Predicting Wildfire Features Based on Weather and Climate Statistics

Authors: Riley Neher, Akshay Srinivasan, and Amando Xu

New York University, New York City, NY
Emails rn50@nyu.edu ats497@nyu.edu acx202@nyu.edu

Abstract. Forest fires are one of the most pressing issues when discussing climate change. A changing climate has led to a warmer, drier climate, leaving forests more susceptible to fires. In the past 32 years, the average annual area burned in Australia grew by 800 percent. Within the 2019-2020 bushfire season alone, more than a billion animals perished. As these devastating fires become more frequent and severe, accurate future predictions become crucial. If we are able to accurately identify the areas most at risk for wildfires, we can take appropriate action to prevent them from occurring or mitigate the damage they cause. Preventing wildfires is also crucial to preventing further climate change and, therefore, further wildfires. In this project, we tested a variety of machine learning models to predict fire area, brightness, radiative power, and the count of fires per day. We believe this project, if successful, can have large impacts for preserving and protecting Australian forest land.

1 Introduction

Wildfires are characteristic of the Australian climate, however climate change has dramatically increased their frequency and severity in recent years. (Hughes and Alexander, 2017) In the past 32 years, the average annual area burned in Australia has grown by 800 percent. (Foley, 2021) This is not a coincidence; there is a clear correlation between the vegetation dryness, air temperature, wind speed and humidity and the amount of area burned yearly from 1930. (Foley, 2021) While the correlation between an increase in these factors and human-induced climate change is hard to determine, researchers have been able to quantify the impact of climate change on the likelihood of wildfires that have occurred in recent years. For example, it was found that anthropogenic climate change increased the likelihood of the 2016 Canadian Fort McMurray fires by 2-4 times. (Kirchmeier-Young et. al., 2019) Furthermore, Kirchmeier-Young et. al. used empirical relationships between weather metrics and fire activity in the region and estimated that 86-91% of the burned area in British Columbia in 2017 was attributable to anthropogenic climate change. (Kirchmeier-Young et. al., 2019) In Australia, Lewis et. al. determined that the extreme temperatures at the time of the 2018 bushfires were 4.5 times more likely due to human-made climate change. (Lewis, et. al., 2021)

In order to fully understand this relationship, the causes of a wildfire must be understood. In order for a wildfire to burn a significant amount of area, it needs three things. The fire needs to be started, fuel to burn, and weather and topography that are conducive to its spread. (Hughes and Alexander, 2017) Many fires are either deliberately or accidentally lit by people, whether that be through powerline faults, uncontrolled campfires, or other accidents. The relationship between weather conditions and fuel for a fire in an area is complex. Warmer, drier conditions can make the fuel drier and more flammable, however

wetter conditions can lead to increased plant growth and create more fuel for the fire. (Hughes and Alexander, 2017) If a wet season is followed by a dry, hot season, the likelihood of wildfires increases. Human-induced climate change also increases the variability of seasons, thus causing very wet seasons where lots of fuel is produced, followed by very dry seasons where this fuel can be burned.

Due to the level of devastation that wildfires can cause and the links between wildfires and climate change, there has been an increasing interest in being able to predict various metrics about wildfires using machine learning tactics. This work can provide fire management teams with more accurate predictive power and identify other connections between climate factors and the prevalence of wildfires. Most of the research up until this point has been concerned with classifying the presence of a wildfire, not predicting metrics about the fire. For example, Sayad et. al. used a variety of different models to classify satellite images of containing a wildfire or not. (Sayad et. al., 2019) They found that a neural network and an SVM model performed the best in classifying the images. Before they could use a model, however, they had to perform a significant amount of data cleaning, interpolation (to normalize the satellite images in one timespan), and extrapolation (getting the data for future dates). This is a substantial limitation in this method of prediction, because it requires more tools and satellite data is not very easy to collect.

However, another set of researchers were able to predict information about wildfire area, rather than wildfire classification. They were able to predict which areas surrounding a burning wildfire have a high risk of imminent wildfire spread. (Radke, et. al., 2019) Their model, FireCast, uses various inputs such as satellite imagery, elevation data, weather data, and historical fire perimeters to identify patterns associated with fire spread in certain environments to produce predictions of wildfire spread. They used a CNN to input all the features and output the areas surrounding the current fire perimeter that are expected to burn within the next 24 hours. This paper, taking in weather features and outputting the future burn area, could guide us in predicting the burned area of the wildfires in our dataset.

Identifying which features were the most influential when predicting variables about these fires is also very important. In a study by Hong et. al., the researchers determined the characteristics that contributed to the most fire-prone areas. (Hong et. al., 2019) This could be helpful for identifying risk factors for fire management teams and finding links between anthropogenic climate change factors and wildfires. We did the same in our study.

Overall, due to the increasing pervasiveness of wildfires across the globe and the anthropogenic causes of these fires, researchers are becoming increasingly interested in predictive modeling and identifying which factors may lead to the fires. Most of this research has been centered around using satellite imagery to classify images containing wildfires or burned area. While this is useful and can lead to accurate classification, this kind of data is not readily available and there is a significant amount of pre-processing of the data needed to feed it into a model. Our method, using simple weather statistics that do not need a satellite to collect, provides more accessibility and ease. We intend to use these features to predict various outcome variables such as fire area, brightness, radiative power, and count of fires per day and identify which features are most helpful in identifying wildfires.

2 Methods

2.1 Data cleaning & Prep

The research data utilized in this study was derived from IBM's Spot Challenge for Wildfires, a datathon competition that tasked participants with creating models to predict wildfires.

The data, collected by NASA through satellite imagery, was processed using IBM PAIRS - Geoscope, a Physical Analytics Integrated Data Repository and Services system that converted the 6 PB of imagery into data-frames for analysis and prediction. This data was organized into several data-sets, including Historical Weather Forecasts, Historical Weather, Vegetation Index, Land Class, and Historical Wildfires, and was further subdivided into seven Australian states.

Most of the data-sets contained minimum and maximum values, means, and variances. The historical weather data also included specific rows for precipitation, relative humidity, soil water content, solar radiation, temperature, and wind speed on a given date. Land class data encompassed percentages of various land cover types, while historical weather forecast data contained predictions and their respective lead times. The historical fire data included estimations of fire area, mean estimated brightness, fire radiative power, and count, which served as the outcome variable for the predictive model.

To conduct the analysis, we first consolidated all data-sets into a single large data-frame. Rows specific to each data-set, such as weather and weather predictions, were pivoted into columns, as each region had multiple rows for each date pertaining to relative humidity, precipitation, etc. We then merged all data-sets into a final combined data-set for analysis. To address missing data, we replaced it with region-specific averages, as filling in zeros would have significantly skewed the results. Region-specific averages provided a more accurate approximation of the conditions compared to zeros or feature-wide averages. For the historical weather forecast data with missing predictions, we substituted the missing data with actual weather conditions and adjusted the lead time, typically ranging from 5 to 15 days, to 0 days.

2.2 Principal Component Analysis

Due to the nature of the data, the features were very likely to be multicollinear. To determine this, we constructed a correlation matrix, shown in Figure 5. As seen in the matrix, many of the variables are collinear with each other. This means that PCA would be useful to extract the important information from the data and reduce any multicollinearity issues that may arise.

An important part of Principal Component Analysis is choosing the number of principal components used in the study. The aim is to use the least number of components possible to capture above a certain threshold of variation in the data. We decided to use a threshold of preserving 95% variation in the data, so that too much variation would not be lost when we reduce dimensionality. We plotted the amount of explained variance with different component numbers and found that 95% of the variance is explained with 9 components. Therefore, for the rest of the models shown below, we pre-processed the data through a principal component analysis with 9 components.

2.3 Models Development

In this study, we employed various machine learning models to predict several fire-related variables, including estimated fire area, mean estimated fire brightness, mean estimated fire radiative power, and daily fire count. Our primary focus was on the estimated fire area as the outcome variable. The models tested were basic linear regression, ridge regression, lasso regression, fully connected neural network (multilayer perceptron), recurrent neural network (LSTM), XGBoost, LightGBM, Random Forest Regressor, and Extra Trees Regressor. To

measure the accuracy of each model, we calculated the root mean squared error (RMSE) and the coefficient of determination (R^2).

Basic Linear Regression A simple linear regression model was developed to serve as a baseline for comparison with more sophisticated models. We fit the linear regression model to the training data and made predictions for the test set.

Ridge Regression Ridge regression is a linear regression model with L2 regularization, which can help reduce overfitting and improve generalization. We used cross-validation to find the optimal regularization parameter (alpha) for the model and applied it to make predictions.

Lasso Regression Lasso regression is a linear regression model with L1 regularization, which promotes sparsity in the resulting coefficients. This allows for automatic feature selection, which can be useful in cases with many correlated features. Like ridge regression, we utilized cross-validation to find the optimal regularization parameter (alpha) and employed the model for prediction.

Fully Connected Neural Network (Multilayer Perceptron) We designed a fully connected neural network with the following architecture:

- Input layer with the number of nodes equal to the number of features (principal components) and a ReLU activation function.
- Hidden layer with 64 nodes and a ReLU activation function.
- Hidden layer with 32 nodes and a ReLU activation function.
- Output layer with a single node and a linear activation function.

The model was trained using the training data for 10 epochs.

Recurrent Neural Network (LSTM) We created a recurrent neural network using Long Short-Term Memory (LSTM) cells to capture temporal dependencies in the data. The architecture was as follows:

- Input layer with LSTM cells, containing 64 nodes and a ReLU activation function.
- Dense hidden layer with 32 nodes and a ReLU activation function.
- Output layer with a single node and a linear activation function.

The artificial neural networks were trained for 10 epochs using the training data.

XGBoost XGBoost is an advanced implementation of gradient boosting, known for its high performance and scalability. We trained the XGBoost model on the training data with default hyperparameters and used it to predict the test data.

LightGBM LightGBM is a gradient boosting framework that employs tree-based learning algorithms. We trained the LightGBM model with default hyperparameters on the training data and made predictions on the test data.

Random Forest Regressor A Random Forest Regressor is an ensemble learning model that constructs multiple decision trees and combines their predictions. We trained the model with default hyperparameters on the training data and used it for prediction.

Extra Trees Regressor The Extra Trees Regressor is another ensemble learning model that constructs multiple decision trees, but it uses a more randomized approach to split selection. We trained the model with default hyperparameters on the training data and applied it for prediction.

Metrics to Determine Model Performance For each of the models, we split the data into training and testing sets with a test size of 0.3. To measure the accuracy, we calculated the root mean squared error (RMSE) and the coefficient of determination (R^2). These metrics allowed us to assess the performance of each model and compare their relative strengths and weaknesses.

3 Results

Table 1: Outcome Variable: Estimated Fire Area

Model	Train RMSE	Train r^2	Test RMSE	Test r^2
Linear Regression	289.31	0.16	288.85	0.15
Ridge Regression	289.31	0.16	288.85	0.15
Lasso Regression	289.31	0.16	288.85	0.15
Fully Connected Neural Network (Multilayer Perceptron)	277.86	0.22	278.03	0.21
Recurrent Neural Network (LSTM)	275.66	0.23	276.30	0.22
XGBoost	142.95	0.79	279.52	0.20
LightGBM	215.30	0.53	272.09	0.24
Decision Tree Regressor	272.05	0.25	279.79	0.20
Random Forest Regressor	**105.67**	**0.89**	**272.40**	**0.24**
Extra Trees Regressor	283.96	0.19	284.08	0.18

When predicting estimated fire area, it is unclear why the random forest regressor performs the best. It is likely that the assumptions which the model operates based off of just coincidentally align with the data. The linear regression models did not perform well against this outcome variable, probably due to the outliers in the data. This outcome variable is very skewed, with very large outliers– which linear regression does not handle well.

Table 2: Outcome Variable: Mean Estimated Fire Brightness

Model	Train RMSE	Train r^2	Test RMSE	Test r^2
Linear Regression	**6.97**	**0.38**	**6.95**	**0.38**
Ridge Regression	**6.97**	**0.38**	**6.95**	**0.38**
Lasso Regression	**6.97**	**0.38**	**6.95**	**0.38**
Fully Connected Neural Network (Multilayer Perceptron)	9.45	-0.14	9.55	-0.16
Recurrent Neural Network (LSTM)	9.46	-0.14	9.59	-0.17
XGBoost	5.41	0.63	7.15	0.35
LightGBM	6.28	0.50	6.97	0.38
Decision Tree Regressor	6.98	0.38	7.08	0.36
Random Forest Regressor	2.68	0.91	7.21	0.34
Extra Trees Regressor	7.08	0.36	7.11	0.35

When predicting mean estimated fire brightness, the linear regression models (basic linear regression, ridge regression, and lasso regression) perform the best. This may be because mean estimated fire brightness, unlike the other outcome variables, has a roughly normal distribution with few outliers (as shown in Figure 2). Linear regression does not perform well with data that contains outliers, which may be why it was not the highest performing among the other outcome variables. It may also just be the case that there was more of a linear relationship between the features and the mean estimated fire brightness than the other outcome variables.

Table 3: Outcome Variable: Mean Estimated Fire Radiative Power

Model	Train RMSE	Train r^2	Test RMSE	Test r^2
Linear Regression	64.47	0.09	64.00	0.09
Ridge Regression	64.47	0.09	64.00	0.09
Lasso Regression	64.47	0.09	64.00	0.09
Fully Connected Neural Network (Multilayer Perceptron)	**64.12**	**0.10**	**63.67**	**0.10**
Recurrent Neural Network (LSTM)	63.96	0.11	63.71	0.10
XGBoost	42.82	0.60	65.50	0.05
LightGBM	56.33	0.31	64.56	0.08
Decision Tree Regressor	61.68	0.17	64.75	0.07
Random Forest Regressor	24.79	0.87	65.79	0.04
Extra Trees Regressor	63.95	0.11	63.92	0.09

With this outcome variable, the neural networks perform the best. The precise reason for this is unknown, however they do not perform much better than the linear regression models. Therefore, it is likely that the input data has a roughly linear relationship with the dependent variable.

Table 4: Outcome Variable: Count of Fires Per Day

Model	Train RMSE	Train r^2	Test RMSE	Test r^2
Linear Regression	137.45	0.17	139.3	0.16
Ridge Regression	137.45	0.17	139.3	0.16
Lasso Regression	137.45	0.17	139.3	0.16
Fully Connected Neural Network (Multilayer Perceptron)	131.87	0.23	133.84	0.22
Recurrent Neural Network (LSTM)	131.41	0.24	133.59	0.23
XGBoost	70.81	0.78	132.92	0.23
LightGBM	**104.03**	**0.52**	**130.70**	**0.26**
Decision Tree Regressor	129.04	0.27	138.64	0.17
Random Forest Regressor	49.95	0.89	131.77	0.25
Extra Trees Regressor	135.03	0.20	137.62	0.18

When predicting count of fires per day, the LightGBM model performed the best. It is interesting that it seems XGBoost suffered some overfitting (a larger difference between training and test set results), when LightGBM is traditionally known for overfitting due to the use of deeper decision trees. However, it is clear that for this outcome variable, the data was best suited for LightGBM.

It should be noted that when we originally ran the decision tree models (decision tree regressor and extra trees regressor), we saw severe overfitting. After hyperparameter tuning, this overfitting was minimized.

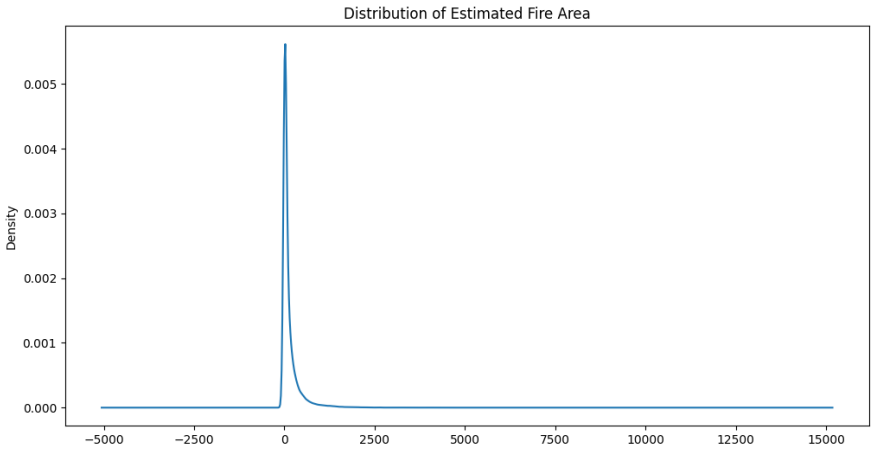

Fig. 1. Distribution of Estimated Fire Area

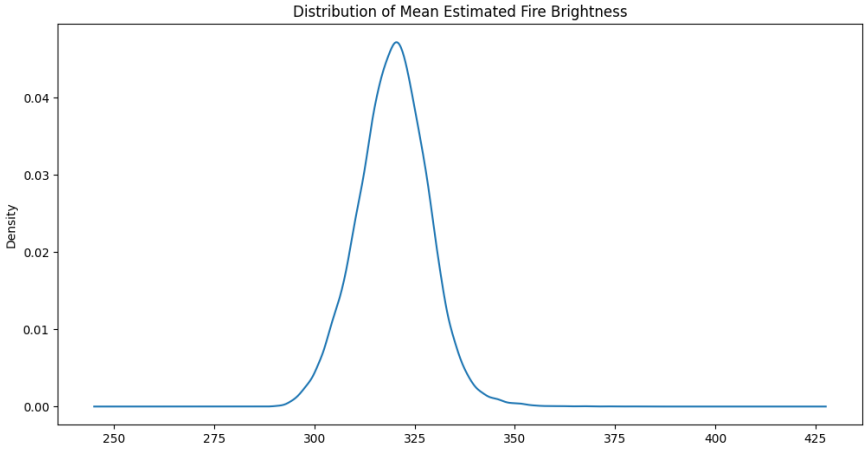

Fig. 2. Distribution of Mean Estimated Fire Brightness

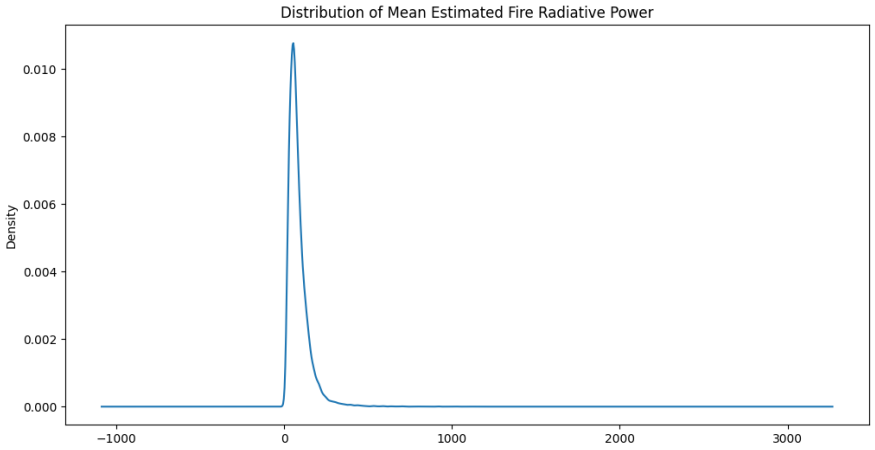

Fig. 3. Distribution of Mean Estimated Fire Radiative Power

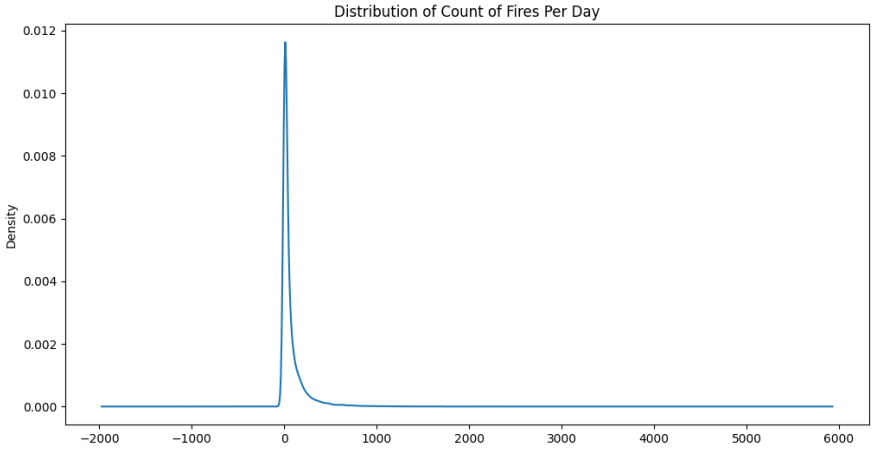

Fig. 4. Distribution of Count of Fires Per Day

4 Discussion

Our study aimed to predict wildfire features based on weather and climate statistics, utilizing machine learning models to better understand the complex relationships between various factors and the increasing frequency and severity of wildfires. By employing easily accessible weather statistics that was given rather than satellite imagery, our approach provides greater accessibility and facilitates the prediction of various outcome variables related to wildfires. The findings of our research contribute to the growing body of work on understanding and predicting wildfires, such as the studies by Hughes, Alexander and Kirchmeier-Young.

In the process of predicting estimated fire area, random forest regressor performed the best, potentially due to the model's assumptions aligning with the data. Linear regression models were less successful in this case, likely due to the presence of outliers in the data, which resulted in a skewed distribution. For mean estimated fire brightness, linear regression models performed well, possibly because of the roughly normal distribution of the data with few outliers, or perhaps because the features had a more linear relationship with the mean estimated fire brightness than with other outcome variables. When predicting fire radiative power, neural networks showed the best performance, although the difference between them and linear regression models was relatively small, suggesting a roughly linear relationship between the input data and the dependent variable. Lastly, the LightGBM model performed the best in predicting the count of fires per day, indicating that this model was most suited to the data for this outcome variable.

In line with the findings by Hughes and Alexander (2017), our study confirmed the significant influence of climate change on the increasing threat of wildfires. By identifying the relationships between various weather and climate factors and the occurrence and characteristics of wildfires, our research can help inform fire management teams and policymakers on effective strategies to mitigate the impacts of climate change and reduce the frequency and severity of wildfires. Moreover, our study extends the current understanding of these relationships by utilizing machine learning models to predict specific wildfire features, such as fire area, brightness, and radiative power, based on weather and climate statistics.

Kirchmeier-Young (2019) highlighted the role of human-induced climate change in exacerbating extreme fire seasons. Our study further explored this relationship by examining the impact of weather and climate factors on various wildfire features, providing a more comprehensive understanding of the complex interplay between climate change and wildfires. The predictive models employed in our study can aid in identifying the areas most at risk for wildfires, which can inform prevention and mitigation efforts, ultimately contributing to the reduction of anthropogenic impacts on climate and wildfires.

Despite the promising results, there are some limitations to our study. The data-set used was specific to Australian wildfires, and our findings may not generalize to other geographical areas with different climatic conditions or vegetation types. Furthermore, while we addressed multicollinearity issues through PCA, additional factors not included in the data-set, such as human activities or land management practices, could enhance the predictive power of our models. Future research could explore the applicability of our approach to other regions and incorporate a broader range of factors to further improve the predictive models.

In conclusion, our study demonstrated the potential of using weather and climate statistics to predict wildfire features, providing valuable insights for fire management teams and helping to identify links between climate factors and the prevalence of wildfires. Our research builds upon and extends the findings of related studies, such as those by Hughes, Alexander and Kirchmeier-Young, and highlights the importance of accessible data and the potential of machine learning models in predicting wildfire characteristics. By further advancing our understanding of the complex relationships between climate change and wildfires, we can contribute to the development of effective strategies for mitigating the devastating impacts of wildfires on ecosystems, communities, and the global climate.

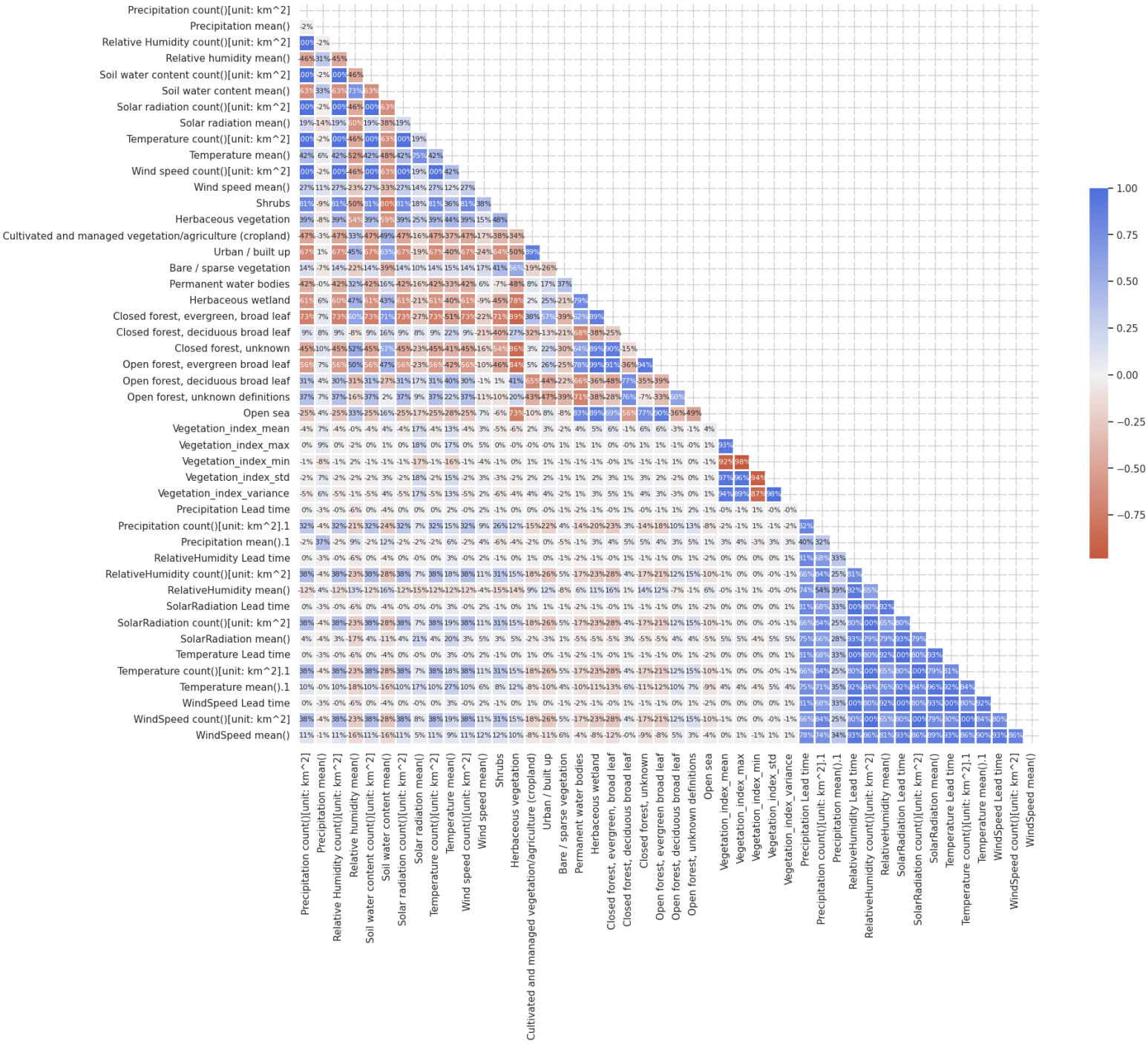

Fig. 5. Correlation Matrix

References

1. Hughes, L., Alexander, D. (2017). Climate change and the Victorian bushfire threat: Update 2017. Sydney: Climate Council of Australia.

2. Foley, M. "CSIRO Study Proves Climate Change Driving Australia's 800% Boom in Bushfires" The Sydney Morning Herald. 2021.

3. Kirchmeier-Young M.C et. al. "Attribution of the Influence of Human-Induced Climate Change on an Extreme Fire Season" Earth's Future. vol. 7. no. 1. 2019.

4. Lewis, S. et. al. "Attribution of the Australian Bushfire Risk to Anthropogenic Climate Change" European Geosciences Union. vol. 21. no. 3. 2021.

5. Sayad, Y. et. al. "Predictive modeling of wildfires: A new dataset and machine learning approach" Fire Safety Journal. vol. 104. 2019.

6. Radke, D. et. al. "FireCast: Leveraging Deep Learning to Predict Wildfire Spread" The Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019.

7. Hong, H. et. al. "Predicting spatial patterns of wildfire susceptibility in the Huichang County, China: An integrated model to analysis of landscape indicators" Ecological Indicators. vol. 101. 2019.