**Swinburne University Of Technology**  
**INF30036**

**Business Analytics and Artificial Intelligence**

**Project Title:**

***Predictive Modelling of Forest Fires***

**Submitted By:**  
**Bhuvan Virmani – 103847174**

**Shenal Fernando – 104198462**

**Le (John) Doan – 102855707**

**Malith Nissanka- 103844858**

Contents

[Contents 1](#_Toc740479310)

[Business Objectives 1](#_Toc1148170506)

[Objective Statement 1](#_Toc378358121)

[Data Exploration 1](#_Toc1102138113)

[Initial Data Overview 1](#_Toc681469579)

[Data Dictionary 1](#_Toc543048353)

[Exploratory Data Analysis (EDA) 1](#_Toc386852753)

[Data Preparation 1](#_Toc1533607989)

[Data Sampling 1](#_Toc2051161707)

[Building the Model 1](#_Toc750401595)

[Models Considered: 1](#_Toc1885014620)

[Model Evaluation 1](#_Toc1448423309)

[Model Evaluation Metrics 1](#_Toc1804561912)

[Model Comparison Table 1](#_Toc964356345)

[Final Model Selection and Justification 1](#_Toc1797839561)

[Conclusion and Recommendations 1](#_Toc1745745325)

[Summary of Findings 1](#_Toc1733930066)

[Model Limitations 1](#_Toc254418309)

[Recommendations 1](#_Toc1388313481)

[Next Steps 1](#_Toc1443139469)

[Appendices 1](#_Toc2036256016)

# Business Objectives

## Objective Statement

The main goal of this project is to create a model that can predict the extent of forest fires by utilizing weather data and fire danger indicators. Forest fires present significant dangers to environmental sustainability, human life, and economic stability. The aim is to help forest management agencies and environmental authorities predict fire damage more accurately, leading to better resource allocation, preparedness, and response strategies. The model will enhance fire prevention methods and reduce risks in at-risk areas by forecasting the extent of fire damage using weather factors like temperature, wind speed, humidity, and rainfall.

This predictive ability will also help enhance fire response efficiency, decrease the environmental impact of forest fires, and enhance safety measures for impacted communities. The main goals of this model are to impact key performance indicators (KPIs) such as reducing response time, optimizing allocation of firefighting resources, and minimizing the area affected by fires. Additionally, using data-driven methods, forest management teams can improve decision-making, leading to decreased financial and environmental repercussions (Chas-Amil et al., 2013).

This model could be incorporated into early warning systems, providing a proactive tool for reducing risk. Insurance companies can use it practically for risk assessment, while governments can utilize it to reduce the socio-economic and environmental impacts of forest fires.

Data Exploration

Initial Data Overview  
The dataset used for analysis contains information about forest fires in the northeast region of Portugal. The key details are as follows:

* **Number of records:** The dataset contains 517 records.
* **Number of variables:** There are 13 variables (columns) in the dataset.

Exploratory Data Analysis (EDA)

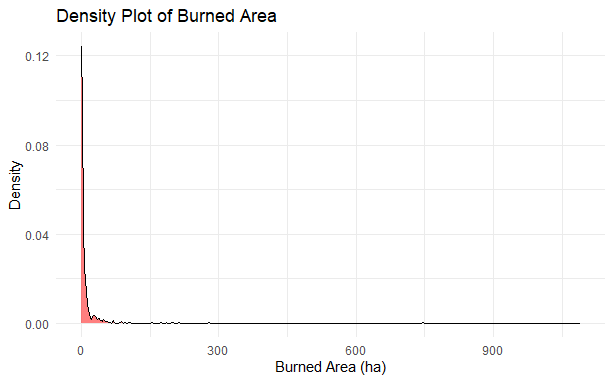
**1. Target Variable Exploration**

Analysing the raw and transformed distributions helps one grasp the distribution and features of the target variable, Burned Area. Understanding this will enable one to choose suitable transformations to prepare the data for modelling.

**1.1 Histogram and Density Plots of Burned Area (Raw)**

The raw value histogram and density displayed to determine the distribution of the burned area and to look for any skewness or outliers.

A graph with a blue line

Description automatically generated

The histogram of the burned area (in hectares) has a distinct right skew, with most observations clustered at 0 and very few values dispersed over the upper range. This indicates that the dataset is dominated by tiny fires, with just a few instances involving significant burnt areas. Most of the burned areas are less than 200 hectares; as the burned area rises, the abrupt drop-off indicates the unusual occurrence of substantial fires.

**Statistical Interpretation:**

* Mean: 12.89 ha (hectares)
* Median: 0.54 ha
* Skewness: 12.73 (extremely right-skewed)
* Kurtosis: 190.09 (significantly peaked distribution)

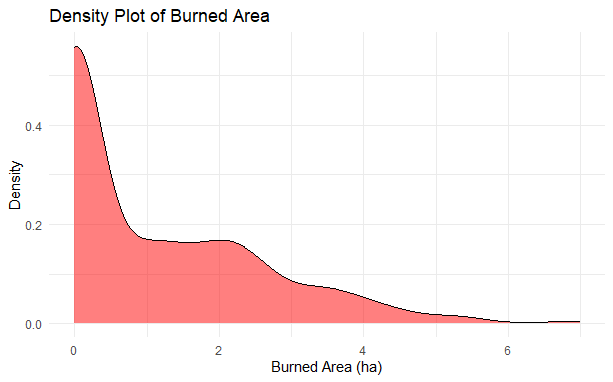
The burnt area is significantly skewed, with several flames burning zero or tiny land. A few vast fires significantly impact the mean value, as seen by the distribution’s tail. This significant skewness implies that prediction models might benefit from putting the burnt area into a more normal distribution. The median (0.54 hectares) is much lower than the mean, showing a significant effect from outliers.

**1.2 Histogram and Density Plots of Log-Transformed Burned Area**

In its raw form, the burned area distribution is severely skewed, which might negatively impact model performance. To remedy this, we logarithmically alter the burned area variable. By lowering skewness, the log transformation improves the data's modelling suitability.

A graph of a distribution of burned area

Description automatically generated



In contrast to the raw burned area, the log-transformed histogram displays a more dispersed and normalised distribution. The change lessens the effect of extreme outliers, such as big fires, but leaves the peak at the left side, representing minor burned areas (around 0), visible. Consequently, the distribution becomes more balanced and more straightforward to predict.

**Statistical Interpretation:**

* Mean: 1.11 (log-transformed units)
* Median: 0.43 (log-transformed units)
* Skewness: 1.21 (moderately skewed)
* Kurtosis: 0.93 (approaching normal distribution)

The log-transformed burned area lowered the skewness from 12.73 to 1.21, making the distribution more controllable and closer to the standard shape. This adjustment greatly enhances the dataset's applicability for predictive modelling since most machine learning algorithms usually need to distribute target variables to make correct predictions.

Despite the transition, a few significant fire episodes (more than 50 hectares) remain, which will be critical in determining the accuracy of the model in forecasting severe fire events.

**2. Exploratory Predictors**

This part emphasises the distribution and significance of essential predictors, mostly numerical variables, for the forecasted burned area. Exploring the distribution of these variables is crucial for finding patterns, outliers, skewness, and possible adjustments to raise model performance.

**2.1 Histogram and Density plots of X (Spatial Coordinate)**

On a horizontal axis, the X coordinate shows the geographical location of the flames. This numerical variable is categorical and has the potential to provide light on the geographical patterns of fire incidents.

A graph of a graph

Description automatically generatedA graph of a function

Description automatically generated

The data don’t show much horizontal spatial bias since the histogram demonstrates a pretty equal distribution of fires across various X coordinates.

**Statistical Interpretation**

* Mean: 4.67
* Skewness: 0.03 (nearly normal)
* Kurtosis: -1.19 (flat distribution

The X-coordinate variable may not be as predictive by itself, given the roughly uniform distribution. It may be used with other geographical factors, such as Y. Because the distribution is approximately normal, no transformation is required to identify possible fire-prone locations.

**2.2 Histogram and Density plots of Y (Spatial Coordinate)**

The Y coordinate indicates the geographical position of flames on a vertical axis. It complements the X coordinate in terms of analysing geographical trends.

A graph of a graph

Description automatically generatedA graph of a function

Description automatically generated

**Statistical Interpretation**

* Mean: 4.39
* Skewness: 0.41 (mild right skew)
* Kurtosis: 1.38 (slightly peaked)

Unlike the X coordinate, the Y coordinate has a slight skew, indicating that particular vertical portions of the study area may be more susceptible to flames. Although no modification is required, integrating this variable with X in a heatmap might assist in revealing spatial patterns.

**2.3 Histogram and Density plots of Fine Fuel Moisture Code (FFMC)**

The FFMC measures the moisture content of refined fuels, such as leaves and twigs. Higher numbers indicate dry fuels, which raise fire danger, whereas lower values suggest wetter conditions.

A graph with a blue rectangular bar

Description automatically generatedA graph with lines and numbers

Description automatically generated

With most data gathered between 85 and 100, the plot has a high right skew that suggests dry conditions for most fire incidents.

**Statistical Interpretation:**

* Mean: 90.64
* Skewness: -6.51 (extremely left-skewed)
* Kurtosis: 65.60 (significantly peaked)

The significant kurtosis and skewness of FFMC expose that most fire-prone scenarios arise in relatively dry environments. Although the significant skewness of this variable makes predicting fire initiation difficult, it is nonetheless important and calls for further research.

**2.4 Histogram and Density plots of Duff Moisture Code (DMC)**

The DMC indicates fire hazards and gauges the moisture content of deeper organic layers. Higher readings indicate drier weather, which raises the possibility of flames that burn deeper.

A graph of blue bars

Description automatically generatedA graph of a chart

Description automatically generated

The DMC distribution reveals a slight right skew, meaning most fire occurrences happen under dry organic layer circumstances.

**Statistical Interpretation:**

* Mean: 110.89
* Skewness: 0.54 (moderate right skew)
* Kurtosis: 0.16 (normal distribution)

DMC’s modest skewness points to dry circumstances ruling the dataset. The low kurtosis, therefore, suggests that extreme values are rare. No transformation is immediately required.

**2.5 Histogram and Density plots of Drought Code (DC)**

To evaluate the fire danger over time, it is essential to quantify the impact of long-term drying conditions using the Drought Code.

A graph of a number of blue bars

Description automatically generatedA graph with a line

Description automatically generated

DC values vary from moderate to high, with concentrations of 500-700, suggesting long-term dry conditions.

**Statistical Interpretation:**

* Mean: 546.89
* Skewness: -1.10 (moderately left-skewed)
* Kurtosis: -0.23 (flat distribution)

DC is significantly left-skewed, indicating that most fires occur during protracted dry conditions. Given the nature of the drought code, no change is required. It is likely to have an essential role in forecasting more extensive and protracted flames, particularly during prolonged dry spells.

**2.6 Histogram and Density plot of Initial Spread Index (ISI)**

The Initial Spread Index is a composite metric used to predict the pace of fire spread, taking into account both wind speed and the moisture content of particulate fuel.

A graph of a number of blue bars

Description automatically generatedA graph of a graph

Description automatically generated

According to the ISI histogram, which displays a distribution bent towards lower values, most fires happen under modest initial spread circumstances.

**Statistical Interpretation:**

* Mean: 9.0
* Skewness: 2.51 (highly right-skewed)
* Kurtosis: 20.98 (highly peaked)

ISI is highly skewed to the right. Hence, most flames start out at a slow pace and only occasionally show a fast spread of fire. This skewness implies that before adding information to the prediction model, transformation (logarithmic) is probably needed. The significantly peaked kurtosis indicates a concentration of values around the mean, meaning that very few occurrences had extreme beginning spread conditions.

**2.7 Histogram and Density Plots of Temperature**

Temperature is a very important indicator of fire strength because higher temperatures make fires more likely to start and spread. Analysing temperature distributions helps us understand their function in fire dynamics.

A graph with blue bars

Description automatically generatedA graph showing a normal plot of temperature

Description automatically generated with medium confidence

With a few exceptions, the majority of temperatures fall between 15°C and 25°C. This very normal distribution indicates that temperature does not need any adjustment for modelling purposes.

**Statistical Interpretation:**

* Mean: 18.89°C
* Skewness: -0.33 (slightly left-skewed)
* Kurtosis: 0.09 (normal distribution)

The temperature displays an almost symmetric distribution with no extreme tails, with the skewness at -0.33 and kurtosis at 0.09. No transformation is required.

**2.8 Histogram and Density Plots of Wind Speed**

Wind speed greatly influences the rate and direction at which a fire spreads, and strong winds accelerate this process.

A graph of a wind speed

Description automatically generatedA graph of a wind speed

Description automatically generated

The majority of wind speeds are grouped between 2 and 6 km/h, which significantly skewed the distribution. There aren’t many high wind observations, which might have a big impact on the spread of the fire.

**Statistical Interpretation:**

* Mean: 4.01 km/h
* Skewness: 0.58 (moderate right skew)
* Kurtosis: 0.03 (normal distribution)

The distribution exhibits few outliers (high wind speeds) with a modest skewness of 0.58 and an almost normal kurtosis of 0.03. These anomalies could affect the propagation of fire. Although handling outliers might help, based on existing data, no change is required.

**2.9 Histogram and Density plots of Relative Humidity (RH)**

The moisture content of a plant is influenced by relative humidity, which in turn impacts how quickly it burns.

A graph of a number of blue bars

Description automatically generated with medium confidenceA graph with a line

Description automatically generated

With a few exceptions for low relative humidity values, the plot shows most relative humidity values falling between 40 and 70%.

**Statistical Interpretation:**

* Mean: 44%
* Skewness: 0.85 (moderate right skew)
* Kurtosis: 0.38 (normal distribution)

Though severe low-humidity circumstances are unusual, relative humidity is skewed to the right, suggesting that fires often occur in dry settings. Given that increased humidity lowers fire probability, RH is predicted to have a modest inverse connection with the burnt area. The mild skewness is fine; transformation may not be required.

**2.10 Histogram and Density Plots of Rain**

Rainfall lessens the chance that a fire will start and spread. This variable aids in determining how wet circumstances affect fire behaviour.

A graph with a graph of rain

Description automatically generatedA graph of rain

Description automatically generated

The right-skewed distribution suggests that most fire occurrences happen with little to no rain.

**Statistical Interpretation:**

* Mean: 0.02 mm
* Skewness: 19.62 (extremely right-skewed)
* Kurtosis: 412.34 (significantly peaked)

The severe skewness and kurtosis imply that most fire incidents occur under dry circumstances, with rare cases of rainfall during fire seasons. Unless combined with other weather-related variables, the dataset’s low rainfall could make this variable less predictive. A logarithmic transformation might reduce the excessive skewness and enhance model performance.

**2.11 Summary Table**

|  |  |  |
| --- | --- | --- |
| **Predictors** | **Skewness** | **Kurtosis** |
| X | 0.03 | -1.19 |
| Y | 0.41 | 1.38 |
| FFMC | -6.51 | 65.60 |
| DMC | 0.54 | 0.16 |
| DC | -1.10 | -0.23 |
| ISI | 2.51 | 20.98 |
| Temp | -0.33 | 0.09 |
| Wind | 0.58 | 0.03 |
| RH | 0.85 | 0.38 |
| Rain | 19.62 | 412.34 |

**FFMC**, **ISI**,and **Rain** may benefit from log transformations because of severe skewness and kurtosis. This will assist to mitigate the effects of outliers and stabilise model performance.

**3. Categorical Variables Exploration**

Critical new perspectives on the temporal distribution of forest fires come from categorical factors like month and day. Through the analysis of these factors, we may identify trends or patterns linked to seasonality, which, therefore, affect the frequency and severity of fires. Understanding possible fire danger periods made possible by this study helps forest management organisations allocate resources and prepare appropriately.

**3.1 Bar and Box Plots for Month**

Seasonal weather fluctuations substantially impact the frequency and intensity of forest fires (Flannigan et al., 2009). Fire occurrences must be studied regularly to identify essential fire seasons and execute effective control methods (Westerling et al., 2006).

A graph of blue bars

Description automatically generatedA graph of green and black squares

Description automatically generated

A bar plot of monthly fire frequencies shows temporal dispersion trends. While winter months show fewer fires due to increased rainfall and humidity, summer months usually see higher temperatures and drier conditions, which allow ignition and spread and hence influence fire activity (Flannigan et al., 2009).

A box plot of the log-transformed burned area by month depicts the yearly fire size and severity fluctuations. Significant increases in burned areas, especially in August and September, point to times when environmental circumstances support the spread of fire (Bradstock, 2010). Extreme fire incidents indicated by outliers could call for further research on contributing causes.

**3.2 Bar and Box Plots for Day**

The number of fires may vary depending on the day of the week, especially if human activity is involved in the ignition events. Examining fire incidents day by day yields useful information for fire control by pointing out patterns in human behaviour.

A graph of blue bars

Description automatically generatedA graph of a chart

Description automatically generated with medium confidence

The bar plot clearly illustrates a surge in fire incidences on weekends (Saturday and Sunday), most likely owing to increased human activity. Fusco et al. (2016) have consistently correlated these activities with unintentional fire starts in arid environments. Given that human presence increases the danger of fire, understanding this pattern might inform targeted fire prevention initiatives and enhanced weekend patrols (Johnston et al., 2020).

The box plot shows that more significant burned areas occur more often on weekends, implying a causal relationship between human activity and more catastrophic fires. Particularly when human ignitions take place, delayed reactions and easily accessible fuel help flames to spread (Keeley & Syphard, 2018).

**4. Correlation Analysis**

A correlation heat map visually shows the correlations among many numerical variables in the dataset. Emphasising the strength and direction of the linear correlations between variables helps one rapidly identify important predictors and multicollinearity.

A screenshot of a grid

Description automatically generated

**Correlation with Log-Transformed Area**

Weak relationships between the log-transformed burned area and most meteorological predictors. With a slight positive correlation (0.06), temperature hardly affects fire size, consistent with actual findings. A weak positive correlation (0.07) for wind speed indicates a limited impact on bigger flames. With a very small association (0.05), the Fine Fuel Moisture Code (FFMC) suggests that drier fine fuels marginally influence fire size. Analogously, the modest positive correlation (0.07) of Duff Moisture Code (DMC) indicates that reduced duff moisture only somewhat increases fire size. Overall, these predictions, taken together, are not very powerful influences.

**Multicollinearity between Predictors**

Several predictor variables have moderate to high correlations, suggesting possible multicollinearity. DC and DMC have a high positive correlation (0.68), indicating that they assess similar moisture conditions. This multicollinearity may influence regression models and must be addressed by regularisation or dimensionality reduction (Dormann et al., 2013). DC and temperature (0.50) and DMC and temperature (0.47) have modest associations, indicating that higher temperatures lead to drier conditions. FFMC and DMC have a lesser connection (0.33), suggesting they measure distinct characteristics of fuel moisture. These connections demonstrate the interdependency of fire-related factors.

**Negative Correlations**

Relative Humidity (RH) shows a small negative association (-0.05) with the log-transformed burnt area, indicating that higher humidity lowers fire size. This is congruent with the results of Jolly et al. (2015), who discovered that increased humidity inhibits fire propagation by retaining moisture in fuels, lowering the risk of ignition and fire intensity. Furthermore, the Initial Spread Index (ISI) has a small negative correlation (-0.01), suggesting that its influence on fire size is modest in this dataset.

**5. Scatter Plots (Lower Triangle)**

Selected critical predictors are temperature (temp), wind speed (wind), Fine Fuel Moisture Code (FFMC), and Duff Moisture Code (DMC). Predictive modelling depends on each of these factors as they significantly influence fire behaviour and the degree of burned areas.

**5.1 Temperature vs. Log-Burned Area**

A graph with blue dots

Description automatically generated

**Positive Relationship:** The red dashed trend line on the scatter plot shows a positive linear trend between temperature and the log-transformed burnt area. This implies that greater burnt areas are often linked to higher temperatures. According to the trend line's positive slope, more giant flames will likely occur as temperatures rise.

**Linear Trend with Confidence:** The confidence interval, highlighted in light pink around the trend line, shows the level of uncertainty in the connection between temperature and the log-burned area. The relatively narrow band supports the trend’s dependability, indicating that the linear model fits the data in this range rather well.

**No Clear Non-Linear Pattern:** There is no evidence of a non-linear relationship between temperature and log-burned area; the linear trend line fits very well. Though further research using non-linear models might confirm this assumption, this implies that linear models may be sufficient in capturing the influence of temperature on fire size.

**5.2 Wind Speed and Log-Burned Area**

A graph of a graph of a wind speed

Description automatically generated with medium confidence

**Positive Linear Relationship:** As shown by the red dotted trend line, there is a positive linear relationship between wind speed and log-transformed burned area. The burning area starts to get bigger as the wind speed goes up. This makes sense because stronger winds tend to feed fires by spreading embers and providing oxygen, which can speed up the spread of the fire and affect more land.

**Moderate Confidence in Prediction:** The trend line’s confidence interval, which is coloured in light pink, is quite small, indicating that the correlation between wind speed and burnt area is generally constant between data points. This suggests that, while there is still considerable fluctuation, particularly at higher wind speeds, wind speed is a good predictor of the amount of burnt regions.

**No Strong Non-linear Trends:** The scatter plot does not clearly show non-linear patterns, and the linear trend line provides a reasonable fit to the data. This implies that a linear model might capture much of the link between wind speed and fire size. Nevertheless, more research using non-linear models could still be helpful, especially at very high wind speeds when the fire's behaviour may vary.

**5.3 FFMC vs. Log-Burned Area**

A graph with red and blue dots

Description automatically generated

**Negative Linear Relationship:** FFMC (Fine Fuel Moisture Code) and the log-transformed burned area exhibit a negative linear connection, as seen by the red dotted trend line. This implies that the log-burned area falls as FFMC rises. Drier fine fuels indicated by a higher FFMC score correspond with bigger flames generally. This negative connection fits the idea that drier fuels produce more burned areas.

**Consistent Predictive Power:** The very small light pink confidence interval around the trend line indicates that the link between FFMC and the log-burned area is constant and dependable across the dataset. The linear trend line matches well, demonstrating that FFMC effectively predicts fire size, particularly when modelling bigger fire occurrences**.**

**Absence of Non-linear Behaviour:** Based on the solid, smooth trend line, the two variables seem to have a linear relationship. The lack of notable deviations or non-linear behaviour suggests that a linear model might represent the connection between FFMC and the log-burned area.

**Variability at Lower FFMC Values:** Although greater burned areas are more often seen at higher FFMC values, the figure also reveals considerable fluctuation at lower FFMC values, where burned areas vary from small to large.

**5.4 DMC vs. Log-Burned Area**

A graph with blue dots and red lines

Description automatically generated

**Positive Linear Relationship:** The dashed red trend line shows a positive linear relationship between DMC and the log-transformed burned area. Thus, the log-burned area usually increases as DMC rises. This implies that bigger flames are associated with drier conditions in the duff layer, that is, with greater DMC.

**Moderate Confidence in Trend:** The light pink confidence interval around the trend line is quite tight over the majority of the DMC value range, demonstrating that the relationship between DMC and the log-burned area is stable and dependable. This implies that DMC is a reliable predictor of burned area size, with considerable fluctuation at higher DMC levels.

**Linear Fit is Appropriate:** There is no clear indication of non-linear behaviour; the smooth trend line matches the data really well. This implies that linear approaches allow one to efficiently describe the link between DMC and log-burned area. Still, further research on non-linear effects at very high DMC values would be advisable, particularly for a more accurate depiction of major flames.

# Data Preparation

**1. Summary Statistics**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Min** | **25%** | **Median** | **Mean** | **75%** | **Max** |
| X | 1 | 3 | 4 | 4.67 | 7 | 9 |
| Y | 2 | 4 | 4 | 4.3 | 5 | 9 |
| FFMC | 18.7 | 90.2 | 91.6 | 90.64 | 92.9 | 96.2 |
| DMC | 1.1 | 68.6 | 108.3 | 110.9 | 142.4 | 291.3 |
| DC | 7.9 | 437.7 | 664.2 | 547.9 | 713.9 | 860.6 |
| ISI | 0 | 6.5 | 8.4 | 9.02 | 10.8 | 56.1 |
| Temp | 2.2 | 15.5 | 19.3 | 18.89 | 22.8 | 33.3 |
| RH | 15 | 33 | 42 | 44.29 | 53 | 100 |
| Wind | 0.4 | 2.7 | 4 | 4.018 | 4.9 | 9.4 |
| Rain | 0 | 0 | 0 | 0.21 | 0 | 6.4 |
| Area | 0 | 0 | 0.52 | 12.85 | 6.67 | 1090.84 |

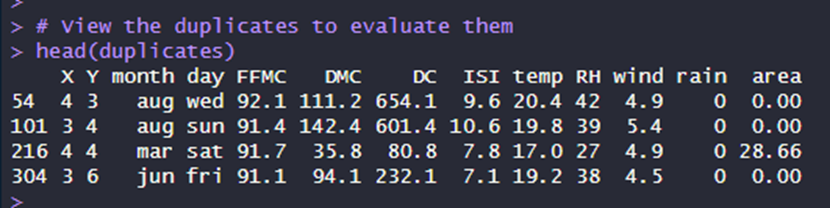
**2. Data Cleaning**

**2.1 Handling Missing Values**

Every dataset column was examined to get the missing values count. Descriptive statistics and R's tools were used in order to precisely find the missing data columns. The below analysis shows that there are no missing values in any of the dataset’s columns.

**2.2 Handling Duplicate Records**

Four duplicate rows were found, and they correspond to the same values for every feature including the burned area.



The above duplicaes must be deleted. Because retaining duplicates may cause model overfit, in which case repeated entries disproportionately affect the model, therefore lowering its generalizability (Han et al., 2011). Eliminating duplicates also guarantees equal weighting of data and increases computing efficiency, hence strengthening the model (Kotsiantis et al., 2006). Furthermore, many duplicates have a burned area of zero, adding no meaningful information.

The total number of rows now has changed to 513 after deleting those duplicates.

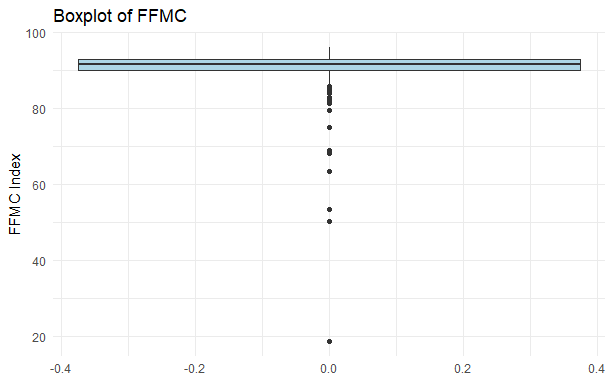
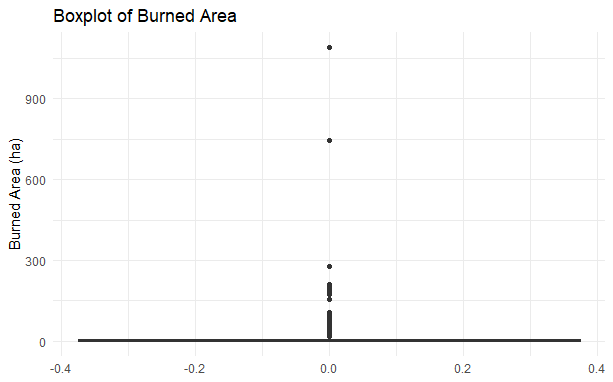


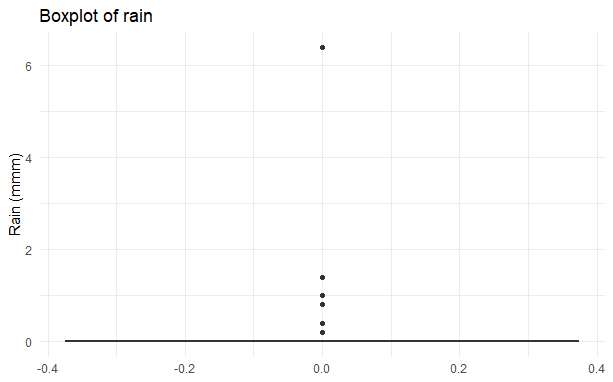
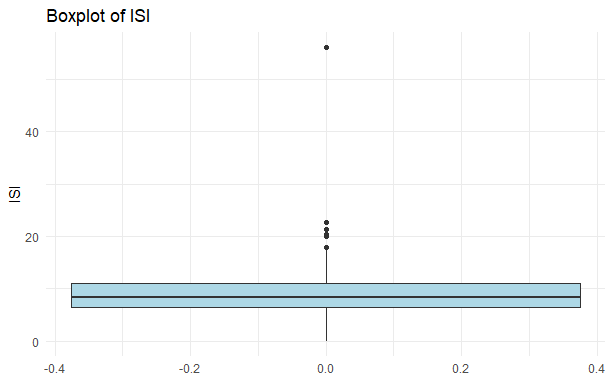
**3. Outlier Detection**

Outliers may considerably impact prediction model accuracy and dependability, mainly when they disproportionately influence statistical parameters like mean and standard deviation. Outlier detection and mitigation are critical for guaranteeing robust model performance since extreme values may distort the outcomes of regression and machine learning algorithms (Aggarwal, 2017). Outliers for crucial predictors such as Burned Area, FFMC, ISI, and Rain were identified and treated methodically in this study.

**3.1 Boxplot Analysis**

Boxplots give a visual depiction of the data distribution that highlights possible outliers. Data points beyond the “whiskers”, which usually indicate 1.5 times the Interquartile Range (IQR) above the third quartile or below the first quartile, are referred to as outliers in a boxplot.





**Burned Area:** The most excellent burned area exceeds 1000 hectares, and the raw burned Area boxplot exposes multiple extreme outliers. While outliers go well beyond the top range, most of the observations fall below 200 hectares.

**FFMC:** The FFMC boxplot also showed numerous lower outliers, most likely resulting from relatively uncommon wet circumstances throughout the fire seasons.

**ISI and Rain:** Both ISI and Rain have significantly right-skewed distributions with multiple high outliers, as verified visually by boxplots.

**3.2 Interquartile Range (IQR) Method**

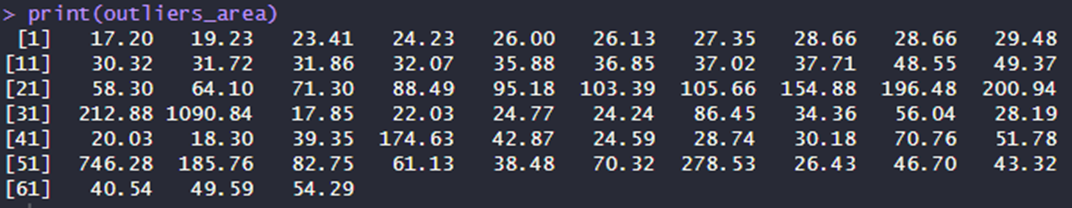
The IQR method was used programmatically to identify outliers for every variable exactly. To determine the IQR for each predictor, the Q1 (25th percentile) and Q3 (75th percentile) were computed:

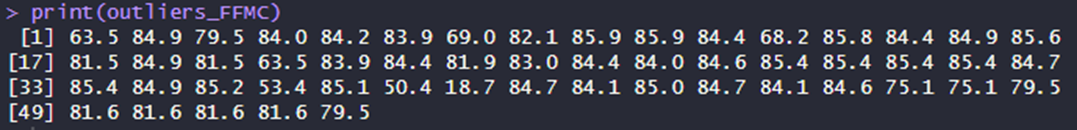
* Lower Bound = Q1 - 1.5 x IQR
* Upper Bound = Q3 + 1.5 x IQR

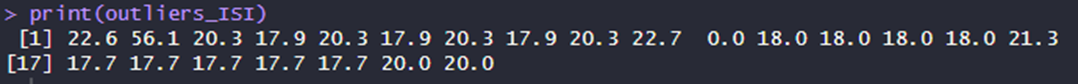
Any value that fell outside of these boundaries was labelled as an outlier.

The following summarises the number of outliers along with a list of those outliers discovered in each crucial variable:

Number of outliers detected in ‘area’: 63

Number of outliers detected in ‘FFMC’: 53

Number of outliers detected in ‘ISI’: 23

Number of outliers detected in ‘rain’: 8



**4. Log Transformation**

To successfully handle outliers and stabilise the data distribution, ‘area’, ‘FFMC’, ‘ISI’ and ‘rain’ were transformed using logarithms. This adjustment compresses the range of high values, minimises skewness, and lessens the influence of severe outliers, making the data more suited for modelling. Log transformation is a commonly established strategy for dealing with skewed data, especially when dealing with variables with substantial magnitude differences (Osborne, 2010). Using this strategy reduces the effect of outliers, resulting in a more normal distribution that enhances the performance of regression models and other statistical analyses.

* **Area, ISI:** The log transformation worked well for area and ISI, significantly decreasing skewness and enhancing data symmetry.
* **FFMC:** A reflection and log transformation were used to address left-skewed. The distribution became right-skewed due to this reflection, making it appropriate for log transformation.

The table below shows the skewness values of each predictor before and after applying log transformation:

|  |  |  |
| --- | --- | --- |
| **Predictors** | **Skewness**   **(before)** | **Skewness**  **(after Log-transformed)** |
| area | 12.73 | 1.21 |
| FFMC | -6.51 | 0.17 |
| ISI | 2.51 | -0.93 |
| rain | 19.62 | 14.07 |

**4.1 Analysis of Log Transformations**

**Area:** Burned Area’s log transformation effectively reduced the excessive skewness and brought the distribution closer to normal.

**FFMC:** Skewness of FFMC was decreased to 0.17 by the transformation, suggesting near-normality, which is advantageous for most models.

**ISI:** The log transformation of ISI effectively decreased skewness from 2.52 to -0.93, considerably improving the distribution’s symmetry.

**Rain:** The Rain variable offered a special difficulty as strong right-skewness (skewness of 14.07 after log transformation) dominated zero values.

**4.2 Handling Rain with Category Binning**

Because the log transformation did not work well for rain, category binning was used instead. The rain variable was divided into three different levels:

* No Rain (0 mm)
* Light Rain (0.1 mm to 5 mm)
* Heavy Rain (> 5 mm)

**4.3 Boxplot Analysis of Transformed Variables**

A graph with a purple line

Description automatically generatedA graph with a purple line

Description automatically generated

A graph with a purple line

Description automatically generated

A graph with a purple bar

Description automatically generated

**Area:** A boxplot shows that the remaining outliers are predicted as the burned variable usually consists of only a few significant fire occurrences that vary significantly from minor ones.

**FFMC:** A boxplot verified the smaller spread; most data were neatly centred, and some predicted outliers were seen. Notwithstanding the outliers, the general distribution became more balanced, improving its predictive modelling applicability.

**ISI:** A boxplot demonstrates that the modified data is centred on a narrower range, which improves predictiveness and reduces the model’s bias. Despite some lingering outliers, the general distribution is now considerably more suited for modelling, especially for models that need a balanced input distribution.

**Rain:** A bar plot suggests that categorical binning is a more effective method for portraying rain in a manner that emphasises significant distinctions while keeping zero values.

**5. Encoding Categorical Variables**

Day and month are nominal categorical variables in the dataset that lack a significant numerical order. Treating them as ordinal (e.g., December as “greater” than January) would be erroneous. To prevent this, each category was transformed into binary columns (such as month\_jan, day\_mon), where the existence of each category is indicated by 0 or 1. This process is known as one-hot encoding.

This strategy is critical for models that use distance calculations, such as Support Vector Machines (SVM), k-Nearest Neighbors (kNN), Artificial Neural Networks (ANN), and tree-based models. One-hot encoding prevents these models from misinterpreting categorical variables as numerical, reducing bias and enhancing model performance (Pedregosa et al., 2011).

**6. Feature Engineering**

**6.1 Temperature and Relative Humidity Interaction**

This depicts the inverse connection in which high temperatures diminish fuel moisture, increasing fire danger, but high humidity lessens it. Division by zero is avoided by adding one to relative humidity. The histogram distribution is right-skewed, suggesting that high-risk circumstances are uncommon yet critical.

A graph of a temperature

Description automatically generated

**6.2 Temperature and Wind Speed Interaction**

Wind accelerates fire propagation by providing oxygen and transporting embers. This interaction phrase describes conditions including high temperatures and strong winds. The histogram’s right-skewed distribution indicates key times of high fire danger.

A graph of a graph

Description automatically generated

**6.3 Composite Fire Risk Index (CFRI)**

Combines standardised indicators (FFMC, DMC, DC, and ISI) to provide a complete fire risk measure. The log-transformed CFRI balances the effect of extreme values, resulting in a more regularly distributed risk representation and improved model prediction.

A graph of a fire risk index

Description automatically generated

**Data Dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** | **Reason for Inclusion** |
| Temp\_RH\_Interaction | Numeric | The product of temperature and relative humidity. | Captures the combined effect of temperature and humidity on fire danger, focusing on dry, hot circumstances. |
| Temp\_Wind\_Interaction | Numeric | Product of temperature and wind speed. | Describes the interplay impact of temperature and wind on fire spread velocity. |
| CFRI | Numeric | The Composite Fire Risk Index combines many fire indexes. | Combines many indicators to deliver a comprehensive fire risk assessment, increasing model predictability. |

# Data Sampling

**1. Split Data into Training and Testing Sets**

**Splitting Methodology**

After cleaning the data, dividing it into training and testing subsets is essential in building a robust machine-learning model. This split enables generalisation, prevents overfitting, and offers a valid evaluation of model performance.

Stratified sampling divided this dataset into 30% testing and 70% training. This technique maintains the target variable **log-transformed burned area (area\_log)** distribution across both subsets. Given the skewed distribution of fire sizes, where tiny fires are more common than bigger ones, stratified selection guarantees that both large and small fire occurrences are covered in both sets, eliminating biases (Kuhn and Johnson, 2013). This method is especially crucial for models trained on unbalanced data.

To enable the model to learn and forecast the whole range of burned areas, stratification on area\_log aids in maintaining a balanced representation of the extremes in the target variable. On the other hand, random sampling runs the danger of underrepresenting outliers, which might result in incorrect estimates for large fires.

**Challenges with Imbalanced Target Variables**

One problem is that the target variable is lopsided, with many cases representing tiny or unburned regions. This means that models may learn disproportionately from the majority class. By guaranteeing that both major and minor fires are included in the training and testing sets, stratified sampling helps to alleviate this and enhances the model’s capacity to learn from and forecast over the whole target distribution (Fernández et al., 2018).

**Potential Data Leakage Concerns**

Data leakage is a typical issue when preparing data for machine learning, particularly when performing transformation such as **scaling**. Early data separation helps prevent biases that can compromise the model assessment (Kaufman et al., 2012). Data leakage often results in unduly optimistic performance indicators, skew model assessment and generalizability.

**2. Normalisation**

Min-Max Normalisation was used to translate the characteristics into a range of [0, 1]. This approach is very successful for models sensitive to feature sizes, such as k-Nearest Neighbours (kNN), Support Vector Machines (SVM), and Artificial Neural Networks (ANN).

Normalisation guarantees that all features are on a similar scale, avoiding characteristics with more comprehensive ranges, such as those of the Drought Code, from controlling the model training process. This stage was vital for models using gradient-based optimisation as it allowed quicker convergence to an optimum solution. Despite the possibility of altering outlier sensitivity, the advantages of Min-Max scaling, such as preventing overfitting and assuring impartial model training, outweighed the risks (Han et al., 2011). To address the skewed nature of variables such as area\_log, the preceding step’s stratified sampling guaranteed that significant and minor fire occurrences were represented fairly.

# Building the Model

In this part, we carefully build several regression models to predict the area\_log (log-transformed burned area) variable. We try out different methods to find the best one. We use Decision Trees, Random Forest, Support Vector Regression (SVR), k-Nearest Neighbors (kNN), and Artificial Neural Networks (ANN) because of the type of data we have and how complex the problem is. The aim is to use various approaches to gather more information and determine which models align most with our expectations for making predictions.

The training dataset (70% of the data) is used to make sure that models learn patterns that make sense without adding any bias. Models may capture associations between predictors and the target variable (area\_log) by training on 70% of the data, with the remaining 30% set aside for testing.

1. **Decision Tree:**

Using the rpart function from the caret package, a decision tree model is created in the code. The decision tree divides the dataset in a recursive manner using input features that best minimize variance when predicting the target variable, which is the burned area of forest fires. The algorithm examines all variables in every node and chooses the one that produces the most similar subgroups in terms of burned area. This procedure goes on until a stopping condition like minimum node size or maximum tree depth is achieved. Decision trees are well suited for this purpose due to their ability to manage intricate, non-linear connections among factors like wind speed, humidity, and temperature, which are crucial for forecasting forest fire behaviour. Moreover, decision trees can be easily understood, helping fire managers grasp the important factors that impact fire propagation. This method aids in creating models that are simple to see and interpret, while delivering accurate forecasts based on past forest fire information.

1. **Random Forest:**

The Random Forest algorithm improves prediction accuracy and decreases overfitting by employing a group of decision trees rather than just one. The Random Forest model in the given code constructs numerous decision trees on various subsets of the dataset by using the train() function with the method parameter set to "rf". In the Random Forest, every tree is trained on a random sample of the training data, and at each node, a random subset of predictors is chosen to determine the optimal split (Parmar et al., 2018).

This method guarantees variation within the trees, avoiding them from overfitting to the noise in the training data, a typical issue in individual decision trees. The Random Forest model offers a more consistent and reliable prediction of forest fire burned area by combining the forecasts of various trees, capturing intricate relationships among temperature, wind, and humidity variables. Furthermore, Random Forests are adept at managing missing data and outliers, which makes them valuable for datasets such as the forest fire dataset that may contain these problems (Parmar et al., 2018). Therefore, utilizing the ensemble method greatly enhances generalization in comparison to single decision trees, leading to reduced mean squared error (MSE) when predicting forest fire damage.

1. **Support Vector Machine (SVM)**

The Support Vector Machine (SVM) algorithm is an effective tool for predicting burned areas in forest fire datasets due to its ability to capture complex, non-linear relationships among features like temperature, wind speed, and humidity. SVM identifies the optimal hyperplane that separates data points in high-dimensional space, allowing it to model intricate patterns relevant to fire occurrence.

One of the key advantages of SVM is its robustness to overfitting, thanks to the regularization parameter that balances training error and model generalization. This is crucial for the forest fire dataset, where noise can significantly impact predictions. Additionally, SVM focuses on support vectors—data points most informative for defining the decision boundary—making it computationally efficient, especially with smaller datasets.

SVM can effectively handle varying data distributions and is robust to feature scaling, which is crucial given the diverse factors influencing forest fires. Additionally, with linear kernels, SVM offers interpretability by highlighting the key variables contributing to fire sizes. This makes SVM a strong choice for the forest fire dataset, as it balances complexity and overfitting while delivering more accurate and reliable predictions of burned areas.

1. **Artificial Neural Networks (ANN):** What architecture was used (layers, activation functions)?
2. **K-Nearest Neighbors (KNN):** Why this model, and how was the nearest neighbors' choice made?
3. **Clustering:**   
   This K-Means clustering analysis aimed to group the data into meaningful clusters and predict the 'area' variable using the centroids of each cluster. The 'area' column was initially excluded from clustering to avoid bias, and only numeric features were retained for the algorithm. To determine the optimal number of clusters, the Elbow Method was employed by plotting the Within-Cluster Sum of Squares (WCSS) for different values of K (1 to 10). Based on the elbow point in the plot, 3 clusters were selected as the optimal solution, ensuring well-separated clusters.

After applying the K-Means algorithm with K=3K=3K=3, the cluster labels were added to the original dataset. The predicted 'area' for each data point was calculated as the mean area of its assigned cluster, and three evaluation metrics were computed to assess the model's accuracy: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R²). These metrics indicated that the model provided reasonable predictions, with the clusters effectively capturing the structure in the data.

Model Evaluation

Model Evaluation Metrics  
MSE-

* MSE is a function that determines the cost by averaging the squares of the errors, which is the average squared gap between estimated and actual values. Imagine we possess a regression model that forecasts the prices of homes. MSE calculates the mean squared error between the true price and the predicted prices by the model (Kumar, 2024).
* If the model anticipates a house to be as an example.  
  The squared error of 320,000 is equal to the square of 400,000. MSE calculates this for every prediction and then computes the average. It focuses on significant errors, especially important in situations such as financial prediction where big errors can have a greater impact (Kumar, 2024).
* RMSE
* RMSE stands for Root Mean Square Error and is a cost function that scales errors to match target values by taking the square root of the mean square error. RMSE rescales the error metric to the price scale in the house pricing example. This simplifies comprehension of the average error based on the real values (Kumar, 2024).
* MAE
* MAE calculates the mean size of the errors in a collection of forecasts, disregarding their orientation. The average absolute difference between the predicted and actual values is calculated. Unlike Mean Squared Error (MSE), it does not square the errors, so it does not penalize larger errors as severely. In the house pricing scenario, MAE handles errors linearly when you deviate by 40,000. This metric becomes handy when you need to prevent imposing additional punishment on significant errors (Kumar, 2024).
* RSquared
* R-Squared represents how much of the variability in the dependent variable can be explained by the independent variables. R squared indicates the accuracy of your predictions in estimating the actual data points. It's comparable to evaluating a test based on a scale of 100%. A high R-Squared value (approaching 1) indicates that your model has a strong ability to accurately forecast the true values. For example, a high R-Squared in forecasting house prices suggests that your model effectively accounts for most of the fluctuations in house prices (Kumar, 2024).

**Assessment of Decision Trees**

The rpart algorithm is used to train the decision tree model in the code, and its evaluation is primarily focused on mean squared error (MSE). Nevertheless, in the case that decision trees are utilized for classification purposes, additional measures like accuracy, precision, recall, F1-score, and the confusion matrix become essential in assessing its efficacy, particularly when the result is categorical.

**Metrics for assessing performance**

* Accuracy: It is the proportion of accurately predicted instances among all predicted instances. It provides a comprehensive evaluation of the model's effectiveness. Nevertheless, accuracy alone might not be enough for imbalanced datasets such as forest fire data, where large fires are rare. Accuracy in classification tasks is calculated based on the information presented in the confusion matrix.
* Precision: a metric that gauges the correctness of positive forecasts. It is especially beneficial when incorrect predictions of fire occurrence come at a high cost. In forest fire prediction, precision indicates the frequency of accurate predictions of burned areas by the model.
* Recall: Recall evaluates how well the model can identify all real positive instances (such as fires). Ensuring high recall is important in fire management because missing a potential fire event can lead to serious consequences, making it crucial to predict most fires.
* F1-Score: It is the harmonic mean of precision and recall. It evenly considers both measures, making it particularly beneficial for situations requiring a compromise between precision and recall, particularly with imbalanced datasets.
* Confusion matrix: displays the model's true positives, false positives, true negatives, and false negatives, giving a visual overview of the decision tree's performance in classification tasks.

The caret package in R can extract these metrics by specifying the model to classify different sizes of burned areas such as small, medium, and large fires. Overfitting can affect the decision tree, but methods such as pruning (like in rpart) can address this issue by simplifying the model, decreasing variance, and keeping predictive accuracy intact.

**Random Forest Evaluation**

The code utilizes the random forest model, which combines several decision trees to produce more reliable and precise predictions through averaging the predictions of each tree. In regression, the evaluation is based on mean squared error (MSE), while in classification tasks, metrics like accuracy, precision, recall, F1-score, and confusion matrix are used, like decision trees.

Accuracy: Random forests typically offer greater accuracy compared to singular decision trees by minimizing overfitting. Accuracy in classification is determined by examining the confusion matrix.  
  
Precision: The precision of random forests is generally greater than that of single trees due to the collective effort of multiple trees in rectifying mistakes made by individual decision trees. By combining the votes of several trees, it decreases the amount of incorrect positive outcomes.

Recall: Random forests tend to enhance recall by being sturdy in detecting true positives. Because the model combines the results of numerous trees, it is more likely to identify a higher number of possible fire occurrences compared to using just one decision tree, which helps in decreasing incorrect classifications of no fire situations.  
  
F1-Score: The F1-Score of the random forest typically surpasses that of individual decision trees because of its ability to effectively balance precision and recall, particularly when dealing with intricate datasets such as forest fire prediction.

Confusion Matrix : The random forest model's confusion matrix assesses classification performance by considering true positives, false positives, true negatives, and false negatives. Random forests generally yield confusion matrices that are more dependable, exhibiting fewer misclassifications than individual trees.

The random forest technique is more resistant to overfitting compared to decision trees as it employs a bootstrapped dataset and randomly picks subsets of features, promoting diversity among trees. This results in increased precision and reduced error rates, which is ideal for predicting forest fires due to the intricate and nonlinear connections between factors such as temperature, humidity, and wind speed.

SVM Evaluation

# Model Comparison Table

| **Model** | **MSE** | **RMSE** | **MAE** | **RSquared** | **Remarks** |
| --- | --- | --- | --- | --- | --- |
| Decision Tree | 85% | 0.82 | 0.84 | 0.83 | Simple but effective |
| Random Forest |  |  |  |  | Most stable performance |
| SVM |  |  |  |  | Best for boundary cases |
| ANN |  |  |  |  | Good with large data |
| KNN |  |  |  |  | Sensitive to neighbours |
| Clustering | N/A | N/A | N/A | N/A | Unsupervised learning |

Final Model Selection and Justification  
Based on the results from the evaluation metrics, justify the selection of the most optimized model. Explain why this model outperforms others and meets the business objectives better.

# Conclusion and Recommendations

Summary of Findings  
Provide a summary of the insights gained from the model building and evaluation. Emphasize how these insights address the business objectives.

## Model Limitations

Discuss any limitations of the models and how these might affect real-world applications

Recommendations  
Offer recommendations for how the business can implement the model and what further steps should be taken for improvement.

Next Steps  
Propose any further work that could improve the model (e.g., additional data, more features, or advanced modeling techniques).

References

# Appendices

* **Appendix A:** Data Dictionary
* **Appendix B:** Code Snippets or Details on Model Building
* **Appendix C:** Visuals and Charts (if not embedded in the main report)