Classifying Australian Bushfires Severity Using Machine Learning

Course: Capstone Project of Applied Data Analytics III

Author: Arya Raj Khadka

Student ID: a1904627

Supervisor: Andrew Merdith, Derrick Hasterok

Date: 24th October 2025

Abstract:

Bushfires are one of the most devastating natural disasters, with the 2019-2020 Black Summer fires highlighting the gaps in national preparedness. This study develops a machine learning framework to classify detected bushfires by satellites into three severity categories- Low, Medium, High- using VIIRS satellite data from 2020-2024. It uses 200,000 stratified fire detection data from 5.3 million MODIS/VIIRS observations. Logistic Regression achieved an accuracy of 0.65 (F1-score:0.69), whereas random forest model reached an accuracy of 0.81(F1-score:0.79), demonstrating its superior predictive power. These findings highlight that machine learning can be used to enhance bushfire severity and support proactive emergency management in Australia.

1. Introduction:

1.1 Problem Context and Motivation:

Australia experiences sever bushfires seasons annually, with fire causing catastrophic impacts on ecosystems, society and economy. For instance, the Australia Black Summer bushfires of 2019–20 were unprecedented in scale and duration, impacting 80% of the continent's population and resulting in AUD 2.4 billion (US$1.5 billion) in insured losses (Shustikova, 2025). Fundamentally, a bushfire is caused by the combination of three conditions –fuel, oxygen, ignition source- which is also known as fire triangle. Though these conditions might cause bushfires, other environmental factors like high fuel load (fallen bark, litter, twigs, and branches), dry fuel moisture, strong wind speed, and high temperature enhance the spread and intensity of it.

While bushfires cannot completely be prevented, we can mitigate the impact by early detection and classification systems. This research aims to develop a Machine Learning (ML) model that can classify Australian bushfire severity to inform emergency response planning.

1.2 Research Objective and Hypothesis:

Primary Hypothesis: Machine learning algorithms can classify Australian bushfire severity using satellite-derived features including fire intensity measurements, temporal pattern, and geographical data.

Secondary Hypothesis:

a) Random Forest will outperform baseline models like Logistic Regression in classification accuracy
b) Combining multiple ensemble algorithms will provide superior performance to single models.
c) Temporal and geographical features will significantly improve classification performance

2. Methodology

2.1 Data Acquisition:

The dataset was collected from NASA FIRMS (Fire Information for Resource Management System), which consisted of dataset from January 1, 2020 – December 31, 2024. Initially, the dataset consisted of 5,303,011 records with 16 features and after feature engineering it ended up with 26 features.

2.2 Data Description and Variables:

Variable	Description	Data Type
Latitude	Geographic latitude coordinate	Continuous
Longitude	Geographic longtitude coordinate	Continuous
region	Broad geographic region	Categorical
Bright_ti4	Brightness temperature I-4 channel (3.55-3.93μm)	Continuous
Bright_ti5	Brightness temperature I-5 channel (10.3-11.3μm)	Continuous
frp	Fire Radiative Power (energy output)	Continuous
confidence_pct	Fire detection confidence level	Ordinal
satellite	Satellite source	Categorical
instrument	Detection instrument	Categorical
scan	Pixel scan size	Continuous
Acq_hour	Hour of acquisition	Integer
year	Year of Detection	Integer
type	Type of fire	Categorical
Acq_date	Date of acquisition	DateTime
Day night		
version	Data processing version	Integer

2.3 Data Preprocessing and Cleaning:

The following preprocessing steps were undertaken to ensure that the data was cleaned and validated.

- **Geographical Filtering:** Records were filtered ensuring that all the records were from Australian region (Latitude: -44 to –10, Longitude: 113 to 154)
- **Coordinate Duplication:** Removed 280 exact coordinate duplicates to prevent biasedness resulting from duplicate values.
- **File Type Filtering:** Only records that are classified as “vegetation fires” (type=0) were retained as the focus of analysis on wildfires and excluded other heat sources.
- **Confidence Score Mapping:** Mapped low confidence=50%, nominal=75%, and high confidence=90% to convert the categorical column to numerical.
- **Temporal Feature Engineering:** The acq_date was converted to date time data type to exact useful features like month, day of year, week of year, season.
- **Spatial Feature Engineering:** Latitude was used to create a region feature, categorizing fires into “South", "Central”, and “North” Australia to explore severity across regions.
- **Binary Indicators:** Features like is_daytime and is_fire_season were created considering they could be a strong predictor of the target variable.
- **Stratified Sampling:** A balanced sample of 200,000 records which preserved the original severity distribution (target variable) was sampled to avoid the computational cost of running the entire dataset.

2.3 Target Variable Creation:

The target variable (severity classification) was engineered based on the fire radiative power (frp), which measures the fire energy output.

a. Class 0 (Low Intensity): frp<10 MW
b. Class 1 (Medium Intensity): 10<=frp<50 MW
c. Class 2 (High Intensity): frp>=50 MW

2.4 Feature Encoding and Scaling:

2.4.1 Categorical Feature Encoding:

A preprocessor using Column Transformer was made for feature encoding and standardization. Features like satellite, day night, region which are nominal were one hot encoded inside the preprocessor, and features like season which has a natural ordering were encoded using Ordinal Encoder. To ensure that multicollinearity is avoided while using One Hot Encoding, only the first column was kept.

2.4.2 Feature Standardization:

StandardScaler from scikit-learn was applied to the numerical features inside the same processor so that all the features can be normalized. The scaling was performed separately in training and testing dataset to prevent data leakage. Though scaling might not be useful for tree-based models like Random Forest, it is crucial for linear models like Logistic Regression.

2.5 Multicollinearity Removal

Two different experiments were run to check the effect of the multicollinearity columns. At first, highly correlated features (r>0.85) were identified and removed. This meant four columns week_of_year,acq_hour,day_of_year and is_afternoon were removed, and the preprocessing steps (encoding and scaling) were applied to the dataset, with a Random Forest Classifier used to predict the output. This gives us a f1_score of 0.7916. A different experiment was conducted where all the preprocessing steps and models were kept constant; however, this time highly correlated features were not removed. This gave us a f1_score of 0.7922. Since keeping four features does not enhance the predictive power and removing those features allows us to better infer the variables, I removed them.

2.6 Machine Learning Modelling:

The dataset from 2020 to 2022 was used as a training set and dataset from 2023 to 2024 was used for testing set. The following model were implemented using Scikit-learn:

- Logistic Regression: A baseline linear model
- Random Forest Classifier: An ensemble method for robust performance

Model performance was evaluated using Accuray and F1-score, with F1-score being more appropriate for multiclass classification with class imbalance.

3. Results

3.1 Exploratory Data Analysis:

3.1.1 Class Distribution:

The target variable consists of imbalances with low-severity fires predominating the distribution:

- Class 0 (low) =72.86%
- Class 1 (Medium)=23.71%
- Class 2 (high)= 3.43%

3.1.2 Bivariate Analysis of Key Relationships:

Our EDA revealed an important relationship between fire severity and both geographical and intensity features.

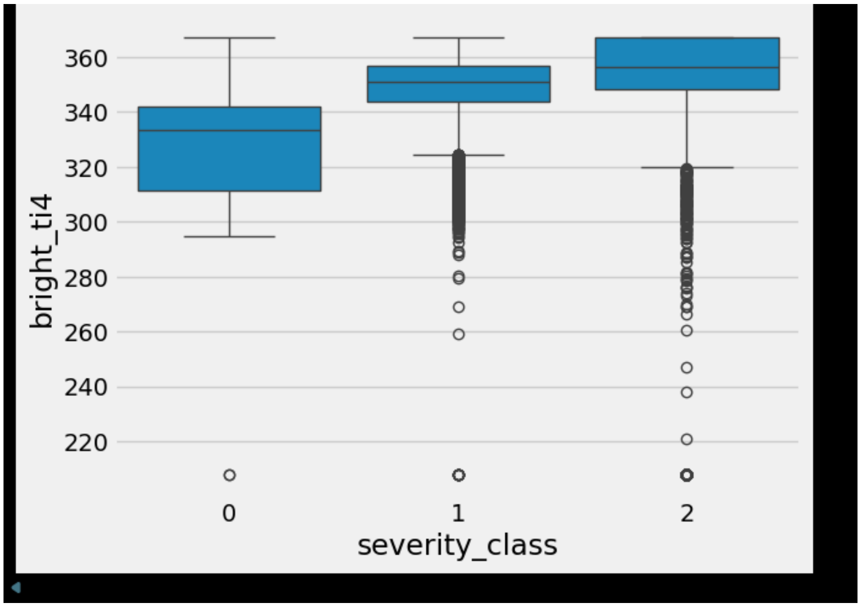
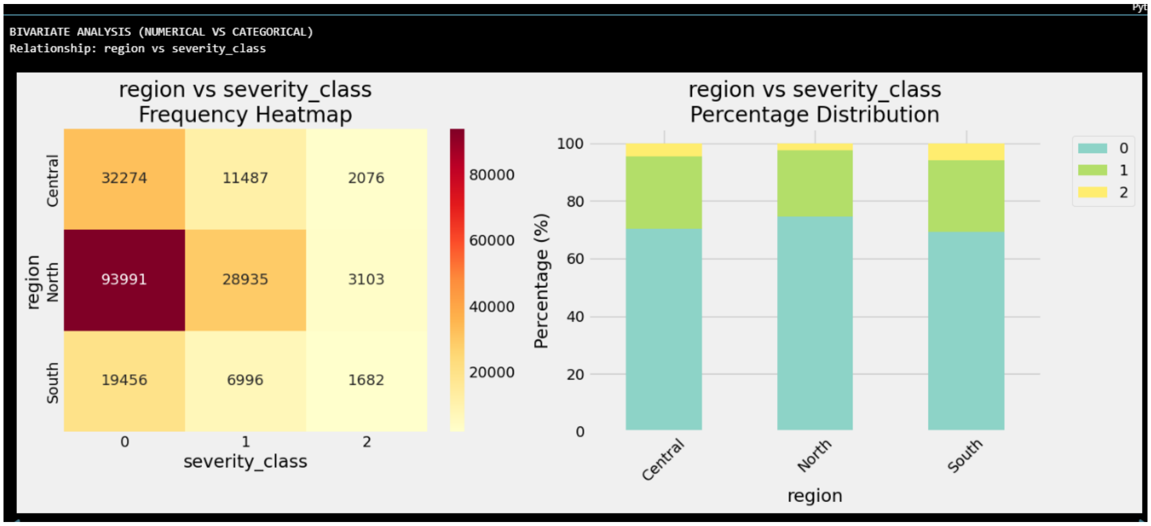


Figure 1: Boxplot showing relationship between severity class and bright_ti4

Figure 1 demonstrates a strong positive relationship between bright_ti4 and fire severity, as expected. The boxplot shows a clear separation between severity classes, with higher severity fires exhibiting significantly higher bright_ti4. Class 2 (High severity) fires show higher median temperatures and extreme outliers. Some of the fires might have less bright_ti4 but still fall in the high severe category which shows that there are other likely features that help in predicting the target variable.



```
BIVARIATE ANALYSIS (NUMERICAL VS CATEGORICAL)
Relationship: region vs severity_class
```

Figure 2: Frequency Heatmap and Stacked bar plot of different region and severity class

Figure 2 reveals significant geographical patterns in fire severity across Australia. The northern region experiences the highest volume of fire across all severity classes. However, the southern region shows the highest proportion of severe fires (Class 2) relative to its total fire count, suggesting that when fires occur in southern Australia, they are more likely to be high severity events.

3.1.2 Statistical Significance Testing:

To quantitively validate the relationships observed in our EDA, we conducted statistical hypothesis tests. For continuous variables like bright_ti4 we conducted one-way ANOVA to test the differences in means across severity classes. For categorical variables like region, we used Chi-square tests of independence.

Table 1: Statistical Test Results for Feature Severity Relationships:

Feature	Test Type	P-Value	Result	Interpretation
bright_ti4	ANOVA	<0.001	Significant	Brightness temperature differs significantly across severity classes
year	ANOVA	9.5*10^-9	Significant	Inter-annual variations in severity are statistically significant
acq_hour	ANOVA	<0.001	Significant	Time of detection relates to fire severity
bright_ti5	ANOVA	<0.001	Significant	Background temperature differs by severity
region	Chi-Square	<0.001	Significant	Geographic region strongly associated with severity
season	Chi-Square	<0.001	Significant	Seasonal patterns in severity are statistically significant
day_night	Chi-Square	<0.001	Significant	Day/night detection relates to severity patterns
is_fire_season	Chi-Square	<0.001	Significant	Fire season status strongly associated with severity

All tested features showed statistically significant relationships with fire severity at the alpha=0.05 level. The extremely small p-values for features like bright_ti4 and region indicate particularly strong association. These results provide statistical confirmation that our feature selection is well-justified.

3.2 Model Performance:

3.2.1 Baseline Model

Table 2: Model Performance Comparison:

Model	Train Accuracy	Test Accuracy	Train F1	Test F1	Consistency (CV)
Logistic Regression	0.691	0.650	0.730	0.700	0.07%
Random Forest	0.838	0.805	0.828	0.794	0.09%

Finding 1: Comprehensive consistency checks across 5 random samples revealed exceptional stability:

- Logistic Regression: 0.07% coefficient of variance (Mean F1:69.7%, Std:0.0005)

- Random Forest: 0.09% coefficient of variation (Mean F1: 79.4%, Std: 0.0008)

Both models demonstrated good consistency ratings, confirming reliable performances across different data samples. This means that the sampling technique that we have used is effective and representative of the entire population.

Finding 2: The Random Forest model significantly outperforms Logistic regression, suggesting that the relationships between features and fire severity are non-linear. So, this hints towards the notion that other nonlinear algorithms could possibly provide us with better results in comparison to linear models.

Finding 3: The gap between the training and testing performance of Random Forest is small (~3.5%), indicating that the model is generalizing well and not severely overfitting.

3.2.2 Advanced Ensemble Modeling: Stacking Classifier

To enhance the predictive power beyond individual models, we implemented an advanced stacking ensemble approach. Stacking classifier leveraged the strengths of multiple algorithms by combining their predictions through a meta-learner, often achieving superior results than single models.

Hyperparameter Optimization with Bayesian Methods:

Traditional hyperparameter tuning methods like Grid Search CV and Random Search CV present significant limitations while searching for the best hyperparameters. Grid Search suffers from curse of dimensionality and becomes computationally expensive as you increase the parameter size. Random Search relies on stochastic sampling without using previous evaluation information.

We employed Optuna, which is a Bayesian optimization framework that intelligently navigates the hyperparameters space by building an objective function, using acquisition function to balance exploration and exploitation and uses optimized sampling technique.

Model Selection and Optimization Process:

The optimization study included four different models:

a) Support Vector Classifier (SVC):
b) Random Forest
c) Gradient Boosting
d) LightGBM

Table 3: Optimized models for stacking ensembles

Model	n_estimator	max_depth	learning_rate	f1_score
LGBM	200	19	0.1020	0.7997
Random Forest	180	20	-	0.7938

LightGBM along with the hyperparameter listed in Table 3 was identified as the best model succeeded by random forest. I furthered my analysis to do more hyperparameter tuning using Optuna for LightGBM, and I was not getting a better f1_score, so I used the same hyperparameter to build the stacking classifier.

Stacking Classifier:

The chosen model was Stacking ensemble, which is a sophisticated meta learner technique that combines the prediction of multiple models (level 0) using a meta-learner (level 1)

- Base Estimators (Level 0):

1) Random Forest: A robust, non-linear model excellent at capturing complex interactions in tabular data. The hyperparameter (n_estimators:180, max_depth:20) was successful in predicting the output efficiently.
2) LGBMCLassifer: A highly efficient gradient bosting framework. The extensive hyperparameter tuning (n_estimators:180, max_depth:20, learning_rate:0.1020) indicated a model optimized for both performance and to avoid overfitting.

- Meta Learner/ Final Estimator (Level 1):

1) Logistic Regression: A linear model was chosen to learn how to best combine the predictions from different models. Its simplicity helps prevent overfitting at the meta-level.

3.3 Feature Importance

Random Forest Classifier also gives a feature importance score to all the features, and below are the top five features as per the random forest model.

Table 3: Feature Importance table from random forest

Feature	Feature Importance	Interpretation

bright_ti4	0.439	Primary fire intensity channel
bright_ti5	0.130	Background temperature reference
confidence	0.085	Detection reliability indicator
is_daytime	0.068	Day/night detection context
scan	0.067	Satellite pixel size parameter

The table clearly shows that bright_ti4 and brigh_ti5 are important predictors of the target variable. The combined importance of bright_ti4 and brigh_ti5 accounts for 56.9% of the model’s predictive power, which highlights the fundamental role of thermal radiation in severity assessment. Moreover, the gap between bright_ti4 and bright_ti5 importance reflects their different physical roles. Bright_ti4 directly measures fire thermal emissions, while bright_ti5 provides background context for atmospheric correction.

4. Discussion

4.1 Interpretation of the Statistical Results

The statistical significance of all tested features validates our feature engineering approach. The high p-values for temporal features confirm the importance of seasonal patterns in Australian wildfire behavior.

4.2 Interpretation of the Results:

Model	Test Accuracy	Train f1_score	Test f1-Score
Logistic Regression	0.650	0.723	0.698
Random Forest	0.804	0.965	0.794
Stacking Classifier	0.806	0.871	0.797

The success of Random Forest model (79% F1-score) demonstrates the feasibility of predicting wildfires' severity from satellite data using non-linear pattern recognition. Also, the minimal gap between training and testing (3.4%) demonstrates effective generalization without overfitting.

The Stacking classifier shows only a 0.3% f1_score improvement in comparison to the Random forest model, which is statistically insignificant. Thus, we can conclude that the standalone model Random Forest can capture the relationship between the features and target variable perfectly.

The Stacking ensemble, however, successfully reduces the overfitting compared to Random Forest (training-test gap: 17.1% for RF VS 7.4% for Stacking) in the f1_score.

4.3 Feature Significance:

The fact that bright_ti4 is a strong predicter of the target variable aligns with the physical expectation. The high feature importance of brightness temperature could possibly reflect a high correlation between FPR and bright_ti5. However, the model's ability to maintain strong performance on the test data from different years suggests genuine physical relationships rather than circular reasoning.

4.4 Interpretation of the Hypothesis:

Hypothesis 1: "Ensemble methods outperform any single model for this complex prediction task"

Interpretation and Support: Partially rejected

- While the Stacking Classifier technically achieved the highest f1_score, the improvement of only 0.3% f1_score is negligible in practical terms. The added complexity of the stacking ensembles does not justify the minimal gain. The standalone Random Forest model proves to be nearly as effective and much more efficient.

4.4 Limitations:

1. Vegetation Data Absence: NDVI/EVI vegetation indices not integrated despite collection, as it was missing critical fuel conditions.
2. Computational Constraints: Analysis limited to 200,000 stratified samples due to processing limitations.
3. Data Scope: Focused on satellite features rather than comprehensive meteorological data.
4. Temporal Scope: Focused of VIIRS data from 2020-2024; historical data excluded to ensure temporal consistency

5. Conclusion:

The project successfully developed a robust machine learning model for wildfire severity classification, achieving an F1-score of 0.79 and exceptional consistency. Feature engineering and preprocessing were crucial steps in achieving these results. Moreover, minimal overfitting is a positive sign of the robust nature of both models. Though Stackijng Classifier is able to achieve slightly better output, the computational complexity and training time undermines its performance gains over the much simple Random Forest.

6. References:

1) Bushfire Severity Modelling and Future Trend Prediction Across Australia: Integrating Remote Sensing and Machine Learning. (2019). Arxiv.org. https://arxiv.org/html/2410.02963v1
2) Henderson, K., & Chakrabortty, R. K. (2024). A machine learning predictive model for bushfire ignition and severity: The Study of Australian black summer bushfires. Decision Analytics Journal, 100529. https://doi.org/10.1016/j.dajour.2024.100529
3) Partheepan, S., Sanati, F., & Hassan, J. (2025). Modelling bushfire severity and predicting future trends in Australia using remote sensing and machine learning. Environmental Modelling & Software, 188, 106377. https://doi.org/10.1016/j.envsoft.2025.106377
4) Shustikova, I. (2025, February 23). *Black Summer five years on: A sobering reminder of Australia bushfire risk*. Moody's. https://www.moodys.com/web/en/us/insights/insurance/black-summer--five-years-on---a-sobering-reminder-of-australia-b.html