Fast**National University of Computer & Emerging Sciences, Karachi  
Fast School of Computing  
Mid Term-II Spring-2023**

**5th April 2021, 08:30 AM – 10:30 AM**

|  |  |  |
| --- | --- | --- |
| **Course Code:** CS 4081 | **Course Name:** Data Science | |
| **Instructor Name / Names:** Dr. Muhammad Nouman Durrani | | |
| **Student Roll No:** | | **Section:** |

Instructions:

* Attempt all the questions.
* Your **Student ID** and a **Secret code** (random number specific to each student) MUST be written on Top of your .**ipynb** or **.py** file. The **Secret code** will be provided by the Invigilator to each student separately at the end of the Exam. If the **Secret code** is missing on top of the submitted file, no points will be awarded.
* Please submit a **single file (.pynb preferred or .py)** containing all your solved questions **with results** in sequence.
* Please avoid submitting your solution in .rar, .zip etc.
* File(s) containing source code(s) **submitted after the Due Time (10:30 AM) will not be considered.**
* **For getting Datasets and other necessary files:**
* In case of any ambiguity, you may make an assumption. But your assumption should not contradict any statement in the question paper.

**Time**: 120 minutes. **Total Points:** 20

**Question 1:**

[**CLO: 2]** **[5 Points]**

Load the **train.csv** file in Pandas and perform the following tasks:

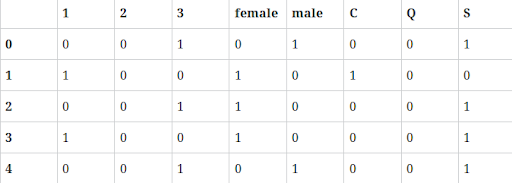
1. Take a look at the data format and drop the columns which won't contribute much to the machine learning model.

cols = ['Name', 'Ticket', 'Cabin']

1. Create dummy variables from the following columns and drop them once their dummy variables are created.

dummies = []

cols = ['Pclass', 'Sex', 'Embarked']



1. **Concatenate** the dummy variables to the original data frame, column-wise.
2. Age, has lots of missing values. Compute a median or interpolate() all the ages and fill those missing age values.

Hint: Pandas has an interpolate() function that will replace all the missing NaNs to interpolated values.

1. Sort the DataFrame by ‘Survived’ in ascending order.

**Question 2:**  **[ Points: 4 + 3.5 = 7.5 ]**

Suppose you are working on the Iris dataset that contains 150 observations and 5 variables. Variables “Sepal\_length, Sepal\_width, Petal\_length, Petal\_width” are quantitative variables describing the length and widths of parts of flowers in cm. Variable “Species” is a categorical variable that consists of three different species namely, Setosa, Versicolor, and Virginica.

We want to see if our classifiers (K-NN in part (a) and Decision Tree in part (b)) are correctly able to predict the Species class a flower belongs to based on the **Sepal\_length, Sepal\_width, and Petal\_width only**.

1. Apply the **K-NN classifier** on the above three selected attributes (where K or n\_neighbors = 1-10) if Manhattan distance (as the distance metric) = https://miro.medium.com/max/276/0*WifHj75MP5XttcQy.png and 5-fold Cross-validation is used. You will apply KNN multiple times putting K = 1-10 (where K is the number of nearest neighbors). Print “highest accuracy” and also print “value of K” (number of nearest neighbors) on which it has the highest accuracy
2. Apply the **Decision Tree classifier** on the above three selected attributes using **Gini** and **Entropy** with pruning = 0.014. You are also supposed to use a holdout method with 30% of data reserved for testing. Finally, print the accuracies along with the classification report for both impurity measures (Gini and Entropy. Print the classifier accuracy.

**Question 3:**  **[ Pointss: 4.5 + 3.5 = 7.5 ]**

Consider the **diabetes dataset**. The dataset consists of eight (8) baseline variables “Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age” and one categorical output variable “Outcome”.

1. Suppose we want to apply **K-Means** clustering on the baseline variables “Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age ONLY.

Also, in the given dataset, there is a lot of variation in the magnitude of the data. Variables like Pregnancies and DiabetesPedigreeFunction have a low magnitude whereas variables like Glucose BloodPressure etc. have a higher magnitude. So you are supposed to **standardize** the dataset as well.

Apply KMeans with the following criterion:

* Pick the initial centroids using kmeans++.
* Draw the elbow curve after applying the K-means algorithm by choosing the number of clusters N =1 to 7.
* After analyzing the elbow curve, apply the K-Means algorithm on the appropriate(best possible) value of N, and visualize the clusters.

1. In this problem, we want to see if there is any linear relationship between the two attributes “Glucose, Diabetes Pedigree Function, and Blood Pressure” in the **diabetes dataset**.

Apply Linear Regression with the following criterion:

* Use the hold-out Cross-Validation with 70 / 30 split.
* Calculate the root mean square error (RMSE).
* Visualize the relationship between “Glucose, Diabetes Pedigree Function and Blood Pressure”.

**Appendix**

# You may use these libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import statsmodels.formula.api as smf

from sklearn.model\_selection import cross\_val\_score, StratifiedShuffleSplit

from sklearn.model\_selection import train\_test\_split,KFold

from sklearn.neighbors import KNeighborsClassifier

from sklearn.tree import DecisionTreeClassifier

from sklearn.linear\_model import LinearRegression

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

from sklearn.preprocessing import LabelEncoder

from sklearn.metrics import classification\_report

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import accuracy\_score