UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Facultad de Estudios Superiores Acatlán

Diplomado de Ciencias de Datos

Examen

Diplomado de Ciencia de Datos, Módulo II

Profesora: Carla Paola Malerva Reséndiz

Alumno: Ricardo Paramont Hernández González

Fecha: Miercoles 21 de abril de 2021

1. Conjunto de Datos

El dataset empleado para el examen corresponde al contenido dentro del archivo "data_examen.csv", que se compone de las variables:

- **key** : Variable que debe indicar el id del viaje, sin embargo no tiene informaci´on correcta , por lo cual no hace caso de dicha variable.
- **fare amount**: Monto de la tarifa asociado a cada viaje, la tarifa incluye gastos de peaje.
- **pickup datetime** : Fecha y hora en la que se comenz´o el viaje
- **pickup longitude** : Longitud donde comenz´o el recorrido
- **pickup latitude** : Latitud donde comenz´o el recorrido
- **dropoff longitude** : Longitud donde concluy´o el recorrido
- **dropoff latitude** : Latitud donde concluy´o el recorrido
- **passenger count** : N´umero de pasajeros durante el trayecto
- **fare class** : Tipo de tafira que se cobr´o, una tarifa baja o una tarifa alta.

2. Calidad de Datos

2.1 Etiquetado

- Se etiquetaron las variables según el tipo de información contenida como: continua, discreta o fecha.
- Columnas etiquetadas:

  continuas: 'fare_amount', 'pickup_longitude', 'pickup_latitude', 'dropoff_longitude', 'dropoff_latitude'.

  discretas: 'passenger_count', 'fare_class'

  fecha: 'pickup_datetime'

2.2 Completitud

La siguiente tabla registra el número de datos faltantes por columna:

	columna	total	completitud
0	c_dropoff_longitude	7	99.999
1	c_dropoff_latitude	7	99.999
2	c_fare_amount	0	100.000
3	d_pickup_datetime	0	100.000
4	c_pickup_longitude	0	100.000
5	c_pickup_latitude	0	100.000
6	v_passenger_count	0	100.000
7	v_fare_class	0	100.000

Se eliminaron los siete registros con datos faltantes.

2.3 Consistencia.

Se buscaron datos inconsistentes dentro del dataset y se eliminaron.

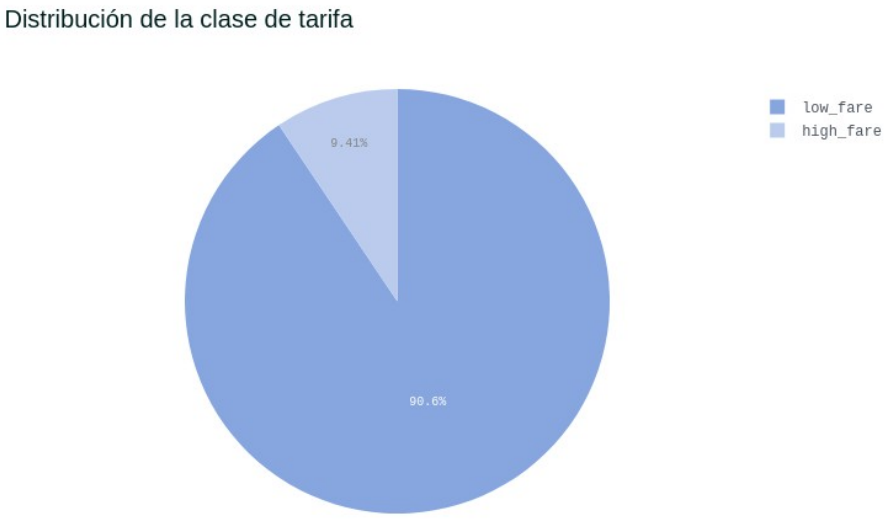
Los criterios co los que se encontraron registros inconsistentes fueron:

- longitud entre –180 y 180
- latitud entre –90 y 90
- número de pasajeros entre 0 y 6

  * Existía un único registro inconsistente con 208. Los registros con 0 pasajeros se aceptan como posibles envíos de mercancías
- tarifa en cantidad mayor a 0

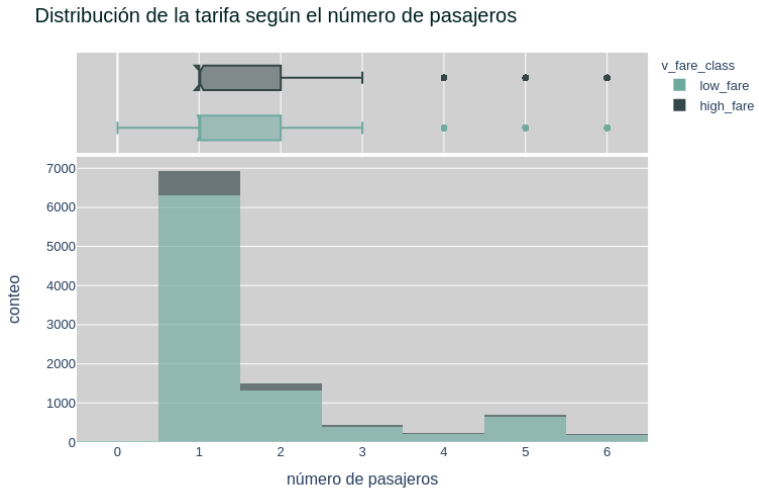
3. Análisis exploratorio

3.1 Distribución de la clase tarifa



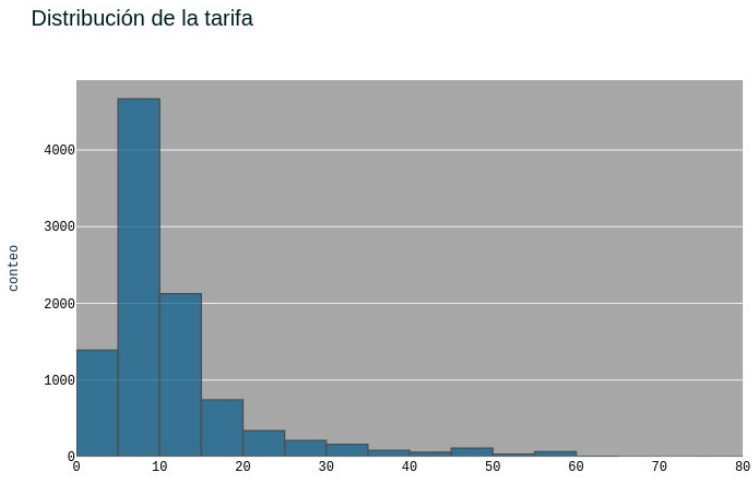
Se observa que la clase de tarifa alta tiene un tasa de ocurrencia de 10%. Se considera que con esa tasa no es necesario aplicar undersampling u oversampling.

3.2 Distribución de la clase tarifa según número de pasajeros



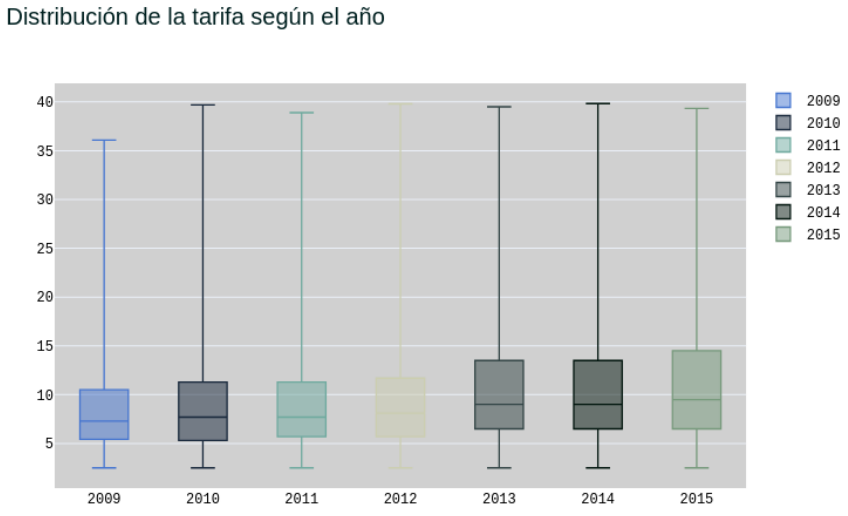
No se alcanza a observar a simple vista una diferencia entre la distribución de tarifas según el número de pasajeros. Sin embargo, la gran mayoría de viajes se realizan por un solo pasajero.

3.3 Distribución de la tarifa



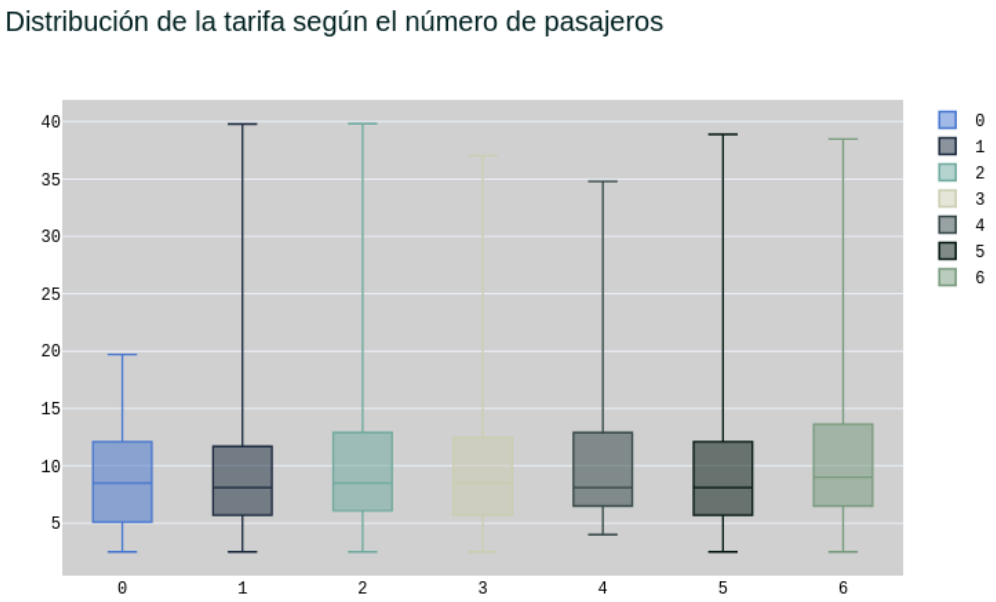
La tarifa de los viajes tiende a valores entre 5 y 10.

3.4 Distribución de la tarifa según el año



Se observa un aumento gradual de la tarifa, siendo que la media tendía aproximadamente a 7 en el año 2009 y aproximadamente a 10 ya para el año 2015.

3.5 Distribución de la tarifa según el número de pasajeros



No se alcanza a apreciar facilmente una diferencia significativa entre el costo del viaje y el número de pasajeros.

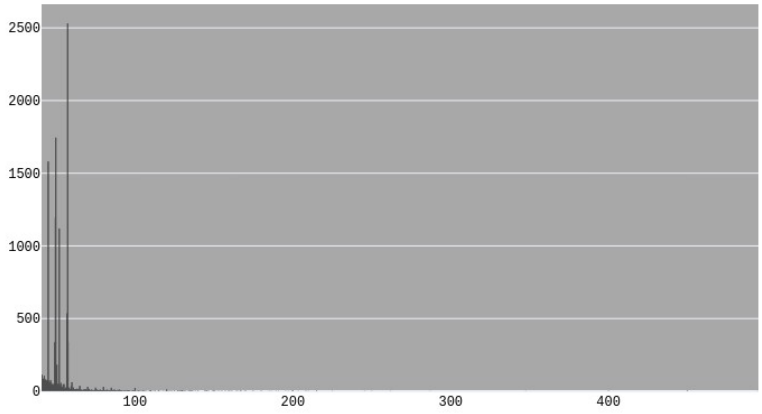
4. Datos Anómalos u Outliers

- Se buscaron datos anómalos para las variables de ubicación: pickup_longitude, pickup_latitude, dropoff_longitude y dropoff_latitude.

	features	n_outliers	n_outliers_%
0	c_pickup_longitude	11748	0.02
1	c_pickup_latitude	11687	0.02
2	c_dropoff_longitude	11749	0.02
3	c_dropoff_latitude	11683	0.02

- Se encontraró que la mayoría de los datos anónalos correspondian a resgitros con valor de cero en todas sus variables de ubicación. Dichos registros son entonces datos faltantes o mal registrados. Se procedió a borrar todos esos registros.

- Se buscaron datos anómalos para la variable objetivo de fare_amount:



- Se encuentran registros con valores de tarifas demasiado elevadas y se procede a borrar dichos registros, los cuales corresponden a 0.03% de los registros totales.

5. Ingeniría de variables

5.1 Fecha

Se crean las siguientes variables, fruto de la desconposición de la variable pickup_date:

- año
- cuarto de año
- mes
- semestre
- día del mes
- día de la semana
- fin de semana: dummy

5.2 variables de ubicación

Se intenta calcular aproximaciones de la distancia recorrida en el viaje usando la distancia euclidiana y la distancia Manhattan entre los puntos de inicio y final del viaje.

5.3 clase de tarifa

Se codifica la variable objetivo de clasificación fare_class:

- high_fare = 1
- low_fare = 0

6. Modelado

6.1 Clasificación

- Se probaron 5 métodos distintos de clasificación y se compararon sus métricas, dandole mayor peso al roc auc. La evaluación de los modelos se encuentran en el archivo excel dentro de la carpeta del examen.
- El mejor modelo encontrado consta de un arbol de decisión. Sus métricas de vealuación para el conjunto de entrenamiento y prueba son:

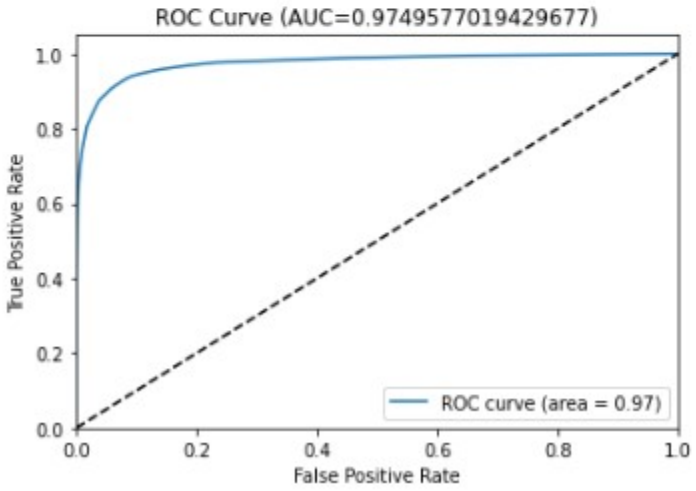
Métricas Train

```
Exactitud : 0.973
Precision : 0.993
Recall : 0.979
f1_score : 0.986
TPR : 0.979
FPR : 0.121
```

Matríz de confusión de Train

	Observacion_low	Observacion_high
Prediccion_low	369988	7960
Prediccion_high	2701	19546

ROC train



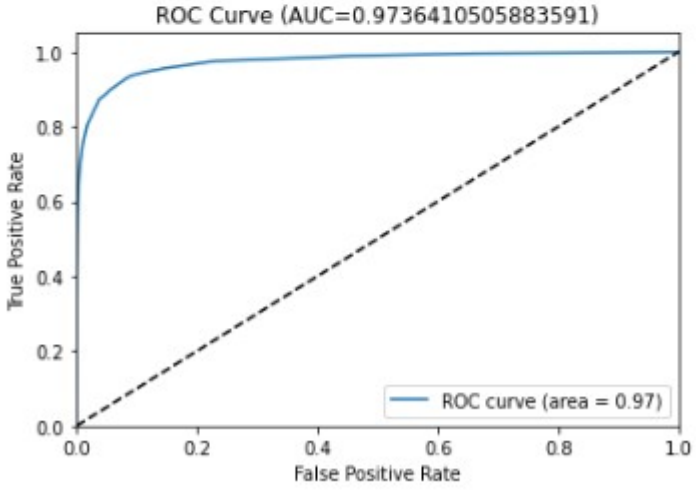
Métricas Test

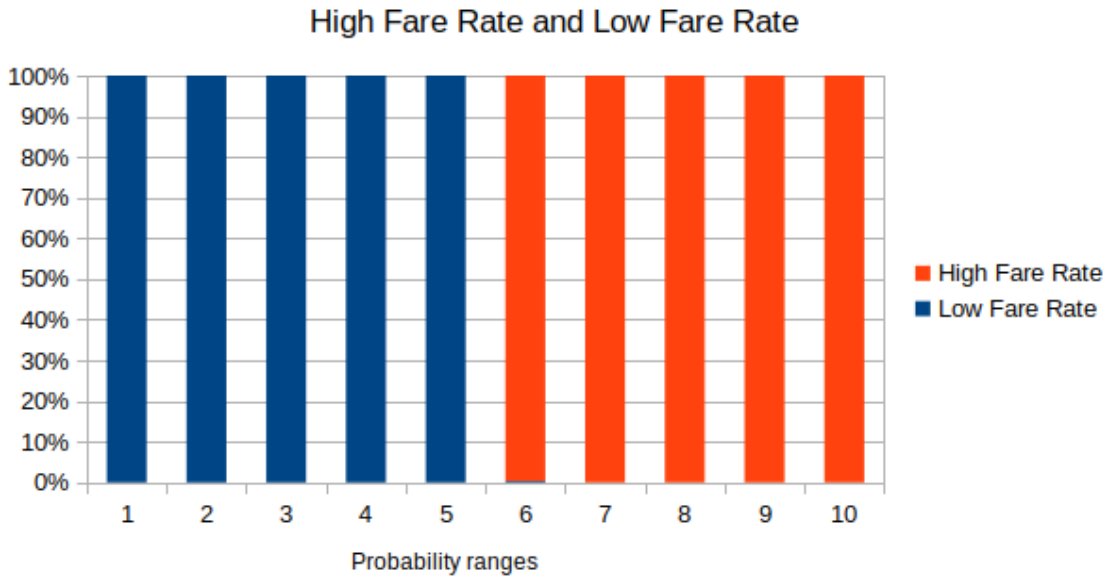
```
Exactitud : 0.973
Precision : 0.992
Recall : 0.979
f1_score : 0.986
TPR : 0.979
FPR : 0.125
```

Matríz de confusión de Test

	Observacion_low	Observacion_high
Prediccion_low	158515	3337
Prediccion_high	1210	8451

ROC test





6.1 Regresión

- Se probaron 5 métodos distintos de regresión y se compararon sus métricas, dandole mayor peso al rmse. La evaluación de los modelos se encuentran en el archivo excel dentro de la carpeta del examen.
- El mejor modelo encontrado consta de un bosque aleatorio. Sus métricas de vealuación para el conjunto de entrenamiento y prueba son:

Métricas Train
Exactitud : 0.968
Precision : 0.992
Recall : 0.974
f1_score : 0.983
TPR : 0.974
FPR : 0.143

Métricas Test
Exactitud : 0.968
Precision : 0.992
Recall : 0.974
f1_score : 0.983
TPR : 0.974
FPR : 0.142

Gráfica de intervalos de confianza del modelo con datos de entrenamiento

acercamiento:

Gráfica de intervalos de confianza del modelo con datos de prueba

acercamiento: