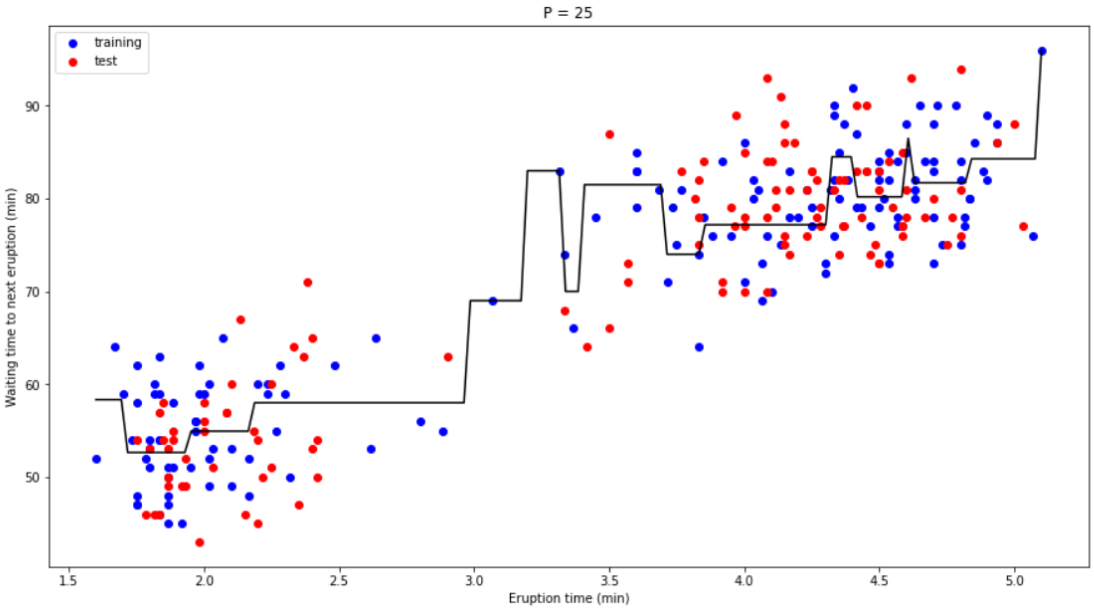
ENGR 421 / DASC 521: Introduction to Machine Learning
Homework 04: Decision Tree

İzel Yazıcı - 77549

I read the csv and split data into two dataset which first 150 rows of data as a train and last 122 rows of the data a test data. The data was a univariate regression dataset. It was about duration of the eruption and waiting time between eruption for a geyser.

I implemented decision tree formula by setting the pruning parameter P = 25. Implementation graph is the following.



The line graph that I have fitted with Decision Tree is not exactly as it should be, I think I am making a small mistake in the calculation, but I could not find the reason.

RMSE of Decision Tree

$$\sqrt{\frac{\sum_{i=1}^{N_{test}}(y_i-\hat{y}_i)^2}{N_{test}}}.$$

```
y_pred = [ Score_DT(X_test[i], is_terminal, node_splits, node_avg_value) for i in range(N_test)]
rmse = np.sqrt(np.mean((y_test - y_pred)**2))
print(f"RMSE is {rmse} when P is {P}")
```

```
RMSE is 6.4991477594218345 when P is 25
```

Decision trees by setting the pre-pruning parameter *P* to 5, 10, 15, …, 50.

RMSE for test data points as a function of *P* is the following. The RMSE values I have calculated for different pruning values give the result that should be.