**CE 475 Machine Learning Final Project Report**

Created by Ögeday ÖZTOPRAK

20150602028

**Identification of the Problem**

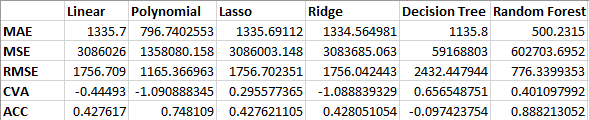
We have a dataset that include 120 data points on the columns named x1,x2,x3,x4,x5,x6 and also we have another column named Y and it only contains 100 data points. The mission was about deciding which regression model is better for estimating last 20 rows of the Y. I have used X columns as my train data.

**Methodology**

I have believed that deciding best model for this problem is trying several estimation methods and getting some result from them.Then comparing the results of the regression models and selecting the best model would be the best for estimating last 20 rows of Y.I had necessary researched about this calculations then I have realized that I need to use a selection algorithm before starting the process.I took a look at the sklearn library and found three selection algorithm named Pipeline,Tree Based Feature Selection,L1 Based Feature Selection and Recursive Feature Elimination, so I deci Recursive Feature Elimination for my calculations.According to results of the algorithm my most efficient first four columns for these calculations are; **x1,x2,x3 and x6.**I have used first 100 rows of the columns till to find the best estimation model.I have tried to use **Linear Regression,Polynomial Regression,Lasso Regression,Decision Tree,Ridge Regression and Random Forest Regression**. I collected their **Mean Absolute Error, Mean Squared Error,Root Mean Squared Error, Cross Validation Accuracy and Accuracy** results by using metrics library of the Sklearn.According the comparison of these results especially about the mean absolute error, Random Forest Regression is the best regression model to estimate 20 values for the Y.

**Results**

(Original version of the results attached to the report file.)

****

According to these calculations it is obvious that using Random Forest Regression to estimate the last 20 Y values is the best model because mostly Random Forest has smaller error values than the other models that I have used.