BLAZE PREDICT

Predictive Modelling of Fire Sizes Using Climate Indicators

Rohan Jovi Thevara 22209135
Alvin Sebastian 22201493

01 INTRODUCTION

Wildfires pose as a formidable menace to both the environment and human well being. Despite the presence of Wildfire prediction and control centeres such as European Forest Fire Information System, NICC, and others, the global landscape continues to witness the occurrence of extensive amount of wildfires. These catastrophic events wreak havoc on ecosystems, causing irreparable damage, and tragically, leading to the loss of numerous lives.

Wildfires also tend to generate substantial volumes of finer particulates recognized as PM2.5, along with even smaller nanoparticles, both of which are well known for their detrimental effects on human health

.

The smoke emerging from forest and peat combustion can persist in the atmosphere for extended periods, spanning weeks, and traverse great distances of thousands of kilometers. This phenomenon adversely affects the well-being of populations residing in regions far removed from the actual wildfire sites.

02 OBJECTIVES

The core purpose of this study is to construct a robust machine learning model with the capacity to effectively predict both the likelihood and potential scale of wildfires. Using historical wildfire data alongside meteorological, geographical, and temporal information, we aim to:

- **Anticipate Wildfire Incidents:** Evaluate the probability of a wildfire occurrence within an area, taking into account the environmental conditions.
- **Gauge Fire Magnitude**: Once a potential wildfire event is identified, being able to offer insights into its potential classification. This knowledge facilitates better distribution of resources and implementation of mitigation strategies.
- **Guide Prevention and Response**: By understanding the factors contributing to wildfire vulnerabilities, we can provide guidance for preemptive measures, dissemination of public alerts, and swift response tactics. This reduces the impact on both the ecosystem and society.
- **Enhance Understanding**: Understand the various factors contributing to wildfires, ultimately paving the way for further research and policy-making in wildfire management.

03 DATA COLLECTION & PRE-PROCESSING

Data collection:
To predict wildfire occurrences and sizes, we integrated data from the US Department of Agriculture[1] with meteorological data from the Meteostat library[2]. While the former provided wildfire records, it lacked weather data. Using coordinates, we mapped wildfire locations to the nearest Meteostat weather station, enabling us to extract relevant weather data for the corresponding years.

Data Pre-processing:
We streamlined the primary dataset, retaining columns such as Fire Year, Discovery Date, Fire Size, Latitude, and Longitude. Focusing on 2008-2018, we also retained all columns from Meteostat. We categorized fire sizes into classes A-G, adding a 'N' class for days without fires. For balance, each fire day was preceded by an 'N' day. Missing data in variables like snow were set to zero. However, missing values in vital parameters like tmax, tmin, and prcp led to row removal. Ultimately, predictors like tmin, tmax, prcp, pres, wspd, month, Latitude, and Longitude were chosen, with Fire Size Class as our target variable for modeling.

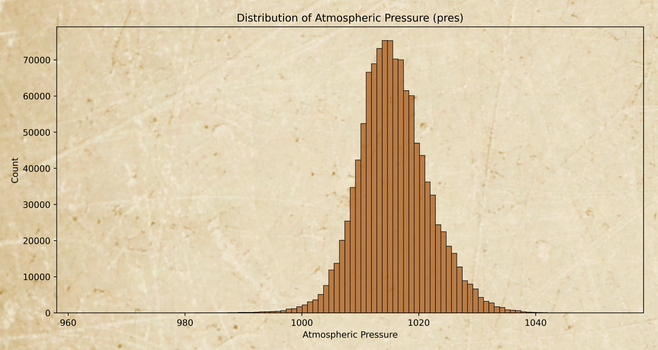
04 EXPLORATORY DATA ANALYSIS

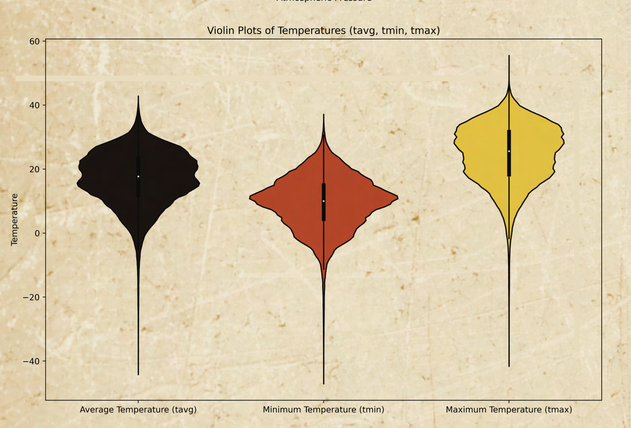
DESCRIPTIVE STATISTICS Upon compiling our dataset, we began with a comprehensive statistical summary. The numerical columns such as Fire Size, Latitude, Longitude, tavg, and weather parameters revealed patterns and distributions crucial for our understanding. For instance, the average fire size stood at approximately 762 hectares, while the mean atmospheric pressure was around 1015 hPa.

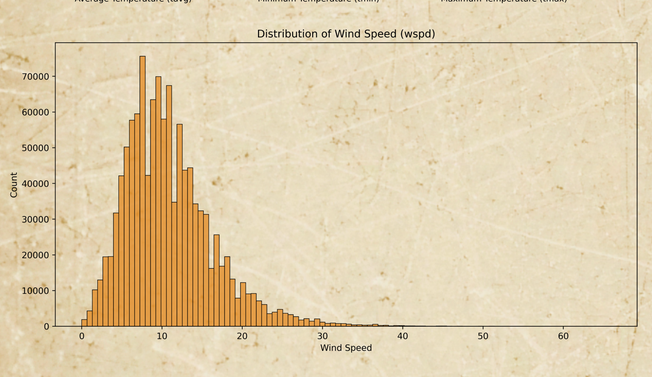
On the categorical front, our dataset consisted of diverse entries. The Fire Class Size column, our primary point of interest, contained 8 unique categories, and the State column highlighted wildfires spread across 51 U.S. states.

GRAPHICAL ANALYSIS To discern patterns and relationships visually:

Weather Parameters: Histograms depicting the distributions of atmospheric pressure, and wind speed. Violin plot depicting distribution of temperature variables. This gave insights into the typical weather conditions on days with wildfires.

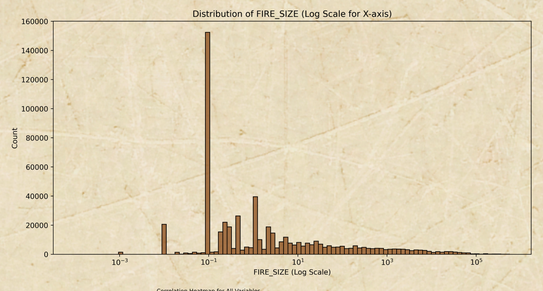


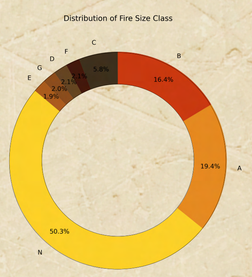


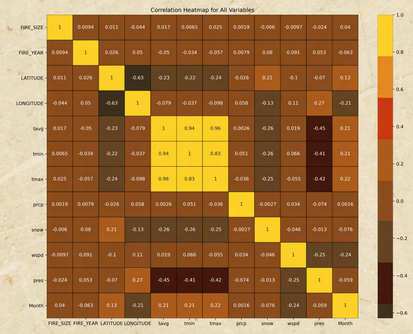


Fire Size: We took the logarithmic scale of the fire sizes to handle its wide range and plotted its distribution.

Fire Size Distribution: Lastly, the distribution of 'FIRE_SIZE_CLASS' was depicted through a pie chart, showing the relative frequencies of different fire sizes.





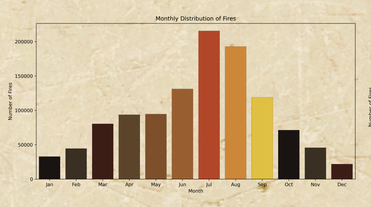


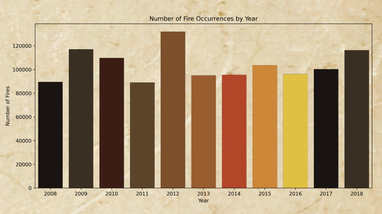
Correlation: A heatmap showcased the relationships between variables, illuminating potential multicollinearity and influential factors.

TEMPORAL ANALYSIS

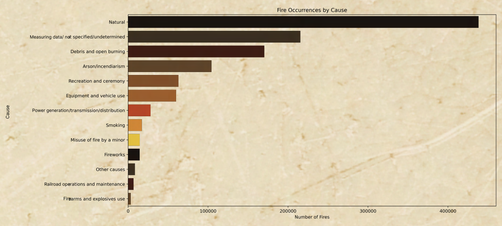
Fire Dynamics: Monthly fire distributions and their causes were charted, revealing seasonality and common ignition sources

Temporal Analysis: The yearly distribution highlighted the number of fire occurrences, emphasizing potential increasing or decreasing trends.

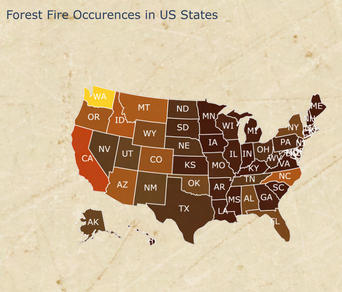


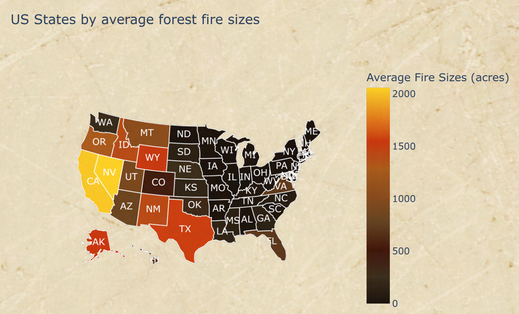


Cause Analysis: Employed a bar plot to describe the primary caused of wildfire revealing human activities and natural events as predominant ignition sources.



State-wise Comparisons: We visualised wildfire occurences across U.S. States using a heat map, highlighting region with frequent outbreaks and areas experiencing intense fires.





05 METHODOLOGY

Our methodology was underpinned by a systematic approach to building and optimizing various machine learning models for wildfire prediction. The steps we undertook are outlined below:

Data Splitting

We partitioned our dataset into training and test sets, allocating 80% of the data for training our models and reserving 20% for evaluation.

Model Selection

We experimented with multiple algorithms to identify the most suitable model for our dataset:

- **Logistic Regression**: As a foundational algorithm, we started with logistic regression to establish a baseline.
- **Decision Tree**: We used decision trees for their interpretability and ability to capture non-linear relationships.
- **Deep Neural Network (DNN)**: Two architectures of DNNs were utilized:
  Initial Model: A 3-layer neural network with 128 and 64 neurons in the first and second layers, respectively. Dropout layers were also integrated to prevent overfitting.
  Tuned Model: This model was more complex, featuring 256, 128, and 64 neurons across its layers, along with increased dropout rates for more robustness. We ultimately decide to go with the selection of 128 neurons as a parsimonious approach.
- **Random Forest:** This ensemble method was implemented in two versions:
  Initial Model: A standard configuration with 100 trees.
  Tuned Model: An optimised version with 200 trees, a maximum depth of 50, and specified criteria for splitting and leaf nodes to enhance performance.

Feature Selection: Initial model training was performed using meteorological data as predictors. However, the best model (tuned Random Forest) yielded an accuracy of only 50%. Recognizing the potential influence of temporal and geographic factors on wildfires, we expanded our feature set to include latitude, longitude, and date. The rationale was to encapsulate the seasonal variations affecting temperatures and the geographical temperature differentials across the country.

Model Evaluation

Following the inclusion of additional predictors, our models were retrained and evaluated. The tuned Random Forest emerged as the top performer yet again, this time achieving an accuracy of 70%. This substantial improvement underscored the importance of feature engineering and selection in model performance.

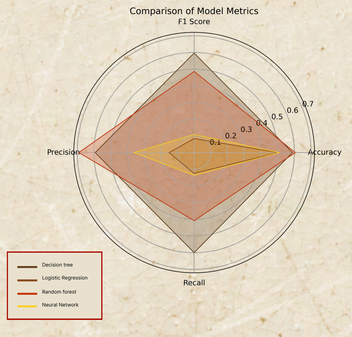
In conclusion, our methodology was iterative, beginning with a basic model and feature set, and gradually expanding based on insights and performance metrics. The tuned Random Forest model, when trained with an extended set of predictors, proved to be the most effective in predicting wildfire occurrences and their sizes.

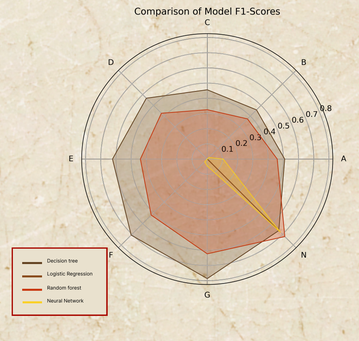
06 RESULTS

Logistic Regression: Starting with the basics, logistic regression didn't quite light up our expectations. With an accuracy just shy of 50% and an average F1 score of 0.08, it's clear this model wasn't the best fit for our data.

Deep Neural Networks (DNN): Venturing into neural networks, our DNN model provided a slight improvement over logistic regression. It achieved an accuracy slightly above 50% and an average F1 score of 0.109. However, for the computational resources it demanded, the returns were not as high as we'd hoped.
Decision Tree vs. Random Forest: This is where things got intriguing. Both these models showed promising results, with the decision tree slightly edging out the random forest. Our decision tree model boasted an accuracy of 59% and an F1 score of 0.596. On the other hand, the random forest model yielded an accuracy of 60.5% but with a lower F1 score of 0.48. Despite the general belief in the robustness of random forests, our decision tree seemed to have better accuracy.

Upon further experiments, tweaking the dataset and removing the 'N' values, the models displayed a significant jump in performance, achieving close to 80% accuracy. Interestingly, the decision tree continued to outshine, even in these altered scenarios.





07 DISCUSSION

The findings of our project presented a series of insightful revelations. Surprisingly, the decision tree emerged as the most effective model. While meteorological factors are undeniably crucial in predicting wildfire occurrences, our analysis underscored the significant influence of geographical location. By integrating data points like latitude, longitude, date, and other geographical attributes, the predictive accuracy experienced a remarkable improvement. This jump in accuracy is a testament to the profound interplay between geographical, temporal, and meteorological factors in the behavior of wildfires.
Seasonal variations were also clearly manifested in our results. For instance, fire occurrences dwindled during winter, aligning with the intuitive understanding of the inhibitive role cold weather plays in fire spread.
From a wildfire protection and management perspective, the implications are manifold. Our findings suggest that merely relying on meteorological predictors might not furnish a holistic picture. Incorporating geographical data can aid in devising more localized and effective fire prevention strategies. Moreover, understanding the temporal patterns, such as reduced fires in winter, can help authorities optimize resource allocation during peak fire seasons. This could lead to better-preparedness and more efficient response mechanisms, ultimately safeguarding ecosystems and communities alike.

08 CHALLENGES

- **Large Dataset Size:** Navigating a dataset of 2.5 million rows required strategic preprocessing, especially given the volume of missing values in columns like 'tmax', 'tmin', and 'prcp'.
- **Missing Values**: Our initial date range (2013-2018) left us with a diminished dataset of just over 400,000 rows after handling missing data. This constraint led us to expand to 2008-2018, yielding 1.4 million usable rows.
- **Computational Constraints:** The sheer size of the dataset, combined with the desire to utilize more complex models like Random Forests, led to computational challenges. The balance between model performance and efficiency became a pivotal aspect. Our computational limitations could have hindered Random Forest from realizing its full potential. Issues such as byte allocation and extended training times were frequent hurdle.
- **Data Imbalance:** Our strategy to distribute 'N' values helped address imbalance. However, omitting 'N' values showed a notable accuracy boost, underscoring the influence of balance on performance.

In essence, we grappled with challenges tied to data volume, missing values, computational limitations, and data balance, each demanding tailored solutions for effective modeling.

09 FUTURE WORK

- **Dataset Expansion**: Our accuracy saw a direct boost with a larger dataset. This advocates for further data augmentation, perhaps extending the date range or integrating additional data sources.
- **Satellite Data Integration:** Satellite imagery offers insights into vegetation health, soil moisture, and other crucial variables. Incorporating this can enhance the model's predictive capability.
- **Real-time Monitoring Systems**: Implementing real-time data capture mechanisms, like sensor networks, can provide timely updates, enhancing predictive accuracy.
- **Climate Change Impact Analysis:** A deeper dive into climate change effects and global warming trends can offer valuable insights. This knowledge can refine hyperparameter tuning, optimizing model performance.
- **Feedback Loop:** Establishing a feedback mechanism is crucial. After predictions are made and interventions implemented, evaluating the outcomes and retraining the model with this feedback ensures its continuous evolution and improvement.

10 CONCLUSION

In our journey to predict wildfire occurrences and sizes using a decade-long dataset, we discovered the critical interplay between meteorological, geographical, and temporal factors. The unexpected success of the decision tree model and the marked improvement upon incorporating geographical data emphasize wildfires' multifaceted nature.
Our work suggests a need for wildfire management to move beyond just meteorological data, incorporating a more holistic approach. The challenges, from data volume to computational limitations, highlighted the intricacies of large-scale environmental modeling.
Ultimately, our study serves as a stepping stone towards a future where wildfires can be more predictably managed, even as we recognize the importance of continuous model refinement in the face of changing climate dynamics.

References

[1] Short, Karen C. 2021. Spatial wildfire occurrence data for the United States, 1992-2018 [FPA_FOD_20210617]. 5th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.5
[2] https://meteostat.net/en/ (Meteostat)
[3] G. E. Sakr, I. H. Elhajj, G. Mitri and U. C. Wejinya, "Artificial intelligence for forest fire prediction," 2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Montreal, QC, Canada, 2010, pp. 1311-1316, doi: 10.1109/AIM.2010.5695809.
[4] R. Laube and H. J. Hamilton, "Wildfire Occurrence Prediction Using Time Series Classification: A Comparative Study," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 4178-4182, doi: 10.1109/BigData52589.2021.9671680.
[5] F. Huot, R. L. Hu, N. Goyal, T. Sankar, M. Ihme and Y. -F. Chen, "Next Day Wildfire Spread: A Machine Learning Dataset to Predict Wildfire Spreading From Remote-Sensing Data," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-13, 2022, Art no. 4412513, doi: 10.1109/TGRS.2022.3192974.
[6] A. K. Sahoo, S. K. Lenka, and S. K. Nayak, "Predicting Forest Fire Using Remote Sensing Data And Machine Learning," arXiv:2101.01975 [cs], Jan. 2021, Accessed: Jun. 16, 2023.
[7] "Predictive modeling of wildfires: A new dataset and machine learning approach".Younes Oulad Sayad a, Hajar Mousannif b, Hassan Al Moatassime;