Methodology

To predict the last 20 rows which does not have labels, I have implemented some Machine Learning algorithms such as Linear Regression, Logistic Regression, Desicion Trees and Neural Networks to predict their label value. To predict label value, we must train these models with first 100 rows. To train these models, we seperate our data as X(labels from x1 to x6) and Y(label) value. We will give these variables to train our model using .fit(X, Y) method provided by scikit-learn. I have used 4 different Machine Learning algorithms for this homework. They are as Linear Regression, Logistic Regression, Desicion Trees and Neural Networks. The reason why did i choose these models is, these models are very popular algorithms and widely common in Machine Learning area. I wanted to test these algorithms by training and predicting using same data. I will talk about pros and cons about these models.

* Linear Regression

Pros: Cons:

1.Easy to implement 1.Can overfit.

2.Less complexity. 2.Assumes data has linear relationship.

* Logistic Regression

Pros: Cons:

1.Easier to implement, train. 1.May overfit with lesser data.

2.No assumtion. 2.Construts linear boundries.

3.Can extend to multiclass. 3.Hard to use for complex problems.

* Desicion Trees

Pros: Cons:

1.Easy to understand, interpret. 1.More time to train model.

2.Can work with numerical and 2.Can overfit easily.  
categorical features. 3.Training is expensive and complex.

3.Requires little data preprocessing.

* Neural Networks

Pros: Cons:

1.Fast predict speed. 1.Expensive and time consuming.

2.Can be used with both 2.Depend on a lot on training data.

Regression and classification.

3.Can trained with any number of

input

Implementation

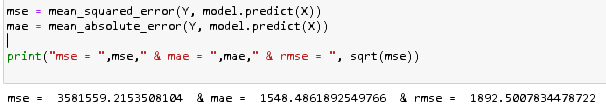
I have used Python as coding environment, .csv files for input and output, I have used pandas for dataframes(better format for models), sklearn module for models and mean squared error and mean absulute error, math module for sqrt. At first i have preprocces my data by using arrays, for loop to iterate over rown in input.csv file(data part of excel file provided to us), in these loop i added first 100 row to train array, remaining 20 row to test array. After that, i have deleted the index numbers in both arrays ( 0th index was their index number) and i deleted 6th index in test array which was “”. After that i changed them into dataframes which is better version of arrays to use in machine learning models.

Results

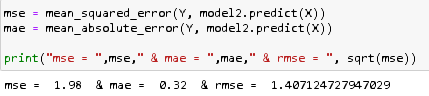
I have used mean square error, mean absolute error , and square root of mean squared error to check my models Which is basically sumtracting prediction value from label for that row and taking square of this and adding it.And then dividing it to number of examples. We dont have data for test output but we have for training data so i have used training data for this. Here are the outputs.

But these are training results I dont know to test results and test results are more important than this. This only gives an idea. The lesser the mse and mae the better the model is this is the generic rule.

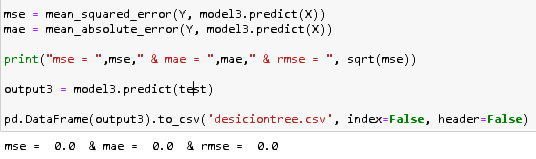
Linear Regression



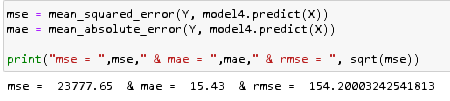
Logistic Regression



Desicion Trees



Neural Networks



The reason why Desicion Trees got 0 is because of how trees work. They make a desicion and and move from that side. So the values are inserted inside this is why it always finds himself in the tree. Other than that it seems like Logistic Regression got the best out of these 4 models because it has lowest mse and mae values compared to other models.

Conclusion

In this homework we expected to estimate last 20 row of given data using first 100 row. To do this we have preprocessed the data, implemented the machine learning models to predict values, different models gave different results. To select the best one I have used performance metrics such as mse and mae.But we dont know the actual results of last 20 row so i have used them on test input. These are training results I dont know to test results and test results are more important than this. This only gives an idea. The lesser the mse and mae the better the model is this is the generic rule.After that I also printed every prediction to different .csv files to check which result is best for unseen cases(last 20 row). From this project, i have learned that some models are better than others in this work but each has its own cons and pros. It should be selected carefuly because each model gives unique results.