Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Maestría en Ciencia de Datos

Materia: Aprendizaje Automático

Práctica 3

Clasificación

Profesor: José Anastacio Hernández Saldaña

Alumno: Francisco David Treviño Bautista

Matrícula: 562795

Grupo: 1

Fecha de entrega: 20 de julio de 2024

Conjunto de Datos

Se eligió el conjunto de datos con la información de viajes de taxis amarillos reportada por la ciudad de Nueva York TLC mensualmente y que es de carácter público.

Para más información se puede consultar la siguiente página:

http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

El conjunto de datos original se descargó en archivo con formato tipo ".parquet", el cuál maneja compresión de datos, y en R Studio se cuenta con una librería para cargar los datos como un dataframe de Pandas.

Variable	Descripción
VendorID	Un código que indica el proveedor de TPEP que proporcionó el registro. 1 = Tecnologías móviles creativas, LLC 2 = VeriFone Inc.
tpep_pickup_datetime	La fecha y hora en que se activó el taxímetro.
tpep_dropoff_datetime	La fecha y hora en que se desconectó el taxímetro.
Passenger_count	El número de pasajeros en el vehículo. Este es un valor ingresado por el conductor del taxi.
Trip_distance	La distancia del viaje transcurrido en millas reportadas por el taxímetro.
PULocationID	Zona de Taxi TLC en la que estaba activado el taxímetro
DOLocationID	Zona de Taxi TLC en la que se desactivó el taxímetro
RateCodeID	El código de tarifa final vigente al finalizar el viaje. 1 = Tarifa estándar, 2 = JFK , 3 = Newark , 4 = Nassau o Westchester 5 = Tarifa negociada, 6 = Viaje en grupo
Store_and_fwd_flag	Este indicador indica si el registro de viaje se mantuvo en la memoria del vehículo antes de enviarlo al proveedor, también conocido como "almacenar y reenviar", porque el vehículo no tenía una conexión con el servidor. Y = viaje de almacenamiento y avance N = no es un viaje de almacenamiento y avance
Payment_type	Un código numérico que indica cómo el pasajero pagó el viaje. 1 = Tarjeta de crédito, 2 = Efectivo, 3 = Sin cargo 4 = Disputa, 5 = Desconocido, 6 = Viaje anulado
Fare_amount	La tarifa de tiempo y distancia calculada por el taxímetro.
Extra	Extras y recargos varios. Actualmente, esto solo incluye los cargos de $0,50 y $1 por hora pico y por noche.
MTA_tax	Impuesto MTA de $0.50 que se activa automáticamente según la tarifa medida en uso.
Improvement_surcharge	Recargo por mejora de $0.30 en viajes evaluados al bajar la bandera. El recargo por mejora comenzó a cobrarse en 2015.
Tip_amount	Monto de la propina: este campo se completa automáticamente para las propinas de tarjetas de crédito. Las propinas en efectivo no están incluidas.
Tolls_amount	Importe total de todos los peajes pagados en el viaje.
Total_amount	El importe total cobrado a los pasajeros. No incluye propinas en efectivo.
Congestion_Surcharge	Monto total cobrado en el viaje por el recargo por congestión del Estado de Nueva York.
Airport_fee	$1.25 para traslados únicamente en los aeropuertos LaGuardia y John F. Kennedy

Posteriormente a la carga de la información, se tuvo que realizar un análisis del conjunto de datos para entender su distribución con estadística descriptiva, así como detectar los valores mínimos y/o máximos.

También hubo actividades de limpieza de los datos, porque se detectaron registros con datos inválidos, negativos o fuera de rango. Por ejemplo, había viajes con montos negativos, con una duración mayor a 12 horas, etc.

El conjunto de datos tiene las siguientes características:

- Viajes de taxi que salen del aeropuerto LaGuardia.
- Tienen como destino cualquier ubicación dentro de Manhattan.
- Los viajes corresponden a la semana del domingo 14 al sábado 20 del mes de abril.

Cabe mencionar que se restringió el conjunto de datos por el tamaño y tiempo de procesamiento, originalmente cuenta con 3.5 millones de registros.

Elección de Variables

Seleccionar las variables para el modelo de clasificación.

Tipo de Variable	Variable	Descripción
Independiente	trip_distance	La distancia del viaje transcurrido en millas reportadas por el taxímetro.
Independiente	trip_duration_minutes	Duración del viaje en minutos. Campo calculado en función a la duración del viaje en segundos.
Independiente	DOLocationID	Zona de Taxi TLC en la que se desactivó el taxímetro y/o terminó el viaje.

Seleccionar las variables objetivo para clasificar.

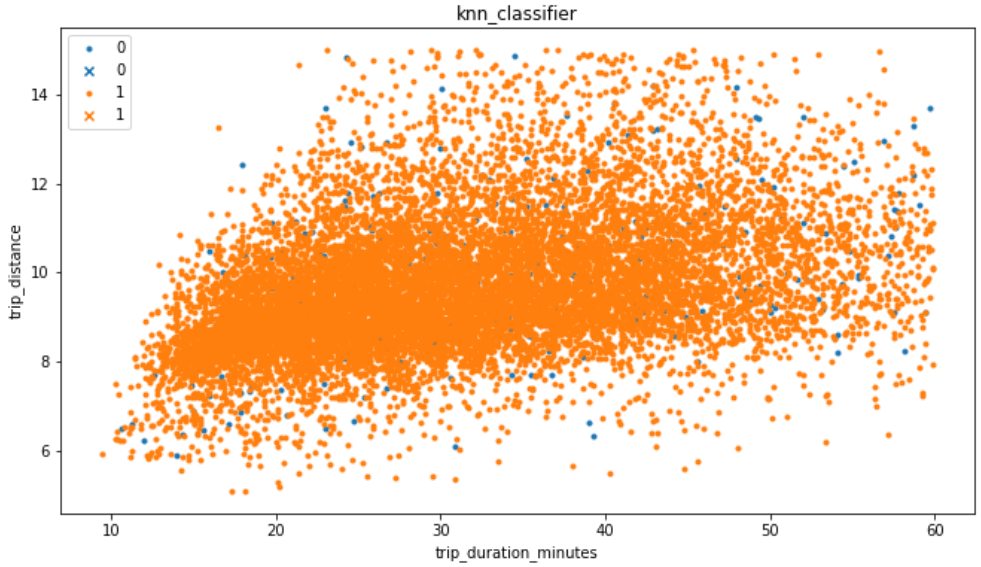
Tipo de Variable	Variable	Descripción
Dependiente	tip_given	Variable binaria que indica si el usuario dio una propina al taxista como parte de su pago total.

Modelos de Clasificación

Se desarrolló código adicional para realizar las pruebas del conjunto de variables de clasificación, así como las variables objetivo, se evaluaron los siguientes modelos:

- Logistic Regression
- Support Vector Machine
- K-Nearest Neighbors
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier

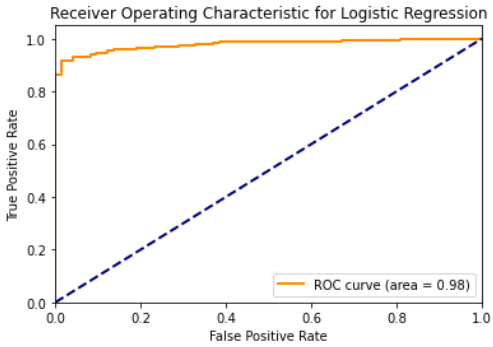
Por otro lado, se realizaron pruebas usando de referencia el código del repositorio del profesor, que se revisó en clase, usando “knn_classification_model “ para las clasificaciones de las variables “trip_duration_minutes” y “trip_distance” en relación con “tip_given”.



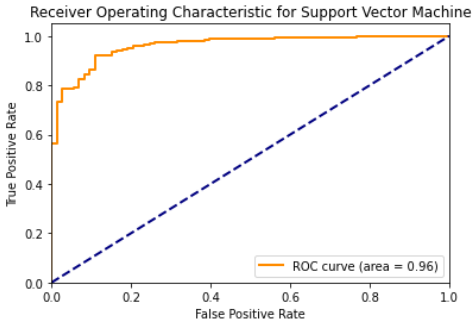
Evaluación del Modelo

Utilizar validación cruzada y escoger un criterio de evaluación (precisión, exactitud, AUC, etc.).

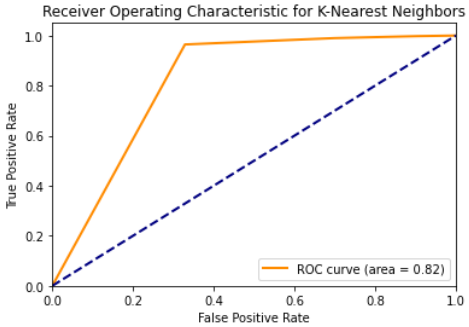
- Logistic Regression



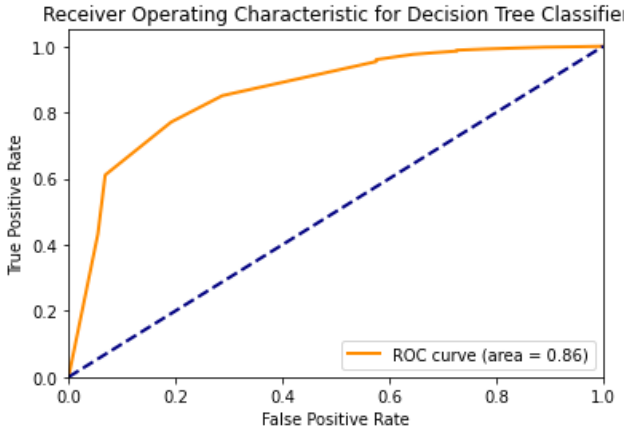
- Support Vector Machine



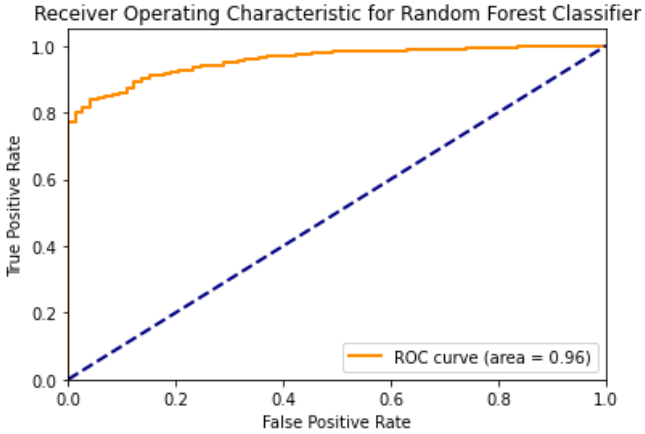
- K-Nearest Neighbors



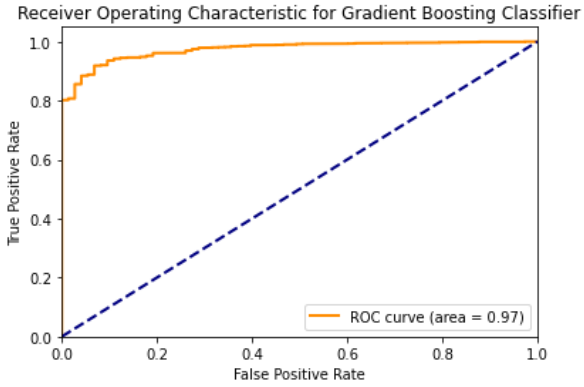
- Decision Tree Classifier



- Random Forest Classifier

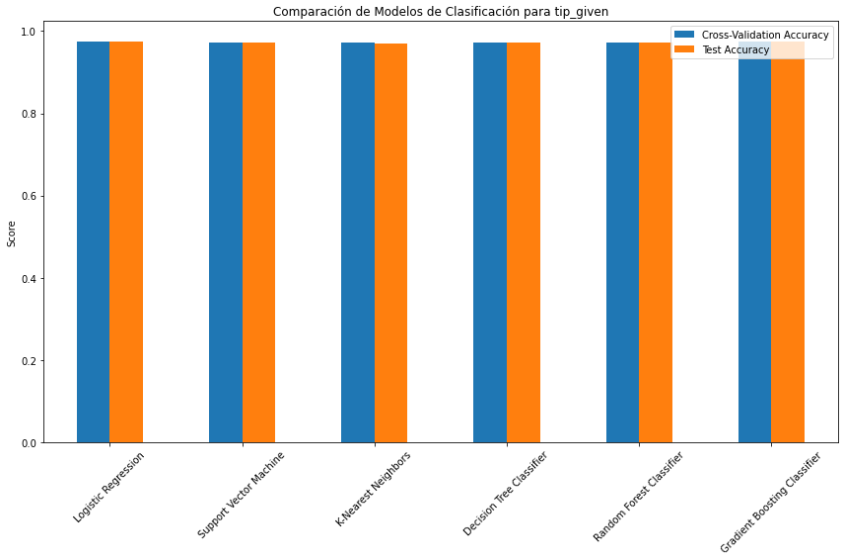


- Gradient Boosting Classifier



Selección del Modelo

Modelo	Cross-Validation Accuracy	Test Accuracy	Test ROC AUC
Logistic Regression	0.972853328	0.971901463	0.527584430
Support Vector Machine	0.972853328	0.971901463	0.536780144
K-Nearest Neighbors	0.968713992	0.965742879	0.509073647
Decision Tree Classifier	0.971024229	0.970746728	0.571325105
Random Forest Classifier	**0.972853328**	**0.971901463**	**0.590644243**
Gradient Boosting Classifier	0.969002731	0.966512702	0.556834396



Los modelos con la mejor precisión en las métricas Cross-Validation Accuracy y Test Accuracy son Logistic Regression, Support Vector Machine, y Random Forest Classifier, todos con valores muy cercanos y altos (alrededor de 0.97).

El modelo **Random Forest Classifier** tiene el mayor valor de ROC AUC (0.590644243), lo que indica que tiene una mejor capacidad para distinguir entre las clases de “tip_given”.

Considerando todas las métricas, el Random Forest Classifier parece ser el mejor modelo porque tiene valores muy altos de precisión en validación cruzada y prueba, y tiene el mayor valor de ROC AUC, lo que indica una mejor capacidad de discriminación.