FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE COMPUTACIÓN Y SISTEMAS

SISTEMA DE RECOMENDACIONES PARA IDENTIFICAR A LOS MEJORES CLIENTES POTENCIALES EN EL PROCESO DE ALQUILER DE LOCALES EN UN CENTRO COMERCIAL

PRESENTADA POR

EDGAR LUIS APESTEGUI INOCENTE

JOAN OSCAR BASILIO ORDOÑEZ

ASESOR

AUGUSTO ERNESTO BERNUY ALVA

CARLOS CHRISTIAN ACUÑA FLORES

TESIS

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE COMPUTACIÓN Y SISTEMAS

LIMA – PERÚ

2020

USMP
UNIVERSIDAD DE SAN MARTÍN DE PORRES

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS

SISTEMA DE RECOMENDACIONES PARA IDENTIFICAR A LOS MEJORES CLIENTES POTENCIALES EN EL PROCESO DE ALQUILER DE LOCALES EN UN CENTRO COMERCIAL

TESIS

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE COMPUTACIÓN Y SISTEMAS

PRESENTADA POR

APESTEGUI INOCENTE, EDGAR LUIS

BASILIO ORDOÑEZ, JOAN OSCAR

LIMA – PERÚ

2020

El presente trabajo está dedicado a nuestros seres queridos quienes nos apoyaron y nunca dejaron de alentarnos en todas las etapas de la carrera.

ÍNDICE

ÍNDICE DE TABLAS

Página

ÍNDICE DE FIGURAS

Página

RESUMEN

Este proyecto tiene como objetivo diseñar un sistema de recomendaciones basado en la generación de pronósticos de ventas que sea económicamente eficiente para la toma de decisiones del Centro Comercial, esta eficiencia se logra a partir de que el sistema está programado para pronosticar las ventas de cada arrendatario actual y las ventas de cada potencial arrendatario del Centro Comercial, con ese conocimiento el arrendador se encuentra en capacidad de tomar mejores decisiones con respecto a la negociación de los contratos de arrendamiento, su renovación a plazo determinado o indeterminado, o su resolución. El proyecto tiene como guía la metodología CRISP-DM, compatible con proyectos de investigación de Machine Learning.

La generación de proyecciones de ventas fue posible utilizando el algoritmo XGBoost de Machine Learning, que es reconocido por elaborar soluciones sobre proyecciones de ventas para tiendas y tener técnicas de optimización de datos que le permite procesarlos diez veces más rápido que otras soluciones de Machine Learning. En este proyecto, para que el algoritmo procese los datos de las ventas fue necesario crear un dataset que pueda ser interpretado por el algoritmo, luego al dataset se le agregaron variables de serie de tiempo para que se generen mejores proyecciones de ventas, las cuales al final de este proyecto llegaron a superar el 70% de acierto, lo que significa que son una herramienta útil para el Centro Comercial.

Palabras clave: proyección, Machine Learning, XGBoost, Centro Comercial, cliente potencial.

ABSTRACT

This project aims to design a recommendation system based on the generation of sales forecast economically efficient for the decision-making of the Shopping Center, this efficiency is achieved from the fact that the system is programmed to forecast the sales of each current and prospective tenant (client) of the Shopping Center, with this knowledge the Shopping Center is able to make better decisions regarding the negotiation of rental agreements, their renewal for specific or indefinite term, or their resolution. The project is guided by the CRISP-DM methodology, compatible with Machine Learning research projects.

The generation of sales projections was possible using the Machine Learning XGBoost algorithm, which is recognized for developing solutions on sales projections for stores and having data optimization techniques that allow it to process data ten times faster than other Machine Learning solutions. In this project, a dataset was developed with the aim of being interpreted by the algorithm, so that it processes the sales data. After this, time series variables were added to the dataset to generate better sales projections. At the end of this project, the aforementioned projections exceeded 70% of success, which means that they are a useful tool for the Shopping Center.

Keywords: projection, Machine Learning, XGBoost, Shopping Center, current and prospective tenant

INTRODUCCIÓN

Hoy en día, tener conocimiento sobre las proyecciones de venta es de vital importancia para las empresas debido a que les permite anticiparse a situaciones inesperadas y eliminar la incertidumbre en el futuro. Los algoritmos de aprendizaje automático (Machine Learning) para empresas pueden usarse para muchos beneficios, como conocer mejor las necesidades de los clientes, establecer una relación entre la satisfacción del cliente y las ventas, optimizar los procesos de selección, predecir tendencias, entre otros.

El proyecto tiene como objetivo optimizar el proceso de alquiler de locales mediante el prototipo de un sistema de recomendaciones que identifica al cliente potencial que logrará un mayor número de ventas, logrando reducir el tiempo que toma conocer qué local venderá más. A fin de cumplir con el objetivo se hará uso de Machine Learning para la proyección de ventas de los locales.

Para el desarrollo de la investigación, se ha decidido separarlo en seis capítulos. El primer capítulo es el planteamiento del problema, donde se detalla el problema a solucionar, los objetivos y la justificación de la investigación. El segundo capítulo corresponde al marco teórico, los antecedentes y la definición de términos básicos. El tercero describe el método utilizado durante la investigación para cumplir con los objetivos. El cuarto plantea el desarrollo en sí del proyecto y muestra la ejecución de la metodología descrita en el capítulo tres. En el capítulo quinto se entregan los

resultados, los hallazgos y la explicación de la información obtenida. En el capítulo sexto, se plantea la discusión, se muestra la interpretación de los resultados del capítulo cinco y se reafirman o contradicen las bases teóricas encontradas en el capítulo dos. Finalmente se entregan las conclusiones de este proyecto, todas ellas responden adecuadamente a los objetivos planteados en esta investigación siendo la conclusión principal la generación de recomendaciones que permitan el apoyo a la toma de decisiones de Gerencia General sobre las empresas que obtendrán mayores ventas a futuro superando el nivel de acierto del 70%.

CAPÍTULO I
PLANTEAMIENTO DEL PROBLEMA

1.1 Situación problemática

El Centro Comercial en estudio genera ingresos mediante el alquiler de sus espacios comerciales (ingreso fijo) y un porcentaje del total de ventas de los locales a quienes alquila (ingreso variable), siendo estas sus principales fuentes de ingresos. Teniendo un total de 322 locatarios en Lima y Arequipa.

Las áreas del Centro Comercial poseen un sistema donde registran la venta diaria y la cantidad de transacciones realizadas por los arrendatarios, a partir de ahora serán denominados locatarios. Esta información es brindada al Centro Comercial, según la cláusula séptima de las Normas Generales de Arrendamiento, mediante correos electrónicos y formatos impresos, la información recibida de parte de los locales es ingresada al sistema al finalizar el día por el área de operaciones.

Cuando un cliente ha tenido una baja tendencia de ventas y su contrato dentro del Centro Comercial está por terminar, Gerencia General, basándose en su experiencia, estima sus ventas mensuales y las compara con el fin de identificar que cliente logrará mejores ventas, teniendo así

la oportunidad de decidir si renovar su contrato o contratar a un nuevo cliente potencial en su lugar.

La decisión de renovar o contratar a un nuevo local puede llegar a tomar cinco días en la estimación de ventas para los clientes vigentes y potenciales. La demora del proceso ya ha perjudicado a la empresa debido a que se contratan clientes que no logran alcanzar los niveles de venta proyectados. Esto causa pérdidas para el Centro Comercial porque existen casos en los que se escoge un cliente potencial y este no logra alcanzar las ventas que se esperaban en el proceso de selección.

1.2 Definición del problema

Demora en la identificación de los mejores clientes potenciales durante el proceso de alquiler de locales causado por el retraso en la generación y poco nivel de confianza de las proyecciones de ventas.

1.3 Formulación del problema

1.3.1 Problema General

Ante la problemática descrita, la pregunta que la investigación busca responder es:

- ¿Cómo optimizar la identificación de los mejores clientes potenciales durante el proceso de alquiler de locales?

1.3.2 Problemas específicos

Para lograr contestar el problema general, se formulan dos preguntas específicas que la investigación buscará responder.

- ¿Qué características son las más importantes para la generación de ventas a futuro?

- ¿Cómo predecir las ventas a futuro respetando el diverso flujo de ventas de cada cliente?

1.4 Objetivo general y específicos

1.4.1 Objetivo General

El objetivo principal de la investigación es:

- Formular un sistema de recomendaciones que permita identificar a los mejores clientes que van a alquilar los locales comerciales, ya sea en el ingreso o en la renovación.

1.4.2 Objetivos específicos

Para lograr el objetivo principal, es necesario alcanzar los siguientes objetivos específicos:

- Identificar los factores más importantes que influyen en las proyecciones de ventas.
- Generar las proyecciones de ventas para los locatarios y clientes potenciales.

1.5 Importancia de la investigación

El desarrollo de una solución para el Centro Comercial requiere de la investigación en el campo interdisciplinario de Data Science, específicamente en el uso de Machine Learning, gracias a su capacidad de aplicar modelos matemáticos a grandes volúmenes de datos. El estudio servirá para identificar de manera más rápida al mejor cliente potencial con mayores ventas para su ingreso al Centro Comercial. La investigación ayudará a las áreas estratégicas de negocio (marketing y Gerencia General) acortando tiempos en la generación proyecciones de ventas.

A diferencia de otros proyectos de investigación, el presente usará identificación de variables para el sector retail aplicadas a

Machine Learning, para reducir la incertidumbre ya sea en las proyecciones de venta por cada cliente como en la selección de clientes potenciales.

1.5.1 Alcance

El proyecto tuvo como alcance los siguientes puntos en mención:

- El proyecto tiene como finalidad pronosticar las ventas de los locatarios.
- El proyecto en mención considerará los locales de las sedes de Lima y Arequipa
- Generar proyecciones de ventas de los locales (excepto tiendas anclas) para el apoyo al proceso de selección de clientes potenciales de los rubros existentes.
- Se desarrollará un módulo de proyecciones en el sistema de gestión de ventas vigente en el Centro Comercial (sistema MAD), por ende, no se tendrán costos de infraestructura.
- El sistema no apoyará al incremento de ventas de cada uno de los locales.

1.5.2 Limitaciones

El proyecto se encontró limitado por los siguientes puntos:

- El tiempo disponible para el desarrollo de la investigación fue de 3 meses, siendo muy corto para ahondar en un análisis más detallado del negocio.
- El periodo de tiempo de datos históricos de las ventas son 2017, 2018 hasta marzo de 2019.

1.5.3 Beneficiados de la investigación

La realización de la investigación traerá beneficios a tres grupos diferentes.

- Área de marketing: Podrán conocer las proyecciones de venta generadas para cada cliente, permitiendo una pronta toma de acciones correctivas, por ejemplo, lanzamiento de campañas publicitarias y activación de eventos con el fin de incrementar las ventas de los locales.
- Gerencia General: Recibirá recomendaciones sobre renovar contrato o dejar ingresar a un nuevo cliente basadas en las proyecciones de ventas de los clientes vigentes y potenciales. Lo cual a su vez se traduce en una mejora en los ingresos del Centro Comercial al escoger a un cliente potencial que generará mayores ventas.
- Clientes potenciales: Al acortarse el proceso de identificación de clientes potenciales, reciben una respuesta más rápida a la solicitud de ingreso.

1.6 Viabilidad de la investigación

Para la realización del sistema es necesario tener los siguientes roles mostrados a continuación en la siguiente tabla.

Tabla 1

Roles del proyecto

Rol	Nombre	Importancia en el proyecto
Product Sponsor	Guillermo Delgado Aparicio	Encargado de financiar el proyecto.
Product Owner	Guillermo Delgado Aparicio	Tiene la labor de dar a conocer todos los requerimientos solicitados para el sistema y de otorgar la conformidad del mismo.

Product Manager	Joan Oscar Basilio Ordoñez	Responsable del cumplimiento de los objetivos trazados junto a la correcta gestión del proyecto.
Tech Team	Edgar Luis Apestegui Inocente, Gustavo Seminario	Equipo capaz de brindar soluciones tecnológicas a las necesidades del sistema por desarrollar.
Test Team	Edgar Luis Apestegui Inocente y Joan Oscar Basilio Ordoñez	Equipo responsable de la validación de las pruebas de cada módulo y de obtener la aceptación del sistema por parte del Product Owner.

Elaborado por: los autores

Tabla 2

Recursos humanos

Nombre - Rol	Tiempo (meses)	Costo (S/mes)	Total
Joan Basilio Product Manager y Tester	4	S/4,000.00	S/16,000.00
Edgar Apéstegui Desarrollador Web y Tester	4	S/3,000.00	S/12,000.00
Gustavo Seminario Desarrollador Python	4	S/4,000.00	S/16,000.00
Total			**S/44,000.00**

Elaborado por: los autores

Tabla 3

Requerimientos de hardware

Recursos	Cantidad	Costo	Total
PC para Desarrollo	2	S/3,500.00	S/7,000.00
PC AI/Machine Learning	1	S/6,000.00	S/6,000.00
Total			**S/13,000.00**

Elaborado por: los autores

Según los recursos mostrados en cada tabla, se tiene como resultado una inversión total requerida para el sistema de S/57,000.00 nuevos soles.

Tabla 4

Resumen de costos del proyecto

Recursos	Total
Recursos Humanos	S/44,000.00
Recursos Infraestructura	S/0.00
Recursos Técnicos	S/13,000.00
Total	**S/57,000.00**

Elaborado por: los autores

CAPÍTULO II
MARCO TEÓRICO

2.1 Antecedentes de la Investigación

Para realizar esta investigación se tomaron en cuenta antecedentes acerca de la aplicación Machine Learning, que demuestran lo versátil de la aplicación y su impacto en diferentes áreas de la vida de los seres humanos.

2.1.1 Machine Learning aplicado a la predicción de ventas en una empresa distribuidora

El presente proyecto realizado por Kreplak (2018) nace frente al problema de una empresa distribuidora de productos alimenticios que debate entre la planificación de compras y las posibles ventas por realizar, con factores como la duración de los productos con los que comercializa y las pérdidas que puede ocasionar una predicción errónea debido a que se realiza de manera subjetiva.

Teniendo como un objetivo obtener un modelo predictivo de ventas, el autor desarrolló una solución en lenguaje Python en base a un histórico de datos acumulados de más de un año con el objetivo de predecir

la venta de ítem por tienda. Kreplak estableció las variables a analizar (número de transacciones, productos comercializados, días festivos, precio del petróleo, clasificación de productos, tiendas, entre otras) así como los pasos a seguir:

- Carga de datos cuya venta sea diferente de cero mediante archivos en formatos csv.
- Cálculo de promedio de ventas en lapsos de tiempo establecidos.
- Análisis de las variables a usar y sus características en relevancia con las proyecciones a realizar.
- Aplicación del algoritmo LGBM, ajustando los parámetros a establecer con el fin de obtener el resultado más cercano a la realidad.

El autor realizó el estudio en base al algoritmo Light Gradient Boosting Machine (LGBM), basado en árboles de decisión, ajustando los parámetros para aumentar la velocidad, precisión, overfitting, entre otras, debido a que la mayoría de los árboles de decisión crecen por niveles, en cambio, LGBM crece por hojas según la condición establecida en cada una y debe ser aplicada en un dataset grande para evitar el overfitting.

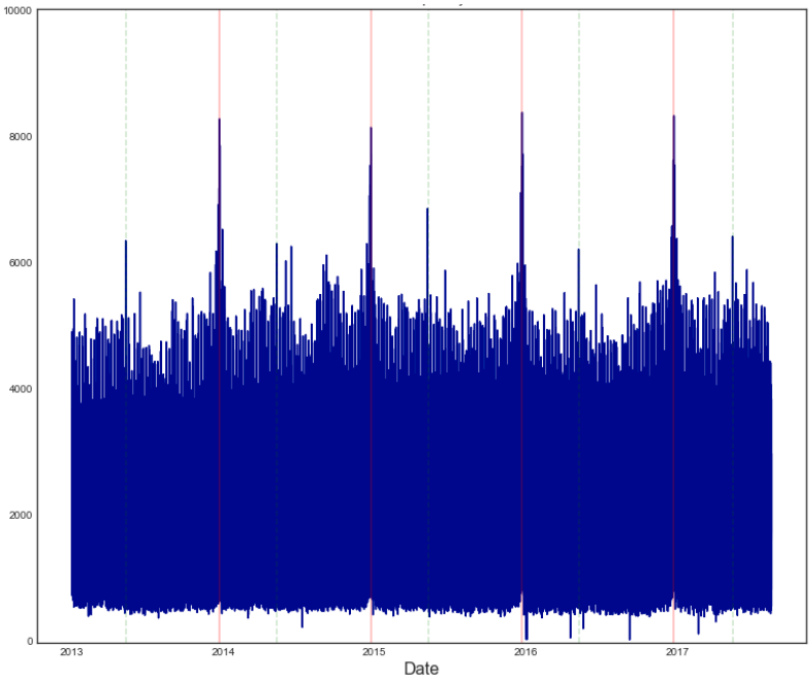


Figura 1. Distribución de transacciones por día

Fuente: Kreplak, 2018

Esta tesis tuvo un enfoque similar al objetivo propuesto en nuestra investigación demostrando la importancia de las variables como eventos, rubros y ventas por ítem, además de comprender los parámetros usados en arboles de decisión. La empresa comparte el mismo rubro comercial, lo que fortalece el enfoque utilizado para atacar el problema de proyecciones de ventas utilizando diversas herramientas de Machine Learning.

2.1.2 Metodología para la aplicación de Machine Learning en la clasificación de pacientes

García (2014) ejecutó un proyecto realizado en el Hospital Infantil Napoleón Franco Pareja de Cartagena en base al problema que se presenta en los centros de salud, tal y como lo menciona la Superintendencia de Salud Nacional de Colombia, específicamente en el área de urgencias en donde tienen la necesidad de una atención rápida de pacientes, los

cuales muchas veces no son correctamente distribuidos según su urgencia y genera el colapso de esta área. La directiva del hospital en estudio detectó sus posibles causas, siendo estas la infraestructura, asignación de recursos y personal no capacitado, los cuales generan el malestar en los pacientes que buscan una pronta ayuda.

De acuerdo con diversos estudios de resultados exitosos en la predicción de diversas enfermedades, la autora realizó un modelo de Machine Learning en redes neuronales, máquinas de soporte vectorial y regresión lógica, para así realizar la evaluación de sus resultados, con el fin de obtener una pronta clasificación de los pacientes en términos de nivel asistencial. Para llevar a cabo este proyecto, la autora catalogó las enfermedades reportadas según su clasificación, escalas y variables, cuyos valores fueron 1 (alta complejidad) o 0 (baja complejidad). A continuación, el autor analizó la información de las historias clínicas, sin embargo, estas no contenían información relevante para el proyecto que evalúe las escalas establecidas, es así como la autora propuso un nuevo formato (aceptado por la Coordinación de Calidad y Epidemiología) y eliminó variables que no favorecían al análisis, posteriormente, obtuvo resultados positivos en la clasificación de pacientes.

Una vez obtenidas las variables (enfermedades, síntomas, datos personales, entre otros.), así como su valorización (porcentaje casos registrados) y escalas, el autor procedió a realizar la depuración de aquellas variables que no aportaron valor a la investigación, la autora utilizó el 80% de la muestra para realizar el entrenamiento y validación, mientras que el 20% para para las pruebas, posteriormente, generó la matriz de transformación en la que asignó valores a cada componente, así como su porcentaje de varianza y acumulado. En la varianza acumulada, creó 3 conjuntos de datos y aquellos con los valores más altos, que llevan a entender la fiabilidad de cada componente, sirvieron para el entrenamiento de las tres técnicas de Machine Learning mencionadas. Del entrenamiento y validación, el modelo de redes

neuronales obtuvo el mejor resultado con una probabilidad de clasificación del 85%, por consiguiente, el autor procedió a realizar las pruebas, donde obtuvo un resultado de 84%, muy parecido a las muestras de entrenamiento.

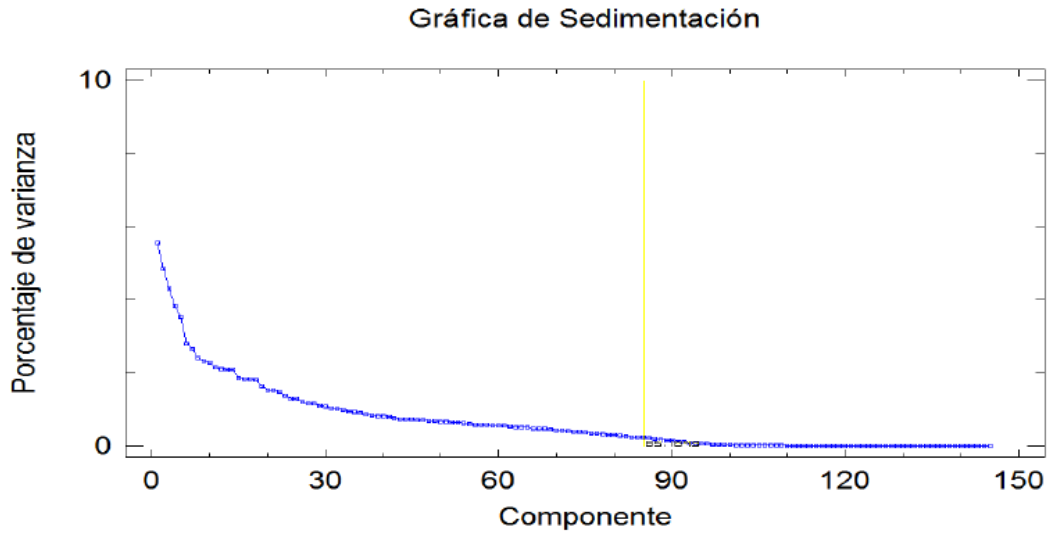


Figura 2. Sedimentación de componentes vs porcentaje de varianza

Fuente: García, 2014

Por último, brindó recomendaciones para tener en cuenta en futuras mejoras del proyecto, como realizar una investigación de los pacientes y los lugares de donde provienen, para así tener variables que ayuden a mejorar la exactitud de los resultados.

El marco de trabajo usado en la investigación de García ayudó a comprender el desarrollo de una solución basada en Machine Learning en cada ejecución de las tareas desarrolladas en este proyecto, como son: la obtención y limpieza de la data; la selección de variables; y, por último, la distribución porcentual de la muestra para el entrenamiento y las pruebas.

2.1.3 Uso de modelos basados en series de tiempo para la predicción de la demanda de habitaciones

Pallares (2014) realizó un proyecto para el sector hotelero, con el fin de apoyar a la toma de decisiones proyectando la cantidad de habitaciones de demanda, cuya muestra es desde el 1 de julio de 2008 hasta el 30 de junio de 2014 de un hotel en Colombia, para ello se basó en diferentes criterios como las estaciones, días de semana, eventos especiales y festivos, y tuvo en cuenta otras variables más complejas, como los alargues y acortes de la estadía de huéspedes. El proyecto es considerado muy importante puesto que el sector hotelero tiende a ser difícil en cuanto a proyecciones, debido a la inestabilidad que se presenta en el rubro en diferentes partes del mundo, lo que conlleva a la cancelación intempestiva de reservas hoteleras y paquetes turísticos.

El autor en su investigación menciona al Revenue Management como una de las herramientas más utilizadas en el mercado hotelero debido a que favorece la maximización de los ingresos basándose en vender un producto o servicio al mejor postor al precio más alto y en el momento exacto; no obstante, el autor también enfatiza que el Revenue Management frecuentemente genera un impacto negativo en los ingresos causado por predicciones erróneas. El autor menciona que los problemas que se pueden presentar al utilizar pronósticos errados generarían un resultado negativo, debido a que, si no es correcto, se puede vender a un precio muy bajo o alto y no obtener la estrategia establecida.

Para el realizar el pronóstico de demanda de habitaciones se tomaron en cuenta dos enfoques, el primero basado en modelos de serie de tiempo que consiste en conocer el comportamiento de la demanda de las habitaciones en intervalos de tiempo anteriores a la llegada del cliente; y

el segundo enfoque está basado en el pronóstico de las reservas realizadas con varios días de anticipación, así como días de la semana, cantidad de días festivos, meses del año y temporada.

Con respecto a la parte técnica de la tesis del autor, se revisaron los resultados de aplicar Machine Learning Ridge Regression, Kernel Ridge Regression y Redes Neuronales (Perceptron Multicapa y Redes Neuronales Función Base Radial) y se observó que el autor logró obtener resultados más cercanos a la realidad con la técnica Ridge Regression, que se basó en una análisis de las reservas hechas con días de anticipación, a las cuales luego se les agregaron las reservas durante las temporadas altas y bajas, los días de semana y meses del año festivos y no festivos.

En la siguiente figura se observan los resultados de los pronósticos en líneas rojas, mientras que los resultados reales obtenidos se presentan en líneas negras.

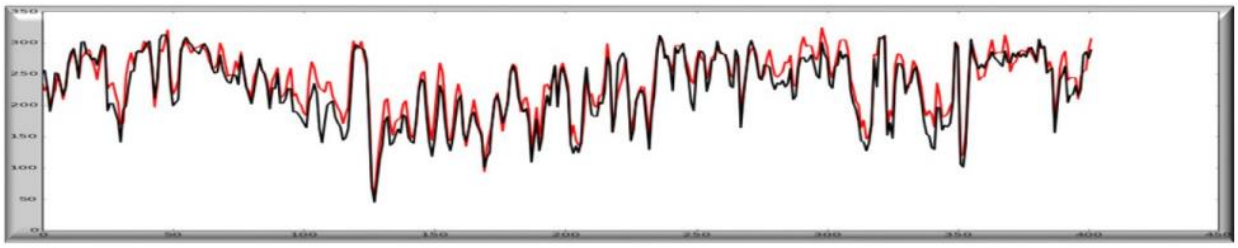


Figura 3. Pronóstico vs Ocupación Diaria

Fuente: Pallares, 2014

Esta investigación demostró que a más criterios de análisis que se adicionen al modelo de series de tiempo, más exacto es el pronóstico que se obtiene. Es por este motivo que la investigación nos genera el aporte necesario para poder aplicar en nuestra investigación variables basadas

en intervalos de tiempo antes de la fecha a pronosticar para la realización de las proyecciones.

2.1.4 Clasificación de empresas utilizando algoritmos de Machine Learning

Carreño realizó un proyecto en el año 2017 en España cuya finalidad era identificar potenciales empresas en el ámbito de fraudes fiscales por evasión del impuesto sobre el valor añadido (IVA). Según el IC3 (Internet Crime Compilant Center), cuya finalidad es investigar y detectar el crimen a través de internet, el número de quejas recibidas y pérdidas en dólares en internet ha ido en incremento, lo que demuestra el alcance global de este problema.

Tabla 5

Estudio de pérdidas a causa del fraude por IC3

Año	Quejas recibidas	Pérdidas en Dolares
2011	314,246	485253871 Millones
2012	289874	581441110 Millones
2013	262813	781841611 Millones
2014	269422	800492073 Millones

Fuente: Carreño, 2017

Para la investigación, el autor utilizó data real proporcionada por la Agencia Estatal de Administración Tributaria (AEAT) comprendida entre los años 2010 y 2015. Primero, trabajó con un conjunto de datos de libre acceso conocido como Mammography Dataset para realizar sus pruebas y medir sus resultados, este contiene data sobre la calcificación en las mamas siguiendo la relación de clase positiva (sana) y clase negativa (enferma).

El autor utilizó el 80% de la data para el entrenamiento y el 20% para las pruebas, posteriormente aplicó el mismo procedimiento sobre los datos proporcionados por la AEAT, dicha data está separada por años: cuatro años se usaron para el entrenamiento y un año de datos restante para las pruebas.

El autor usó diversas técnicas Machine Learning como regresión logística, árbol de decisión, random forest, y otras más. Tras realizar los experimentos, los resultados obtenidos fueron analizados mediante las métricas de una matriz de confusión para encontrar la precisión y exactitud del modelo. Llegaron a la conclusión que el algoritmo capaz de obtener los mejores resultados para identificar a las empresas potenciales fue el Random Forest, gracias a que obtuvo una precisión cercana a 99%, además de una especificidad y sensibilidad cercanos al 100%.

Tabla 6

Comparación de técnicas de aprendizaje

	LR	Dtree	RF	NB
Accuracy	94,497	94,137	99,836	66,173
F-Measure	94,191	94,057	99,836	68,961
Sensibility	89,219	92,784	99,845	75,154
FPR	0,225	4,509	0,171	42,806
Specificity	99,775	95,49	99,828	57,193
Precision	99,749	95,365	99,828	63,711
AUC	95	97	99	74

Fuente: Carreño, 2017

El objetivo alcanzado por el autor nos permite conocer que utilizando algoritmos de Machine Learning es posible identificar a empresas potenciales, además, bajo la investigación del autor logró reconocer la variable más determinante para cumplir con su objetivo, guiándonos en la

identificación de los factores más influyentes para la generación de las proyecciones.

2.1.5 Predicción de resultados con técnicas de Machine Learning aplicado a la ciencia del deporte

Calderón (2017) realizó un proyecto en Madrid, España, partiendo del hecho de la dificultad de predecir los resultados en el fútbol, y teniendo como base algunos intentos de predecir resultados usando la estadística como herramienta. El proyecto busca predecir dichos resultados mediante el uso de Machine Learning y demostrar que existen factores con mayor importancia a otras que pueden determinar la victoria.

El autor utilizó la base de datos de European Soccer Database ubicada en Kaggle, dicha base incluye partidos desde la temporada 2008 hasta la 2016 de once ligas distintas de fútbol, además, incluye atributos de los jugadores y equipos recuperados de EA Sports FIFA como la capacidad de reflejos de los porteros, definición de cara a puerta, pie favorito, entre otras más. Posteriormente, realizó una limpieza de la data eliminando las filas que no contenían valor bajo un criterio previo, con el fin de determinar que data debe mantenerse para el modelo de Machine Learning. El lenguaje de programación escogido fue Python, gracias a que contiene librerías que ayudan a la transformación y el análisis de los datos para el entrenamiento y la predicción.

Partiendo de que cada nodo es un equipo, el autor procesó la información por un software para el análisis y visualización de redes a través de grafos. Posteriormente, estableció los parámetros (rating de los jugadores) para obtener las predicciones hallando cierta tendencia de los equipos locales a ganar.

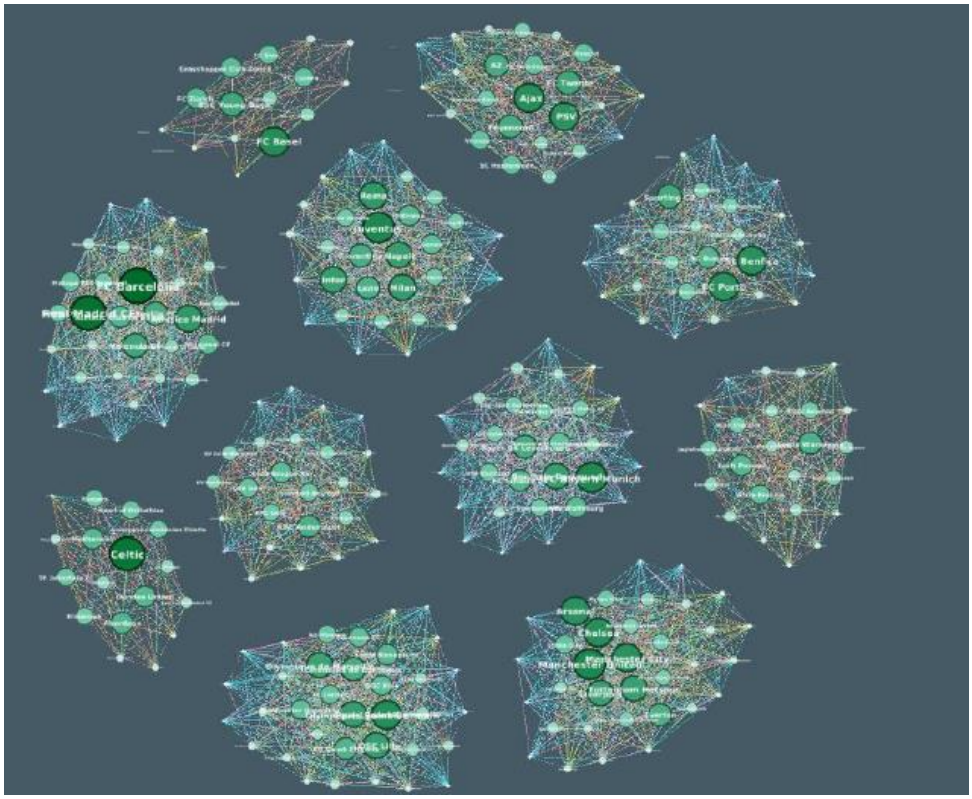


Figura 4. Grafo de las ligas usando el software Gephi

Fuente: Calderón, 2017

El autor logró superar el nivel de acierto de las casas de apuestas de 53% al utilizar el algoritmo de Logistic Regression, además, demostró que su modelo tenía una mayor eficiencia para determinar al ganador con tan solo cinco características, finalmente, recomendó que para obtener mejores resultados se pueden añadir más variables que no consideraron en esta investigación como los suplentes, posiciones en el campo, entre otras.

La tesis nos enseña que, a pesar de tener una base de datos confiable, se debe de aplicar un proceso de limpiado al dataset, el cual consiste en establecer criterios previos para la eliminación de los datos, para así obtener mejores resultados como respuesta del algoritmo de Machine Learning.

2.2 Bases teóricas

2.2.1 Machine Learning

Los inicios del término Machine Learning comenzaron cuando Alan Turing plantea una prueba de habilidad de una máquina y la opción de que esta sea capaz de pensar como un ser humano (Turing, 1950). En 1952 se realiza el primer programa de aprendizaje automático (juego de damas) por Samuel Arthur, quien fue pionero en juegos de computadora (Schaeffer y van den Herik, 2002). Machine Learning se define como la aplicación de algoritmo y técnicas con la capacidad de mejorar de forma autónoma, esto ocurre gracias a los siguientes grupos: fuentes de información, el algoritmo que procesa la información, el autoaprendizaje y el uso del software que servirá de intermediario entre la máquina y el usuario (ManagementSolutions, 2018).

2.2.2 Categorías de Machine Learning

Las categorías del Machine Learning han sido desarrolladas por varios autores, entre ellos Jeff Barnes, quien el 2015 definió 3 categorías, dos de los más utilizadas en Machine Learning son las siguientes: el aprendizaje supervisado y el no supervisado.

a) Aprendizaje supervisado

Es una categoría de Machine Learning que utiliza datos conocidos para construir un modelo capaz de hacer predicciones. Lo datos están compuestos por datos de entrada y de salida y son los que permiten al algoritmo la generación de predicciones (Barnes, 2015).

b) Aprendizaje no supervisado

La computadora deberá identificar patrones y relaciones para generar un nuevo modelo predictivo, puesto que los datos de entrada, así como los de

salida son desconocidos. En resumen, la computadora deberá reconocer sus propios grupos de datos semejantes entre sí. (Barnes, 2015).

2.2.3 Conceptos de Machine Learning Supervisado

a) Clasificación

Muchas veces usados con datos confidenciales con protocolos establecidos, para así obtener resultados que se encuentren dentro de un conjunto de datos conocidos, dentro de los ejemplos más comunes se encuentran las clasificaciones binarias (si/no) donde el objetivo del modelo es determinar si el objeto en estudio pertenece a una categoría o no. Los principales aportes se evidencian en predicciones médicas, así como en predicciones financieras, detección de correos spam, entre otros (Bost, Popa, Tu, y Goldwasser, 2015).

b) Regresión

A diferencia del modelo clasificación que tiene como resultado categorías, los modelos de regresión presentan sus resultados en cantidades o valores numéricos (variables dependientes) en función a sus entradas (variables independientes), estos resultados ayudan a identificar la relación entre las entradas y su salida. (Perez, 2019).

2.2.4 Algoritmos de Machine Learning Supervisado

Los algoritmos de Machine Learning Supervisado han sido definidos por varios autores, entre ellos Leo Breiman, (2001), Jeff Barnes (2015), Yan-yan Song y Ying Lu (2015) y Jose Martinez (2018), quienes describieron cada algoritmo desde su propia experiencia en el campo. Los algoritmos más importantes definidos por los citados autores y que fueron estudiados para la comprensión general de los algoritmos de Machine Learning, son los siguientes.

a) Regresión lineal (Lineal Regression)

Es un tipo de análisis estadístico que analiza las relaciones de una variable independiente respecto a otras variables dependientes. Se utiliza para obtener resultados basados en muestras aleatorias, como, por ejemplo, la relación entre la altura de una persona y la talla del calzado que usa (Barnes, 2015).

b) Regresión logística (Logistic Regression)

Referido a proyectos de clasificación, predice la probabilidad de una variable dependiente (esta variable presenta valores binarios), genera probabilidades entre 0 y 1 (valor mínimo y máximo respectivamente) y determina a que clasificación pertenece, por ejemplo, se utiliza para establecer la probabilidad de que un correo sea spam. Es muy usado cuando se tiene una gran cantidad de datos y relaciones no complejas (Martinez, 2018).

c) Árboles de decisión (Decision trees)

Es un algoritmo que predice los valores de respuesta mediante el cumplimiento de reglas de decisión, evalúa todos los datos de entrada y todos los puntos de decisión posible, eligiendo al que obtenga el mejor resultado. Los componentes principales de este algoritmo son los nodos y ramas, y los pasos en la construcción del modelo dividir, detener y podar. Se inicia con el nodo raíz, representa una decisión (si esta es positiva o negativa) y genera la subdivisión de todos los conjuntos en dos o más. Los nodos internos o de prueba son todas las condiciones que determinan la estructura del árbol, y los nodos finales o hojas; son los que determinan el resultado final después de una serie de eventos cumplidos. Las ramas son las direcciones que se tomaron como resultado de las decisiones, parten desde el nodo raíz hasta la hoja o resultado (Song y Lu, 2015).

d) Selvas aleatorias (Random forest)

Usado para tareas de clasificación o regresión, consiste en la creación de múltiples árboles de decisión generados de forma aleatoria en cada nodo (cada conjunto de entrenamiento es distinto de los ya entrenados), evitando así el sobreajuste conforme se generan más árboles y alcanzando un valor límite del error de generalización (Breiman, 2001).

2.2.5 Overfitting

Se genera cuando el modelo presenta un buen rendimiento con la data entrenada, mas no con los data destinada a la evaluación. Dicho escenario ocurre cuando el modelo memorizó la data de entrada y no es capaz de analizar correctamente otra data destinada al mismo propósito (Amazon, 2016).

2.2.6 Underfitting

Se refiere al desajuste o aprendizaje no concluido, generalmente cuando se obtiene el mismo resultado, el modelo no identifica la relación de los datos de entrada con los resultados esperados, se evita determinando más variables que ayuden al aprendizaje (Amazon, 2016).

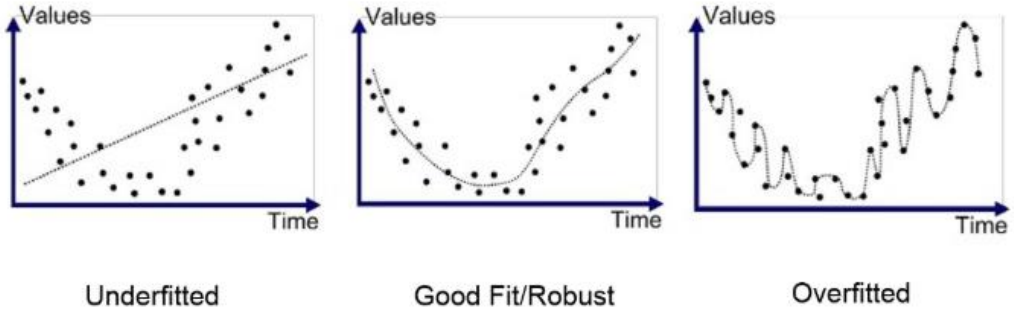


Figura 5. Posibles resultados del ajuste de un modelo

Fuente: Román, 2019

2.2.7 Gradient Boosting Machine (GBM)

Es una técnica de aprendizaje, cuya estrategia es crear nuevos modelos de forma secuencial en las que el aprendizaje se va adaptando, de este modo genera una mejor precisión en los resultados. El algoritmo se basa en una función de perdida de todo el conjunto, el cual será determinado por el investigador facilitando así que los GBM sean flexibles para cualquier proyecto en particular (Natekin y Knoll, 2013).

2.2.8 XGBoost

Los autores Chen y Guestrin (2016) mencionan a XGBoost como un sistema de Machine Learning escalable para algoritmos de árboles potenciados ampliamente reconocido. Este sistema mostró resultados importantes para problemas de predicciones de ventas de tiendas, clasificación de textos, categorización de productos y más, participando en grandes concursos como el KDDCup 2015 donde fue utilizado en cada uno de los diez mejores equipos ganadores. El factor más importante de XGBoost es su gran escalabilidad de escenarios debido a sus sistemas embebidos y algoritmos de optimización. Gracias a ello, XGBoost es capaz de procesar hasta diez veces más rápido comparado a otras soluciones populares, como Gradient boosting Machine, sobre una misma máquina como recurso y escala a billones de ejemplos en configuraciones de memoria limitada.

2.2.9 Feature engineering

Según Zheng y Casari (2018), definen al feature engineering como la actividad en la cual se extraen representaciones numéricas de la data sin procesar para convertirlos en datos capaces de ser interpretados por un modelo de Machine Learning. Este paso es muy importante debido a que un buen manejo de las variables permitirá que el modelo genere resultados de mayor calidad.

2.2.10 Data Science

Según la definición otorgada por Provost y Fawcett (2013) se define como Data Science al conjunto de procesos, técnicas y principios que permiten la generación de conocimiento a partir del estudio de los datos. Para esto es necesario el acceso a datos que hayan sido trabajados previamente por herramientas de procesamiento de datos, de estas se obtiene un mejor beneficio frente a datos que no han sido tratados. Los nuevos conocimientos adquiridos por parte de Data Science permiten a las empresas tener un mayor soporte a la hora de tomar decisiones de negocio, esto hace que incluso existan empresas que tomen sus decisiones en gran parte basadas en los conocimientos obtenidos del Data Science frente a su propia intuición.

Los autores a su vez plantean un escenario en donde se muestra el soporte en la toma de decisiones que pueden otorgar soluciones que utilizan Data Science en una empresa utilizando como ejemplo a una vendedora que desea contratar servicios de anuncios y debe optar entre su intuición basada en la experiencia en el campo o el análisis de los datos que muestran las reacciones de los consumidores a diversos anuncios obtenidos por mediante el uso de Data Science. Ella puede escoger por cualquier de las dos soluciones: una basado exclusivamente en su experiencia u otra que utiliza únicamente los conocimientos obtenidos por el Data Science, pero a su vez la vendedora puede optar por una tercera opción, que es utilizar ambas opciones como soporte para tomar la mejor decisión sobre sus anuncios. Este corto ejemplo planteado por el autor muestra que las soluciones de Data Science no se basan en la práctica del todo o nada, sino, que el conocimiento recibido por el Data Science apoya a la toma de decisiones y no obliga a la que dicho conocimiento sea el único que se utilice para determinar una decisión de negocio.

Según Hitt (2011), el apoyo a la toma de decisiones basadas en conocimientos a partir de los datos impacta en las empresas. Para demostrarlo, desarrollaron un estudio donde demostraron

estadísticamente que las empresas que toman decisiones de negocio basados en el conocimiento otorgado por los datos estuvieron asociadas al incremento del 5% a 6% en su productividad.

2.2.11 CRISP-DM

Los autores Wirth y Hipp (2000) definen a la metodología CRISP-DM como procesos que permite crear un marco de trabajo para proyectos que implican minería de datos, con el objetivo de hacer un proyecto mejor gestionado y rápido. La metodología CRISP-DM divide el proyecto en 6 fases principales:

a) Entendimiento empresarial

   Esta fase consiste en comprender los objetivos del proyecto desde una perspectiva empresarial, aterrizando la problemática a objetivos específicos para el proyecto.

b) Comprensión de datos

   Se realizan las actividades de recopilación, adaptación e identificación de problemas para así descubrir la información de los datos.

c) Preparación de datos

   Aquí se realizarán todas las actividades que darán como resultado los datos que alimentarán a la herramienta de modelado. Entre estas actividades se encuentran el registro de datos, limpieza de datos y construcción de nuevos datos.

d) Modelado

   Se aplican las técnicas de modelado calibrando los parámetros designados por la herramienta.

e) Evaluación

   Se evalúa el modelo para estar seguros de que cumple lo determinado con los objetivos del comercio.

f) Despliegue

Puede presentarse como un informe o la implementación de un proceso de datos, esto dependerá de lo establecido con el usuario.

2.2.12 Modelo Matemático

Pollak (2011) define al modelo matemático como la versión idealizada de la situación del mundo real teniendo en cuenta la relación de variables, esto permite analizar el comportamiento y comprobar los resultados obtenidos.

2.2.13 Retail

Según la definición de Turienzo (2017), son los productos y servicios ofrecidos al cliente final, siendo muchos sectores parte de este concepto, desde empresas del sector alimenticio y bebidas hasta el sector de suministros de electricidad, gas y agua. Dejando claro que el elemento clave del retail es el consumidor o cliente.

2.3 Definición de términos básicos

2.3.1 Centro Comercial

La Asociación Española de Centros Comerciales y Parques Comerciales (2016) definió a los centros comerciales como un conjunto de establecimientos comerciales que comparten un criterio de unidad y una gestión unitaria.

2.3.2 Tienda Ancla

Eric López Pastor, profesor de Marketing de la Universidad ESAN, junto a Fernando de la Flor, director ejecutivo del Grupo Inmobiliario Penta, comentaron para el diario El Comercio en el año 2011 sobre la definición de una tienda ancla y su importancia dentro de un Centro Comercial.

Pastor (2011) brindó la definición de tienda ancla como la tienda más grande en un Centro Comercial reconocida por su gran capacidad para atraer compradores, variedad de productos y grandes descuentos.

Además, las tiendas ancla pueden ubicarse a extremos o ampliamente separadas de otras tiendas ancla dentro del Centro Comercial con la finalidad de que el consumidor tenga que movilizarse entre las tiendas más pequeñas para poder llegar a ellas (Pastor, 2011).

Para contribuir a la definición de ancla y su importancia, De la Flor (2011) agregó que las anclas son importantes porque generan un gran flujo de visitas que contribuye a las ventas de las tiendas menores.

2.3.3 Proyección de ventas

"Es la cantidad de ingresos que una compañía espera obtener en el futuro procedentes de las ventas" (García, 2017).

2.3.4 Cliente Potencial

Valdivia (2015) se refiere al cliente potencial como una persona o empresa que, sin ser cliente, cumple las condiciones de convertirse en un cliente a futuro.

2.3.5 Dato

Según De Pablos, López-Hermoso, Martín-Romo y Medina (2004) describen al dato como "un carácter o cualquier conjunto de caracteres, pudiendo ser estos datos de tipo numérico, alfabético o alfanumérico, según la naturaleza del carácter que le de forma" (p.16).

2.3.6 Dataset

Según United Nations Department of Economic and Social Affair (2020), se define al dataset como “una colección lógica de valores u objetos de base de datos relacionados a un solo tema” (p.185).

2.3.7 Variable

Niño, Cobos y Mendoza (2001) afirman que “Se considera variable a una zona de memoria referenciada por un nombre de variable, donde se puede almacenar el valor de un dato, que puede cambiarse cuando lo deseemos” (p.27).

2.3.8 Muestra

Hernández, Fernández y Batista (2010) presentaron su definición la palabra como: “La muestra es, en esencia, un subgrupo de la población. Digamos que es un subconjunto de elementos que pertenecen a ese conjunto definido en sus características al que llamamos población” (p.175).

2.3.9 Algoritmos

Schaeffer (2009) describe a los algoritmos como “un método de solución para un cierto problema que, dada una instancia (o sea, los datos particulares del problema), realiza una sucesión finita de pasos deterministas de cómputo y siempre llega al resultado correcto” (p.2).

2.3.10 Pronósticos

Según el diccionario de la legua española, se define la palabra pronóstico como “Acción y efecto de pronosticar” y esta última a su vez como “predecir algo futuro a partir de indicios”. Según estas definiciones, se puede comprender la palabra pronósticos como el conjunto de

acontecimientos que ocurrirán en el futuro respaldados por evidencias, pruebas o manifestaciones.

Según Hoshmand (2010), el objetivo de pronosticar es proporcionar información que ayude a los responsables a una mejor toma de decisiones. Menciona que los pronósticos son una parte importante de la planificación de un sistema y que las organizaciones los necesitan para poder tomar decisiones de forma efectiva y en el momento justo. Además, Hoshmand dice que, si bien los encargados de realizar los pronósticos no tendrán la certeza total sobre lo que sucederá, son capaces de reducir el margen de error sobre las decisiones de negocio.

Según Evans (2003), para hacer pronósticos que ayuden a los negocios, se deben tener en cuenta dos conceptos: la ciencia; al utilizar herramientas estadísticas que permitan mejorar la precisión de los pronósticos, y el arte; en el momento de escoger relaciones para determinar la ecuación más precisa para los pronósticos basándose el conocimiento producto de la experiencia en el negocio.

Para Alon, Qi y Sadowski (2001) “los pronósticos son especialmente útiles para grandes empresas retail quienes posiblemente tienen una mayor cuota de mercado” (p.147). Además, mencionan que “los pronósticos de demanda precisos son cruciales para las operaciones de retail rentables porque sin un buen pronóstico, ya sea demasiada o muy poca mercadería permanecería, afectando directamente los ingresos y la posición competitiva” (p.148).

2.3.11 Percentil

Suárez y Tapia (2012) mencionan sobre el percentil que “Son cada uno de los 99 valores P1, P2, P3,……..P99 que dividen atribución de los datos en 100 partes iguales ” (p.109).

2.3.12 Z Score

Según Mcleod (2019) se refiere a la posición de una puntuación expresada en desviaciones estándar respecto a la media, si el Z se ubica por encima o debajo de la media, esta será positiva o negativa respectivamente.

2.3.13 Rubro

Pérez y Merino (2013) definen a los rubros como conceptos bajo los cuales se agrupan empresas con el fin de clasificarlos según el giro de negocio al cual se dedican.

CAPÍTULO III
METODOLOGÍA

Para el desarrollo de las proyecciones de ventas, se decidió utilizar como guía la metodología CRISP-DM. Dicho método comprende un conjunto de tareas que sirven como base para las mejores prácticas. La sucesión de fases no es necesariamente rígida. Cada fase es estructurada en varias tareas generales que se proyectan en tareas específicas, pero en ningún momento se propone como realizarlas. A continuación, se muestran las fases de CRISP-DM que se utilizarán durante el transcurso del proyecto:

3.1 Comprensión del negocio

Durante esta etapa se reconocen de forma clara los objetivos que se abordarán durante el proyecto. Las tareas incluidas durante esta etapa.

3.1.1 Evaluación de la situación

Durante esta tarea se describe la situación actual del problema antes de la generación del modelo.

3.1.2 Determinar los objetivos de DM

Una vez establecidos los objetivos del negocio, se procede a establecer el objetivo de DM, el cual una vez cumplido debe de soportar el objetivo del negocio.

3.1.3 Criterios de aceptación para DM

Se describe cual será el mínimo necesario por parte del DM para que se implemente a futuro.

3.1.4 Evaluación inicial de herramientas

En esta tarea, se planifica que herramientas serán las necesarias para generar una solución para el negocio.

3.2 Comprensión de los datos

La fase abarca la obtención de la data, interpretar cada uno de los campos y verificar su calidad con el fin de comprender la naturaleza de los datos.

3.2.1 Recolectar los datos iniciales

Se recogen los datos iniciales que serán utilizados durante la investigación

3.2.2 Descripción de los datos

Una vez que se tienen los datos, se debe de describir cada campo (descripción, tipo de dato, valor por defecto).

3.2.3 Verificación de la calidad de los datos

Aquí se determina las acciones que se harán para aquellos valores que pueden hacer ruido en el proceso como valores nulos, valores fuera de rango, entre otros; de esta forma se planificará que acción tomar cuando se encuentren dichos casos.

3.3 Preparación de los datos

En esta fase se estructuran los datos obtenidos mediante técnicas que permitirán usarlos en el proceso DM.

3.3.1 Seleccionar datos

Para esta tarea se determinarán los datos que serán ingresados al proceso basándose en la recolección realiza en la fase anterior.

3.3.2 Limpiar datos

Se procede a ejecutar las actividades de mejora de la calidad de los datos diseñada anteriormente. Después de la tarea el volumen de la data disminuirá dependiendo de las validaciones establecidas.

3.3.3 Estructurar los datos

Durante esta tarea se agregan registros o variables adicionales a los campos. Es posible agregar datos externos a los recogidos con el fin de enriquecer la data y mejorar los resultados esperados.

3.4 Modelado

Se determinan las técnicas de modelado que se utilizarán para el procesamiento de los datos. Estos se realizan bajo los siguientes criterios:

- Ser apropiada para el problema
- Disponer de datos adecuados
- Cumplir con los requisitos del problema
- Tiempo adecuado para generar el modelo
- Conocimiento de la técnica

3.4.1 Selección de la técnica de modelado

Se elige la técnica a utilizar para resolver el problema cumpliendo con los objetivos planteados en la primera etapa.

3.4.2 Generación del plan de prueba

Se decide cual será el procedimiento para determinar la validez del modelo.

3.4.3 Construcción del modelo

En esta etapa se describe la construcción del modelo de Machine Learning, incluyendo la descripción de los parámetros utilizados para su creación.

3.4.4 Evaluación del modelado

Se evalúa el modelo generado con la data de “test” y se generan gráficos o tablas que muestren cómo se comporta el modelo con la data.

3.5 Evaluación

Durante dicha fase se evalúa el modelo generado comparándolo con los criterios de éxito establecidos inicialmente.

3.6 Despliegue

Con el modelo construido y validado, se procede a su aplicación en el negocio permitiendo ver el resultado del modelo.

CAPÍTULO IV
DESARROLLO

4.1 Comprensión del negocio

4.1.1 Evaluación de la situación

A continuación, se mostrará el proceso de selección de cliente potencial que se realiza en el Centro Comercial actualmente. En dicho proceso interviene Gerencia General demorando un tiempo aproximado de cinco días.

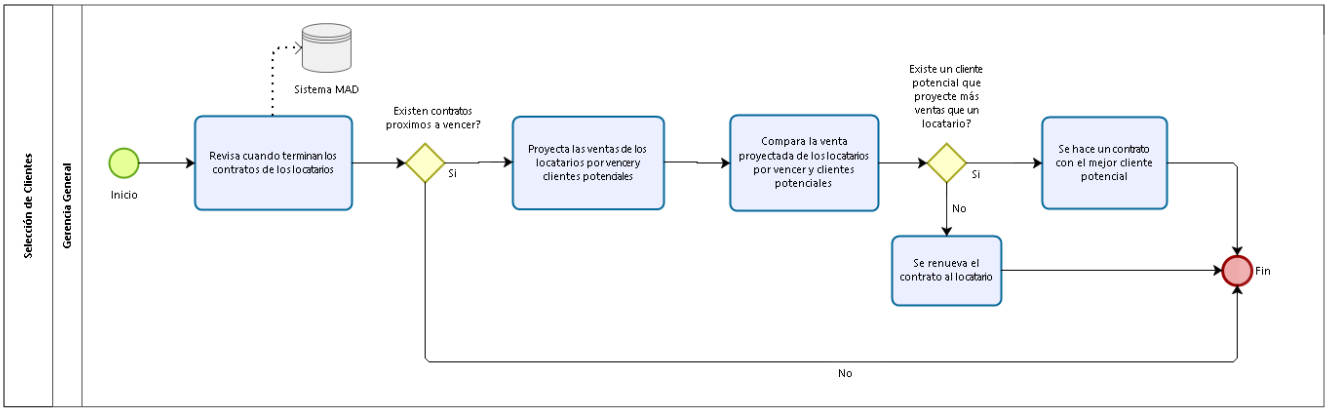


Figura 6. Situación actual del proceso

Elaborado por: los autores

Con la propuesta del proyecto, el proceso de selección de clientes mejorará recortando el tiempo en la toma de decisión. Esto liberará a Gerencia General de la proyección de ventas para identificar al mejor cliente potencial para el Centro Comercial. Cuando el objetivo establecido se alcance, las proyecciones apoyarán a la generación de recomendaciones de los clientes potenciales.

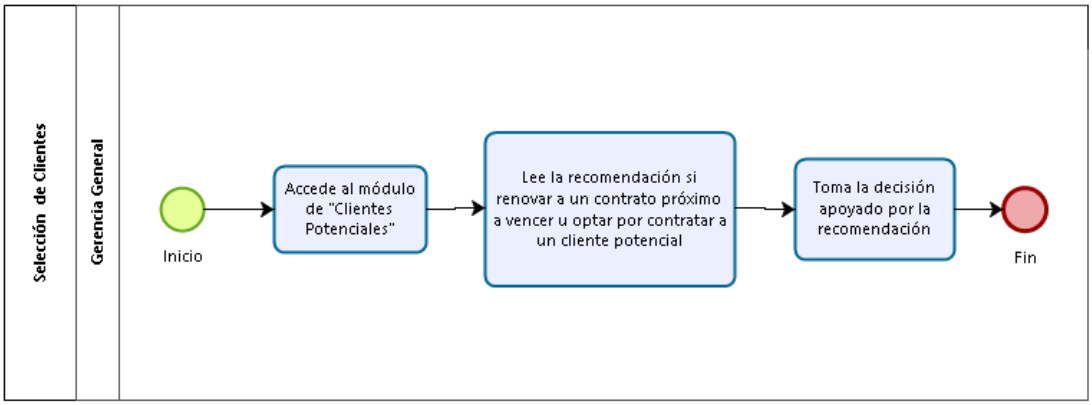


Figura 7. Mejora propuesta del proceso

Elaborado por: los autores

4.1.2 Determinar los objetivos de DM

El objetivo de DM es generar proyecciones de venta para cada locatario.

4.1.3 Criterios de aceptación para DM

Se tiene como criterio de aceptación que las proyecciones tengan un nivel de acierto mayor o igual al 70%.

4.1.4 Evaluación inicial de las herramientas

A continuación, se mostrará un cuadro comparativo evaluando que lenguaje es el mejor para el propósito de la investigación.

Tabla 7

Comparación de las características entre R y Python

	R	**Python**
Objetivo	Usado principalmente para modelos estadísticos	Utilizado para varios propósitos como la programación y la ciencia de los datos
Integración	Complejidad al utilizarse dentro de un sistema	Fácil implementación en sistemas gracias a su enfoque orientado a la programación
Capacidad de manejo de base de datos	Genera problemas con el manejo de grandes volúmenes de datos	Puede manejar muchos datos sin tener problemas

Elaborado por: los autores

Bajo estos puntos, se puede determinar que la tecnología que se utilizará será el lenguaje de programación Python gracias a su gran versatilidad, fácil integración y mejor manejo de los datos.

4.2 Comprensión de los datos

4.2.1 Recolectar los datos iniciales

Los datos que se utilizarán en modelo han sido extraídos de la base de datos del sistema usado actualmente por el Centro Comercial, en este se encuentran las ventas de los locatarios de forma diaria, así

como los rubros, contratos, la lista de locatarios, fechas en que se realizó la venta, entre otros, obteniendo un total de 322 locatarios con un total aproximado de 149 mil registros de ventas.

4.2.2 Descripción de los datos

De los datos extraídos, se utilizarán los siguientes campos:

- MontoSinIgv: Es el monto de la venta registrada por el locatario en la fecha indicada, siendo de tipo Numeric (16,2) y teniendo como valor por defecto Null.
- Idlocatario: Es el identificador único del locatario en el sistema, siendo de tipo Int y teniendo como valor por defecto Null.
- CodigoContrato: Es el código que se le asignó al contrato cuando ingresó al Centro Comercial, siendo de tipo Varchar(50) y teniendo como valor por defecto Null.
- FechaVenta: Indica la fecha de venta en la cual se realizó el registro, siendo de tipo Datetime y teniendo como valor por defecto Null.
- Rubro: Es el código que asigna un giro de negocio al locatario, siendo de tipo Int y teniendo como valor por defecto Null.

4.2.3 Verificación de la calidad de los datos

Para que un dato de ventas se considere erróneo debe cumplir con las siguientes condiciones:

a) Valores perdidos
   - Que la venta del día del locatario este vacía
   - Que no se reconozca a que locatario pertenece la venta registrada
   - Que no se tenga la fecha de venta del registro

b) Valores atípicos
   - Valores demasiado alejados de la realidad de ventas de un locatario

4.3 Preparación de los datos

En esta fase se estructuran los datos obtenidos mediante técnicas que permitirán usarlos en el proceso DM.

4.3.1 Seleccionar datos

Se creó un dataset con los datos proporcionados y se escogieron las variables que se usarán para el modelo, las cuales fueron:

- CodigoContrato
- FechaVenta
- IdLocatario
- Rubro_serie

Luego, se agregaron 2 variables de fechas en las que pueden generarse mayor número de ventas:

- Días_Pago (días más comunes en los cuales se realizan los pagos por trabajos o remuneraciones)
- Feriado

4.3.2 Limpiar los datos

Al recibir los datos de las ventas de los contratos se aplica una limpieza de los registros con el fin de mejorar la calidad de la data que se utilizará en el sistema. Para esto se realizaron las siguientes tareas.

- Limpieza de valores perdidos: No se encontraron ventas diarias con valores “null” ó 0.
- Limpieza de valores atípicos: Se encontró la presencia de valores atípicos para las ventas de 68 locatarios, la identificación se hizo en base al z score que es una medida de lejanía con respecto a la media. A continuación, se mostrará el ejemplo de uno de los locatarios identificado con el código 2399 donde se realizó dicha limpieza.

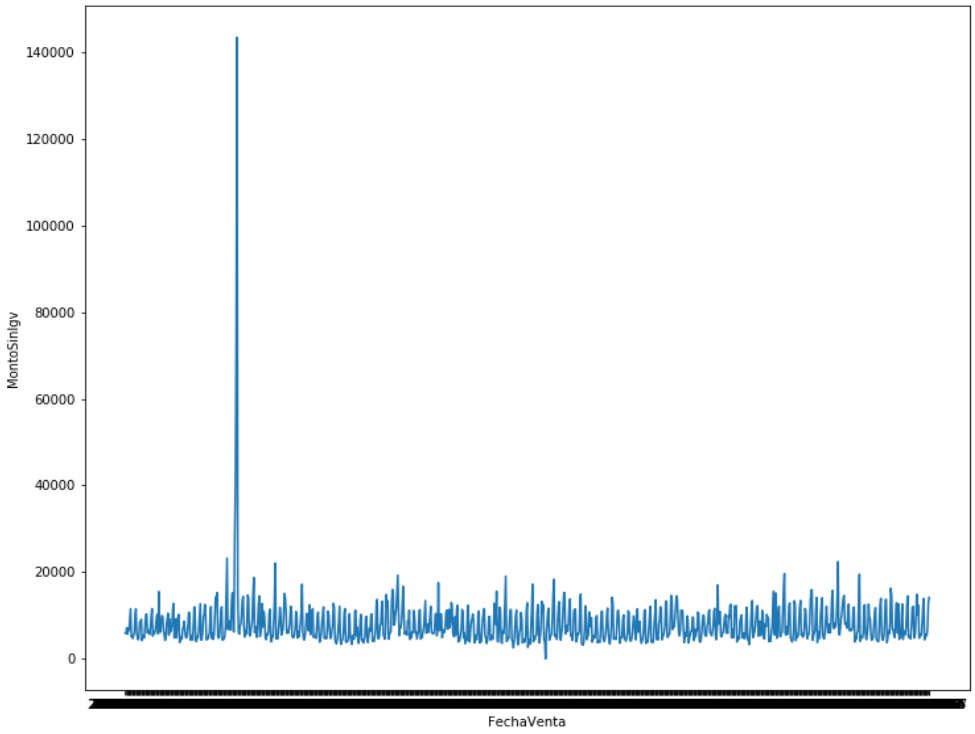


Figura 8. Valores atípicos del locatario 2399

Elaborado por: los autores

En este gráfico observamos un pico por encima de 140000 soles en ventas, el cual está muy alejado de sus ventas normales. esto se debió a que este locatario ingreso un acumulado de sus ventas por mes.

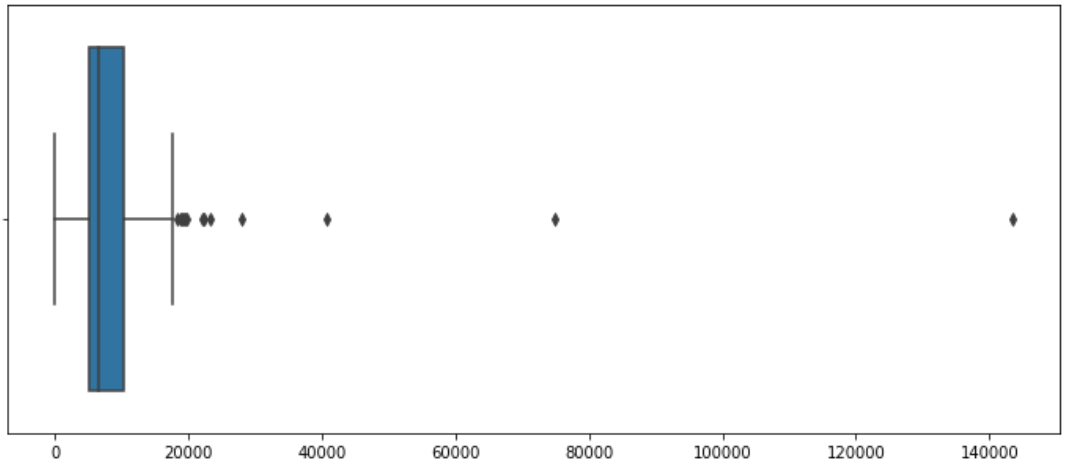


Figura 9. Valores de venta del locatario 2399

Elaborado por: los autores

Este registro fue eliminado porque no se conocía el monto generado en ventas y solo perjudicaba el pronóstico real de las ventas.

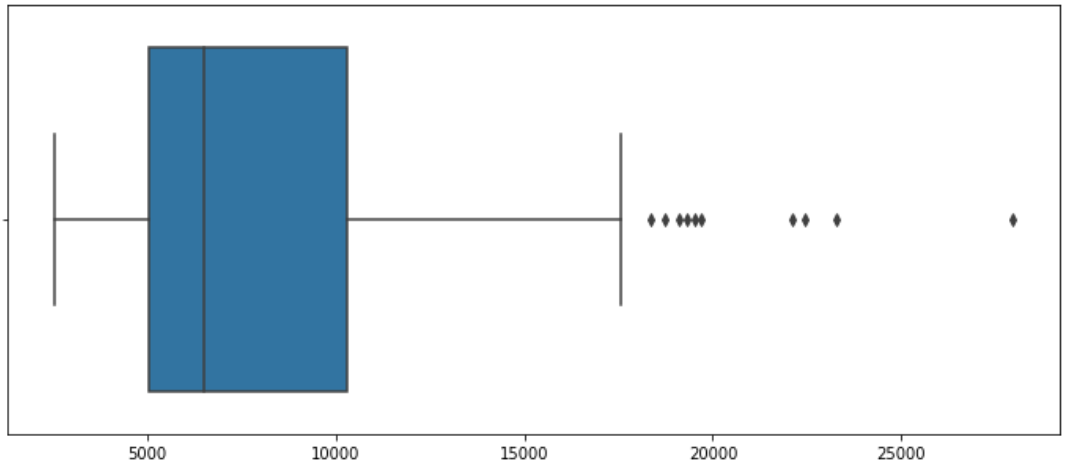


Figura 10. Valores de venta del locatario 2399 después de la limpieza

Elaborado por: los autores

Para la limpieza de los valores atípicos se utilizó una técnica de intercuartiles, en donde lo valores mayores al percentil 75 no se incluían en el dataset de entrenamiento para la preparación del modelo.

4.3.3 Estructurar los datos

Para que los datos ingresen al modelo, se deben de ordenar previamente a su uso. El campo de fecha se dividió en “Anio”, “Mes” y “Día” para su mejor uso en el modelo.

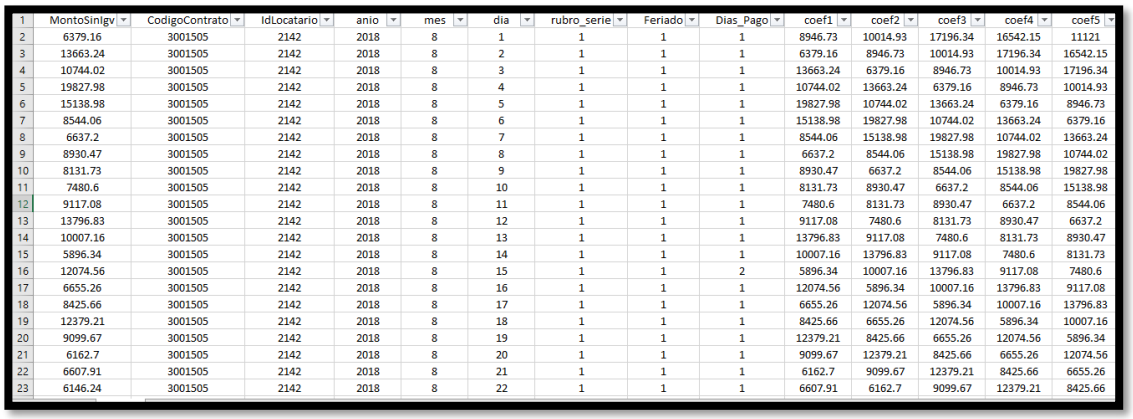
Como aporte de la investigación, se añadieron 5 variables de series de tiempo que representan las ventas de 5 días anteriores a la fecha registrada. Estas variables se ingresaron debido a que reflejan cuál es la tendencia de ventas de los últimos días y el distinto nivel de ventas por temporadas de cada locatario.

- Coef1 (venta de un día anterior al registro de venta)
- Coef2 (venta de dos días anteriores al registro de venta)
- Coef3 (venta de tres días anteriores al registro de venta)
- Coef4 (venta de cuatro días anteriores al registro de venta)
- Coef5 (venta de cinco días anteriores al registro de venta)

Finalmente, se logró elaborar un dataset que permita almacenar tanto la data obtenida de la base de datos como la propia agregada para poder entrenar con más variables el modelo. La estructura final del dataset está compuesta por los siguientes campos:

- MontoSinIgv
- CodigoContrato
- IdLocatario
- Anio
- Mes
- Dia
- Rubro_serie
- Feriado

- Días_Pago
- Coef1
- Coef2
- Coef3
- Coef4
- Coef5



	MontoSinIgv	CodigoContrato	IdLocatario	anio	mes	dia	rubro_serie	Feriado	Dias_Pago	coef1	coef2	coef3	coef4	coef5
2	6379.16	3001505	2142	2018	8	1	1	1	1	8946.73	10014.93	17196.34	16542.15	11121
3	13663.24	3001505	2142	2018	8	2	1	1	1	6379.16	8946.73	10014.93	17196.34	16542.15
4	10744.02	3001505	2142	2018	8	3	1	1	1	13663.24	6379.16	8946.73	10014.93	17196.34
5	19827.98	3001505	2142	2018	8	4	1	1	1	10744.02	13663.24	6379.16	8946.73	10014.93
6	15138.98	3001505	2142	2018	8	5	1	1	1	19827.98	10744.02	13663.24	6379.16	8946.73
7	8544.06	3001505	2142	2018	8	6	1	1	1	15138.98	19827.98	10744.02	13663.24	6379.16
8	6637.2	3001505	2142	2018	8	7	1	1	1	8544.06	15138.98	19827.98	10744.02	13663.24
9	8930.47	3001505	2142	2018	8	8	1	1	1	6637.2	8544.06	15138.98	19827.98	10744.02
10	8131.73	3001505	2142	2018	8	9	1	1	1	8930.47	6637.2	8544.06	15138.98	19827.98
11	7480.6	3001505	2142	2018	8	10	1	1	1	8131.73	8930.47	6637.2	8544.06	15138.98
12	9117.08	3001505	2142	2018	8	11	1	1	1	7480.6	8131.73	8930.47	6637.2	8544.06
13	13796.83	3001505	2142	2018	8	12	1	1	1	9117.08	7480.6	8131.73	8930.47	6637.2
14	10007.16	3001505	2142	2018	8	13	1	1	1	13796.83	9117.08	7480.6	8131.73	8930.47
15	5896.34	3001505	2142	2018	8	14	1	1	1	10007.16	13796.83	9117.08	7480.6	8131.73
16	12074.56	3001505	2142	2018	8	15	1	1	2	5896.34	10007.16	13796.83	9117.08	7480.6
17	6655.26	3001505	2142	2018	8	16	1	1	1	12074.56	5896.34	10007.16	13796.83	9117.08
18	8425.66	3001505	2142	2018	8	17	1	1	1	6655.26	12074.56	5896.34	10007.16	13796.83
19	12379.21	3001505	2142	2018	8	18	1	1	1	8425.66	6655.26	12074.56	5896.34	10007.16
20	9099.67	3001505	2142	2018	8	19	1	1	1	12379.21	8425.66	6655.26	12074.56	5896.34
21	6162.7	3001505	2142	2018	8	20	1	1	1	9099.67	12379.21	8425.66	6655.26	12074.56
22	6607.91	3001505	2142	2018	8	21	1	1	1	6162.7	9099.67	12379.21	8425.66	6655.26
23	6146.24	3001505	2142	2018	8	22	1	1	1	6607.91	6162.7	9099.67	12379.21	8425.66

Figura 11. Dataset mostrado desde una hoja Excel

Elaborado por: los autores

4.4 Modelado

4.4.1 Selección de la técnica de modelado

Para el proyecto se determinó por utilizar la técnica de árboles de decisión llamado XGBoost.

4.4.2 Generación del plan de prueba

La técnica de evaluación del modelo será la de validación simple, que consiste en dividir los datos históricos desde enero de 2017 hasta diciembre del 2018, en 80% para el entrenamiento y el 20% restante como pruebas, teniendo en cuenta que se deberá de detener el entrenamiento cuando se alcance un mínimo de error o ya no se vea una diferencia de error considerable entre iteraciones. La razón de realizar la validación según esta

proporción es que para el presente proyecto es importante mantener el orden consecutivo de las fechas de las ventas registradas, asignando de esta forma la proporción de entrenamiento a las primeras ventas registradas y para las pruebas, las últimas.

4.4.3 Construcción del modelo

A continuación, se muestra parte del uso del XGBoost en el código de Python para la creación de los hiperparámetros, así como la división de la data en entrenamiento (“train”) y pruebas (“valid”).

```
# use DMatrix for xgbosot
dtrain = xgb.DMatrix(train_dataset[every_column_except_y2], label=train_dataset['MontoSinIgv'])
dtest = xgb.DMatrix(test_dataset[every_column_except_y2], label=test_dataset['MontoSinIgv'])

watchlist = [(dtrain, 'train'), (dtest, 'valid')]

param2 = {
    'eta': 0.4,  # the training step for each iteration
    'silent': 0,  # logging mode - quiet
    'colsample_bytree':0.7,
    'nthread': 6 ,
    'max_depth':5,#6
    'nrounds':100,
    'subsample':0.7,
    'gamma': 0.3,
    'seed':42,
    'early_stopping_rounds':10
}
```

Figura 12. Extracto de código, data e hiperparámetros

Elaborado por: los autores

Los parámetros que se utilizan para el XGBoost están detallados a continuación:

- Eta: Permite disminuir la probabilidad de caer en overfitting reduciendo la importancia del árbol anterior, posee un rango de 0 a 1.
- Silent: Se usa 0 si se desea ver mensajes de ejecución y 1 si no se desea mostrar, su valor predeterminado es 0.

- Colsample_bytree: Proporción de la submuestra a usar en el siguiente árbol, tiene un rango de 0 a 1, tiene un valor predeterminado de 1.
- Nthread: Es el número de subprocesos paralelos utilizados para ejecutar XGBoost y tiene como valor predeterminado el máximo número de procesos que soporta la computadora.
- Max_depth: Es la profundidad máxima de un árbol. El aumento de este valor hace que el modelo sea más complejo pueda caer en sobreajuste, puede tener un valor de 0 en adelante siendo 6 su valor por defecto.
- Nrounds: Es el número de árboles que se utilizarán en XGBoost, tiene como valor defecto 100.
- Subsample: Es la proporción de cantidad de instancias de datos que recopilará de forma aleatoria evitando el sobreajuste, puede ser un valor entre 0 a 1 y tiene un valor por defecto de 1.
- Gamma: Es el valor de pérdida mínima permitida para crear una partición en un nodo, puede tener un número de 0 en adelante siendo esta su valor por defecto.
- Seed: Se define como el número aleatorio de reproductibilidad, su valor por defecto es 0.
- Early_stopping_rounds: es el número de límite de iteraciones permitido para reducir el error de validación, este no tiene valor predeterminado.

4.4.4 Evaluación del modelo

A continuación, se muestra un ejemplo del locatario con código 2407 del proceso de entrenamiento y prueba a través de cada iteración.

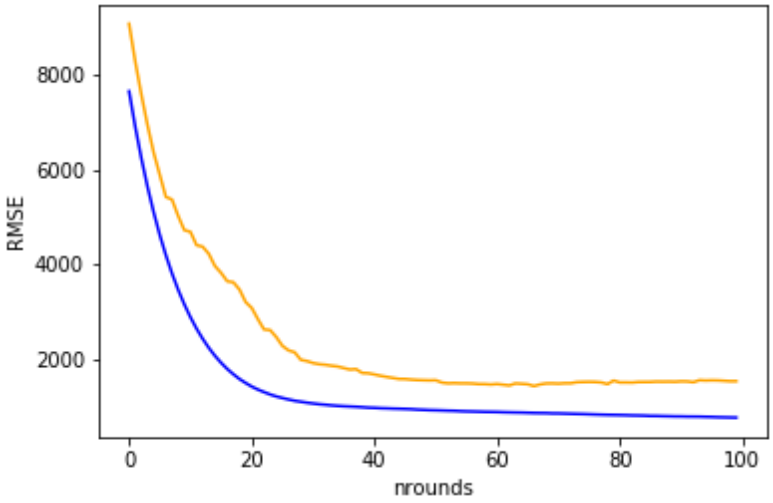


Figura 13. Entrenamiento y prueba del locatario 2407

Elaborado por: los autores

nrounds	train_score	test_score
90	814.01398	1550.94922
91	810.80493	1549.04004
92	810.03503	1532.84766
93	808.90698	1575.04785
94	805.51440	1568.61035
95	802.57599	1572.40039
96	797.62549	1569.95117
97	795.23303	1562.28418
98	792.17712	1553.67969
99	789.67578	1554.45020

Figura 14. Disminución del margen de error en soles por iteración

Elaborado por: los autores

Se observa que el error cuadrático medio disminuye conforme avanzan las iteraciones, debemos tener en cuenta que mientras más bajo sea el valor del RMSE (error cuadrático medio), mejor será el ajuste, así también se observa que no hay presencia de overfitting pues las

curvas del entrenamiento y validación son distintas. Este proceso se realizó para cada locatario.

4.5 Evaluación

Los resultados obtenidos del uso del modelo como solución a la problemática del negocio han sido descritos en el capítulo “Resultados”.

4.6 Despliegue

Una vez que se tiene un modelo para cada locatario, es posible realizar las proyecciones de venta para el año 2019. Cabe recordar que siempre se recomienda reajustar el modelo con data más reciente, permitiendo realizar mejores proyecciones de ventas para los locatarios durante el año 2020.

Para esto se deberá de seguir las fases realizadas durante la investigación y poner atención a las mejoras sugeridas en la parte final de la investigación.

CAPÍTULO V
RESULTADOS

En este capítulo se mostrarán los resultados de la investigación y como estos cumplen con los objetivos trazados.

- Identificar los factores más importantes que influyen en las proyecciones de ventas

A continuación, se muestra el ranking de importancia de las variables en la proyección de ventas (feature importance). Esta permite conocer cuáles son las variables que contribuyen con la predicción del modelo.

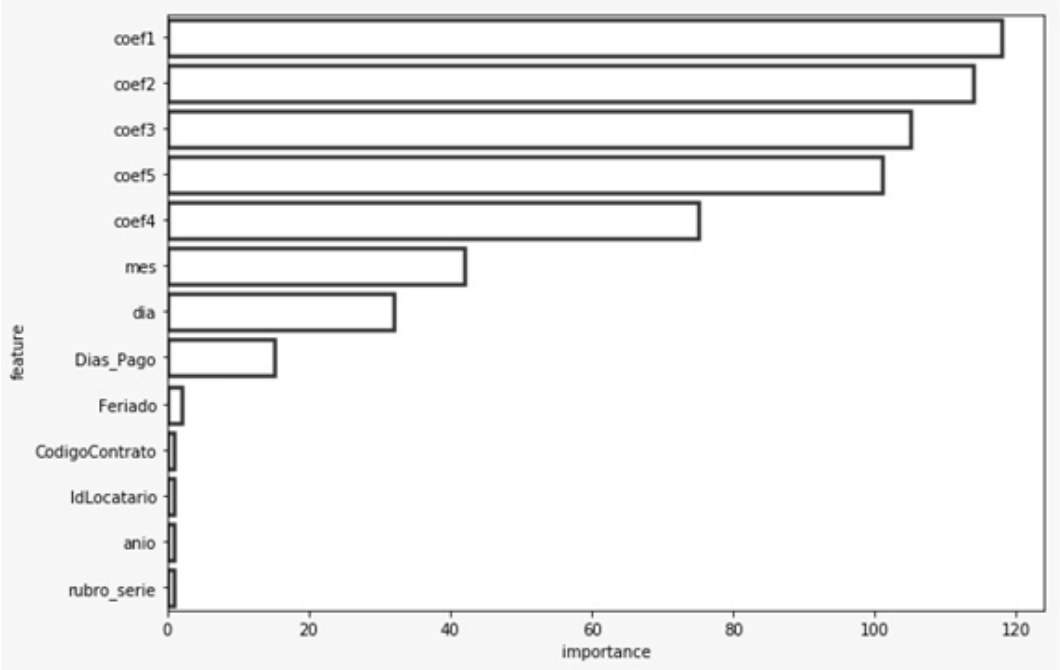


Figura 15. Importancia de cada variable dentro del árbol

Elaborado por: los autores

Encontradas las variables que contribuyen con las proyecciones de ventas y habiendo analizado que los resultados obtenidos superaron el 70% de acierto (comparación de las proyecciones de ventas contra las ventas obtenidas de los locales actuales), se procedió a asignar ventas proyectadas basadas en un promedio por rubro a los clientes potenciales que se encuentran a la espera de una respuesta por parte del Centro Comercial, generando así la recomendación de los mejores clientes potenciales, en otras palabras, se recomendará a aquellos que generen mejores resultados de ventas en los próximos tres meses de entre la cartera de clientes potenciales que tiene el Centro Comercial y los locales que actualmente arriendan un espacio.

- Generar las proyecciones de ventas para los locatarios y clientes potenciales

Como parte de la evaluación de los resultados obtenidos se realiza la comparación de las ventas obtenidas en el mes de enero, febrero y marzo del año 2019; con las proyecciones obtenidas de los mismos meses (en semanas), se observa que el porcentaje de acierto de las proyecciones superan el 70%, y siendo el valor mínimo observado 74%.

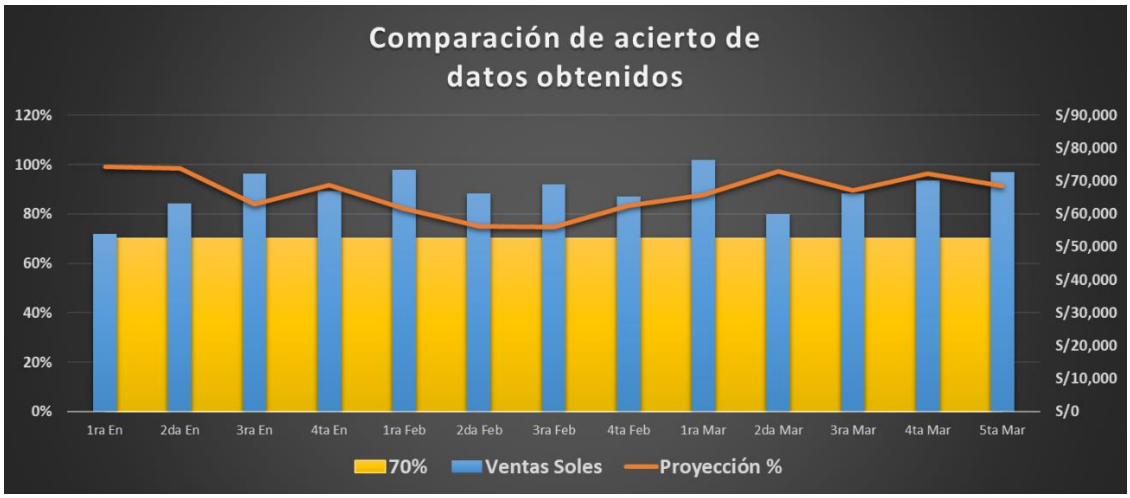


Figura 16. Validación de criterio de aceptación locatario 2399

Elaborado por: los autores

El área en color amarillo representa el 70% de criterio de aceptación, las barras en color azul son las ventas obtenidas en semanas en los meses de enero, febrero y marzo de 2019, así también se observa la proyección obtenida en porcentaje en los periodos de las ventas mostradas en la línea de color naranja. De manera sencilla se puede observar que las proyecciones superan el 70% de criterio, siendo su punto más bajo la tercera semana de febrero (74.58%).

A continuación, se evalúa el caso de un locatario cuyo contrato fue renovado en diciembre 2018 ya que las proyecciones planteadas por Gerencia

General indicaban un crecimiento en sus ventas, sin embargo, los montos obtenidos por el locatario en el periodo de enero a marzo del 2019 no llegaron a cumplir el 100% de lo proyectado, siendo su punto más bajo de 71.93% y el más alto del 97%. Estos resultados se traducen en pérdidas, llegando a alcanzar el monto de 11449 soles acumulado en el periodo evaluado, el detalle se muestra en la siguiente tabla:

Tabla 8

Caso locatario 2788 pérdidas por metas establecidas

	VENTAS	PROYECCIONES		COMPARACIÓN	
Semanas	Ventas reales	Machine Learning	Gerencia General	Ventas vs Proyección G.G.	Pérdida
1ra Ene	S/6,114.4	S/4,615.4	S/8,500.0	71.93%	S/2,385.6
2da Ene	S/2,786.7	S/3,038.1	S/3,300.0	84.45%	S/513.3
3ra Ene	S/3,633.0	S/4,157.6	S/4,600.0	78.98%	S/967.0
4ta Ene	S/5,399.4	S/4,623.2	S/6,800.0	79.40%	S/1,400.6
1ra Feb	S/5,515.9	S/4,538.3	S/6,300.0	87.55%	S/784.1
2da Feb	S/4,443.2	S/4,340.3	S/5,400.0	82.28%	S/956.8
3ra Feb	S/5,032.2	S/5,535.0	S/5,500.0	91.49%	S/467.9
4ta Feb	S/3,853.4	S/4,141.1	S/4,200.0	91.75%	S/346.6
1ra Mar	S/4,293.2	S/4,520.7	S/4,900.0	87.62%	S/606.8
2da Mar	S/5,335.2	S/4,461.8	S/5,500.0	97.00%	S/164.8
3ra Mar	S/3,429.6	S/3,805.0	S/4,500.0	76.21%	S/1,070.4
4ta Mar	S/5,271.2	S/4,989.2	S/5,800.0	90.88%	S/528.8
5ta Mar	S/3,243.7	S/3,105.0	S/4,500.0	72.08%	S/1,256.3
TOTAL	S/58,351.0	S/55,870.6	S/69,800.0	83.60%	S/11,449.0

Elaborado por: los autores

Se puede observar las ventas reales obtenidas por el locatario, los montos proyectados por el Machine Learning y las proyecciones de Gerencia General. Se realizó la comparación de las ventas reales contra las proyecciones de Gerencia General, observando así los resultados mencionados anteriormente (menor a 100%), se calcula la diferencia de estos en soles, para así obtener los montos traducidos en pérdidas de alcance a las metas establecidas, alcanzando un total de 11,449 soles.

Así mismo, se muestran los datos de las proyecciones realizadas por Gerencia General y las obtenidas usando Machine Learning, comparándolas con las ventas obtenidas en el periodo mencionado anteriormente, observando así que las proyecciones usando Machine Learning como total tienen una diferencia de valor en absoluto de 4.25%, mientras que las de Gerencia General en un 19.62%, siendo así los valores más acertados los de Machine Learning para el caso evaluado, de haber contado con la herramienta el Centro Comercial habría podido evaluar otras opciones con mejores resultados en la toma de decisiones en el proceso de alquiler de espacios.

Tabla 9

Caso locatario 2788 ventas reales vs proyección

Semanas	VENTAS	PROYECCIÓN	
	Ventas reales	Machine Learning	Gerencia General
1ra En	S/6,114.4	75.48%	139.02%
2da En	S/2,786.7	109.02%	118.42%
3ra En	S/3,633.0	114.44%	126.62%
4ta En	S/5,399.4	85.63%	125.94%
1ra Feb	S/5,515.9	82.28%	114.21%
2da Feb	S/4,443.2	97.68%	121.53%
3ra Feb	S/5,032.2	109.99%	109.30%

4ta Feb	S/3,853.4	107.47%	109.00%
1ra Mar	S/4,293.2	105.30%	114.13%
2da Mar	S/5,335.2	83.63%	103.09%
3ra Mar	S/3,429.6	110.94%	131.21%
4ta Mar	S/5,271.2	94.65%	110.03%
5ta Mar	S/3,243.7	95.72%	138.73%
TOTAL	S/58,351.0	95.75%	119.62%

Elaborado por: los autores

CAPÍTULO VI
DISCUSIÓN

Luego de haber realizado la investigación en el área de Machine Learning, se ha logrado aprender y conocer sobre su uso y eficacia en el sector retail, específicamente en centros comerciales. Como se mencionó en las bases teóricas, la utilización de Machine Learning para generar un modelo predictivo y así pronosticar ventas futuras ha sido de gran utilidad, llegando así generar proyecciones de ventas con un acierto mayor al 70%, estos resultados permitirán tomar mejores decisiones entre los clientes potenciales que solicitan su ingreso al Centro Comercial. Para este caso se optó por un modelo basado en algoritmos de árboles de decisión llamado XGBoost, que puede llegar a ser mejor que el Random Forest el cual esta explicado en las bases teóricas, debido a que los modelos subsecuentes que genera aprenden de los errores realizados por los otros submodelos, sin necesidad de expandir los árboles que tienen mayor probabilidad de error. Además, la selección de las variables tomadas en los antecedentes de investigación sirvió de guía para nuestro modelo, pero a su vez, en nuestro análisis consideramos la utilización de variables diferentes como los últimos 5 días de venta de los locatarios representando cada una como una variable para el modelo, aplicando así variables de series de tiempo por medio del feature engineering.

CONCLUSIONES

1) Sí se logró formular un sistema de recomendaciones que permita identificar a los mejores clientes que van a alquilar los locales comerciales. Se consiguió identificar a las empresas locatarias, actuales y postulantes, que tendrán más posicionamiento en el mercado a corto plazo, es decir las que crean y crearán mayor fidelización con los clientes y por ende mayores ventas, esto generó certeza a los administradores del Centro Comercial sobre sus futuros contratos de arrendamiento. El logro de este objetivo se basa en el empleo de Machine Learning, que contó para su proyección con el histórico de ventas desde enero del 2017 hasta marzo del 2019.

2) Se observó un alto nivel de importancia en las variables de series de tiempo creadas, estas son las ventas de los cinco días previos del día a evaluar y así identificar patrones en el flujo de ventas de cada locatario.

3) Se ha logrado generar proyecciones de venta por locatario con un nivel de error menor del 30%, tomando de 2 a 3 horas para que el modelo genere las proyecciones de ventas, consecuente a esto se podrán realizar consultas de forma automática de los datos generados.

RECOMENDACIONES

1) En un mercado tan competitivo como el retail, conocer el futuro para realizar una planeación óptima es una necesidad, en el proyecto se tomaron datos de venta de clientes internos; sin embargo, contar con información de fuentes externas mejorarán las recomendaciones hacia la identificación de clientes potenciales con mayores ingresos, por ejemplo la venta de rubros no existentes en la data entrenada, venta por rubros de los principales competidores, demanda por niveles socioeconómicos, PBI, inflación, entre otros.

2) Si bien se obtuvo la mayor importancia en las variables de series de tiempo (venta de los cinco días previos al día evaluado), se debería enriquecer el modelo con las últimas ventas de los locales además de nuevas variables para volver a entrenar el modelo de Machine Learning y mejorar las proyecciones a futuro, debido a que probablemente existan tendencias de ventas que no se encuentran reflejadas en la data entrenada en el modelo.

3) Si se dispone de un mayor tiempo y más recursos destinados a la investigación se lograrían obtener mejores proyecciones usando variables generadas por el mismo Centro Comercial, por ejemplo, la ubicación del local

en el Centro Comercial, cantidad de competidores, publicidad, nivel de atención al cliente, limpieza y orden, entre otros.

FUENTES DE INFORMACIÓN

Alon, I., Qi, M. y Sadowski, R. (2001). Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services*, 8, 147-148. doi: 10.1016/S0969-6989(00)00011-4

Amazon. (2016). Amazon Machine Learning: Guía para desarrolladores. Recuperado de: https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/machinelearning-dg.pdf

Asociación Española de Centros Comerciales y Parques Comerciales. (2016). *Terminología.* Recuperado de http://directoriocentroscomerciales.aedecc.com/terminologia

Barnes, J. (2015). *Azure Machine Learning. Microsoft Azure Essentials*. Recuperado de https://download.microsoft.com/download/0/9/6/096170E9-23A2-4DA6-89F5-7F5079CB53AB/9780735698178.pdf

Bost, R., Popa, R. A., Tu, S., Goldwasser, S. (2015). Machine Learning classification over encrypted data. *NDSS*, 15, 1. doi: 10.14722/ndss.2015.23241

Breiman, L. (2001). Random Forests. *Springer*. 45, 7. doi: 10.1023/A:1010933404324

Calderón, S. (2017). *Predicción de resultados deportivos con técnicas de Machine Learning aplicado al fútbol*. (Tesis de grado). Universidad Carlos III de Madrid, Madrid, España.

Carreño, A. (2017). *Detección de sucesos raros con Machine Learning*. (Tesis de maestría). Universidad Politécnica de Madrid, Madrid, España.

Chen, T., Guestrin, C. (2016). XGboost: A scalable tree boosting system. *Association for Computing Machinery*, 185. doi: 10.1145/2939672.2939785

De la Flor, F. (9 de diciembre de 2011). Tiendas con gancho. *El Comercio.* Recuperado de https://www.esan.edu.pe/sala-de-prensa/2011/12/09/el_comercio_eric_lopez_12_02_final.pdf

De Pablos, C., López-Hermoso, J., Martín-Romo, S. y Medina, S. (2004). *INFORMÁTICA Y COMUNICACIONES EN LA EMPRESA*. Recuperado de https://books.google.com.pe/books?hl=es&lr=&id=U0MXWtqjxtsC&oi=fnd&pg=PA9&dq=informatica&ots=D9bkKN4y8A&sig=9riPJaxJxjY60cMMk7R-DCecUns&redir_esc=y#v=onepage&q&f=false

Evans, M. (2003). Practical Business Forecasting. Recuperado de http://www.untag-smd.ac.id/files/Perpustakaan_Digital_1/BUSINESS%20Practical%20business%20forecasting.pdf

García, G. (2014). *Modelo de Machine Learning para la clasificación de pacientes en términos del nivel asistencial requerido en una urgencia pediátrica con Área de Cuidados Mínimos*. (Tesis de maestría) Universidad Tecnológica de Bolívar, Cartagena, Colombia.

García, I. (2017). Definición de Proyección de ventas [Información referente a la proyección de ventas]. Recuperado de https://www.economiasimple.net/glosario/proyeccion-de-ventas.

Hernández, R., Fernández, C. y Baptista, M. (2010). Metodología de la investigación. Recuperado de https://www.esup.edu.pe/descargas/dep_investigacion/Metodologia%20de%20la%20investigaci%C3%B3n%205ta%20Edici%C3%B3n.pdf

Hitt, L. (2011). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?. *MIT Initiative on the digital economy,* 1. Recuperado de http://ide.mit.edu/sites/default/files/publications/2011.12_Brynjolfsson_Hitt_Kim_Strength%20in%20Numbers_302.pdf

Hoshmand, A. (2010). Business Forecasting: A Practical Approach. Recuperado de http://pith-edu.weebly.com/uploads/7/1/8/0/71808217/business_forecasting_-_a_practical_approach_by_a._reza_hoshmand_2e.pdf

Kreplak, G. (2018). *Predicción de Ventas Comestibles Corporación Favorita*. (tesis de maestría). Universitat Oberta de Catalunya, Barcelona, España.

López Pastor, E. (9 de diciembre de 2011). Tiendas con gancho. *El Comercio.* Recuperado de https://www.esan.edu.pe/sala-de-prensa/2011/12/09/el_comercio_eric_lopez_12_02_final.pdf

ManagementSolutions. (2018). *Machine Learning, una pieza clave en la transformación de los modelos de negocio*. Recuperado de https://www.managementsolutions.com/sites/default/files/publicaciones/esp/Machine-Learning.pdf

Martinez, J. (2018). *Regresión Logística para Clasificación*. España: IArtificial. Recuperado de https://iartificial.net/regresion-logistica-para-clasificacion

McLeod, S. (2019). *Z-score: definition, calculation and interpretation.* Simply Psychology. Recuperado de https://www.simplypsychology.org/z-score.html

Natekin, A., Knoll, A. (2013). Gradient boosting Machines, a tutorial. *Frontiers in neurorobotics*. 7(21), 1. doi: 10.3389/fnbot.2013.00021

Niño, M. A., Cobos, C. A., Mendoza, M. E. (2001). *Introducción a la Informática: Algoritmos y Programación en Lenguaje C.* Recuperado de https://www.researchgate.net/profile/Miguel_Nino_Zambrano/publication/251740349_Introduccion_a_la_Informatica_Algoritmos_y_Porgramacion_en_Lenguaje_C/links/0046352b872ad8ddfa000000.pdf

Pallares, F. (2014). *Desarrollo de un modelo basado en Machine Learning para la predicción de la demanda de habitaciones y ocupación en el sector hotelero*. (Tesis de maestría). Universidad Tecnológica de Bolívar, Cartagena, Colombia.

Perez, A. (2019). *Trabajando con Python: Machine Learning y la Regresión*. Medium. Recuperado de https://medium.com/qu4nt/el-Machine-Learning-y-la-regresión-python-c94db2968a4e

Pérez, J. y Merino M. (2010). *Definición de rubro*. Definición.de. Recuperado de https://definicion.de/rubro/

Pollak, H. O. (2011). What is Mathematical Modeling?. *Journal of Mathematics Education at Teachers College*, 2, 64. Recuperado de https://journals.library.columbia.edu/index.php/jmetc/article/download/694/140

Provost, F., Fawcett, T. (2013). Data Science for Business. Recuperado de https://www.oreilly.com/library/view/data-science-for/9781449374273/

Real Academia Española. (2014). *Diccionario de la Lengua Española*. España. Recuperado de https://dle.rae.es

Román, V. (2019). *Machine Learning: Cómo desarrollar un Modelo desde Cero*. Recuperado de https://medium.com/datos-y-ciencia/Machine-Learning-cómo-desarrollar-un-modelo-desde-cero-cc17654f0d48

Schaeffer, E. (2009). Computación aleatorizada-probabilidad y algoritmos. *RISCE-Revista Internacional de Sistemas Computacionales y Eléctrónicos*, 2. Recuperado de https://repositoriodigital.ipn.mx/bitstream/123456789/8068/1/Risce_n0.pdf#page=8

Schaeffer, J., van den Herik, H. J. (2002). Games, computers, and artificial intelligence. *Elsevier*, 1. doi: 10.1016/S0004-3702(01)00165-5

Song, Y., Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2), 131. doi: 10.11919/j.issn.1002-0829.215044

Suárez, M. O. y Tapia, F. A. (2012). *Interaprendizaje de Estadística Básica.* Recuperado de http://repositorio.utn.edu.ec/handle/123456789/2341

Turienzo, L. (2017). ¿QUÉ ES RETAIL?. España: *RetailNewsTrends*. Recuperado de https://retailnewstrends.me/que-es-el-retail-2/

Turing, A. M. (1950). Computing Machinery and intelligence. *Mind*, 59(236), 433-435. doi: 10.1093/mind/LIX.236.433

United Nations. (2000). *Handbook on geographic information systems and digital mapping* (79). Recuperado de https://unstats.un.org/unsd/publication/SeriesF/SeriesF_79E.pdf

Valdivia, J. (2015). Comercialización de productos y servicios en pequeños negocios o microempresas. Recuperado de https://play.google.com/store/books/details/Juan_Alfonso_Valdivia_Garc%C3%ADa_Comercializaci%C3%B3n_de_p?id=y8LIBgAAQBAJ

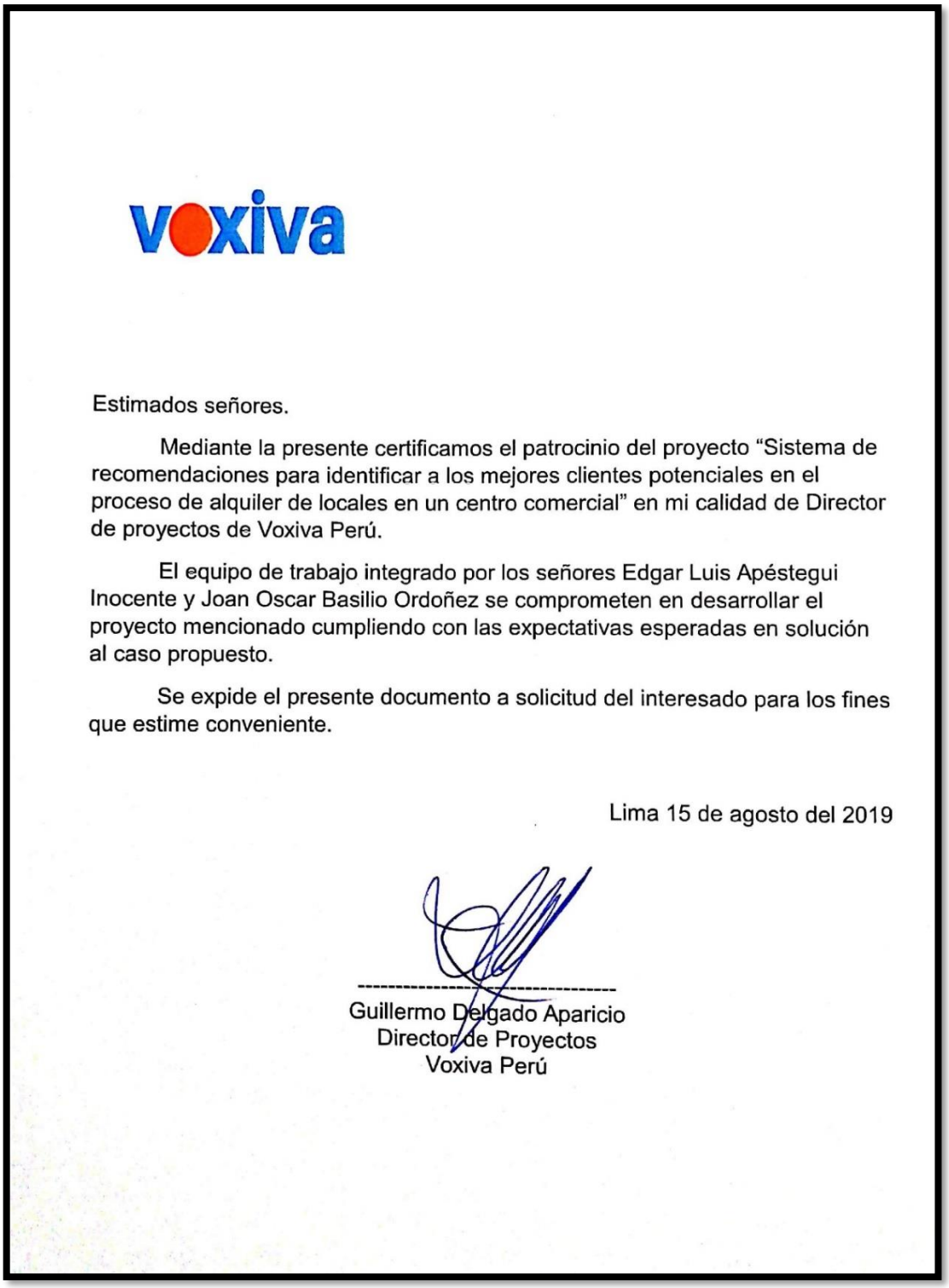
Wirth, R., Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining*. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 29-39. doi: 10.1.1.198.5133

Zheng, A., Casari, A. (2018). *Feature engineering for Machine Learning: principles and techniques for data scientists*. Recuperado de https://www.repath.in/gallery/feature_engineering_for_machine_learning.pdf

ÍNDICE DE ANEXOS

ANEXOS

Anexo 1 Acta de patrocinio



voxiva

Estimados señores.

Mediante la presente certificamos el patrocinio del proyecto "Sistema de recomendaciones para identificar a los mejores clientes potenciales en el proceso de alquiler de locales en un centro comercial" en mi calidad de Director de proyectos de Voxiva Perú.

El equipo de trabajo integrado por los señores Edgar Luis Apéstegui Inocente y Joan Oscar Basilio Ordoñez se comprometen en desarrollar el proyecto mencionado cumpliendo con las expectativas esperadas en solución al caso propuesto.

Se expide el presente documento a solicitud del interesado para los fines que estime conveniente.

Lima 15 de agosto del 2019

Guillermo Delgado Aparicio
Director de Proyectos
Voxiva Perú

Anexo 2 Acta de conformidad de pruebas

Acta de aceptación de pruebas de usuario

Certificado de aceptación de pruebas de usuario.

Proyecto	**SISTEMA DE RECOMENDACIONES PARA IDENTIFICAR A LOS MEJORES CLIENTES POTENCIALES EN EL PROCESO DE ALQUILER DE LOCALES EN UN CENTRO COMERCIAL**	
Fecha	06/11/2019	
Estado	Aceptado	✓
	Aceptado en condición de mejora	
	No aceptado	

Nombre/Cargo	**Firma**
Delgado Aparicio, Guillermo Product Sponsor Product Owner	
Basilio Ordoñez, Joan Product Manager Test Team	
Apestegui Inocente, Edgar Tech Team Test Team	

Anexo 3 Acta de aceptación de proyecto

Acta de aceptación del proyecto

Proyecto
SISTEMA DE RECOMENDACIONES PARA IDENTIFICAR A LOS MEJORES CLIENTES POTENCIALES EN EL PROCESO DE ALQUILER DE LOCALES EN UN CENTRO COMERCIAL
Nombre del Sponsor
Guillermo Delgado Aparicio

DECLARACIÓN DE LA ACEPTACIÓN FORMAL
Por la presente se deja constancia de que el Proyecto **"Sistema de recomendaciones para identificar a los mejores clientes potenciales en el proceso de alquiler de locales en un centro comercial"** ha sido aceptado y aprobado por el Sponsor del Proyecto Guillermo Delgado Aparicio, director de proyectos, por lo que concluye que el proyecto ha sido culminado exitosamente.
Tesistas
Edgar Luis Apestegui Inocente Joan Oscar Basilio Ordoñez
Módulos entregados
- Indicadores - Reporte ventas - Mapa de calor - Locatarios - Selección de potenciales clientes - Proyecciones
Observaciones adicionales
El proyecto ha sido desarrollado dentro de los tiempos planificados, siendo la fecha de término el 7 de noviembre de 2019.

Guillermo Delgado Aparicio
Director de Proyectos
Voxiva Perú