## **Assignment 1 report (maximum 5 pages)**

Full Name: Jingyun He

Student ID: 530416562

**1. Data Preprocessing (2 points)**

Describe any data preprocessing techniques that you conducted, with justification (and data visualisation if needed)

In the data preprocessing section, I use the MinMaxScaler function in the sklearn.Preprocessing library to keep the normalization of dataset.

I need normalization to help me to transform the features (X\_train, X\_test) to be on a similar scale. That can improve the performance and stability of the training later.

**2. Methodology (6 points)**

***2.1. K-nearest Neighbor***

Describe the theory behind your chosen method

K-nearest Neighbor is a supervised learning classifier to make classifications or predictions. It is based on the Euclidean and Manhattan Distance formula to predict the labels by finding k nearest labels.

***2.2. Logistic Regression***

Describe the theory behind your chosen method

Logistic Regression is a supervised learning classifier to predict the probability of a target variable. It essentially assumes that the data obeys the distribution, and then use the most likelihood to estimate the parameters.

***2.3. Decision Tree***

Describe the theory behind your chosen method

Decision Tree is a supervised learning classifier, which is utilized for both classification and regression tasks. It is a tree structure in which each non-leaf node corresponds to a test for the values of an attribute, each branch corresponds to an attribute value, and each leaf node assigns a class. It makes new data get its own classification through the structure.

***2.4. Random Forest***

Describe the theory behind your chosen method

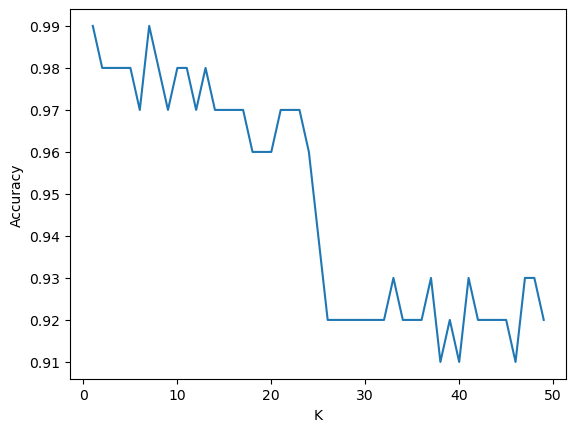
Random Forest is composed of many decision trees, and there is no correlation between different decision trees. It judges and classifies all the decision trees in the forest, and then votes for the classification with the most results.

**3. Performance, Evaluation & Comparison (9 points)**

***3.1. K-nearest Neighbor***

Explain the result and justify your answer (with graphs or tables or some reference)

I use cross\_val\_score function to test the value of k from 1 to 50, and get the line chart as below:



According to the chart above, the accuracy will be the highest when k = 8. Then, I use this hyper-parameter and K-nearest Neighbor classifier to make the prediction of y. Finally, I get the accuracy about the KNN is 0.84.

***3.2. Logistic Regression***

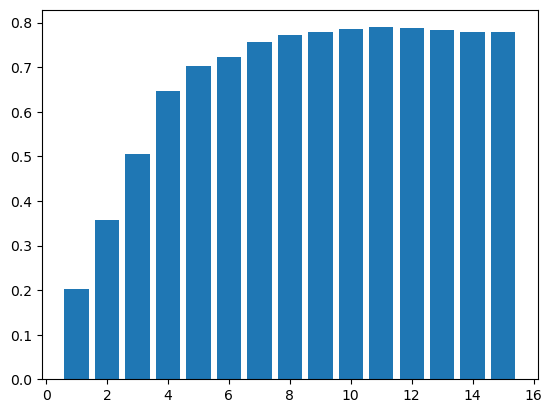
Explain the result and justify your answer (with graphs or tables or some reference)

I use cross\_val\_score function to repeat testing data 10 times and get the accuracy about Logistic Regression is 0.85.

***3.3. Decision Tree***

Explain the result and justify your answer (with graphs or tables or some reference)

I use cross\_val\_score function to test the max\_depth from 1 to 15, and get the bar chart as below:



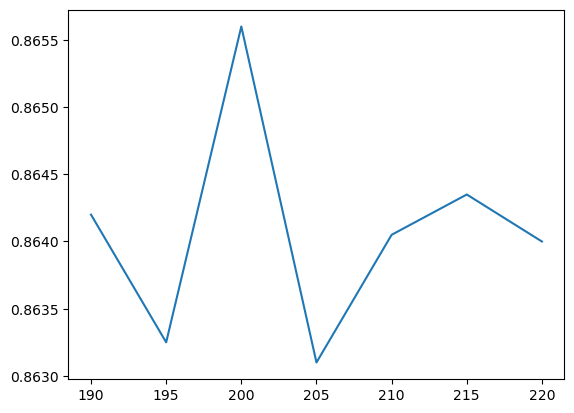
According to the bar chart above, the accuracy will be the highest when max\_depth = 11. So, I use this hyper-parameter and Decision Tree classifier to make the prediction of y. Finally, I get the accuracy about the Decision Tree is 0.79.

***3.4. Random Forest***

Describe the theory behind your chosen method

I use cross\_val\_score function to test the n\_estimators from 190 to 220 and step to 5, and the result chart will be displayed after the paragraph.

According to the chart, the accuracy will be the highest when n\_estimators = 200. So, I use this hyper-parameter and Random Forest classifier to make the prediction of y. Finally, I get the accuracy about the Random Forest is 0.88



***3.5. Comparison***

Compare the results and justify your answer (with graphs or tables or some reference)

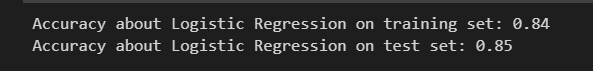
After the calculation:

the accuracy about K-nearest Neighbour is 0.84

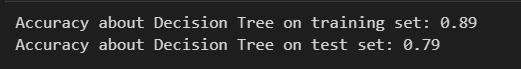
Graphical user interface

Description automatically generated

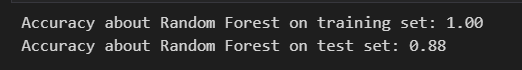
the accuracy about Logistic Regression is 0.85



the accuracy about Decision Tree is 0.79



the accuracy about Random Forest is 0.88



**4. Conclusion (1 points)**

Summarize your methodology and results, select the best model and justify your selection.

In conclusion, the accuracy about Random Forest with the n\_estimators = 200 is the highest, which is 0.88. Hence, Random Forest with n\_estimators = 200 is the best model.

**\*Writing Style (2 points)**

*Your writing is concise, clear and free of grammatical and spelling errors. You use appropriate technical terminology. Your paragraphs and sentences are well connected and follow a clear logic. There is a distinction between the essential parts of the report and less important material (use the appendix).*