INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO
GERENCIA DE CAPITAL HUMANO
POSGRADOS

"EFICIENCIA EN LOS PROCESOS DE COMPRA Y VENTA DE UNA EMPRESA DEL SECTOR DE REFACCIONES MEDIANTE EL USO DE CIENCIA DE DATOS"

REPORTE ANALÍTICO DE EXPERIENCIA LABORAL
Que para obtener el grado de MAESTRO EN CIENCIA DE DATOS E INFORMACIÓN

Presenta:

José Alejandro Solís Marrufo

Asesor:

Dr. Raúl Arturo Alvarado López

Ciudad de México, septiembre de 2019.

Autorización de Impresión

INFOTEC
POSGRADOS

AUTORIZACIÓN DE IMPRESIÓN Y NO ADEUDO EN BIBLIOTECA

MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

Ciudad de México, 10 de febrero de 2020
INFOTEC-DAIC-GCH-SE0116/2020.

La Gerencia de Capital Humano / Gerencia de Investigación hacen constar que el trabajo de titulación intitulado

EFICIENCIA EN LOS PROCESOS DE COMPRA Y VENTA DE UNA EMPRESA DEL SECTOR DE REFACCIONES MEDIANTE EL USO DE CIENCIA DE DATOS

Desarrollado por el alumno **José Alejandro Solís Marrufo** y bajo la asesoría del **Dr. Raúl Arturo Alvarado López**; cumple con el formato de biblioteca. Por lo cual, se expide la presente autorización para impresión del proyecto terminal al que se ha hecho mención.

Asimismo se hace constar que no debe material de la biblioteca de INFOTEC.

Vo. Bo.

Mtra. Julieta Alcibar Hermosillo
Coordinadora de Biblioteca

Anexar a la presente autorización al inicio de la versión impresa del trabajo referido que ampara la misma.

C.p.p Servicios Escolares

Agradecimientos

Agradezco primeramente a mis papás, por su amor, paciencia, esfuerzo, apoyo y motivación que siempre me han brindado a lo largo de mi vida.

Agradezco a la empresa y mis compañeros del trabajo de los cuales aprendí mucho y me dieron consejos para ser mejor profesionista.

Finalmente, agradezco a todos mis seres queridos, amigos y familia que me apoyaron y dieron ánimos para terminar la maestría.

Tabla de contenido

Índice de figuras

Índice de gráficos

Índice de tablas

Siglas y abreviaturas

(IDC): *International Data Corporation* (Corporación de Datos Internacionales).
(GLM): *General Lineal Model* (Modelo General Lineal).
(GPR): *Ground Penetrating Radar* (Georradar).
(POC): *Proof of Concept* (Prueba de Concepto).
(TI): Tecnología de la Información.
(CRISP – DM): *Cross Industry Standard Process for Data Mining* (Proceso Estándar Inter-Industrias de Minería de Datos).
(SPSS): *Statistical Package for the Social Sciences* (Paquete Estadístico para las ciencias sociales).
(DSS): *Decision Support System* (Sistema de Soporte de Decisiones).
(VP): Verdaderos Positivos.
(VN): Verdaderos Negativos.
(FP): Falsos Positivos.
(FN): Falsos Negativos.
(XGBoost): *Extreme Gradient Boosting* (Impulso de Gradiente Extremo).
(GBT): *Gradient Boosted Tree*. (Árboles Impulsados por Gradiente).
(MAE): *Mean Absolute Error* (Error Absoluto Medio).
(MAPE): *Mean Absolute Percentage Error* (Error Porcentual Absoluto Medio).
(MSE): *Mean Square Error* (Error Cuadrático Medio).
(RMSE): *Root Mean Square Error* (Raíz del Error Cuadrático Medio).
(RMSLE): *Root Mean Square Logarithmic Error* (Raíz del Error Logarítmico Cuadrático Medio).
(SQL): *Structured Query Language* (Lenguaje de Consulta Estructurado).

Introducción

Planteamiento de problema

Con el crecimiento de las tecnologías de la información y comunicación, cada vez se genera mayor cantidad de datos, pues los datos son el corazón de la transformación digital los cuales deben ser aprovechados por las empresas y la industria para lograr mejores objetivos y en algunos casos seguir siendo competitivos (Reinsel, Gantz, & Rydning, 2018). La transformación digital es un cambio cultural y estratégico, mediante el cual las empresas u organizaciones se orientan a mejorar la experiencia de sus clientes y/o a la creación de nuevos modelos de negocios, a través de la incorporación de tecnologías digitales, para ofrecer soluciones más eficaces, innovadoras, rápidas y rentables (PMG Business Improvement, 2019).

Este cambio cultural se debe dar ya que el crecimiento de la generación de datos se está dando de manera exponencial, de acuerdo con la consultora IDC para 2025 se prevee que se estarán generando 175 zettabytes de datos, que vendría siendo 5 veces mayor a los 33 zettabytes generado en el 2018 (Reinsel, Gantz, & Rydning, 2018).

Esto ha permitido que las empresas e industrias y otras organizaciones como universidades y el mismo gobierno recurran a los datos y el análisis para lograr objetivos y tareas que antes se creían que no se podían realizar, sin embargo, esto genera la problemática de encontrar el talento, conocimiento, herramientas y seguridad para la realización de estas tareas (Deoras, 2019).

La ciencia de datos se enfoca en la generación de conocimiento a partir de los datos (Dhar, 2013). Es una disciplina emergente y de gran pertinencia para todas las organizaciones que deseen codificar el valor oculto e intangible de sus datos. Hoy más que nunca estamos más conectados con personas y dispositivos, tenemos acceso a mejores redes y servicios, y sin duda consumimos y producimos mayores cantidades de datos e información. Por lo que se requiere contar con las habilidades, conocimientos, experiencias y técnicas de los científicos de datos para procesar, analizar y visualizar de formas más inteligentes los datos en información,

promoviendo así, más y mejores conocimientos de nuestra realidad en sus contextos (Moreno Salinas, 2017). La ciencia de datos surge en la industria por la necesidad de analizar mayor cantidad de información, es interdisciplinaria y está basado principalmente en 3 habilidades: matemáticas y estadística, programación y experiencia en el tema.

Las habilidades de estadística permiten la creación de modelos y manejo de los dataset, las habilidades de programación permiten el diseño de algoritmos de manera eficiente para su almacenaje, procesamiento y visualización de los datos; y las habilidades de experiencia en el tema permite la comprensión del problema y hacer las preguntas adecuadas para encontrar las respuestas que se desean obtener.

Cuando se combinan las habilidades de programación con las de matemáticas y estadística surge el aprendizaje de máquina, cuando se combina las habilidades matemáticas y estadísticas con la experiencia en el tema surge la investigación tradicional, pero si combinamos las habilidades de programación con las de experiencia se entra en una zona de peligro debido a que los resultados obtenidos no pueden ser confiables ya que no cuenta con un sustento matemático que lo valide ocasionando que los modelos al ejecutarlos no se acerquen a los resultados esperados, esto se puede observar en la figura 1.

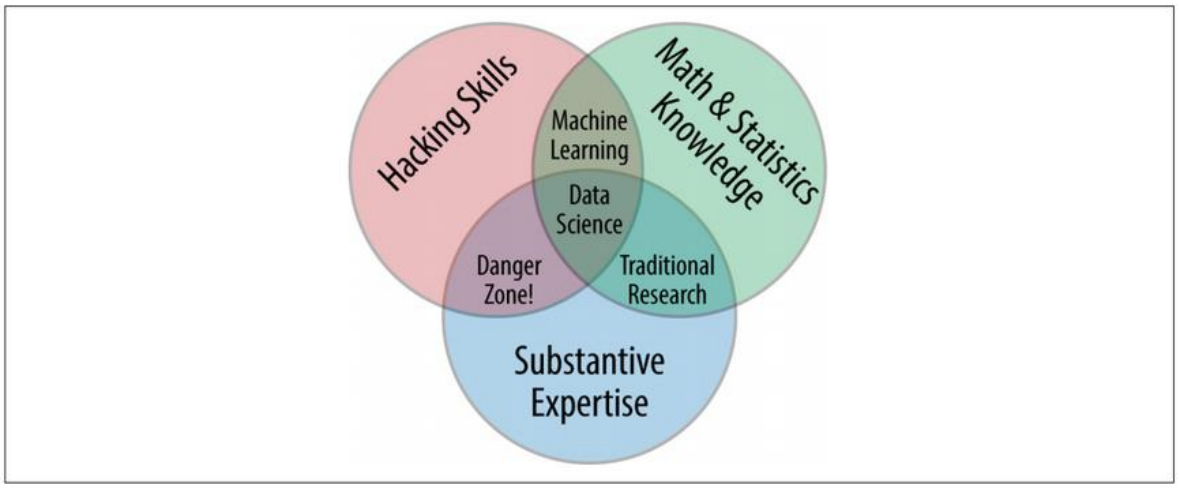


Figura 1.- Diagrama de Venn de habilidades de Ciencia de Datos. Fuente: Jake (2017, pág. 4).

El aprendizaje automático (*machine learning*) es un campo de estudio de la inteligencia artificial que consiste en desarrollar técnicas que permiten a las computadoras aprender, utilizando información estructurada o no estructurada (Camargo, Camargo, & Joyanes, 2015). Se divide en aprendizaje supervisado y aprendizaje no supervisado.

El aprendizaje supervisado implica modelar de alguna manera la relación entre lo medido características de los datos y alguna etiqueta asociada con los datos; una vez que este modelo es determinante extraído, se puede usar para aplicar etiquetas a datos nuevos y desconocidos. Se subdividen tareas de clasificación y tareas de regresión: en la clasificación, las etiquetas son categorías discretas, mientras que, en regresión, las etiquetas son cantidades continuas. Algunos ejemplos de algoritmos de regresión son: Regresión Lineal, *GLM* (Modelo Lineal Generalizado), Árboles de Decisión, Máquinas de Soporte Vectorial, Redes Neuronales, etc. Algunos ejemplos de algoritmos de clasificación son: Máquinas de Soporte Vectorial, Naive Bayes, Vecinos más Cercanos, Análisis de Discriminante, etc (Jake, 2017).

El aprendizaje no supervisado implica modelar las características de un conjunto de datos sin referencia a cualquier etiqueta, y a menudo se describe como "dejar que el conjunto de datos hable por sí mismo". Estos modelos incluyen tareas como el agrupamiento y la reducción de dimensionalidad. Algoritmos de agrupamiento identifican distintos grupos de datos, mientras que los algoritmos de reducción de dimensionalidad buscan representaciones más sucintas de los datos (Jake, 2017). Algunos ejemplos de algoritmos de agrupamiento son: K-Medias, Cadenas de Markov escondidas, Redes Neuronales, etc.

En la figura 2 se muestra como se dividen y clasifican estos algoritmos.

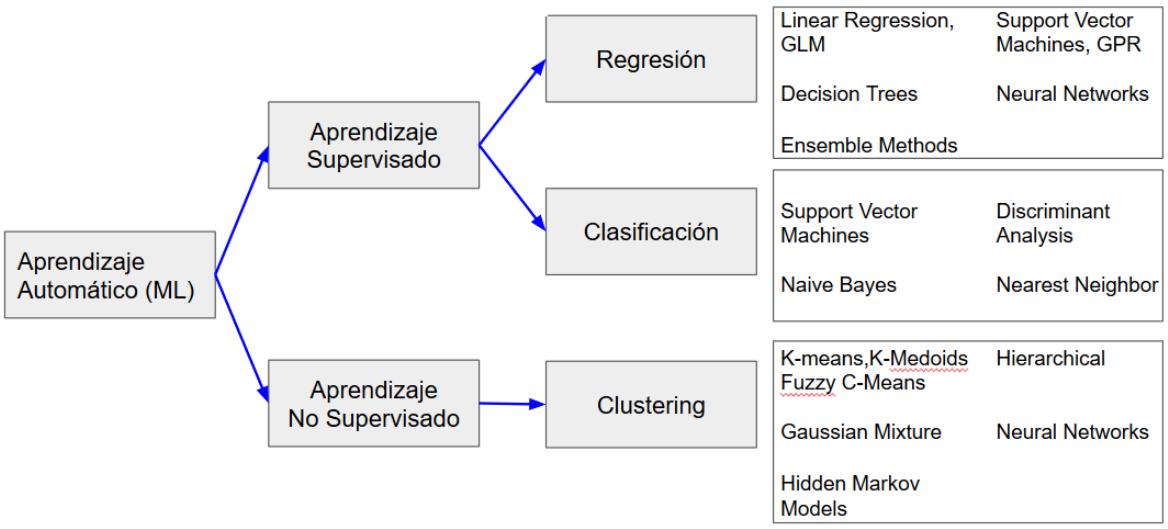


Figura 2.- División y algoritmos del Aprendizaje Automático. Fuente: Plenumsoft (2018)

Objetivos

Objetivo general

- Documentar la experiencia laboral sobre proceso que se lleva a cabo para que una empresa del sector de refacciones conozca la importancia de la análitica computacional en la toma de decisiones de las empresas para incrementar su productividad y la toma de decisiones.

Objetivos específicos

- Describir el Proyecto de predicción de Obsolescencia del stock usando un algoritmo de clasificación para la eficiencia de su proceso y adopción de la plataforma de Ciencia de Datos Dataiku.
- Describir el Proyecto de predicción de Compras y Ventas usando un algoritmo de regresión para la eficiencia de su proceso y adopción de la plataforma para Ciencia de Datos Dataiku.
- Explicar los beneficios que se pueden obtener al utilizar ciencia de datos y las recomendaciones a seguir.

Metodología

Para realizar este reporte analítico de experiencia laboral se utilizaron bases de datos brindadas por un cliente para realizar distintas pruebas de concepto mejor conocidas como *POC´s* que le permitieran saber la capacidad de la herramienta Dataiku DSS y asegurarse de que sea la herramienta adecuada para implementar en la empresa, cabe mencionar que este reporte se centra en el análisis de un caso de estudio el cual fue elegido por ser realizado para empresa de gran envergadura en el sector y ser de los proyectos más maduros con los que se contaba. Para más detalle sobre la herramienta ver el Anexo 1.

La metodología a utilizar fue estudio de tipo cuantitativo pues las *POC´s* que se detallan en este reporte buscan predecir la obsolescencia, ventas y compras que se tendrán de los productos de la empresa utilizando variables numéricas y categóricas.

Al final se presentan los hallazgos derivados de la demostración al cliente con los resultados obtenidos al observar el funcionamiento de la herramienta, se resuelvan dudas respecto a requerimientos y conocimientos necesarios para su uso y la capacidad del algoritmo creado mediante métricas que al compararse con los procesos que se llevan a cabo por parte del cliente, tiene mejor eficiencia.

El presente reporte analítico de experiencia laboral es realizado para obtener el título de Maestro en Ciencia de Datos e Información por parte del INFOTEC. El documento consta de cuatro capítulos, primeramente, se presenta los conceptos generales que aborda el documento, los cuales incluye una descripción de la empresa y la metododología que se utiliza, luego, en los capítulos 2 y 3 se describen los proyectos mencionados en los objetivos específicos, proyecto de obsolescencia y proyecto de ventas y compras respectivamente; el capítulo 4 contiene los beneficios y recomendaciones que brinda a las empresas el utilizar ciencia de datos. Finalmente, se presenta la conclusión del documento y el anexo donde se presenta la plataforma de ciencia de datos Dataiku DSS.

Capítulo 1

Descripciones Generales

Capítulo 1. Descripciones Generales

Este capítulo aborda la descripción de la empresa Plenumsoft, la metodología utilizada para llevar a cabo las pruebas de concepto y la metodología de investigación de este documento.

1.1 Descripción de la empresa

Grupo Plenum es un corporativo mexicano fundado el 23 de junio de 1995, cuenta con 24 años de experiencia, especializado en proveer servicios de consultoría y soluciones para automatización de procesos de Tecnologías de la Información (TI) y telecomunicaciones en diversos sectores, conformado por empresas de alta especialidad como Plenumsoft, Plenumsoft Marina, Plenumsoft Energy, Cytron Medical y Avanx.

Plenumsoft es la empresa de Grupo Plenum especializada en el desarrollo e innovación de soluciones integrales para varios grupos de negocio, enfocados en resolver problemas interdisciplinarios complejos a través de la detección, elaboración de cotizaciones y generación de soluciones. Entre los servicios que ofrece están consultoría en Ciencia de Datos, soluciones con Inteligencia Artificial, desarrollo de productos y proyectos, modernización de aplicaciones y gestión de seguridad, información y recursos empresariales.

En la empresa trabajo en el área encargada de Ciencia de Datos, la cual se enfoca principalmente en la venta de la plataforma para Ciencia de Datos y Big Data Dataiku DSS, ya que la empresa es la encargada de su venta y distribución para toda Latinoamérica y de brindar soluciones en Ciencia de Datos e Inteligencia Artificial para las empresas que lo requieran. La metodología que se utiliza para la venta de la herramienta Dataiku DSS son las *POC´s*, estas son un pequeño ejercicio o demostración para probar la idea de diseño o supuesto. El objetivo principal del desarrollo de una *POC* es demostrar la funcionalidad y verificar cierto concepto o teoría que se puede lograr en el desarrollo (Singaram & Jain, 2018). Normalmente son gratuitas para que el cliente se interese por conocer más de la herramienta y tenga mayor seguridad de que esta cumple con lo esperado.

Las *POC´s* pueden ser confundidas con el desarrollo de prototipos, sin embargo, la principal diferencia es que mientras que una *POC* muestra que se puede desarrollar un producto o una característica, un prototipo muestra cómo se desarrollará (Singaram & Jain, 2018).

Se neceistan 2 cosas principalmente para plantear una *POC* adecuada para el cliente:

- Conocer el giro de negocio.
- Razón por la cual están interesados en una herramienta en Ciencia de Datos.

Conociendo estas 2 cosas es posible plantear un proyecto en el que se demuestre las capacidades de la herramienta para su realización, o plantear un proyecto o solución a un problema que no estaba considerado debido a que no se tenía en mente el alcance que puede tener la Ciencia de Datos.

Cabe destacar que al ser una *POC* el cliente puede ver los resultados y la capacidad que se quería probar, pero no los algoritmos y el proceso que se llevó a cabo para obtenerlos al menos que la *POC* sea comprada por el cliente.

Para llevar a cabo la *POC* de Ciencia de Datos se usa principalmente una metodología conocida como *CRISP – DM (Cross Industry Standard Process for Data Mining)* la cual adaptamos para resolver los proyectos de este tema, ya que es un método probado para orientar los trabajos de minería de datos e incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas. Por otro lado, como modelo de proceso, *CRISP-DM* ofrece un resumen del ciclo vital de minería de datos, se puede observar en la figura 3.

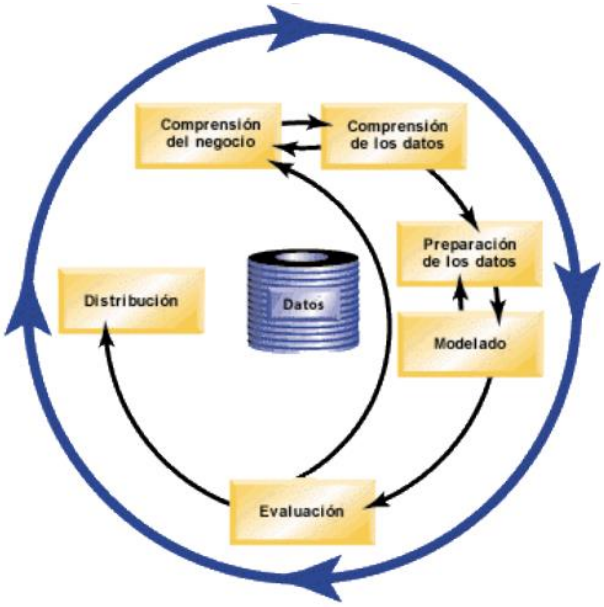


Figura 3.- Proceso CRISP-DM. Fuente: SPSS (2012, pág. 1).

Las etapas que contiene son la comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y distribución. Como se puede observar es un modelo muy flexible, en el que si es necesario es posible regresar a una etapa previa.

En la etapa de comprensión del negocio se determinan los objetivos comerciales de la empresa, los cuales vendrían siendo los objetivos de la *POC*, también se obtiene información acerca del cliente y del problema a resolver.

En la etapa de comprensión de los datos implica el acceso a ellos, descripción, exploración y visualización mediante tablas, gráficos, mapas, etc.

La etapa de preparación de los datos es normalmente la más larga, consta de la selección de datos, limpieza de datos, construcción de nuevos datos, transformación de datos, integración, etc.

La etapa de modelado incluye la selección de la técnica de modelado, parámetros, selección de métricas bajo las que hay que optimizar el modelo y generación del modelo.

La etapa de evaluación consta de la revisión del desempeño de los modelos de acuerdo con los criterios establecidos con el objetivo de seleccionar el mejor de ellos.

Finalmente, la etapa de despliegue consta de 2 partes, la planificación y control del despliegue de los resultados y la presentación al cliente. (SPSS, 2012, pág. 1)

El modelo adaptado utilizado en la empresa consta de las mismas etapas que el modelo original, la diferencia es que extraemos algunas actividades relevantes para los proyectos de Ciencia de Datos y las convertimos en etapas como se observa en la figura 4, una la denominamos como adquisición de los datos en la cual se encuentra el proceso a seguir para obtener los datos que se encuentran en los servidores del cliente o repositorios, si deben ser descargados o realizar una conexión a su instancia y como transformarlos para poderlos ingresar en la plataforma Dataiku DSS; la otra de las actividades es la que se encuentra entre las etapas de Evaluación y Despliegue y es la Visualización de los Resultados, ésta se refiere a la manera en la que se presentará los resultados obtenidos al cliente.

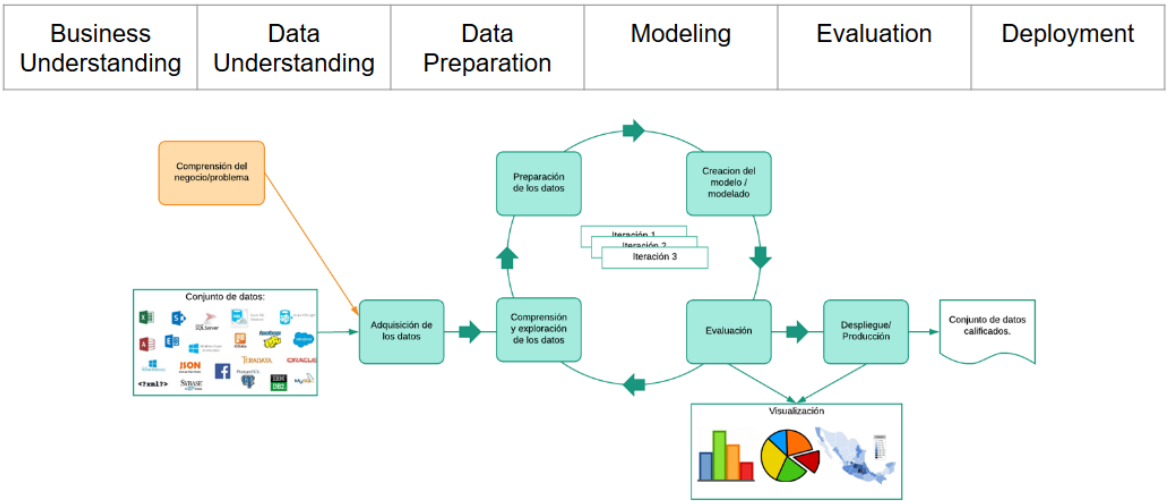


Figura 4.- Metodología Plenumsoft. Fuente: Plenumsoft (2018, pág. 2).

Me incorporé a la empresa Plenumsoft en marzo del 2017 en mi último año de la licenciatura en Actuaría de la Facultad de Matemáticas de la Universidad Autónoma de Yucatán, en el área de Planeación Estratégica. En un inicio, mis actividades consistían identificación de objetivos, construcción de mapas

estratégicos y creación de indicadores. Por la misma época se empezó a desarrollar el departamento de Ciencia de Datos en la empresa, y fue hasta que terminé la licenciatura e ingresé a la maestría en Ciencia de Datos e Información que fui trasladado a este departamento. Actualmente llevo 2 años (noviembre 2017 a la fecha) en el departamento de ciencia de datos con el puesto de científico de datos realizando actividades como impartición de cursos de ciencia de datos, realización de *POC´s*, análisis exploratorio de datos, modelado y minería de datos. Aproximadamente he participado o estado a cargo de 9 proyectos en este período de tiempo.

Actualmente vivimos en una era digital, donde en casi todo momento de nuestra vidad diaria estamos en contacto con dispositivos o plataformas que generan datos continuamente de nuestra persona, actividades, sentimientos, gustos y el medio que nos rodea, y el poder aprovechar estos datos y convertirlos en información permitirá generar diferenciadores en los productos de las empresas, mejores estrategias de marketing y de dirección para la toma de decisiones de manera que se consiga el logro de los objetivos o soluciones a los problemas de las empresas. De aquí que la empresa Plenumsoft impulsa el aprendizaje, capacitación y proyectos relacionados con Ciencia de Datos.

Capítulo 2

Proyecto “Predicción de Obsolescencia”

Capítulo 2. Proyecto “Predicción de Obsolescencia”

Para este reporte se decidió abordar sobre la aplicación de ciencia de datos realizados para una empresa que tiene gran importancia para el sureste del país, pues se trabajaron tanto algoritmos de regresión como de clasificación y el análisis y desarrollo de estos algoritmos abarca la metodología y procesos que realizo como trabajador de la empresa Plenumsoft.

Por cuestiones de privacidad de los datos no se dará a conocer el nombre de la empresa ni los datos personales de los clientes; por lo tanto, para este proyecto de “Predicción de Obsolescencia” diremos que fue realizada para la empresa “A”.

La empresa “A” se encarga de la comercialización, arrendamiento y servicio en el sector automotriz, incluyendo automóviles, camiones, motores, maquinaria, equipo de construcción. Adicionalmente, se encarga de la distribución y venta de refracciones a nivel nacional. Por ello debe administrar de forma adecuada el inventario de sus productos con base en la demanda teniendo en cuenta el suministro de forma puntual y eficiente del material requerido por cada una de sus sucursales. Sin embargo, existen productos clasificados como “obsoletos” que ya no se encuentran en constante petición, y por lo tanto el espacio que ocupan en almacén podría aprovecharse de mejor manera.

La *POC* de la plataforma Dataiku DSS se enfoca en implementar una propuesta para satisfacer esta necesidad, es decir, se propondrá un flujo de datos para la clasificación múltiple de las piezas con el fin de identificar aquellas que pudieran potencialmente convertirse en refacciones obsoletas aplicando algoritmos de Aprendizaje Automático.

2.1 Objetivo del Proyecto de Obsolescencia de la empresa “A”

Identificar las piezas potencialmente obsoletas mediante el diseño de un flujo de datos, en específico, el entrenamiento y evaluación de modelos de aprendizaje automático de modo que permitan coadyuvar a la empresa “A” en el diseño de estrategias relacionadas con la administración oportuna de sus inventarios. Durante el proceso se definirá e implementará un flujo de trabajo que permita la limpieza,

pre-procesamiento y exploración del histórico de transacciones refaccionarias (más de 10 años) de la empresa “A” para la identificación de patrones relacionados con la clasificación de las refacciones, incluyendo las que se consideran obsoletas. También se seleccionará, entrenará y evaluará modelos de aprendizaje automático para la correcta clasificación de las refacciones de la empresa “A”, en específico para la detección de refacciones obsoletas, a partir de un conjunto de entrenamiento tomado a partir del histórico de transacciones de dicho corporativo.

2.2 Solución propuesta

La solución propuesta a llevar a cabo en la *POC* consiste en desarrollar un flujo de trabajo para la clasificación de las refacciones de la empresa “A” aplicando las etapas del proceso de Ciencia de Datos.

El diagrama general de la *POC* se presenta en la figura 5, la solución incluye el desarrollo de las siguientes características:

1. Conexión con la tecnología de Base de Datos proporcionada por la empresa “A”.

2. Limpieza y pre-procesamiento de las Bases de Datos propuestas para el entrenamiento del Modelo de Clasificación.

3. Entrenamiento de Modelos de Aprendizaje Automático para la Clasificación de las Refacciones.

4. Diseño e implementación de un Escenario para la automatización de la aplicación del Flujo de Trabajo, incluyendo la clasificación de nuevos datos mediante el Modelo de Clasificación con mejor eficiencia encontrado.

5. Despliegue visual de los resultados mediante un Panel de Visualización (*Dashboard*).

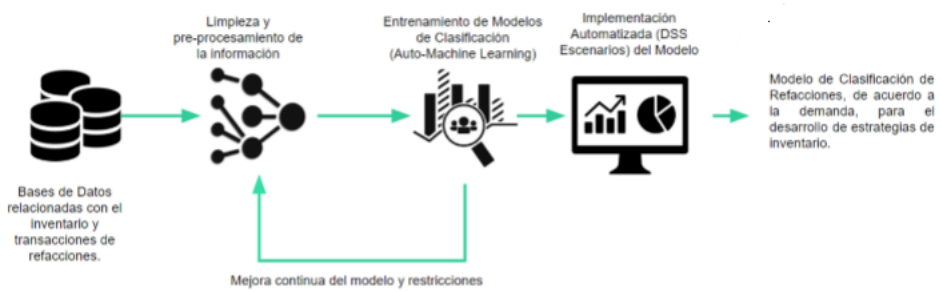


Figura 5.- Diagrama del Flujo de Trabajo para el proyecto Predicción de Obsolescencia. Fuente: Plenumsoft (2019).

2.3 Desarrollo del proyecto

Antes de describir el proceso llevado a cabo para la *POC* se explicará la metodología utilizada por la empresa “A” para definir los productos que se convertía en obsoletos.

En la empresa “A” existen 7 clasificaciones para los productos del inventario:

- O = Obsolescencia.
- N = Nuevo
- P = Pendiente
- A, B, C, D = Nivel de acuerdo a la demanda del producto por parte de los clientes.

Las reglas de la clasificación son las siguientes:

- Un producto mantendrá su clasificación Pendiente o “Nueva” de acuerdo al número de meses configurados (7 meses). Después de ese tiempo se clasifica de acuerdo con su nivel de demanda donde A implica mayor demanda.
- Se cambiará la clasificación "N" (Nuevo) a "P" (Pendiente) cuando un producto sea "N" en una sucursal, pero ya haya sido clasificada en otra sucursal.
- Se clasifica como “O” (Obsolescencia) un producto el cual el número de meses desde la última entrada a la fecha de clasificación supere la

cantidad de meses configurados como “meses para O” y no tenga ventas en este periodo (24 meses).

- Se clasifica como D cuando el número de meses con ventas sea menor o igual a lo configurado como "meses para D". O cuando no haya sido clasificada anteriormente y se rebase su periodo de observación (clasificación N), siempre y cuando se respete el punto anterior.

En este caso para la *POC*, el objetivo es predecir los productos que obtendrán la clasificación de obsolescencia los siguientes 3 meses, es decir, los productos que no tendrán demanda o entradas en los siguientes 3 meses y tengan un acumulado de al menos 21 meses en el que no se hayan demandado. También la *POC* fue acotada para que se realice sobre 2 de las empresas que conforman a la empresa “A”.

2.3.1 Integración de Base de Datos

Las bases de datos constituyen un sistema de proceso de datos cuyo objetivo básico es el de conservar información y mantenerla disponible para su acceso de forma eficiente. El interés de los usuarios por la información contenida en una base de datos es debido, normalmente, a su significación en los procesos de toma de decisiones. Las aplicaciones de bases de datos tienen cuatro componentes principales: datos, programas, dispositivos de almacenamiento y usuarios (Alonso Martínez, 1992).

Para ambos proyectos que se abordan en este documento, los datos fueron otorgados por la empresa “A”, estos se encuentran almacenados en su servidor en fomato sql y para poder accesar a ellos se creo una conexión entre su servidor y el servidor de la empresa que contiene la herramienta Dataiku DSS para poder descargar los dataset.

Para este proyecto la empresa “A” proporcionó 4 dataset que podemos observar en la figura 6:

- Almacenes

- Empresas
- Histórico de demandas
- Artículos

Del dataset almacenes tenemos las siguientes variables: IdAlmacen, IdEmpresa, SucursalId, Descripción, Activo, FechaOperacion, ServerName, BDName, AreaId, TipoAlmacen, DireccionId, ProveedorId, SucProvId, InterEmpresas, AlmacenAlterno, GrupoAlmacen, SurteRemision, ObsDemandas, FechaDemandas, SurtePDA, ConfDemandas, pToleranciaEstadístico, confABC_O, confABC_P, ActualizaABC, DeptoRefId, UUA, FUA, Consigna, FechaImpresionPiso, UsrImpresionPiso, TipoAlmacen, ZonaId, OrdenZona, WAREHOUSE_ID, TipoAlmacenId, Ocurrencias_D6, Ocurrencias_D12, Ocurrencias_D24, Calculo24Meses.

Del dataset empresas tenemos las siguientes variables: EmpresaId, RazonSocial, NombreCorto, RFC, CURP, Direccion, Ciudad, Poblacion, Estado, CodPost, Telefono, Email, Servidor, NomBaseDatos, RutaBD, MascCta, NumNiveles, DesoElimOpc, Decimales, Secuencial, FolioUnico, ContabEnLinea, AltaCtasEnTras, PerAjuste, AltaCtasEnPol, Deptos, Multimoneda, SugSigNumCta, UtilCatRanCta, UltPeriodoCerrado, PeriodoActivo, EjercicioActivo, UltPeriodoAbierto, FolSolCheque, EmpRefacciones, CuadreId, FolioProgPago, UtilidadPerdidaID, CuotaFija, PorcRetIVA, NumMaxLineasxFac, EmpAfectadas, PorcUtilEmpServ, DeServicios, SociedadCivil, PorcSupISPT, PorcInfISPT, ConAcumISPTid, SectorID, SegContable, Colonia, EmpaqueSurtIdeal, ActivoBI.

Del dataset histórico demandas tenemos las siguientes variables: EmpresaId, SucursalId, AlmacenId, ArticuloId, Anio, Mes, Inicial, Final, Demandas, Perdidas, Negadas, Devoluciones, CostoIni, CostoFin, Entradas, Salidas, CantCompra, CantDevolucion, CostoCompra, CostoDevolucion, NivelABCId, Cantconsignas.

Del dataset artículos tenemos las siguientes variables: ArticuloId, CveArt, Descripcion, codBarras, LineaId, GrupoId, SubGrupoId, AntecesorId, NivelABCId, Peso, Precio, Costo, MonedaId, UnidadCompraId, UnidadVentaId, ProveedorClave, Obsoleto, TipoSoporteId, CoreId, CantXUnidad, TipoArt, Comentarios, FechaAlta,

Usuarioid, FechModificado, IVA, TipoIVA, IDN, Nacional, ArtNewEst, CoreId2, UUA, FUA, Aplicaciones, NivelCCPID, TipoNivelCCPID, UsrAplicaABCID, FechaAplicaABC, UnidadTransfeID, PaqSinPrecio, PVC_CODE, PVC_CODE_DESCRIPTION, CLASIFICATION, LTA, CORE_PART_NUMBER, MARKETINGPROGRAM, WPT.

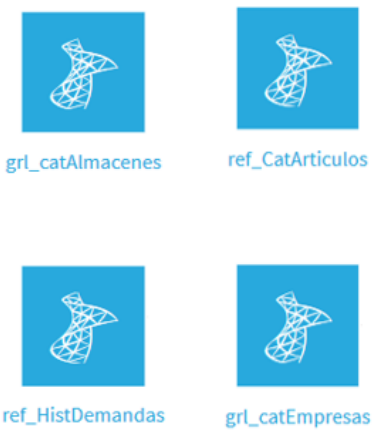


Figura 6.- Datasets predicción de obsolescencia. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

2.3.2 Limpieza y preprocesamiento

Lo primero que se debe hacer al recibir los datos es explorar los *dataset* enviados para poder ver su estructura, las variables con las que contamos, el porcentaje de dato vacíos que se tengan y de esta manera poder definir las variables que servirán para entender el conjunto de datos y crear los modelos de predicción.

También hay que tomar en cuenta la metodología que se maneja en la empresa “A” al momento de hacer las clasificaciones de obsolescencia de sus productos del inventario.

Los pasos que se siguieron para la limpieza y preprocesamiento de los dataset mencionados anteriormente, fueron los siguientes:

1. En el *dataset* Almacenes, hay una variable Activo la cual es booleana (verdadero y falso) e indica si el almacén se encuentra actualmente en funcionamiento, por lo que se procedió a eliminar las filas cuyo valor es falso

para esa variable, es decir, ya no se encuentra en funcionamiento. Luego, se mantuvieron los registros en el que el TipoAlmacen tenga el valor “R” y que en la variable Descripcion no se encontrara la palabra “EQ LIGERO”, ya que por instrucción del cliente son los almacenes que le interesaba con un manejo de producto que no sea el mencionado anteriormente. Finalmente, se removió las variables que no se necesitaban para el análisis y modelado y solo se mantuvieron las variables: AlmacenId, EmpresaId, SucursalId, FechaDemandas, Ocurrencias_D6, Ocurrencias_D12, Ocurrencias_D24, confABC_O, confABC_P y Descripcion. confABC_O y confABC_P indican la configuración de apartir de cuantos meses son clasificados en esta categoría.

2. Se realizó un *inner join* o unión interna del *dataset* preparado de Almacén con el *dataset* de Histórico de Demandas, para obtener como resultado un *dataset* que contiene todas las variables del *dataset* preparado con la agregación de las variables ArticuloId, Anio, Mes, Demandas, Entradas y NivelABCId que proviene del dataset Histórico de Demandas. Este Nuevo *dataset* de denominará dataset_1. Se usó como criterio de unión las variables AlmacenId, EmpresaId y SucursalId.

3. Con el dataset_1 se concatenó las variables Anio y Mes en una nueva variable denominada Fecha, se parseó la variable Fecha y la variable FechaDemandas. La variable Demandas y la variable Entradas contenía valores negativos por lo que se procedió a transformar todos los valores menores o iguales a cero en cero y los mayores a cero en uno para ambos casos creando las variables Demandas_Bin y Entradas_Bin respectivamente.

4. Se realizó un split con el dataset_1 para crear 2 nuevos *dataset*: Train y Test. El Train contiene los datos a partir de diciembre del 2013 a marzo del 2018 de acuerdo con la variable Fecha, esto corresponde a un 85% de los registros aproximadamente. El Test contiene los datos de abril del 2018 a julio 2018

de acuerdo con la variable Fecha, lo cual corresponde al 15% de los registros aproximadamente.

5. Se realizó una agrupación por ArticuloId en el *dataset* de Train y se obtuvo el mínimo, máximo, desviación estándar y promedio de las variables Demandas y Entradas.

6. Con el nuevo dataset creado de la agrupación anterior se realizó una unión con el *dataset* Train conservando los registros de este último y agregando los valores obtenidos de mínimo, máximo, promedio y desviación estándar de las variables Demandas y Entradas; también se realizó un join del *dataset* surgido de la agrupación con el *dataset* Test conservando los registros de este último y agregando las variables mencionadas anteriormente. A estos 2 nuevos dataset los llamaremos Train_joined y Test_joined respectivamente.

2.3.3 Entrenamiento de modelos de clasificación.

Se realizaron 2 entrenamientos, uno para predecir las entradas de los siguientes 3 meses y otro para predecir las demandas de los siguientes 3 meses.

Para entradas:

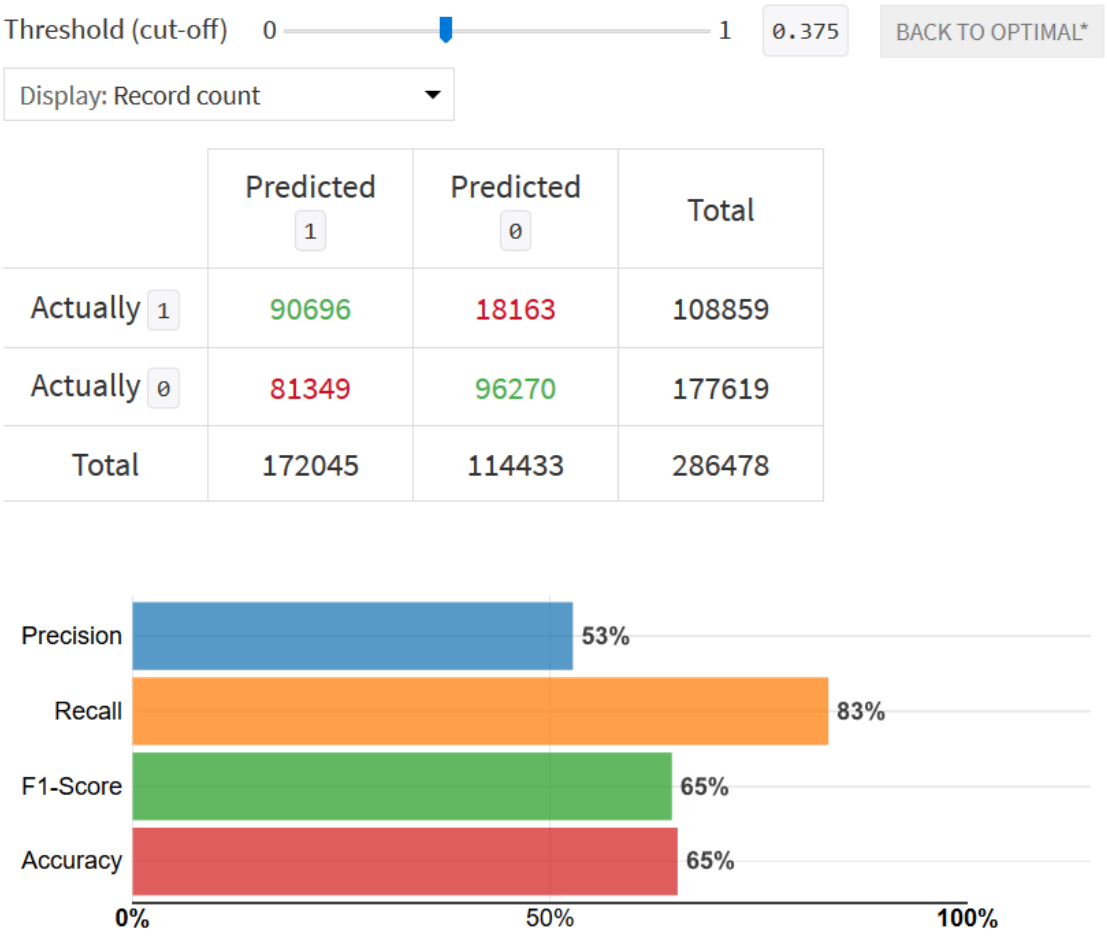


Figura 7.- Matriz de confusión y métricas principales para predicción de entradas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

En la figura 7 podemos observar varias métricas importantes, en primer lugar, el treshold (umbral) el cual establece la separación o límite que que debe cumplir la predicción para poder pertenecer a una clase. En este caso se obtuvo que el *Threshold* óptimo para la métrica para la cual fue optimizada el modelo es 0.375, es decir, que si la probabilidad de que una refacción pertenezca a la clase 1, es decir, que suceda al menos una entrada es mayor a 0.375 será clasificada como uno, si es menor será clasificada como cero, es decir no ocurrió entradas.

Luego tenemos la matriz de confusión, esta es una herramienta que permite la visualización del desempeño de los algoritmos de aprendizaje automático donde las columnas de la matriz indican las clases de predicciones y las filas indican las

clases reales de los datos. Dentro de la matriz se encuentran 4 tipos de datos, los Verdaderos Positivos (VP) son aquellos que realmente eran uno y fueron clasificados correctamente como uno, los Falsos Positivos (FP) son aquellos que realmente eran cero y fueron clasificados como uno, los Falsos Negativos (FN) son aquellos que realmente eran uno y fueron clasificados como negativos y los Verdaderos Negativos (VN) son aquellos que realmente eran cero y fueron clasificados como cero.

Después tenemos cuatro métricas para evaluar el desempeño del algoritmo de aprendizaje de máquina: *Precision* (Precisión), *Recall* (Sensibilidad), *F1-Score* (Valor-F) y *Accuracy* (Exactitud).

La precisión es la proporción de las predicciones positivas que realmente eran positivas, es decir:

$$Precisión = \frac{VP}{VP + FP}$$

La sensibilidad es la proporción de los valores reales positivos que fueron predichos como positivos, es decir:

$$Sensibilidad = \frac{VP}{VP + FN}$$

Es la métrica bajo la cual se optimizó el algoritmo de aprendizaje de máquina, esto se debe a que se prefiere disminuir el error por predecir una refacción que no tendrá ventas cuando en realidad si las tendrá, ya que es menos costoso conservar pocas piezas en el almacén a no tener piezas y generar costos de pedido por pocas piezas o en un momento dado perder el cliente.

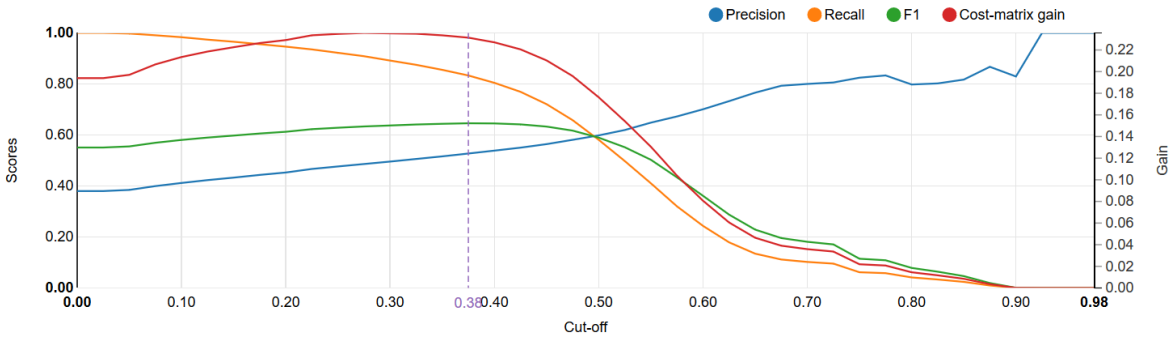
El valor-F es una media harmónica entre la Presición y la Sensibilidad, llega a su mejor valor mientras más se acerque a 1 y lo contrario ocurre al acercarse al cero, está dado por la fórmula:

$$Valor - F = \frac{2 * Precisión * Sensibilidad}{Precisión + Sensibilidad}$$

La exactitud mide la proporción de las predicciones correctas, tanto de positivos como de negativos, es decir:

$$Exactitud = \frac{VP + VN}{VP + FP + FN + VN}$$

A pesar de que el algoritmo *XGBoost* resultante tiene porcentajes no tan altos de Precisión, Exactitud y Valor-F resulta ser un mejor modelo a que dejarlo al azar o por probabilidad frecuentista. Además, hay que tomar en cuenta que que las predicciones están hechas a tres meses por lo que la sensibilidad adquiere un papel más importante pues el criterio para volverse obsoletos es que 24 meses se encuentren sin vender una refacción, por lo tanto, las refacciones que el algoritmo diga que se volverán obsoletas tienen una alta probabilidad de que sea así.



Gráfica 1.- Desempeño de las métricas de acuerdo al Threshold del modelo de predicción de entradas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La gráfica 1 muestra los resultados que tomaría la Precisión, la Sensibilidad, el Valor-F y la matriz de costos cuando se va cambiando el valor del *Threshold*. La línea puntuada ubica el corte óptimo para el *Treshold* de acuerdo con la métrica por la que se quería optimizar considerando no afectar de manera muy significativa al resto de las métricas.

Threshold-dependent (current threshold = `0.3750`)	
Accuracy Proportion of correct predictions (positive and negative) in the test set	0.6526
Precision Proportion of positive predictions that were indeed positive (in the test set)	0.5272
Recall Proportion of actual positive values found by the classifier	0.8332
F1 Score Harmonic mean between Precision and Recall	0.6457
Hamming loss Fraction of labels that are incorrectly predicted (the lower the better)	0.3474
Matthews Correlation Coefficient Correlation coefficient between actual and predicted values. +1 = perfect, 0 = no correlation, -1 = perfect anti-correlation	0.3718

Tabla 1.- Desempeño métricas dependientes del Threshold para el modelo de predicción de entradas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La tabla 1 muestra las métricas que son dependientes del *Threshold*, entre ellas se encuentran las 4 antes mencionadas y dos más, la Pérdida de Hamming el cual es la probabilidad complemento de la exactitud, es decir, la proporción de valores que fueron predichos de manera incorrecta; y el coeficiente de correlación de Matthews el cual indica la existencia de correlación entre los valores reales y los predichos, donde mientras más cercano a uno es una correlación perfecta, el cero indica la inexistencia de correlación, es decir, el algoritmo no es mejor que una predicción aleatorioa y cercano a menos uno indica una correlación imperfecta, es decir, que hay un total desacuerdo entre la predicción y el valor real. En este caso se tiene un coeficiente de correlación mayor a cero por lo que el modelo es aceptable.

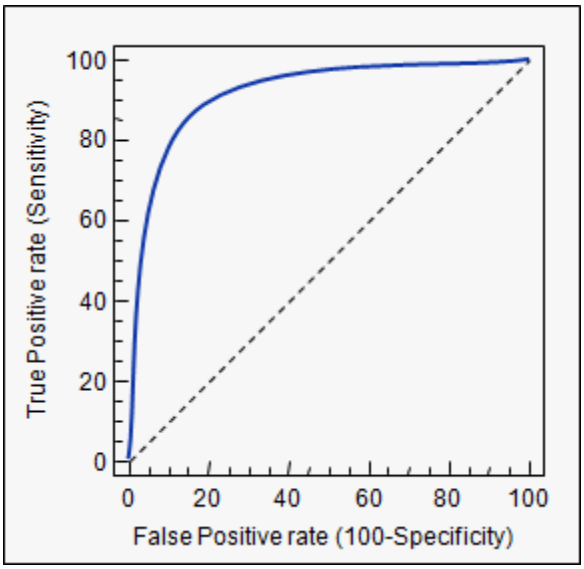
Threshold-independent	
Log loss Error metric that takes into account the predicted probabilities (the lower the better)	0.5671
ROC - AUC Score Area under the ROC; from 0.5 (random model) to 1 (perfect model)	0.7535
Calibration loss Average distance between calibration curve and diagonal. From 0 (perfectly calibrated) up to 0.5.	0.0275

Tabla 2.- Desempeño métricas independientes del Threshold para el modelo de predicción de entradas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La tabla 2 muestra las métricas que se pueden obtener que no dependen del *Threshold*, el *Log loss* (Pérdida Logarítmica) es una métrica de error que mide el rendimiento del algoritmo de aprendizaje de máquina y toma valores entre cero y uno, mientras más pequeño sea es mejor, en este caso toam un valor de 0.5671 el cual está un poco elevado y no es el ideal; el *ROC – AUC Score* (Área Debajo de la Curva ROC) es una métrica que se utiliza para los modelos de clasificación y utiliza para su construcción las métricas de sensibilidad y de especifidad, esta última tiene la siguiente fórmula:

$$Especifidad = \frac{VN}{VN + FP}$$

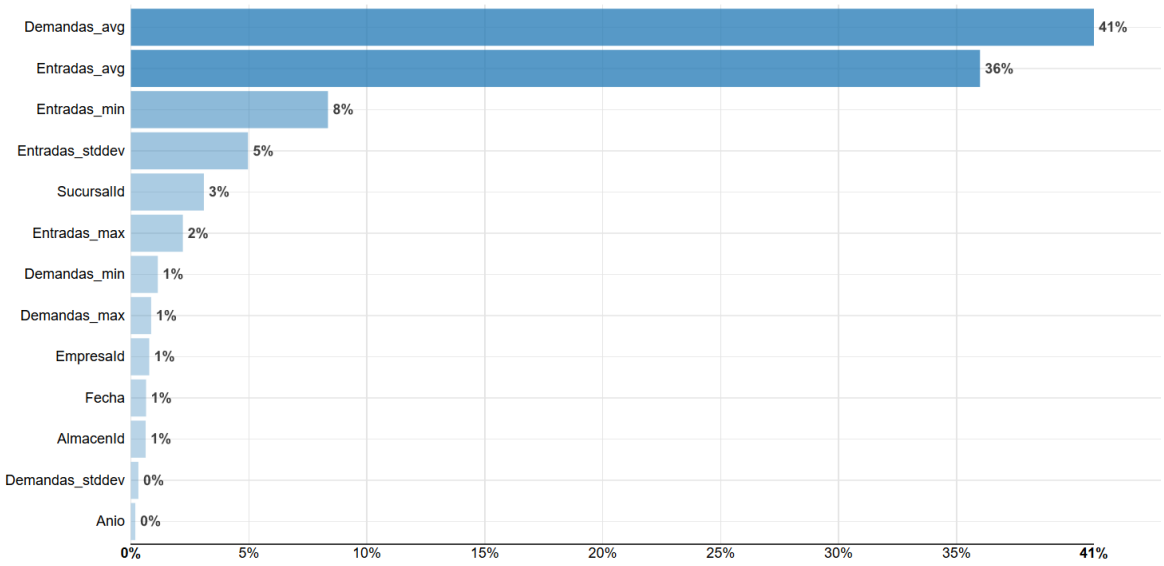
Indica la proporción de valores reales negativos que fueron clasificados o predichos como negativos. Para crear la curva ROC se grafica la Sensibilidad en el eje y y (1 – Especifidad) en el eje x, esto generará una curva y lo que se encuentre debajo de ella es la métrica que obtenemos, por ejemplo, la gráfica 2.



Gráfica 2.- Curva ROC AUC. Fuente: MEDCALC (2019)

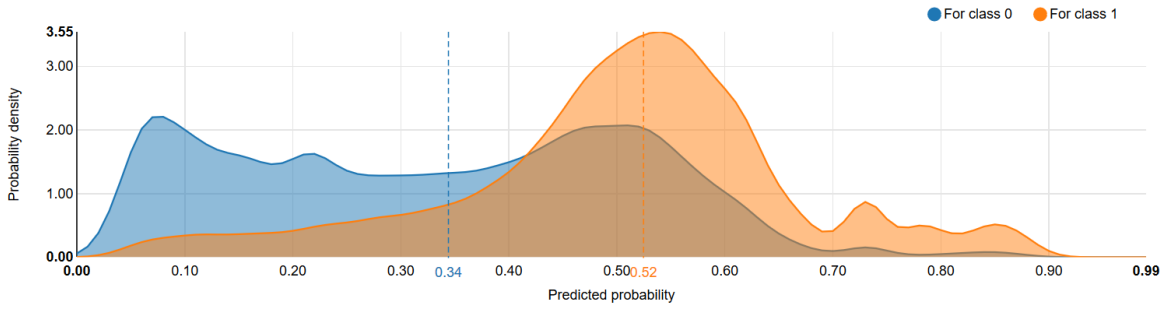
Para el modelo toma un valor de 0.7535 el cual es bueno, ya que mejor mientras más se acerque a uno.

Finalmente está la métrica de la *Calibration Loss* (Pérdida de Calibración) que es la distancia promedio entre la curva de calibración y la diagonal, toma valores de 0 a 0.5 y mientras más cercano a cero es mejor, en este caso se tiene un bueno resultado de esta métrica con 0.0275.



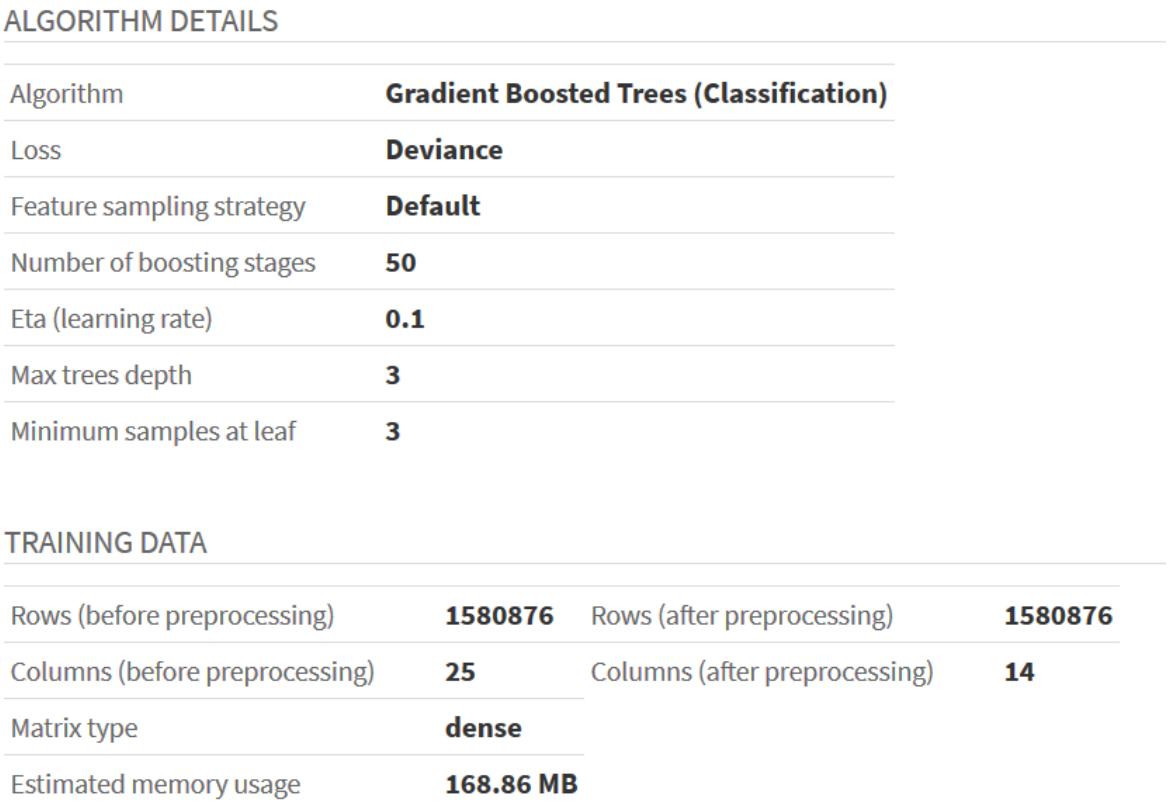
Gráfica 3.- Variables importantes modelo de predicción de entradas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La gráfica 3 muestra las variables más importantes del modelo de aprendizaje de máquina, de acuerdo con esta el promedio de demandas anuales del año anterior, el promedio de entradas anuales del año anterior y el número de entradas mínimas que hubo de refacciones del año anterior son las 3 variables más importantes que considera el modelo para poder predecir si habrá o no ventas en el mes. También podemos observar que con las primeras 6 variables se tiene el 95% de la importancia de las variables, las otras 3 variabes que no han sido mencionadas son la desviación estándar del número de entradas del año anterior, la sucursal a la que se le está haciendo la predicción y las entradas máximas que hubo en el año anterior.



Gráfica 4.- Densidad de probabilidad vs Predicción de probabilidad para la diferenciación de clases para el modelo de predicción de entradas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La gráfica 4 muestra que tan bien el algoritmo reconoce el grupo al que pertenece cada registro, mientras más dispersos se encuentran las gráficas, es decir, los colores estén menos sobrepuestos indica que el algoritmo es capaz de clasificar de una mejor manera. En este caso si se aprecia una diferencia en las clases, pero hay un sector relevante que está sobrepuesto.



ALGORITHM DETAILS

Algorithm	**Gradient Boosted Trees (Classification)**
Loss	**Deviance**
Feature sampling strategy	**Default**
Number of boosting stages	**50**
Eta (learning rate)	**0.1**
Max trees depth	**3**
Minimum samples at leaf	**3**

TRAINING DATA

Rows (before preprocessing)	**1580876**	Rows (after preprocessing)	**1580876**
Columns (before preprocessing)	**25**	Columns (after preprocessing)	**14**
Matrix type	**dense**		
Estimated memory usage	**168.86 MB**		

Figura 8.- Detalles algoritmo de predicción de entradas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La figura 8 nos muestra un resumen del algoritmo utilizado para obtener la predicción de entradas de refacciones, fue un algoritmo de clasificación *Gradient Boosted Trees* (Árboles Impulsados por Gradiente), utiliza como pérdida la desviación, tiene una estrategia de muestreo default, es decir, se utilizó todo el *dataset* Train para el entrenamiento y fue evaluado con el *dataset* Test que contiene los últimos 3 meses de entradas, utiliza 50 etapas de impulso, una tasa de aprendizaje de 0.1, una profundidad de árboles de 3 y un mínimo de 3 muestras por hoja.

Este algoritmo crea un modelo predictivo con base en varios modelos “débiles”, por ejemplo, los árboles de decisión o árboles aleatorios, y utiliza métodos de impulsamiento mediante una variable de aprendizaje ajustando los nuevos modelos para generar una estimación más precisa de la variable de respuesta y la función de pérdida utilizada (Natekin & Knoll, 2013).

En la parte de los datos de entrenamiento muestra el número de registros que se utilizaron, en este caso fueron 1,580,876 registros, de las 25 variables iniciales para el modelo se terminaron usando 14, indica que el tipo de matriz es densa, es decir, se tiene un mayor número de elementos no – cero, y finalmente dice la memoria utilizada para la generación del algoritmo.

Para demandas:

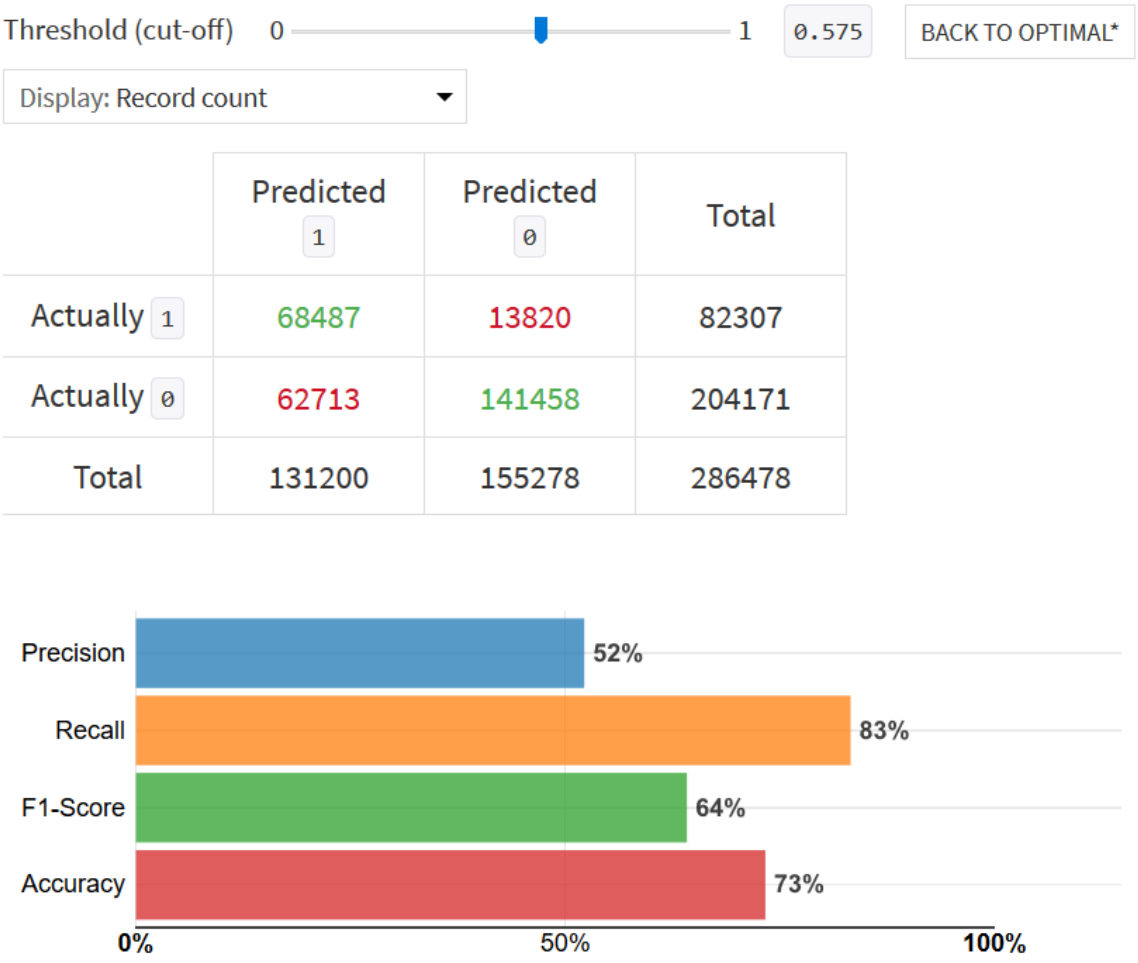
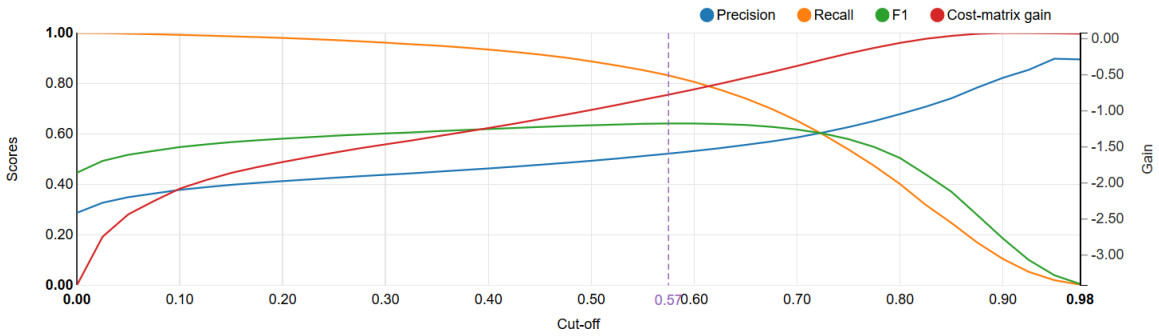


Figura 9.- Matriz de confusión y métricas principales para predicción de demandas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

En la figura 9 podemos observar que el *Threshold* para demandas es de 0.575, el cual es un resultado cercano a lo usualmente usado que es 0.5, también

fue optimizado para obtener el mejor resultado de Sensibilidad el cual fue de 83% como el algoritmo de entradas, su precisión es un 1% más bajo con un 52%, su Valor-F es un 1% más alto con un 64% y su Exactitud aumenta significativamente llegando a un 73%; por lo que podríamos decir que se obtuvo un algoritmo de clasificación más eficiente para demandas que para entradas.



Gráfica 5.- Desempeño de las métricas de acuerdo al Threshold del modelo de predicción de demandas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

En la gráfica 5 muestra que hay un punto donde puede haber mayor Precisión, sin embargo, se perdería mucha Sensibilidad, al igual que hay un sector en el que puede ser mayor la Sensibilidad, pero menor la Precisión, esto provocaría una ineficiencia en el modelo por lo que el *Threshold* utilizado es el ideal.

Threshold-dependent (current threshold = `0.5750`)	
Accuracy Proportion of correct predictions (positive and negative) in the test set	0.7328
Precision Proportion of positive predictions that were indeed positive (in the test set)	0.5220
Recall Proportion of actual positive values found by the classifier	0.8321
F1 Score Harmonic mean between Precision and Recall	0.6415
Hamming loss Fraction of labels that are incorrectly predicted (the lower the better)	0.2672
Matthews Correlation Coefficient Correlation coefficient between actual and predicted values. +1 = perfect, 0 = no correlation, -1 = perfect anti-correlation	0.4768

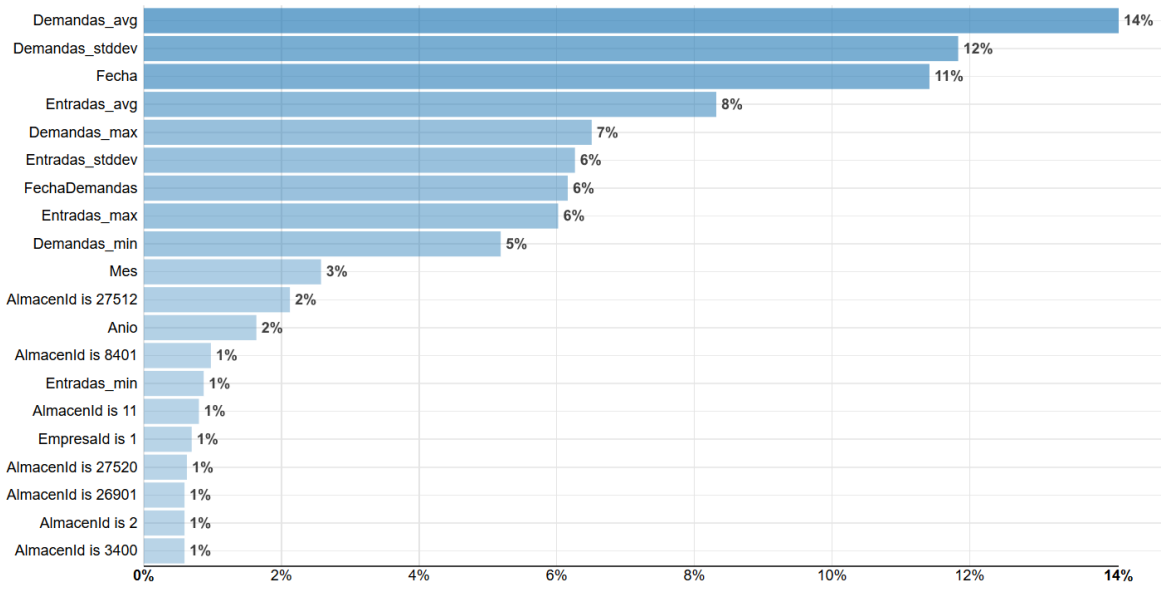
Tabla 3.- Desempeño métricas dependientes del Threshold para el modelo de predicción de demandas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

Se puede observar en la tabla 3 que las métricas dependientes del *Threshold* son buenas, especialmente la Sensibilidad, la Exactitud, el Coeficiente de Correlación de Matthews, el Valor-F y la pérdida de Hamming dan resultados aceptables y sólo la Precisión es la que se debería buscar mejorar.

Threshold-independent	
Log loss Error metric that takes into account the predicted probabilities (the lower the better)	0.5247
ROC - AUC Score Area under the ROC; from 0.5 (random model) to 1 (perfect model)	0.8406
Calibration loss Average distance between calibration curve and diagonal. From 0 (perfectly calibrated) up to 0.5.	0.1696

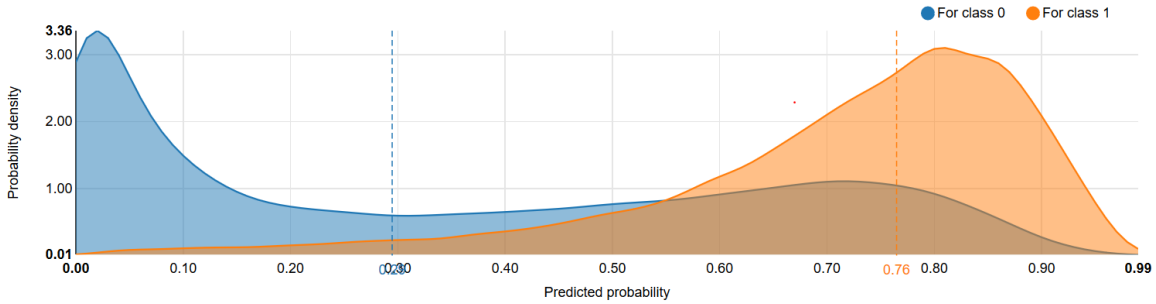
Tabla 4.- Desempeño métricas independientes del Threshold para el modelo de predicción de demandas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

En la tabla 4 se puede ver que para las métricas que son independientes al *Threshold* se tiene una curva ROC muy buena de 0.84, la pérdida de calibración salió un poco alta a comparación del algoritmo de entradas, sin embargo, sigue siendo un buen resultado y la pérdida logarítmica es muy parecida al del modelo anterior.



Gráfica 6.- Variables importantes algoritmo de predicción de demandas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

Para las variables más relevantes del algoritmo de clasificación se puede apreciar en la gráfica 6 que conserva un mayor número de variables y los porcentajes de importancia están distribuidos de una manera equitativa pues la variable más relevante es la de demandas promedio del año anterior con un 14%, siguiéndole la desviación estándar de las demandas mes con mes en el año anterior, la fecha en la que ocurrió la demanda, las entradas promedio del año anterior y el número de demandas máximas que hubo en un mes del año anterior conforman las 5 más importantes par el algoritmo dando un total del 52%. En este caso también podemos apreciar que en vez de usar el Id de la sucursal, utiliza el almacén del que se quiere predecir la demanda.



Gráfica 7.- Densidad de probabilidad vs Predicción de probabilidad para la diferenciación de clases para el modelo de predicción de demandas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

Podemos observar en la gráfica 7 que las medias de los grupos están muy separadas indicando que el algoritmo clasifica bien los grupos la mayoría de las veces, ya que aún se conserva una parte relevante que está sobrepuesta, la cual se puede deber a registros incompletos en los meses o una mala clasificación de la refacción.

ALGORITHM DETAILS

Algorithm	**XGBoost**	
Booster	**gbtree**	
Actual number of trees	**100**	
Total iterations computed	**100**	⚠ Increasing max number of trees may improve performance (but increases training time)
Max trees depth	**5**	
Eta (learning rate)	**0.5**	
Alpha (L1 regularization)	**0**	
Lambda (L2 regularization)	**1**	
Gamma (Min loss reduction to split a leaf)	**0**	
Min sum of instance weight in a child	**0**	
Subsample ratio of the training instance	**0.75**	
Fraction of columns in each tree	**1**	
Replace missing values with	**NaN**	

Figura 10.- Detalles algoritmo de predicción de demandas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

El algoritmo utilizado fue un *XGBoost (Extreme Gradient Boosting),* como se muestra en la figura 10, el cual tiene las mismas propiedades que un algoritmo de Árboles Impulsados por Gradiente (GBT), la diferencia es que busca utilizar todos los recursos computacionales para el impulsado de árboles y controla de mejor manera el sobre entrenamiento. Podemos observar que el método de impulsado es el impulsado por gradiente, utiliza 100 árboles con una profundidad de 5, proporción de aprendizaje de 0.5, valores de sus parámetros alfa y gamma de cero y su lambda de 1, el cero en la suma mínima de un hijo indica que no hay un mínimo para el pesado. Se hizo 100 iteraciones con una proporción en las submuestras de 0.75.

2.3.4 Flujo de trabajo.

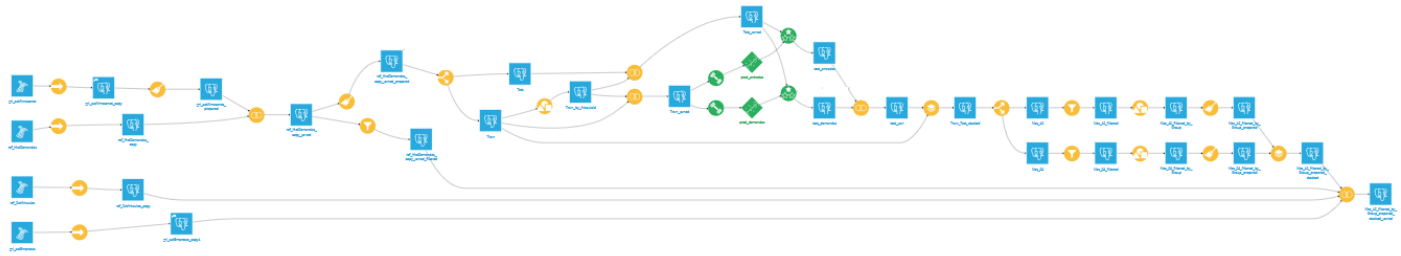
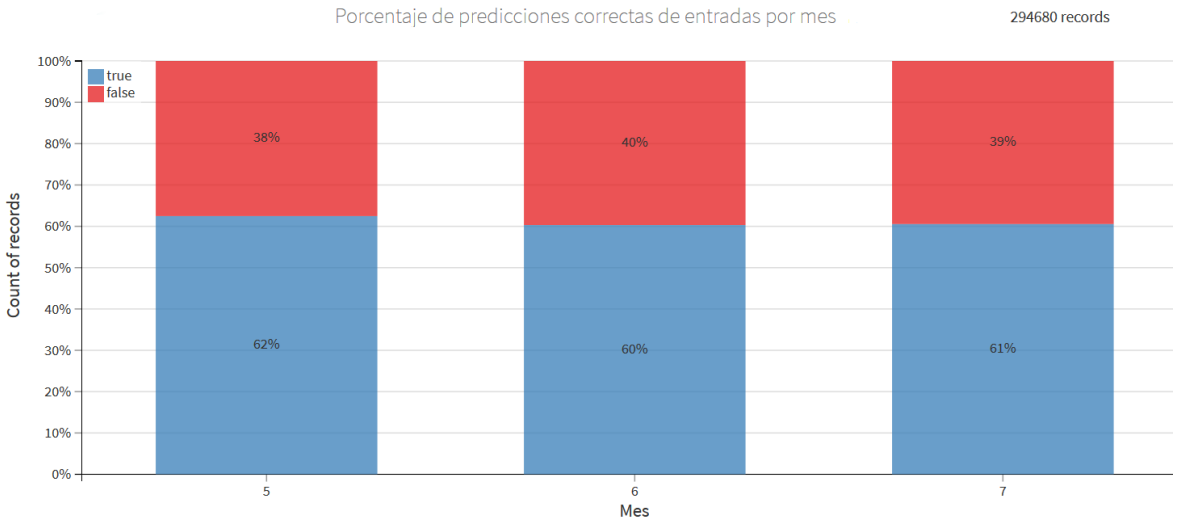


Figura 11.- Flujo de trabajo modelo de predicción de entradas y demandas para obsolescencia. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

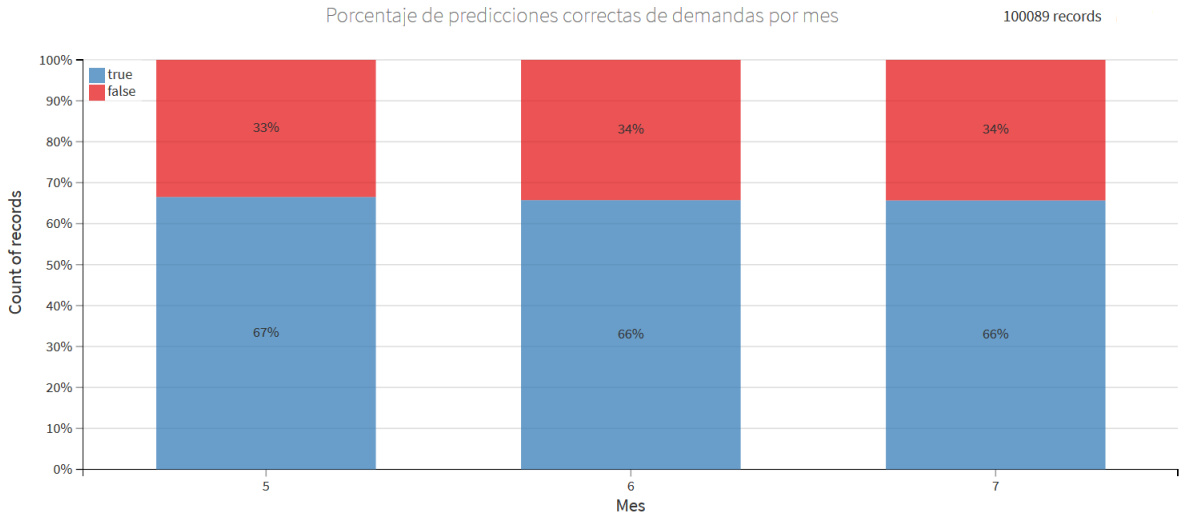
La zona verde de la figura 11 muestra las predicciones de clasificación realizadas, podemos observar que después de que se realizaron y se crearon los dataset de las predicciones se unieron los dataset para realizar análisis posteriores de resultados específicos que solicitaba la empresa “A”, los cuales por confidencialidad no serán incluídos en este reporte.

2.3.5 Visualización de resultados.



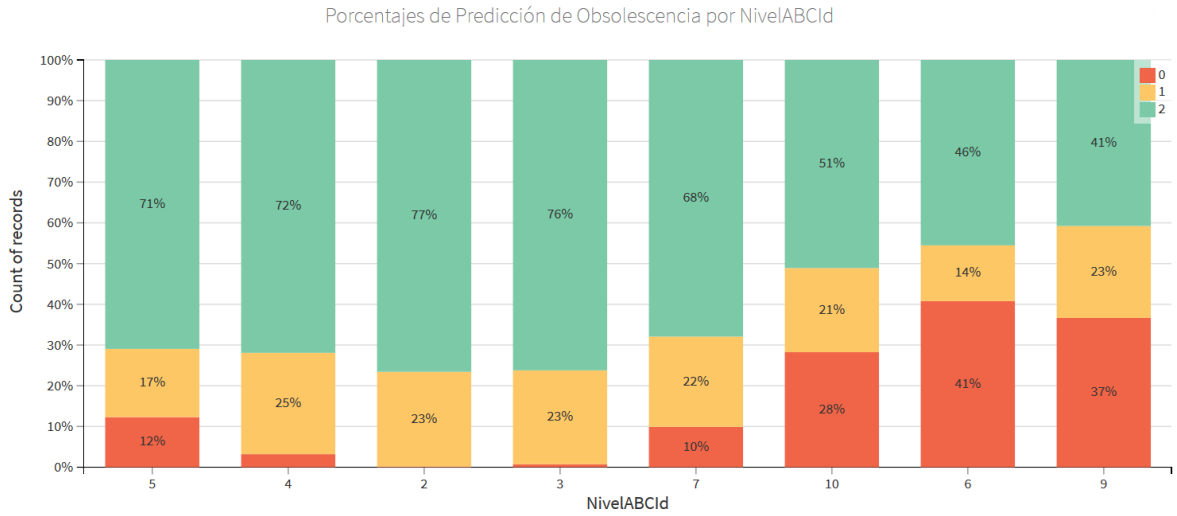
Gráfica 8.- Porcentaje de predicciones correctas de entradas por mes. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La gráfica 8 muestra los resultados obtenidos al evaluar el modelo de predicción de entradas en los siguientes 3 meses (mayo, junio y julio) del 2019. Esta gráfica evalúa la métrica de Exactitud y se puede observar que está alrededor del 61% lo cual es un poco más bajo a lo obtenido en el entrenamiento, pero se encuentra dentro de lo esperado. Otro aspecto importante a resaltar es que para los 3 meses predice de manera muy similar, no hay alguno que tenga mayor Exactitud que otro.



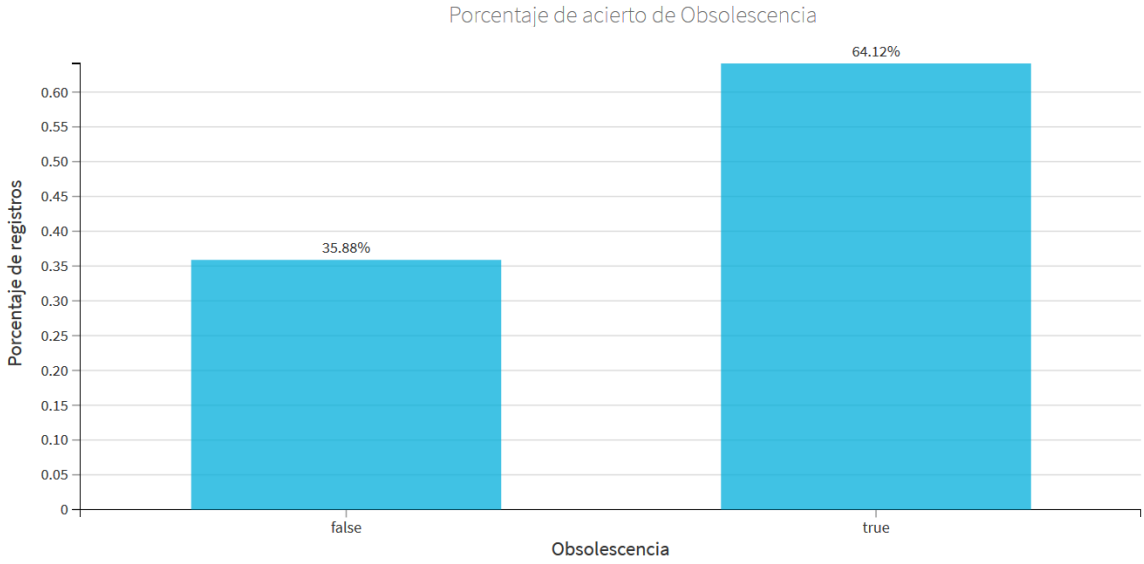
Gráfica 9.- Porcentaje de predicciones correctas de demandas por mes. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La gráfica 9 muestra los resultados obtenidos al evaluar el modelo de predicción de demandas de los siguientes 3 meses (mayo, junio y julio) del 2019. Al igual que la gráfica anterior se evalúa la métrica de Exactitud obteniendo un resultado aproximadamente de 66% que también está un poco bajo del 73% que decía el algoritmo de entrenamiento, pero también indica que predice un poco mejor que el de entradas. No existe diferencia relevante entre de la Exactitud entre los 3 meses predichos.



Gráfica 10.- Porcentaje de Predicción de Obsolescencia por NivelABCId. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La gráfica 10 muestra el porcentaje de refacciones que serán obsoletas (rojo), las que tendrán alguna demanda o entrada (amarillo) o que tendrán ambas (demanda y entrada, color verde) de acuerdo con su nivel ABCId. Como se puede observar los niveles 2, 3 y 4 tienen un porcentaje bajo de obsolescencia mientras que los noveles 6, 9 y 10 tienen porcentajes muy elevados de obsolescencia.



Gráfica 11.- Porcentaje de acierto de Obsolescencia. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La gráfica 11 muestra el porcentaje de refacciones obsoletas que fueron predichas correctamente por el algoritmo y las que fueron predichas de manera errónea. Como podemos observar el 64% aproximadamente de las refacciones que se volvieron obsoletas en mayo, junio y julio fueron correctamente predichas el cual es un porcentaje muy cercano a los resultados obtenidos de manera individual de las demandas y las entradas y que indica la eficiencia del algoritmo.

A partir de este modelo de predicción y del análisis realizado, la empresa pudo localizar los niveles ABCId que no tienen el comportamiento deseado o esperado de acuerdo con sus características, esto les permitió realizar ajustes en la asignación de niveles y operacionales. El algoritmo de predicción les permitió tener una mejor aproximación de lo que sucedería en los siguientes 3 meses para poder tomar medidas preventivas o correctivas.

Capítulo 3

Proyecto “Predicción de Compras y Ventas”

Capítulo 3. Proyecto: “Predicción de Compras y Ventas”

Este proyecto fue realizado también para la empresa “A”, pero con un objetivo distinto, predecir las cantidades que deben de comprar de cada producto del inventario y cuantos serán vendidos en los siguientes 3 meses por cada almacén.

La *POC* de la plataforma Dataiku DSS se enfocará en implementar una propuesta para satisfacer esta necesidad, es decir, se propondrá un flujo de datos para la predicción de piezas de inventario que se comprarán y venderán en cada almacén de la empresa en los siguientes 3 meses utilizando algoritmos de Aprendizaje Automático.

3.1 Objetivo del Proyecto de Compras y Ventas de la empresa “A”

Estimar las cantidades de compra y venta de productos del inventario de cada mes mediante el diseño de un flujo de datos, en específico, el entrenamiento y evaluación de modelos de aprendizaje automático de modo que permitan coadyuvar a la empresa “A” en el diseño de estrategias relacionadas con la administración oportuna de sus inventarios y la toma de decisiones de la alta dirección. Durante el proceso se definirá e implementará un flujo de trabajo que permita la limpieza, pre-procesamiento y exploración del histórico de transacciones refaccionarias (más de 10 años) de la empresa “A” para la identificación de patrones de compras y ventas. También se seleccionará, entrenará y evaluará modelos de aprendizaje automático para el correcto desempeño de los modelos de predicción, minimizando el Error Absoluto Medio (MAE) el cual será la métrica utilizada para medir la eficiencia del algoritmo.

3.2 Solución propuesta

La solución propuesta a llevar a cabo en la *POC* consiste en desarrollar un flujo de trabajo para la predicción de los productos del inventario de la empresa “A” aplicando las etapas del proceso de Ciencia de Datos.

El diagrama general de la *POC* se presenta en la figura 12, la solución incluye el desarrollo de las siguientes características:

1. Conexión con la tecnología de Base de Datos proporcionada por la empresa “A”.

2. Limpieza y pre-procesamiento de las Bases de Datos propuestas para el entrenamiento del Modelo de Clasificación.

3. Entrenamiento de Modelos de Aprendizaje Automático para predicción de cantidad de productos del inventario que se comprarán y venderán.

4. Diseño e implementación de un Escenario para la automatización de la aplicación del Flujo de Trabajo, incluyendo la predicción de nuevos meses mediante el modelo de Regresión con mejor desempeño encontrado.

5. Despliegue visual de los resultados mediante un Panel de Visualización (*Dashboard*).

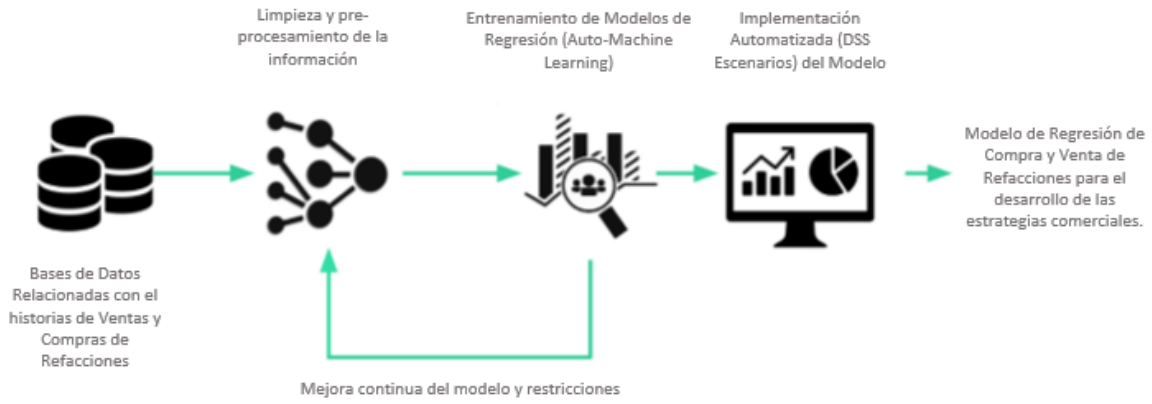


Figura 12.- Diagrama del Flujo de Trabajo para el proyecto Predicción de compras y ventas.
Fuente: Plenumsoft (2019)

3.3 Desarrollo del proyecto

En este caso para la *POC* se seleccionó que se harían con datos de 2 empresas que son subsidiarias de la empresa “A”, el objetivo es predecir las ventas y compras de refacciones los siguientes 3 meses, se realizará prediciendo la cantidad en lugar de monto.

3.3.1 Integración de Base de Datos

Para este proyecto la empresa “A” proporcionó 5 dataset los cuales podemos observar en la figura 13:

- Cmp_detFactura_LP
- Cmp_encFactura_LP
- fac_encFacturas_LP
- fac_detFacturas_LP
- Almacenes

Del dataset almacenes se tienen las mismas variables del proyecto presentado en el capítulo anterior.

Del dataset cmp_detFactura_LP tenemos las siguientes variables: FacturaId, EmpresaId, SucursalId, AlmacenId, ArticuloId, Cantidad, Precio, Prrorateo, OrdenCompraId, PaisId, CantDevuelta, CostAran, CostoEnt, CantCoreDev, CantFaltante, RequisicionId, ID, procesadoWMS.

Del dataset cmp_encFactura_LP tenemos las siguientes variables: MovimientoId, EmpresaId, SucursalId, AlmacenId, cmpFacturaId, ConceptoId, ProveedorId, SucProvedorId, Folio, FecFactura, CargXEmpaque, Status, FecCaptura, UsuarioId, Subtotal, ImpIVA, Descto, TransferenciaId, InfantCare, Emergencia, MonedaId, TipoCambio, Aduana, VendedorId, FecAplica, TipoCambioImp, garantía, SucursalDesID, AlmacenDesID, facturable.

Del dataset fac_encFacturas_LP tenemos las siguientes variables: SucursalID, FacturaID, Serie, Folio, Fecha, StatusID, TablaID, ReferenciaID, TipoFacturacion, TipoImpresion, MovimientoID, ClienteID, Cliente, TipoClienteID, RFC, Consignatario, DireccionID, Direccion, Credito, DiasCredito, MonedaID, TipoCambio, Subtotal, Descuento, Impuesto, Retencion, Observaciones, UsuarioID, EncPolID, UUA, FUA, Core, ImportarI32, SelloDigital, CancelaSelloDigital, TarjetaLealtad, FormaPago, FolioTicket, Uuid, FacUnificada, FacSumarizada, UsoCFDI_Id.

Del dataset fac_detFacturas_LP tenemos las siguientes variables: FacturaID, TipoRenglonID, ObjetoID, AlmacenID, AreaID, DepartamentoID, serAreaID, Cveart,

Descripcion, Cantidad, CantidadDev, Costo, Precio, Descuento, Impuesto, Retencion, SRT, KitID, VendedorID, ReconID, RenglonID, ArtPaqueteID, CantPaquete, ConfigPaq.

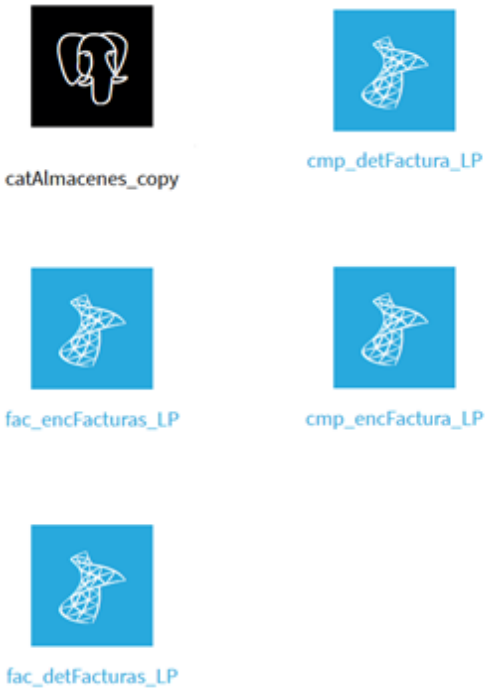


Figura 13.-Datasets predicción de compras y ventas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

3.3.2 Limpieza y preprocesamiento

Los pasos que se siguieron para la limpieza y preprocesamiento de los dataset mencionados anteriormente, fueron los siguientes:

1. Del *dataset* cmp_detFactura_LP se eliminaron variables que no contenían información, que tenía muy pocos registros, aquellos cuya respuesta fuera la misma y las que no tuvieran sentido para el objetivo de predicción: CantDevuelta, CostAran, CantCoreDev, CantFaltante, procesadoWMS, RequisicionID y PaisId. Luego se eliminaron los

registros que no pertenecen a la empresa 1 o 2. Finalmente se redondearon los valores de la variable Cantidad.

2. Del *dataset* cmp_detFactura_LP se eliminaron variables utilizando el criterio antes mencionado: InfantCare, facturable, TransferenciaId y Aduana. Se mantuvieron los registros que pertenecieran a la empresa 1 o 2. Luego se generó una nueva variable llamada Costo_Total la cual es la suma del subtotal y el IVA menos los descuentos. Finalmente, se parsearon las variables de fechas y se extrajeron el año y mes de la variable FecFactura.

3. Se realizó un join por medio de la variable cmpFactura de los 2 *dataset* anteriores con la limpieza de datos realizada.

4. Se agrupó el nuevo *dataset* con la combinación de variables ArticuloId, EmpresaId, SucursalId, AlmacenId, FecFactura_year (año de la fecha de la factura) y FecFactura_month (mes de la fecha de la factura) para obtener la suma de Cantidad del artículo comprado mes con mes, la fecha de la primera compra del mes de la refacción, la suma del subtotal del mes de cada refacción y la suma del costo total de cada mes para cada refacción.

5. Por medio de una receta de Python se calculó el mínimo, máximo, desviación estándar y promedio de cantidad de compras realizadas de la refacción de un año anterior. Finalmente se volvieron a limpiar los datos y se dividió el dataset en 2, uno para entrenamiento del modelo de predicción de compras que contiene todos los registros exceptuando los últimos 3 meses que corresponde al 90% de los registros aproximadamente y el otro para evaluación el cual contiene esos últimos 3 meses que corresponde al 10% de los registros aproximadamente.

6. Para crear el modelo de predicción de ventas se usaron los *dataset* fac_encFacturas_LP, fac_detFacturas_LP y Almacenes. Primero con el dataset fac_encFacturas_LP se eliminaron las variables con los criterios ya mencionados: MovimientoID, ReferenciaID, Folio, Credito, FacSumarizada, FacUnificada, CancelaSelloDigital, SelloDigital. ImportarI32, Core, UsuarioID, Observaciones, DireccionID, Direccion, RFC, Consignatario, TipoImpresion, TipoFacturacion, EncPolID, TarjetaLealtad, FormaPago, FolioTicket, UsoCFDI_Id, UUA, FUA, Uuid, Retencion. Despué se parseó la variable Fecha y se extrajo el año y el mes.

7. Para el *dataset* fac_detFacturas_LP se eliminaron las variables: KitID, ConfigPaq, CantPaquete, ArtPaqueteID, AreaID, DepartamentoID, serAreaID, Cveart, VendedorID, ReconID y RenglonID. Luego se creó la variable Ventas la cual surge de la multiplicación de Precio por Cantidad.

8. Se realizó una unión de los *dataset* fac_detFacturas_LP y fac_encFacturas_LP en base a la variable FacturaID y después se unió con el *dataset* Almacenes con base en la variable SucursalID.

9. Se agrupó el nuevo *dataset* con la combinación de variables ObjetoID, AlmacenId, EmpresaID, SucursalID, Fecha_year (año de la fecha de la venta) y Fecha_month (mes de la venta) para obtener la suma de Cantidad del objeto comprado mes con mes, la fecha de la primera compra del mes de la refacción y la suma de las ventas del mes de cada refacción.

10. Por medio de una receta de Python se calculó el mínimo, máximo, desviación estándar y promedio de cantidad de ventas realizadas de la refacción de un año anterior. Finalmente se volvieron a limpiar los datos y se dividió el *dataset* en 2, uno para entrenamiento del modelo

de predicción de ventas que contiene todos los registros exceptuando los últimos 3 meses que corresponde al 90% de los registros aproximadamente y el otro para evaluación el cual contiene esos últimos 3 meses que corresponde al 10% de los registros aproximadamente.

3.3.3 Entrenamiento de modelos de clasificación y visualización de resultados.

Se realizaron 2 algoritmos predictivos, uno para cantidad de ventas de los siguientes 3 meses y otro para cantidad de compras de los siguientes 3 meses.

Para Compras:

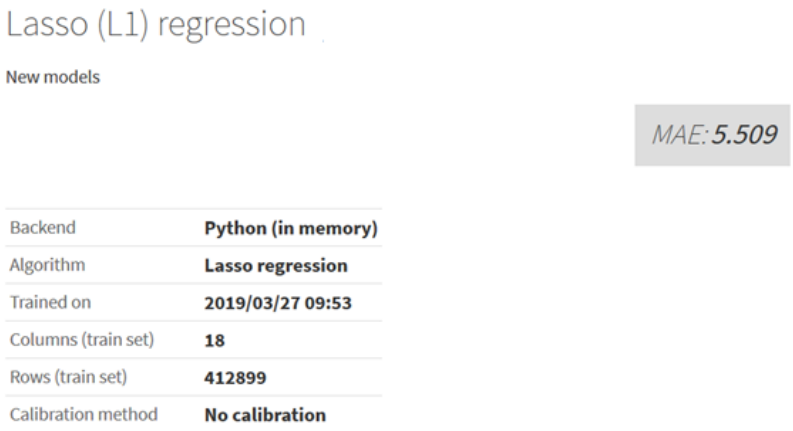


Figura 14.-Detalles algoritmo de predicción de compras. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La figura 14 muestra que el mejor algoritmo para la predicción de cantidad de compras que se harán mes con mes los siguientes 3 meses fue una Regresión Lasso (L1).

La regresión Lasso como su nombre lo indica es un algoritmo que combina un modelo de regresión con un procedimiento de contracción de algunos parámetros hacia cero y selección de variables, imponiendo una restricción o penalización sobre los coeficientes de regresión. En este caso se usa una penalización o regularización de tipo L1 la cual agrega una penalización igual al

valor absoluto de la magnitud de los coeficientes, en otras palabras, limita el tamaño de los coeficientes (Ramos Castillo, 2018).

Este algoritmo obtuvo un MAE (Error Medio Absoluto) de 5.509 el cual fue bueno y aceptado por el cliente, es la métrica bajo la cual se optimizaron los algoritmos. La fórmula del MAE es la siguiente:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |x_i - \hat{x_i}|$$

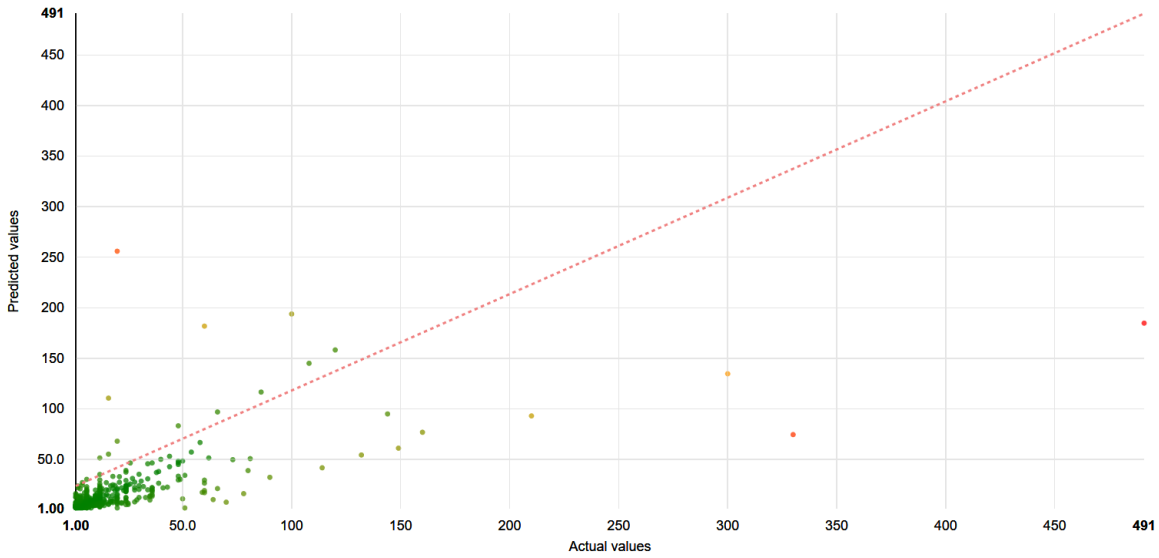
Es decir, es el promedio del valor absoluto de la diferencia del valor predicho menos el valor real. Se introdujeron 18 variables para entrenar el algoritmo y no hubo algún método de calibración utilizado.

Variable	Coefficient
Cantidad_mean	0.6930
Cantidad_min	0.1599
Intercept	-3.9346

Tabla 5.- Variables utilizadas en el algoritmo de predicción de compras. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

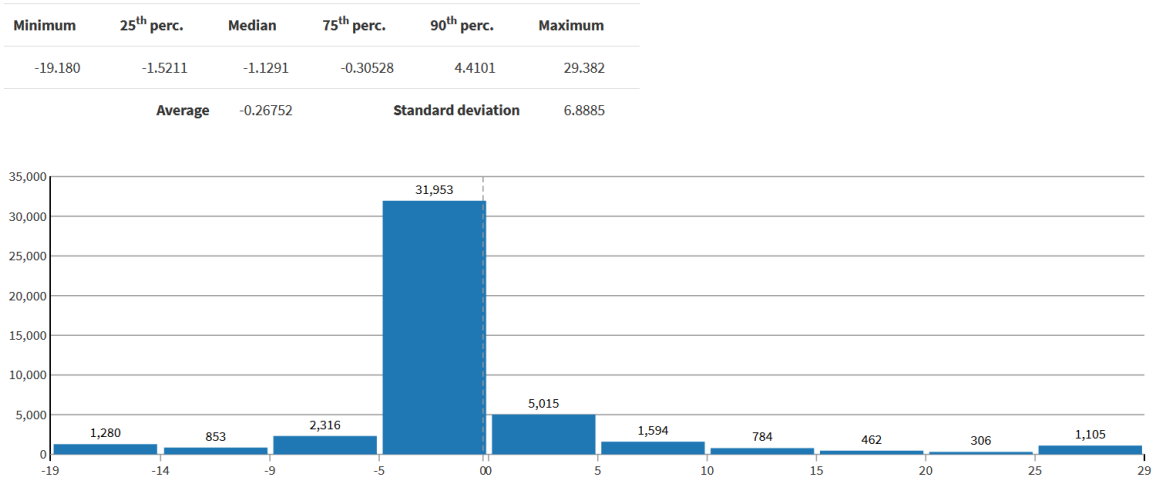
De acuerdo con la tabla 5 el algoritmo solo usó 2 variables de las 18 introducidas, las cuales fueron: cantidad promedio que se compró de la refacción el año anterior con un coeficiente de 0.693 y la cantidad mínima de lo que se compró de la refacción el año anterior con un coeficiente de 0.1599. También se incluyó un intercepto en el algoritmo con un coeficiente de -3.9346.

La razón de tener pocas variables en el algoritmo es por la penalización L1 la cual tiende a eliminar variables.



Gráfica 12.- Valores reales vs Valores predichos del algoritmo de predicción de compras. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La gráfica 12 muestra que tan cercano están los valores predichos de los valores reales. Se puede observar que clasifica mejor los resultados los valores más cercanos a cero que los más lejanos.



Gráfica 13.- Distribución de los errores para el algoritmo de predicción de compras. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La gráfica 13 da algunas estadísticas sobre los errores (diferencia del valor predicho vs el valor real) como el mínimo, máximo, mediana y percentiles, también muestra la distribución de los errores e indica que deben tener un comportamiento normal aproximado a cero para indicar que es un buen modelo de predicción, en este caso cumple con esas características.

Explained Variance Score Best possible score is 1.0, lower values are worse	0.60542
Mean Absolute Error (MAE) Average of the absolute value of the regression error	5.5089
Mean Average Percentage Error Average of the absolute value of the regression error	94.6%
Mean Squared Error (MSE) Average of the squares of the errors	6748
Root Mean Squared Error (RMSE) Root of the above mesure	82.143
Root Mean Squared Logarithmic Error (RMSLE) Root of the average of the squares of the natural log of the regression error	0.56447
Pearson coefficient Correlation coefficient between actual and predicted values. +1 = perfect correlation, 0 = no correlation, -1 = perfect anti-correlation	0.78034
R2 Score (Coefficient of determination) regression score function	0.60540

Tabla 6.- Métricas del algoritmo de predicción de compras. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La tabla 6 muestra otras métricas del algoritmo de predicción de compras como el score de la varianza explicada el cual mide que tan grande es la dispersión de los errores y tiene la siguiente fórmula:

$$explained\_variance(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}}$$

Mientras más cercano a uno quiere decir que los errores no son muy dispersos, por el contrario, mientras más cercano a cero implica que la dispersión es muy grande.

También muestra el MAE previamente mencionado y el MAPE (Error Porcentual Absoluto Medio) que indica de manera porcentual la media de los errores absolutos, su fórmula es:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i - \widehat{x_i}|}{x_i}$$

Se obtuvo un resultado del 94.6% el cual es aparentemente alto, sin embargo, los valores que se tenían eran pequeños por lo que es normal tener un porcentaje alto ya que al fallar en la predicción por pocas unidades en cuestión porcentual es muy alto la equivalencia.

El MSE (Error Cuadrático Medio) mide el promedio de los errores al cuadrado tiene la siguiente fórmula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})^2$$

Esta métrica hace los errores pequeños (menores a 1) más chicos y los errores grandes más grandes, también se le conoce como varianza. El RMSE es la raíz del MSE y es conocida como la desviación estándar y mide la dispersión de los errores. Y el RMSLE saca la raíz del logaritmo natural de los errores, esta es usada cuando no se quiere penalizar los errores muy grandes.

El Coeficiente de Pearson indica que tan bien están relacionados los valores predichos de los valores verdaderos donde mientras más cercano a uno se obtiene una correlación perfecta, a menos uno una correlación imperfecta y el cero implica que no existe correlación. Se tiene un coeficiente de 0.78034 el cual es bueno.

La métrica R2 indica que tan bien el algoritmo modela la variabilidad de los datos, se tiene un resultado de 0.60540 el cual es bajo.

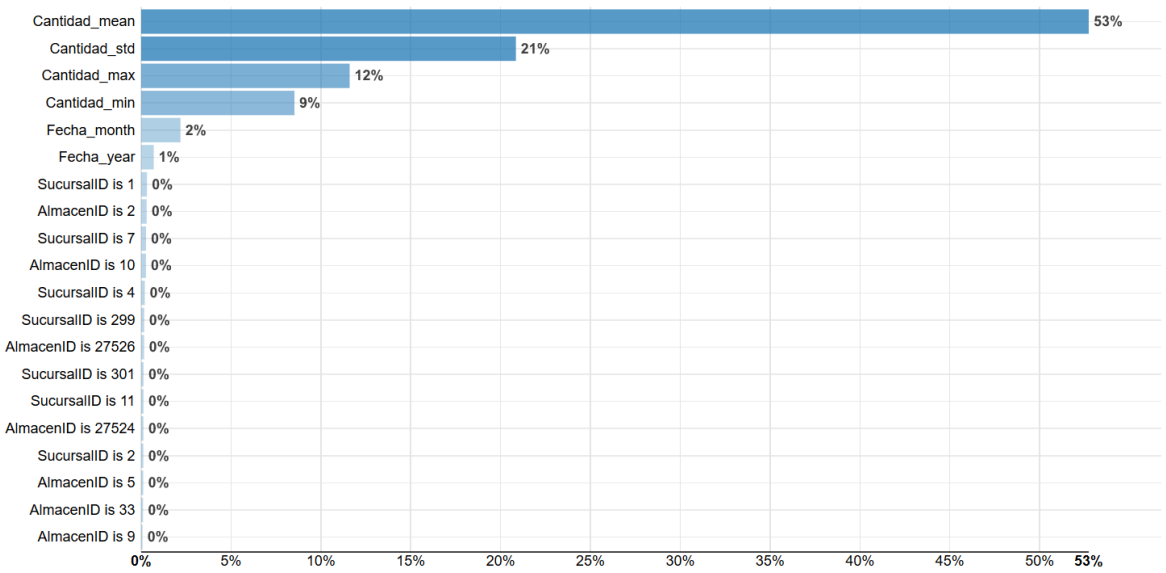
Para Ventas:

Figura 15.- Algoritmo de predicción de ventas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La figura 15 muestra que el mejor algoritmo para la predicción de cantidad de ventas que se harán mes con mes los siguientes 3 meses fue un Bosque Aleatorio.

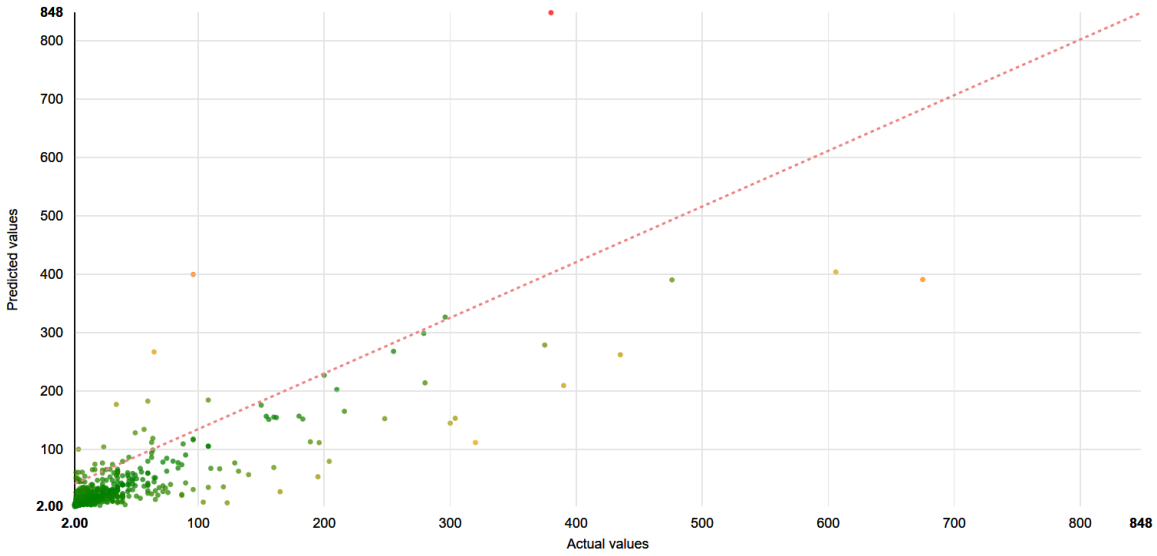
El bosque Aleatorio forma una familia de métodos que consisten en construir un conjunto de árboles de decisión que surgen de una variable aleatoria del algoritmo de inducción del árbol. Por lo general tienen bajo sesgo y alta varianza y son ideales para métodos de ensamblaje (Gilles, 2015).

El algoritmo fue optimizado con la métrica del MAE el cual tuvo un resultado de 14.005, se utilizaron 18 variables para el entrenamiento.



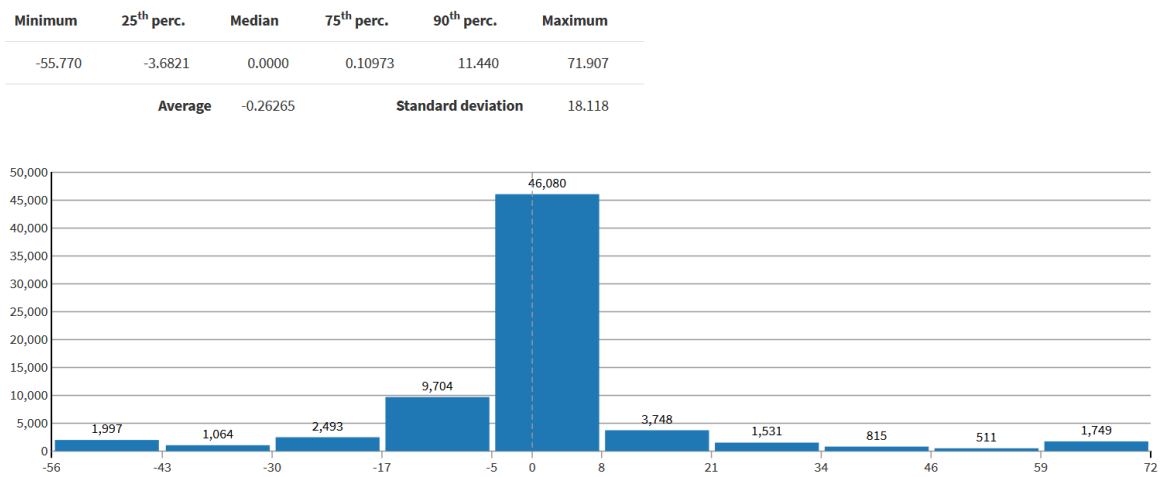
Gráfica 14.- Variables importantes algoritmo de predicción de ventas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La gráfica 14 muestra que tan importantes o relevantes fueron las variables introducidas para el modelo, se puede observar que la de mayor relevancia fue el de cantidad promedio de ventas de la refacción que se obtuvo del año anterior (también era de las variables importantes para el algoritmo de compras), luego se tiene la variable de la desviación estándar de las ventas del año anterior, cantidad máxima de ventas y mínima del año anterior. Estas 4 variables es el 95% de la relevancia para el algoritmo.



Gráfica 15.- Valores reales vs Valores predichos del algorimo de predicción de ventas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

Se puede observar en la gráfica 15 que se clasifica mejor los resultados los valores más cercanos a cero, sin embargo, aquellos valores más lejanos no son tan distintos como para considerarlos valores atípicos.



Gráfica 16.- Distribución de los errores para el algoritmo de predicción de ventas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

Se puede apreciar en la gráfica 16 que los errores tienen un comportamiento normal tomando un mínimo de -55.77 y un máximo de 71.907.

Explained Variance Score Best possible score is 1.0, lower values are worse	0.67120
Mean Absolute Error (MAE) Average of the absolute value of the regression error	14.005
Mean Average Percentage Error Average of the absolute value of the regression error	62.4%
Mean Squared Error (MSE) Average of the squares of the errors	6140
Root Mean Squared Error (RMSE) Root of the above mesure	78.358
Root Mean Squared Logarithmic Error (RMSLE) Root of the average of the squares of the natural log of the regression error	0.56822
Pearson coefficient Correlation coefficient between actual and predicted values. +1 = perfect correlation, 0 = no correlation, -1 = perfect anti-correlation	0.81960
R2 Score (Coefficient of determination) regression score function	0.67120

Tabla 7.- Métricas del algoritmo de predicción de ventas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

En la tabla 7 se puede notar que en muchas métricas se tienen mejores resultados que el algoritmo de compras, por ejemplo, el Score de Varianza Explicada, el Coeficiente de Pearson y el Score R2 son más altos lo que implica que es mejor modelo, para el MAPE, MSE, RMSE y RMSLE se tienen resultado menores lo cual quiere decir que tiene un mejor rendimiento. Para el caso del MAPE el algoritmo de ventas fue más alto, pero esto se debe a que las cantidades que se venden son mayores a las que se compran.

ALGORITHM DETAILS

Algorithm	**Random forest regression**	Split quality criterion	**MSE**
Number of trees	**100**	Use bootstrap	**Yes**
Max trees depth	**20**	Feature sampling strategy	**auto**
Min samples per leaf	**1**		
Min samples to split	**3**		

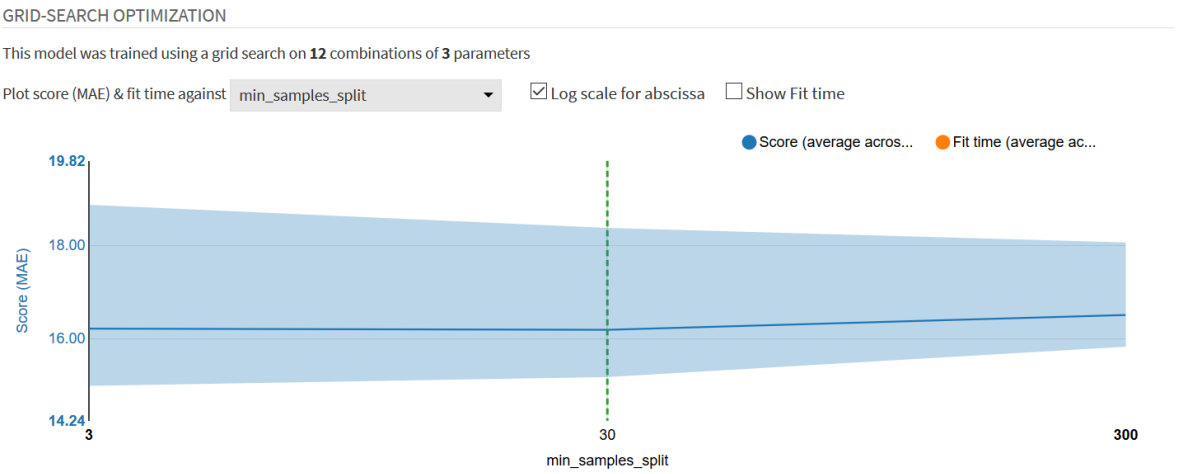
TRAINING DATA

Rows (before preprocessing)	**630597**	Rows (after preprocessing)	**630597**
Columns (before preprocessing)	**18**	Columns (after preprocessing)	**101**
Matrix type	**dense**		
Estimated memory usage	**485.92 MB**		

Figura 16.- Detalles algoritmo de predicción de ventas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

Se puede observar en la figura 16 que el algoritmo de regresión de Bosque Aleatorio utilizó 100 árboles, con una profundidad de 20, un mínimo de muestras por hoja de 1 y 3 muestra para indicar que se puede dividir es hoja, para el criterio de división se utilizó el error cuadrado promedio y se usó el método de bootstrap para las estimaciones.

También se puede apreciar que de las 18 variables iniciales se convirtieron en 101 después del procesamiento del algoritmo, esto se debe a la transformación que se les hacen a las variables. Finalmente muestra que es una matriz densa lo cual implica que la mayoría de los registros son no cero.



Gráfica 17.- Optimización búsqueda por cuadrícula para el algoritmo de predicción de ventas. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

La gráfica 17 muestra el rendimiento del algoritmo respecto al MAE cuando se cambia el número de muestras mínimas que necesita la hoja antes de dividirse, se puede observar que el mejor resultado se obtiene cuando son 30.

3.3.4 Flujo de trabajo.

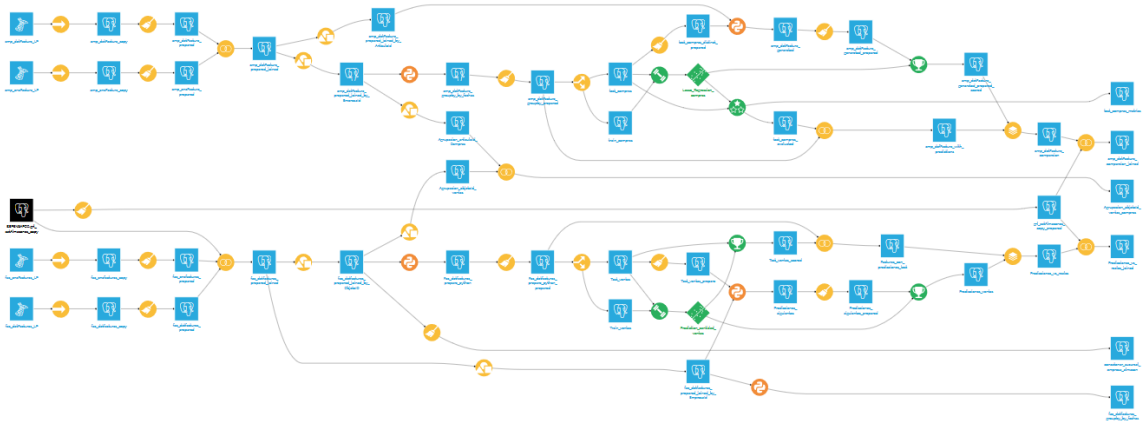


Figura 17.- Flujo de trabajo modelo de predicción de ventas y compras. Fuente: Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS (2019).

Al igual que en el capítulo anterior se muestra el flujo de trabajo completo realizado, sin embargo, lo explicado en este documento llega hasta los íconos verdes que son

donde se realizaron los algoritmos como se puede apreciar en la figura 17. A diferencia del otro flujo se tiene un ícono negro el cual indica que es un dataset que se utilizó de otro proyecto (proyecto del capítulo 2).

A partir de este algoritmo de predicción y el análisis realizado la empresa obtuvo un mejor conocimiento de lo que esperaría vender y comprar los siguientes meses, lo cual les permitió tomar medidas de mercadotecnia y de atención a clientes para poder colocar mejor sus productos y medidas para anticiparse a la venta de refacciones, por ejemplo, realizar pedidos de refacciones en grandes cantidades a sus proveedores de los productos que se esperan haya mucha demanda para obtener mejor precio.

Capítulo 4

Beneficios y Recomendaciones del uso de la Ciencia de Datos

Capítulo 4. Beneficios y Recomendaciones del uso de la Ciencia de Datos

4.1 Beneficios

Un uso correcto y efectivo de la Ciencia de Datos puede traer múltiples beneficios para las empresas como detección de fraudes y mitigación de riesgos, predicción de pérdida de clientes, segmentación de usuarios, personalización de la experiencia de los clientes, mejora y optimización de procesos, además también puede haber beneficios los cuales dependerán del sector al cual pertenezcan, por ejemplo, en el sector salud se puede predecir enfermedades o encontrar el mejor tratamiento posible, en educación se puede predecir el abandono escolar, medir el aprovechamiento escolar y mejorar las técnicas de educación, para los gobiernos les puede servir para predicción de fenómenos meteorológicos, incendios o contaminación, petrolización preventiva de las calles, mantenimiento preventivo de edificios, etc.

En este reporte se hizo mucho énfasis en la descripción de proyectos en los que se realizaron *POC´s,* se debe a que es la principal actividad que realizo en la empresa ya que es el primer acercamiento que busca realizar la empresa con el cliente. Esto se debe a que las *POC´s* deben demostrar el mayor valor de un sistema, asegurando que esté alineado con el reenvío de los objetivos estratégicos a largo plazo de la empresa, también demostrará que toda la infraestructura adecuada está en su lugar y, en última instancia, eliminará cualquier riesgo asociado con avanzar a toda velocidad con proyectos de ciencia de datos a escala. Ya que deben demostrar no solo que una solución resuelve un problema específico en particular, sino que un sistema proporcionará un valor generalizado a la empresa: que es capaz de llevar una perspectiva basada en datos a gama de objetivos estratégicos del negocio. (Dataiku, 7 Steps to Driving an Efficient Data Science POC: Move Past Simple Evaluation & Bring Value, 2017)

4.2 Recomendaciones

Para poder aprovechar al máximo la Ciencia de Datos en los objetivos que se plantean cumplir por parte de las empresas se necesita contar con un equipo de científicos de datos capaces y con los conocimientos pertienentes.

De acuerdo al documento "*Nurturing a Productive Data Team*" del año 2017 los mejores equipos de datos encuentran la manera de lograr un equilibrio entre soluciones nuevas, innovadoras y creativas y tener cuidado para garantizar que todas las soluciones basadas en datos estén bien pensadas. Tener una cultura de datos también genera beneficios para las empresas. Crear una cultura de datos significa compartir y fomentar ampliamente las siguientes ideas:

- Estar impulsado por los datos y tomar decisiones basadas en los datos es una responsabilidad compartida entre todos en la empresa (no es solo responsabilidad del equipo de datos).
- Todos deberían estar al tanto de los proyectos de datos en curso y futuros y de cómo podría afectar su trabajo o su trabajo.
- La automatización y el análisis predictivo son una evolución positiva e importante para la empresa, no una forma de eliminar trabajos o responsabilidades de otros empleados. (Dataiku, Nurturing a Productive Data Team: How to Grow Data Teams from Infancy to Maturity, 2017)

Para poder llevar a cabo una buena *POC* se necesita:

- Escoger un caso real y concreto: de esta manera se pueden evaluar los resultados y el cliente relacionar de mejor manera el proyecto con las actividades que se realizan en su empresa.
- Restringir la *POC* a un tiempo razonable: en general 60 días máximo es suficiente.
- Definir claramente los objetivos: de esta manera se sabe cuando fue un éxito o no.
- Involucrar a las personas adecuadas: esto incluye a los científicos de datos que se encargará en llevar a cabo la *POC,* pero también el

equipo de TI que lo pondrá en producción, así como a los usuarios que utilizarán la solución.

- Considerar si el proyecto se pone a producción: al poner en producción el proyecto este se integra en las operaciones de la compañía lo cual permite evaluar todo el potencial que tiene y si permititrá obtener los objetivos del Retorno de Inversión (ROI) establecidos por la empresa.
- Asegurar la autonomía: trabajar con los expertos con los que cuenta el cliente para poder aprovechar de mejor manera sus conocimientos y aprender de ello, pero sin perder la autonomía del equipo de ciencia de datos.
- Ser ágil pero enfocado: las mejores soluciones provienen de los equipos que pueden direccionar distintas soluciones en el menor tiempo posible, pero es importante nunca perder de vista los objetivos de la *POC* (Dataiku, 7 Steps to Driving an Efficient Data Science POC: Move Past Simple Evaluation & Bring Value, 2017).

Conclusiones

Conclusiones

La ciencia de los datos es un campo interdisciplinario que puede ser utilizado en todas las empresas, desde encontrar mejoras en sus procesos operacionales, identificar problemas, recolección de datos y crear estrategias mejor enfocadas con el análisis de la información hasta procesos más complicados o específicos como detección de fraudes, pérdida de clientes, detección de enfermedades, comportamientos de aprendizaje, etc.

A partir del éxito de los algoritmos creados para la detección de obsolescencia, predicción de compras y predicción ventas realizados, la empresa “A” solicitó propuestas de otros proyectos que se pudieran realizar con la herramienta Dataiku DSS. Se propusieron los siguientes proyectos:

Optimización de almacenes e inventarios

- Estimación de los tiempos para llenar el inventario: Con base al historial de inventarios y pedidos se puede predecir el tiempo o cantidad de producto mínimo que el almacén necesita para generar una alerta que indique que es necesario llenar el inventario. Esto permite contar con los productos cuando el cliente lo necesite, optimizar costos de pedidos al poder tener una planificación de los productos que están próximos a agotarse.
- De acuerdo con la capacidad del almacén optimizar las cantidades de producto de acuerdo con la venta esperada, mínimo y máximo posible: Con base a los m2 de capacidad del almacén, historial de ventas, y reglas de mínimos y máximos establecidos por la empresa, se determina las cantidades adecuadas que deben de tenerse de cada producto en cada almacén para obtener la ganancia máxima posible.

Mejorar cadena de suministro

- Mantenimiento predictivo: Con base al histórico de mantenimientos y características de los vehículos se hace una predicción del momento en el que los vehículos deben entrar a mantenimiento para evitar una falla futura, esto permite una mejor satisfacción del cliente.

- Elección de los mejores proveedores: Con base al historial de pedidos a los proveedores, los requerimientos y beneficios de cada uno predecir el mejor proveedor para el siguiente pedido de acuerdo con sus características. Esto permite reducir riesgos de atraso por parte de proveedores y reducir costos.
- Optimización de procesos de requisiciones de compras: Con base al proceso que se lleva a cabo para hacer las requisiciones de compras se puede utilizar IA para hacer más ágil el proceso.
- Medir la satisfacción del cliente: Mediante encuestas y redes sociales, esto permite conocer las áreas para mejorar, mantener clientes y fidelidad.
- Optimización de rutas para la logística de distribución para y entre los almacenes (toma en cuenta georreferencia): Mediante las rutas de distribuciones crear una web app que permita encontrar la ruta óptima para todos los puntos a los que necesita llegar, de acuerdo con las condiciones que se establezcan.
- Predicción de aumento de siniestros (ocupación del taller): Con base al histórico de ocupación del taller, predecir los momentos en el que habrá un aumento de clientes. Esto permite planificar y establecer estrategias para ser capaces de atender a todos los clientes y no perderlos.
- Inbound marketing (ofertas y contenidos personalizados): Permite la eficiencia en la obtención de clientes y aumento de leads.
- Seguimiento y control de rutas: Mediante dispositivos de rastreo se puede monitorizar en todo momento los camiones que llevan la mercancía.
- Optimizar la atención al cliente (chatbots, mesa de servicio, identificación de principales problemas): Permite la prevención de problemas, mayor satisfacción del cliente y disminución de costos de atención.
- Churn prediction: Mediante el histórico de compras de los clientes se puede predecir cuales están próximos a irse. Permite establecer estrategias de retención de clientes, prevención de pérdidas e identificación de causas de churn.

- Segmentación de clientes: Con base a los datos que se tengan de los clientes o prospectos se pueden clasificar en grupos para envío de marketing, potenciales clientes, clientes riesgosos, etc.

En conclusión, de un solo giro comercial se pueden realizar múltiples proyectos con ciencia de datos o *POC´s* como las que se describieron en ese reporte de experiencia laboral, lo único que se necesita es contar con un histórico de datos o tener un banco de datos del que se pueda obtener la información requerida para el proyecto de acuerdo a los objetivos planteados.

Entonces, si se pueden lograr grandes cosas al usar Ciencia de Datos, ¿por qué muy pocas empresas lo utilizan?

Esto se debe por múltipes razones, ya que es una disciplina relativamente nueva y para formar a un científico de datos se requiere de muchos conocimientos, sin embargo, en la empresa (Plenumsoft), se ha identificado que principalmente se debe a una falta de conocimiento respecto al tema, porque se cree que para poder aplicarla se necesita ser una empresa “grande” y contar con muchos datos o pagar los servicios de un experto en Ciencia de Datos, pues no se conoce el beneficio monetario que les puede traer en invertir en uno, otra de las causas es que se desconoce lo que uno puede lograr al implementar Ciencia de Datos; ya que las empresas grandes saben que les traerá beneficios pero no comprenden cómo.

Por lo tanto, el primer paso es dar a conocer a las empresas, lo que puede esperar de un proyecto de Ciencia de Datos, de esta manera los planteamientos de estos proyectos serán mejor enfocados en resolver las necesidades primordiales de la empresa y darle un valor agregado a los productos o servicios que ofrezcan, esto es un aspecto muy relevante ya que una de las habilidades primordiales al realizar Ciencia de Datos que no tienen que ver con estudios es el expertise que se tiene sobre el tema o giro de la empresa y no hay nadie mejor que comprenda de su negocio que el mismo dueño o los trabajadores de esa empresa. Y esta es una forma en la que el proceso para el desarrollo de proyectos de Ciencia de Datos se vuelve más eficiente y adquiere mejores resultados, que es cuando el cliente se involucra en el proyecto y lo asimila como suyo, al estar pendiente de los avances, resolver los problemas con los que se topa el equipo de Científicos de Datos y dar

a conocer a los demás miembros de la empresa lo que se pretende lograr para que ellos también se sumen al proyecto y se culturize a la empresa.

Actualmente Plenumsoft realiza 2 acciones para resolver estos problemas: imparte cursos de Inteligencia Artificial (IA) y Ciencia de Datos, y tiene un programa llamado Células Plenum el cual forma estudiantes que se encuentran en sus últimos años de licenciatura de carreras a fines a computación y matemáticas en temas de Ciencia de Datos mediante la participación de proyectos con clientes y de la capacitación de expertos en Ciencia de Datos.

Bibliografía

Alonso Martínez, M. (1992). *Conocimiento y bases de datos: una propuesta de integración inteligente.* Obtenido de https://www.tesisenred.net/bitstream/handle/10803/31767/3de3.MAMcap5_conclusiones_bibliograf%C3%ADa.pdf?sequence=4&isAllowed=y

Camargo, J. J., Camargo, J. F., & Joyanes, L. (2015). Knowing the Big Data. *Facultad de Ingeniería*, 63-77.

Dataiku. (2017). 7 Steps to Driving an Efficient Data Science POC: Move Past Simple Evaluation & Bring Value.

Dataiku. (2017). Nurturing a Productive Data Team: How to Grow Data Teams from Infancy to Maturity.

Dataiku. (s.f.). *Dataiku*. Obtenido de https://www.dataiku.com/

Deoras, S. (2019). 10 Challenges That Data Science Industry Still Faces. *Analytics India Magazine*.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 64-73.

Elaboración propia a partir de los datos de la empresa "A" y el uso de software Dataiku DSS. (2019).

Elaboración propia con el software Dataiku DSS. (2019).

Gilles, L. (3 de Junio de 2015). *University of Liege.* Obtenido de https://arxiv.org/pdf/1407.7502.pdf

Jake, V. (2017). *Python Data Science Handbook : Essential tools for working with data.* California, EUA: O´Reilly.

MEDCALC. (2019). Obtenido de https://www.medcalc.org/manual/roc-curves.php

Moreno Salinas, J. G. (2017). Científico de datos: codificando el valor oculto e intangible de los datos. *RDU: Revista Digital Universitaria. UNAM*.

Natekin, A., & Knoll, A. (Diciembre de 2013). *ResearchGate*. Obtenido de ResearchGate: https://www.researchgate.net/publication/259653472_Gradient_Boosting_Machines_A_Tutorial

Plenumsoft. (2018). *Algoritmos de Aprendizaje Automático*.

Plenumsoft. (2018, pág. 2). Metodología Plenumsoft.

Plenumsoft. (2019). Propuesta de POC Predicción de Obsolescencia.

Plenumsoft. (2019). Propuesta POC Predicción de compras y ventas.

PMG Business Improvement. (2019). *Desafíos en la evolución de las empresas hacia la Transformación Digital.* Obtenido de https://www.ecommerceccs.cl/wp-content/uploads/2019/06/El-desafio-hacia-la-Transformaci%C3%B3n-Digital.pdf

Ramos Castillo, L. (Junio de 2018). *Universidad de Sevilla.* Obtenido de https://idus.us.es/xmlui/bitstream/handle/11441/77576/Ramos%20Castillo%20Laura%20TFG.pdf?sequence=1

Reinsel, D., Gantz, J., & Rydning, J. (Noviembre de 2018). *SEAGATE.* Obtenido de SEAGATE: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Singaram, M., & Jain, P. (13 de Enero de 2018). *What is the Difference between Proof of Concept and Prototype ?* Obtenido de Enrepenuer India: https://www.entrepreneur.com/article/307454

SPSS, I. (2012, pág. 1). Obtenido de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

Anexos

Anexo I

Características centrales que integran la plataforma Dataiku DSS

Dataiku DSS es una plataforma colaborativa que potencia tanto el análisis de autoservicio como el de operacionalización de modelos de aprendizaje automático puestos a producción. (Dataiku, Dataiku, s.f.)

Características de Dataiku:

- Herramienta *end-to-end* para todo el equipo: científicos datos, analistas, y ejecutivos.
- Interfaz compartida basada en colaboración.
- Posibilidad de replicar scripts y reutilizar avances previos.
- Integración nativa con tecnología *Big-Data*.
- Ciclos rápidos de evolución, con inclusión nativa de nuevos estándares y conectores.
- Habilidad nativa de manejar datos estructurados y no-estructurados, con o sin *datalake*.
- Basado e integrado con algoritmos *open-source*: iPython, scikit-learn, R, github, etc.
- Uso nativo de Version-Control y *Audit-trail*.
- Innovación transparente, anti-"*Black-Box*": algoritmos y predicción le pertenece al usuario.

Dataiku DSS permite tener usuarios "*clickers*" y "*coders*". Los primeros son aquellos que pueden desarrollar un proyecto de Ciencia de Datos en base a las recetas visuales que tiene Dataiku DSS que permiten desarrollar procesos que se usan en un proyecto mediante clicks sin tener la necesidad de saber programar, las cuales como se muestra en la figura anexo 1 son:

- **Receta de Preparación**: permite crear scripts de limpieza, normalización y enriquecimiento de datos de forma visual e interactiva.
- **Receta de Sincronización**: permite sincronizar un conjunto de datos a otro. La sincronización puede ser global o por partición. Uno de los principales casos de uso de la receta de sincronización es copiar un conjunto de datos

entre *backends* de almacenamiento donde es posible realizar diferentes cálculos.

- **Receta de Agrupación**: permite realizar agregaciones en cualquier conjunto de datos en Datatiku DSS, ya sea un conjunto de datos SQL o no. Esto es el equivalente de una sentencia "agrupar por" de SQL. La receta ofrece herramientas visuales para configurar las agregaciones (personalizadas) y los alias.
- **Receta de Ventana**: permite realizar funciones de análisis en cualquier conjunto de datos en Dataiku DSS, ya sea un conjunto de datos SQL o no. Esto es el equivalente de una declaración "sobre" de SQL. La receta ofrece herramientas visuales para configurar las ventanas y los alias.
- **Receta de Distinción**: permite deduplicar filas en un conjunto de datos recuperando filas únicas. Las filas se comparan utilizando las columnas que especifique. También puede optar por obtener el número de duplicados para cada combinación. Se puede realizar en cualquier conjunto de datos. La receta ofrece herramientas visuales para configurar las ventanas y los alias.
- **Receta de Unión**: permite las uniones entre dos o más conjuntos de datos. Dataiku DSS maneja las combinaciones internas, las combinaciones externas izquierdas y las combinaciones externas completas.
- **Receta de Separación:** permite enviar filas de un conjunto de datos a otro conjunto de datos, según las reglas establecidas.
- **Receta de los Top N:** permite recuperar la primera N y la última M filas de subconjuntos con los mismos valores de claves de agrupación. Las filas dentro de un subconjunto están ordenadas por las columnas que especifique. Se puede realizar en cualquier conjunto de datos, ya sea un conjunto de datos SQL o no. La receta ofrece herramientas visuales para configurar las especificaciones y los alias.
- **Receta de Apilamiento:** combina varios conjuntos de datos en un conjunto de datos. Esta receta es el equivalente de una declaración SQL "unir todos".
- **Receta de Muestreo: t**iene el doble propósito de muestrear y / o filtrar conjuntos de datos

- **Receta de Ordenamiento:** permite ordenar un conjunto de datos. Se especifica una lista de columnas, cada una con orden ascendente o descendente. Se puede realizar en cualquier conjunto de datos, ya sea un conjunto de datos SQL o no. Sin embargo, para que la receta sea útil, el conjunto de datos de salida debe conservar el orden de escritura. Los más comunes son los sistemas de archivos y HDFS. De este modo, al crear una nueva receta de clasificación, el conjunto de datos de salida se configurará para conservar el orden por escrito, si es posible. La receta también ofrece herramientas visuales para configurar las especificaciones y los alias.
- **Receta de Pivote:** permite crear tablas dinámicas, con más control sobre las filas, columnas y agregaciones de lo que ofrece el procesador de pivote. También ejecutar el pivote de forma nativa en sistemas externos, como bases de datos SQL o Hive.
- **Receta de Descarga:** permite descargar archivos de conexiones basadas en archivos y almacenarlos en una carpeta administrada por Dataiku DSS. La carpeta administrada en sí puede almacenarse en cualquier conexión basada en archivos que acepte carpetas administradas. La receta de descarga solo trata con archivos: no interpreta los archivos y no crea un conjunto de datos que se pueda utilizar directamente. Sólo crea una carpeta administrada. Una vez que se haya creado la receta de descarga y su carpeta administrada de salida asociada, se puede crear un conjunto de datos "Archivos en carpeta" basado en la carpeta administrada de salida. Este conjunto de datos de "Archivos en carpeta" se ocupa del análisis de los archivos en la carpeta administrada, el formato de manejo, la configuración y el esquema.

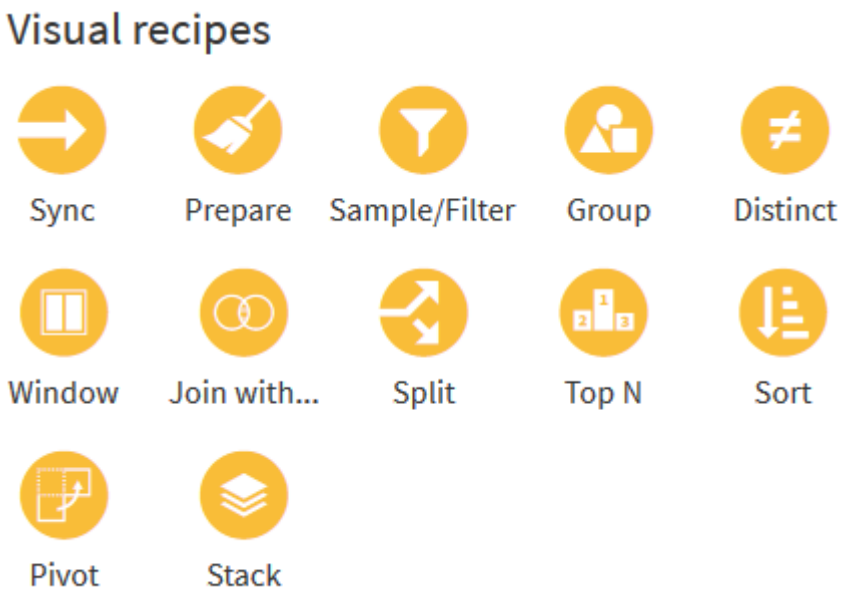


Figura Anexo 1.- Recetas Visuales de Dataiku. Fuente: Elaboración propia con el software Dataiku DSS (2019).

Por otro lado, se encuentran los "*coders*" que son aquellas personas que saben programar y por lo tanto no están limitados a usar las recetas de visualización ya que Dataiku DSS cuenta con recetas de codificación como se puede ver en la figura anexo 2, las cuales permite la programación en los siguientes lenguajes: Python, R, Spark Scala, SQL, Hive, Pig, Impala, PySpark, Spark R, Spark SQL y Shell.

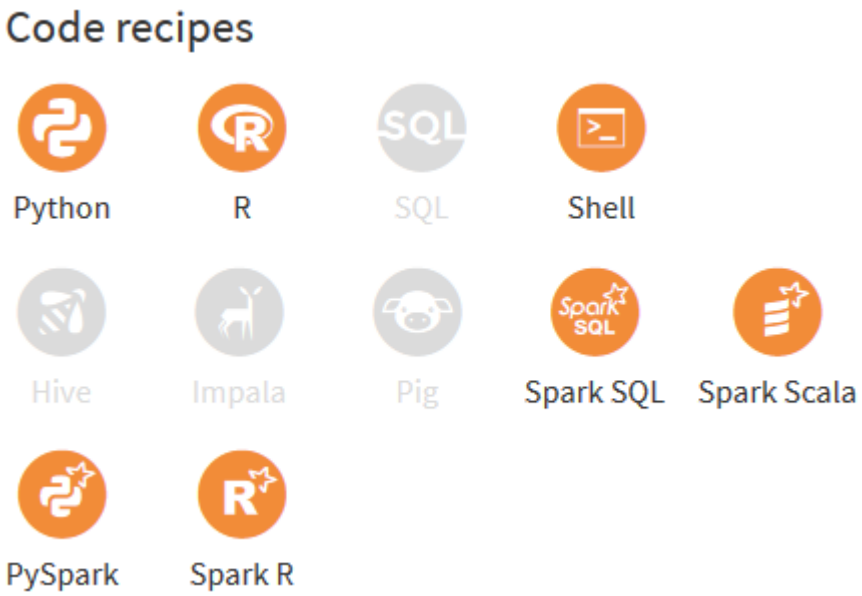


Figura Anexo 2.- Recetas de Código de Dataiku. Formato: Elaboración propia con el software Dataiku DSS (2019).