Classifying Wildfire Size Using Generalized Linear Models: An Analysis of Spatial and Climatic Influences

Varsha Vijay

Received December 29, 2024
Accepted July 27, 2025
Electronic access August 15, 2025

Wildfires are a significant cause of property destruction, loss of life, and deforestation in the United States. Various factors, including geospatial location, time of year, temperature, relative humidity, precipitation, wind speed, vegetation type, aridity, and tree cover, influence the size and severity of a wildfire. Although wildfire risk assessment, climate change impacts, and fire spread prediction have been widely studied, there is limited research on using machine learning (ML) models to predict wildfire size for practical early-warning systems. This study aims to evaluate the effectiveness of several ML classification models in predicting the size of wildfires categorized into five distinct classes. I conducted experiments on a comprehensive dataset containing historical wildfire records, applying five machine learning models: Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Decision Tree, and Multilayer Perceptron. The experimental results demonstrate that these algorithms can accurately classify wildfire size based on environmental and geographic variables. Notably, the Multilayer Perceptron and Logistic Regression models each achieved an accuracy exceeding 82% in size classification. However, challenges with class imbalance, relevant feature selection, and model interpretability remain, indicating a need for more detailed evaluation metrics and highlighting directions for future improvement. Overall, this study highlights the potential of machine learning to support wildfire management and mitigation strategies through more precise size prediction.

1 Introduction

Wildfires present a significant threat to both natural environments and human populations, with increasing frequency and severity observed globally, particularly in the United States. Various factors influence wildfire size and behavior, including geospatial location, time of year, temperature, relative humidity, precipitation, wind speed, vegetation type, aridity, and tree cover. The interaction between these factors makes predicting wildfire size a complex challenge.

While wildfire spread prediction and risk assessment have been well studied, accurately classifying wildfire size at the point of ignition is just as vital for effective disaster response and resource planning. Understanding the potential size of a wildfire allows firefighting agencies to allocate resources efficiently, prioritize high-risk fires, and mitigate catastrophic damage. Unlike spread models, which require continuous updates as fires evolve, size classification offers an early estimate that can inform initial containment strategies. Better wildfire size classification can strengthen early intervention, lower suppression costs, and ultimately reduce damage to ecosystems and communities.

In this study, I utilized the USDA's spatial wildfire occurrence dataset, which includes records of wildfires across the United States from 1992 to 2020 (Short, 2022)[1]. It provides detailed geospatial information, fire sizes and classifications, occurrence dates and times, and cause classifications. Temperature, wind speed, and precipitation data were sourced from the Meteostat API (Meteostat, nd)[2]. By integrating these datasets, I tested various prediction algorithms to classify wildfire sizes at ignition, evaluated key features influencing size, and compared multiple machine learning models to determine the most effective approach.

The strong performance of the Multilayer Perceptron and Logistic Regression models suggests that neural networks and linear models can effectively classify wildfire size given the right conditions. However, the observed biases point to a need for more balanced datasets and diverse feature sets.

Future research should explore methods to enhance classification accuracy, such as incorporating additional climatic and geographical features and employing data augmentation techniques to address class imbalance.

While factors like vegetation type and humidity play a crucial role in wildfire behavior, their absence is a limitation; however, classifying wildfire size with the available data still provides valuable insights. This study aims to evaluate the effectiveness of machine learning models in predicting wildfire size classes based on spatial, temporal, and climatic data. By focusing on this classification task, we seek to provide a foundation for improved decision-making in wildfire management, even when faced with data limitations.

2 Related Work

Numerous studies have examined various aspects of wildfire behavior and risk. For instance, Westerling and Bryant (Westerling and Bryant, 2006)[3] analyzed wildfire risks in California under different climate change scenarios, revealing that rising temperatures and changing precipitation patterns significantly affect fire frequency and property damage, particularly in areas where wildlands meet urban development. Keeping, Harrison, and Prentice (Keeping et al., 2024)[4] developed a generalized linear model to predict daily wildfire occurrence across the contiguous United States, achieving high predictive accuracy and providing insights into key wildfire predictors.

Shadrin et al. (Shadrin et al., 2021[5]) employed a deep neural network model using the Multi-scale Attention Net architecture to predict wildfire spread on a large scale, demonstrating excellent predictions in the 1 to 5 days following the start of a fire. Additionally, Westerling et al. (Westerling et al., 2011)[6] projected future wildfire occurrence and burned areas in California under various climate and development scenarios, anticipating substantial increases in burned areas by 2085.

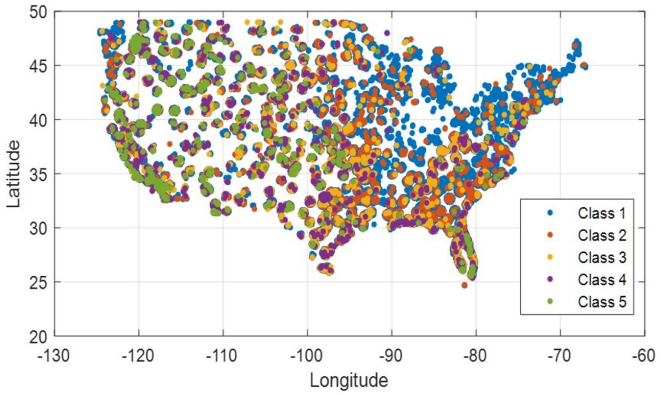

Fig. 1 Overlay of 72,029 wildfire geospatial data points on Lat-Lon grid based on size

Although much research has focused on predicting wildfire occurrence and spread, less attention has been given to predicting wildfire size at ignition. Predicting wildfire size at ignition is a critical early-stage task, distinct from spread prediction, and helps guide initial containment and resource deployment. Recent machine learning approaches have shown promise in related areas, but further exploration is needed to assess their effectiveness in early size classification using comprehensive spatial, climatic, and temporal data.

3 Methodology

3.1 Dataset Description

The USDA metadata provides a comprehensive spatial database of wildfires in the United States from 1992 to 2020, encompassing over 2.2 million wildfires and representing a total of 180 million acres burned. Each fire is classified according to its size using a code from A to G, where A represents the smallest fires and G denotes the largest, as detailed in Table 1. For this study, smaller fires classified as codes A and B were excluded, as they typically occurred in areas where they could be more easily suppressed. The remaining classes, C to G, were renamed to classes 1 through 5 for simplicity in analysis. Matching climatic variables (temperature, wind speed, precipitation) from the Meteostat API with the USDA wildfire records was challenging due to missing data. For the scope of this study, records containing incomplete or missing values for any of the selected features were excluded, resulting in a curated dataset of 72,029 data points that were successfully matched and used for analysis. This approach ensured that all included data points had complete information for model training, though it led to a reduction in the total number of available records. Figure 1 shows the overlay plot of these data points on a Lat-Lon grid.

3.2 Data Preprocessing

To prepare the dataset for analysis, I applied several preprocessing steps. First, I filtered the USDA spatial wildfire occurrence dataset to focus solely on wildfires within California, excluding records from other regions in the United States. This decision was made to refine the study's scope and ensure the analysis was region-specific, as California experiences a high frequency of wildfires with diverse characteristics, providing a suitable case study for this research.

Next, I integrated climatic data, including temperature, wind speed, and precipitation, sourced from the Meteostat API, matching these variables with the corresponding wildfire records. Given the comprehensive nature of the USDA dataset, which included a wide array of features, I further refined the data by selecting only the most relevant features for this study. Specifically, I retained DayofYear, Temperature, Precipitation, WindSpeed, and geospatial features (gridX and gridY) that were critical to predicting wildfire size.

The original wildfire size data had a naturally skewed distribution, with many small fires and fewer large ones, reflecting real wildfire patterns. To manage this imbalance and provide a more granular analysis, the data was re-categorized into five classes (1-5). While using all the original size classes would have caused even greater imbalance, this recategorization still left some uneven distribution, but to a lesser extent.

These preprocessing steps reduced data noise and ensured that the model features were meaningful and manageable. This

streamlined dataset enabled more effective model training and evaluation, allowing the analysis to focus on key factors influencing wildfire size in California.

3.3 Machine Learning Algorithms

Five machine learning algorithms—Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Decision Tree, and Multilayer Perceptron—were employed to predict wildfire size. These algorithms were selected for their diversity in approach and suitability for classification tasks. Logistic Regression and Naïve Bayes offer effective baseline models, providing insight into simpler, interpretable methods. K-Nearest Neighbors was included for its non-parametric flexibility in capturing decision boundaries, while Decision Tree and Multilayer Perceptron provide more complex decision-making processes, with the latter offering potential for capturing intricate patterns in data.

Hyperparameter optimization was performed systematically to maximize each model's predictive performance. This process employed a Grid Search strategy combined with 5-fold cross-validation to achieve optimal results. Although Gradient Boosting methods, such as XGBoost and Random Forest, could offer improved performance, these models were not included to focus on a range of simpler, interpretable models that allow for a comparative evaluation of feature effectiveness.

3.4 Geospatial Encoding

For feature engineering, geospatial encoding was employed to transform latitude and longitude information into a grid-based format. Geographic coordinates were converted into gridX and gridY values starting at zero. This transformation enables more effective spatial analysis by consistently structuring location data, helping the model learn spatial patterns and relationships.

3.5 Feature Set

The selection of features — DayofYear, Temperature, Precipitation, WindSpeed, gridX, and gridY—were chosen for their relevance to wildfire behavior and to capture key influencing factors. DayofYear accounts for seasonal variations in wildfire risk. Temperature, Precipitation, and WindSpeed are critical climatic variables influencing fire intensity and size. The gridX and gridY variables, derived from geospatial encoding, represent the spatial location of each fire in a structured grid format. Together, these features provide a robust framework for predicting wildfire size by incorporating temporal, climatic, and spatial dimensions.

Although this feature set captures key aspects of wildfire behavior, additional factors, such as vegetation type, soil moisture, and humidity could further improve the model's performance. These features were not included in the current analysis due to data limitations. Additionally, past fire history could help in understanding wildfire behavior, but is more complex to quantify in a way that can be directly used by machine learning models.

Fire Size Class Code	Final Fire Size in Acres
A	<1/4
B	1/4 - 9
C	10 - 99
D	100 - 299
E	300 - 999
F	1000 - 4999
G	5000+

Table 1: Original Fire class codes based on size

3.6 Partitioning Data

The dataset was divided using a 3:1 split, with 75% of the data designated for training the model and the remaining 25% reserved as the test set.

4 Results and Discussion

4.1 Model Evaluation

Table 2 presents the average accuracy of each algorithm across the 5-fold cross-validation. The Multilayer Perceptron and Logistic Regression models achieved the highest accuracy, showcasing their effectiveness in classifying the size of wildfires based on the selected features. Figure 2 visually compares accuracies of the models. Despite the high accuracies, the confusion matrices plotted in Figure 3 reveal that both the multilayer perceptron and the logistic regression models predominantly predicted all fires as class 1, indicating a bias due to the skewness of the data set towards class 1. Although these results provide an initial comparison, a more rigorous evaluation, including statistical significance tests, such as confidence intervals, to assess the reliability and variability of these accuracy measures, will be performed in future work.

4.2 Feature Importance Analysis

Figure 4 illustrates the feature importance of each algorithm. The analysis indicates that the Decision Tree and K-Nearest Neighbors algorithms performed better at classifying across different size classes compared to the Multilayer Perceptron and Logistic Regression models. The feature importance analysis highlights the need for a broader set of features to improve classification across all size classes.

Model	Accuracy
Naïve Bayes	76.8%
K-Nearest Neighbors	80.7%
Decision Tree	74.9%
Multilayer Perceptron	82.1%
Logistic Regression	82.2%

Table 2: Algorithm Model Accuracy Values

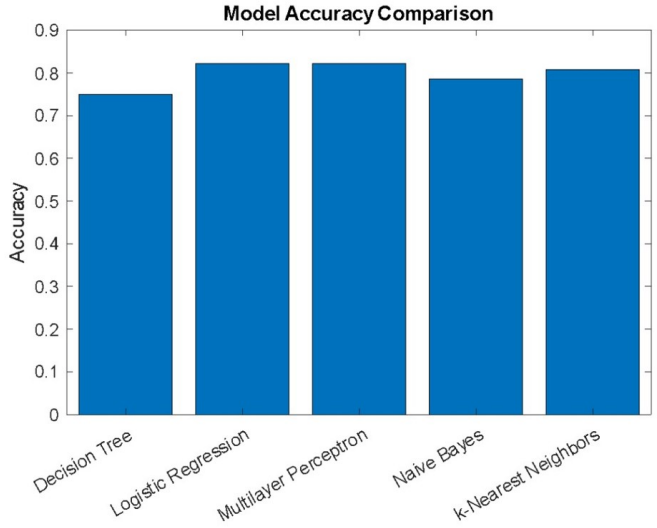

Fig. 2 Algorithm Model Accuracy Visualization

4.3 Discussion

The analysis shows that although the Multilayer Perceptron and Logistic Regression models achieved high accuracy, their performance was affected by dataset skewness, causing them to predominantly predict all fires as class 1. This suggests that these models struggle with imbalanced data, which impacts their ability to effectively differentiate between smaller and larger fires.

In contrast, the Decision Tree and K-Nearest Neighbors models performed better across different size classes, suggesting they may be better suited to handle variability in wildfire sizes. However, the feature importance analysis reveals that the selected features may not fully capture the complexities of wildfire behavior. This points to the potential benefit of incorporating additional or alternative features to improve model performance and robustness.

The overall accuracy provides a general measure of performance, but it is important to note that in datasets with significant class imbalance, such as the wildfire size data analyzed here, this overall accuracy can be a less informative metric. A more comprehensive evaluation involves examining metrics that assess performance per class, such as Precision (accuracy of positive predictions), Recall (ability to find all positive instances), and F1-score (harmonic mean of precision and recall), along with overall measures like the Area Under the Receiver Operating Characteristic curve (AUC-ROC). These metrics would provide a clearer picture of how well each model identifies instances of the less frequent, larger wildfire size classes. While these per-class metrics are not explicitly tabulated here, the confusion matrices provide a foundational visual of per-class performance, clearly showing where misclassifications happen.

4.4 Resampling Limitations

Although resampling techniques such as oversampling (e.g. SMOTE) or undersampling are commonly used to address class imbalance, their application in this study presents challenges. Oversampling the minority class (larger fires) risks overfitting and poor generalization due to learning from potentially unrepresentative synthetic or duplicated data. Undersampling the majority class (smaller fires) may discard valuable information. To address class imbalance without these drawbacks, future work could explore ensemble methods, such as combining predictions from multiple models trained on different subsets of the data or with different weighting schemes.

4.5 Training/Inference time and Efficiency

Beyond predictive accuracy, the practical use of wildfire size classification models in realtime decision-making depends on computational efficiency, including both training and inference times. While not the primary focus of this study, observations during analysis indicated notable differences across the models. Simpler models like Naïve Bayes and Decision Trees generally trained faster, while the Multilayer Perceptron (MLP) required longer training times due to its iterative nature. For deployment in real-time warning systems, future work should optimize model architectures and explore techniques like model compression or distributed training to reduce inference latency and improve efficiency, especially for complex models like the MLP.

4.6 Comparison with Existing Work

While direct quantitative comparison with existing wildfire size classification models is challenging due to variations in datasets, feature sets, and methodologies, it is important to contextualize our findings within the broader literature. Many studies focus on wildfire occurrence or spread prediction, but fewer specifically

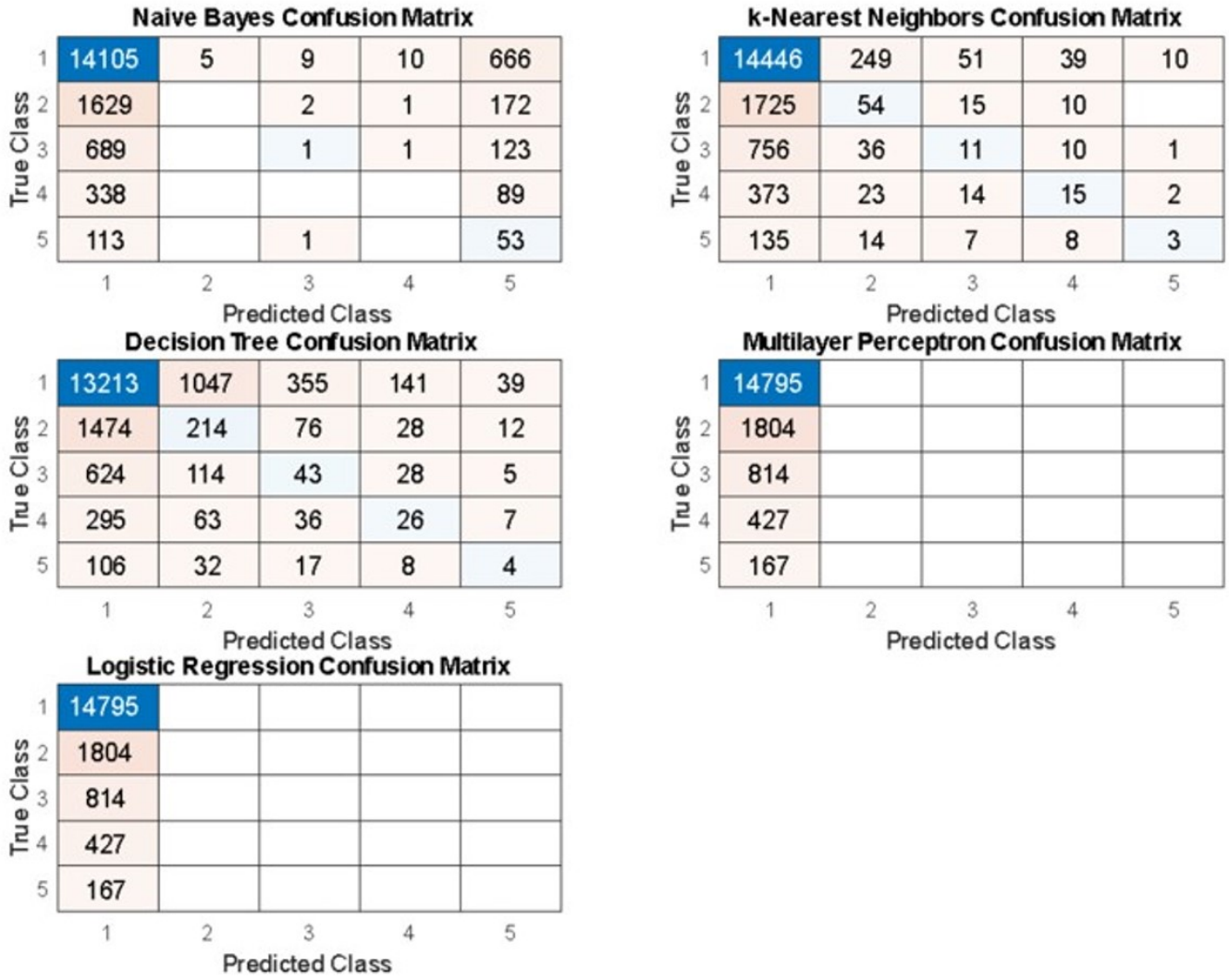

Fig. 3 Confusion Matrix comparing the True vs Predicted Values of Test Data Set by Algorithm

address size classification using traditional machine learning approaches suitable for direct benchmarking.

For instance, (Angelov, 2021)[7] employed a range of machine learning models, including LightGBM, Random Forest, K-Nearest Neighbors, Gaussian Naïve Bayes, Support Vector Machine, Decision Tree, and Multilayer Perceptron, for the binary classification of wildfire size (Class B vs. Classes $>$ B) in the United States. Their study used a comprehensive set of characteristics, including historical weather information (humidity, precipitation, temperature, and wind speed), historical vegetation data, and spatio-temporal information (latitude/longitude, date, remoteness, elevation). Their best-performing model, LightGBM, achieved an accuracy of 69.5% and an F1-score of 71.39% on their full dataset, representing a significant improvement over a 50% accuracy baseline.

My Logistic Regression model, while not directly comparable due to differences in the number of output classes (five in our study vs. two in Angelov's) and the specific feature set (lacking direct vegetation data and humidity in our case), achieved an average accuracy of 82.2% across five folds. The performance of my models, particularly Logistic Regression and Multilayer Perceptron, demonstrates competitive accuracy within the context of similar classification tasks. Angelov's findings also highlighted the importance of meteorological and geospatial features, aligning with my observations on the influence of Temperature, Precipitation, WindSpeed, gridX, and gridY. This reinforces the potential of these models to provide valuable initial assessments even with a more limited feature set.

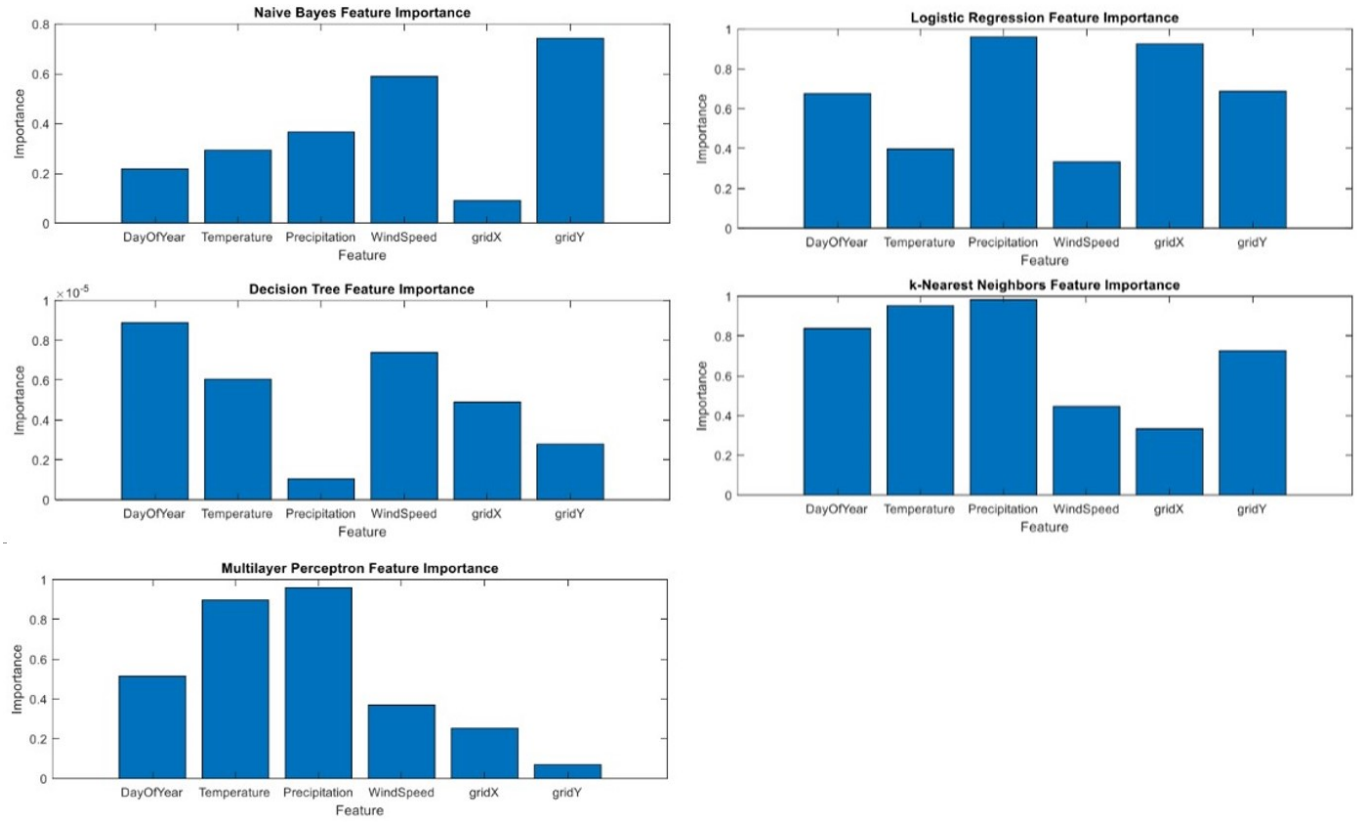

Fig. 4 Feature importance by Algorithm

5 Conclusion

This study successfully investigated the effectiveness of various machine learning algorithms in classifying wildfire size based on spatial and climatic influences. By leveraging historical wildfire records integrated with meteorological and geospatial data, I demonstrated that models such as Logistic Regression and Multilayer Perceptron can achieve high overall accuracy in predicting wildfire size categories. My findings highlight the significant impact of factors like 'Day of Year', 'Temperature', 'Precipitation', 'WindSpeed', and geospatial coordinates ('gridX', 'gridY') on wildfire size, underscoring their important role in wildfire behavior.

Despite the observed high overall accuracy, the inherent class imbalance in wildfire datasets presented a notable challenge, often leading to a bias towards predicting the majority class. This emphasizes the need for evaluation using metrics beyond simple accuracy. The work presented provides a foundational analysis, showcasing the potential of machine learning to inform initial containment strategies and resource allocation for wildfires. By offering an early-stage assessment of potential fire size, these models contribute to enhancing early intervention efforts, reducing suppression costs, and ultimately minimizing environmental and infrastructural damage. This research reinforces the value of data-driven approaches in predicting wildfire size, which can support more precise and effective disaster response.

6 Further Work

Future research could concentrate on several key areas to enhance the robustness, interpretability, and practical applicability of wildfire size classification models.

First, a more comprehensive evaluation framework could be implemented. This could include calculating and reporting per-class metrics such as Precision, Recall, and F1-score to provide detailed insights into the models' ability to classify minority wildfire size classes. Additionally, the AUC-ROC score can be calculated, which informs how well a model can distinguish between different wildfire sizes (e.g., small vs. large, medium vs. large) across all possible decision points. Addressing class imbalance through techniques such as resampling (oversampling, undersampling, SMOTE) could also be explored to improve the accuracy and fairness of these models. Statistical significance testing (e.g., reporting confidence intervals for accuracy and

other metrics) can be conducted to provide a more robust and statistically sound evaluation of model performance.

Second, the feature set can be expanded and analyzed. This would involve incorporating additional environmental variables such as relative humidity, various vegetation types, aridity indices, and tree cover, which are known to influence wildfire behavior and could significantly enhance model accuracy and robustness. Furthermore, a thorough feature importance analysis using more interpretable, model-agnostic methods like SHAP (SHapley Additive exPlanations) values or permutation importance can be conducted. This would provide a deeper understanding of how each feature quantitatively and qualitatively influences wildfire size classification.

Third, the generalizability and dynamic behavior of wildfires warrants further focus. The inherent spatial correlation in wildfire occurrences suggests the exploration of advanced spatial machine learning techniques, such as Convolutional Neural Networks (CNNs) for gridded data or Gaussian Process Regression, to better capture these complex dependencies. Investigating temporal trends, including the impact of climate change on wildfire size, is also crucial; therefore, incorporating sequential models like Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) models may help capture these temporal dynamics more effectively and assess model robustness in projected future climate scenarios. A key step would be to evaluate these models across diverse geographic regions beyond California to confirm their generalizability and practical utility under different environmental conditions.

Finally, exploring advanced machine learning techniques, such as transfer learning, ensemble models, and attention-based models, in conjunction with hyperparameter optimization using advanced techniques like Bayesian Optimization, could refine predictions and improve a model's performance in complex wildfire scenarios. Discussions on the training time, inference time, and overall model efficiency, particularly for more complex architectures like neural networks, are important and could be provided, as these factors are paramount for real-time deployment and decision-making in operational wildfire management systems.

Acknowledgments

Thank you for the guidance of Dr. Weiwei Sun, Department of Computer Science and Technology, Cambridge University, in the development of this research paper.

References

1 K. Short, *Spatial wildfire occurrence data for the united states, 1992–2020.*

2 Meteostat, *Meteostat weather data*, `https://meteostat.net/en/`., Retrieved August 1, 2024, from.

3 A. Westerling and B. Bryant, *Climate change and wildfire in and around California: Fire modeling and loss modeling*.

4 T. Keeping, S. Harrison and I. Prentice, *Modelling the daily probability of wildfire occurrence in the contiguous united states*.

5 D. Shadrin, S. Illarionova, F. Gubanov, K. Evteeva, M. Mironenko, I. Levchunets, R. Belousov and E. Burnaev, *Wildfire spreading prediction using multimodal data and deep neural network approach*.

6 A. Westerling, B. Bryant, H. Preisler, T. Holmes, H. Hidalgo, T. Das and S. Shrestha, *Climate change and growth scenarios for california wildfire*.

7 D. Angelov, *Machine learning for binary classification of wildfire size*.