último, ya con todos los servicios de una respuesta en estado APROBADO O RECHAZADO se termina el flujo inicial y se da respuesta a la solicitud del microservicio.

Los modelos procesan a nivel de insumo de las ordenes, pero el comportamiento de este es analizado a nivel de Orden.

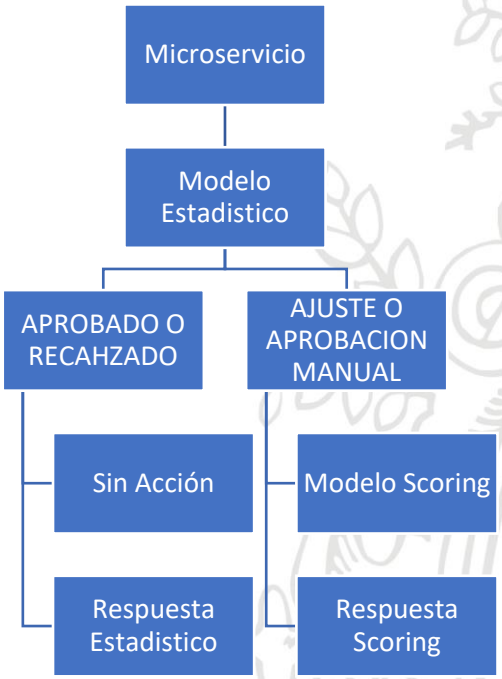


Figura 29. Diagrama de decisión del BOT

5.6.4. Puesta en productivo (Microservicio)

El BOT es una aplicación API REST la cual se llama o se utiliza por un método POST. Machine Learning Classic permite crear una API fácilmente, que permite publicación y consumo. Una vez se tiene listo y probado el flujo del BOT, se puede crear el microservicio.

Para hacer un consumo del modelo se debe hacer un llamado al END-POINT creado desde la aplicación usando una llave única de autenticación, y la información a procesar se le debe enviar en formato JSON con los campos mínimos necesarios.

6. Resultados

Los Resultados del proyecto se presentan en un tablero de control donde allí se encuentran consolidados todos los resultados respecto al comportamiento del BOT.
Las cifras de interés para “La Empresa” son, la cantidad de insumos aprobado y rechazados que componen una orden, cantidad de ordenes aprobadas completas por el BOT (Sin intervención de los analistas) y el gasto total de esas órdenes.

Una orden está aprobada en su totalidad cuando cada uno de los insumos que la componen tienen un estado de aprobado, y si una orden tiene consigo un insumo rechazado, la orden debe reprocesarla la entidad que presta el servicio o de ser necesario ser intervenida por un analista.

Los 6 tableros presentados a continuación muestran cifras porcentuales reales, pero a solicitud de "La Empresa", cifras de control de gasto (Pesos Colombianos) y cantidad de insumos u ordenes procesadas, deben ser eliminadas por políticas de confidencialidad.

Todos los tableros muestran un fragmento de este, estos incluyen otros gráficos relevantes para "La Empresa" que no serán divulgadas en este documento, los mostrados aquí son con fines académicos y algunas fueron alteradas.

La información presentada se construyó en conjunto con "La Empresa" basados en la necesidad de interpretación de la información.

6.1. Tablero de Control - Proceso de Gestión de Insumos de servicio (PowerBI)

6.1.1. Indicadores Generales

Este tablero muestra las cifras relacionadas a la cobertura a nivel de insumo y de orden y analizar cuantas ordenes se procesaron por el BOT y cuantas tuvieron que intervenir los analistas.

Es importante también revisar los falsos rechazos y falsas aprobaciones, son ordenes que el BOT en su decisión final rechazo el insumo, pero por razones internas en la compañía el analista debe reversar ese rechazo y aprobar el insumo. Por ejemplo, un insumo que la empresa prestadora del servicio intercambio los valores entre cantidad y valor, la cantidad se fue en 200 y el valor en $1.

Y, por último, y no menos importante, los tiempos de respuesta. La razón de ser de este BOT es tener una respuesta rápida a las órdenes de servicio para tener la flota el mayor tiempo posible disponible para su uso. Por ende, en la figura 30 se muestran indicadores de tiempo de respuesta de aprobación y rechazo y cuánto tiempo estuvo en ajuste por parte del cliente o por un ajuste del analista.

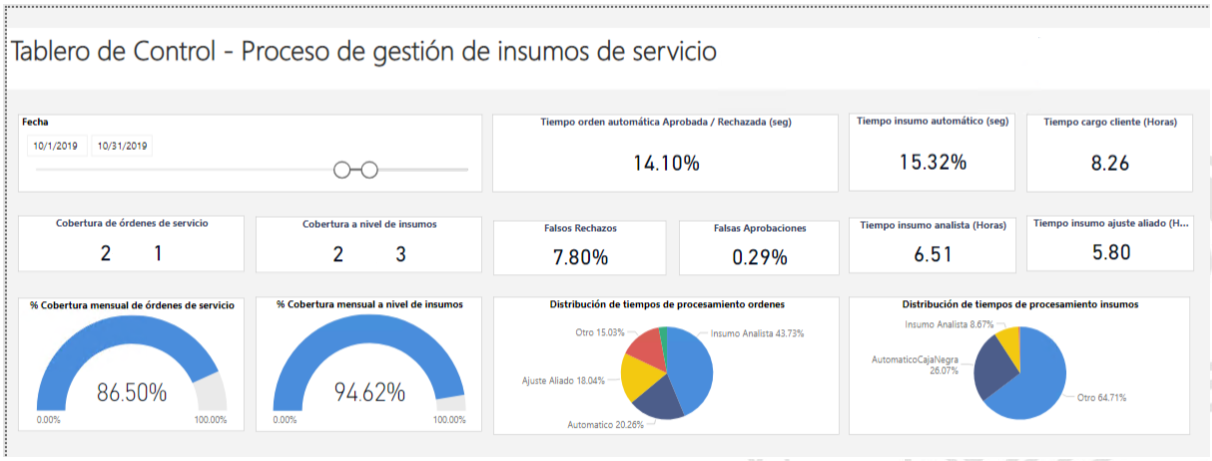


Figura 30. Tablero Indicadores Diarios

Las coberturas a nivel de orden y de insumo fueron eliminadas a solicitud de la empresa contratante iData.

6.1.2. Indicadores Diarios

En este informe se presentan el promedio del tiempo que se tarda el BOT en procesamiento de una orden para su diario operar, desde que tiene su origen en el sistema del proveedor de servicio hasta recibir la respuesta de la solicitud.

Adicionalmente cual es la participación o promedio de cobertura del BOT por cada uno de los insumos y de ordenes de todos los proveedores a nivel nacional, recordar que una orden está compuesta por uno o varios insumos.

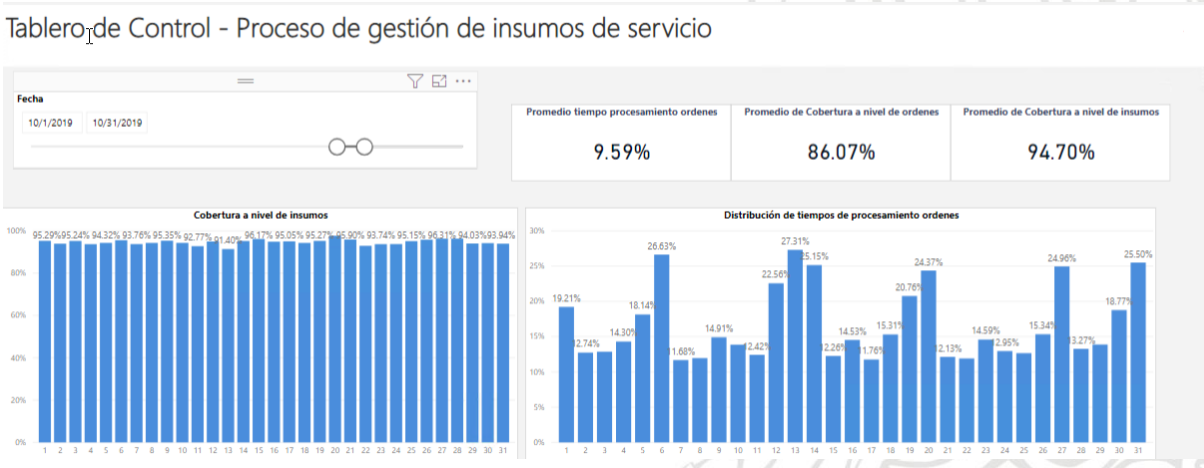


Figura 31. Tablero Indicadores Diarios

6.1.3. Indicadores de Ordenes

Este tablero tiene el detalle de la cantidad de Ordenes de servicio atendidas y la cobertura de insumos, nuevamente debieron ser eliminados por políticas de “La Empresa”.

Adicionalmente en la figura 32 se presenta el porcentaje de Falsos rechazos, Falsas aprobaciones y el promedio de tiempo que tarda un analista en responder en caso de que requiera la intervención de este.

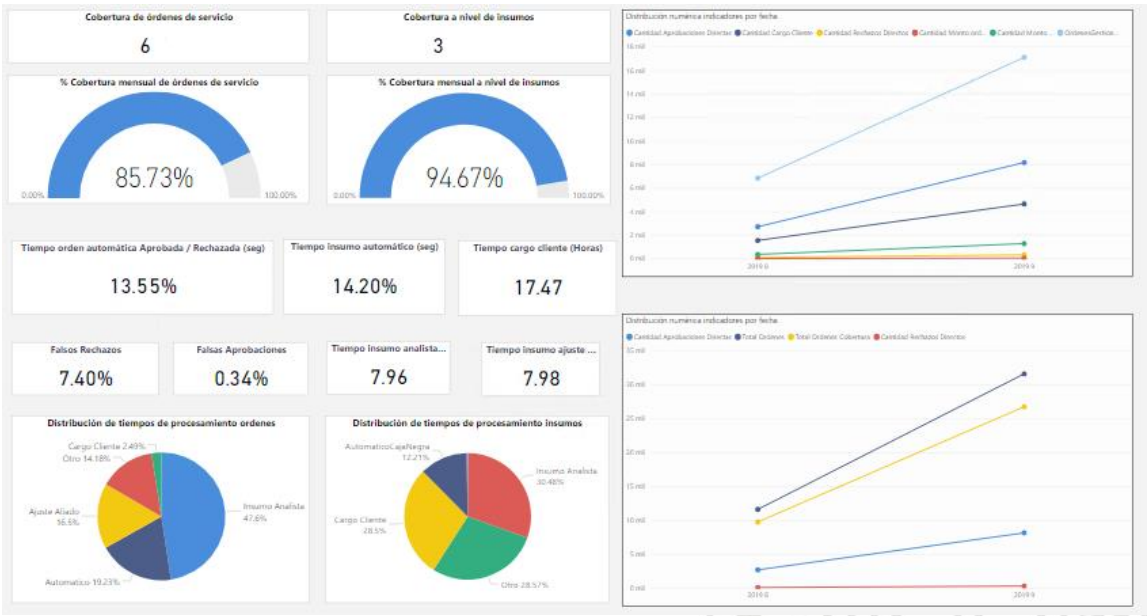


Figura 32. Tablero Indicadores de Ordenes

6.1.4. Indicadores de Dinero

Para el contenido de este tablero se presenta a modo ilustrativo ya que todas las cifras fueron borradas y/o adulteradas.

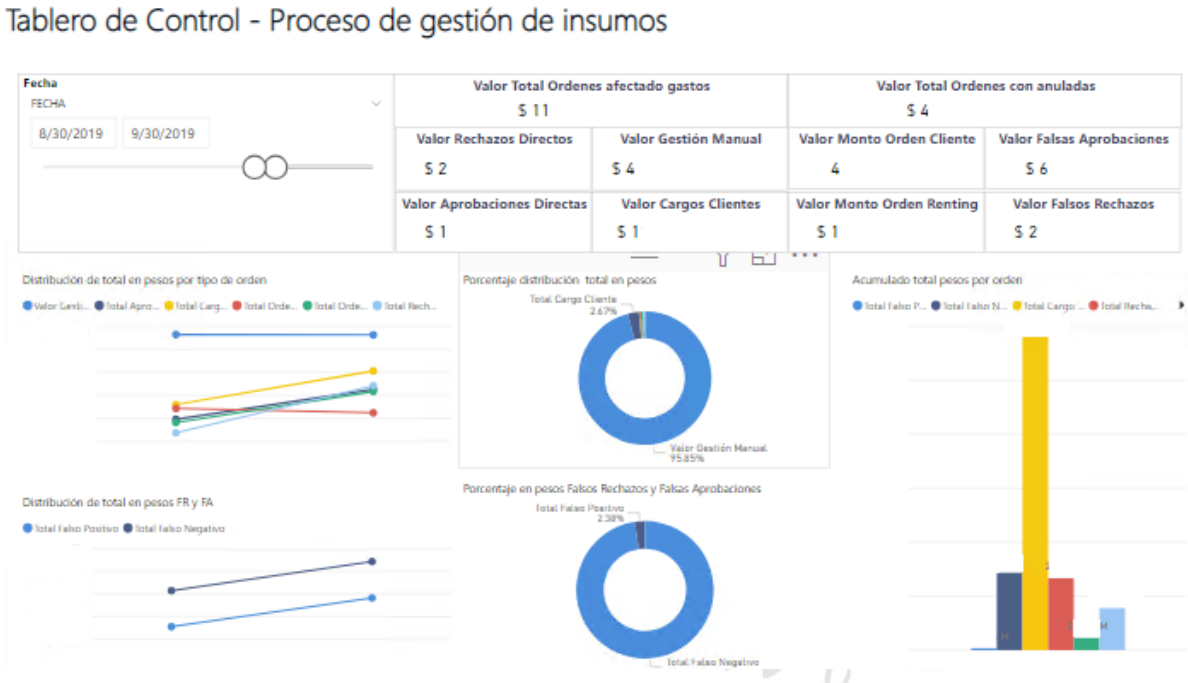


Figura 33. Tableros indicadores de Dinero

6.2. Análisis detallado comportamiento BOT

Como parte fundamental de los resultados del proyecto a petición de "La Empresa" se muestra un cuadro resumen del mes de octubre con:

- Cantidad insumos aprobados y rechazados.
- Participación de cada modelo a nivel de insumos.
- Aprobados y rechazados a nivel de Estadístico y Scoring.
- Cantidad de ordenes aprobadas completas a nivel de Estadístico y Scoring.
- Gasto ejecutado por cada modelo como el total.

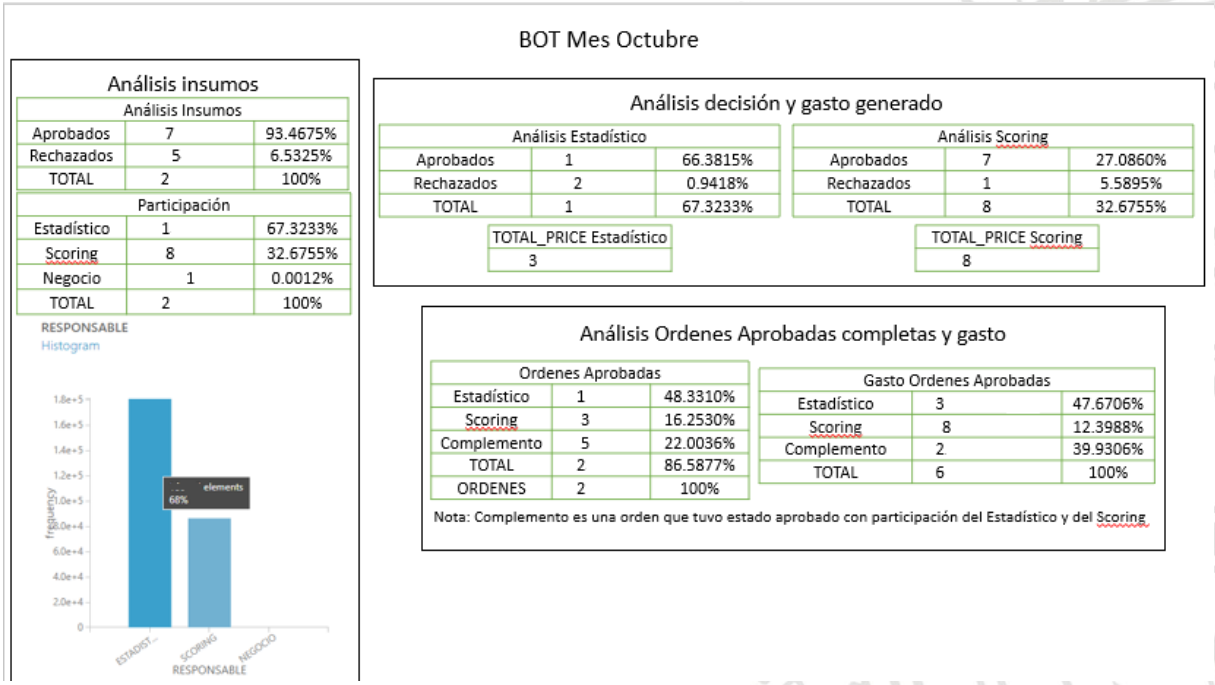


Figura 34. Resultados BOT octubre 2020

6.2.1. Análisis insumos

Basados en la Figura 34 vemos que la cantidad de insumos aprobados fue de un 93.46% y rechazados de 6.53%, aunque la cifra parece elevada al comparar con la operación normal de aprobaciones por parte de los analistas hace sentido al comportamiento del BOT.

La participación respecto a los modelos, "La Empresa" fue enfática en que la participación del modelo estadístico debía ser superior al Scoring, ya que este se basa en el comportamiento histórico de la flota y el otro es basado en la similitud con otro insumo que si posea dicho histórico.

6.2.2. Análisis de decisión

Para analizar la participación de los modelos se presenta el cuadro en la figura 34, donde se muestra la cantidad de aprobados y rechazados de cada modelo, vemos que el total de insumos aprobados o rechazados por cada modelo

coincide con la participación de cada modelo sobre el total de insumos procesados.

También se presenta una cifra de cuanto gasto representó para "La Empresa" cada una de las aprobaciones de cada modelo, cifra que debió ser eliminada.

6.2.3. Análisis de Ordenes completas

Como cifra de calificación y aprobación del proyecto "La Empresa" midió el porcentaje de participación, aprobación o rechazo y gasto generado sobre las ordenes completas.

Por ejemplo, si el BOT Aprobó una orden, pero un insumo tuvo que ser intervenido por un analista, esa orden hace parte del conteo de los analistas y se quita de las cifras de participación del BOT.

En la figura 34 se muestra entonces la cantidad de ordenes aprobadas por el modelo estadístico y el Scoring. Se presenta una cifra adicional que es llamada complemento, esta hace referencia a una orden que tuvo participación en los insumos tanto el modelo supervisado como el no supervisado

Y el ultimo cuadro y más importante es el gasto generado a "La Empresa", de las aprobaciones de cada uno de los modelos como también el complemento, explicado anteriormente.

7. Conclusiones

Se presentan las conclusiones más relevantes del proyecto realizado a nivel de practica de semestre de industria.

- Los análisis descriptivos de las variables son el proceso más importante a la hora de la implementación de cualquier modelo de Machine Learning, pero en conjunto con un conocimiento experto puede representar gran avance sobre las variables a implementar en la ejecución de modelo.
- Analizar el comportamiento de algunas distribuciones de los KPI, brindó información de la cantidad de valores atípicos presentadas en las muestras, como serian tratados y acotados fue uno de los procesos de mayor interés por parte de "La Empresa".
- La posibilidad de disponibilizar un microservicio facilitó la conectividad con el CI (Centro de Inteligencia), ya que de otra manera se tendría que hacer un desarrollo completo de un sitio web o buscar diferentes alternativas para consumo del BOT.
- En las primeras etapas de pruebas de funcionamiento no se alcanzaron grandes porcentajes de participación del modelo, esto debido a que se cogían pequeñas muestras de datos. Se alcanzo el índice de participación mostrado cuando se entrenó el modelo Estadístico con muestras de dos años y medio hacia atrás y un año para el entramiento del modelo de Scoring.
- Para los resultados de las pruebas del mes de octubre mostradas en la sección 5.6 "La Empresa" quedo satisfecha respecto a la participación y control del

gasto, cumpliendo el objetivo general el proyecto, que busca reducir la carga operativa de los analistas, enfocándose en otras actividades que permitan escalar la productividad de la compañía.

- Las visualizaciones permiten una interpretación rápida y sencilla de indicadores relevantes para "La Empresa", se configuraron con actualización automática que permita revisar cifras del último mes cargado en base de datos.
- Los tiempos de respuesta y carga operativa de los analistas se redujo, ahora ordenes que tardaban hasta 40 minutos en recibir respuesta, se tiene una respuesta inmediata de parte del BOT. Los Analistas dejaron de procesar casi el 80% de las ordenes diarias para desempeñar otro tipo de actividades.

8. Anexos

Anexo 1: Analisis_Combinatoria.xlsx

9. Bibliografía

A., M. M. (Febrero de 2015). *anestesiar.org*. Obtenido de https://anestesiar.org/2015/una-tarea-imposible-la-tecnica-de-booststrapping

D., R. (13 de Julio de 2018). *Analyticslane Aprendizaje Supervisado y no Supervisado*. Obtenido de https://www.analyticslane.com/2018/07/13/aprendizaje-supervisado-y-aprendizaje-no-supervisado/

DataCamp. (8 de noviembre de 2019). *Data Camp*.

Freidman J., H. T. (2008). Elements of Statistical Learning Data Mining, Inference and Prediction. 1.

Microsoft. (11 de 04 de 2019). *Microsoft Docs*. Obtenido de https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-ml

Microsoft Azure. (Enero de 2021). *¿Que es Azure?* Obtenido de https://azure.microsoft.com/es-es/overview/what-is-azure/

PWC and Microsoft. (Marzo de 2018). *PWC*. Obtenido de https://www.pwc.es/es/publicaciones/tecnologia/assets/pwc-ia-en-espana-2018.pdf

Sergas. (Octubre de 2014). *Epidat 4: Ayuda de Distribuciones de probabilidad*. Obtenido de https://www.sergas.es/Saude-publica/Documents/1899/Ayuda_Epidat_4_Distribuciones_de_probabilidad_Octubre2014.pdf

Yadav, D. (6 de Diciembre de 2019). *Towards Data Science*.