National University of Computer & Emerging Sciences, Karachi

Spring 2021 CS-Department

Midterm # 2, 20th April 2021, 11:15 am – 12:45 pm

Course Code: CS481	**Course Name:** Data Science
Instructor Name: Dr Muhammad Atif Tahir	
Student Roll No:	**Section No:**

Instructions:

- Return the question paper.
- You are allowed to use PCs but all programs should be written in the answer sheet
- Read each question completely before answering it. There are 2 **questions and 2 pages**

Time: 90 minutes. **Max Marks**: 12.5 points

Question 1 [5 Points]: Complete the following program

```
import pandas as pd
data = {'names' : ['muhammad','faisal','','ayesha', 'f#araz'],
 'number' : ['2','3','4','', '6'], 'dob' : ['1/1/1990','2/1/1995','5/13/1996','1/2/2000','31/12/2001'],
     'salary' : ['20000','4000','4thousand','500000','2000'] }
```

(a) Convert data into dataframe object "frame1". Output should be as follows [0.5 Points]

	names	number	dob	salary
0	muhammad	2	1/1/1990	20000
1	faisal	3	2/1/1995	4000
2		4	5/13/1996	4thousand
3	ayesha		1/2/2000	500000
4	f#araz	6	31/12/2001	2000

(b) Replace empty string in column names with "FAST" [0.5 Points]

(c) Remove # from faraz [0.5 Points]

(d) Convert column number into int and replace empty field with mean [1 Point]. Hint: look for help("pandas.to_numeric")

(e) Represent 4thousand as numeric 4000 [0.5 Points]

(f) Convert dob from string into datetime format and out of range should be displayed as "NaT" [1 Point]. Hint: look for help('pandas.to_datetime') Expected output after (b), (c), (d) , (e) and (f) is as follows

	names	number	dob	salary
0	muhammad	2.00	1990-01-01	20000
1	faisal	3.00	1995-02-01	4000
2	FAST	4.00	1996-05-13	4000
3	ayesha	3.75	2000-01-02	500000
4	faraz	6.00	NaT	2000

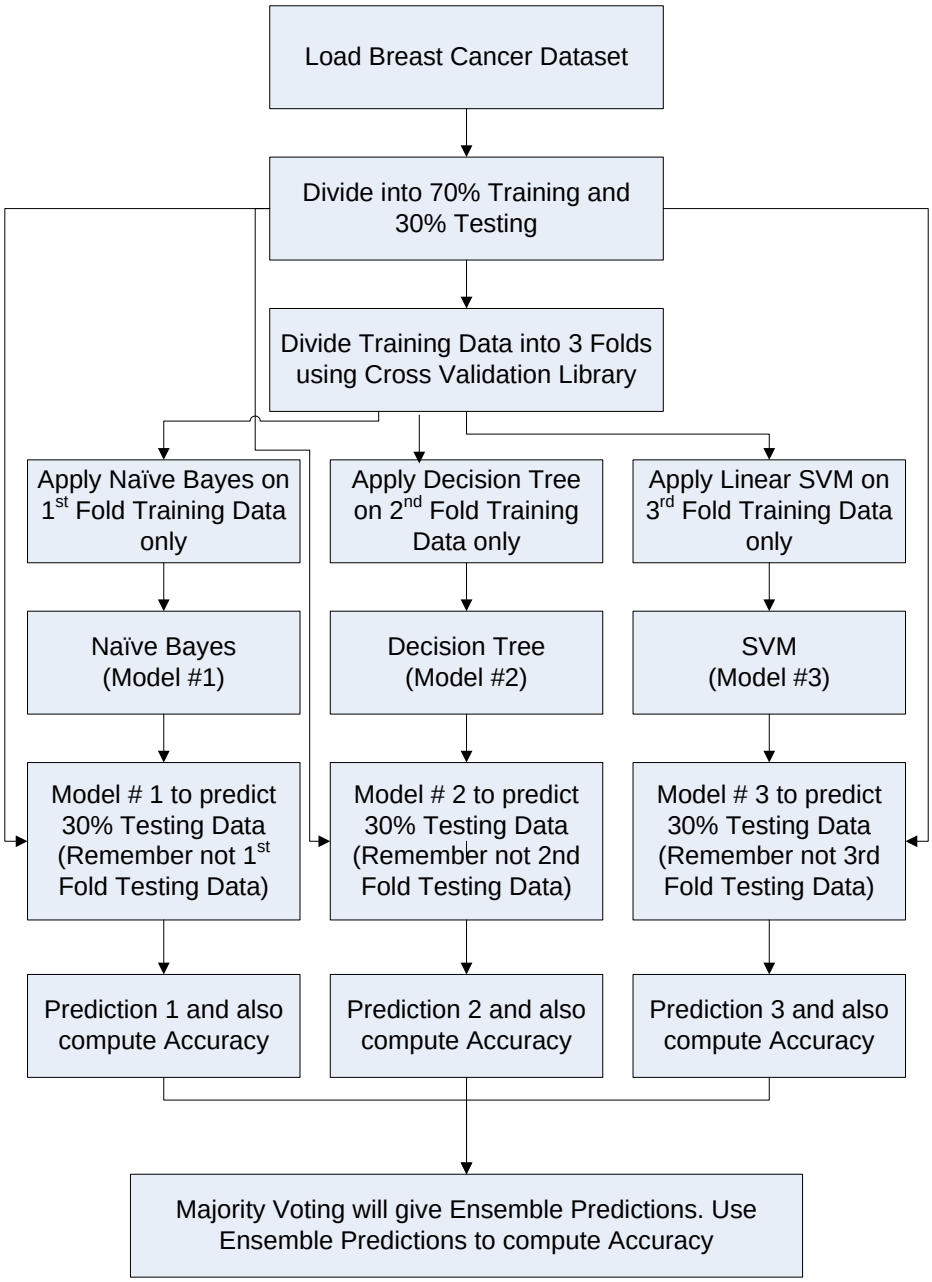
(g) Print the mean, standard deviation, median and maximum value in column salary [1 Point]

Question 2 [7.5 Points]: Implement the model shown in Figure below. Output should be as follows although accuracies can vary

```
Size of Training Data
(398, 30)
Size of Testing Data
(171, 30)
Fold1: Accuracy using Naive Bayes: 0.935672514619883
Fold2: Accuracy using Decstion Tree 0.8830409356725146
Fold3: Accuracy using SVM: 0.9590643274853801
Accuracy using Majority Voting: 0.9590643274853801
```

Consider the following Initial Coding:

```
# Load libraries
from sklearn.metrics import accuracy_score
from sklearn import tree
from sklearn import svm
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.datasets import load_breast_cancer
breast_data = load_breast_cancer()
#…… Complete the program as explained in block diagram.
```



For majority voting, see help about voting classifier
help('sklearn.ensemble.VotingClassifier')

BEST OF LUCK!