Programming for Data Analytic
SOFT8032
First Assessment

November 2021

1 Third Assessment. Second Project

This project contributes 50% in your final mark. This is an individual project and has to be all done by yourself. You may be called for a zoom meeting to explain different parts of your submission, if needed.

Any question regarding the project should be communicated with farshad.toosi@mtu.ie or Canvas message.

1.1 Dataset Overview

For this project we are going to perform a number analytic tasks on the **weatherAUS.csv** file.

1.2 Project Specification

The objective of this project is to provide an insight into the underlying pattern of the dataset such as relationship between features, feature prediction and etc. Please perform the following tasks:

1. **Task1:** Consider a sub-dataFrame with four attributes: *MinTemp*, *WindGustSpeed*, *Rainfall* and *RainTomorrow*. Consider *RainTomorrow* as the class attribute. Create four datasets as follows:

   (a) *MinTemp*, *WindGustSpeed*, *Rainfall* and *RainTomorrow* as class attribute.

   (b) *MinTemp*, *WindGustSpeed*, and *RainTomorrow* as class attribute.

   (c) *MinTemp*, *Rainfall* and *RainTomorrow* as class attribute.

   (d) *WindGustSpeed*, *Rainfall* and *RainTomorrow* as class attribute.

   For each of the above subsets, run Decision Tree Classifier 35 times. At each run, try a different value for *max_depth*, ranging from 1 to 35. Collect the test accuracy at each run and visualize them using simple plot for each dataset and answer the following questions:

   (a) Which dataset has a better accuracy in general?

   (b) Which attribute has a more important role in predicting the *RainTomorrow*.

(c) What value approximately is an appropriate value for *max_depth* and why?

Use comment in the body of Task1 and answer each of the above questions. Note that, accuracy has to be measured on the test data where each dataset is divided into train (67%) and test (33%).

The visualization should contain 4 visual objects (each for one dataset). X axis specifies the values of *max_depth* ([1, 35]) and Y axis specifies the test data accuracy.

Data Cleansing: Only remove the rows where there is at least one empty cell in one of the 4 aforementioned columns.

2. **Task2:** Create a dataset with three attributes as follows:

   (a) *Pressure*. This attribute does not exist in the csv file and you need to create it by making the average of *Pressure9am* and *Pressure3pm* attributes. E.g., The average of the first cell of *Pressure9am* and the fist cell of *Pressure3pm* would be the first cell of *Pressure* and so on.

   (b) Humidity. This attribute does not exist in the csv file and you need to create it by making the average of *Humidity9am* and *Humidity3pm* attributes. E.g., The average of the first cell of *Humidity9am* and the fist cell of *Humidity3pm* would be the first cell of *Humidity* and son on.

   (c) *RainToday*. This attribute exists in the csv file.

   Use *Pressure* and *Humidity* as oridnary attributes and *RainToday* as the class attribute.

   Apply the following learning algorithms on your dataset:

   - KNeighborsClassifier
   - DecisionTreeClassifier
   - GaussianNB
   - SVM
   - RandomForestClassifier

   Apply cross-validation technique, when test data is %33 of the whole data, and conclude which technique is the best one and why? Show the averaged train and test accuracy for each model over all the cross-validation iterations separately using a bar chart see Figure 1 as a sample. Please add appropriate visualization features, e.g., legend, etc. Please do NOT use any argument for any of the above learning algorithms e.g., *max_depth* etc.

   Data Cleansing: Only remove the rows where there is at least one empty cell in the initial 5 aforementioned columns.

   Note that this task might take relatively long time to execute.

3. **Task3:** Create a dataset with five attributes as follows: *WindSpeed9am*, *WindSpeed3pm*, *Humidity9am*, *Humidity3pm* and *MinTemp*. The first four attributes are ordinery attributes (*WindSpeed9am*, *WindSpeed3pm*, *Humidity9am*, *Humidity3pm*) and *MinTemp* is the class attribute. Since *MinTemp* contains continuous values, therefore, discretization process is required for *MinTemp*.

Apply a k-nearest neighbours classifier where *n_neighbors* is equal to 5 and explore the following question: What is the best number of bins for discretization over the following number of bins: 2, 3, 4, 5, 6

Your judgment to answer the above question should be based on the absence or minimum overfitting and high test accuracy. For this question we assume there is no overfitting if the difference between train and test accuracy is less than 10%.

Use comment in the question and interpret your finding and applied methodology.

Data Cleansing: Only remove the rows where there is at least one empty cell in the initial 3 aforementioned columns.

4. **Task4:** Create a dataset with four attributes as follows: *Temp9am*, *Temp3pm*, *Humidity9am* and *Humidity3pm*. Apply an unsupervised algorithm on this dataset (K-Means) as follows:

   (a) All attributes need to be descritized into two bins.

   (b) Apply K-Means on the dataset using different number of clusters: ([2, 3, 4, 5, 6, 7, 8]).

   (c) Use an appropriate visualization technique and report which number would be the best number of clusters.

   (d) Repeat the last two steps and this time, descritize all the attributes into three bins.

   Note: This task need to have two separated plots. Use comment and explain how many clusters is more appropriate when the number of bins is two and when the number of bins is three.

   Data cleansing: Only remove the rows where there is at least one empty cell or non numerical value in any of the four aforementioned attributes.

Note that visualization plots need to have proper labels and annotations.

1.3 Submission and Deadline

Please use the python file template that is provided for you and complete your project in that file. Each task needs to be implemented as a separated function with interpretation as a comment below the function.

Please write your name and student ID as a comment in the designated area in the provided template python file.

The template file should be re-named at the end using your student ID followed by letter s, for example if your student ID is: 1234567 the the python file should be named: s1234567.py

The deadline for this project is 15th Dec 2021. One-week late submission with 10 marks penalty would be accepted and the deadline would be 22nd Dec 2021. Two weeks late submission with 20 marks penalty would be accepted and the deadline would be 29th Dec 2021.

Any question about this project should be communicated with Farshad Ghassemi Toosi farshad.toosi@mtu.ie or via Canvas.

Please submit your project via Canvas.

1.4 Rubric

This rubric is subject to change.

1. Correct task implementation (model training, accuracy reporting, visualization if needed etc). (100%)

2. Relatively correct task implementation (model training, accuracy reporting, visualization if needed etc). (70%)

3. Partly correct task implementation (model training, accuracy reporting, visualization if needed etc). (40%)

4. Wrong task implementation. (0%)

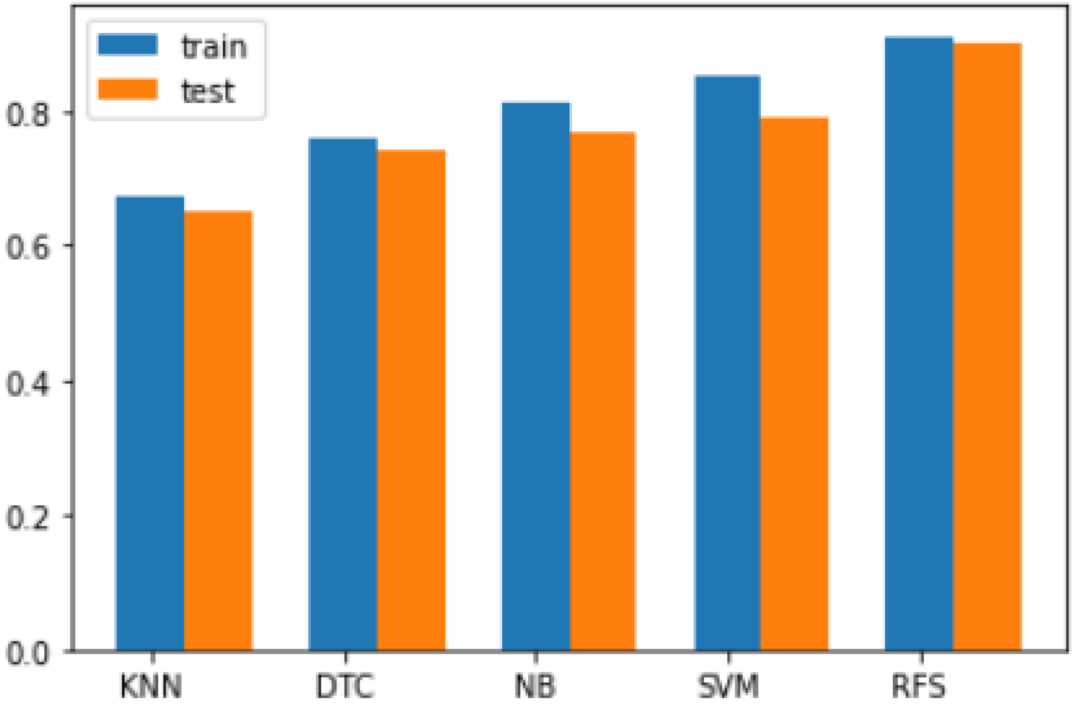

Figure 1: Bar Chart