FUNDATEC

Estudiante: Jaime Villegas Gallardo

Curso: Big Data

Profesor: Juan Manuel Esquivel Rodriguez

Tarea 3

Instrucciones

Descomprimir el archivo tarea3.zip

Construir el contenedor y copiar los archivos

```
$ ./build_image.sh
```

Ejecutar el contenedor de la base de datos

```
$ cd db/
$ ./run_image.sh
```

Levantar jupyter - en la raiz del contenedor

```
$ ./load_jupyter_notebook.sh
```

Documento de jupyter

Una vez levantado el jupyter acceder a la carpeta "tarea3" y abrir el documento 'tarea3.ipynb' donde de la carpeta deberan estar el jar del postgress y el csv para los datos

Problemas encontrados en el dataset

El principal problema del dataset constaba que cada uno de los features estaba expreado en strings por lo que se tuvo que mapear estos y pasar de 23 a 123 features, luego de esto se entreno el modelo y se agrego una seccion extra con un PCA para disminuir la cantidad de features, sin embargo el analisis con los dos modelos se aplico solamente a los datos sin el PCA, este ultimo solamente se entreno un modelo y se jugo con la cantidad de features para encontrar un numero optimo de estos y obtener resultados adecuados

Analisis de resultados

Estos datos son en base a la ultima corrida del codigo, por lo que al correrlos nuevamente estos pueden cambiar

Los datos fueron divididos 70% para entrenamiento 30% para pruebas

Se utilizaron los siguientes modelos para entrenar los datos * Binomial Logistic Regression * Naive Bayes

Ambos modelos de clasificacion binaria, en los cuales se obtuvieron los siguietes resultados de 'exactitud' de acuerdo a las pruebas

- Binomial Logistic Regression
  - Sin K-fold cross validation: 0.9979364424267437
  - Con K-fold cross validation: 0.9979364424267437
- Naive Bayes
  - Sin K-fold cross validation: 0.9467602146099876
  - Con K-fold cross validation: 0.998349153941395

Como se puede observar se obtiene mejor resultado al implementar el modelo Binomial Logistic Regression, tanto en el test con los kfolds y el testing directo.

Con lo que respecta al modelo Naive Bayes muestra una mejoria cuando se realiza la prueba con los kfolds por lo que podemos ver que al ser puesto a prueba de diferentes formas los datos de training se mejora la prediccion del mismo

En el caso del modelo entrenado con la PCA con 20 componentes principales - mejor escenario con 35

- Binomial Logistic Regression
  - Sin K-fold cross validation: 0.9793187347931873
  - Con K-fold cross validation: 0.9979364424267437