CMPE 442/CS 542 - Assignment 3

NOTE: *This assignment is equivalent to the project so requires you to write a report which includes all the sections described below. You do not have to submit your codes and you can use any libraries.* ***Follow the steps that are given in Chapter 2 of the Aurelien Gerone book (uploaded to Moodle).***

This work is to be done individually. Each of you should select data where the associated task is a classification task with at least 10 attributes. You can select the data from the following site or any other place: http://archive.ics.uci.edu/ml/datasets.php

In your report plot all the graphs that you obtain, the methods that you use and discuss the results based on the notes listed below. Listed notes are general and might be missing some remarks. Basically you will not be mistaken if you discuss about everything that you do and also justify every decision that you make within the process.

1. **Problem Definition**
   Discuss about the problem at hand. The problem statement should be clarified. It should include the final goal of the project.

2. **Data Analysis**

   In this section discuss about the features that you have in your data.
   - What are these features?
   - What do they represent?
   - What are the types?
   - What are the statistical descriptions of the data?
   - Use scatter plots to visualize numerical features
   - What are the values that the categorical attributes take?
   - Are there missing values?
   - Which of these features do you think are redundant and should not be used for the model training?

3. **Data Processing**
   - Create a test set and put it aside (discuss about the method that you used to split the data)
   - How do you handle missing values?
   - How do you handle categorical features?
   - Do you need to normalize your data and how you normalize it?

4. **Model Selection and Training**
   Select different models and using K-fold cross-validation select the one with the best results. Report on the performance results you obtain for different models.

5. **Fine-tune Your Model**
   Fine tune your model selected from previous section. This step involves selecting the best hyperparameters for the model (use GridSearch, etc.)

6. **Testing**
   Test your model on the test set.
   Report the performance measurements (accuracy, precision, recall, F-score, etc).
   Analyse the test sample for which your model is not able to correctly classify.
   Why do you think this is happening?

7. **Summary**
   Summarize your findings.
   Discuss about the ways you could improve your mode.