Vigilada Mineducación

# MODELAMIENTO PREDICTIVO DEL NÚMERO DE VISITANTES EN UN CENTRO COMERCIAL

Forecasting the number of shopping center visitors using predictive modeling

RAMON DAVID RUA JARAMILLO

Proyecto de grado

Asesores

Paula Maria Almonacid Hurtado

Henry Laniado Rodas

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS

MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA

MEDELLÍN

2022

**CONTENIDO**

## LISTA DE FIGURAS

## RESUMEN

La capacidad para realizar predicciones acerca del número de clientes o visitantes en un centro comercial es un insumo muy importante en la planeación y el uso eficiente de los recursos físicos y humanos en este tipo de empresas. Asimismo, es relevante entender cuáles son los factores que influyen en su comportamiento.

Basados en los datos históricos del número de visitantes, así como variables externas (ambientales) y tendencias de consulta en línea, se plantea un modelo predictivo del comportamiento de visitas diarias al centro comercial. Los datos históricos corresponden a los ingresos peatonales y vehiculares (carros y motos) de los últimos 6 años en un centro comercial ubicado en la ciudad de Medellín.

Este proyecto inicia con la revisión de literatura referente a modelos predictivos en diferentes lugares como museos, aeropuertos, parques naturales, centros comerciales y restaurantes, entre otros, con el fin de explorar metodologías en dichos casos y posibles opciones de solución.

Por medio de análisis de series de tiempo y algoritmos de aprendizaje automático, se seleccionan las variables más representativas y el modelo mejor ajustado para predecir el número de visitantes.

Se espera que este modelo se fortalezca con algoritmos de estimación, mejorando el rendimiento a lo largo del tiempo y permitiendo ser aplicado en otros entornos empresariales o educativos.



Palabras clave: predicción; visitantes; clientes; centro comercial; series de tiempo; aprendizaje automático.

## ABSTRACT

The ability to make predictions about the number of customers or visitors in a shopping center is a very important input in the planning and efficient use of physical and human resources in this type of company. Also, it is important to understand what aspects influences their behavior.

Based on historical data on the number of visitors, as well as external (environment) variables and online search trends, a forecasting model of the behavior of daily visits to the shopping center is suggested. The historical data correspond to the pedestrian and vehicular entries (cars and motorcycles) of the last 6 years in a shopping center located in the city of Medellín.

This project begins with a literature review regarding forecasting models in different places such as museums, airports, natural parks, shopping centers and restaurants, among others, in order to explore methodologies in such cases and possible solution options.

Through time series analysis and machine learning algorithms, the most representative variables and the best-fit model are selected to predict the number of visitors.

This model is expected to be strengthened with estimation algorithms, improving performance over time and allowing it to be applied in other business or educational environments.



Keywords: forecasting; visitors; customers; shopping center; time series; machine learning.

# INTRODUCCIÓN

Uno de los indicadores económicos relevantes para los países es el número de centros comerciales (Chebat, Sirgy, & Grzeskowiak, 2010), ya que estos constituyen uno de los eslabones importantes en el crecimiento del mercado y del consumo. Con el incremento de estos en las ciudades, la competencia entre ellos ha estado aumentando y como resultado, han empezado a diferenciar la oferta de servicios y el desarrollo de campañas para atraer a más clientes.

La administración de los centros comerciales ha trabajado a lo largo del tiempo en identificar factores que afecten la imagen, la ocupación de locales (marcas) y el número de visitantes (Chebat, Sirgy, & Grzeskowiak, 2010) . Desde la perspectiva estratégica, un centro comercial que pueda generar más tráfico (visitantes) es aquel que puede cobrar un mayor valor a los locales, ya que al tener más tráfico existe una mayor probabilidad de tener más ventas (Perdikaki, Kesavan S, & J., 2012).

Los centros comerciales son además el punto de encuentro de diversos sectores empresariales como recreación, comidas, comercio, servicios financieros, etc. En estos, cada marca tiene sus propias estrategias de atracción de clientes, pero el aporte de la administración de la propiedad horizontal ayuda a generar ambientes agradables y fomentar un mayor tráfico.

Desde el punto de vista de la operación de un centro comercial, el número de visitantes influye en el número personas asignadas en vigilancia, aseo y mantenimiento, que se requieren de acuerdo con la ocupación. De la misma manera, determinar el número de parqueaderos a habilitar está relacionado con el volumen de visitantes. Los eventos generadores de tráfico que la administración del centro comercial realiza, también le dan un gran impulso al número de visitantes.

Por lo anterior, generar un modelo predictivo que identifique oportunamente el número de visitantes al centro comercial, permitirá crear estrategias a nivel operativo y de mercadeo que mejoren la asignación y uso de los recursos físicos, humanos y económicos.

# OBJETIVOS

## GENERAL

Realizar un modelo predictivo que permita determinar el número de visitantes diarios a un centro comercial en la ciudad de Medellín.

## ESPECÍFICOS

- Realizar estadística descriptiva y analítica de los datos entregados por el centro comercial, correspondientes al número de visitantes diarios registrados por las cámaras de conteo (peatonal) y al número de vehículos por placa (carros y motos).

- Identificar las variables más representativas que afectan el comportamiento del número de visitantes al centro comercial, agregándolas al modelo e infiriendo sobre su impacto económico y estadístico

- Implementar métodos de series de tiempo y de aprendizaje automático para realizar la predicción de los visitantes diarios al centro comercial.

- Fortalecer la capacidad de predicción del modelo mediante la calibración del ajuste y precisión de los algoritmos de estimación implementados.

## REVISIÓN DE LITERATURA Y MARCO TEÓRICO

Históricamente las personas han podido satisfacer sus necesidades de compra en las tiendas y comercios de los alrededores a sus residencias o sitios de permanencia. El objetivo principal de los centros comerciales que se crean en las ciudades es atraer clientes y ofrecer una variedad de productos y servicios que cubran la demanda (Ozdemir, Cevik Onar, & Bagriyanik, 2020).

Para los centros comerciales, como en otros sectores, los clientes son de gran relevancia para la generación de ingresos, por lo cual atraerlos y retenerlos es un objetivo crucial para las marcas o comercios. Considerando diversas aplicaciones y modelos de predicción de visitantes, se muestran soluciones de ciencia de datos (y sus evaluaciones) para entender cuáles se pueden emplear en este trabajo de grado. En la revisión de literatura se identifican varios modelos, algoritmos y parámetros de predicción, desde clásicos como regresión lineal, hasta avanzados como aprendizaje profundo.

El desarrollo de este marco teórico considera las técnicas implementadas en diversos estudios (papers) para predecir el número de clientes, visitantes o turistas. A continuación, se resumen los temas abarcados en cada uno, los algoritmos implementados y las medidas de desempeño utilizadas, que serán la base para la implementación del trabajo de grado.

Una primera aproximación para estimar el número de visitantes consiste en obtener variables ambientales, como temperatura y lluvia, y entender cuáles afectan el número de visitantes a los centros comerciales de forma significativa (Ozdemir, Cevik Onar, & Bagriyanik, 2020). Con esta información se generan ***modelos de regresión lineal*** que permitan predecir el número de visitantes, con datos de entrenamiento y de testeo. Uno de los problemas para realizar este análisis es encontrar datos estructurados y limpios, que puedan ser procesados e incluidos en el modelo. Para evaluar la eficiencia del modelo de predicción se utilizan medidas como $\boldsymbol{R^2}$, ***MSE***, ***tasa de éxito***, ***estadístico F***, ***multicolinealidad*** y ***autocorrelación***.

Este estudio fue realizado a través de una empresa proveedora de servicios de tecnología y comunicaciones, en donde se utilizaron las siguientes variables: número total de clientes, temperatura diaria, temperatura diaria promedio, cantidad de lluvia diaria, datos del mercado financiero, densidad promedio de tráfico por hora, densidad de tráfico promedio diario y tendencias de búsqueda en internet. En los resultados se encontró que factores ambientales como temperatura, precipitaciones

y densidad del tráfico tenían un efecto significativo en la cantidad de clientes que iban a los centros comerciales.

Los datos corresponden a 217 días, los cuales fueron particionados en 80% para entrenamiento y 20% para prueba del modelo. Se analizaron 7 regresiones lineales utilizando valores aleatorios con las variables seleccionadas en cada una. Todos presentaron autocorrelación positiva y R2 superiores al 95%.

Otro estudio para predecir el número de visitantes se realizó en restaurantes, utilizando métodos de aprendizaje automático y análisis estadístico con datos internos (sistema POS -Point Of Sale) y externos (clima, eventos, etc.) (Tanizaki, Hoshino, Shimmura, & Takenaka, 2018). Las técnicas de aprendizaje automático empleadas fueron ***Bayesian Linear Regression*** (Regresión Lineal Bayesiana), ***Boosted Decision Tree Regression*** (Regresión con Árboles de Decisión Ampliado) y ***Decision Forest Regression*** (Regresión con Bosques de Desición). A continuación, se describirán brevemente estas tres técnicas.

La ***Regresión Lineal Bayesiana*** es un método para aplicar Redes Bayesianas en aprendizaje automático. La Red Bayesiana es un modelo probabilístico en el que las dependencias condicionales entre múltiples variables aleatorias se expresan mediante una estructura gráfica y las relaciones de dependencia entre variables aleatorias se expresan mediante probabilidades condicionales (Motomura & Hara, 2000). Esta se define por tres variables: variable aleatoria, condición de dependencia entre variables aleatorias y probabilidad condicional.

La ***Regresión con Árboles de Decisión Ampliado*** es un método de aprendizaje que utiliza impulso (boosting), para mejorar la precisión de un algoritmo de aprendizaje, y que maneja múltiples dispositivos de aprendizaje (Freund & Schapire, 1999). En este método, el número de tiempos de aprendizaje del caso pronosticado incorrectamente se incrementa con el fin de mejorar la precisión del aprendizaje, aumentando el peso de ese caso.

La ***Regresión con Bosques de Decisión*** es un método de aprendizaje que utiliza Random Forest, para construir un bosque utilizando múltiples árboles de decisión e integrando los resultados de aprendizaje para cada árbol de decisión (Habe, 2012). El sesgo extremo en el aprendizaje de cada árbol de decisiones se puede prevenir incorporando aleatoriedad al extraer los datos de aprendizaje que se utilizarán en cada árbol de decisiones. Como resultado, se puede prevenir el aprendizaje excesivo.

Los resultados de los modelos anteriores se evaluaron con la ***tasa de predicción*** y la correlación entre variables (***$R^2$***). No se encontró una gran diferencia entre en la tasa de pronóstico utilizando el método Bayesiano y los métodos de decisión, con resultados superiores al 85%.

Por otra parte, se han realizado aplicaciones para predecir el número de turistas que visitan un país, a través de aprendizaje profundo con redes neuronales (***Deep Learning Neural Network***) (Chang & Tsai, 2017). Esta aplicación se basa en el efecto positivo que tiene la industria del turismo en el crecimiento económico, ya que involucra servicios de acomodación, catering, transporte aéreo y terrestre, alquiler de vehículos, servicios de arte y literatura, y ventas del comercio. El uso de aprendizaje profundo con redes neuronales incluyó la selección de características, porque no todas las 69 variables originales son importantes, algunas pueden ser engañosas para el algoritmo de pronóstico y otras pueden llevar a un overfitting (sobreajuste).

El flujo básico del sistema analizado consiste en normalización de los datos, división en entrenamiento y testeo, selección de características, entrenamiento del modelo y evaluación del modelo en entrenamiento y testeo, cómo se observa en la Figura 1.

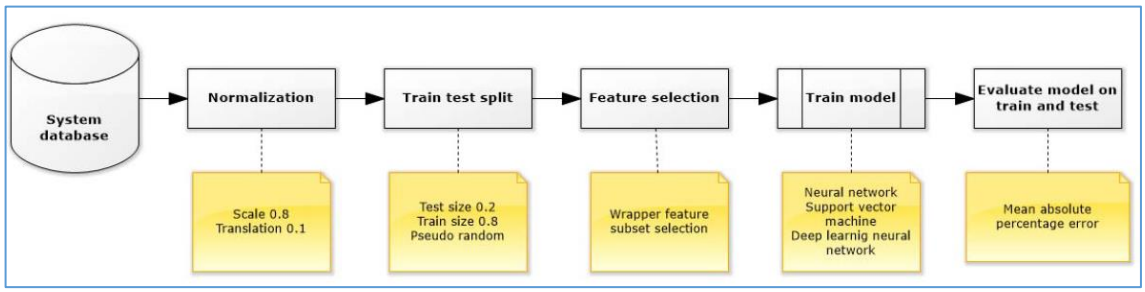



*Figura 1. Resumen del flujo del sistema (Chang & Tsai, 2017).*

Los resultados obtenidos en estos algoritmos muestran que el error absoluto medio porcentual (***MAPE*** - Mean Absolute Percentage Error) fue menor de 10% con redes neuronales y máquinas de soporte vectorial.

Otro ensayo sobre predicción de visitantes se enfoca en los museos, los cuales buscan cada vez más aprovechar el valor de la analítica para tomar decisiones relacionadas con la duración de las exposiciones, las campañas de marketing, la planeación de recursos y la optimización de los ingresos (Yap, Gongy, Naha, &

Mahanti, 2020). Al igual que en otros sectores, los museos se enfrentan a la competencia de actividades bajo techo y al aire libre, incluidos jardines botánicos, zoológicos, bibliotecas, observatorios y rutas de senderismo, entre otros.

El objetivo principal de esta investigación fue comprender el impacto que tiene el clima en la asistencia de visitantes a los museos e investigar el impacto comparable de otros efectos. Las variables de clima a tener en cuenta se agruparon en 3 componentes: térmicos (temperatura, humedad y radiación solar), físicos (lluvia, nieve y viento) y estéticos (nubes, niebla y soleado). La Figura 2 muestra la matriz de correlación entre las variables predictoras y la variable respuesta (visitantes).

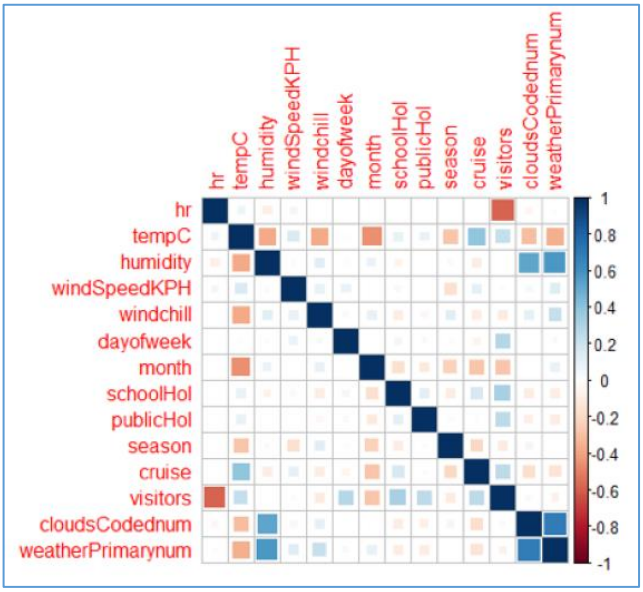



Figura 2. Matriz de correlación de predictores y variable respuesta *(Yap, Gongy, Naha, & Mahanti, 2020)*.

Teniendo en cuenta tanto el tiempo de entrenamiento del modelo como el ajuste (***RMSE*** y $\boldsymbol{R^2}$), el modelo de aumento de gradiente extremo (***Extreme Gradient Boost***) tuvo el mejor rendimiento, con un rango de precisión en el pronóstico del 93%, seguido de bosque aleatorio (***Random Forest***) con 91%, luego red neuronal (***Neural Net***) con 90% y por último, regresión lineal (***Linear Regression***) con 84%. La Figura 3 relaciona los diferentes modelos de regresión implementados.

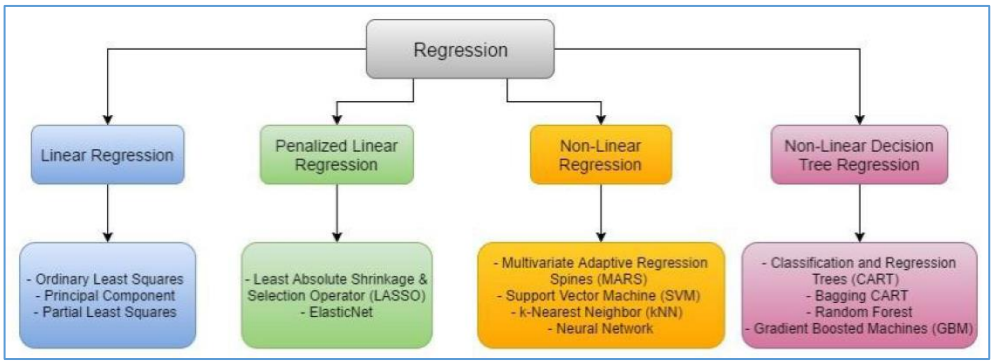



Figura 3. Diagrama de flujo de tipos de regresiones *(Yap, Gongy, Naha, & Mahanti, 2020)*.

En términos de la ***diferencia de variabilidad***, el modelo lineal empleado tuvo un rango del 28%, en red neuronal fue del 37%, en bosque aleatorio del 22% y en aumento de gradiente extremo del 17%, siendo el menor rango de predicción. En conclusión, es muy importante comprender cuáles son los factores que promueven la asistencia de visitantes, mediante la preparación de los datos, la estadística descriptiva, la correlación entre variables y la aplicación de diversos tipos de regresión.

Otra aplicación de series de tiempo con metodología ***ARIMA*** (***Autoregressive and Moving Average***) fue utilizada para pronosticar el número de visitantes internacionales al aeropuerto de Bali (Indonesia) (Mangindaan & Krityakierne, 2018). En este, se tuvieron en cuenta observaciones que fueron afectadas por interrupciones inesperadas producidas por grandes cambios, como políticas gubernamentales, protestas, inestabilidad económica, desastres naturales y terrorismo. Dichos eventos son llamados *intervenciones* en series de tiempo y el modelo cambia por ***ARIMAX*** (ARIMA con intervención), o modelo de función de transferencia, en donde se utiliza una variable *dummy* para representar la presencia de una intervención en el tiempo T.

En este modelo se aplicó la transformación logarítmica para estabilizar la serie de tiempo, ya que exhibía evidencia de no estacionariedad. Los modelos de series de tiempo que mejor se ajustaron al número de llegadas de visitantes internacionales se estimaron en función de los siguientes criterios: mayor $\boldsymbol{R^2}$, menor error cuadrático medio (***RMSE***) y menor error de porcentaje absoluto medio (***MAPE***).

Por otra parte, el uso de métodos de aprendizaje profundo (***Deep Learning***) y en particular de memoria a corto y largo plazo (***LSTM - Long Short-Term Memory***), ha venido aumentando debido a su excelente desempeño en pronósticos (Peng,

Wang, Ai, & Zeng, 2020). En este caso, el modelo LSTM fue utilizado para predecir el número de turistas que ingresan a Beijing y al valle de Jiuzhaigou.

En el proceso anterior se utiliza ***Random Forest*** (RF) para reducir la dimensionalidad de los datos y seleccionar un subconjunto de características con la información más relacionada con las llegadas de turistas. El algoritmo de evolución diferencial (***Differential Evolution*** - DE) se incluye para elegir las longitudes de retraso de cada índice de consulta de búsqueda y los datos históricos de llegada de turistas, para reconstruir la entrada de pronóstico. ***LSTM*** se utiliza para modelar la relación no lineal entre las llegadas de turistas y los datos del índice de consultas de búsqueda. En la figura 4 se presenta un diagrama de flujo de RF-DE-LSTM.

Los resultados muestran que los enfoques de aprendizaje profundo superaron a los métodos más utilizados comúnmente, como ARIMA, VAR (Vector Autoregression Model), MLR (Multiple Linear Regression), SVR (Support Vector Regression) y BPNN (Back-Propagation Neural Network). Los criterios de evaluación empleados incluyen ***MSE***, ***MAE***, ***RMSE*** y ***MAPE***.

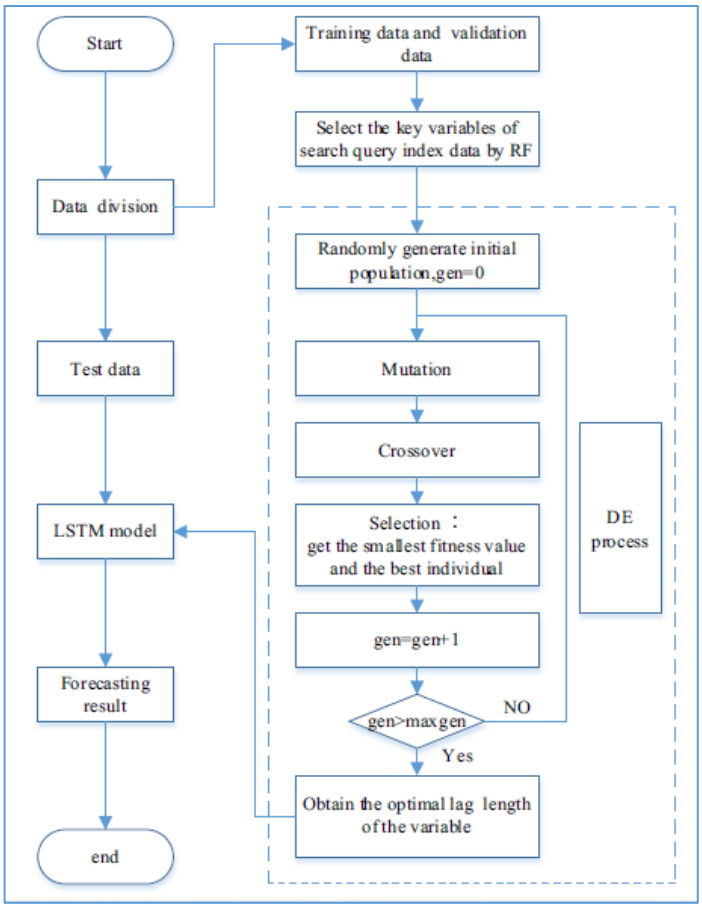



Figura 4. Estructura de predicción del modelo LSTM *(Peng, Wang, Ai, & Zeng, 2020)*.

Una última aplicación de referencia consiste en la predicción de la demanda hotelera diaria (Huang & Zheng, 2021). Este estudio tiene en cuenta el efecto de aglomeración y propone un modelo basado en aprendizaje profundo con correlaciones espaciales y temporales (***DLM-ST - Deep Learning Model with Spatial and Temporal correlations***), el cual se basa en el LSTM estándar.

Las mejoras que incorpora el modelo DLM-ST incluyen:

- Incremento en la eficiencia para capturar dependencias espaciales entre secuencias y entre secuencias temporales mediante la adición de una celda espacial-temporal en las bases LSTM.
- Incorpora un mecanismo de atención para centrarse en ciertas características importantes en la serie temporal que se puede aplicar a la predicción de la demanda de hoteles.
- Los hiperparámetros se determinan sobre la base del método de optimización bayesiano para mejorar el rendimiento de la predicción.

Para el pronóstico de la demanda diaria se empleó una muestra de 210 hoteles en Xiamen, China, con datos entre enero y diciembre de 2019. El set de entrenamiento incluyó 304 días (83%) y el set de testeo incluyó 61 días (17%).

Los modelos empleados para comparar el resultado de DLM-ST incluyen ARIMAX, VAR y LSTM. El desempeño fue evaluado utilizando 5 medidas: MAE, RMSE, MAPE, sMape (symmetric mean absolute percentage error) y Ln Q (log accuracy ratio). DLM-ST obtuvo un RSME más pequeño que ARIMAX, VAR y LSTM, con significancia estadística.

Como se puede observar en los estudios realizados para diferentes sectores, la creación de modelos predictivos de visitantes se resume en los siguientes algoritmos y medidas de desempeño (Tabla 1):

*Tabla 1. Resumen de modelos y medidas de desempeño (elaboración propia).*

| Modelos | Medidas de desempeño |
|---|---|
| Regresión Lineal | $R^2$ |
| Regresión Lineal Bayesiana | MSE |
| Regresión con Árboles de Decisión Ampliado | RMSE |
| Regresión con Bosques de Decisión | Tasa de éxito (predicción) |
| Aprendizaje profundo con redes neuronales | Estadístico F |
| Aumento de gradiente extremo | Multicolinealidad |
| ARIMA | Autocorrelación |
| ARIMAX | MAE |
| LSTM | MAPE |
| VAR | sMAPE |
| DLM-ST | Ln Q |

Cada uno de los modelos presenta resultados que difieren entre sí en las medidas de desempeño, debido al tipo de datos, variables tenidas en cuenta, comportamiento de la serie de tiempo y adecuación del algoritmo. Estos serán la base de predicción para el comportamiento del número de visitantes al centro comercial, que se evaluarán de acuerdo con algunas de las medidas identificadas.

## DEFINICIÓN DE VARIABLES

El conjunto de datos de visitantes corresponde a una serie de tiempo multivariable, que describe la cantidad de visitantes de un centro comercial durante seis años. Los datos se seleccionaron entre enero de 2016 y diciembre de 2021 y las observaciones de los clientes dentro del centro comercial se recopilaron todos los días. Dicha serie multivariada está compuesta por diez variables (sin incluir la fecha), las cuales se describen a continuación:

- **Visitors**: corresponde al número total de **visitantes** al centro comercial por día, teniendo en cuenta clientes que ingresan a pie, en carro o en moto. Esta es la variable a predecir.
- **Weekday**: es una variable categórica (**1** a **7**) de los días de la semana, en donde **1** es domingo, **2** en lunes, ..., hasta el **7** que es sábado.
- **Holiday**: es una variable binaria para la identificación de festivos en el calendario de Colombia, donde **0** es día no festivo y **1** es día festivo.
- **Covid**: es una variable binaria para la marcación por pandemia Covid-19, en donde **0** son los días sin restricciones de movilidad y **1** son los días con restricciones por picos de la pandemia (confinamiento).
- **Payday**: es una variable binaria para el día de pago (quincena), es decir los 15 y 30 de cada mes, donde **0** no es día de pago y **1** es día de pago.
- **Event**: es una variable binaria para los eventos realizados por el centro comercial para generar visitantes, donde **0** corresponde a no hay evento y **1** a si hay evento.
- **Temp**: variable numérica de la temperatura máxima diaria registrada en la ciudad de Medellín, Colombia (en grados Centígrados).
- **Rain**: es una variable binaria para la ocurrencia de lluvia en un día para la ciudad de Medellín, donde **0** es día sin lluvia y **1** es día con lluvia.
- **GoogleTrends01**: variable numérica de la importancia relativa de la búsqueda de "El Tesoro Medellín" en Google todos los días, el valor registrado está entre **0%** y **100%**.
- **GoogleTrends02**: variable numérica de la importancia relativa máxima de las búsquedas diarias en Google de "El Tesoro Parque Comercial", "Centro Comercial El Tesoro" y "El Tesoro Medellín", el valor registrado está entre **0%** y **100%.**

En este proyecto se utilizan los datos del número de visitantes diarios que registra y almacena el centro comercial a través de las cámaras de conteo instaladas. Esta información es de uso interno del centro comercial y sólo será utilizada para el desarrollo del trabajo de grado.

Las variables de temperatura y lluvia fueron tomadas de bases de datos abiertas dispuestas por entidades territoriales o gubernamentales. En la página web de Datos Abiertos se encuentran las variables **Temp** (Ministerio de Tecnologías de Información y Comunicaciones, 2022) y **Rain** (Ministerio de Tecnologías de Información y Comunicaciones, 2022).

La variable **Covid** se tomó de las fechas en donde el gobierno local decretó los picos de pandemia que implicaron el confinamiento o las medidas restrictivas a la movilidad en la ciudad de Medellín.

Para las tendencias de búsqueda en **Google**, se consultaron los términos más relevantes sobre el centro comercial durante el rango de fechas definido (Google, 2022). Es de aclarar que para que los datos fueran diarios, esta consulta se realizó en rangos de seis meses, para horizontes mayores el resultado que arrojaba era semanal.

## VISUALIZACIÓN Y ANÁLISIS EXPLORATORIO DE DATOS

Con el Análisis Exploratorio de Datos (**EDA** - Exploratory Data Analysis) buscamos entender las características de los datos y crear resúmenes y visualizaciones para identificar posibles relaciones entre las variables (Wikipedia, 2022).

La información recolectada contiene **2.192** registros del número de visitantes al centro comercial, con las variables descritas previamente y no presenta valores nulos.

### VISUALIZACIÓN DE LA SERIE DE TIEMPO

La primera gráfica por considerar es la serie de tiempo para la variable a predecir (**visitantes**), en dónde a simple vista se observan picos al finalizar cada año y registros muy bajos con los confinamientos por la pandemia (ver Figura 5).

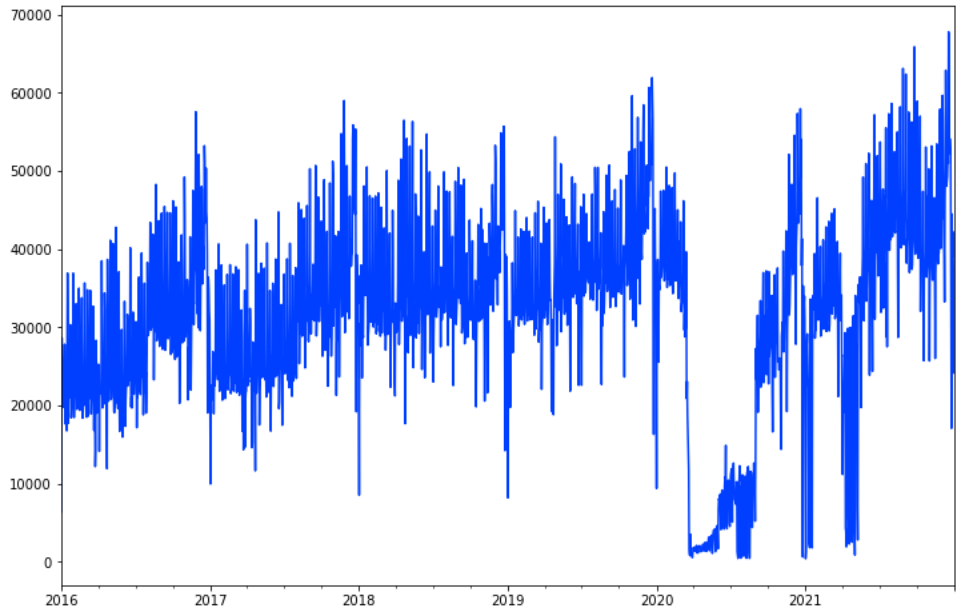



*Figura 5. Serie de tiempo del número de visitantes al centro comercial.*

Al consolidar el número de visitantes de forma mensual podemos comparar visualmente esta variable entre los diferentes años (ver Figura 6), en donde los años 2016 y 2017 son similares, el 2020 es el de menores valores y el 2021 tiene las mejores cifras en el segundo semestre. También se observa estacionalidad en los meses de abril, noviembre y diciembre.

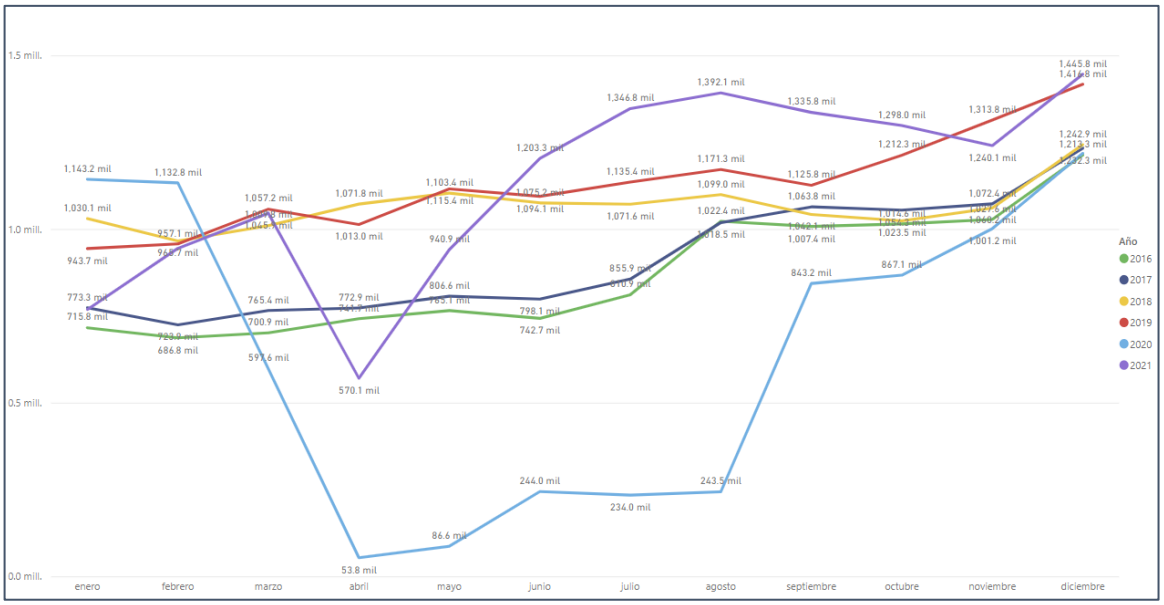



*Figura 6. Visitantes mensuales al centro comercial.*

De forma similar, comparamos el número de visitantes por mes a lo largo de los años en el gráfico de Box plot de la Figura 7. La menor dispersión se presenta en los meses de febrero y septiembre, y la mayor se da en abril y mayo.

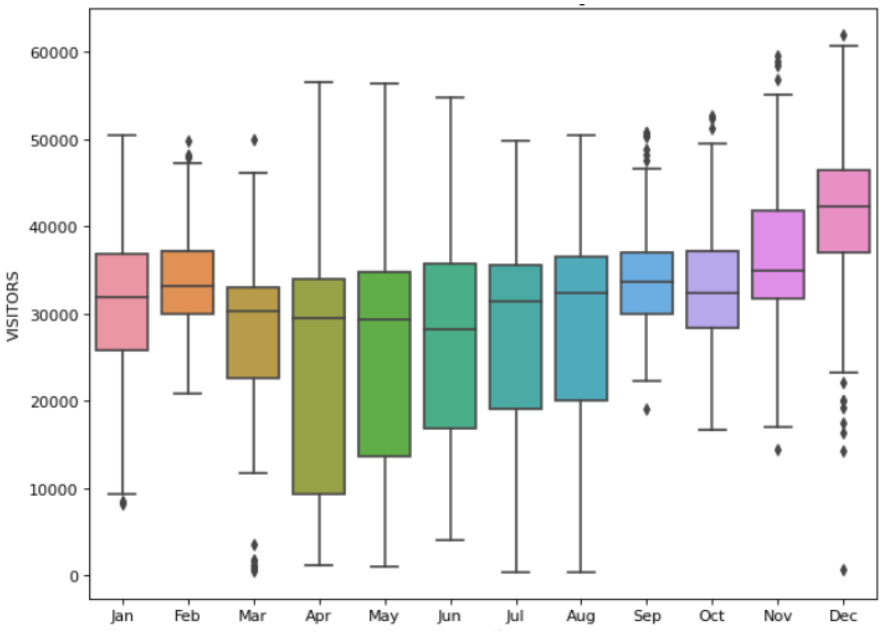



*Figura 7. Box plot de visitantes por mes.*

## ESTADÍSTICA DESCRIPTIVA DE LAS VARIABLES

Para tener un concepto general de los datos realizamos una tabla de estadística descriptiva para las variables numéricas (Tabla 2), en donde se resume la tendencia central, la dispersión y la forma de la distribución del conjunto de datos.

*Tabla 2. Estadística descriptiva de las variables numéricas.*

| | VISITORS | TEMP | GoogleTrends01 | GoogleTrends02 |
|---|---|---|---|---|
| **count** | 2192.00 | 2192.00 | 2192.00 | 2192.00 |
| **mean** | 31448.06 | 28.06 | 0.14 | 0.18 |
| **std** | 11714.95 | 2.16 | 0.18 | 0.19 |
| **min** | 419.00 | 16.20 | 0.00 | 0.00 |
| **25%** | 25317.25 | 26.80 | 0.00 | 0.00 |
| **50%** | 32588.50 | 28.30 | 0.00 | 0.17 |
| **75%** | 38126.50 | 29.70 | 0.23 | 0.27 |
| **max** | 67805.00 | 33.60 | 1.00 | 1.00 |

La Figura 8 muestra los gráficos de densidad y boxplot para la variable Visitantes. En la densidad se pueden observar dos picos, uno cercano a cero y otro en la mediana. En el boxplot se identifican los cuartiles de Visitantes.

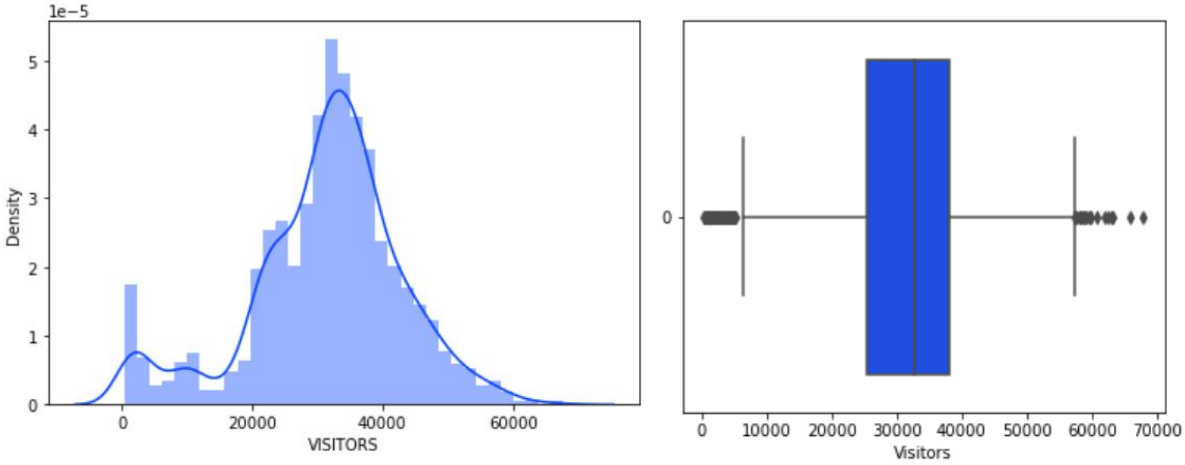



*Figura 8. Diagramas de densidad y dispersión de Visitantes.*

Los gráficos de densidad y boxplot para la variable Temperatura (TEMP) se ven en la Figura 9, con un pico central en 28,3°C y un primer cuartil con temperaturas inferiores a 26,8°C. La menor temperatura observada es de 16,3°C, aunque es un dato atípico está dentro de los históricos de la ciudad, por lo cual se considerará en los datos.

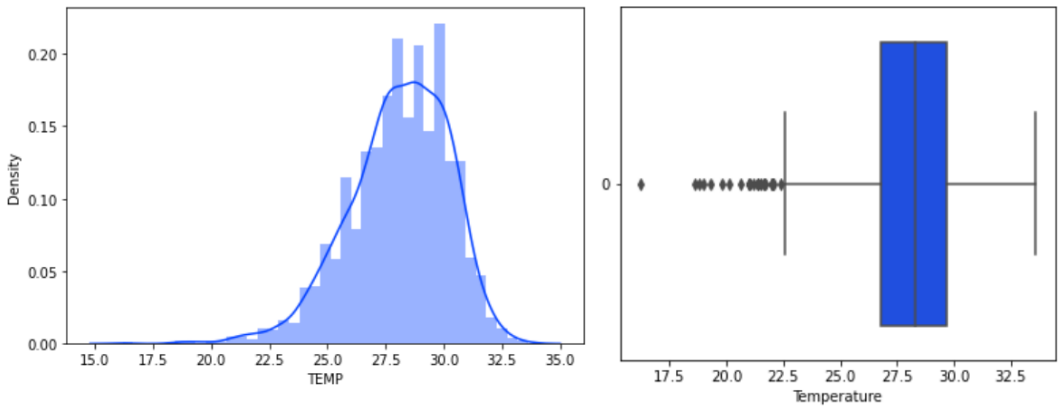



*Figura 9. Diagramas de densidad y dispersión de Temperatura.*

La Figura 10 muestra la tendencia de la serie de tiempo para la variable Visitantes, en donde los valores más altos y más bajos se pueden observar fácilmente. Se distingue una clara estacionalidad desde 2016 hasta el 2020, en donde cada año presenta un pico en diciembre y una caída en enero. Adicionalmente, hay un comportamiento muy similar y creciente en los últimos meses de cada año. La tendencia ayuda a identificar la estacionalidad de la serie, a excepción del año 2020, lo que representa una oportunidad para la modelación con series de tiempo.

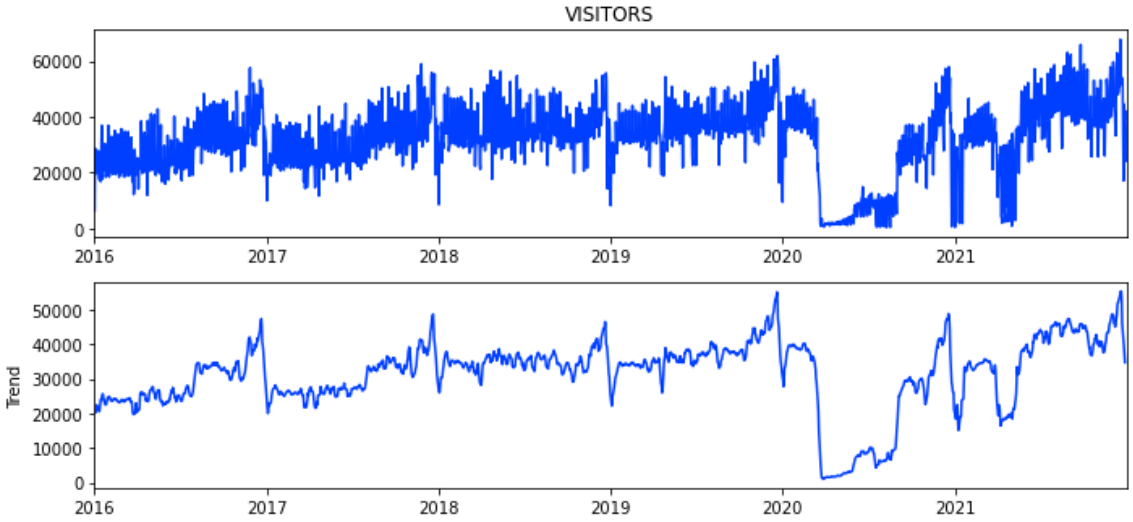



*Figura 10. Tendencia de la variable visitantes.*

Para entender mejor la relación de Visitantes con las variables categóricas, se tienen los boxplot de la Figura 11, en donde las mayores medianas se presentan los días sábado, no festivo y sin Covid-19. Para las variables de quincena, lluvia y evento no se observa una diferencia significativa en los datos medios. La variable binaria Covid muestra un impacto relevante en la afluencia de visitantes, ya que al

estar activa corresponde a los días con restricciones de movilidad en la ciudad por la pandemia.

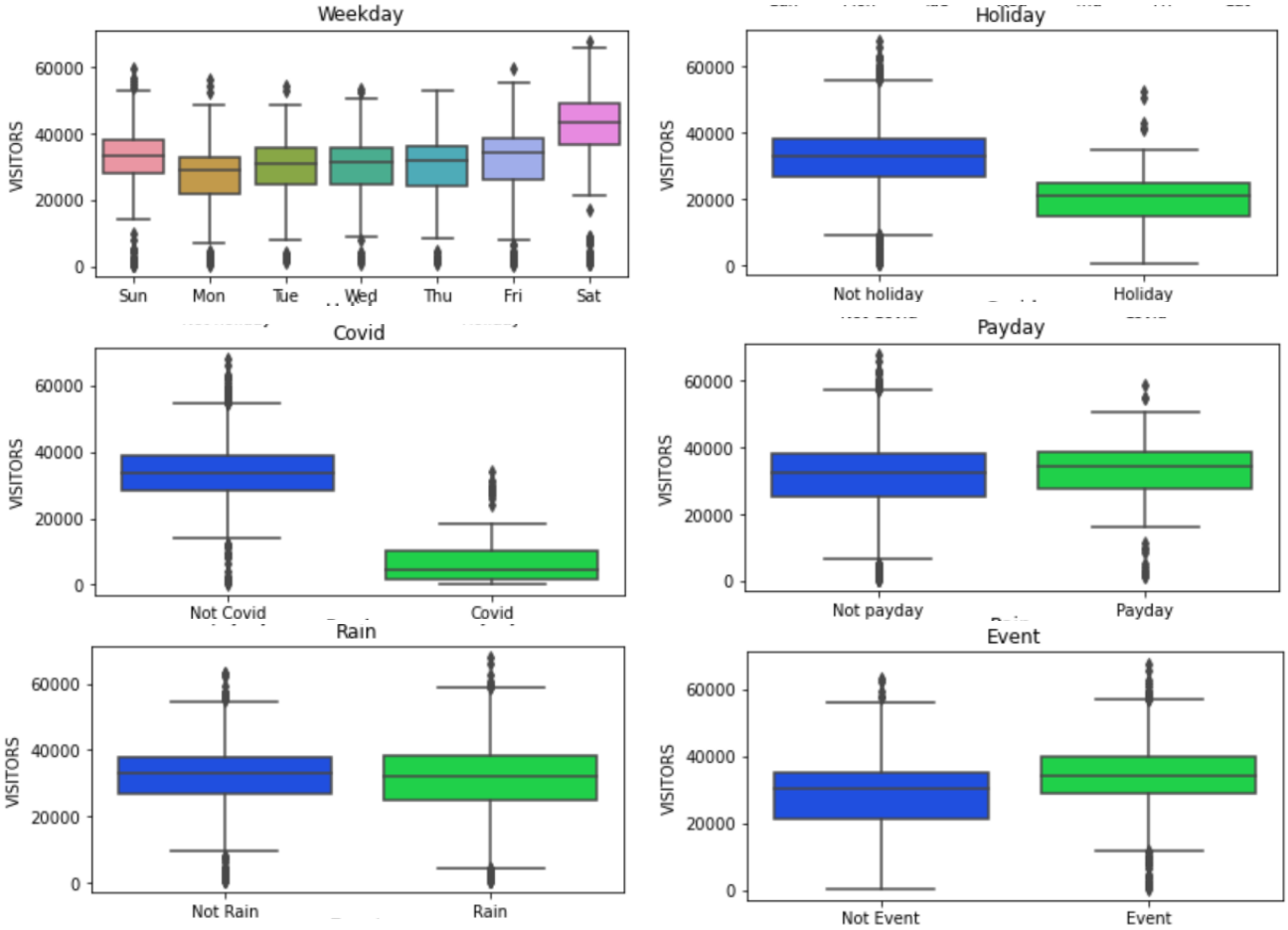



*Figura 11. Dispersión de Visitantes por variables categóricas.*

Considerando el porcentaje de visitantes por día de la semana (Tabla 3), se puede observar que el día de mayor tráfico es el sábado con el 18,24% del total, seguido por el viernes con el 15,50%. Sólo el 3% de los visitantes totales se observa en días festivos y el 6,7% corresponde a días de quincena. No se observa una gran diferencia entre el porcentaje de visitantes con y sin ocurrencia de lluvia.

*Tabla 3. Tablas de contingencia por porcentaje de Visitantes.*

| Weeekday | Holiday | Not Holiday | **Total** |
|---|---|---|---|
| Sat | 0.23% | 18.01% | **18.24%** |
| Fri | 0.22% | 14.28% | **14.50%** |
| Sun | 0.22% | 14.03% | **14.26%** |
| Thu | 0.21% | 13.57% | **13.77%** |
| Wed | 0.24% | 13.38% | **13.62%** |
| Tue | 0.15% | 13.23% | **13.38%** |
| Mon | 1.76% | 10.48% | **12.24%** |
| **Total** | **3.03%** | **96.97%** | **100.00%** |

| Weeekday | Covid | Not Covid | **Total** |
|---|---|---|---|
| Sat | 0.17% | 18.07% | **18.24%** |
| Fri | 0.28% | 14.21% | **14.50%** |
| Sun | 0.10% | 14.16% | **14.26%** |
| Thu | 0.43% | 13.34% | **13.77%** |
| Wed | 0.41% | 13.20% | **13.62%** |
| Tue | 0.42% | 12.96% | **13.38%** |
| Mon | 0.34% | 11.90% | **12.24%** |
| **Total** | **2.16%** | **97.84%** | **100.00%** |

| Weeekday | Not Payday | Payday | **Total** |
|---|---|---|---|
| Sat | 17.03% | 1.21% | **18.24%** |
| Fri | 13.53% | 0.96% | **14.50%** |
| Sun | 13.34% | 0.91% | **14.26%** |
| Thu | 12.81% | 0.96% | **13.77%** |
| Wed | 12.68% | 0.94% | **13.62%** |
| Tue | 12.52% | 0.87% | **13.38%** |
| Mon | 11.39% | 0.85% | **12.24%** |
| **Total** | **93.30%** | **6.70%** | **100.00%** |

| Weeekday | Not Rain | Rain | **Total** |
|---|---|---|---|
| Sat | 8.01% | 10.23% | **18.24%** |
| Fri | 6.43% | 8.06% | **14.50%** |
| Sun | 7.00% | 7.26% | **14.26%** |
| Thu | 6.27% | 7.50% | **13.77%** |
| Wed | 5.87% | 7.75% | **13.62%** |
| Tue | 5.76% | 7.62% | **13.38%** |
| Mon | 6.19% | 6.05% | **12.24%** |
| **Total** | **45.53%** | **54.47%** | **100.00%** |

| Weeekday | Event | Not Event | **Total** |
|---|---|---|---|
| Sat | 12.27% | 5.96% | **18.24%** |
| Fri | 9.70% | 4.79% | **14.50%** |
| Sun | 9.74% | 4.52% | **14.26%** |
| Thu | 8.72% | 5.05% | **13.77%** |
| Wed | 8.30% | 5.32% | **13.62%** |
| Tue | 8.13% | 5.26% | **13.38%** |
| Mon | 7.77% | 4.46% | **12.24%** |
| **Total** | **64.63%** | **35.37%** | **100.00%** |

La autocorrelación (ACF) y la autocorrelación parcial (PACF) para la variable Visitantes (Figura 12) muestra que los datos podrían estar correlacionados con registros hasta 100 días atrás.

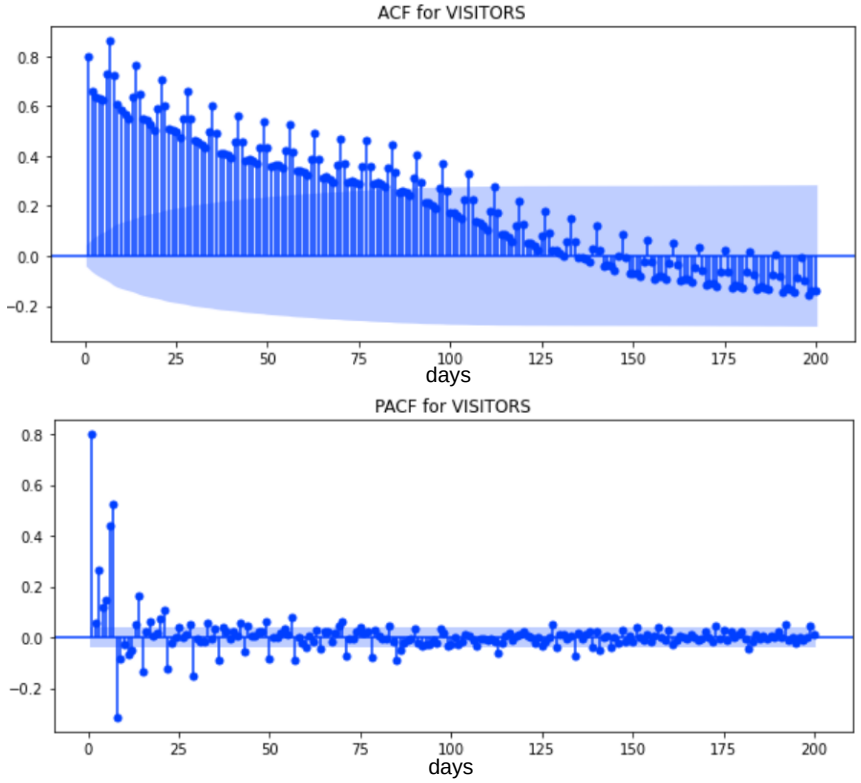
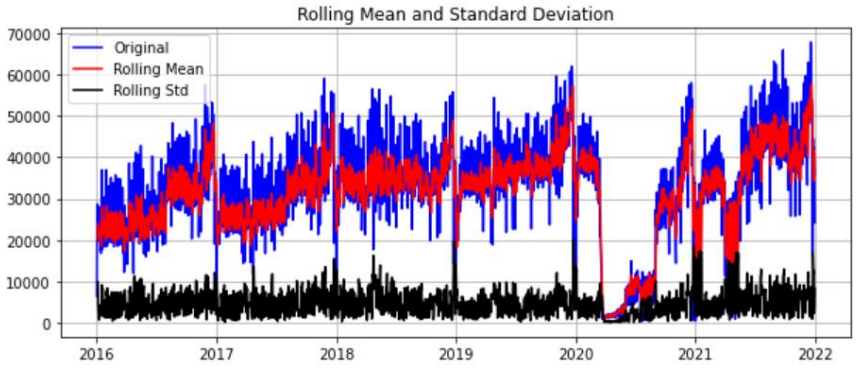



*Figura 12. Autocorrelación y Autocorrelación parcial de Visitantes.*

Al realizar la prueba estadística de Dickey-Fuller Aumentada (ADF) vemos que el valor-p es de 0,014 (Figura 13), por lo cual se rechaza de la hipótesis nula de que existe una raíz unitaria para la variable Visitantes. Adicionalmente, con una significancia del 5% la serie de tiempo de Visitantes es estacionaria, es decir que la media y varianza son aproximadamente constantes a lo largo del tiempo.




```
VISITORS
Test Statistic                -3.314073
p-value                        0.014261
Lags Used                     21.000000
Observations Used           2170.000000
Critical Value (1%)           -3.433367
Critical Value (5%)           -2.862873
Critical Value (10%)          -2.567480
dtype: float64
```

*Figura 13. Test Dickey-Fuller aumentada de estacionariedad para Visitantes.*

## VARIABLES REDUNDANTES Y REZAGOS

Al analizar las variables relacionadas con búsquedas en internet (GoogleTrends) podemos considerar que una de ellas es redundante en el conjunto de datos (ver Tabla 4). La correlación entre Visitantes y GoogleTrends02 es mayor (0,175) que con GoogleTrends01 (0.143), por lo cual se puede incluir sólo GoogleTrends02 en el modelo.

*Tabla 4. Matriz de correlaciones entre Visitantes y GoogleTrends.*

| | VISITORS | GoogleTrends01 | GoogleTrends02 |
|---|---|---|---|
| **VISITORS** | 1.000000 | 0.143426 | 0.175384 |
| **GoogleTrends01** | 0.143426 | 1.000000 | 0.852681 |
| **GoogleTrends02** | 0.175384 | 0.852681 | 1.000000 |

Cuando tenemos búsquedas en internet en el conjunto de datos es importante validar si existe un rezago entre la consulta y las visitas al centro comercial, por lo cual realizamos la correlación entre visitantes y GoogleTrends02 con rezagos hasta de 10 días, obteniendo el mejor resultado con 7 días. La correlación obtenida fue de

***0.212*** y la nueva variable que se incluye en el modelo es GoogleTrends03, reemplazando a GoogleTrends02.

La correlación entre Visitantes y GoogleTrends03 corresponde al coeficiente de Pearson, que es una medida de dependencia lineal. Al emplear coeficientes de correlación no paramétricos, se obtuvo una correlación de Spearman de 0,182 y de Kendall de 0,129 entre las variables, los cuales son similares en signo y magnitud respecto al coeficiente de Pearson.

## ANÁLISIS DE REGRESIÓN LINEAL

Una última consideración para el análisis exploratorio es validar si los datos se pueden modelar con una regresión lineal. Para esto se ejecutó un primer modelo con las variables independientes Weekday, Holiday, Covid, Payday, Event, Temp, Rain y GoogleTrends03, y variable dependiente Visitors. El resultado de la regresión arrojó un $R^2$ de 0,537 y todas las variables con significancia estadística, excepto Payday y Rain. Luego se ejecutaron 3 regresiones adicionales, 2 sin considerar cada una de estas variables individualmente y luego una sin considerarlas simultáneamente.

El modelo de regresión lineal final fue el siguiente:

$$Visitors = 3{,}80x10^4 + 1.183{,}78*Weekday - 1{,}07\ x10^4*Holiday - 2{,}43\ x10^4*Covid + 2.496{,}85*Event - 398{,}01*Temp + 6.372{,}07*Googletrends03$$

Por último, realizamos pruebas de diagnóstico a la regresión lineal y de acuerdo con los tests de multiplicador de Lagrange, Jarque-Bera, Kurtosis, Sesgo y análisis de residuales, encontrando que los datos presentan heterocedasticidad, autocorrelación y que los residuales no se distribuyen normal. Esto implica que la regresión lineal no es un modelo adecuado para el conjunto de datos y sugiere implementar un modelo de series de tiempo.

# MODELADO Y EVALUACIÓN

Para la modelación de los datos se tienen en cuenta dos aproximaciones desde la ciencia de datos, la primera consiste en modelos clásicos de series de tiempo univariados y multivariados, y la segunda corresponde a modelos de aprendizaje de máquina, tanto de regresión como de clasificación.

## MODELOS DE SERIES DE TIEMPO CLÁSICOS

Los métodos clásicos de pronóstico de series de tiempo pueden estar enfocados en relaciones lineales y no lineales, pueden explicar sistemas complejos y capturar comportamientos dinámicos y causalidades de los procesos subyacentes, además proporcionan un medio manejable para predecir y monitorear la evolución del estado de un sistema (Cheng, Beyca, & Le, 2015).

Tomando en cuenta los componentes de las series analizadas, se incluyen procesos de modelación para series de tiempo clásicos univariados, como autoregresión y media móvil, y multivariados como Sarimax y Varmax. Los mejores resultados se obtuvieron con VARMA (Vector Autoregression Moving-Average), AR (Autoregression) y SARIMA (Seasonal Autoregressive Integrated Moving-Average).

En esta aproximación con modelos clásicos se consideraron 11 métodos de predicción para series de tiempo:

- Autoregression (AR).
- Moving Average (MA).
- Autoregressive Moving Average (ARMA).
- Autoregressive Integrated Moving Average (ARIMA).
- Seasonal Autoregressive Integrated Moving-Average (SARIMA).
- Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX).
- Vector Autoregression (VAR).
- Vector Autoregression Moving-Average (VARMA).
- Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX).
- Simple Exponential Smoothing (SES).

- Holt Winter's Exponential Smoothing (HWES).

El pronóstico considera un horizonte de 30 días y la comparación de desempeño se realizó con RMSE, en donde el menor RMSE se obtuvo con VARMA (1.450) y el mayor con SES (5.374). Ahora observemos los dos mejores modelos y sus predicciones.

El método VARMA es la generalización de ARMA a múltiples series de tiempo, en donde se modela el siguiente paso de la serie utilizando el modelo ARMA. Con este se obtuvo un RMSE de 1.450 visitantes diarios y el orden (p, q) fue de (1, 1).

El método AR modela el siguiente paso en la secuencia como una función lineal de las observaciones en pasos de tiempo anteriores. El RMSE para AR fue de 2.357 visitantes diarios, con rezagos (lags) de 7 días. La Figura 14 muestra las predicciones de los dos métodos para enero de 2022, en donde se puede ver como tienden a aplanarse en un valor esperado a medida que aumenta el tiempo.

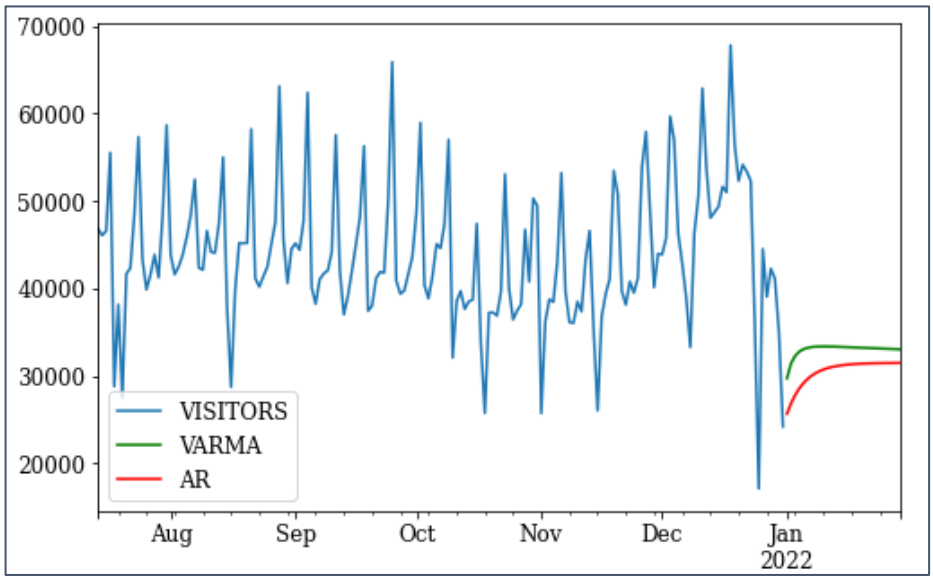



*Figura 14. Predicción de Visitantes con VARMA y AR.*

Para el modelo SARIMA se optimizaron los hiperparámetros mediante un algoritmo de descomposición estacional (seasonal_decompose) de la librería *statsmodels* de Python, encontrando que los mejores fueron (1,0,1)(1,1,1)[7], con un RMSE de 3.204 y la predicción resultante se puede ver en la Figura 15.

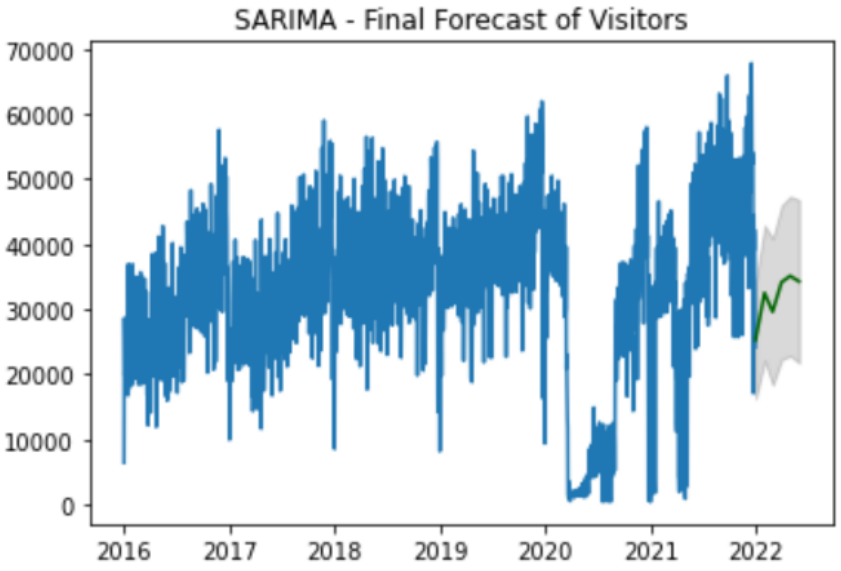



*Figura 15. Predicción de Visitantes con SARIMA optimizado.*

De forma similar, se optimizaron los hiperparámetros del modelo SARIMAX, dando como resultado los siguientes parámetros: (3,0,2)(0,1,1)[7], con un RMSE de 3.552. La predicción resultante se puede ver en la Figura 16.

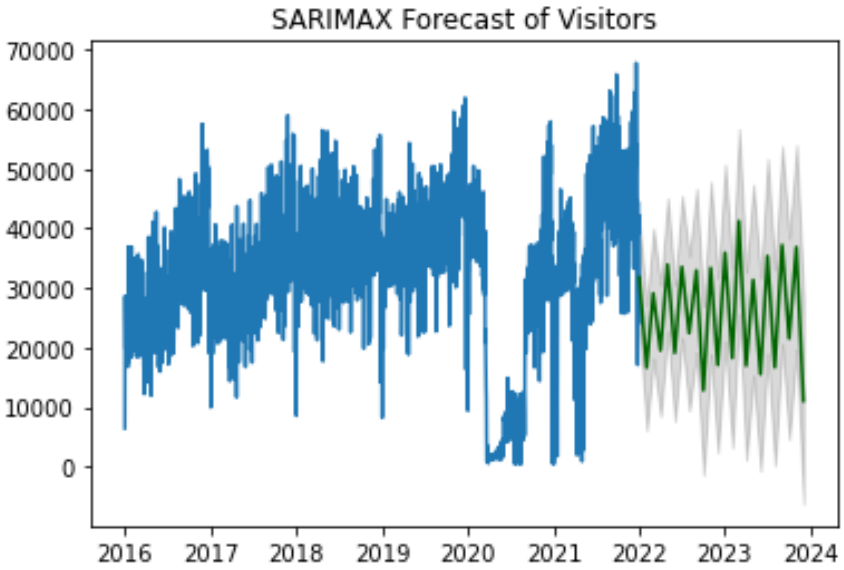



*Figura 16. Predicción de Visitantes con SARIMAX optimizado.*

Ambos modelos tuvieron un buen desempeño, con todos sus parámetros estadísticamente significativos. Dado que los datos tienen un comportamiento poco volátil, para la variable (Visitantes), los modelos clásicos tienen un ajuste adecuado y son un buenos predictores.

Un aspecto relevante en el modelo es la ocurrencia de la pandemia, ya que perturbó la tendencia y estacionalidad del número de visitantes. Si consideramos un horizonte temporal desde 2016 hasta 2019 como datos de entrenamiento de un modelo SARIMA y predecimos para los años 2021 y 2022, podemos observar un comportamiento esperado para el número de visitantes del centro comercial sin Covid-19 (ver figura 17).

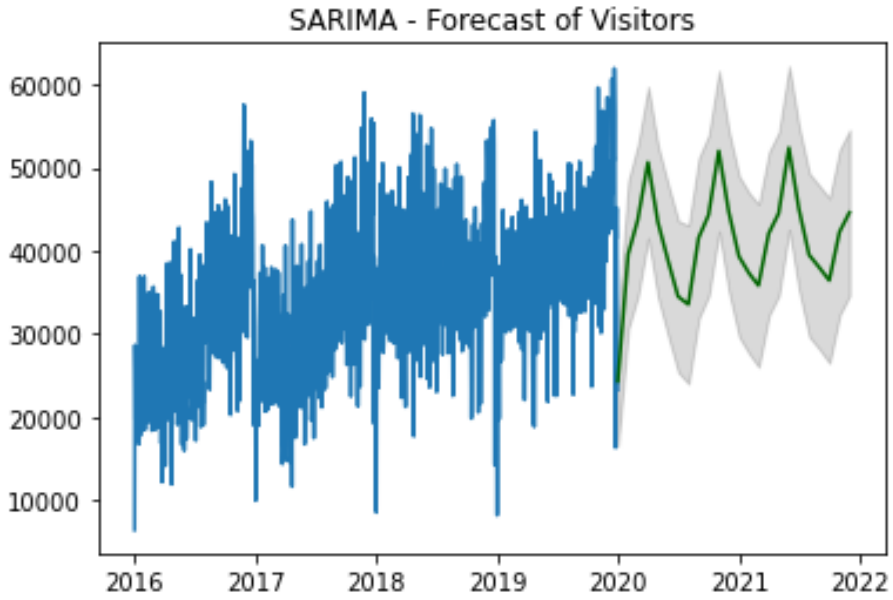



*Figura 17. Predicción de Visitantes 2020 y 2021 con SARIMA.*

Aunque un cierre generalizado económico y social por pandemia no estaba entre las proyecciones a finales de 2019, esta nos mostró que debemos estar preparados para cambios abruptos en el mercado y generar modelos adaptables que consideren nuevas variables explicativas.

## MODELOS MEDIANTE APRENDIZAJE DE MÁQUINA

Los modelos de aprendizaje de máquina (Machine Learning), también llamados modelos de caja negra o basados en datos, son ejemplos de modelos no lineales no paramétricos que usan solo datos históricos para aprender la dependencia estocástica entre el pasado y el futuro (Bontempi, Ben Taieb, & Borgne, 2012).

En esta implementación se pretende predecir el número diario de visitantes al centro comercial, para ello se utilizaron en primer lugar algoritmos de aprendizaje de máquinas para regresión y posteriormente algoritmos para clasificación.

El procedimiento aplicado consiste en entrenar el modelo con una porción de los datos, que le permiten al algoritmo hacer su aprendizaje inicial. De esta forma es capaz de plantear correlaciones en los datos de entrenamiento y proporcionar un modelo, para luego ser validado con los datos de prueba. La métrica utilizada para evaluar el desempeño de los modelos fue MAPE.

En esta aproximación con modelos de aprendizaje de máquina se consideraron 11 métodos de **regresión** para predicción con series de tiempo, con el fin de predecir tendencias en el número diario de visitantes:

- Ridge.
- Lasso (Least absolute shrinkage and selection operator).
- Elasticnet.
- XGBoost (eXtreme Gradient Boosting).
- LightGBM (Light Gradient Boosting Machine).
- KNN (K-Nearest Neighbors).
- RF (Random Forest).
- SGD (Stochastic gradient descent).
- Bagging (Bootstrap Aggregation).
- MLR (Multiple Linear Regression).
- Stacking (Stacked Generalization).

Todos los modelos se ejecutaron con el paquete *scalecast* (Dynamic Forecasting at Scale) de Python, que utiliza un enfoque de previsión escalable con modelos comunes de *scikit-learn* y *stats*, así como modelos de Facebook Prophet, Microsoft LightGBM y LinkedIn Silverkite, para pronosticar series de tiempo (Keith, 2022).

El pronóstico considera un horizonte de 60 días, con un test split de 20% en el conjunto de datos y un periodo de validación de 60 días para ajustar el modelo. La comparación de desempeño de los modelos se realizó con MAPE InSample (de prueba) y de Test (validación).

La Tabla 5 muestra los resultados de los diferentes métodos empleados, en donde LigthGBM, SGD y XGboost tuvieron los mayores MAPE de prueba. Se destaca también que en todos los modelos se presentó overfitting (sobreajuste), si se comparan los MAPE entre los datos de prueba y de validación. Ahora consideremos dos de los mejores modelos, con sus hiperparámetros y predicciones.

*Tabla 5. Resultados de modelado con aprendizaje de máquina.*

| Model | TestSetMAPE | InSampleMAPE |
|---|---|---|
| **LightGBM** | 0.8839 | 0.1919 |
| **SGD** | 0.8189 | 0.2556 |
| **XGboost** | 0.8922 | 0.0114 |
| **RF** | 0.9410 | 0.1108 |
| **KNN** | 0.9739 | 0.3435 |
| **Stacking** | 0.9217 | 0.2323 |
| **Bagging** | 1.1393 | 0.2561 |
| **Ridge** | 0.7825 | 0.2481 |
| **Lasso** | 0.7966 | 0.2478 |
| **Elasticnet** | 0.8588 | 0.2478 |
| **MLR** | 7.95E+09 | 0.2475 |

La optimización de hiperparámetros en los modelos se realizó con grid search, evaluando y comparando los MAPE, y considerando entre otros ‘max_depth’ desde 2 hasta 12, ‘n_estimators’ desde 10 hasta 250, ‘max_features’ 'auto', 'sqrt' y 'log2', y ‘max_samples’ desde 0.75 hasta 1.

La figura 18 muestra los resultados del set de prueba para el método LightGBM, con un intervalo de confianza del 95%, cuyos hiperparámetros fueron: n_estimators=250, boosting_type=goss, max_depth=2, learning_rate=0.1, reg_alpha=0.0, reg_lambda=1.0 y num_leaves=45. El $R^2$ de prueba obtenido fue de 0.9083 y el de validación fue de 0.3124. La predicción resultante con LigthGBM se observa en la Figura 19.

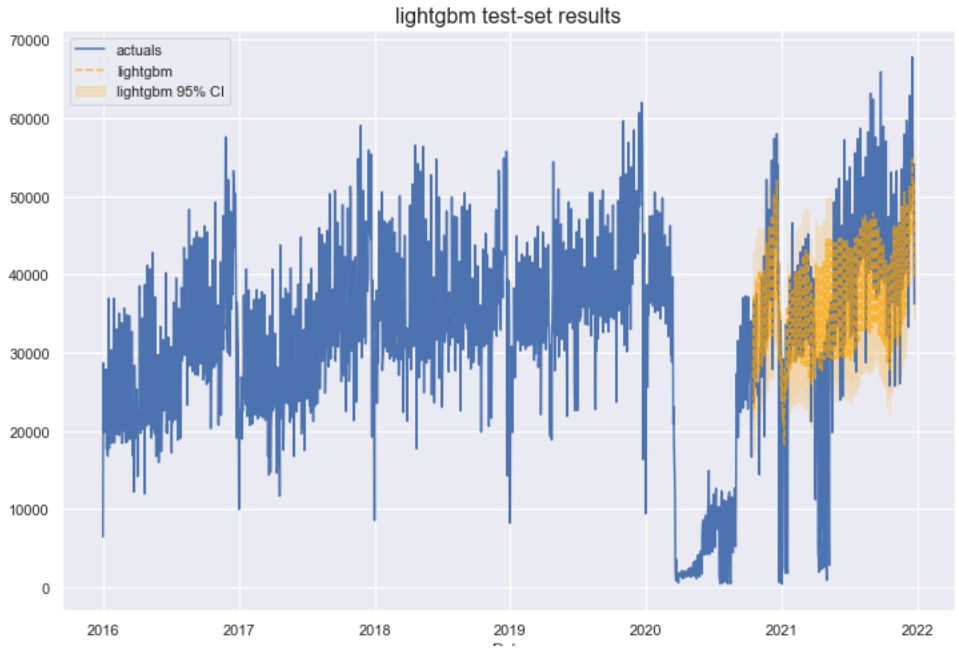



*Figura 18. Resultados de test-set con modelo LigthGBM.*

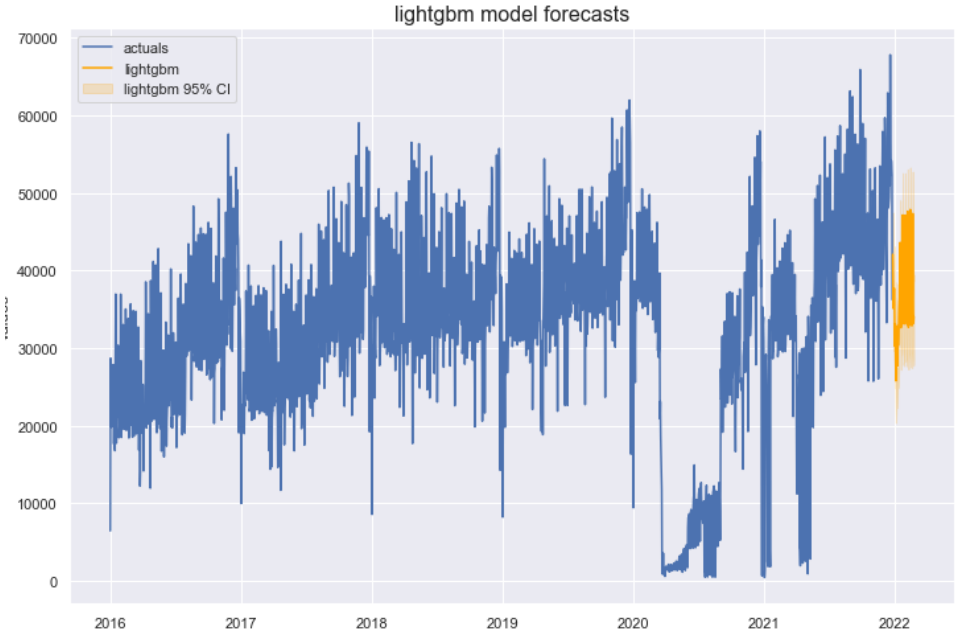



*Figura 19. Predicción con modelo LightGBM.*

La figura 20 muestra los resultados del set de prueba para el método SGD, con un intervalo de confianza del 95%, cuyos hiperparámetros fueron: penalty=elasticnet, L1_ratio=0.15 y learning_rate=constant. El MAPE de prueba obtenido fue de 0.2556 y el de validación fue de 0.8189. La predicción resultante con SGD se observa en la Figura 21.

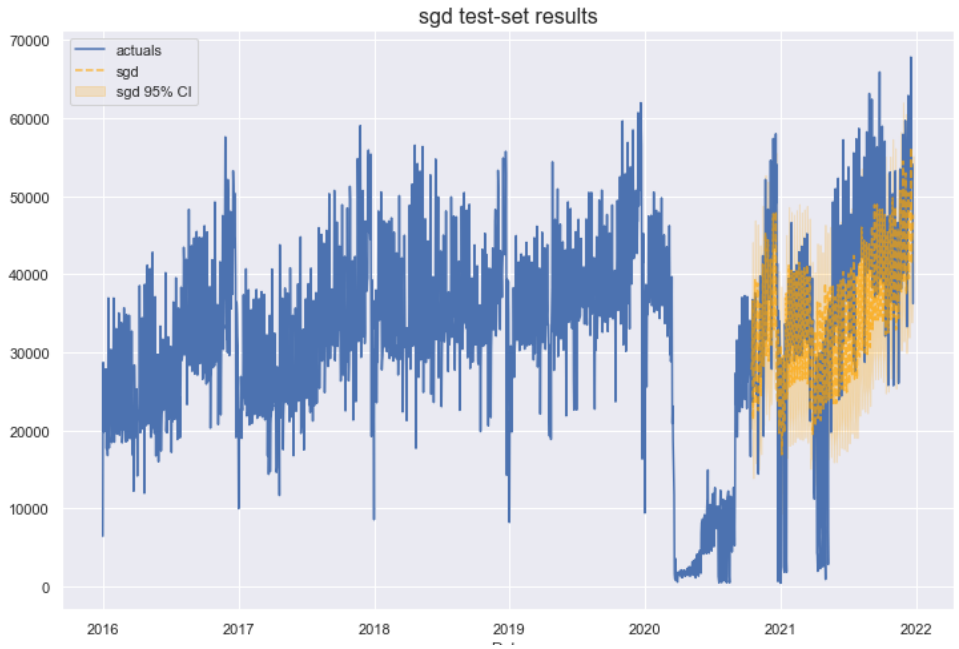



*Figura 20. Resultados de test-set con modelo SGD.*

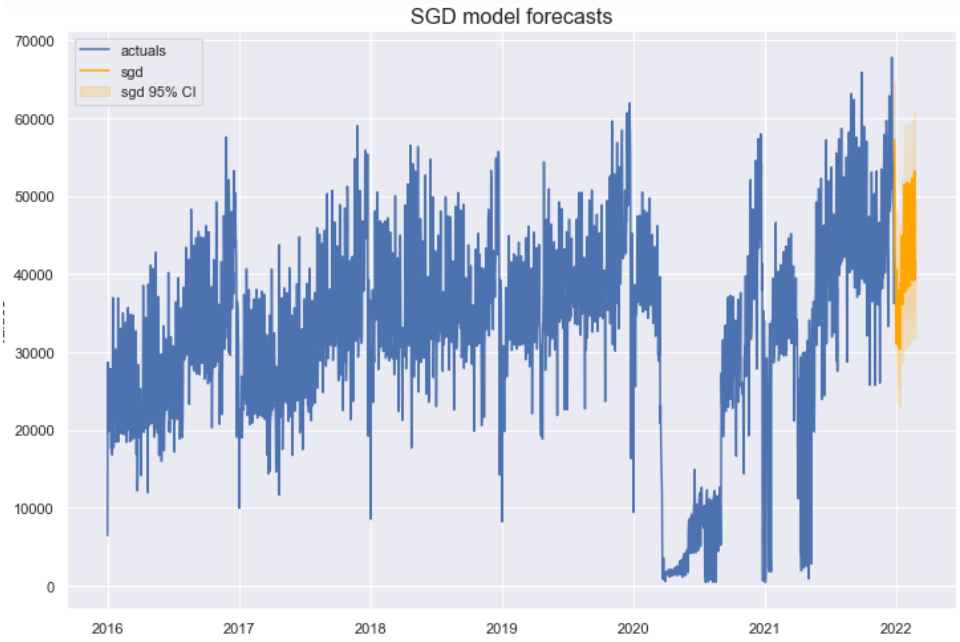



*Figura 21. Predicción con modelo SGD.*

Además de los hiperparámetros y de acuerdo con las características identificadas en el EDA (como estacionalidad y autocorrelación), los regresores de series de tiempo incluidos fueron:

- 7 términos autorregresivos.
- 2 términos autorregresivos estacionales separados por 12 períodos.
- Mes, trimestre, semana y día del año con una Transformación de Fournier.
- Año bisiesto y semana como variables ficticias.
- Año.

Otra aproximación con aprendizaje de máquina consistió en realizar modelos de **clasificación** al conjunto de datos, con optimización de los hiperparámetros implementando *cross validation* y comparando las medidas de desempeño con MAPE. Esta opción se puede considerar en predicción al utilizar como etiqueta el número de visitantes.

En esta aproximación se consideraron métodos no lineales y algoritmos de ensamble:

- k-Nearest Neighbors (KNN).
- Classification and Regression Tree (CART).
- Support Vector Machine (SVM).
- Naive Bayes.
- Bagged Decision Trees.
- Random Forest (RF).
- Extra Trees (ET).
- Gradient Boosting Machine (GBM).

La siguiente tabla (6) muestra los resultados de estos modelos de clasificación con aprendizaje de máquina.

*Tabla 6. Resultados de modelado con clasificación.*

| Model | MAPE |
|---|---|
| **Bayes** | 0.0168 |
| **GBM** | 0.0198 |
| **RF** | 0.0213 |
| **ET** | 0.0244 |
| **Bag** | 0.0244 |
| **CART** | 0.0290 |
| **KNN** | 0.0396 |
| **SVM** | 0.0777 |

Podemos ver que tanto el método de Naive Bayes como Gradient Boosting Machine logran un MAPE inferior al 2% en el conjunto de prueba.

Por último, desarrollamos un modelo de red neuronal convolucional (**CNN**) para el conjunto de datos, utilizando la biblioteca de *Keras* y activaciones *Relu* y *Softmax*.

El modelo se ajustó a un número fijo de épocas, en este caso 10, y se utilizó un tamaño de lote (batch) de 32 muestras, donde se exponen 32 ventanas de datos al modelo antes de que se actualicen los pesos del modelo. El resultado obtenido es un MAPE de 0.0181, el cual es muy bueno comparándolo con los anteriores.

## CONCLUSIONES

Este proyecto es de mucha utilidad tanto en centros comerciales como en grandes superficies, en donde la planeación y asignación de los recursos físicos, humanos y económicos dependerán del número de clientes o visitantes a lo largo del tiempo. Igualmente, con las variables exógenas incluidas en el modelo se pueden crear estrategias de mercadeo para atraer visitantes en empresas similares.

La selección de variables explicativas en el modelo dependerá del tipo de centro comercial, su ubicación geográfica, variedad de servicios o producto ofertados y del perfil de cliente. Para este caso, las variables día de pago (quincena) y lluvia no fueron significativas respecto al número de visitantes, aunque no necesariamente se descarten para otro centro comercial.

El número de visitantes al centro comercial es una variable que presenta heterocedasticidad, autocorrelación y sus residuales no se distribuyen normal, por lo cual un modelo lineal no se ajusta a los datos. Los modelos clásicos de series de tiempo tuvieron un buen desempeño al no ser una variable volátil.

Para la consecución de datos meteorológicos fue necesaria la combinación de varias fuentes, como el Siata, Datos Abiertos y páginas web de clima. Es importante considerar si las variables corresponden a mínimos, máximos, promedios o medidas por minutos u horas, para procesar adecuadamente esta información y que sea de utilidad en el modelado. Además, pueden existir otras variables que influyan el número de visitantes en diversas circunstancias, como calidad del aire, radiación, humedad y viento. Debido a la ubicación geográfica de la ciudad de Medellín y al nicho de mercado o perfil de clientes del centro comercial, estas variables climáticas no tuvieron un mayor efecto en el modelo.

El análisis exploratorio de datos es de suma importancia en cualquier modelado, ya que ayuda a identificar variables significativas, rezagos, datos medios, tendencias y dispersión de los datos. Para esta actividad se cuenta con diversidad de métodos analíticos y descriptivos que facilitan la inferencia estadística.

Aunque los modelos de aprendizaje de máquinas suelen presentar muy buenos desempeños en términos de predicción en comparación con los modelos de series de tiempo clásicas, es necesario disponer de tiempo y dedicación para la realización

de la calibración de los hiperparámetros y la regularización de los modelos. Lo anterior con el fin de poder llevar a cabo una buena generalización de los resultados.

Como se pudo ver en los diversos modelos, el ajuste de los hiperparámetros en los modelos de aprendizaje de máquina es muy importante para el buen resultado de los mismos y de sus predicciones, por lo cual deben ejecutarse y validarse adecuadamente hasta hallar los óptimos.

El número de modelos clásicos y de aprendizaje de máquina disponibles actualmente es muy amplio, así como la capacidad de procesamiento y almacenamiento computacional, lo cual facilita el trabajo predictivo. Sin embargo, es de suma importancia conocer con antelación el tipo de algoritmos que cada modelo ejecuta, para ejecutar aquellos que más se acomoden al tipo de datos.

Para una futura implementación de este modelo en el centro comercial se puede considerar el número de visitantes mensuales, ya que la asignación de recursos humanos como personal de vigilancia y de aseo generalmente se realiza por mes, mientras que la apertura de entradas (vehiculares y peatonales) y de lotes de parqueadero se puede hacer diariamente e incluso por franjas horarias.

Dado el número de datos en este modelado, el hardware utilizado fue un equipo básico (Procesador Core i5 @2.40GHz, 8GB RAM) con sistema operativo de 64 bit. Asimismo, los tiempos de procesamiento de los algoritmos fueron bajos, con una duración de 470 segundos en la optimización de hiperparámetros más demorada.

## REFERENCIAS


Bontempi, G., Ben Taieb, S., & Borgne, Y.-A. (2012). Machine learning strategies for time series forecasting. *European business intelligence summer school*, 62-77.

Chang, Y.-W., & Tsai, C.-Y. (2017). Apply deep learning neural network to forecast number of tourists. *31st International Conference on Advanced Information Networking and Applications Workshops.* Taipei.

Chebat, J.-C., Sirgy, J., & Grzeskowiak, S. (2010). How can shopping mall managment best capture mall image. *Journal of Business Research*, 735-740.

Cheng, C., Beyca, O., & Le, T. (2015). Time series forecasting for nonlinear and non-stationary processes: a review and comparative study. *Iie Transactions ,vol. 47, no 10,*, 1053-1071.

Freund, Y., & Schapire, R. E. (September de 1999). A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, págs. 771-780.

Google. (2022). Obtenido de Google Trends: https://trends.google.es/trends/?geo=CO

Habe, H. (2012). *Random Forests.* IPSJ SIG Technical Report.

Huang, L., & Zheng, W. (2021). Novel deep learning approach for forecasting daily hotel demand with agglomeration effect. *International Journal of Hospitality Management, 98*(103038), 11.

Keith, M. (30 de 1 de 2022). *Forecast dynamically at scale with this unique package*. Obtenido de PythonRepo: https://zzun.app/repo/mikekeith52-scalecast

Mangindaan, J. M., & Krityakierne, T. (2018). Analysis of international visitor arrivals in Bali: modeling and forecasting with seasonality and intervention. *3rd International Conference on Mathematical Sciences and Statistics.* Thailand.

Ministerio de Tecnologías de Información y Comunicaciones. (2022). *Datos Hidrometeorológicos Crudos - Red de Estaciones IDEAM : Temperatura*. Obtenido de Datos Abiertos: https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Datos-Hidrometeorol-gicos-Crudos-Red-de-Estaciones/sbwg-7ju4

Ministerio de Tecnologías de Información y Comunicaciones. (2022). *Precipitación*. Obtenido de Datos Abiertos: https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Precipitaci-n/s54a-sgyg

Motomura, Y., & Hara, I. (2000). Bayesian Network Learning System based on Neural Networks. *International Symposium on Theory and Applications of Soft Computing.* Tokio.

Ozdemir, C., Cevik Onar, S., & Bagriyanik, S. (2020). *Estimating shopping center visitor numbers based on various environmental indicators.* Istanbul: Easy Chair Preprint.

Peng, L., Wang, L., Ai, X.-Y., & Zeng, Y.-R. (2020). Forecasting Tourist Arrivals via Random Forest and Long Short-term Memory. *Cognitive Computation*, 125-138.

Perdikaki, O., Kesavan S, & J., S. (2012). Effect of Traffic on Sales and Conversion Rates of Retail Stores. *Effect of Traffic on Retail Sales Performance Manufacturing & Service Operations Management*, 145–162.

Tanizaki, T., Hoshino, T., Shimmura, T., & Takenaka, T. (2018). Demand forecasting in restaurants using machine learning and statistical analysis. *12th CIRP Conference on Intelligent Computation in Manufacturing Engineering.* Naples.

Wikipedia. (2022). *Exploratory data analysis*. Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Exploratory_data_analysis

Yap, N., Gongy, M., Naha, R. K., & Mahanti, A. (2020). Machine Learning-based Modelling for Museum Visitations Prediction. *2020 International Symposium on Networks, Computers and Communications (ISNCC).* Montreal.