ITNPBD6 Assignment # 1. Student Number: XXXXXX

**Project Methodology:**

* Import Libraries(pandas,numpy,matplot,sklearn,seaborn) and Import Dataset
* Performed Data Cleaning and checked for any Missing Values
* Plotted Histograms for every Variable before and after cleaning
* Converted time.of.day from strings to a numeric datatype.
* Scaled the features using MinMaxScaler
* Checked Correlation between the Variables
* Splitted the Dataset into 70 / 30.
* Performed Logistic Regression and traced accuracy,precision,recall and f1 score
* Constructed a Heatmap for the Correlation Matrix, displaying results for the model.
* Checked if the model was making predictions right or wrong.
* Trained the Model using “Decision Tree Classifier”, constructed a correlation matrix and a heatmap displaying the results for the model.

**Variables:**

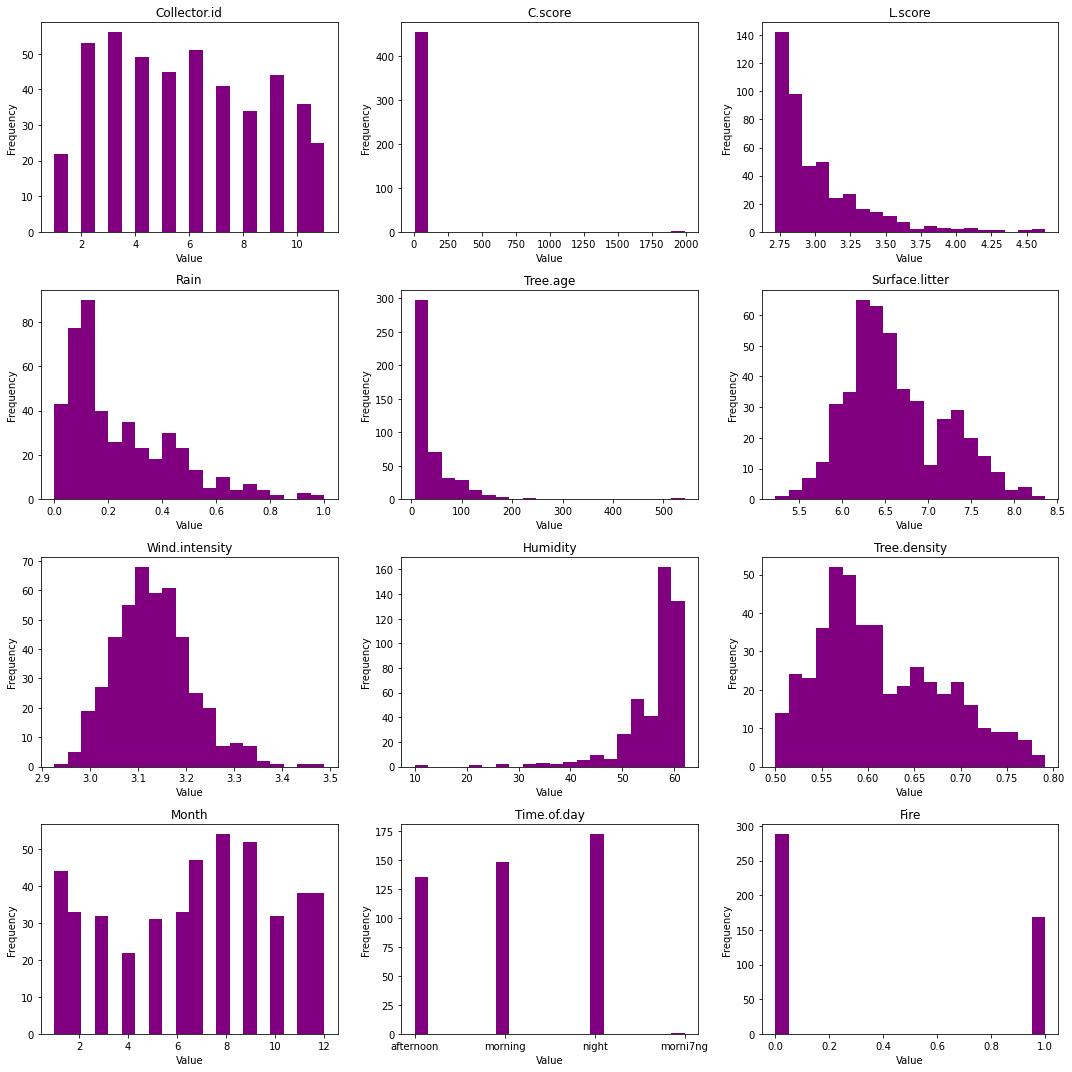
| **Variable** | **Type** |
| --- | --- |
| Variable: c.score | Numerical |
| Variable: I.score | Numerical |
| Variable: tree.age | Numerical |
| Variable: surface.litter | Numerical |
| Variable:wind.intensity | Numerical |
| Variable: tree density | Numerical |
| Variable: month | Categorical |
| Variable: Time of day | Categorical |

**Impact:**

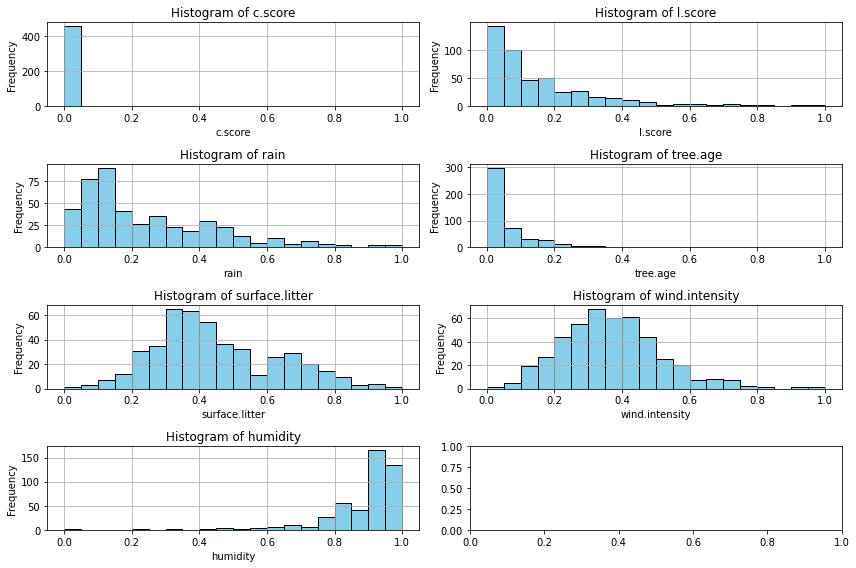
Well upon observing all the variables i had come to a conclusion as removing 3 of the variables from my dataset which include: “collector.id”, “rain”,”humidity”. Collector.id clearly has no impact nor any relation with fire. Humidity and rain had a weak correlation due to which I considered not adding them. All variables were initially Numerical in Nature, apart from “time.of.day” which was in string format and converted to integers ranging from (0 to 2).

Month and time are categorical because they are discrete in nature.

**Before Cleaning:**

****

**After Cleaning:**

****

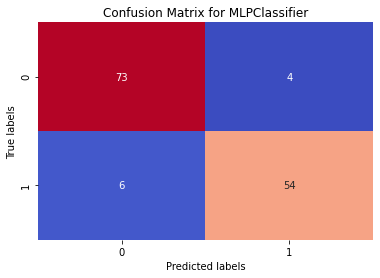
**Model Training and HyperParameters:**

| **Model** | **Hyperparameter** | **Metric** | **Metric Value** | **Confusion Matrix** |
| --- | --- | --- | --- | --- |
| Logistic regression | Solver: lbfgs,C:1 | Accuracy,  Precision | 81.75% 90.69% | [[73 4]  [21 39]] |
| Decision Tree | Max\_depth  Criteria:gini | Accuracy,  Precision | 88.32%,  91.025% | [[73 6]  [ 8 49]] |
| Neural Network(MLP) | Hidden\_layer\_sizes | Accuracy,  Precision | 93.0% ,  93.10% | [[73 4]  [ 6 54]] |

Chose Solver, due to small dataset and binary classification task.

Chosen via hyperparameter tuning based on cross-validation.

**Confusion Matrix for the Final Model-(Neural Network: (MLPClassifier)):**

****

The Model tells:

* There are 73 instances correctly predicted as fire unlikely (True Negatives).
* There are 54 instances correctly predicted as fire likely (True Positives).
* There are 4 instances incorrectly predicted as fire likely when they are fire unlikely (False Positives).
* There are 6 instances incorrectly predicted as fire unlikely when they are fire likely (False Negatives).
* Out of all Confusion Matrices for the Models This Model had the Highest Accuracy and Precision

**References:**

1. **ChatGpt**
2. **Youtube (**[ML Project | Prediction using DecisionTree](https://www.youtube.com/watch?v=l8Jv7D_Rrd4)**)**
3. **Youtube(**[Multilayer Perceptron Algorithm | Theory & Code Explanation | 1510049](https://www.youtube.com/watch?v=C6J1cgW6i9Y)**)**
4. **Doc** [**https://scikit-learn.org/stable/modules/neural\_networks\_supervised.html**](https://scikit-learn.org/stable/modules/neural_networks_supervised.html)**)**
5. **Hyperparameter Tuning (https://levelup.gitconnected.com/a-comprehensive-analysis-of-hyperparameter-optimization-in-logistic-regression-models-521564c1bfc0)**

**Comment:**

ChatGPT was used to:

* Clarify concepts and instructions related to data preprocessing, model building, and evaluation techniques.
* Answer questions and provide insights on various aspects of the assignment, such as feature selection, model evaluation, and visualization techniques.
* Validate approaches and suggest improvements to the methodologies proposed in the assignment.