CSCU9M5 Assignment 1. Student Number 3344671

2. Project Methodology

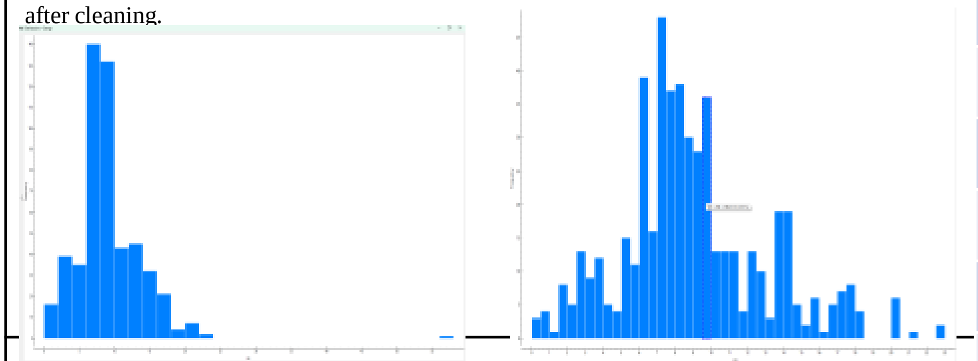
1. Loaded dataset into Orange and identified feature types numerical/categorical and looked for missing, outliers and errors in the data
2. Removed one extreme ISI outlier. Inspected remaining data and confirmed no issues.
3. Features X and Y changed to categorical as Orange assigned numerical by default
4. Split data into 70% training, 15% validation and 15% test data.
5. Trained 6 models, Random Forest, Gradient Boosting, Logistic Regression, KNN, Neural Networks and Decision Trees, on the 70% training data.
6. Tested against the validation data and tuned hyperparameters.
7. Selected Gradient Boosting as it achieved the best validation score for CA and F1. Recorded the validation scores for each model to a table.
8. Retrained the selected model on to the training data.
9. Tested the selected model against the test data and got the confusion matrix
10. Ranked features for their importance and checked for bias using the confusion matrix.

3 Variables/ Data Understanding

Variable	Type	Used	Impact
X	Categorical	yes	Used as categorical to avoid false numerical order which misleads linear models. This feature helps model identify location-based fire
Y	Categorical	yes	Used as categorical to avoid false numerical order which misleads linear models. This feature helps model identify location-based fire.
Month	Categorical	yes	Helps models detect seasonal fire risk. It is categorical to avoid false numerical order
Day	Categorical	No	Day of the week has no physical influence on fire behaviour so removed as it just adds noise without value.
FFMC	Numeric	yes	Strong predictor for fire. Few instances of near outliers but kept as they are in a scientifically possible range
DMC	Numeric	yes	Important predictor for fire. Numeric so linear/tree models can learn trends.
DC	Numeric	yes	Important predictor for fire. Numeric so linear/tree models can learn trends.
ISI	Numeric	yes	Important predictor for fire. One extreme outlier removed for model stability.
TEMP	Numeric	yes	Important predictor for fire. Numeric so linear/tree models can learn trends.
RH	Numeric	yes	Important predictor for fire. Numeric so linear/tree models can learn trends.
Wind	Numeric	yes	Important predictor for fire. Numeric so linear/tree models can learn trends.
Rain	Numeric	yes	509 of 517 entries are zero. May weaken linear models but tree-models handle Zero dominated data well.
Area	Categorical	Target	Defines the classification task. Binary as it is just true or false

4. Data Preparation

In the ISI variable histogram showed there is one extreme outlier far outside the normal range. Removing it helps tree-based models which are sensitive to outliers. The outlier was trimmed using the select rows widget in Orange. Histograms below show the ISI distribution before and after cleaning.



5. Model Training and Hyper Parameters

This table below is showing the different models and hyper parameters that were trained, along with the correct metric for each.

Model	Hyper Parameters	Validation Metric
Random Forest	150 trees, attributes at each split 5, Do not split subsets smaller than 5	CA: 0.909, F1: 0.909
Gradient Boosting	150 trees, 0.1 learning rate, depth limit 3, subsampling 1.00, Do not split subsets smaller than 2	CA: 0.922, F1: 0.922
Tree	Do not split subsets smaller than 5, Maximum tree depth 10, Stop when class majority reaches 95%, Min leaves 2	CA: 0.896, F1: 0.896
Neural Networks	Hidden layers: 5, 10. Activation: ReLU Solver Adam, Regularisation: 0.01 Iterations: 200	CA: 0.727, F1: 0.719
Logistic Regression	Regularisation: L2 (Ridge)strength C = 10	CA: 0.740, F1: 0.738
KNN	5 neighbours, metric: Euclidean, Weight: Uniform	CA: 0.831, F1: 0.832

As the data was slightly imbalanced and had more False values, F1 was used with the CA when choosing the best hyperparameter settings. For each model, I experimented with each settings and the setting with the best validation for both F1 and CA scores were chosen. When tuning hyperparameters tree based models performed best with 150 trees which makes sense for data about 500 rows as this number is not too overfitting or underfitting. Linear models performed best with the least changes with the standard settings. Gradient Boosting hyperparameters that were chosen were 150 trees, 3 for depth limit and a learning rate of 0.1 to avoid overfitting.

6. Final Model and Results

Gradient Boosting achieved the highest validation score for both F1 score and Classification Accuracy although Random Forest performed better with training data it achieved lower validation accuracy, pointing out that Random Forest is overfitting. For this reason, Gradient Boosting was chosen as the final model. Below is the confusion metric of the Gradient Boosting model.

Actual / Predicted	False	True
False	29	2
True	5	41

The confusion matrix tells us that the model correctly predicts most cases. We have 29 true negatives and 41 true positives which are correct predictions. We also have 7 errors for this model 2 are false positives and 5 false negatives. This means the model predicts that 5 areas which burned more than 4% did not burn which can be cause a safety risk. However, the model performs well with more correct prediction than incorrect predictions on unseen test data.

7. Insight about data or models gained

I learn that tree-based models like Random Forest , Decision Tree and Gradient Boosting outperformed linear models which shows that relationship between features is nonlinear. With ranking importance of features, it is shown that DC, Month, DMC, Y, and Temp are the five most important features. This means drought conditions, seasonal patterns and location are important indicators of fire. The model seems unbiased between the classes as the confusion matrix showed it had five false negatives and two positive negatives which does not suggest that this model favours one more than the other. It is important to note that in this case false negatives, 5, represent areas which burned more than 4% which are incorrectly predicted. This could mean areas with higher fire risk identified safe.

Brownlee, J. (2020) ‘How to choose a feature selection method for machine learning’, *MachineLearningMastery.com*, 20 August. Available at: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/ (Accessed: 28 November 2025).

Chen, F. et al. (2014) ‘The impact of precipitation regimes on forest fires in Yunnan Province, Southwest China’, *ScientificWorldJournal*, 2014, Article ID 326782. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC4163318/ (Accessed: 28 November 2025).