The 3rd International Seminar of Science and Technology
ISST 2023 Vol 3 (2024) 03 003
Trends in Science and Technology for Sustainable Living
Faculty of Science and Technology
Universitas Terbuka

Enhancing wildfire risk assessment through weather modeling and machine learning: a case study in Orange County, USA

Lulus Kahono*

Universitas Terbuka, Statistics Department, South Tangerang, Banten, Indonesia, 15437

Abstract - Forest wildfire is a significant threat, causing extensive environmental damage and loss of life. Accurate and effective risk assessments are essential for better understanding and mitigating this risk. In this study, we propose a novel approach that combines weather modeling and machine learning for improved wildfire risk assessment. Our case study focuses on Orange County, USA, known for its high wildfire susceptibility. The methodology involves collecting historical weather and wildfire data in the region. We employ a machine learning model that integrates weather data with other factors, including vegetation and topography, to predict wildfire risk. Through machine learning techniques, we analyze the relationships among these variables and generate accurate risk predictions for Orange County. Results indicate a significant enhancement in wildfire risk assessment with this approach. The developed model provides valuable insights into the factors influencing wildfires, aiding authorities, and policymakers in implementing effective mitigation strategies. By leveraging weather modeling and machine learning, we can enhance our understanding and management of wildfire risks, safeguarding the environment and communities. This research highlights the potential of these technologies to improve wildfire mitigation efforts. The findings contribute to a proactive approach to addressing the serious threat of forest wildfires.

Keywords: machine learning, weather modeling, wildfire

1 Introduction

California has a subtropical Mediterranean climate, which means the region experiences long and dry summers as well as cool and wet winters. Typically, there are four seasons in California: spring, summer, autumn, and winter. Wildfires in California mostly occur during the summer season, between May and October [1].

We chose Orange County as our research site because of its substantial representation concerning climate, topography, and the junction between urban and wild areas. Located in Southern California, Orange County displays a varied topographical range encompassing coastlines and mountains. This ecological diversity, combined with its Mediterranean climatic conditions, renders it prone to both spontaneous and anthropogenic wildfires. Our objective was to comprehensively decipher the patterns and shifts in wildfire incidents within this dynamic setting by scrutinizing fire events spanning from 2013 to 2022.

*Corresponding author: 044085171@ecampus.ut.ac.id

Our investigation aimed to not only assess the prevalent fire seasonality between May and October but also to discern the potential for wildfires in other months, see Fig.1. While the majority of incidents coincide with the traditional fire season, our findings underscore the importance of understanding the underlying factors that contribute to the risk of wildfires throughout the entire year. This broader perspective allows for more effective fire management strategies and disaster preparedness efforts.

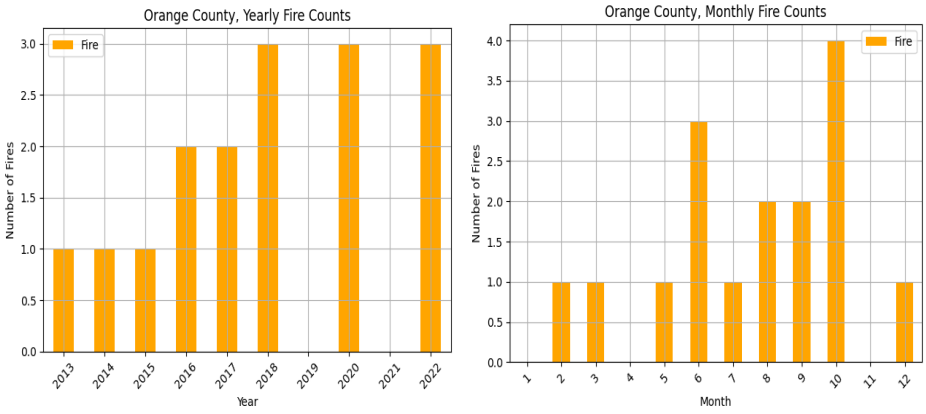

Fig. 1. Orange County, Fire Counts (2013 – 2022)

In the realm of fire ignition, a triad of elements converges to facilitate the captivating phenomenon [2]. Initially, the fuel source provides combustible material, offering energy-rich molecules like hydrocarbons. Second, an oxidizing agent, typically oxygen from the air, sustains the combustion process by fostering chemical reactions. Lastly, the ignition factor, often heat or a spark, triggers the energy exchange. These elements harmonize, resulting in a self-sustaining exothermic reaction, releasing heat, light, and various byproducts. Understanding this triadic interplay unlocks insights into fire management and safety, enriching our grasp of combustion dynamics.

Through this study, we sought to provide valuable insights into wildfire behavior and its temporal variations, contributing to the development of informed strategies for fire prevention, response, and mitigation in Orange County and similar regions facing comparable challenges. By comprehensively evaluating the historical wildfire data, our research contributes to enhancing the resilience of communities against the ever-present threat of wildfires.

2 Materials and methods

The integration of Python and R via the Rpy2 package represents a potent strategy to harness the distinctive attributes of both programming languages within a unified project. Rpy2 facilitates the reciprocal invocation of functions, enabling the seamless amalgamation of R's specialized libraries and Python's broader ecosystem [3]. The computational capabilities of the R library are harnessed to streamline and facilitate the intricacies associated with the numerical computations at hand.

We employ a hybrid approach by leveraging both fire and weather data to gain insights into the prevalent wildfire patterns within Orange County. Our methodology entails the utilization of a Machine Learning Framework that encompasses various stages, namely Data Collection, Data Preprocessing, Feature Selection, Model Selection, and Train Test Validation, as seen on Fig.2. This framework serves as the backbone of our efforts to categorize and analyze the intricate relationship between weather conditions and fire occurrences.

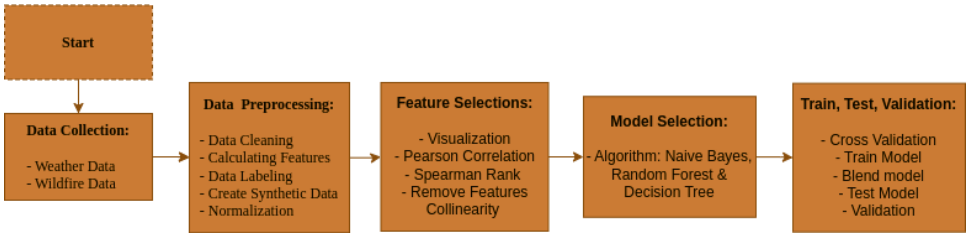

Fig. 2. Machine Learning Frameworks

2.1 Data Collection

By means of a thoroughgoing process of data compilation, we systematically accumulate a varied spectrum of information pertaining to fires and meteorological conditions encompassing the temporal span from 2013 to 2022. This meticulous undertaking encompasses the collection of multifaceted data concerning fire occurrences as well as climatic phenomena, facilitating a comprehensive exploration of the interrelationships between these factors during the specified period. The acquired dataset thus stands as a valuable resource for empirical analyses aimed at discerning patterns, trends, and potential causal connections within the context of fire dynamics and atmospheric influences.

2.1.1 Weather data

VisualCrossing.com serves as the predominant data repository for meteorological information [4]. It furnishes comprehensive weather data, as delineated in Table 1.

Table 1. Features in Weather Data.

No	Element	Description	US	Metric	UK
1	tempmax	Maximum Temperature	F	C	C
2	tempmin	Minimum Temperature	F	C	C
3	temp	Temperature (or mean temperature)	F	C	C
4	dew	Dew Point	F	C	C
5	feelslike	Feels like	F	C	C
6	precip	Precipitation	inches	mm	mm
7	precipprob	Precipitation chance	%	%	%
8	precipcover	Precipitation Cover	%	%	%
9	preciptype	Precipitation type	–	–	–
10	snow	Snow	inches	cm	cm
11	snowdepth	Snow Depth	inches	cm	cm
12	windspeed	Wind Speed	mph	kph	kph
13	windgust	Wind Gust	mph	kph	kph
14	winddir	Wind Direction	degrees	degrees	degrees
15	visibility	Visibility	miles	km	km
16	cloudcover	Cloud Cover	%	%	%
17	humidity	Relative Humidity	%	%	%
18	pressure	Sea Level Pressure	mb	mb	mb
19	solarradiation	Solar Radiation	W/m2	W/m2	W/m2
20	solarenergy	Solar Energy	MJ/m2	MJ/m2	MJ/m2
21	uvindex	UV Index	–	–	–
22	severerisk	Severe Risk	–	–	–

No	Element	Description	US	Metric	UK
23	sunrise	Sunrise time	–	–	–
24	sunset	Sunset time	–	–	–
25	moonphase	Moonphase	–	–	–
26	icon	A weather icon	–	–	–
27	conditions	Short text about the weather	–	–	–
28	description	Description of the weather for the day	–	–	–
29	stations	List of weather station sources	–	–	–

The accrued advantages extend to researchers, enterprises, and individuals alike, facilitating astute decision-making through the furnished dataset. VisualCrossing.com markedly amplifies the depth of insights within the purview of meteorological contemplation and also provides access to weather prediction information using 50 years historical data.

2.1.2 Wildfire data

Wildfire-related data is obtained from CAL FIRE, a reliable and authoritative source for wildfire-related information [5]. The information contains the incident name, incident date, incident administrative unit, incident location, incident acres burned, and more.

2.2 Data Preprocessing

Data preprocessing involves cleaning, transforming, and preparing data for model training. It aims to enhance data quality and usability, ensuring accurate predictions. Techniques like data cleaning, transformation, feature engineering, handling imbalanced data, normalization, and handling outliers are employed. By selecting relevant features, addressing data inconsistencies, and dividing data into training and test sets, preprocessing improves model accuracy, convergence, and generalization. A well-preprocessed dataset enables algorithms to extract meaningful patterns, leading to more effective and robust machine-learning models for real-world applications.

2.2.1 Data Cleaning

Data cleaning encompasses the identification and correction of errors and inconsistencies in the dataset. Its goal is to enhance data quality, ensuring reliable and unbiased model training. Tasks include handling missing data through removal or imputation, removing duplicates, detecting, and treating outliers, and formatting data correctly. By conducting thorough data cleaning, machine learning models can learn effectively from the data and make accurate predictions on new, unseen instances.

2.2.2 Calculating Features

Within this phase, an assemblage of weather indices is compiled through diligent sourcing from diverse references. The calculation of these indices is accomplished via both mathematical calculation and the utilization of the firebehavioR Library in the R programming environment [6]. The roster of indices encompassed herein comprises the subsequent elements, as described in Table 2.

Table 2. Features Description.

No	Features	Description
1	Heat Index	The heat index, also known as the apparent temperature, is what the temperature feels like to the human body when relative humidity is combined with the air temperature. This has important considerations for the human body's comfort [7], [8].
2	VPD (Vapor Pressure Deficit)	VPD stands for Vapor Pressure Deficit. It is a measure of the difference between the amount of moisture in the air and the maximum amount of moisture that the air can hold at a given temperature. In simpler terms, VPD is a way to measure how dry the air is. The paper proposes a method for controlling the environment of a greenhouse using VPD to optimize plant growth [9].
3	KBDI (Keetch Byram Drought Index)	This index is cumulative and requires daily temperature and daily and annual precipitation as input data. It is aimed to reflect the dryness - and thus the flammability - of organic material in the ground by considering the effects of rainfall and evapotranspiration on moisture deficiency in deep duff and upper soil layers [10], [11].
4	Fosberg KBDI	The Fosberg Fire Weather Index is a fire danger index developed by Fosberg (1978). It is based upon the equilibrium moisture content and wind speed and requires hourly observations of temperature, relative air humidity, and wind speed as input data [12]. This recent study suggested a modification to the Fosberg fire weather index that includes the impact of precipitation is proposed, which improves the utility of the index as it offers an improved relationship between the index and area burned [13].
5	Forest Mark 5	The Mark 5 forest fire danger index was developed to assess fire danger and behavior in eucalypt forest fuel types and has been widely used in Eastern Australia. The FFDI requires temperature, relative humidity, wind speed, and a fuel availability index (i.e., a drought factor) measured at 15:00 as input variables [14].
6	Fuel Moisture KBDI	The index is based on the principle that fire danger is determined by the combined influence of wind speed and fuel moisture content. It increases when wind speed increases and decreases when fuel moisture content increases [15].
7	Drought Factor	The function of the drought factor utilizes daily weather data (temperature, rainfall, relative humidity, and wind speed) and calculates a humidity index that represents atmospheric drying potential. This humidity index is then used to estimate the fuel moisture content of surface fuels based on an empirical model that relates fuel moisture content to the humidity index [16].
8	Nesterov	It is a simple daily fire danger rating index requiring daily air temperature, dew point temperature, and precipitation as input data. This index is cumulative and resets to zero when daily precipitation exceeds 3 [mm]. The dryness of the fuel is estimated assuming that the drying rate is related to vapor saturation and therefore takes dew point temperature into account. It is therefore particularly appropriate for capturing fine fuel moisture, and thus predicting fire ignition, but is less appropriate for predicting fire behaviour or spread [17]. The Nesterov index is merged with the TVDI (temperature/vegetation dryness index) to create a map of fire risk, which can be used to identify areas prone to fire outbreaks as suggested by this recent study [18].
9	Nesterov Mod	A modified version of the Nesterov index, which includes wind velocity and daily values of rainfall, is in use in Portugal [17].
10	Zdenko	Calculation of the index value based on temperature, humidity level, and rainfall [6].

2.2.3 Data Labeling

Data labeling involves assigning relevant tags or labels to data points for supervised training. It relies on manual labeling by human annotators, semi-supervised and active learning methods, or crowdsourcing.

2.2.4 Create Synthetics Data

Synthetic data in machine learning is artificially generated data resembling real-world observations, created through algorithms or statistical methods rather than direct sourcing [19], [20]. Used when obtaining sufficient diverse real data is challenging, it aids dataset augmentation, class distribution balance, and sensitive information protection during model development. Successful synthetic data relies on faithful replication of the original data's traits. Recent article explores the use of synthetic data in conjunction with real-world data to improve the training efficiency of CNNs for weather classification.

However, caution is needed to avert bias introduction or model compromise. Gaussian copula, a statistical model, is used to generate multivariate data adhering to predefined distributions and correlations [21]. Its application enhances synthetic data's authenticity, mitigates biases, and maintains integrity in practical machine learning.

The Synthetic Data Vault (SDV) package in Python is an advanced software module designed for generating synthetic data sets with enhanced privacy protection. Leveraging sophisticated algorithms facilitates data augmentation while preserving sensitive information, fostering robust research practices.

As visually depicted in Fig.3, 16 instances of wildfire occurrences are recorded within the temporal span from 2013 to 2022, engendering a pronounced state of data imbalance within the dataset. To rectify this inherent imbalance, synthetic data generation is being undertaken.

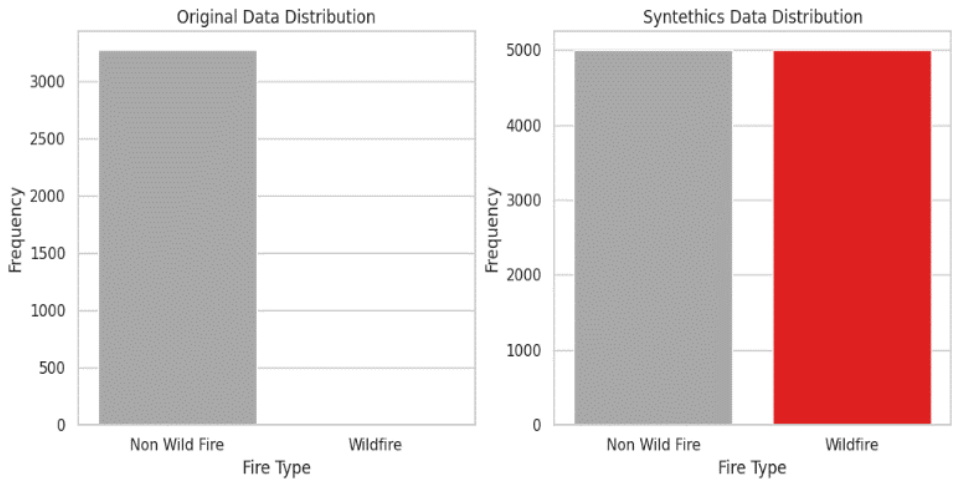

Fig. 3. Synthetics Data Distribution Plot

2.2.5 Normalization

Normalization, an integral preprocessing approach in the realm of machine learning, serves to meticulously rescale and standardize numerical attributes present within a dataset. This transformative process yields a cohesive range, effectively circumventing the undue influence exerted by specific features due to their pronounced magnitudes, thus facilitating equitable learning.

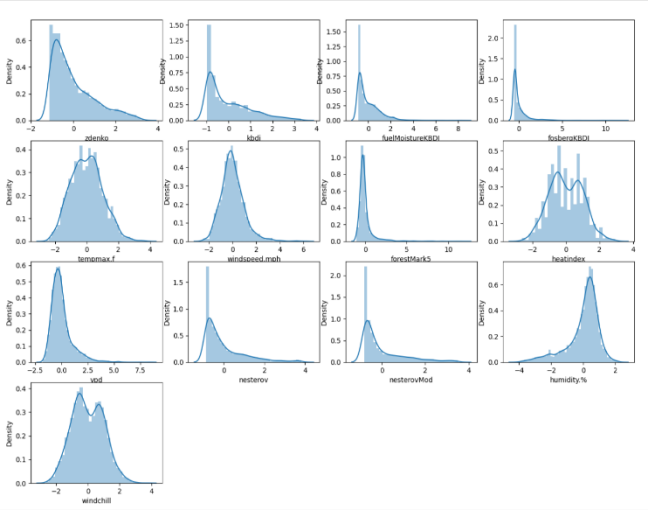

Fig. 4. Z-score Normalization Plot.

Z-score normalization as seen in Fig.4, whereby data undergoes a recalibration, rendering it to possess a mean of 0 and a standard deviation of 1, thus establishing a principled framework for enhanced comparability and model convergence. This procedure contributes substantively to the mitigation of bias and the robustness of analytical outcomes.

2.3 Feature Selection

Feature selection addresses the "curse of dimensionality," a common challenge in machine learning where an excessive number of features can lead to overfitting, increased computational burden, and reduced generalization ability [22].

By selecting a subset of features based on their relevance to the target variable, feature selection enhances the model's ability to extract meaningful patterns and relationships from the data. Correlation is the most important thing related to feature selections, from correlation, we know which features are important to build the classification model, as seen in Fig. 5.

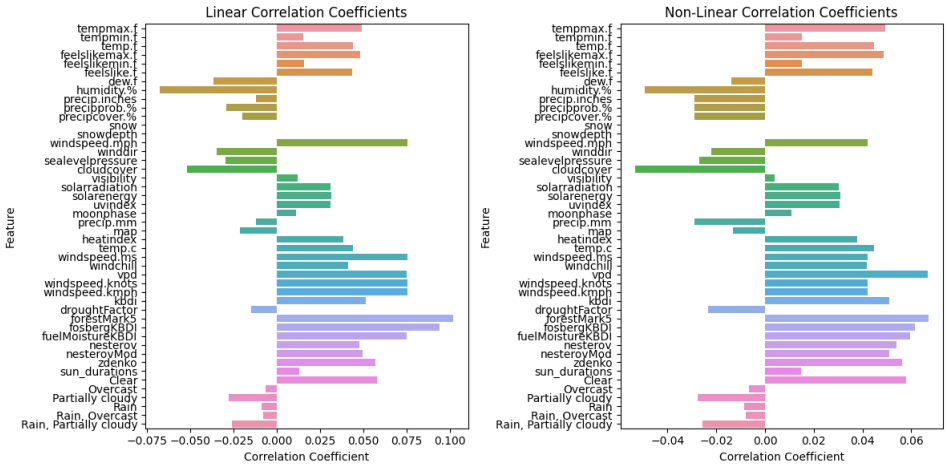

Fig. 5. Linear and Non-Linear Correlation Coefficient

Univariate feature selection works by selecting the best features based on univariate statistical tests. It can be seen as a preprocessing step for an estimator. Scikit-learn exposes feature selection routines as objects that implement the transform method, SelectKBest removes all but the k highest-scoring features. From this process, the most important features that are generated from the method is forestMark 5 as also seen in Fig.5. We only use 1 feature (forestMark5), as the variable to build the fire weather classification model for Orange County.

2.4 Train, Test, Validation

The training set is used for model learning, adjusting internal parameters to minimize prediction errors. The validation set assesses the model's performance on unseen data, measuring its generalization ability. Meanwhile, the test set helps fine-tune hyperparameters and prevents overfitting by evaluating the model on newly labeled data during development. Finely dividing the data and choosing suitable evaluation metrics are essential for building reliable and effective machine learning models.

Within this conceptual framework, the data for the year 2022 was deliberately employed as unobserved data. Subsequently, the synthetically generated dataset was partitioned into distinct training and testing subsets, the former constituting 70% and the latter 30%, thereby adhering to a prescribed ratio.

2.4.1 Cross Validation

Stratified K-Fold Cross-Validation: Stratified K-Fold Cross-Validation is similar to K-Fold, but it ensures that each fold maintains the same class distribution as the original dataset. This is particularly useful when dealing with imbalanced datasets where one class is significantly more prevalent than the others.

2.4.2 Model Selection

Model Selection involves choosing the appropriate machine learning algorithm to effectively capture the complex patterns within the data. The chosen model undergoes rigorous Train Test Validation to fine-tune its performance and ensure its ability to accurately classify weather patterns associated with fire incidents.

Within the scope of this research, the classification framework has been constructed, employing algorithms of distinct methodological underpinnings. Specifically, the selected algorithms include Naive Bayes, esteemed for its probabilistic foundation, alongside the ensemble-based Random Forest, and the individually decisive Decision Tree, each chosen for their unique attributes and relevance to the study's objectives [23–27].

The present scholarly article [28], introduces an innovative recommendation algorithm rooted in blended learning principles. This approach entails the adept utilization of a multi-model ensemble strategy, amalgamating anticipated ratings from diverse foundational models.

Empirical findings derived from rigorous experimentation substantiate the newfound blending methodology's supremacy over alternative techniques delineated within the study.

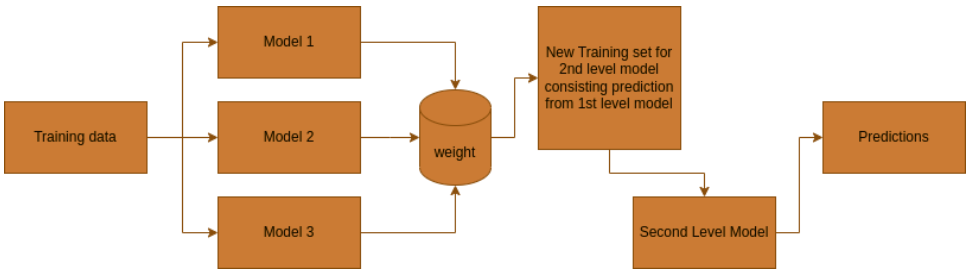

Fig. 6. Blend Model.

3 Results and discussion

3.1 Confusion Matrix

The confusion matrix, a pivotal tool in classification evaluation, yields insightful results. It assesses model performance through four essential metrics: true positives (correctly identified positives), true negatives (correctly identified negatives), false positives (incorrectly labeled as positives), and false negatives (incorrectly labeled as negatives).

These metrics facilitate the calculation of performance indices like accuracy, precision, recall, and F1-score. A well-balanced matrix reflects a model's ability to distinguish between classes accurately, while imbalanced figures highlight specific strengths or weaknesses. By interpreting the confusion matrix, practitioners can refine models, rectify misclassifications, and optimize decision-making processes.

3.1.1 3.1.1 Naïve Bayes

The confusion matrix plot shows true positives (correctly identified positives) = 3, true negatives (correctly identified negatives) = 0, false positives (incorrectly labeled as positives) = 44, and false negatives (incorrectly labeled as negatives) = 318.

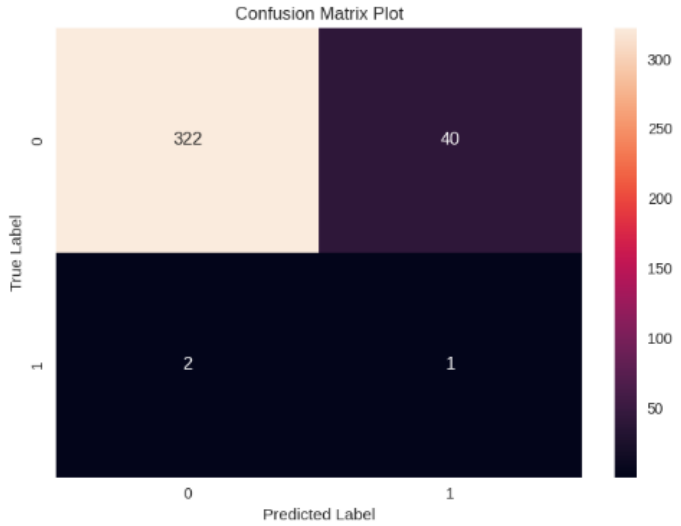

Fig. 7. Naïve Bayes Confusion Matrix Plot

The classification performance analysis based on the confusion matrix indicates significant challenges. With 3 true positives and 0 true negatives, the model correctly identified both positives

and negatives. The high numbers of false positives (44) and false negatives (318) underline the need for improvement in accuracy and precision.

3.1.2 Random Forrest

The confusion matrix plot shows true positives (correctly identified positives) = 3, true negatives (correctly identified negatives) = 0, false positives (incorrectly labeled as positives) = 74, and false negatives (incorrectly labeled as negatives) = 288.

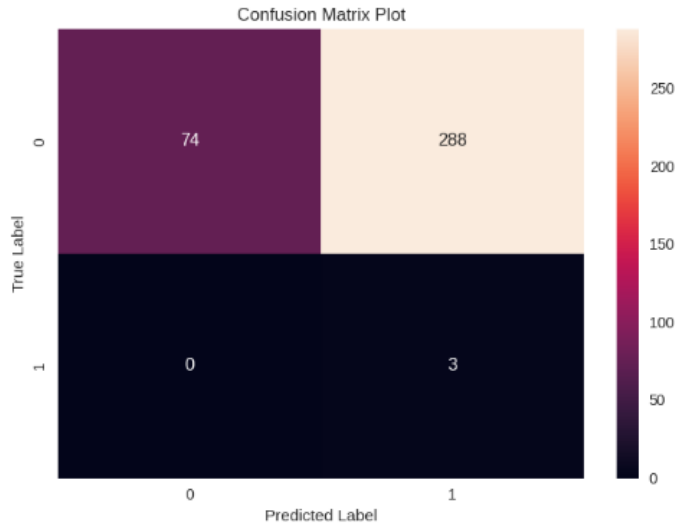

Fig. 8. Random Forrest Confusion Matrix Plot

The classification performance analysis based on the confusion matrix reveals concerning outcomes. With 3 true positives and 0 true negatives, the model correctly identified both positive and negative instances. The substantial count of 74 false positives and 288 false negatives underscores the need for urgent enhancement to achieve accuracy and precision.

3.1.3 Decision Tree

The confusion matrix plot shows true positives (correctly identified positives) = 2, true negatives (correctly identified negatives) = 1, false positives (incorrectly labeled as positives) = 254, and false negatives (incorrectly labeled as negatives) = 108.

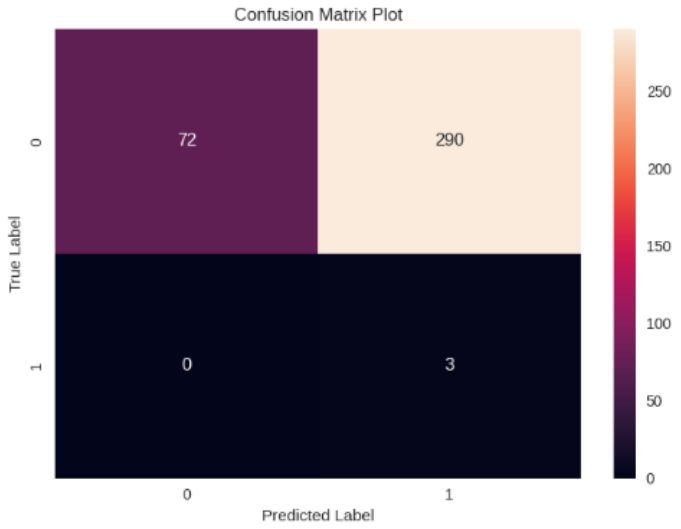

Fig. 9. Decision Tree Confusion Matrix Plot

The classification performance analysis based on the confusion matrix reveals concerning outcomes. With 2 true positives and 1 true negative, the model does not correctly identify both positive and negative instances. The substantial count of 254 false positives and 108 false negatives underscores the need for urgent enhancement to achieve accuracy and precision.

3.1.4 Blend Models

The blend model exhibits a noteworthy enhancement in the performance of the classification model. The confusion matrix plot shows true positives (correctly identified positives) = 3, true negatives (correctly identified negatives) = 0, false positives (incorrectly labeled as positives) = 235, and false negatives (incorrectly labeled as negatives) = 127 as seen in Fig.10.

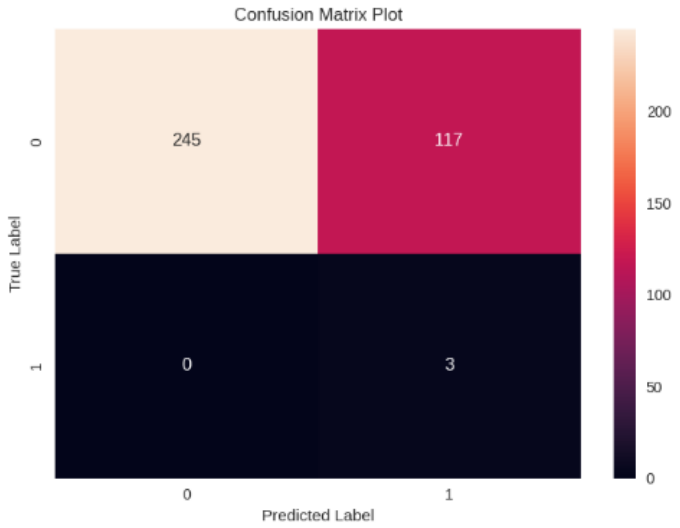

Fig. 10. Blend Models Confusion Matrix Plot

3.2 Calendar Plot

3.2.1 Wildfire in 2022

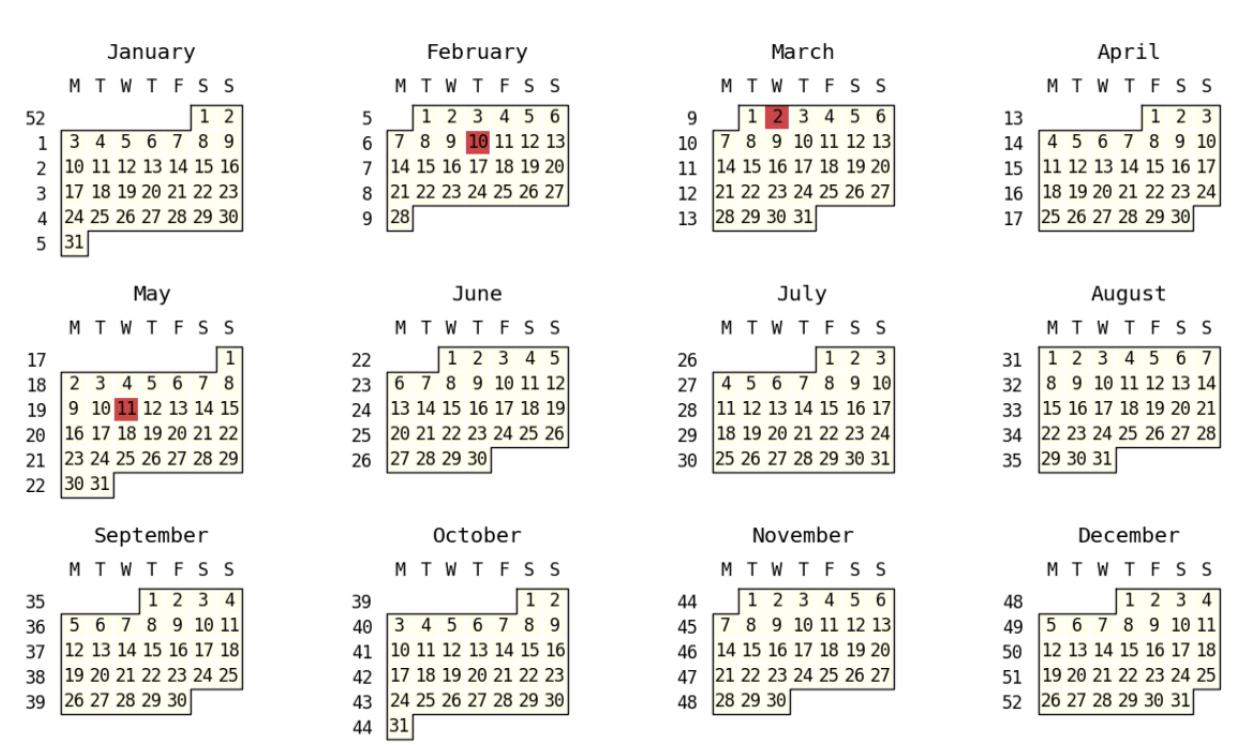

Fig. 11. Wildfire Events in 2022

Fig.11 visually portrays three distinct wildfire incidents transpiring within Orange County during the year 2022. Observationally, none of these occurrences unfolded between May to October, thereby delineating a notable absence of wildfire events during this specific period.

3.2.2 How Forest Mark 5 classified the weather conditions in 2022

Illustrated in Fig 12. Is the classification of wildfire weather using calculation of Forest Mark 5 index. Categories or fire danger class represented by colors, where yellow equal to low, brown equal to moderate, pink equal to high, and red equal to very high. Forest Mark 5 classified wildfire weather into 5 categories, as seen on Table 3.

Table 3. Forest Mark 5 fire danger class.

FFDI Range	Fire Danger Class
0 – 5	Low
5 – 12	Moderate
12 – 25	High
25 – 50	Very high
> 50	Extreme

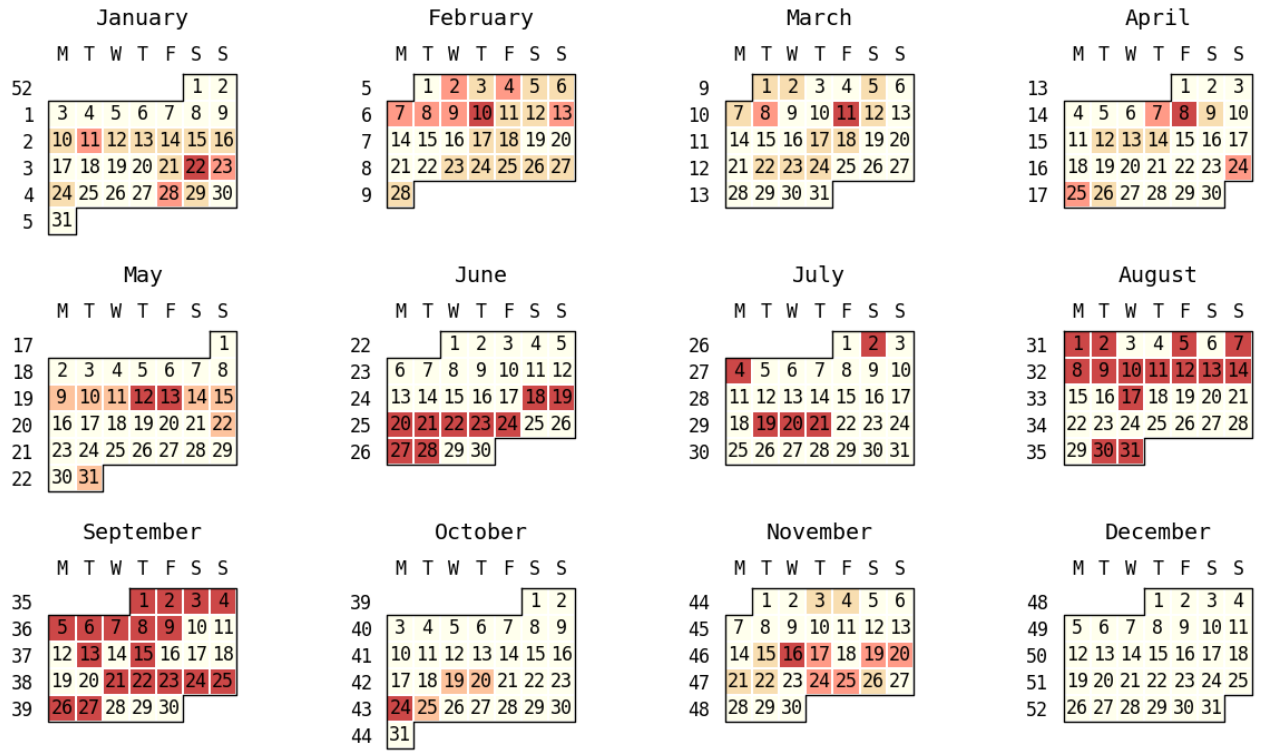

Fig. 12. Forest Mark 5 fire danger class in 2022

Observation of the result is undertaken to find connection and relationship of the fire danger class and wildfire events for the specific region. From the results we clearly see that wildfire events commonly happened in high and very high fire danger class, even though it is still difficult to precisely predict when the wildfire will happen in the future if we use these fire danger classifications for predictions due to the massive output.

3.2.3 How the model classified the weather conditions in 2022

llustrated in Fig.13 is the classification of wildfire weather by the ensemble models. In conjunction with the wildfire events occurring on the 10th of February, 2nd of March, and 11th of May, as presented in Fig.11, a consistent pattern emerges. Wildfires consistently manifest over durations of seven or more consecutive days marked by red alert.

The observations furnish the foundation for our hypothesis, which posits the temporal connection between wildfire occurrences and prolonged episodes of heightened risk signified by red alerts. These empirical insights serve as pivotal parameters for model selection, optimizing congruence with the empirical data and geographical context.

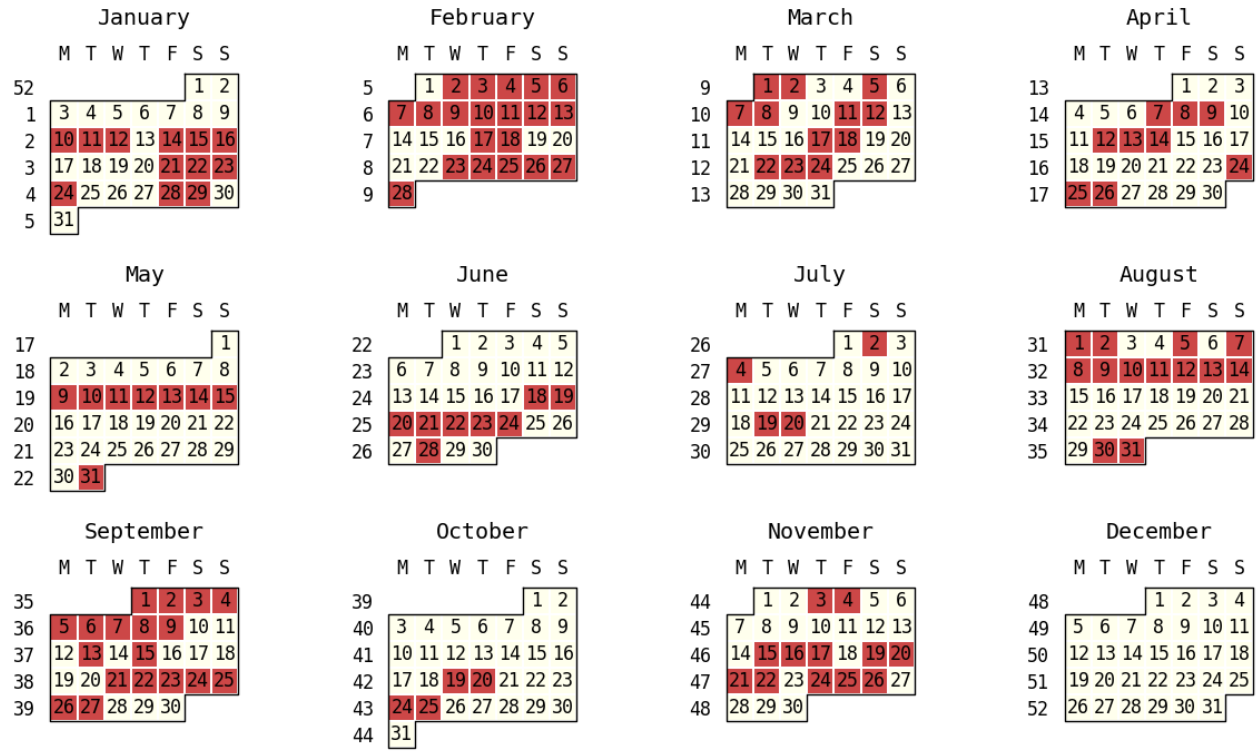

Fig. 13. Calibrated forest mark 5 index and historical data

4 Conclusion

Not all fire incidents occur during the period from May to October, thus explaining that fire events may indeed transpire outside of this specific period. There is a temporal correlation between the occurrence of wildfires and extended periods characterized by heightened risk, as denoted by the red alerts. These empirical insights serve as pivotal determinants in the process of model selection, allowing for the optimization of congruence between the chosen model and both the empirical data and the geographical context. Utilizing the present classification framework, it becomes feasible to recalibrate the extent of the fire weather categorization outcome through the amalgamation of the Forest Mark 5 index and retrospective data derived from the specific region.

References

[1] "California Fire Season: In-Depth Guide," wfca.com, Jul. 05, 2022. https://wfca.com/articles/california-fire-season-in-depth-guide/

[2] "Fire triangle," wikipedia.org, Jul. 29, 2023. https://en.wikipedia.org/w/index.php?title=Fire_triangle&oldid=1167654742

[3] "rpy2: Python interface to the R language (embedded R)." https://rpy2.github.io

[4] Visual Crossing, "Weather Data Documentation." https://visualcrossing.com/resources/documentation/ (accessed Feb. 19, 2023).

[5] "Incidents | CAL FIRE." https://www.fire.ca.gov/incidents

[6] J. P. Ziegler, C. M. Hoffman, and W. Mell, "firebehavioR: An R Package for Fire Behavior and Danger Analysis," Fire, vol. 2, no. 3, p. 41, Jul. 2019, doi: 10.3390/fire2030041.

[7] "What is the heat index?" weather.gov. https://www.weather.gov/ama/heatindex

[8] C. Vitolo, C. Di Napoli, F. Di Giuseppe, H. L. Cloke, and F. Pappenberger, "Mapping combined wildfire and heat stress hazards to improve evidence-based decision making," Environ Int, vol. 127, pp. 21–34, Jun. 2019, doi: 10.1016/j.envint.2019.03.008.

[9] A. P. Williams et al., "Observed Impacts of Anthropogenic Climate Change on Wildfire in California," Earths Future, vol. 7, no. 8, pp. 892–910, Aug. 2019, doi: 10.1029/2019EF001210.

[10] "Fire Weather Indices Wiki | Keetch-Byram drought index." https://wikifire.wsl.ch/tiki-indexa61f.html?page=Keetch-Byram+drought+index&structure=Fire

[11] K. Dolling, P.-S. Chu, and F. Fujioka, "A climatological study of the Keetch/Byram drought index and fire activity in the Hawaiian Islands," Agric for Meteorol, vol. 133, no. 1–4, pp. 17–27, Nov. 2005, doi: 10.1016/j.agrformet.2005.07.016.

[12] "Fire Weather Indices Wiki | Fosberg fire weather index." https://wikifire.wsl.ch/tiki-indexb1d5.html?page=Fosberg+fire+weather+index&structure=Fire

[13] S. L. Goodrick, "Modification of the Fosberg fire weather index to include drought," Int J Wildland Fire, vol. 11, no. 4, p. 205, 2002, doi: 10.1071/WF02005.

[14] "Fire Weather Indices Wiki | McArthur Mark 5 forest fire danger index." [Online]. Available: https://wikifire.wsl.ch/tiki-index4fe9.html?page=McArthur+Mark+5+forest+fire+danger+index&structure=Fire

[15] "Fire Weather Indices Wiki | Sharples fuel moisture and fire danger rating indices." https://wikifire.wsl.ch/tiki-index91e2.html?page=Sharples+fuel+moisture+and+fire+danger+rating+indices&structure=Fire

[16] K. L. R. Y. L. C. H. L. Jeremy S. Littell David L. Peterson, "A review of the relationships between drought and forest fire in the United States," vol. 22, no. 7. Glob Chang Biol. doi: 10.1111/GCB.13275.

[17] "Fire Weather Indices Wiki | Nesterov ignition index." https://wikifire.wsl.ch/tiki-indexfa8e.html?page=Nesterov+ignition+index&structure=Fire

[18] M. Onderka and I. Melicherčík, "Spatial Discretization of the Nesterov Fire Rating Index using Multispectral Satellite Imagery," Aug. 2023.

[19] S. Chatterjee and Y.-C. Byun, "A Synthetic Data Generation Technique for Enhancement of Prediction Accuracy of Electric Vehicles Demand," Sensors, vol. 23, no. 2, p. 594, Jan. 2023, doi: 10.3390/s23020594.

[20] R. P. Supervisor, K. Veeramachaneni, N. Boumal, and D. Liu, "Synthetic data assessment based on model improvement."

[21] "GaussianCopula Model — SDV 0.18.0 documentation." [Online]. Available: https://sdv.dev/SDV/user_guides/single_table/gaussian_copula.html

[22] "Article title - 1.13. Feature selection." https://scikit-learn/stable/modules/feature_selection.html

[23] F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," International Journal of Information Technology, vol. 13, no. 4, pp. 1503–1511, Aug. 2021, doi: 10.1007/s41870-020-00430-y.

[24] A. Gupta, M. C. Lohani, and M. Manchanda, "Financial fraud detection using naive bayes algorithm in highly imbalance data set," Journal of Discrete Mathematical Sciences and Cryptography, vol. 24, no. 5, pp. 1559–1572, Jul. 2021, doi: 10.1080/09720529.2021.1969733.

[25] P. J. Pereira, A. Pereira, P. Cortez, and A. Pilastri, "A Comparison of Machine Learning Methods for Extremely Unbalanced Industrial Quality Data," 2021, pp. 561–572. doi: 10.1007/978-3-030-86230-5_44.

[26] E. B. B. Palad, M. J. F. Burden, C. R. Dela Torre, and R. B. C. Uy, "Performance evaluation of decision tree classification algorithms using fraud datasets," Bulletin of Electrical Engineering and Informatics, vol. 9, no. 6, pp. 2518–2525, Dec. 2020, doi: 10.11591/eei. v9i6.2630.

[27] D. Shaohui, G. Qiu, H. Mai, and H. Yu, "Customer Transaction Fraud Detection Using Random Forest," in 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), IEEE, Jan. 2021, pp. 144–147. doi: 10.1109/ICCECE51280.2021.9342259.

[28] F. A. Malik, W. Ye, Q. Chen, and D. Li, "Recommendation Algorithm based on Blending Learning," in Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference, New York, NY, USA: ACM, Jun. 2019, pp. 113–117. doi: 10.1145/3341069.3342983.