Evaluación Final:

PDE MACHINE LEARNING CON PYTHON

Prof. Abraham Zamudio

Agosto 2022

Pregunta 1

Considere el conjunto de datos ubicado la siguiente url:

https://www.mef.gob.pe/contenidos/archivos-descarga/Reactiva_Peru_Lista_de_empresas_al_30102020.xlsx

Para cada uno de los siguientes items realice comentarios acerca de las conclusiones que puede obtener de los gráficos realizados.

- Realice un diagrama de barras para observar como se distribuyo el fondo del proyecto en los diferentes rubros del mercado (columna : SECTOR ECONÓMICO).
- Muestre mediante un diagrama de barras el total de monto prestado (columna : MONTO PRÉSTAMO (S/)) por cada entidad otorgante (columna : NOMBRE DE ENTIDAD OTORGANTE DEL CRÉDITO)
- Muestre mediante un diagrama de barras el total de monto prestado (columna : MONTO PRÉSTAMO (S/)) para cada departamento (columna : DEPARTAMENTO)
- Realice un histograma para la columna MONTO COBERTURADO (S/) para cada uno de los departamentos.

Pregunta 2

Considere el conjunto de datos : DatosViernesNegro.csv

1. Muestre en un diagrama de barras como se distribuye el numero de elementos de la variable Gender.
2. Considere la siguiente codificación para la variable Age

Age	Age_Cod
0-17	Stage1
18-25	Stage2
26-35	Stage3
36-45	Stage4
46-50	Stage5
51-55	Stage6
55+	Stage7

Muestre en un diagrama de barras como se distribuye el numero de elementos de la variable Age_Cod

3. Elimine las variables con mas de 2022 valores faltantes.
4. Elimine las variables User_ID y Product_ID
5. Muestre un histograma de la variable Purchase separando los datos por sexo (Variable Gender)
6. Muestre un histograma de la variable Purchase separando los datos por la variable Age_Cod
7. Cree un modelo de clasificacion (diferente a una regresión logistica) considerando que la variable de interés es el Sexo (Variable : Gender).
8. Cree un RandomForestRegressor (haciendo un proceso de GridSearchCV) considerando como variable dependiente a Purchase.
9. Cree un dataset de nombre **VN_Cities_A_C** donde solo se considere las ciudades A y C (Variable City_Category). Para este nuevo dataset cree un modelo de regresión logística (haciendo un proceso de GridSearchCV) considerando a la variable City_Category como la variable de interes.
10. Implemente una red neuronal de clasificacion para pronosticar la variable Gender. Realice los cinco primeros pasos descritos en la sesion 12. Es decir :
    1. 1er Paso : Cargar datos
    2. 2do Paso : Definir un modelo en keras
    3. 3er Paso : "Compilamos" el modelo del paso 2
    4. 4to Paso : Ajustar el modelo a nuestro conjunto de datos (proviene del paso 1)
    5. 5to Paso : Evaluación del modelo

Pregunta 3

Considere el siguiente conjunto de datos :

https://raw.githubusercontent.com/robintux/Datasets4StackOverFlowQuestions/master/BIG_MART_SALES_PREDICTION.csv

- Preprocesamiento :
  - Considere las columnas con datos faltantes, muestre y almacene en disco duro un diagrama de barras que el numero de datos faltantes para esas columnas.
  - A partir del contenido de las columnas, separe estas en columnas con datos categóricos (nombre del dataframe : BM_cat) y columnas con datos numéricos (BM_Cuan). Para cada uno de estos dataframes realice una imputación adecuada para los valores faltantes.
  - Muestre un histograma de la columna Item_Outlet_Sales para cada nivel de la variable Outlet_type ('Supermarket Type1', 'Supermarket Type2', 'Grocery Store','Supermarket Type3')
- Modelamiento : Considere que la variable objetivo es Item_Outlet_Sales
  - Construya modelos de
    - regresión: Sin hacer grid search
    - bosque aleatorio : Usando grid search. Usted plantee el espacio de hiperparametros.
    - Proponga una red neuronal buscando mejorar los indicadores de calidad obtenidos con los dos modelos anteriores.

Pregunta 4

Considere el siguiente conjunto de datos

https://raw.githubusercontent.com/robintux/Datasets4StackOverFlowQuestions/master/cardiotocograms_2000.csv

Este conjunto de datos contiene 2126 registros de características extraídas de exámenes de cardiotocogramas, que luego fueron clasificados por tres obstetras expertos en 3 clases:

- Normal
- Suspect
- Pathological

Realice los siguientes items :

- Muestre un diagrama de barras para la cantidad de observaciones en cada nivel de la variable fetal_health
- Considere dividir el conjunto de datos en función de la variable objetivo : fetal_health. Para cada uno de estos 3 conjuntos de datos realice un estudio descriptivo de las variables :
  - 'baseline value'
  - 'accelerations'
  - 'fetal_movement'
  - 'uterine_contractions'
  - 'light_decelerations'
- Considere la variable objetivo fetal_health y construya los siguientes modelos
  - Una regresión logística usando grid search con el siguiente conjunto de hiperparametros
    - param_grid = {'penalty': ['l1', 'l2'], 'solver': ('newton-cg', 'lbfgs', 'liblinear'),'C': [1, 2, 5, 10, 20]}
  - Un bosque aleatorio usando grid search con el siguiente conjunto de hiperparametros
    - param_grid = {'n_estimators': [50, 100, 250, 500, 1000],
      'criterion': ('gini', 'entropy'),
      'min_samples_leaf': [1, 2, 5, 10, 20, 50],
      'min_samples_split' : [5, 10, 25, 50],
      'max_depth': [1, 3, 5, 7, 10]
      }
  - Una red neuronal adecuada para el problema

  Para cada uno de estos grid search considere cv=5.