Lab8-Decision Tree or Random Forest

You can select one of these options and selecting between Decision Tree or Random Forest models (please note only 8 marks will be given for this assignment even if you submit both of them)

Option 1 – Decision Tree) Total Marks: 8 **Marks + 2 (individual assessment) = 10 Marks**

In this part you will use Python to analyze the heart disease data set (the link and explanation is included here) by training and building a model with **Decision Tree**.

Heart Disease Dataset: Here, is the link for heart disease dataset of patients.
http://archive.ics.uci.edu/ml/datasets/Heart+Disease
After going to this link you will find two folders: One: Data Folder and two: Dataset description. Data folder that has the dataset. It is better to use processed cleveland data. In the dataset description folder, you will find the description about the columns' names referring to the14 column of the dataset as the following: The last one attribute (number 14) is the result. Include your R source code of regression analysis, training and generating results. Here are the example of attributes and their Information (please see data set documents for more details)

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps) 5. #12 (chol)

6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
.....
13. #51 (thal)
--
14. #58 (num) ---------->result

Test your model and discuss the result of your test with performance metrics. Make sure you separate training set and testing data properly (Train:80 and Test:20) and implement the decision tree using Python for the followings:

1. Use **gini** measure quality of split and build the decision tree. Record the model accuracy.
2. Visualize the decision tree built in Question-1
3. Change max_depth=3 and rebuild the decision tree. Record the model accuracy.

4. Visualize the decision tree built in Question-3
5. Use **entropy** measure quality of split and build the decision tree. Record the model accuracy.
6. Visualize the decision tree built in Question-5
7. Compare all three decision trees using their accuracy

Tips:

- Refer lecture notes titled "Decision Tree" before starting this assignment
- Use Python library "sklearn" to construct the decision tree
- For visualization, you need to the followings:
    - To install **graphviz**, download the Windows executables from from https://graphviz.org/download/
    - To install **pydotplus**, you need to open the *Command Prompt* window by clicking the Start button. In the search box, type *Command Prompt*, and then, in the list of results, click Command Prompt. At the command prompt, type pip install Graphviz.

OR Option2: Random Forest

Total Marks: **8 Marks**

In this part you will use Python to analyze the heart disease data set (the link and explanation is included here) by training and building a model with **Random Forest**.

Heart Disease Dataset: Here, is the link for heart disease dataset of patients.
http://archive.ics.uci.edu/ml/datasets/Heart+Disease
After going to this link you will find two folders: One: Data Folder and two: Dataset description. Data folder that has the dataset. It is better to use processed cleveland data. In the dataset description folder, you will find the description about the columns' names referring to the14 column of the dataset as the following: The last one attribute (number 14) is the result. Include your R source code of regression analysis, training and generating results. Here are the example of attributes and their Information (please see data set documents for more details)

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps) 5. #12 (chol)

6. #16 (fbs)
7. #19 (restecg)

8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
.....
13. #51 (thal)
--
14. #58 (num) ---------->result

Test your model and discuss the result of your test with performance metrics. Make sure you separate training set and testing data properly (Train:80 and Test:20) and implement the decision tree using Python for the followings:

1. Use **gini** measure quality of split and build the random forest model. Record the model accuracy.
2. Use the random forest feature_importance to find the top 5 important features
3. Rebuild the random forest using the top 5 features and record the model accuracy
4. Using the max_depth=3 and the top 5 features, rebuild the random forest model. Record the model accuracy.
5. Visualize the random forest built in Question-4
6. Use **entropy** measure quality of split, max_depth=3 and top 5 features, build the random forest. Record the model accuracy.
7. Visualize the random forest built in Question-6
8. Compare all random forest models so far you constructed using their accuracy

Tips:

- Refer lecture notes titled “Random Forest” before starting this assignment
- Use Python library “sklearn” to construct the random forest
- For visualization, you need to the followings:
    - To install **graphviz**, download the Windows executables from from https://graphviz.org/download/
    - To install **pydotplus**, you need to open the *Command Prompt* window by clicking the Start button. In the search box, type *Command Prompt*, and then, in the list of results, click Command Prompt. At the command prompt, type pip install Graphviz.