Homework – 1. A Journey with Titanic

Frederik Darwin
M10601836

- **Explain what cause the difference of results between two algorithms which you choose?**

    I use Decision Tree Classifier and Gaussian Naïve Bayes as two of my algorithms used in the machine learning process. Decision Tree Classifier (DTC) by theory, is using rectangle-shaped separator to classify data. I won't dive deep into the theory, but DTC start in one root node and make its way down creating multiple end-node and leaf node. From this description you could easily see that every category of predictor that determines the target on DTC is dependent on each other, because a data can't go to specific node without other predictor helping it out.

    On the other hand, Gaussian Naïve Bayes is a classification techniques based on Bayes' Theorem with an assumption of independence among it's predictors. . In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

    Hence, the difference in treating the predictor data directly causes the different result between two of those algorithms and the score results. We can observe that DTC have higher compatibility in predicting the titanic dataset as it have higher score than GNB prediction score.

- **Compare your prediction results on training data against the prediction result on the testing set. Show the accuracy results and explain why they are different.**

    Test data is a set of data representing a real world "messy" data, which our trained model try to tackle after learning the data through training data we provide. The wilderness of real world data probably has plenty of different set of predictors that the training data hasn't seen so the model couldn't possibly predict correctly what is the right classification, in this particular problem.

- **Choose the best of your algorithms and tune it with best performance, and briefly specify how you do it.**

    Using DTC and GNB, firstly I observe the data using describe function and data visualization to see if there is any correlation between the column to each other and so forth. After seeing the correlation, I move to cleaning and completing the training and dataset, searching if there is any null data that is needed to classify the data, filling it with random number or simply just fill it with mode of the data provided. Cleaning the column that won't be used in the model is done too. After that, I make a new column from the existing column based on the correlation they both have that will have to do with the model learning. I transform some of the column that is needed for the model learning from alphabetical to numerical one so the model can easily learn it. Lastly, I use the machine learning algorithm to learn from training dataset and run the test dataset.