**Project 2:** Linear Regression & Classification Tree Report

**CSC 177-01**: Data Warehousing and Data Mining

**Professor**: Jagan Chidella

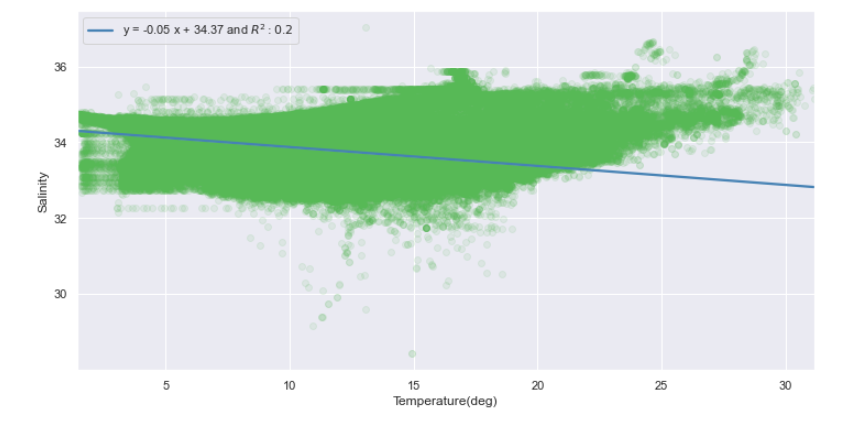
**Group**: Jason Phillips, Mohammad Ameri, Ryon Faroughi, Youser Alalusi, Yusran Sadman

**Introduction:**

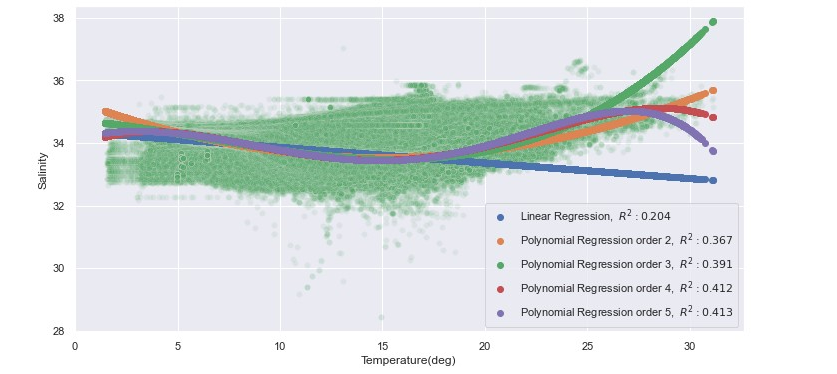
In this project, we took two datasets and applied both linear and multiple regression models to fit our data. We split our data into two sets in order to train the model and produce predicted values from our test set. Originally we were going to use the dataset that we produced in project one, however that dataset was inadequate for this assignment. There appeared to be no correlations between any of the attributes and transforming the data creatively would not make the dataset a better candidate for this project. This project will be broken down into three sections. The first section we will apply linear regression techniques to a water salinity and temperature dataset that is a better dataset for this project. The second section will cover the same linear regression techniques on a provided dataset for admissions. The final section will cover classification and the analysis of the decision tree.

**Regression on Water and Sanity Dataset:**

As mentioned previously, our first dataset was not suitable for this project because there was little to no correlation between any of the attributes. So as a result, we decided to move to an alternative dataset that focuses on the relationship between salinity and the temperature of water.

We began the linear regression portion of the assignment by splitting the data into the testing and training sets. Once the data was split, we were then able to fit the data to a line and produce a linear regression line along with the R-Squared value.

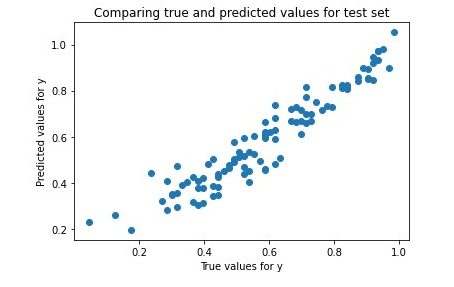
Once we established the relationship between the salinity of water with a single variable, we then observed the polynomial relationship between temperature and the water’s salinity. We can see that they provide a better R-Squared, but they result in a lower fit.



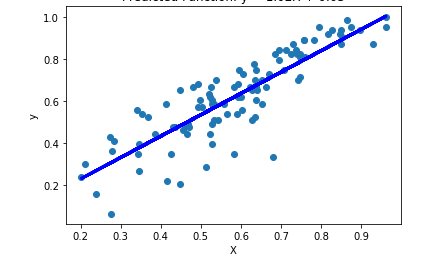
After we completed the first portion of the assignment we came to a conclusion that there was a correlation value of .45, which shows that there isa relationship there.

**Regression on Admissions Dataset:**

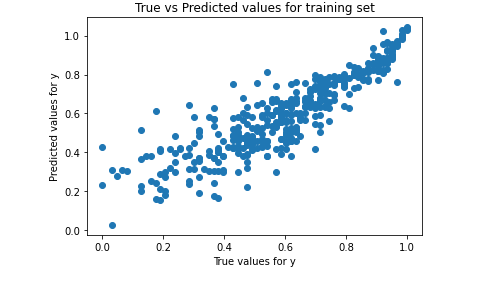
For this part we manually studied the data and made random observations. For example, serial numbers are in ascending order, TOEFL, university rating on a scale of 1-5, SOP, letters of recommendation, and research - either 0 if they have not done research and 1 if they have. For regression we first split the data into training data and test data. The data was split into 80:20 since the distribution of the data in the rows was unknown and we performed shuffling data frames on the data set before partitioning the data into the training data and test data. Then, we applied linear regression and multiple regression on both training and test data, after performing that on both datasets we got some results. Finally, when a simple regression was performed on both training data and the test data it can be seen that the curve that is created is very similar. There are more data points seen in the training data curve due to it being 4x the amount of data points. It's significant that the curves look similar because what can be seen is that with more data points the more accurate the linear regression line is.

**Simple Linear Regression:**

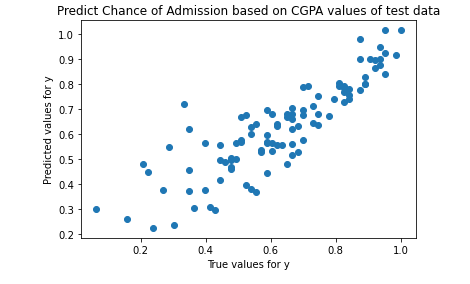
**Fit our training data and tested with test data. Predicted future y values from x values of the test set.**

****

**Here we plotted the test data, and the line is x variable of test and the predicted y values from the test data.**

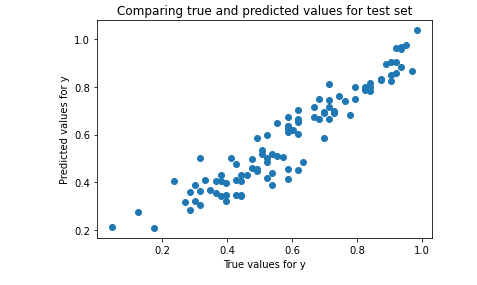
****

**Next we predicted y values from the training data.**

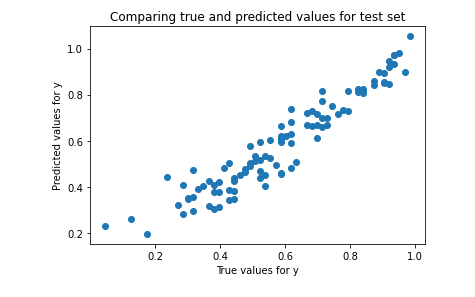
****

**Here we fit our test data and predicted y values from x values of our test data.**

**Multiple Linear Regression:**

****

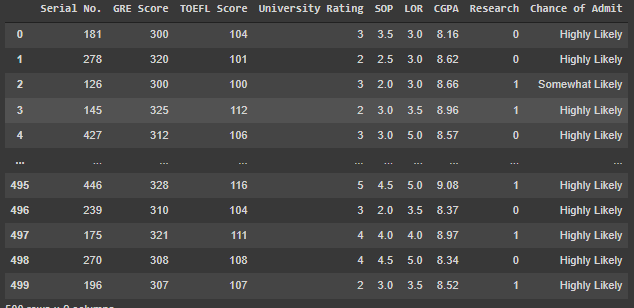
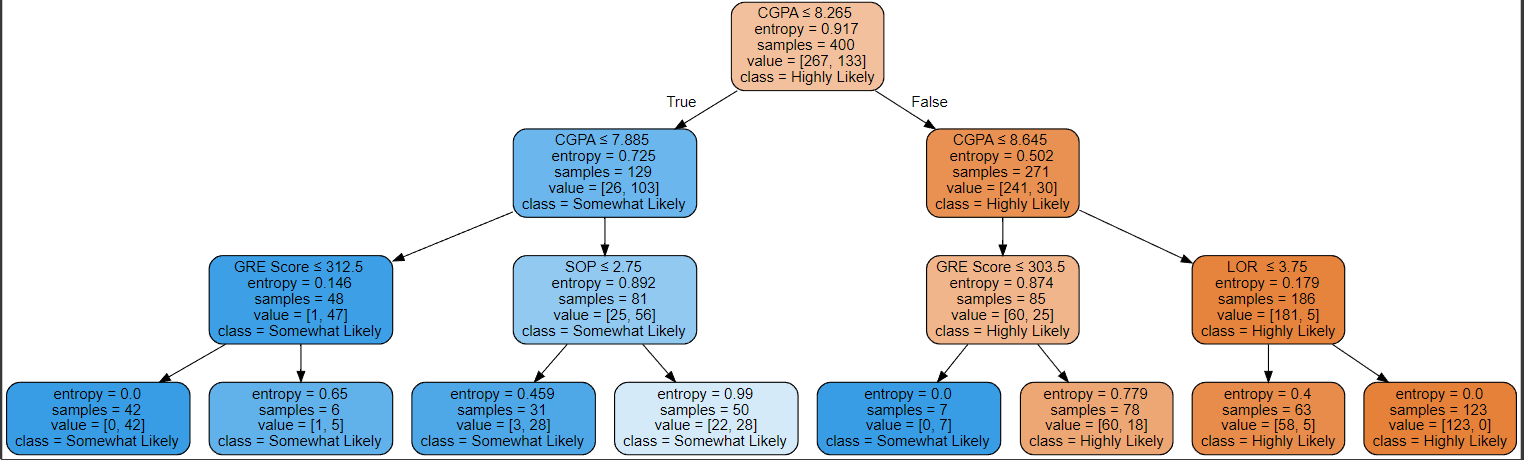
**Here we fit our training data and predicted y values from x values of our test data.**

****

**Lastly, we fit our test data and predicted y values from the x values of our test data.**

**Classification on Admissions Dataset:**

First, we discretize the last column which is “Chance of Admit” into three classes after shuffling. We classify 0-0.33 as low, 0.34-0.66 as mid, and 0.67-1.00 as high.

Then, the data was split 80:20 for training and testing. We dropped the serial number column since it is of no use and since the column we want to predict is chance of admit we set the column to Y and the rest to X. The tree’s criterion is entropy with max depth of 3 and the tree is also displayed with graphviz. 

Also, we were able to predict the classes based on our test data from the original dataset. The accuracy on the test data is 82%. This tree has three possible outputs; unlikely, somewhat likely, and highly likely. These outputs are generated based on the inputs from 5 other attributes; Gre scores, toefl score, Cgpa, Lor, and sop.

**Conclusions:**

After applying linear regression on both datasets and analyzing the results it becomes apparent how important it is to understand the relationship between values in the data and the impact that this relationship has on other attributes.The second portion of the project in particular does a great job at showing how the data has an effect on the admissions attribute. Since there is a strong correlation between these attributes, we can construct a reliable and accurate decision tree that will likely predict the proper outcomes of Admission given the other 5 data points.