Spring 2022 BBL536E Homework 2

Remarks:
Write the code yourself. ***Cheating is strictly forbidden***.
For each problem write your code in the function format and give the names of the functions as problem numbers, for example for the solution of problem1:

```
def problem1(input):
  return something
```

Put the codes for all problems into one file (jupter notebook file) and name that file using your student username in the following format: badays_bbl536e_homework2.ipynb. The notebook file should definitely contain the outputs of the functions, if applicable. Sample solution file (sample_solution.ipynb) is given to you to show how to organize your solutions.

Give as much as documentation for your script using comments.
Create a report ("homework2_report.pdf") for the results you are asked in the problems.

Problem.1 (35 Points)

You are given "fitbit.csv" file containing information activities carrried out during the day and burned calories estimate throughout the day. We are going to select important features affecting calories burned.

First, using mutual information score find the top-4 features having the highest relationship with the target ("calories burned" column).

Second, using Recursive Feature Elimination method with Ridge regressor find top-4 feature combination.

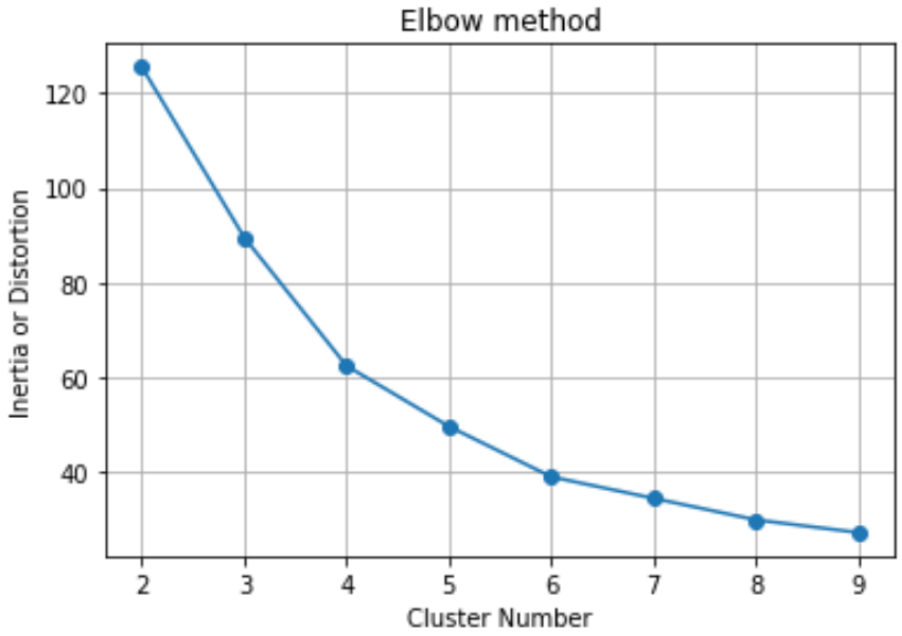
Note that you should drop the "Date" column. Your output should look like the following:

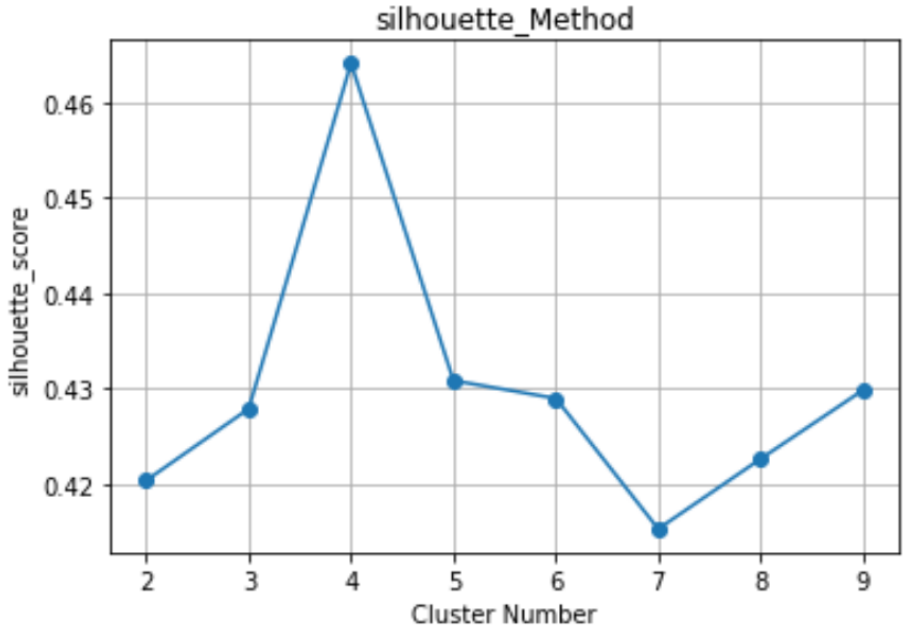
```
Selected features having top mutual information scores
['Steps' 'Distance' 'Minutes Fairly Active' 'Activity Calories']
Selected features by Recursive Feature Elimination
['Distance' 'Minutes Lightly Active' 'Minutes Fairly Active'
 'Minutes Very Active']
```

Problem.2 (35 Points)

You are given "customer.csv" file containing information to be used for the segmentation of customers. Cluster customer information using K-means algorithm. Produce Elbow and silhouette coefficient plots for cluster numbers ranging from 2 to 9. Before clustering apply standart scaling to the attributes. You can drop the "ID" column.

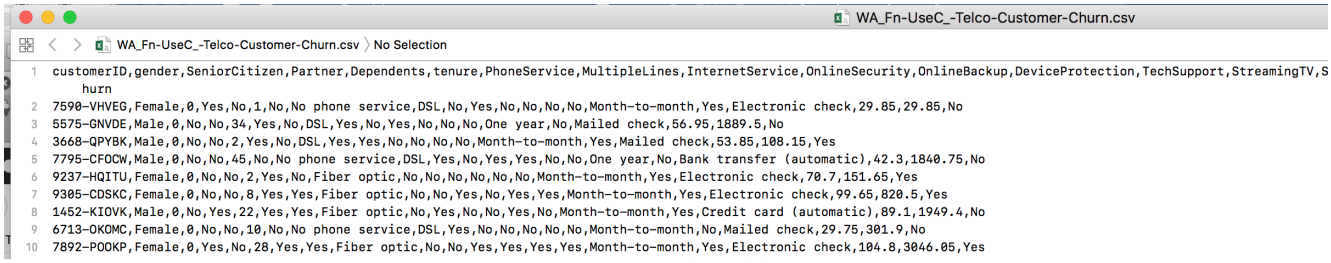
```
problem2("homewor2-data/customer.csv")
```





Problem.3 (30 Points)

You are given "WA_Fn-UseC_-Telco-Customer-Churn.csv" file on Telcom customer churn taken from https://www.kaggle.com/blastchar/telco-customer-churn. You can find more information about the columns in the link provided.


WA_Fn-UseC_-Telco-Customer-Churn.csv

WA_Fn-UseC_-Telco-Customer-Churn.csv 〉 No Selection

```
1  customerID,gender,SeniorCitizen,Partner,Dependents,tenure,PhoneService,MultipleLines,InternetService,OnlineSecurity,OnlineBackup,DeviceProtection,TechSupport,StreamingTV,S
       hurn
2  7590-VHVEG,Female,0,Yes,No,1,No,No phone service,DSL,No,Yes,No,No,No,No,Month-to-month,Yes,Electronic check,29.85,29.85,No
3  5575-GNVDE,Male,0,No,No,34,Yes,No,DSL,Yes,No,Yes,No,No,No,One year,No,Mailed check,56.95,1889.5,No
4  3668-QPYBK,Male,0,No,No,2,Yes,No,DSL,Yes,Yes,No,No,No,No,Month-to-month,Yes,Mailed check,53.85,108.15,Yes
5  7795-CFOCW,Male,0,No,No,45,No,No phone service,DSL,Yes,No,Yes,Yes,No,No,One year,No,Bank transfer (automatic),42.3,1840.75,No
6  9237-HQITU,Female,0,No,No,2,Yes,No,Fiber optic,No,No,No,No,No,No,Month-to-month,Yes,Electronic check,70.7,151.65,Yes
7  9305-CDSKC,Female,0,No,No,8,Yes,Yes,Fiber optic,No,No,Yes,No,Yes,Yes,Month-to-month,Yes,Electronic check,99.65,820.5,Yes
8  1452-KIOVK,Male,0,No,Yes,22,Yes,Yes,Fiber optic,No,Yes,No,No,Yes,No,Month-to-month,Yes,Credit card (automatic),89.1,1949.4,No
9  6713-OKOMC,Female,0,No,No,10,No,No phone service,DSL,Yes,No,No,No,No,No,Month-to-month,No,Mailed check,29.75,301.9,No
10 7892-POOKP,Female,0,Yes,No,28,Yes,Yes,Fiber optic,No,No,Yes,Yes,Yes,Yes,Month-to-month,Yes,Electronic check,104.8,3046.05,Yes
```

Predict customer churn with Logistic Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbor and Neural Network methods. For the parameters of these models use sklearn default parameters. For cross-validation use 5-Fold cross validation with shuffling. When preparing the data for prediction you can drop the customer ID. TotalCharges column has some missing data, drop the rows having missing data in the totalcharges column. Print the average accuracy scores for training and test splits of cross validation as given below. Note that you may get different scores since we use shuffling the data in splitting step.

model	train	test
LogisticRegression	0.806	0.805
DecisionTree	0.998	0.727
LinearSVC	0.742	0.735
KNN	0.832	0.763
MLPClassifier	0.797	0.795