FAST- National University of Computer & Emerging Sciences, Karachi.
FAST School of Computing,
Mid-1, Spring 2021
15th March, 2021, 09:00 am – 10:00 am

Course Code: CS481	**Course Name:** Data Science
Instructor Names: Dr. Muhammad Nouman Durrani, Muhammad Sohail Afzal	
Student Roll No:	**Section No:**

Instructions:

- Return the question paper
- Read each question completely before answering it. There are 5 questions on 2 pages
- In case of any ambiguity, you may make an assumption. But your assumption should not contradict any statement in the question paper.

Time Allowed: 60 minutes **Maximum Points**: 12.5 points

==========

Question 1: Briefly answer the following questions. **[0.5 x 5 = 2.5 Points]**

(i) The first step of a data science process is setting the research goals. What does it explain?

- In this step, we define the what, why and how of project which is basically the project charter. Understanding buiseness goals and context is essential so we do it in this phase. We also involve people with more expertise.

(ii) How do you deal with outliers if you find them in your dataset?

- Can replace with mean value of column or may remove them.

(iii) Discuss the most common transformations in the data preparation phase?

- Dimensionality reduction (reducing number of variables)
- Creating Dummy variables
- Derived measures
- Aggregating
- Extrapolating

(iv) Why cross validation techniques are used to gauge the effectiveness of a machine learning model?

- It measures performance of model by taking every instance one time in the test set. Because it might be possible that your model is good on some test instances but not many other so it gives you fair idea by checking model performance through putting all instances one by one in the test set.

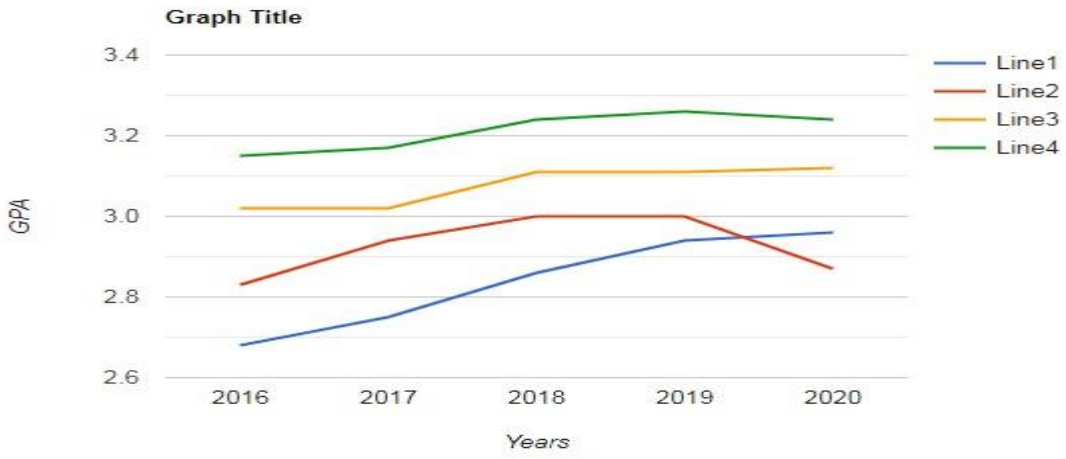
(v) How a classification problem is different from a regression problem?

- Both are types of supervised learning. Fundamentally, **classification** is about predicting a label/class **and regression** is about predicting a quantity. Eg. Iris flower type categorization is classification problem while while predicting house prices over the years is regression problem.

Question 2: Answer all the questions for the following Undergrade CGPA data of all CS batches over the last five years: **[1+ 0.5 + 0.5 = 2 Points]**

Batch	**2016**	**2017**	**2018**	**2019**	**2020**
Freshmen	2.68	2.75	2.86	2.94	2.96
Sophomores	2.83	2.94	3.00	3.00	2.87
Juniors	3.02	3.02	3.11	3.11	3.12
Seniors	3.15	3.17	3.24	3.26	3.24

(i) EDA usually employs summary statistics and graphical representations to get a better understanding of data. Use a suitable graphical exploratory data analysis technique to plot the performance trends of different batches.



(ii) Suppose a DataFrame df of the above data is already created. Display basic statistical characteristics (count, mean, std, min, max, etc.) of the numerical features of the above DataFrame using a python code.

- df.describe()

(iii) Does the given data show any indication of grade inflation over the years? Explain.

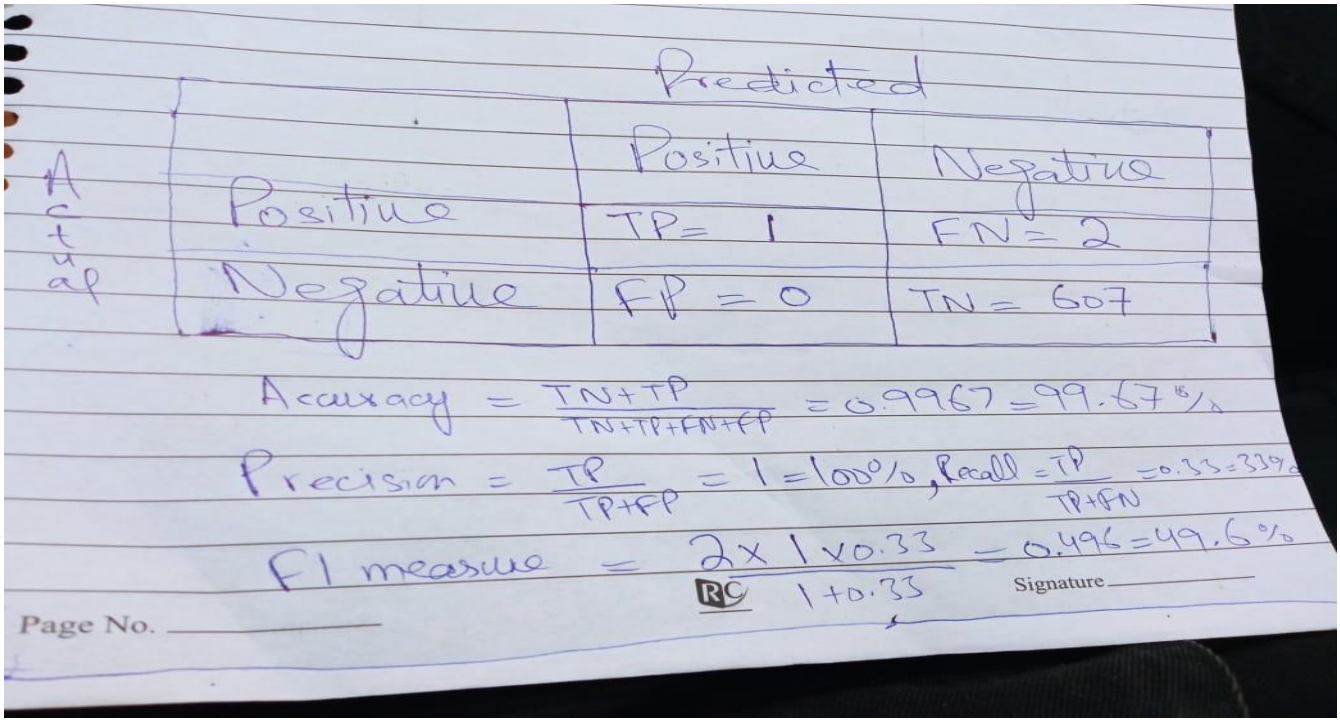
- Graph is inflating from 2016 to 2019 for all batches. However it keeps on increasing for freshmen and Juniors in 2020 as well but for sophomores and seniors, it decreased from 2019 to 2020.

Question 3: Assume that you have developed an ML model to predict whether an individual is at risk of having meningitis without carrying out dangerous intrusive tests. After an initial field testing of the model, you get the following results:

- All 607 people who did not have meningitis were identified as not having meningitis.
- Out of 3 who had meningitis, 2 were identified as not having meningitis and 1 as having Meningitis.

Create a confusion matrix of the above data and calculate accuracy, precision, recall, and F1 measures. **[2 Points]**

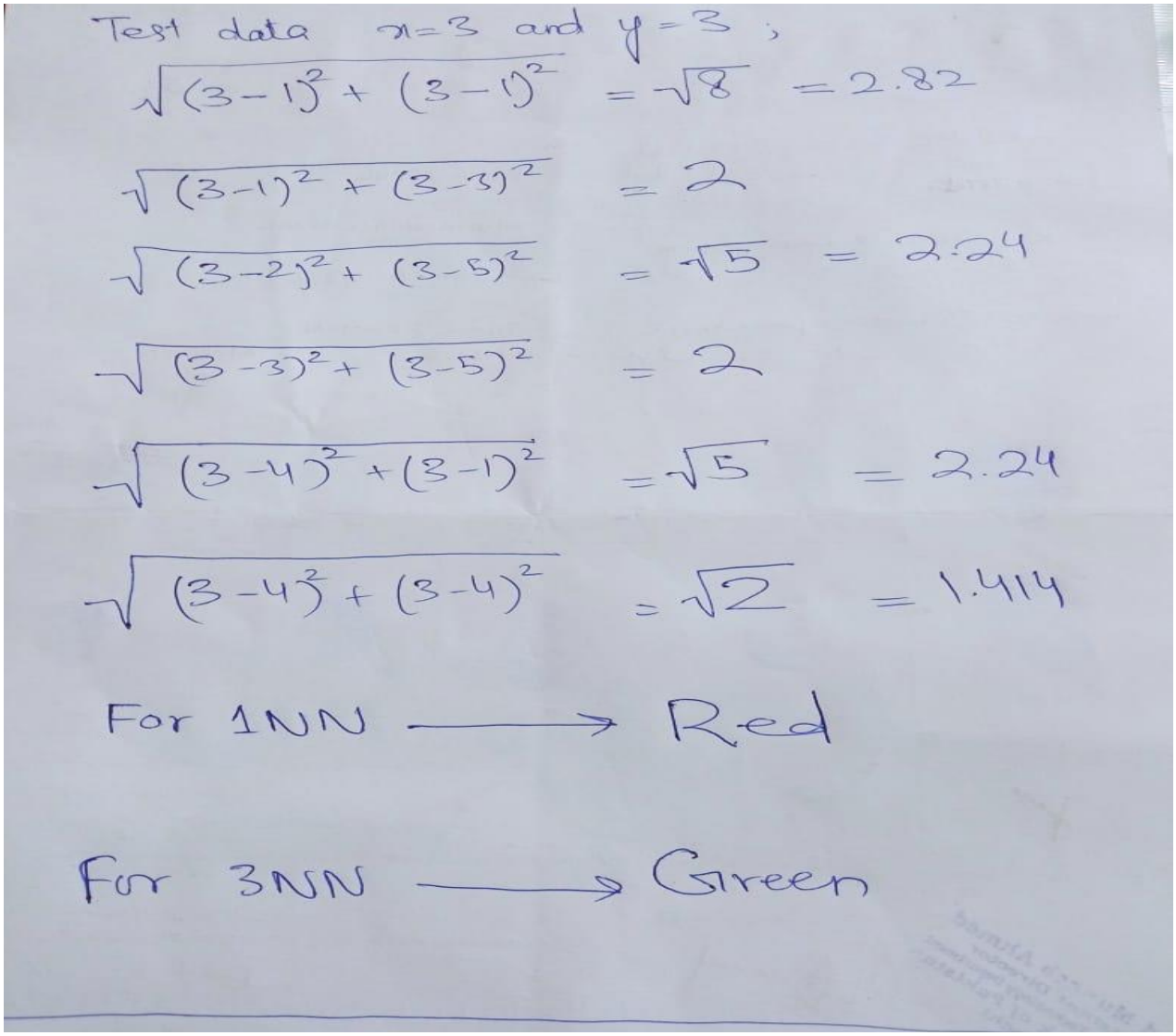
Solution:



Question 4: Suppose, you have the following data where x and y are the 2 input variables and color is the dependent variable. Suppose, you want to predict the most likely color of a new data point x = 3 and y = 3 using the Euclidian distance function n ($\sqrt{((x_1 - x_2)^2 + (y_1 - y_2)^2)}$) in a 1-NN and 3-NN. **[2 Points]**

x	y	color
1	1	red
1	3	green
2	5	blue
3	5	green
4	1	blue
4	4	red

Solution:



Question 5: Consider the following Table and answer all the questions using Python code:

[0.5*6 + 1 = 4 Points]

Name	Age	Mathematics	Physics	Chemistry	English
Ali	23	72	NaN	NaN	64
Mohsin	21.26	?	82.0	NaN	92
Muneeb	22	64	74.0	NaN	NaN
Zohaib	24	2000	96.0	NaN	76
Salman	26	81	79.0	NaN	81

Solutiion:

```
#part i
#Create a DataFrame out of the above data and take a look at the first 5 rows
import pandas as pd
import numpy as np
a = {"Name":["Ali","Mohsin","Muneeb","Zohaib","Salman"], "Age":[23,21.26,22,24,
26],"Mathematics":[72,"?",64,2000,81],"Physics":[np.NaN,82,74,96,79],"Chemistry
":[np.NaN,np.NaN,np.NaN,np.NaN,np.NaN],"English":[64,92,np.NaN,76,81]}
df = pd.DataFrame(a)
df.head()
```

```
#part ii
#dropping columns with all values as null values
df.dropna(how='all',inplace=True,axis=1)
df
```

```
#part iii
#Replace other missing values, outliers, and invalid values with a mean of thei
r respective columns.
df = df.replace('?',0)
df['Mathematics']=df.Mathematics.astype('int32')
df = df.replace(0,np.NaN)
df.loc[df.Mathematics > 100, 'Mathematics'] = np.nan
df = df.fillna(df.mean())
df
```

```
#part iv
#Change the data type of 'Physics' to int64
df['Physics']=df.Physics.astype('int32')
df
```

```
#part v
#Sort the DataFrame by 'Age' in ascending order.
df.sort_values('Age',inplace=True)
df
```

```
#part vi
#Add a new column "Total Marks' at the end of the DataFrame
df['Total Marks'] = df['Mathematics']+df['Physics']+df['English']
df
```

```
#part vii
#Extract marks of Ali and Salman in Mathematics and English only
df.set_index('Name',inplace=True)
print("Marks of Salman in English : ",df.loc['Salman']['English'])
print("Marks of Ali in English : ",df.loc['Ali']['English'])
```