MIDTERM REPORT
COMPARISON ALGORITHMS USING -MSE & -R2

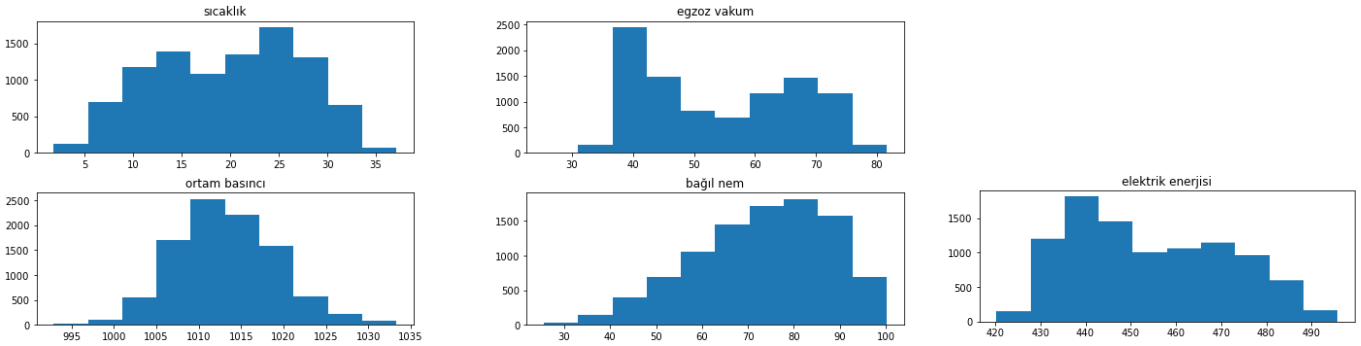
Group Members:

171805047 MURAT SAHİLLİ
171805065 ABDULLAH TAŞ

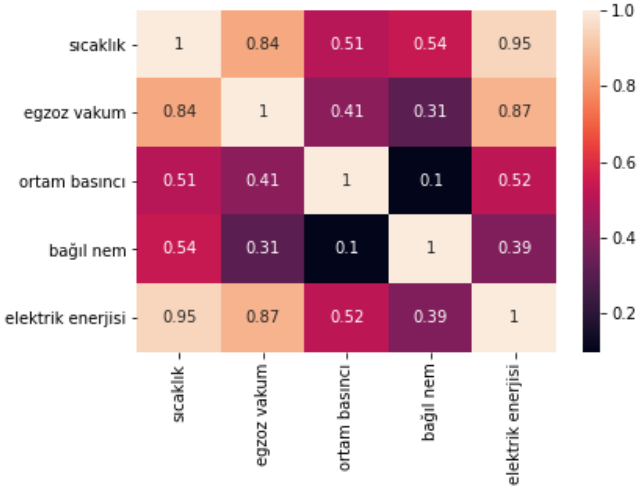
DATASET INFORMATION:

The dataset contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load. Features consist of hourly average ambient variables Ambient Temperature (AT), Vacuum (Exhaust Steam Pressure, V), Atmospheric Pressure (AP) Relative Humidity (RH) to predict the net hourly electrical energy output (PE) of the plant.
AT= Sıcaklık , V= egzoz vakum, AP= ortam basıncı, RH= bağıl nem, PE= elektrik enerjisi

Distribution of Data:

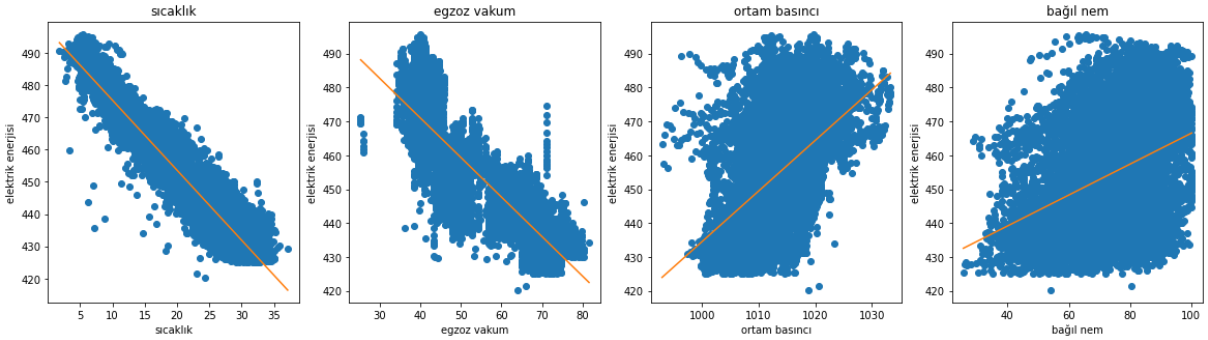


Heatmap of the variables in the dataset:



Features that most associated with the output value:
- sıcaklık
- egzoz vakum

Plotting data against output value:

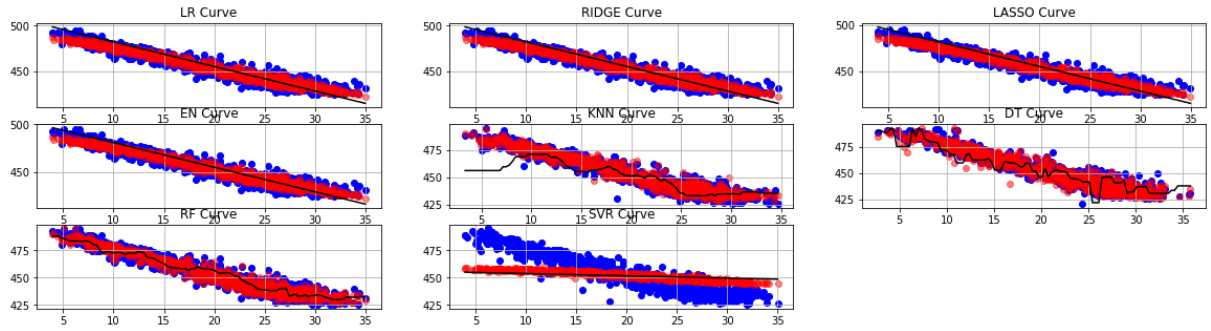


QUESTION 1: How did the use of raw and preprocessed data affect the learning outcome?

K-fold results with raw data:

The mse and r2 mean of each algorithm with using 10 fold cross validation.

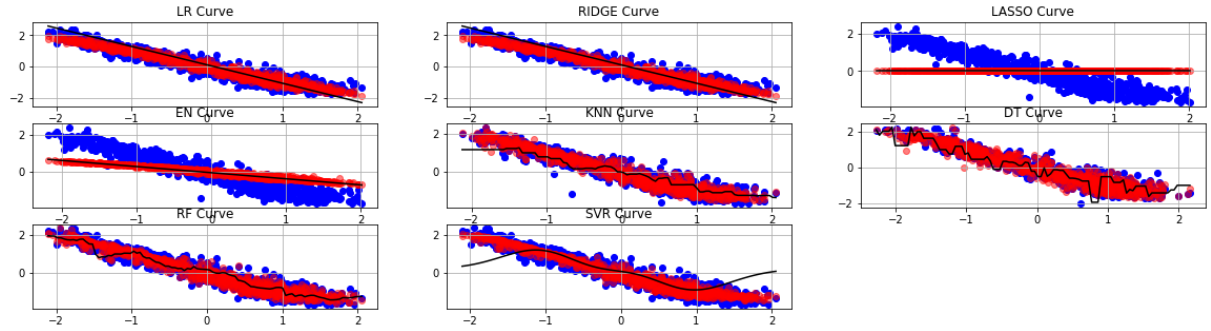
K-Fold	LR	RIDGE	LASSO	EN	KNN	DT	RF	SVR
mse-mean	20.9249	20.9249	20.9907	21.0735	16.3883	20.301	11.5755	188.65
r2-mean	0.927596	0.927596	0.927374	0.92709	0.943354	0.927063	0.959948	0.348451



K-fold results with preprocess data:

The mse and r2 mean of each algorithm with using 10 fold cross validation.

K-Fold-PrePro	LR	RIDGE	LASSO	EN	KNN	DT	RF	SVR
mse-mean	0.0718446	0.0718446	0.995125	0.473154	0.0519646	0.0704557	0.0396141	0.056203
r2-mean	0.927596	0.927596	-0.000827252	0.524242	0.947681	0.92875	0.960121	0.943351



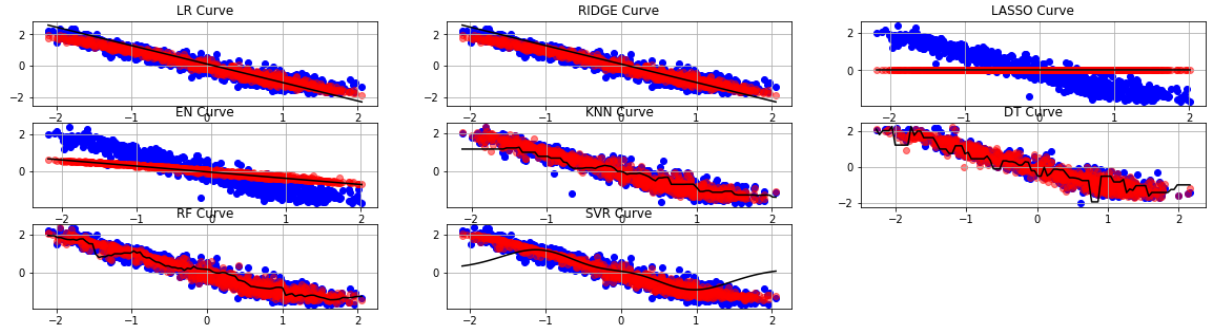
While there were not much changes in Linear, Ridge, KNN, Dtree and RandomForest between raw and preprocessed datas, there were big differences in Lasso, ElasticNet and SVR. Accuracy decreased in EN and Lasso with using preprocessed data, while in SVR the accuracy increased greatly.

QUESTION 2: How much learning performance did you achieve with preprocessed datasets?

K-fold results with preprocess data(all data):

The mse and r2 mean of each algorithm with using 10 fold cross validation.

K-Fold-PrePro	LR	RIDGE	LASSO	EN	KNN	DT	RF	SVR
mse-mean	0.0718446	0.0718446	0.995125	0.473154	0.0519646	0.0704557	0.0396141	0.056203
r2-mean	0.927596	0.927596	-0.000827252	0.524242	0.947681	0.92875	0.960121	0.943351

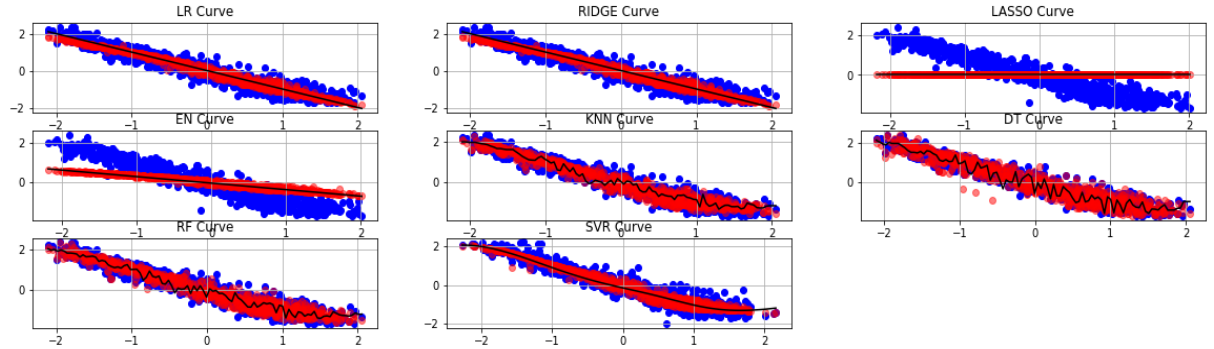


K-fold results with preprocess data(half of the features with the highest cross-correlation coefficient):

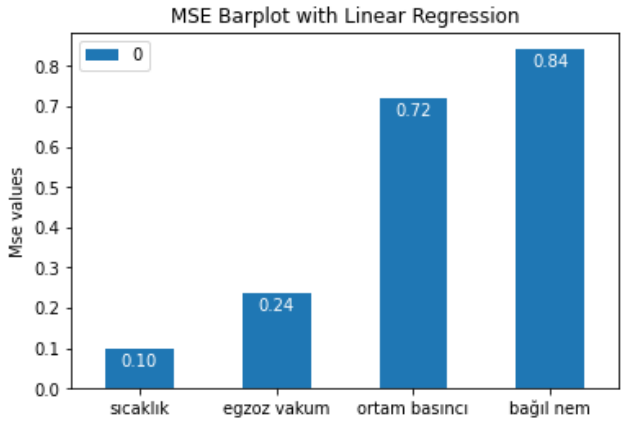
The fatures with highest cross correlation coefficient:
- sıcaklık
- egzoz vakum

The mse and r2 mean of each algorithm with using 10 fold cross validation.

K-Fold-Cross-PrePro	LR	RIDGE	LASSO	EN	KNN	DT	RF	SVR
mse-mean	0.0846132	0.0846132	0.995125	0.473154	0.0694932	0.084628	0.0498432	0.0694381
r2-mean	0.914759	0.914759	-0.000827252	0.524242	0.929983	0.916294	0.949695	0.929989



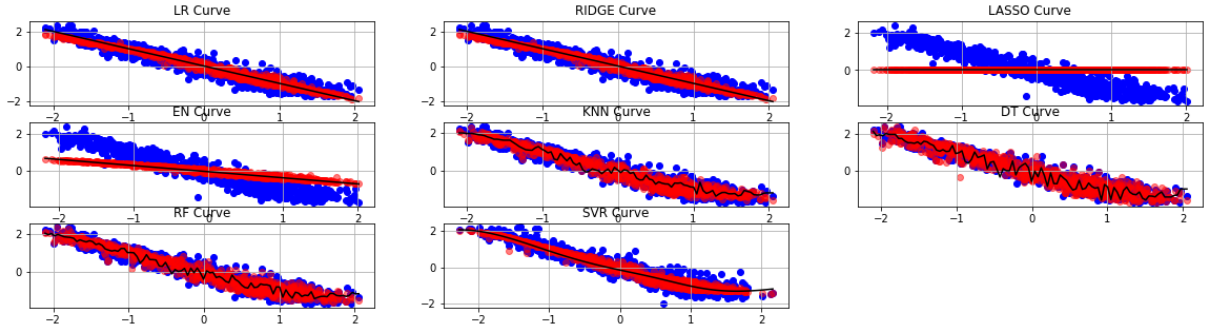
K-fold results with preprocess data(half of the features with the lowest mean square error):



The features with the lowest MSE:
- sıcaklık
- egzoz vakum

The mse and r2 mean of each algorithm with using 10 fold cross validation.

K-Fold-MSE-PrePro	LR	RIDGE	LASSO	EN	KNN	DT	RF	SVR
mse-mean	0.0846132	0.0846132	0.995125	0.473154	0.0694932	0.0833173	0.0496488	0.0694381
r2-mean	0.914759	0.914759	-0.000827252	0.524242	0.929983	0.91474	0.949851	0.929989



When all data are used in Linear, Ridge, KNN, DTree, RandomForest and SVR, the accuracy value is higher than the data divided in half. Taking features with high Cross Correlation coefficient values or features with low MSE values did not increase the accuracy.

QUESTION 3: Which algorithm learned better as a result of all operations ?

At the end of all operations, Random Forest algorithm realized the best learning. When all the data were used and the data was divided in half, the Random Forest algorithm realized the best learning.