# Modelos de aprendizaje Supervisado para predecir la cantidad de pasajeros que saldrán de la Terminal de Transporte Norte de Medellín a otras regiones de País

Marcos Manuel Bru Diaz

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Tutor

Efraín Alberto Oviedo Carrascal, Magíster (MSc) en TICs

Universidad de Antioquia
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2021

| | |
|---|---|
| **Cita** | (Bru Diaz Marcos, 2021) |
| **Referencia**<br><br>**Estilo APA 7 (2020)** | (Bru Diaz Marcos, 2021) . Modelos de aprendizaje supervisado para predecir la cantidad de pasajeros que saldrán de la Terminal de transporte norte de Medellín a otras regiones del País [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia. |




Especialización en Analítica y Ciencia de Datos, Cohorte II.

## Dedicatoria

Quiero dedicarle este proyecto primero a Dios, quien en su ínfima misericordia y soberanía me dio el espíritu de sabiduría para afrontar cada reto que se me presento al largo de este proyecto. Así como también dedicarles este proyecto a mis padres Waldyr Bru García, y a mi madre Nelly Rosa Días Mejía, por su infinito apoyo y por su valentía para sacar adelante a sus hijos.
También quiero dedicarle este trabajo de grado a mi hija Isabella Bru Carreño, siempre estuvo para apoyar a su padre y entenderlo en los momentos que faltaba el tiempo.
También dedicatoria para mi compañera de camino Lina Fernanda Lopez Martínez, quien siempre estuvo presente para darme animo y apoyo en los momentos cruciales. Y en general dedicatoria a mi familia por su inmenso amor y a todos mis amigos.

## Agradecimientos

Agradecimientos infinitos a mi asesor Efraín Alberto Oviedo Carrascal, quien siempre estuvo atento a mis requerimientos, aportándome siempre con profesionalismo sus comentarios para culminar con éxito este proyecto, también quiero agradecerles a todos mis profesores de la Especialización en Analítica y Ciencia de Datos cohorte II, y en general a la Universidad de Antioquia por darme el privilegio de cursar mis estudios de especialización en esta Alma Mater.

# Tabla de contenido

Dedicatoria ..... 2
Agradecimientos ..... 2
Listas de tablas ..... 5
Listas de Figuras ..... 6
1. Resumen ejecutivo ..... 7
2. Descripción del problema ..... 8
2.1 Problema de negocio ..... 8
2.3 Origen de los datos ..... 10
2.4 Métricas de desempeño ..... 10
3. Datos ..... 11
3.1 Datos originales ..... 11
3.2 Dataset ..... 12
3.2.1 *Exploración de los datos* ..... 15
3.3 Descriptiva ..... 16
3.3.1 *Descriptiva datos después reactivación económica* ..... 21
4.Proceso de analítica ..... 27
4.1 Preprocesamiento ..... 28
4.1.2 *Preprocesamiento para primera iteración* ..... 28
4.1.3 *Preprocesamiento para segunda iteración* ..... 29
4.1.4 *Preprocesamiento para tercera iteración* ..... 29
4.1.5 *Transformación de los datos para todas las iteraciones* ..... 30
4.2 Modelos ..... 30
4.2.1 *Regresión Lineal* ..... 30
4.2.2 *Bosques Aleatorios (RandomForestRegressor)* ..... 31
4.2.3 *K vecinos más cercanos regresión (KNNeighborsRegressor)* ..... 31
4.3 Métricas ..... 31
5.Metodología ..... 32
5.1 Baseline ..... 32
5.2 Validación ..... 32
5.3 Iteraciones y Evolución ..... 33
5.3.1 *Iteración Baseline* ..... 33
5.3.2 *Primera iteración* ..... 34

5.3.3 *Segunda iteración* ..... 35
5.3.4 *Tercera iteración* ..... 36
5.3.5 *Cuarta iteración* ..... 37
5.4 Herramientas ..... 38
6.Resultados ..... 39
6.1 Métricas ..... 40
6.2 Consideraciones de producción ..... 41
7. Conclusiones ..... 42
Referencias ..... 43

## Listas de tablas

Tabla 1 Descripción datos originales ..... 11
Tabla 2 Columnas eliminadas del Dataset ..... 13
Tabla 3 Resultado exploración de las variables ..... 15
Tabla 4 Estadísticos 1 ..... 16
Tabla 5 Estadísticos 2 ..... 21
Tabla 6 Métricas Baseline evolución ..... 34
Tabla 7 Métricas primera Iteración evolución ..... 35
Tabla 8 Métricas segunda Iteración evolución ..... 36
Tabla 9 Métricas tercera Iteración evolución ..... 37
Tabla 10 Métricas cuarta Iteración evolución ..... 38
Tabla 11 Primera iteración: ..... 40
Tabla 12 Segunda iteración: ..... 40
Tabla 13 Tercera iteración: ..... 40
Tabla 14 Cuarta iteración: ..... 41

## Listas de Figuras

Figura 1 Formula matemática error absoluto medio ..... 10
Figura 2 Formula matemática error porcentual absoluto medio ..... 11
Figura 3 Filtro por Terminal de Transporte Medellín Norte ..... 12
Figura 4 Exportar Nuevo Dataset ..... 13
Figura 5 Lectura Dataset almacenados en GitHub ..... 13
Figura 6 Cambio formato columna fecha ..... 14
Figura 7 Fecha como índice y agrupación registros por día ..... 14
Figura 8 Dataset resultante ..... 15
Figura 9 Grafico comportamiento variable pasajeros en el tiempo ..... 17
Figura 10 Comportamiento variable despachos en el tiempo ..... 18
Figura 11 Grafico de dispersión variables DESPACHOS Y PASAJEROS ..... 18
Figura 12 Grafico de barras y boxplot variable PASAJEROS ..... 19
Figura 13 Grafico de barras y boxplot variable DESPACHOS ..... 20
Figura 14 Grafico comportamiento en el tiempo variable PASAJERO desde oct 2020 a sep 2021 ..... 22
Figura 15 Grafico comportamiento en el tiempo variable DESPACHOS desde oct 2020 a sep 2021 ..... 23
Figura 16 Grafico de barras y boxplot variable PASAJEROS desde oct 2020 a sep 2021 ..... 23
Figura 17 Grafico de barras y boxplot variable DESPACHOS desde oct 2020 a sep. 2021 ..... 24
Figura 18 Autocorrelación parcial variable PASAJEROS ..... 25
Figura 19 Autocorrelación parcial variable DESPACHOS ..... 26
Figura 20 fases del proceso de machine learning ..... 27
Figura 21 Normalización de los datos ..... 28
Figura 22 Código para crear columnas DAY , HOLIDAY ..... 29
Figura 23 Filtrar registros a partir de octubre de 2020 ..... 30
Figura 24 División de los datos entrenamiento y validación ..... 32
Figura 25 Datos para entrenamiento y validacion ..... 32
Figura 26 Grafico de los datos de entrenamiento y validación ..... 33
Figura 27 Grafico error medio absoluto de cada modelo baseline ..... 33
Figura 28 Grafico error medio absoluto de cada modelo primera iteracion ..... 34
Figura 29 Grafico error medio absoluto de cada modelo segunda iteración ..... 35
Figura 30 Grafico error medio absoluto de cada modelo tercera iteración ..... 36
Figura 31 Grafico error medio absoluto de cada modelo cuarta iteración ..... 37
Figura 32 Despliegue del modelo en Azure ..... 41

## 1. Resumen ejecutivo

El proyecto pretende a partir del análisis de datos, predecir la cantidad de pasajeros que se saldrán de la Terminal de Transporte Norte de Medellín en determinada fecha, empleando diferentes modelos de machine learning como LinearRegression, RandomForestRegressor, KNeighborsRegressor, los cuales basan sus predicciones , en dadas unas variables etiquetadas de entrada(predictoras) predecir un número , para realizar este análisis predictivo, se utilizó el Dataset por nombre "*Operación de pasajeros y despacho de vehículos en la modalidad de transporte de pasajeros por carretera"*, el conjunto de datos describe la salida de pasajeros por hora de las terminales del país desde 01 agosto de 2019 al 30 septiembre de 2021, como la finalidad de este proyecto es predecir salida de pasajeros por día que saldrán de la terminal del norte de Medellín, se procedió a agrupar los registros de cada día, y a filtrar solo por esta terminal, otras de las disposiciones fue eliminar algunas variables y solo quedándonos con las variables DESPACHOS Y PASAJEROS.

Pero aun así , nos enfrentábamos a un problema de series de tiempo, y como el objetivo de este trabajo se enfoca en realizar predicciones aplicando algoritmos de machine learning, se vio en la necesidad de transformar los datos de una serie de tiempo a un problema de aprendizaje supervisado, pero antes de esto se implementaron técnicas de análisis de series tiempo como la autocorrelación parcial , que nos da una medida de la correlación entre observaciones de una serie de tiempo que se encuentran separadas por k unidades de tiempo, esto nos dio una autocorrelación significativamente alta cuando se analizan las 7 observaciones anteriores, a partir de este análisis se procedió a transformar el Dataset para un problema de aprendizaje supervisado, etiquetando los 7 días de rezago en el tiempo en variables de entrada, y etiquetando como salida el 8.° día , de igual manera se estableció una escala entre 0 y 1 para las variables de entrada .

También se crearon dos columnas que, a partir de la fecha de cada registro, que nos permiten identificar qué día de la semana es y si es festivo o fin de semana, toda vez que analizado el mercado se encuentra que la dependiendo del día así es el comportamiento de la cantidad de pasajeros.

En la cuarta iteración donde se trabajaron con los registros a partir de octubre de 2020, el modelo que mejor se ajustó a los datos y obtuvo las mejores métricas fue el algoritmo de LinearRegression, dando como error absoluto medio 766, lo que nos indica la diferencia promedio entre el valor real y el estimado , expresado en porcentaje el error seria de 3.6 % ; durante todas las iteraciones fue el modelo con mejores métricas de desempeño con respeto a RandomForestRegressor y KNeighborsRegressor

## 2. Descripción del problema

### 2.1 Problema de negocio

La Terminal de Transporte del norte de la ciudad de Medellín, es una entidad de carácter público que tiene como objetivo principal prestar servicios adecuado a sus usuarios directos que serían las empresas transportadoras, así como también al usuario externo que serían pasajeros.

Ante esto, se detecta que el panorama actual está marcado por el deficiente servicio prestado por la terminal de transporte, debido que no cuentan con suficiente personal para la atención y guía adecuada de los pasajeros, así como tampoco cuentan con una planificación clara, en la cual se apoyen para mejorar servicios prestados dentro de la misma, como tampoco un adecuado plan de capacitación al personal que atienden a los pasajeros.

Por otro lado, también se encuentran deficiencias en la venta de productos y servicios por parte de los locales comerciales (cafeterías y almacenes) dentro de la terminal de norte, debido que muchas veces no cuentan con la capacidad operativa para atender la gran afluencia de pasajeros, en el caso de las cafeterías muchas veces se quedan sin disponibilidad de los productos que venden.

También se evidencia un déficit en el servicio de parqueadero que presta la terminal, debido que muchas veces por a la gran afluencia de pasajeros, la terminal se queda sin disponibilidad de parqueaderos, generando así mucha congestión dentro y fuera de la terminal, lo cual genera una percepción de desorden e inseguridad por parte de los usuarios.

Otros de los problemas presentados en el terminal cunado existen gran afluencia de pasajeros es la seguridad de los bienes de los usuarios, esto lo genera la baja disponibilidad de cuerpos de seguridad dentro y a las afueras de las instalaciones de la terminal de transporte.

Por tanto, tener una estimación de la cantidad de pasajeros que salgan de la terminal en cierta fecha, se podrían tomar medidas de mejoramiento de locaciones, restaurantes, casilleros para guardar equipaje, parqueaderos, suficiente personal capacitado, y no menos importante contar con un cuerpo de seguridad suficiente para contrarrestar cualquier situación de inseguridad.

## 2.2 Aproximación desde la analítica de datos

Desde la analítica de datos, se propone implementar diversos modelos de machine learning, para predecir la cantidad de pasajeros que saldrán la terminal de norte de Medellín en determinada fecha. El aprendizaje supervisado utiliza un conjunto histórico de datos, donde se tienen los registros previamente catalogados (entradas y salida), para crear un modelo de predicción, el cual aprende de los datos históricos hasta obtener la capacidad de predecir lo que pasará con nuevos conjuntos de datos, estos modelos de predicción se caracterizan por disponer de una variable objetivo, que es justamente lo que se quiere predecir, la cual para este proyecto es una variable numérica.

Pero al momento de implementar los modelos de aprendizaje supervisado al set de datos originales , nos encontramos que están organizados cronológicamente siendo así un problema de series de tiempo, como el objetivo de este trabajo se enfoca en realizar predicciones aplicando algoritmos de machine learning , se vio en la necesidad de transformar los datos de una serie de tiempo a un problema de aprendizaje supervisado, pero antes de esto, se implementaron técnicas de análisis de series tiempo como la autocorrelación parcial , que nos da una medida de la correlación entre observaciones de una serie de tiempo que se encuentran separadas por k unidades de tiempo, esto nos dio una autocorrelación significativamente alta cuando se analizan las 7 observaciones anteriores, a partir de este análisis se procedió a transformar el Dataset para un problema de aprendizaje supervisado, etiquetando los 7 días de rezago en el tiempo en variables de entrada , y etiquetando como salida el 8.° día.

El aprendizaje supervisado consta de tres fases:

1. Modelamiento. Consiste en construir el modelo que permita predecir la variable objetivo.
2. Evaluación. Se evalúa el modelo predictivo construido para ver qué tanto se puede confiar en él.
3. Validación. Una vez el modelo ha sido evaluado y el resultado es el esperado, se someten datos nuevos al modelo para realizar la predicción.

### 2.3 Origen de los datos

Para realizar las predicciones, se utilizó el data set por nombre "*Operación de pasajeros y despacho de vehículos en la modalidad de transporte de pasajeros por carretera"* , disponible en el portal https://www.datos.gov.co/ , el cual según la descripción en la misma, corresponde a los datos de movilidad de pasajeros y despachos de vehículos desde las terminales terrestres de pasajeros habilitadas y/o homologadas del país, presentados en el tablero de control publicado en el portal logístico de Colombia del Ministerio de Transporte. Del periodo de tiempo de 2019-08-01 al 2021-09-30.

### 2.4 Métricas de desempeño

Debido, que la variable que pretendemos predecir "PASAJEROS" es numérica, el error de predicción que arroja el modelo es la diferencia entre el valor real con el valor predicho.

Para evaluar los diferentes modelos predictivos implementados en este proyecto, se utilizaron dos métricas, que se usan comúnmente para evaluar y reportar el desempeño de un modelo de regresión, que son:

- **Error absoluto medio**: Es el promedio de la diferencia absoluta entre el valor observado y los valores predichos. El error absoluto medio o MAE es un puntaje lineal, lo que significa que todas las diferencias individuales se ponderan por igual en el promedio. Por ejemplo, la diferencia entre 10 y 0 será el doble de la diferencia entre 5 y 0. la formula matemática se describe en la Figura 1:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

*Figura 1 Formula matemática error absoluto medio*

- **Error Porcentual Absoluto Medio (MAPE o *Mean Absolute Percentage Error*)** es un indicador del desempeño que mide tamaño del error (absoluto) en términos porcentuales. El hecho de que el error sea estimado en porcentaje lo hace una métrica más fácil de entender para el usuario. la fórmula matemática se describe en la Figura 2:

$$MAPE = \frac{\sum_{t=1}^{n} \frac{|A_t - F_t|}{|A_t|}}{n}$$

*Figura 2 Formula matemática error porcentual absoluto medio*

## 3. Datos

### 3.1 Datos originales

Se cuenta con datos "*Operación de pasajeros y despacho de vehículos en la modalidad de transporte de pasajeros por carretera",* disponible en el portal https://www.datos.gov.co/, corresponde a los datos de movilidad de pasajeros y despachos de vehículos desde las terminales terrestres de pasajeros habilitadas y/o homologadas del país, presentados en el tablero de control publicado en el portal logístico de Colombia del Ministerio de Transporte. Del periodo de tiempo de 2019-08-01 al 2021-09-30. El cual cuenta con 10.435.336 registros y 10 columnas, en la Tabla 1, se describe cada columna.

*Tabla 1 Descripción datos originales*

| Nombre de Columna | Tipo | Descripción |
|---|---|---|
| TERMINAL | object | Corresponde a la sede de la terminal desde donde se realizan y reportan los despachos |
| CLASE_VEHICULO | object | Identificador numérico dado por el RUNT |
| NIVEL_SERVICIO | object | 1 BASICO - CORRIENTE 2 LUJO 3 CORRIENTE DIRECTO |
| MUNICIPIO_ORIGEN_RUTA | int64 | Código Divipola del municipio de origen de la ruta |
| MUNICIPIO_DESTINO_RUTA | int64 | Código Divipola del departamento de destino de la ruta |

| | | |
|---|---|---|
| FECHA_DESPACHO | object | Fecha de realización del despacho. date (DD/MM/YYYY) |
| HORA_DESPACHO | int64 | Hora 24H (hh) |
| TIPO_DESPACHO | object | 1 = Si el despacho de la ruta es en origen, y 2 = Si el despacho de la ruta es en transito |
| DESPACHOS | int64 | Número vehículos de transporte de pasajeros por carretera que salen desde la terminal |
| PASAJEROS | int64 | Número de pasajeros que abordaron viaje desde la terminal de transporte. Las terminales en tránsito sólo deben reportar los pasajeros que abordaron el vehículo en dicha terminal. |

### 3.2 Dataset

El archivo descargado en formato CSV, fue cargado a un Dataframe de pandas, a partir de ahí, se procedió a revisar el tamaño del mismo (filas y columnas ) :

```
Data=pd.read_csv('/gdrive/My Drive/Dataset proyecto especialización/Operaci_n_de_pasajeros_y_despacho_de_veh_culos_en_la_modalidad_de_transporte_de_pasajeros_por_carretera (1).csv')
 # Se lee DataSet Terminales
```

El data set original está conformado por 10.435.336 filas y 10 columnas

Como el objetivo de este proyecto se centró en predecir la cantidad de pasajeros que saldrán de la Terminal del Norte de Medellín ORIGEN, se seleccionó solo esta terminal de origen eliminando así los registros de las otras terminales del país Figura 3

```
1 # Se filtra por terminal de transporte de orrigen y se selcciona solo la Terminal de Transporte de Medellin Norte
2 Data=filtro_terminal_origen(Data,origen='T.T. DE MEDELLÍN NORTE')
```

*Figura 3 Filtro por Terminal de Transporte Medellín Norte*

Después de filtrar solo por la Terminal de transporte norte de Medellín, se procedió a exportar el Dataset, que se utilizara en todas las iteraciones del proyecto, se exporto como se puede ver en la Figura 4:

```
Data=df.to_csv('/gdrive/My Drive/Dataset proyecto especialización/Dataset_pasajeros_TT_norte_Medellin_limpio.csv')
```

*Figura 4 Exportar Nuevo Dataset*

Posteriormente se cargó el Dataset final en el repositorio de GitHub, se procedió cargándolos a un Dataframe de pandas con la siguiente línea de código Figura 5:

```
url= 'https://raw.githubusercontent.com/MarcusBruDiaz/Monografia/main/Dataset_pasajeros_TT_norte_Medellin_limpio.csv'
Data= pd.read_csv(url)
```

*Figura 5 Lectura Dataset almacenados en GitHub*

Se procedió a eliminar las variables relacionadas en la Tabla 2,en la columna "**Motivo eliminación**" se da una explicación del motivo por lo cual se tomó esta decisión para cada una de ellas:

*Tabla 2 Columnas eliminadas del Dataset*

| **Nombre de Columna** | **Tipo** | **Motivo eliminación** |
|---|---|---|
| TERMINAL | object | Se elimina toda vez que sabemos previamente que la terminal de transporte de origen es T.T. de Medellín Norte |
| CLASE_VEHICULO | object | Fue eliminada debido que el objetivo del trabajo es predecir la cantidad de pasajeros indiferente mente de que tipo de vehículos se transporten |
| NIVEL_SERVICIO | object | Fue eliminada debido que el objetivo del trabajo es predecir la cantidad de pasajeros indiferentemente del nivel de servicio del vehículo |
| MUNICIPIO_ORIGEN_RUTA | int64 | Se elimina teniendo en cuenta que solo se trata de un municipio de origen Medellín |
| MUNICIPIO_DESTINO_RUTA | int64 | Es indiferente para este proyecto el destino debido que el objetivo del |

| | | |
|---|---|---|
| | | trabajo es predecir la cantidad de pasajeros que saldrán de la terminal del norte |
| HORA_DESPACHO | int64 | Se elimina debido que se quiere predecir la cantidad de pasajeros por día, al momento de agruparlos solo queda un registro por día |
| TIPO_DESPACHO | object | Fue eliminada debido que el objetivo del trabajo es predecir la cantidad de pasajeros indiferentemente del tipo de despacho. |

Se evidencio que la columna FECHA_DESPACHO, es de tipo object, el cual presenta muchas incompatibilidades a la hora de trabajar con esta columna, por esto se decidió cambiar el formato como se muestra en la Figura 6. a uno más compatible con técnicas y funciones de la librería pandas:

```
1 # Se cambia tipo de formato columna "FECHA_DESPACHO"
2 Data["FECHA_DESPACHO"]= pd.to_datetime(Data["FECHA_DESPACHO"])
```

*Figura 6 Cambio formato columna fecha*

En Dataset original está conformado por varios despachos realizados en el mismo día, así que se procedió a agrupar por días y se sumaron las variables numéricas, incluida la variable PASAJEROS como se muestra en la Figura 7 , quien a su vez con la variable a predecir:

```
1 # Se estable la fecha como indice
2 Data = Data.set_index('FECHA_DESPACHO')
3 # se agrupan por dias los registros y se suman las variables numerias
4 Data=Data.resample('D').sum()
```

*Figura 7 Fecha como índice y agrupación registros por día*

Quedando así un Dataset final conformado por 2 columnas que son:DESPACHOS y PASAJEROS, con 792 registros, como se muestra en la siguiente Figura 8:

1 Data

| FECHA_DESPACHO | DESPACHOS | PASAJEROS |
|---|---|---|
| 2019-08-01 | 1986.0 | 24116.0 |
| 2019-08-02 | 2133.0 | 28542.0 |
| 2019-08-03 | 2218.0 | 32093.0 |
| 2019-08-04 | 1900.0 | 28381.0 |
| 2019-08-05 | 2084.0 | 28555.0 |
| ... | ... | ... |
| 2021-09-26 | 1557.0 | 19800.0 |
| 2021-09-27 | 1797.0 | 20965.0 |
| 2021-09-28 | 1718.0 | 17965.0 |
| 2021-09-29 | 1701.0 | 17383.0 |
| 2021-09-30 | 1720.0 | 17348.0 |

792 rows × 2 columns

*Figura 8 Dataset resultante*

### 3.2.1 *Exploración de los datos*

Se inicia explorando los tipos de cada unas de las variables, en la Tabla 3 . en la columna **TIPO** se puede ver la tipos de cada variable, en la columna # **Null** se registran la cantidad de registros vacíos , en la columna **Fecha final** se puede ver la primera fecha del Dataset y así mismo en la columna **Fecha final** se puede ver la última fecha del Dataset, en la columna **Falta alguna fecha entre inicial y final** se puede observar que los registros están completos entre la fecha inicial y la final no faltando ninguna fecha por registrar.

*Tabla 3 Resultado exploración de las variables*

Explorando las variables

| **Variable** | **TIPO** | **# Null** | **Fecha inicial** | **Fecha final** | **Falta alguna fecha entre inicial y final** |
|---|---|---|---|---|---|
| FECHA_DESPACHO | Object | 0 | 01-08-2019 | 30-09-2021 | NO |
| DESPACHOS | Flotat64 | 0 | | | |
| PASAJEROS | Flotat64 | 0 | | | |

### 3.3 Descriptiva

Con el fin de analizar el comportamiento de la información estadística relacionada con las variables en estudio “DESPACHOS” Y “PASAJEROS” se procedió al análisis de la estadística descriptiva para cada una de las variables como se muestra en la Tabla 4.

*Tabla 4 Estadísticos 1*

| | **DESPACHOS** | **PASAJEROS** |
|---|---|---|
| **count** | 792.000000 | 792.000000 |
| **mean** | 1465.151515 | 16054.343434 |
| **std** | 650.435540 | 10051.400196 |
| **min** | 15.000000 | 52.000000 |
| **25%** | 1253.000000 | 9089.750000 |
| **50%** | 1677.500000 | 15903.500000 |
| **75%** | 1944.000000 | 23171.250000 |
| **max** | 2696.000000 | 52447.000000 |
| **cv** | 44,39% | 62,61% |

Con base en la información estadística observada en la tabla anterior se procede al análisis descriptivo de las variables “DESPACHOS” y “PASAJEROS”, donde se cuenta con un total de 792 observaciones para cada una de las variables en estudio en el periodo comprendido entre 2019 08-01 y 2021-09-30.

Así mismo, se encuentra que la menor cantidad de despachos corresponde a 15 por día, mientras que para la variable “PASAJEROS” el valor mínimo es de 52 pasajeros por día. Por otra parte, se tiene que la cantidad de despachos máxima es de 2696, mientras que la mayor cantidad de pasajeros registrada es de 52447.

La mediana de los datos, correspondiente al cuartil 2 o 50% de los datos, muestra un valor de 1677.50 despachos y 15903.50 pasajeros, mientras que el primer cuartil arroja un valor de 1253 despachos y 9089.75 pasajeros.

Al analizar las medidas de variabilidad o dispersión que experimentan las variables, inicialmente se puede observar un valor promedio o media en los datos de 1465.15 en el número de despachos, mientras que el promedio o media en la cantidad de pasajeros es de 16054.34. Si analizamos la desviación estándar de los datos con respecto a la media, se puede observar un valor de 650.43 en los despachos y 10051.40 en el número de pasajeros.

Esta variabilidad es más evidente al estimar el coeficiente de variación de los datos, el cual muestra que la variabilidad en los datos con respectos a la media para la variable despachos es de 44,39% mientras que para la variable pasajeros es de 62,61%. Por tanto, se puede afirmar que la variable pasajeros presenta una mayor dispersión en los datos con respecto a la media.

**Pasajeros**:

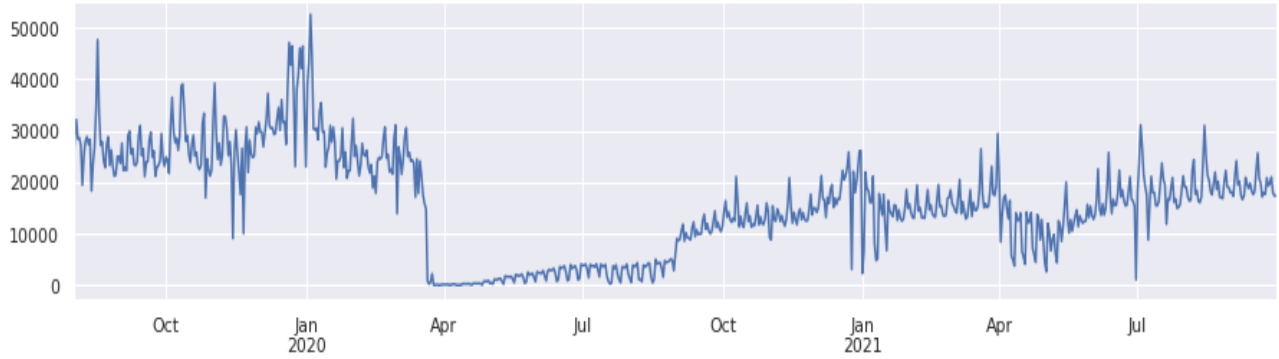


*Figura 9 Grafico comportamiento variable pasajeros en el tiempo*

En la gráfica anterior correspondiente a los datos de la variable “PASAJEROS”, en el periodo entre 2019 08-01 y 2021-09-30, se puede identificar un comportamiento entre los meses de agosto de 2019 y abril de 2020 en un rango promedio entre 20000 y 50000 despachos, sin embargo, la tendencia de los datos experimenta una caída abrupta a partir del mes de marzo de 2020 cuando la variable” PASAJEROS” cae a causa de la pandemia.

Entre mayo y septiembre el rango de datos no llega a superar los 10000 pasajeros, sin embargo, debido a las medidas del gobierno nacional a finales de septiembre e inicios de octubre de 2020, se empieza a ver un incremento en la cantidad de pasajeros.

**Despachos**:

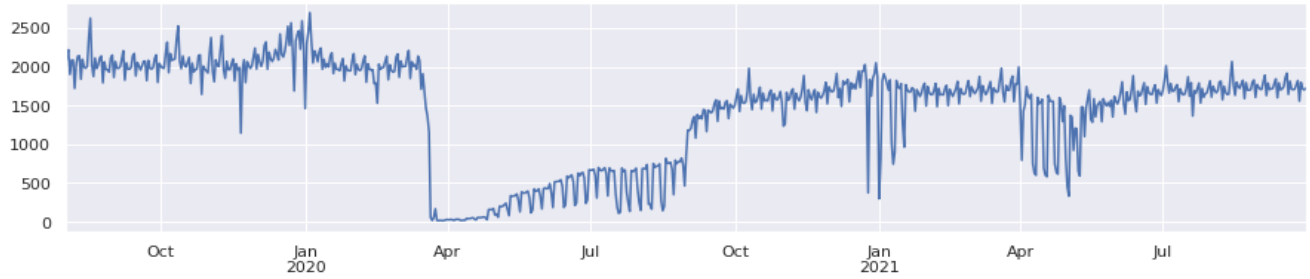


*Figura 10 Comportamiento variable despachos en el tiempo*

En cuanto a la variable "DESPACHOS", en el periodo entre 2019 08-01 y 2021-09-30, se puede evidenciar una tendencia igual que en la gráfica para la variable "PASAJEROS", donde el rango para el comportamiento de los datos varía entre 1500 y 2500 despachos aproximadamente. Sin embargo, se experimenta la caída abrupta en los datos a causa de la pandemia entre marzo y abril de 2020 como se observa en la gráfica. Entre mayo y septiembre el rango de datos no llega a superar los 1000 pasajeros.

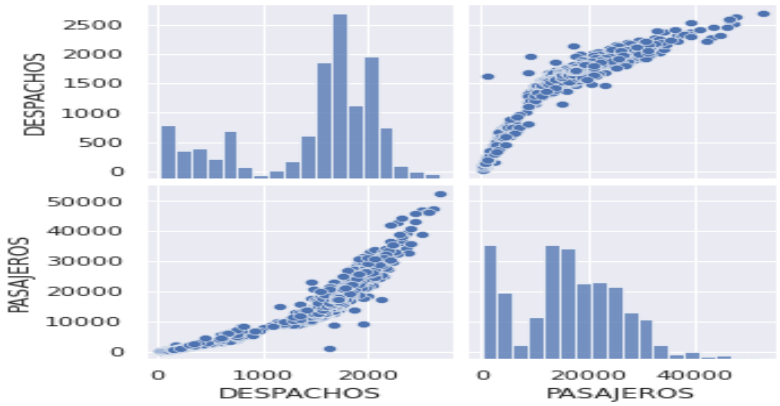


*Figura 11 Grafico de dispersión variables DESPACHOS Y PASAJEROS*

El grafico de dispersión muestra el comportamiento de los datos para las variables “DESPACHOS” y “PASAJEROS” con una tendencia creciente y directa, es decir, que a medida que incrementa la cantidad en una, también se evidencia un aumento en la otra variable, de igual manera se puede observar un buen agrupamiento de los datos a pesar, de que se pueden observar datos atípicos.

Esto se confirma con el indicador o coeficiente de correlación, el cual arroja un valor de 0.92, es decir, que existe una alta correlación directa entre las variables en cuestión.

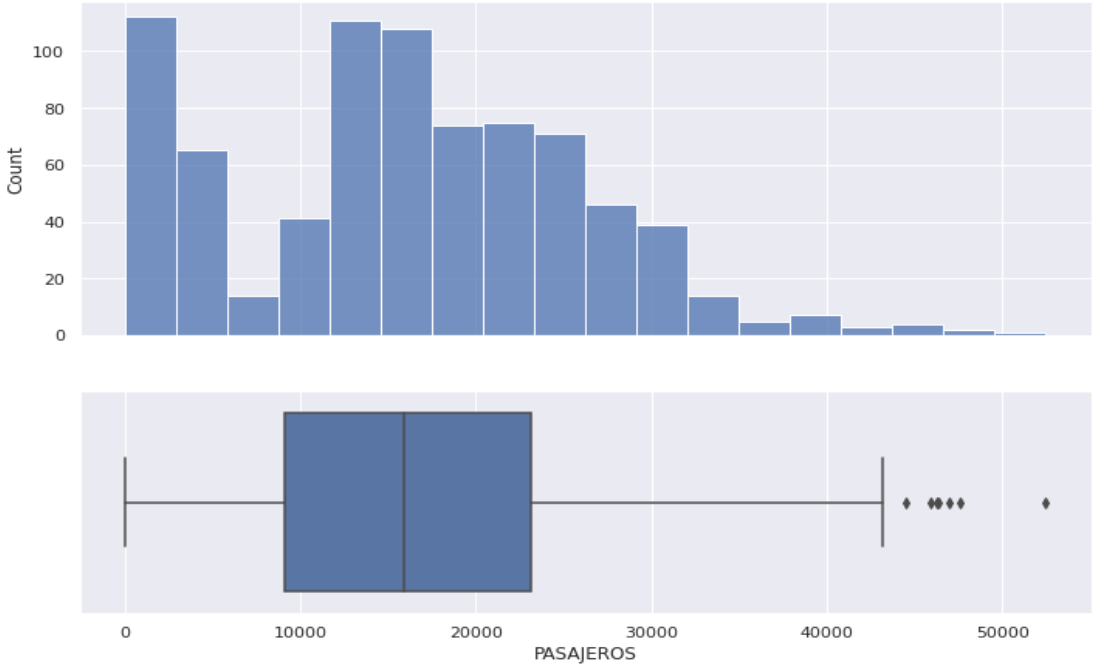


*Figura 12 Grafico de barras y boxplot variable PASAJEROS*

Con el histograma y grafico de cajas y bigotes o boxplot se puede relacionar el comportamiento de los datos correspondientes a la variable “PASAJEROS” conjuntamente, lo que da una idea más clara visualmente del comportamiento de los datos. El 50% de los datos correspondiente a esta variable es de 16054.34 pasajeros, que se puede observar en el boxplot con línea negra ubicada en el centro de la caja y en el histograma en la barra ubicada casi en medio de la distribución, de igual manera, se pueden observar la cantidad mínima de 52 pasajeros y máxima de 52447 pasajeros que corresponden en el boxplot a la delimitación de los bigotes inferior y superior.

Así mismo se puede ubicar el primer y tercer cuartil correspondiente al 25% de los datos que equivale a 9089.75 pasajeros y 75% de los datos que es de 23171.25 pasajeros, mediante el límite inferior y superior de la caja.

Mediante el análisis conjunto de los gráficos, también es posible identificar un comportamiento sesgado de los datos hacia la derecha.

En el boxplot se pueden observar los datos atípicos de manera más clara. Estos datos se consideran atípicos dado que se encuentran alejados del conjunto de datos o cantidad de pasajeros como se puede observar en el boxplot, para este caso, se ubican por encima del límite del bigote superior. Este comportamiento corresponde a que la cantidad de pasajeros, cuando no fue afectada por la pandemia, alcanzó altas cifras en cuanto a la cantidad de pasajeros, en comparación con el resto del periodo de tiempo en análisis.

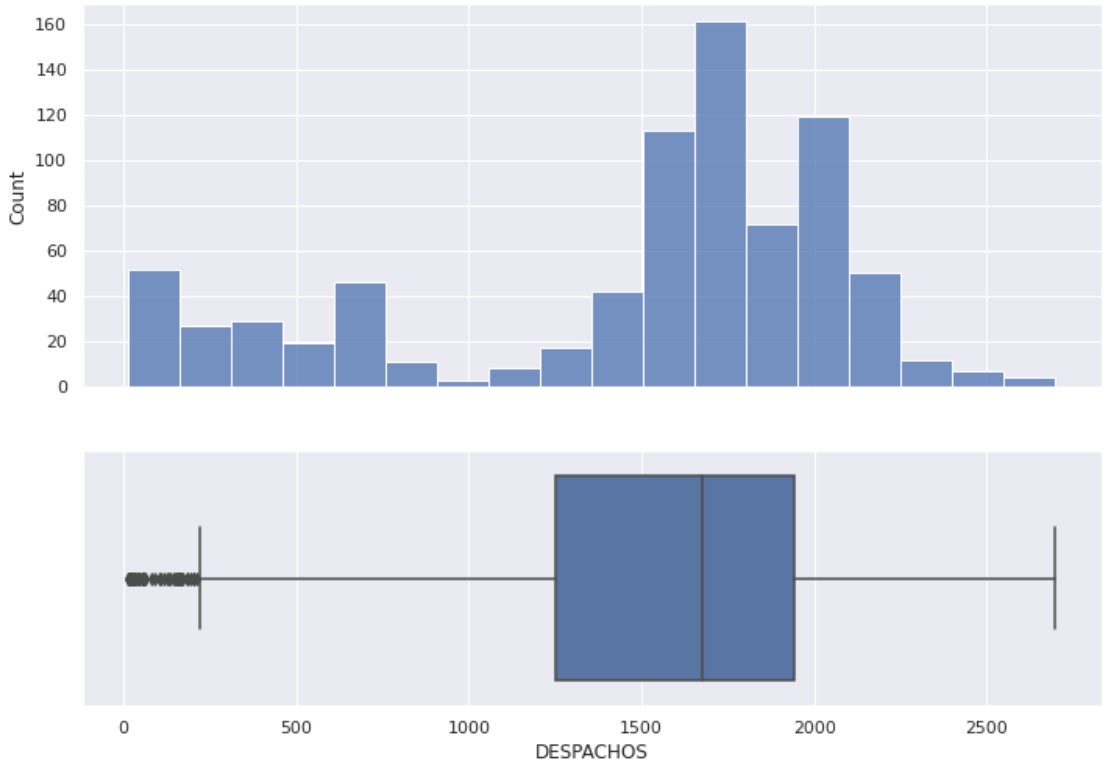


*Figura 13 Grafico de barras y boxplot variable DESPACHOS*

En cuanto al análisis del comportamiento de la variable "DESPACHOS" mediante el histograma y boxplot, se puede observar que los datos se encuentran sesgados a la izquierda, es decir, que se presenta una gráfica asimétrica hacia la izquierda.

De igual manera, se observa en el boxplot el 50% o mediana de los datos correspondientes a la cantidad de despachos levemente corrida a la derecha de la caja. Sobre los valores datos atípicos relacionados con la cantidad de despachos, se puede observar que a diferencia de la variable "PASAJEROS" se ubican por debajo del límite del bigote inferior, esto muestra que son datos muy bajos que se encuentran alejados del conjunto de datos agrupados o de la media y cuya explicación

se debe a que la cantidad de despachos durante el periodo afectado por la pandemia fue notoriamente bajo con respecto al resto del periodo de estudio analizado.

### 3.3.1 *Descriptiva datos después reactivación económica*

Análisis de datos a partir de octubre de 2020

A causa del impacto de la pandemia en el mundo y más exactamente por la influencia de esta en el comportamiento de los datos, se decidió analizar las variables “PASAJEROS” y “DESPACHOS” a partir del periodo de tiempo posterior a la pandemia, el cual se inicia desde el mes de octubre del 2020.

A continuación, se analiza la estadística descriptiva de las variables en estudio para el periodo de tiempo posterior a la pandemia, con el fin de evaluar el comportamiento de los datos y contrastar los resultados con el comportamiento durante la pandemia.

*Tabla 5 Estadísticos 2*

| | **DESPACHOS** | **PASAJEROS** |
|---|---|---|
| **count** | 365.000000 | 365.000000 |
| **mean** | 1601.526027 | 15624.909589 |
| **std** | 282.509608 | 4572.669502 |
| **min** | 298.000000 | 1125.000000 |
| **25%** | 1555.000000 | 13057.000000 |
| **50%** | 1666.000000 | 15506.000000 |
| **75%** | 1741.000000 | 18197.000000 |
| **max** | 2067.000000 | 31043.000000 |
| **cv** | 0,1764 | 0,2926 |

El análisis estadístico para las variables en estudio en el periodo posterior a la pandemia, reflejan una variación en los estadísticos estimados. Para 365 observaciones, se tiene que el valor mínimo de despachos corresponde a 298, mientras que el menor número de pasajeros reportados corresponde a 1125. Mientras que el valor máximo para despachos y pasajeros es de 2067 y 31043 respectivamente.

Al observar la media, se encontró que el promedio de despachos aumentó a 1601 y el promedio de pasajeros se ubica en 15624.

En cuanto a la variabilidad de los datos, se puede observar un coeficiente de variación para los despachos de 0,1764 y para la cantidad de pasajeros de 0,2926. Esto muestra una variabilidad de los datos con respecto a la media o promedio tanto para la variable "DESPACHOS" como para la variable "PASAJEROS" mucho menor que en el periodo de tiempo durante la pandemia analizado anteriormente.

**Pasajeros**

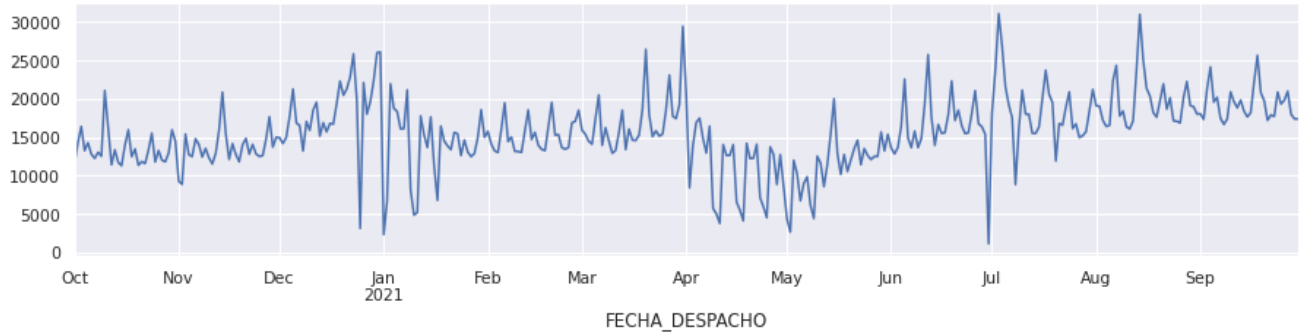


*Figura 14 Grafico comportamiento en el tiempo variable PASAJERO desde oct 2020 a sep 2021*

El grafico correspondiente a la variable "PASAJEROS" muestra un comportamiento con evidentes fluctuaciones en el periodo posterior a la pandemia, el cual se toma a partir de octubre del año 2020. Se puede observar en el mes de enero como luego de alcanzar un pico máximo, la tendencia cae drásticamente a un punto de los más bajos del periodo y aunque se recupera rápidamente, la tendencia fluctuante se mantiene y presenta un crecimiento gradual hasta alcanzar un pico alto en abril, donde luego cae nuevamente drásticamente la cantidad de pasajeros hasta mediados del mes de mayo cuando se comienzan a observar aumentos en la cantidad de pasajeros, sin embargo, en julio nuevamente cae drásticamente y se recupera fuertemente incluso alcanzando el pico máximo en el periodo de tiempo analizado.

## Despachos

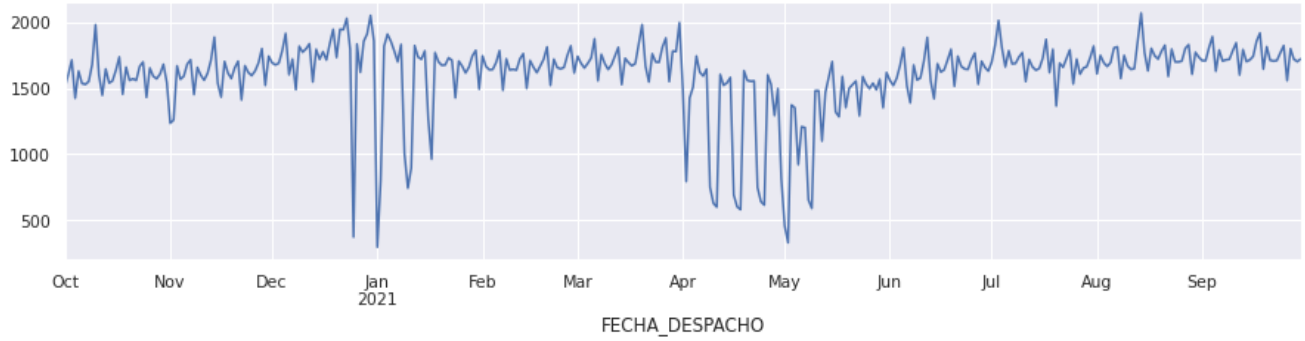


*Figura 15 Grafico comportamiento en el tiempo variable DESPACHOS desde oct 2020 a sep 2021*

En cuanto al comportamiento de los datos para la variable "DESPACHOS" se puede evidenciar que durante la mayoría del periodo de tiempo analizado los datos fluctúan en el rango entre los 1500 y 2000 despachos, sin embargo, se puede observar fluctuaciones con caídas en el mes de enero y durante los meses de abril y mayo, donde se observa la caída y recuperación en la cantidad de despachos de manera repetida hasta que desde el mes de junio, nuevamente retoma con una recuperación en la cantidad de despachos y un comportamiento estable en la tendencia de los datos.

## Pasajeros

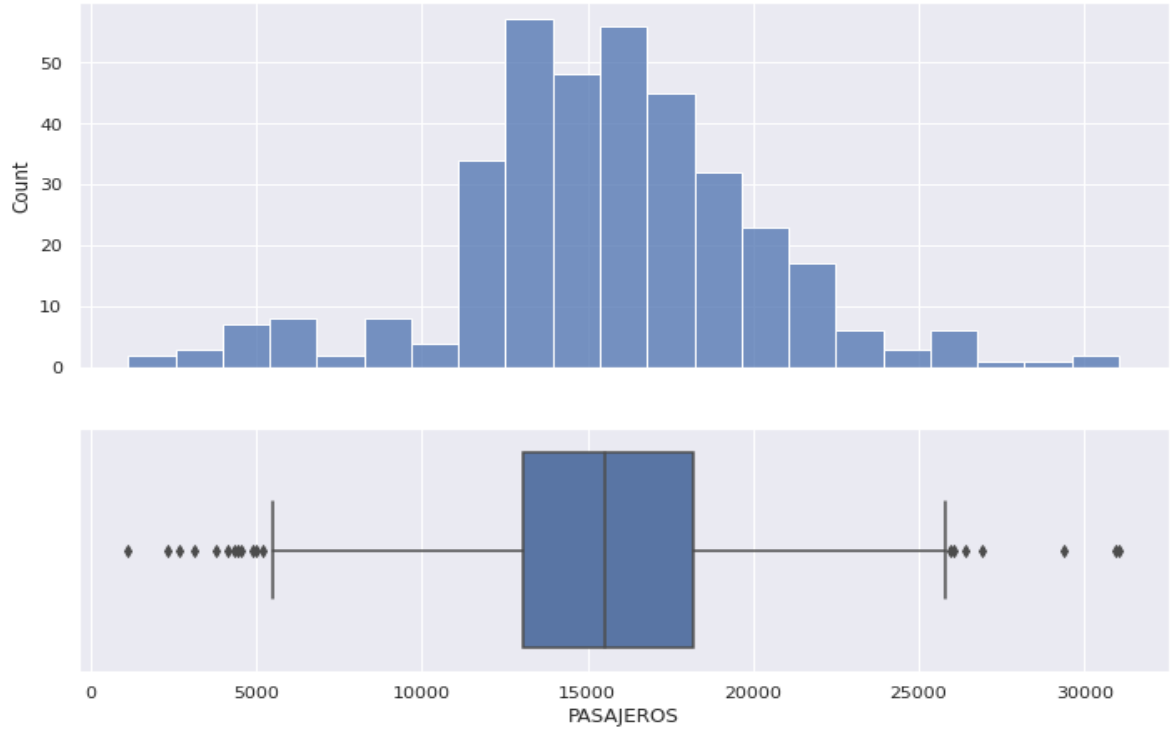


*Figura 16 Grafico de barras y boxplot variable PASAJEROS desde oct 2020 a sep 2021*

El histograma y boxplot nos permiten evidenciar el comportamiento de los datos y la distribución de estos en el periodo de estudio. Como se puede observar, la media, mediana y moda se encuentran en el punto medio del histograma, esto quiere decir que los datos tienden a seguir una distribución normal y se puede observar que el histograma tiende a formar una campana.

El boxplot muestra que la mediana se encuentra equidistante del cuartil 1 y 3 y que la caja se encuentra a su vez equidistante de los límites de los bigotes superior e inferior. Sin embargo, se pueden observar datos atípicos en el boxplot tanto por debajo del límite del bigote inferior como por encima del bigote superior, que corresponde a valores en la cantidad de pasajeros alejados del promedio y mediana de los datos.

Podemos observar que los datos en la gráfica están algo dispersos, sin embargo, al observar un coeficiente de variación de 29,26% se puede considerar aceptable y decir que la media de la cantidad de pasajeros es aceptablemente representativa.

**Despachos**

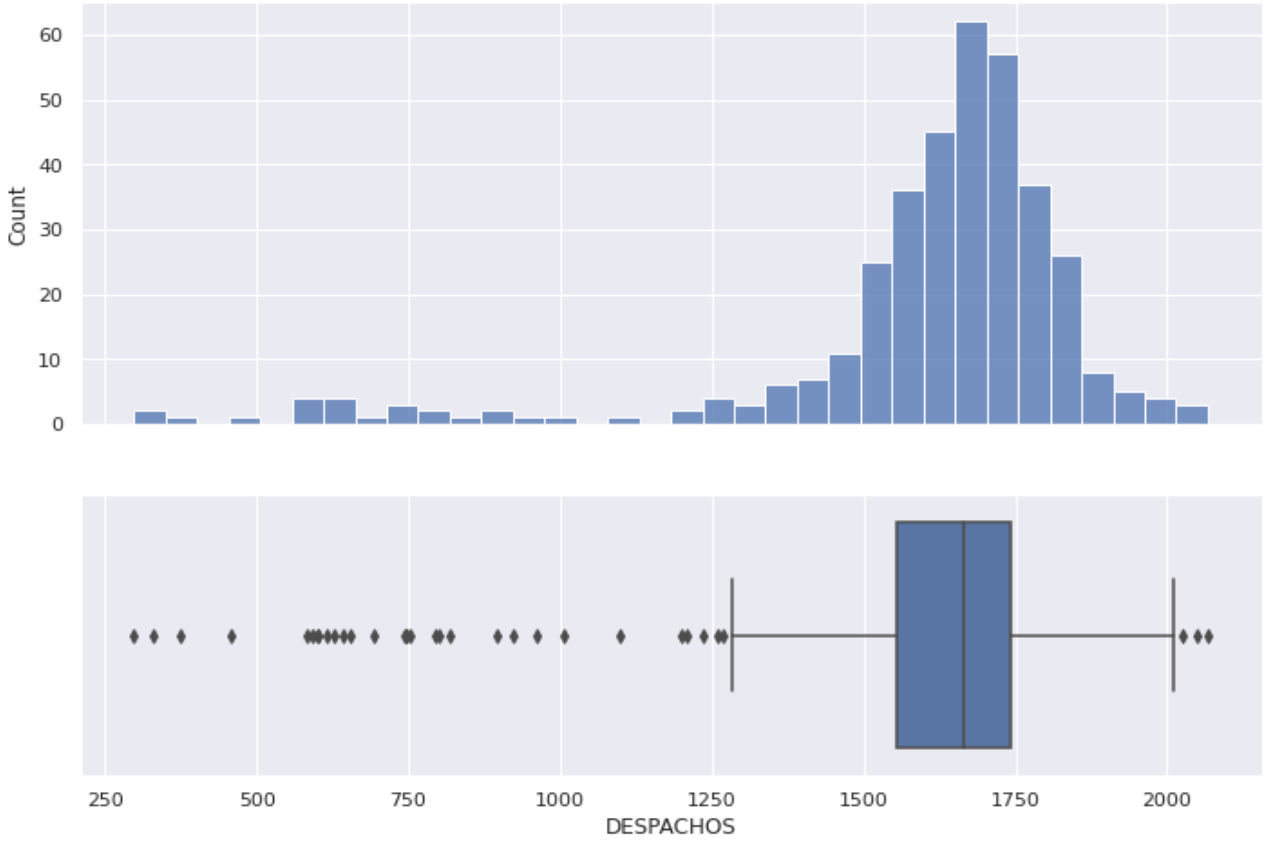


*Figura 17 Grafico de barras y boxplot variable DESPACHOS desde oct 2020 a sep. 2021*

En cuanto a la variable “DESPACHOS” se observa un histograma sesgado hacia la izquierda. Se puede ver que los datos se agrupan mayormente en el rango de valores entre 1500 y 2000 despachos.

En el gráfico de boxplot se puede observar datos atípicos por debajo del límite del bigote inferior en mayor medida que los datos atípicos ubicados por encima del límite del bigote superior.

Esto se explica porque a comienzos del mes de octubre el comportamiento de los datos para esta variable fue bajo debido a que empezaron a eliminarse restricciones, pero el aumento en la cantidad de despachos era gradual y por tanto las cantidades son notoriamente bajas en comparación con el periodo donde se agrupan mayormente los datos.

De igual manera, el comportamiento de los valores atípicos por encima del bigote superior muestra valores para la variable muy por encima que obedece a medidas y fechas específicas que incrementaron muy por encima del promedio las cantidades de despachos.

Es importante remarcar que aun en periodo de recuperación o reactivación de las actividades relacionadas con las variables en estudio, se presentan fluctuaciones por medidas del gobierno que afectan directamente en los valores provocando picos o declives como se puede observar en los gráficos anteriores.

Ahora, teniendo en cuenta que nuestros datos son una serie de tiempo, y con el fin de transformarlos para un problema de aprendizaje supervisado, se implementaron técnicas de análisis de series tiempo como la autocorrelación parcial , que nos da una medida de la correlación entre observaciones de una serie de tiempo que se encuentran separadas por k unidades de tiempo, se analizó en la Figura 18 la correlación parcial de la variable PASAJEROS

**Pasajeros**

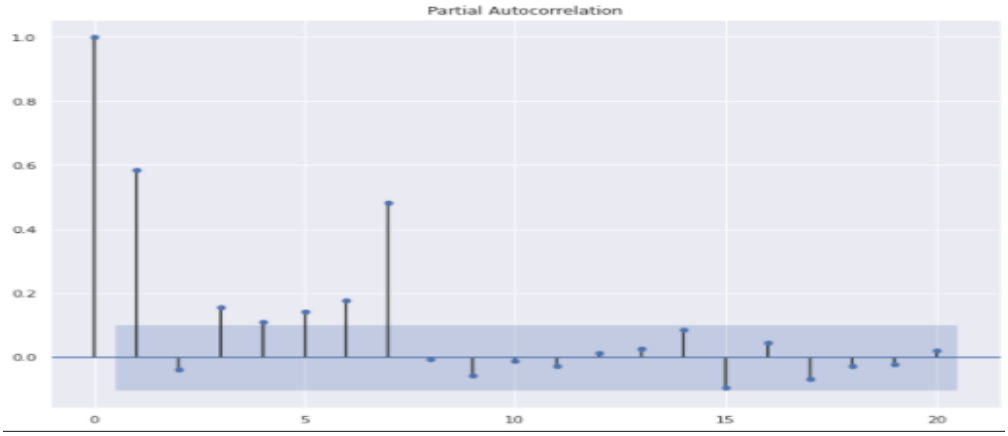


*Figura 18 Autocorrelación parcial variable PASAJEROS*

Esta grafica nos muestra que en el desface 7 hay un pico alto que sobrepasa los límites de significancia, indicándonos que existe una alta correlación de las 7 observaciones anteriores que en nuestro caso serian 7 días anteriores debido que nuestra unidad de tiempo es en días.

De igual manera se hace el análisis para la variable DESPACHOS, también mostrándonos la existencia de una alta correlación en el desface 7, Figura 19.

**Despachos**

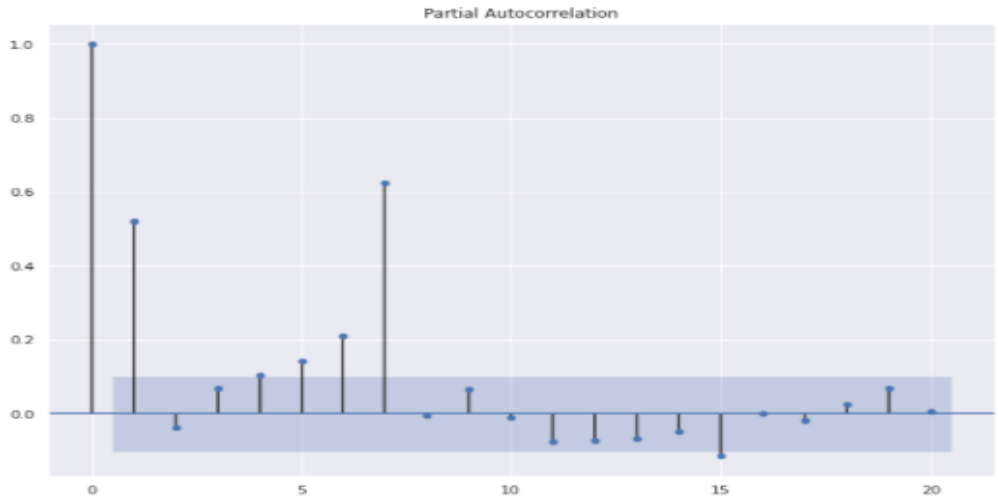


*Figura 19 Autocorrelación parcial variable DESPACHOS*

**4.Proceso de analítica**

En esta parte se muestra un diagrama de las fases en las que generalmente se ven inmersos los proyectos de machine learning Figura 20:

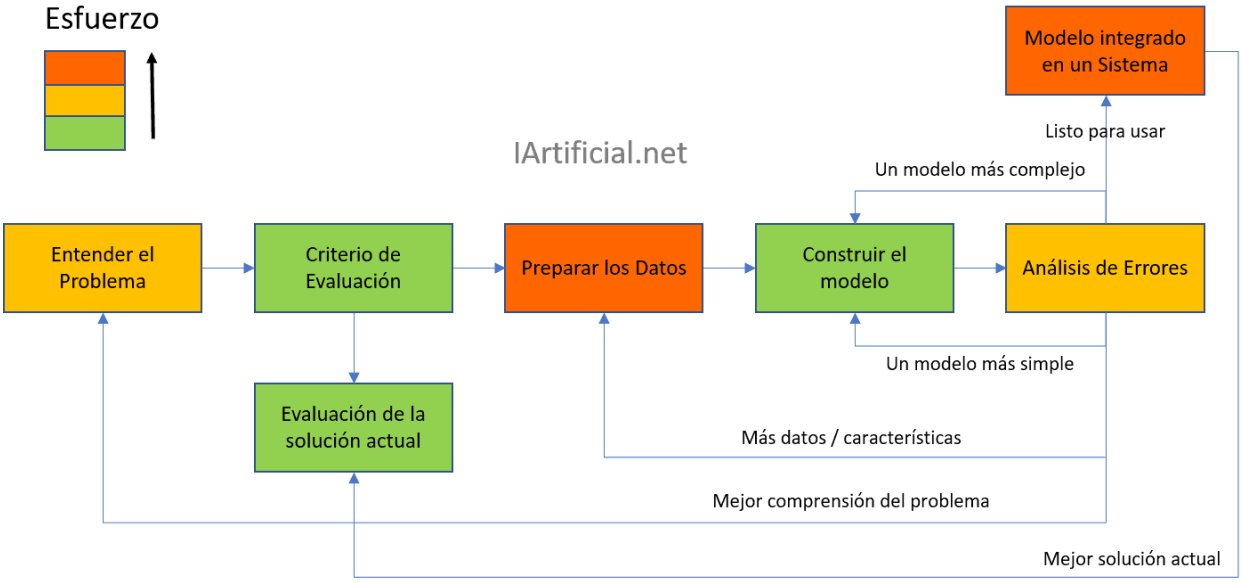


*Figura 20 fases del proceso de machine learning*

A partir del esquema de la Figura 20 se comienza a explicar cada fase de nuestro proyecto:

1. **Entender el problema:** Es de vital importancia entender a qué problemas nos enfrentamos y desde el análisis de datos como darles unas soluciones, por eso es importante entender cómo funciona el negocio y hacernos preguntas de como a partir de los datos podemos solucionar.
2. **Criterio de evaluación:** En importante definir métricas de evaluación de los modelos para conocer si está funcionando bien según el problema de negocio y cómo evoluciona, esto entiendo que lo que no se mide no se controla. Es por esto que para nuestro modelo se utilizaron dos métricas de desempeño que son **error absoluto medio** y el **e**rror **Porcentual Absoluto Medio**
3. **Preparar los datos:** Esta fase en nuestro proyecto se basó en cargar los datos, explorarlos, conociendo así sus características, así como también haciendo un análisis descriptivo de cada una de las variables, también en este punto se hizo algo de preprocesamiento de datos

como cambiar formatos de las variables, agrupar registros por fechas, normalización de los datos, y creación de nuevas características o variables predictoras

4. **Construir el modelo:** en esta fase del proyecto, es importante definir los modelos que más se ajusten al objetivo que tengamos, para nuestro caso se construyeron modelos de regresión que hacen parte de la librería de Sklearn que son LinearRegression, RandomForestRegressor, KNeighborsRegressor.
5. **Análisis de Error:** En esta parte se utiliza para análisis la capacidad predictiva de nuestro modelo, y si es capaz de generalizar y tener buenas métricas con datos nuevos, también esta fase del análisis de error nos permite saber la necesidad de hacer experimentos o iteraciones nuevas con modificaciones significativas en búsqueda de mejorar las métricas de desempeño.
6. **Modelo integrado a un sistema:** Una vez tengamos un modelo con métricas de desempeño aceptables para el negocio, es necesario integrarlo con un sistema transaccional.

## 4.1 Preprocesamiento

La etapa de preparación de los datos o preprocesamiento consiste en analizar y aplicar diversas operaciones de limpieza de datos, es decir, corregir las inconsistencias en los tipos de datos, atributos mal escritos, datos duplicados, cómo tratar los datos faltantes o nulos, transformación, codificaciones, entre otros

### 4.1.2 *Preprocesamiento para primera iteración*

En la primera iteración y subsiguientes, se decidió normalizar los datos en un rango de 0 a 1, con esto se prevé que, aunque algunas características sean muy grandes los modelos no las usaran como predictor principal, se decidió esto toda vez que los 7 días anteriores se convertirían en variables predictoras y los valores de cada día podían estar en rangos muy distintos. Para este escalado se utilizado el código de la Figura 21:

```
# SE ESTANDARIZAN LOS DATOS
X= Data.drop(['PASAJEROS'],axis=1)
scaler = MinMaxScaler(feature_range=(0, 1))
X[:] = scaler.fit_transform(X)
X["PASAJEROS"] = Data_iteracion1["PASAJEROS"]
```

*Figura 21 Normalización de los datos*

### 4.1.3 *Preprocesamiento para segunda iteración*

Para la segunda iteración y posteriores, se utilizaron las variables predictoras del numeral anterior, pero se agregaron nuevas variables, debido que se realizaron algunas investigaciones, arrojando que muchas personas viajan los fines de semana o en su defecto los días festivos, por lo anterior se implementó con el código descrito en la Figura 22

```
Fechas=list(Data.FECHA_DESPACHO)
DiaSemana=[]
Holiday=[]
for i in range(len(Fechas)):
  dt=Fechas[i]
  year, month, day = (int(x) for x in dt.split('/'))
  ans = datetime.date(year, month, day)
  if holidays_co.is_holiday_date(ans)==True:
    Holiday.append(1)
  else:
    Holiday.append(0)
  if ans.weekday()==6:
    Holiday[i]=2
  if ans.weekday()==5:
    Holiday[i]=3
  DiaSemana.append(ans.weekday())
```

Se crea dos colunas con la que se identifica el dia y se es fin desemna o festivo

```
[355]  Data.insert(1,'Day',DiaSemana,True)
       Data.insert(2,'Holiday',Holiday,True)
```

*Figura 22 Código para crear columnas DAY , HOLIDAY*

La cual, a partir de la fecha, puede identificar qué día es, si es fin de semana o festivo, se crearon así dos columnas:

**Día**: quien a su vez tiene las siguientes clasificaciones 0: lunes, 1: martes ,2: miércoles, 3: jueves, 4: viernes, 5: sábado, 6: Domingo.

**Holiday**: quien a su vez tiene las siguientes clasificaciones 1: Festivo, 2: Domingo, 3: sábado

### 4.1.4 *Preprocesamiento para tercera iteración*

No obstante, a partir de la tercera iteración se procedió a trabajar con datos a partir del mes octubre de 2021, teniendo en cuenta las disposiciones del análisis descriptivo de los datos, realizados en el punto anterior DESCRIPTIVA DE LOS DATOS. este filtro se realizó con el código dispuesta en la Figura 23:

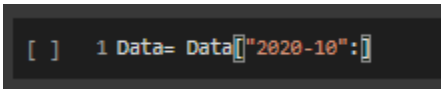


*Figura 23 Filtrar registros a partir de octubre de 2020*

### 4.1.5 *Transformación de los datos para todas las iteraciones*

Se realizó una transformación del Dataset para cada una de las iteraciones, de tal manera que analizando los 7 días de rezago de cada variable se predijera la cantidad de pasajeros del día siguiente, con esta transformación se creaban para el caso del Baseline y la primera iteración un Dataset de las siguientes con dimensiones de 785 filas y 16 columnas.

Para la segunda iteración aplicando la misma transformación quedo un Dataset con las siguientes dimensiones de 785 filas y 32 columnas.

Para la tercera iteración y subsiguientes, se aplicaron las mismas transformaciones quedo un Dataset con las siguientes dimensiones de 358 filas y 32 columnas.

Estas trasformaciones, fueron aplicadas, para que así no se tratase de un problema de series de tiempo, sino más bien de un problema de aprendizaje supervisado y así implementar modelos de machine learning.

## 4.2 Modelos

Como el objetivo es predecir la cantidad de pasajeros que se movilizaran en la Terminal de Transporte Norte de Medellín, y este corresponde a un valor numérico. Para este objetivo, se tiene que los modelos de Machine Learning dispuestos para este fin son los modelos de regresión, los cuales hacen parte de los modelos supervisados, por lo anterior, los modelos utilizados en este proyecto fueron los siguientes:

### 4.2.1 *Regresión Lineal*

Para un modelo regresión lineal, tenemos unas variables predictoras(explicativas) que en nuestro caso serían los 7 días anteriores de cada teniendo en cuenta lo descrito en el numeral 4.1.4, así como también se cuenta con una variable objetivo(resultado), con estos datos el modelo de regresión lineal intenta encontrar la relación entre las diferentes variables que nos permite predecir un resultado continuo.

Este proceso de predecir se trata, que dadas unas variables predictoras (x) y la variable objetivo (Y), el modelo calcula una recta que minimice la distancia entre los puntos de muestra y la recta ajustada, utilizando algunos métodos para estimar los coeficientes (como mínimos cuadrados o gradiente decreciente). Después, utilizaremos la recta obtenida para predecir el resultado con datos nuevos.

### 4.2.2 *Bosques Aleatorios (RandomForestRegressor)*

Un Ramdom forest es un conjunto de árboles combinados entre si con bagging, al usar bagging, lo que sucede, es que los distintos arboles del conjunto, ven distintas porciones de datos, es importante recalcar, que ningún árbol ve todos los datos del set de entrenamiento, esto hace que cada árbol se entrene con muestras de datos distintas para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor. (Heras, iartificial, 2020)

No obstante, este modelo tiene la tendencia de sobre ajustar, esto quiere decir que se aprende muy bien los datos de entrenamiento, y al momento de entregarle datos de validación no es capaz de generalizarlos, arrojando así métricas menos favorables en los datos nuevos

### 4.2.3 *K vecinos más cercanos regresión (KNNeighborsRegressor)*

Un método de aproximación simple no paramétrica es el basado en la regla del vecino más cercano, que consiste en estimar el valor de un dato desconocido a partir de las características del dato más próximo, según una medida de similitud o distancia, el método del vecino más cercano se puede extender utilizando no uno, sino un conjunto de datos más cercanos para predecir el valor de los nuevos datos, en lo que se conoce como los k-vecinos más cercanos (k-NN o k-Nearest Neighbors) (Torres, 2008, pág. 2)

## 4.3 Métricas

De la librería Metrics de scikit-learn, se utilizó la métrica de MEAN_ADSOLUTE_ERROR, y también se utilizó la funciona mean_adsolute_porcentage_error , la cual expresa el error obtenido en porcentaje.

# 5.Metodología

## 5.1 Baseline

Se escogen tres algoritmos de regresión que son: LinearRegression, RandomForestRegressor y KNeighborsRegressor con sus hiperparametros por defecto para tener un valor base (Baseline), esto con el fin de tener un punto de referencia y así compararlo con otros modelos más robustos y búsqueda de hiperparametros.

## 5.2 Validación

Una operación que es común en todos los modelos de aprendizaje supervisado es la división de nuestro conjunto de datos, en -al menos- dos partes: una parte *Train*, de entrenamiento, que corresponderá a la mayor parte de nuestro Dataset y que usaremos para entrenar nuestro modelo y un parte *Test*, de menor tamaño, sobre la que evaluaremos nuestro modelo entrenado.

Esta división **se puede realizar de forma aleatoria o lineal**, y siempre elegimos qué porcentaje queremos para cada división, para nuestro caso en particular se realizó la división de los datos de forma lineal. (Gliese710, 2019)

Para implementar lo anterior, se utilizó la función **train_test_split** , a la cual se le pasa la Data, y con el parámetro **test_size** se le indica al porcentaje de datos que se va a dejar para test, dejando así para entrenamiento de los modelos 80 % y para test el 20 % , para que la división sea lineal se utilizó el parámetro **shuffle** iniciado en **False. Como se muestra en la** Figura 24**:**

```
1 # DIVISION DE LOS DATOS
2 Train, Test = train_test_split(Data, test_size = 0.20, shuffle = False)
```

*Figura 24 División de los datos entrenamiento y validación*

Después de tener los datos de entrenamiento se crearon los datos **X_Train** sin la variable objetivo y **y_Train** correspondiente a la variable objetivo, así como también se creó para el test X**_Test** sin la variable objetivo y **y_Test** correspondiente a la variable objetivo, dejando así los dos primeros sets de datos para entrenamiento y los últimos dos para prueba.

```
1 # SE CREAN LOS DATOS DE ENTRENAMIENTO Y TEST
2 X_Train = Train.drop(columns='PASAJEROS')
3 y_Train = Train['PASAJEROS']
4 X_Test= Test.drop(columns='PASAJEROS')
5 y_Test= Test['PASAJEROS']
```

*Figura 25 Datos para entrenamiento y validacion*

Se muestra en la siguiente Figura 26 los datos de entrenamiento**(negro)** y de prueba**(rojo)**

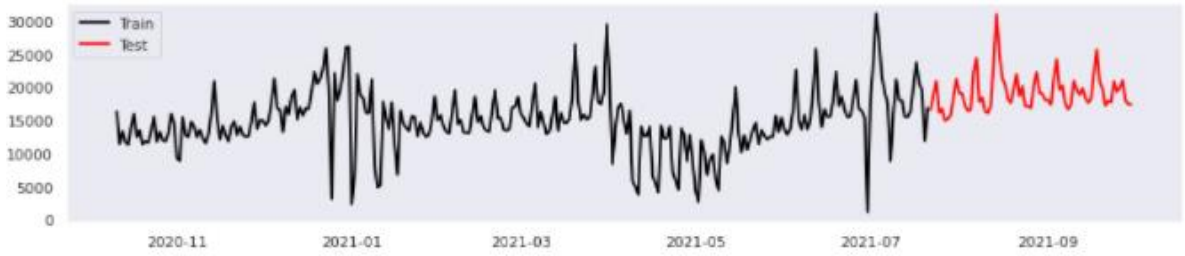


*Figura 26 Grafico de los datos de entrenamiento y validación*

## 5.3 Iteraciones y Evolución

### 5.3.1 *Iteración Baseline*

En la ejecución de le Baseline, se utilizaron como variables de entrada los valores de los 7 rezagos en el tiempo de la variable DESPACHOS Y PASAJEROS, con el fin de predecir la variable objetivo que es el 8. ° día, utilizando todos los registros del Dataset, se obtuvo los siguientes resultados de las métricas Figura 27:

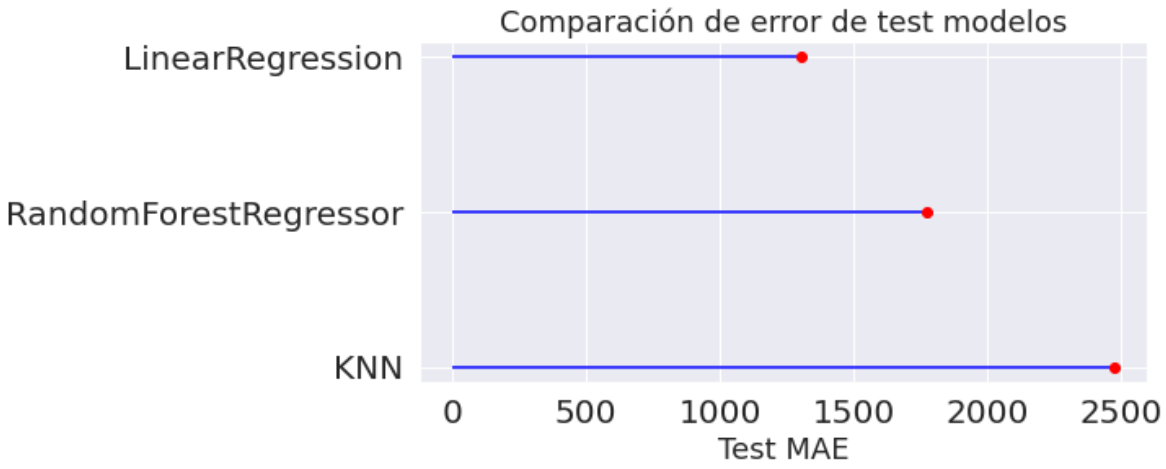


*Figura 27 Grafico error medio absoluto de cada modelo Baseline*

*Tabla 6Métricas Baseline evolución*

Métricas Baseline tabuladas para cada uno de los modelos

| Modelo | Error Medio Absoluto (MAE) | Error Porcentual Absoluto Medio **(MAPE)** |
|---|---|---|
| LinearRegression | 1306 | 16 % |
| RandomForestRegressor | 1794 | 17.12 % |
| KNeighborsRegressor | 2478 | 25.18 % |

### 5.3.2 *Primera iteración*

En un primea iteración de los modelos, se utilizaron las mismas disposiciones del numeral anterior, pero esta vez se escalaron los datos, con lo cual se obtuvo los siguientes resultados de las métricas Figura 28:

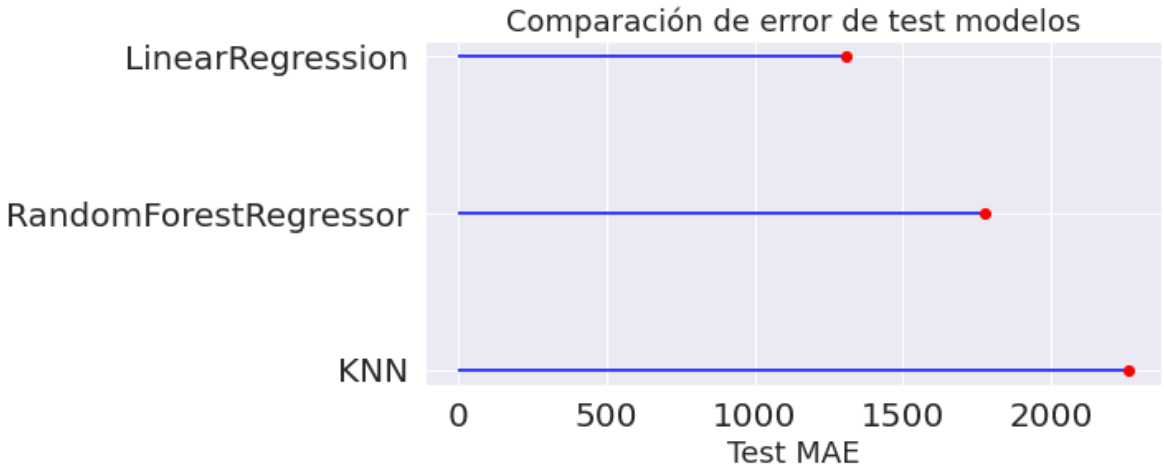


*Figura 28 Grafico error medio absoluto de cada modelo primera iteracion*

*Tabla 7Métricas primera Iteración evolución*

Métricas primera Iteración tabuladas para cada uno de los modelos

| Modelo | Error Medio Absoluto (MAE) | Error Porcentual Absoluto Medio **(MAPE)** |
|---|---|---|
| LinearRegression | 1306 | 16 % |
| RandomForestRegressor | 1769 | 17 % |
| KNeighborsRegressor | 2262 | 20.8 % |

### 5.3.3 *Segunda iteración*

En una segunda iteración, se utilizaron las mismas disposiciones del numeral anterior, pero esta vez, se implementaron, con algo de ingeniera de características, que consistió en agregar dos nuevas variables predictoras que fueron DAY que identifica que día de la semana es y Holiday que clasifica si es sábado(3), domingo(2) o festivo (1), y en el caso de que se tratara de un día de semana lo clasifica como 0, con lo cual se obtuvo los siguientes resultados de las métricas Figura 29:

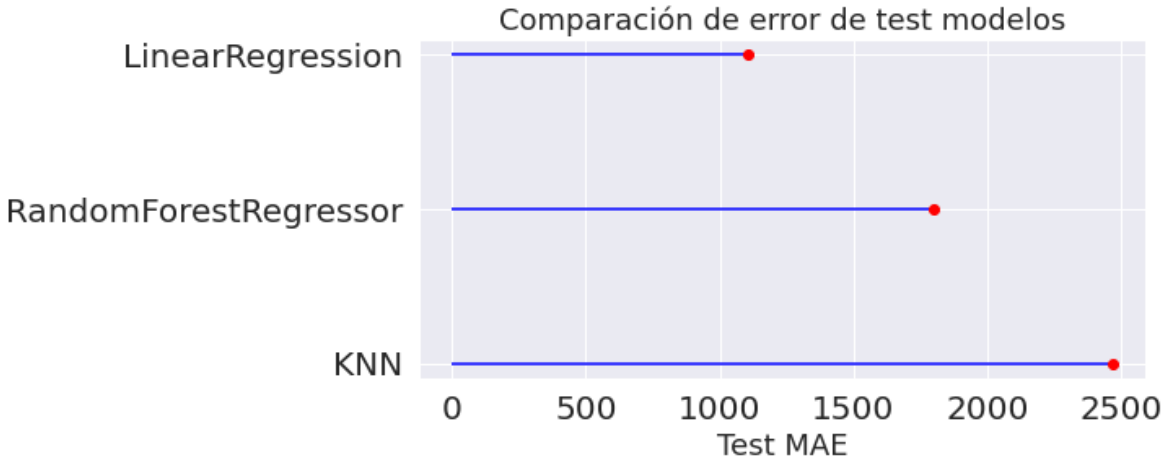


*Figura 29 Grafico error medio absoluto de cada modelo segunda iteración*

*Tabla 8Métricas segunda Iteración evolución*

Métricas segunda Iteración tabuladas para cada uno de los modelos

| Modelo | Error Medio Absoluto (MAE) | Error Porcentual Absoluto Medio **(MAPE)** |
|---|---|---|
| LinearRegression | 914 | 5.46 % |
| RandomForestRegressor | 1565 | 8 % |
| KNeighborsRegressor | 2233 | 12 % |

### 5.3.4 *Tercera iteración*

En la tercera iteración, se utilizaron las mismas disposiciones del numeras anterior, pero esta vez, se trabajó los modelos con los datos a partir de octubre de 2020, con lo cual se obtuvo los siguientes resultados de las métricas Figura 30:

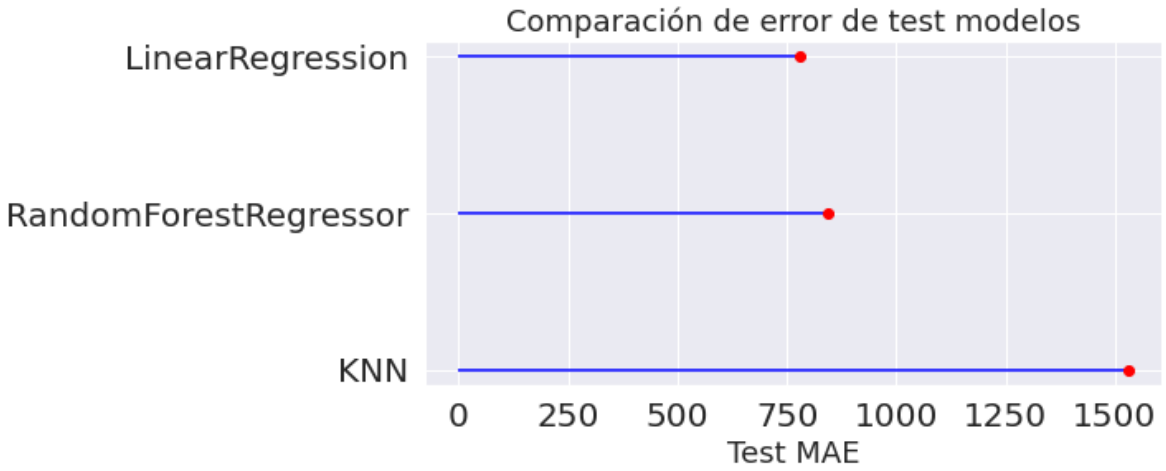


*Figura 30 Grafico error medio absoluto de cada modelo tercera iteración*

*Tabla 9Métricas tercera Iteración evolución*

Métricas tercera Iteración tabuladas para cada uno de los modelos

| Modelo | Error Medio Absoluto (MAE) | Error Porcentual Absoluto Medio **(MAPE)** |
|---|---|---|
| LinearRegression | 780 | 4.12 % |
| RandomForestRegressor | 928 | 4.72 % |
| KNeighborsRegressor | 1531 | 8.12 % |

### 5.3.5 *Cuarta iteración*

En la cuarta iteración, se utilizaron las mismas disposiciones del numeras anterior, con las mismas características en el Dataset, pero esta vez, en el modelo de LinearRegression se implementó con la técnica de validación cruzada, y para los modelos de RandomForestRegressor y KNeighborsRegressor, se implementó igual validación cruzada y búsqueda de hiperparametros, dando como resultado lo siguiente Figura 31:

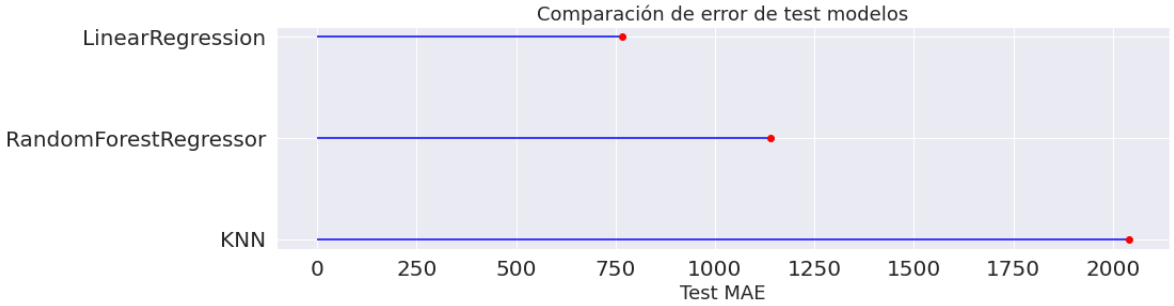


*Figura 31 Grafico error medio absoluto de cada modelo cuarta iteración*

*Tabla 10Métricas cuarta Iteración evolución*

Métricas cuarta Iteración tabuladas para cada uno de los modelos

| Modelo | Error Medio Absoluto (MAE) | Error Porcentual Absoluto Medio **(MAPE)** |
|---|---|---|
| LinearRegression | 765 | 3.92 % |
| RandomForestRegressor | 1139 | 5.58 % |
| KNeighborsRegressor | 2041 | 10 % |

## 5.4 Herramientas

La herramienta principal fue la extensión de Colab de Google, la cual nos permitió interactuar con las siguientes librerías, lenguajes de programación y herramientas:

**Numpy**: es una librería para realizar cálculo numérico en Python. La usaremos principalmente porque nos permite crear y modificar matrices, y hacer operaciones sobre ellas con facilidad.

**Python**: es un lenguaje de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código, se trata de un lenguaje de programación multiparadigma, ya que soporta parcialmente la orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, dinámico y multiplataforma, es uno de los lenguajes más populares en la analítica de datos

**Matplotlib**: es una biblioteca para la generación de gráficos a partir de datos contenidos en listas o arrays en el lenguaje de programación Python y su extensión matemática Numpy.

Sklearn: es una biblioteca para aprendizaje automático de software libre para el lenguaje de programación Python.1, se utilizaron de esta librería los modelos y las meticas.

**Pandas**: Es una librería de software libre para la manipulación de datos en Python, es muy frecuente en el análisis de datos, es capaz de trabajar con múltiplos formatos de origen como csv o xls, en nuestro proyecto la utilizamos para cargar los datos, manipularlos para posterior análisis con los modelos supervisados.

## 6.Resultados

Analizando las métricas de desempeño de cada uno de los modelos implementados en este proyecto, en sus diferentes iteraciones, se puede concluir que el modelo que mejores predicciones realizó y con mejores métricas de desempeño fue el modelo de regresión lineal, es importante destacar de este modelo se capacidad de generalizar, toda vez que entre todos los modelos fue el menos sobre ajustado cuyo MAE fue de 846 para los datos de entrenamiento y de 871 para los datos de prueba. Los resultados más notables fueron a partir de la segunda iteración, cuando se crearon nuevas variables predictoras que se guardan cierta relación con la variable a predecir, las métricas fueron mejorando, para la tercera iteración cuando se decidió trabajar con los datos a partir de octubre de 2021, las métricas mejoraron considerablemente, esto se pudo haber dado, a que los datos a partir de esta fecha  tanto la variable respuesta como los predictores toman distribuciones normales, para la cuarta iteración cuando se utilizaron todas las disposiciones de las iteraciones pasadas, pero esta vez se utilizado validación cruzada en el entrenamiento del modelo, se pudo observar que la métrica mejoro.

Mientras que el modelo de RandomForestRegressor en su cuarta iteración, tuvo sobreajuste cuyo MAE fue de 390 para los datos de entrenamiento y de 1215 para los datos de prueba, y las métricas fueron menos favorables con respecto al algoritmo de regresión lineal

El modelo de KNeighborsRegressor fue el modelo más sobre ajustado cuyo MAE fue de 0 para los datos de entrenamiento y de 2356 para los datos de prueba.

### 6.1 Métricas

Los resultados numéricos de las métricas para las iteraciones más relevantes fueron los siguientes:

*Tabla 11Primera iteración:*

| Modelo | Error Medio Absoluto (MAE) | Error Porcentual Absoluto Medio **(MAPE)** |
|---|---|---|
| LinearRegression | 1771 | 10.16 % |
| RandomForestRegressor | 2136 | 12.11 % |
| KNeighborsRegressor | 4155 | 21 % |

*Tabla 12Segunda iteración:*

| Modelo | Error Medio Absoluto (MAE) | Error Porcentual Absoluto Medio **(MAPE)** |
|---|---|---|
| LinearRegression | 913 | 4.48 % |
| RandomForestRegressor | 1236 | 6.1 % |
| KNeighborsRegressor | 6812 | 34 % |

*Tabla 13Tercera iteración:*

| Modelo | Error Medio Absoluto (MAE) | Error Porcentual Absoluto Medio **(MAPE)** |
|---|---|---|
| LinearRegression | 871 | 4.30 % |
| RandomForestRegressor | 1215 | 5.82 % |
| KNeighborsRegressor | 2356 | 11.42 % |
| Red Neuronal | 1482 | 7.83 % |

*Tabla 14 Cuarta iteración:*

| Modelo | Error Medio Absoluto (MAE) Validación | Error Porcentual Absoluto Medio **(MAPE)** **Validación** |
|---|---|---|
| LinearRegression | 765 | 3.92 % |
| RandomForestRegressor | 1139 | 5.58 % |
| KNeighborsRegressor | 2041 | 10 % |

## 6.2 Consideraciones de producción

El modelo fue desplegado como un servicio en la nube de Azure, como consideración para poner en producción el modelo, seria construir un sitio web que se conecte a la base de datos del tablero de control de despachos de pasajeros de la terminal, a partir de estos datos el administrador del sitio seleccionara las días a analizar que en nuestro caso serian los 7 días anteriores, y el sitio le mostraría la cantidad de pasajeros estimadas para el siguiente día.

La documentación del despliegue, como los resultados de este se encuentran en el siguiente repositorio de GitHub:

https://github.com/MarcusBruDiaz/Despliegue_modeo_pasajeros.git

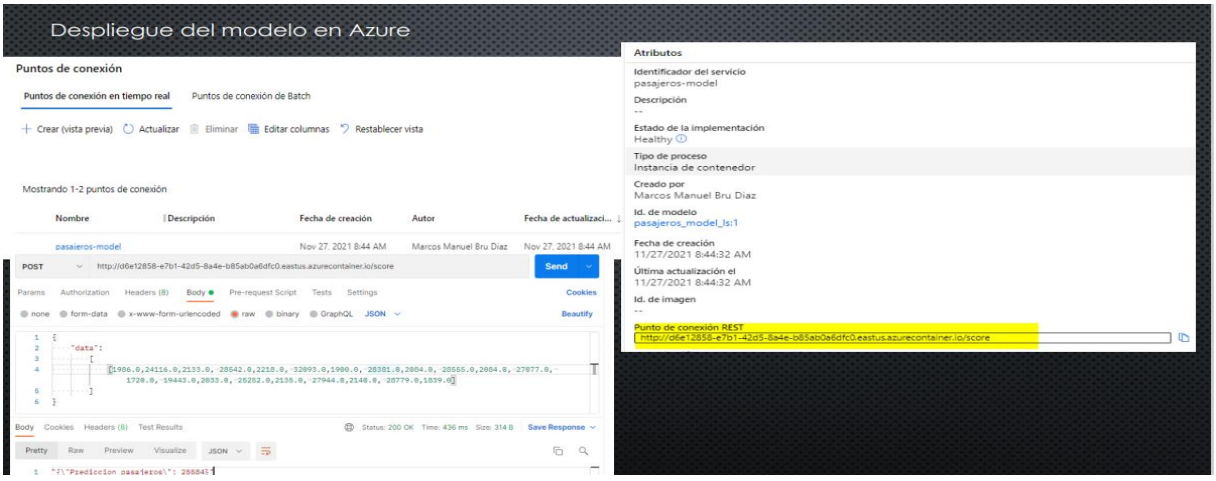


*Figura 32 Despliegue del modelo en Azure*

## 7. Conclusiones

Esta investigación nace con el fin de hacer frente desde el análisis de datos, a los problemas de negocio que presenta la terminal del norte de Medellín, en la prestación adecuado de los servicios. Lo cual es derivado de la falta de planeación y proyección de demanda.

Es necesario precisar en la importancia de saber en promedio la demanda de pasajeros que saldrán de la terminal, para el negocio es un dato muy importante para toma de decisiones, porque a partir de estas proyecciones se pueden tomar medidas tales como adecuar la infraestructura y zonas comunes, habilitar zonas de parqueadero suficiente para la demanda esperada, implementar estrategias con los clientes directos que son los comerciantes que tienen sus negocios dentro de la terminal, y así se abastezcan lo suficiente para atender la demanda de pasajeros, de esta manera se verían beneficiadas estas personas, así como también implementar medidas de seguridad apoyados en la fuerza pública solicitando un mayor acompañamiento en los días de mayor demanda de pasajeros, otras disposiciones serian la implementación de capacitaciones a los funcionarios de la terminal para prestar un mejor acompañamiento a los usuarios.

Así fue como este proyecto se apoyó en datos históricos disponibles por la terminal de transporte, con los cuales se puede hacer inferencias de la cantidad de pasajeros que saldrán de la terminal en determinada fecha, y utilizando las herramientas pertinentes de análisis de datos como el Machine Learning, se puede generar una información precisa con respecto a la demanda de pasajeros.

De cara a la resolución de los problemas de negocio de la terminal, se puede decir que, con la implementación de los modelos predictivos desarrollados, los administradores pueden tomar mejores decisiones, que vayan encaminadas con los objetivos trazados por la entidad, lo cual la posesionaria como unas de las mejores Terminales de transporte del país.

Con respecto a la confiabilidad de predicción del modelo, se puede decir que los resultados obtenidos son considerados buenos para el negocio, debido la terminal en un día pueden tener miles de pasajeros que salen a los diferentes destinos del país, y el mejor modelo tiene un porcentaje de error solo un 3.6 % y un error promedio de 766 entre el valor real y el estimado por el modelo, siendo muy poco con respecto a la gran cantidad de pasajeros que moviliza la terminal del Norte en un día.

No obstante, a lo anterior, y con el fin de mejorar las métricas de desempeño del modelo, se puede considerar la idea de construir nuevas variables predictoras que guarden una buena relación con la demanda de pasajeros.

# Referencias


Brownlee, J. (28 de 08 de 2020). *machinelearningmastery*. Obtenido de machinelearningmastery: https://machinelearningmastery.com/multi-step-time-series-forecasting-long-short-term-memory-networks-python/

Gliese710. (5 de 10 de 2019). *exponentis*. Obtenido de exponentis: http://exponentis.es/como-dividir-un-conjunto-de-entrenamiento-en-dos-partes-train-test-split

Heras, J. M. (19 de 09 de 2020). *iartificial*. Obtenido de iartificial: https://www.iartificial.net/fases-del-proceso-de-machine-learning/

Heras, J. M. (18 de 09 de 2020). *iartificial*. Obtenido de iartificial: https://www.iartificial.net/random-forest-bosque-aleatorio/

*https://www.datos.gov.co/*. (01 de 10 de 2021). Obtenido de https://www.datos.gov.co/: https://www.datos.gov.co/Transporte/Operaci-n-de-pasajeros-y-despacho-de-veh-culos-en-/eh75-8ah6

LONDOÑO, N. V. (01 de 01 de 2013). *https://repository.udem.edu.co/.* Obtenido de https://repository.udem.edu.co/: https://repository.udem.edu.co/bitstream/handle/11407/148/Dise%C3%B1o%20del%20modelo%20de%20servicios%20para%20las%20terminales%20de%20transporte%20de%20Medell%C3%ADn.pdf?sequence=1&isAllowed=y---

Medellin, T. d. (27 de 11 de 2021). *terminalesmedellin*. Obtenido de terminalesmedellin: https://terminalesmedellin.com/mision-y-vision/

mintransporte. (01 de 01 de 2020). *mintransporte*. Obtenido de mintransporte: https://www.mintransporte.gov.co

Torres, G. M. (1 de 9 de 2008). *scielo.* Obtenido de scielo: http://www.scielo.org.co/pdf/rfiua/n45/n45a09.pdf

Tutoriales, G. (26 de 01 de 2015). *gestiondeoperaciones*. Obtenido de gestiondeoperaciones: https://www.gestiondeoperaciones.net/proyeccion-de-demanda/error-porcentual-absoluto-medio-mape-en-un-pronostico-de-demanda/