**SUPPORTING DOCUMENT**

**DATASET**

As my dataset was old, I couldn’t find any paper related to it. So, I selected a paper that has a different dataset and tried working on similar methodologies.

**INTERMEDIATE RESULTS INCLUDING ANY NEGATIVE RESULTS**

**Exploratory Data Analysis**

* We noticed in the dataset that the target column was not binary, hence we applied label encoder and converted Ps and Ns into 1s and 0s respectively.
* Standardization was done in MATLAB manually to convert the data as the original dataset had a year, month, day, and 11 meteorological stations(knots). So, each of them had different measurement units. Hence, we did standardization. (1 Knot = 0.514444 m/s)
* Box plot was built using all the features.

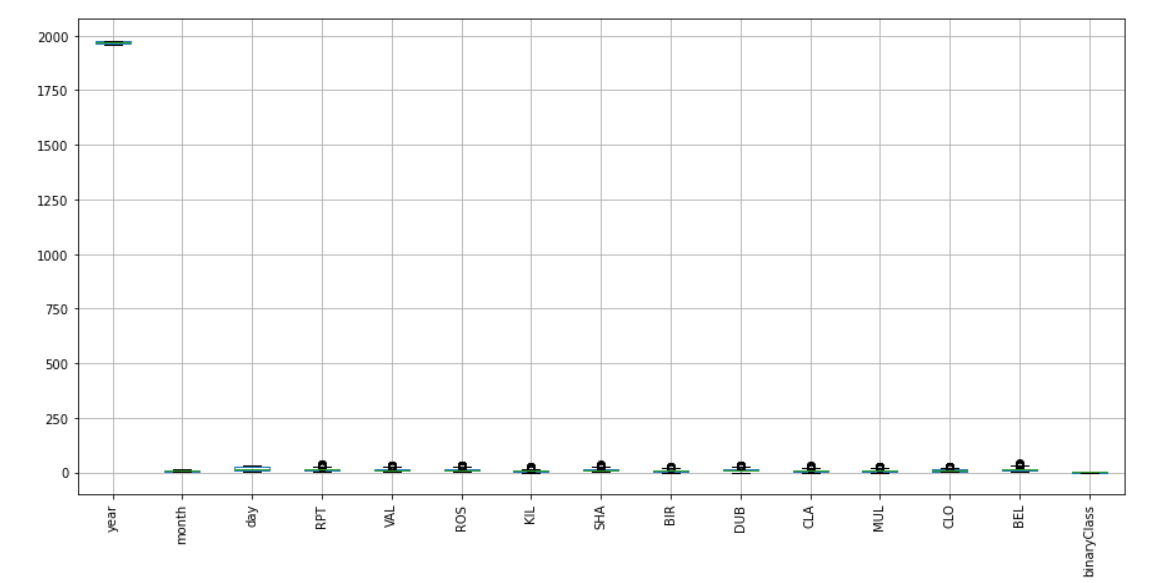


Figure 1

* As we can see above the box plot that all the features do not have major outliers. So, our data is quite good.

**Models**

* We have tried making two models, one with hyperparameters included and the other without.
* For both the models, we used the command optimisehyperparameter, and store the value in a different variable. So, every time you run the model, the value generated is stored in the variable and is used by the model to get better accuracy than the baseline model.
* This was also done to decrease the depth of the tree as we can see in figure 2.

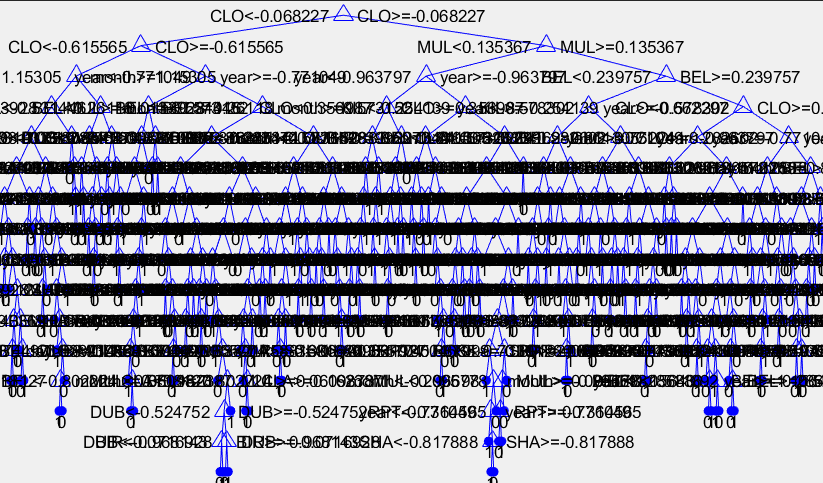


Figure 2

* Figure 3 shows the testing confusion matrix for both the model after the hyperparameters were added.

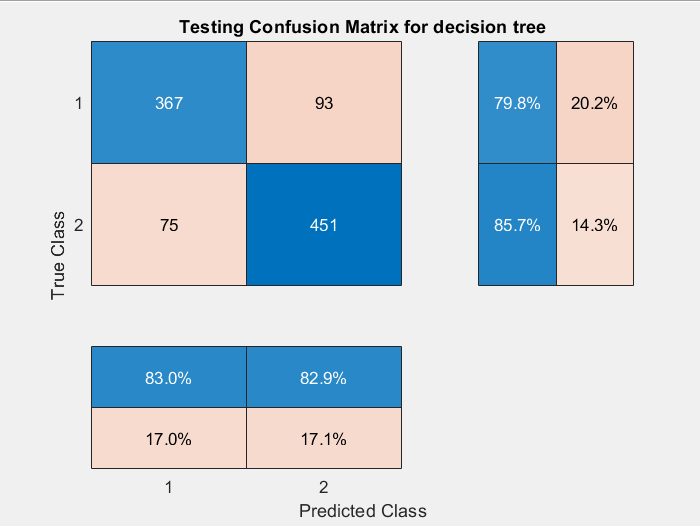
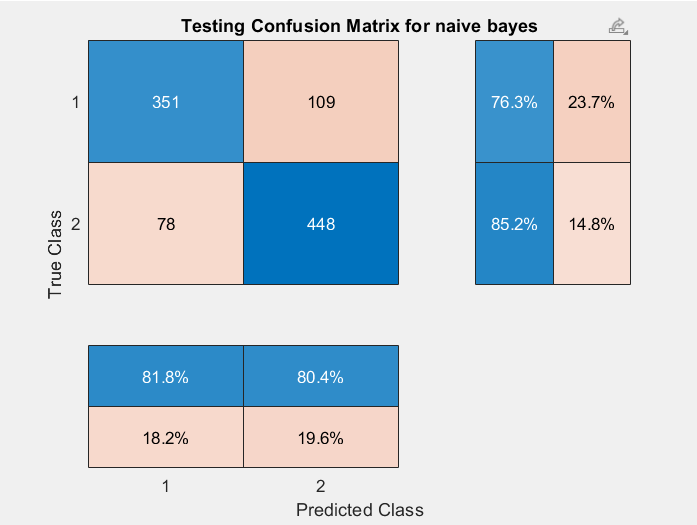
** **

Figure 3

* Even after adding few hyperparameters for the naïve bayes model, we couldn’t increase the accuracy of it. This usually measn that optimiser has found local minimum for the loss.

**Implementation details include a brief description of main implementation choices**

* For both the models, we are keeping 15% of the data unseen, so that we can see what’s the accuracy of it when the data has competed against the training.
* Data did give good accuracy, even before any hyperparameter was optimized or added.
* We have added hyperparameters because we need to accuracy to increase but it is seen that, in the DT classifier, accuracy increases roughly around 3% but there is no significant change in the NB classifier even after optimizing the hyperparameters.
* We calculated the test error of both the models which is the error to show how well the model will work on future data. This has been calculated for both with and without hyperparameters. Visual presentation is given in figure 4.

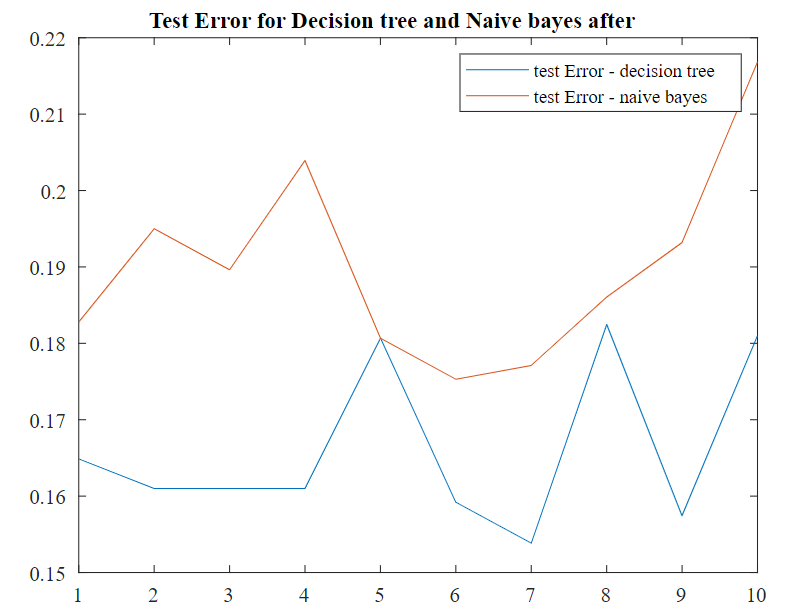
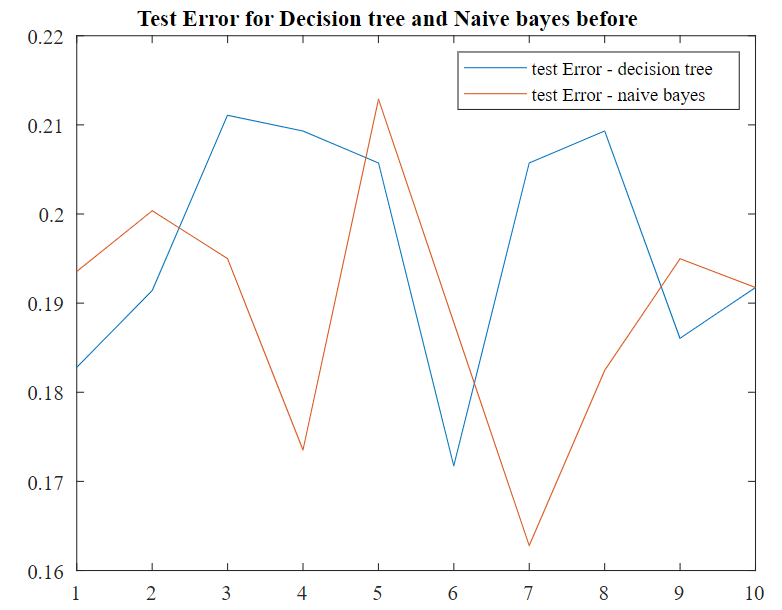
****

Figure 4

* Furthermore, I have not added +1 for naïve bayes as my model does not need it but for other models, it may be used to avoid posterior probability turning into 0.

**References:**

* Few references are added here as I couldn’t fit it in poster.

Few websites have been used for reference such as analytics vidhya, geeksforgeeks etc.

* Mathworks descriptions are used to explain what have we done in the code.

https://www.mathworks.co.uk/

**GLOSSARY**

|  |  |  |
| --- | --- | --- |
| **SL.NO** | **TERM** | **DEFINITION** |
| 1. | Model | State of the machine learning that allows the transformation from input data to calculated output. |
| 2. | Decision Tree (DT) | Data is split depending upon the parameters. |
| 3. | Naïve Bayes (NB) | Used for solvinmg classifiaction problem and is based on bayes theorm. |
| 4. | Training | The action of selecting the ideal parameters for the model to be trained on. |
| 5. | Testing | The action of evaluating the model’s performance |
| 6. | Validation | The action of using the subset of the data which is distinct from training and used to adjust the model’s parameter. |
| 7. | **Optimized Hyperparameters** | Lists the values of the optimized hyperparameters |
| 8. | cvpartition | Sets the partition for cross validation. |
| 9. | holdout | Holds the data specified and uses the rest for the model. |
| 10. | K-fold | It is number of folds for cross-validation |
| 11. | Precision | The number of correct positive results divided by postive results predicted by the classifier. |
| 12. | Recall | The number of correct positive ones divided by all samples |
| 13. | F1 score | It is used to measure a test’s accuracy |
| 14. | ROC curve | It’s a graph showing the perfomance of the model at all thresholds. |
| 15. | AUC (Area under curve) | It is equal to the proability that the classifier will rank a randomly chosen positive value higher than a randomly chosen negative value. |
| 16. | Confusion Matrix | Gives matrix as output and complete perfromance of the model. |