Pablo Saurina Martínez

Jaime De la Vega

Jaime Bernardo Martínez

Práctica Spark

Los objetivos principales han sido dos, el primero, la realización de análisis descriptivo y estadistico para la base de datos junto con la implementación de un modelo estadistico para predecir el tiempo de viaje. Por otro lado, se ha aplicado machine learning con el modelo de clasificación RandomForest para predecir el grupo de edad del usuario según el uso que haga. Para ello hemos utilizado la librería spark de Python, y para la ejecución en archivos grandes hemos empleado el cluster al que tenemos acceso en esta asignatura. En el notebook de Python LinearRegression.ipynb contamos más detalladamente los procesos que hemos seguido para realizar un primer análisis de los datos, filtrar las variables y la información que nos interesa y por último la realización de una regresión lineal, posteriormente, se realiza otra aproximación a través del machine learning con RandomForest model for binary or multiclass classification.

Ejecución de los programas

Los programas principales de la práctica son:

* LinearRegression.ipynb

Este jupyter-notebook de Python contiene el proceso de análisis de datos, pero por comodidad y gestión de tiempo solo se ejecuta con tres archivos de prueba. Sería necesario tener estos tres archivos (sample\_10e2.json, sample\_10e3.json y sample\_10e4.json) en la misma ubicación que el notebook.

* Random Forest

Este jupyter-notebook de Python contiene el proceso de análisis de datos, pero por comodidad y gestión de tiempo solo se ejecuta con tres archivos de prueba. Sería necesario tener estos tres archivos (sample\_10e2.json, sample\_10e3.json y sample\_10e4.json) en la misma ubicación que el notebook.

* LR\_cluster.py

Este archivo de Python es el verdadero programa de la práctica, ya que replica el análisis hecho en el jupyter-notebook, pero esta vez no se limita a los archivos de prueba. Al ejecutar este archivo se ha de utilizar como argumento de entrada un archivo .txt que contenga las direcciones de las bases de datos que utilizaremos para el análisis de datos. En nuestro caso empleamos el archivo bases\_de\_datos.txt, que esta preparado para ejecutarse en el cluster, donde emplear grandes cantidades de datos es más rápido.

Las imágenes que hay en el repositorio son capturas de pantalla de la ejecución de este archivo .py con bases\_de\_datos.txt como argumento de entrada.

Datos empleados

Los datos empleados en esta práctica los hemos obtenido de la página web de BiciMat en el enlace: <https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)> . Para nuestro análisis hemos empleado los datos desde Enero de 2020 hasta Junio de 2021. Estos archivos los hemos subido a hadoop, y sus direcciones estan guardadas en bases\_de\_datos.txt (pues leemos este archivo en la ejecución del programa).

Los archivos empleados son:

/user/psaurina/202001\_movements.json

/user/psaurina/202002\_movements.json

/user/psaurina/202003\_movements.json

/user/psaurina/202004\_movements.json

/user/psaurina/202005\_movements.json

/user/psaurina/202006\_movements.json

/user/psaurina/202007\_movements.json

/user/psaurina/202008\_movements.json

/user/psaurina/202009\_movements.json

/user/psaurina/202010\_movements.json

/user/psaurina/202011\_movements.json

/user/psaurina/202012\_movements.json

/user/psaurina/202101\_movements.json

/user/psaurina/202102\_movements.json

/user/psaurina/202103\_movements.json

/user/psaurina/202104\_movements.json

/user/psaurina/202105\_movements.json

/user/psaurina/202106\_movements.json

Resultados y Conclusiones

El análisis llevado a cabo no nos ha proporcionado resultados positivos acerca de la relación entre la variable a predecir y las predictoras, el modelo lineal que hemos diseñado no predice de forma correcta el tiempo de viaje.

La conclusión a la que llegamos es principalmente al hecho de que los grupos de edad, la fecha de inicio del viaje y la hora de inicio del viaje no estan bien correlacionadas con la duración del viaje. Es decir que si quisiéramos predecir la duración de un viaje deberíamos recabar datos distintos para que esta predicción fuera más precisa, sin embargo, a la hora de predecir la variable de edad con las demás si resulta mucho más sencillo ya que hay una cantidad menor de grupos, luego se puede aplicar Machine Learning con RandomForest que realiza una predicción con 73% de acierto en el conjunto test, resultando una cuota de acierto alto y por lo tanto demostrando un método efectivo