1. ITNPBD6 Assignment Student Number 3338986

2. **Project Methodology**

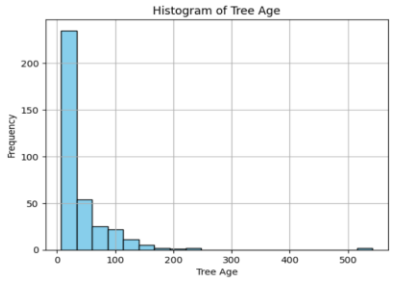
1. Reading the forest fire csv file
2. Separating the file into train and test by proportion of 80 and 20
3. Finding the null values and dropping those rows in train data
4. Updating the outliers with mean values in train data
5. Finding distinct values in categorical columns and updating the errors in train data
6. Separating the data into input and target columns on train ,test
7. Calculating the correlation between input columns and target , then removing the columns having near zero correlation(collecter_id,tree_age)
8. Performing normalisation and one hot encoding for both categorical column as they are nominal data.
9. Oversampling the train data as there is imbalance in proportion of group of target variable
10. Initialising the logistic regression, Decision tree and Neural networks model.
11. Finding the best parameters for each models by grid search with relevant parameter range using train data and doing cross validation
12. Using the updated models comparing the recall score for each model and choosing the model which have highest recall.
13. Model evaluation for the test data with the chosen model and Displaying the confusion matrix for the final model.

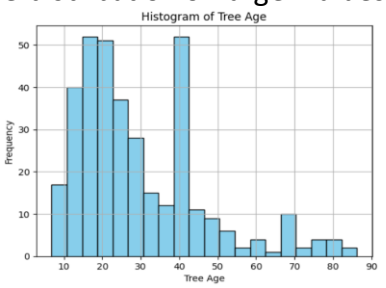
3.**Variables**

Variable	Type	Treatment
c.score	Numeric (continuous)	It affects the model Highly
l.score	Numeric(continuous)	It affects the model moderately
rain	Numeric(continuous)	It affects the model Highly
tree.age	Numeric(continuous)	Age of tree can increase susceptibility of fire
surface.litter	Numeric(continuous)	Increase in litter will increase the chance of fire
wind.intensity	Numeric(continuous)	High wind speed will increase the spread of fire
humidity	Numeric(continuous)	Increase in humdity will increase the chance of fire
tree.density	Numeric(continuous)	It affects the model Highly
month	Categorical(nominal)	One-Hot Encoding
time.of.day	Categorical(nominal)	One-Hot Encoding
Fire	Categorical(ordinal)	Target

4. **Data Preparation**

Two histograms of tree age before and after cleaning. Updating the outliers by mean value of the column decrease the distribution of larger values.





5. Model Training and Hyper Parameters

Model	Hyper Parameters	Validation Metric
Logistic regression	C: 0.1, penalty: l2, solver: lbfgs	85%,Recall
Decision tree	max_depth: 5, max_features: Sqrt, min_samples_leaf: 6	97%,Recall
Neural networks	hidden_layer_sizes=(128, 64), activation='tanh', alpha=0.0001, learning_rate_init=0.01	98%,Recall

1.**Logistic regression:**
Solver: To minimize the loss function
Penalty and C:To regularise function for preventing over fitting
2.**Decision tree**:
Max_depth: Finding the depth of tree to have balance between overfitting and underfitting
Max_features: Choosing the number of relevant features
Min_samples_leaf: The minimum number of leaf nodes to prevent overfitting
3.**Neural Networks**:
Hidden_layer_sizes: To understand the complex patterns in the model
Activation: Mathematical function that's represent nonlinear patterns in the data
Alpha: Control the strength of regularisation
Learning_rate_init: To determine size of weights and influence the convergence

- For each model finding the important parameters required to enhance the predictions performance . The reason for choosing the parameters are given above. Providing the list of values that is permissible within the range of parameters. Providing those values in the dictionary and passing it in gridsearchcv along with train data. This will go through all the combination of parameters and find the values which gives the best score for the model. Then it prints the parameters for each model.

6. Final Model and Results

Actual / Predicted	No fire	Fire
No fire	47	5
Fire	4	34

- In providing train data for each model with best parameters, neural networks provides better recall with percentage 98. However on the test data, Decision tree gives better percentage with 92 where as NN is 89. Possible reason for neural networks to get low recall on test data would be over fitting happened on NN with train data as the sample is very small and oversampling made multiple duplicate records so It captures all variations for train data and cannot perform well with test data.
- The model is chosen based on recall metrics because we need to reduce the case when there is fire and it is not predicted correctly(False Negative) . So we have to choose to maximum metric which is showing True positive(actual fire and predicted fire which is recall.

7. **Reference**.

A global wildfire dataset for the analysis of fire regimes and fire behaviour. https://www.nature.com/articles/s41597-019-0312-2