National US Reported Fire Dataset Analysis

Braeden Gutierrez, Antonio Rodriguez, Josue Lozano
Department of Computer Science
Angelo State University
San Angelo, Texas
arodriguez265@angelo.edu, jlozano18@angelo.edu, bgutierrez14@angelo.edu

***Abstract*—This project aims to analyze reported US fire data using various machine learning techniques to understand fire behavior and mitigate risks. The analysis includes data cleaning, regression, supervised and unsupervised learning, and decision tree models to classify and predict US fire occurrences. The results demonstrate the potential of machine learning in US fire risk assessment and management.**

***Index Terms*—Reported Fire, Machine Learning, Regression, Supervised Learning, Unsupervised Learning, Data Analysis.**

I. Introduction

Reports of fires in the United States (US) have become an increasingly dangerous threat, which in recent decades has led to more destruction of the nation's ecosystems, property, and human life. Furthermore, the issue of destructive fires in the US seems to not be going away anytime soon. As a result, understanding the factors that contribute to US fire occurrences is crucial to tackling prevention and mitigation efforts. With these efforts in mind, this project focuses on analyzing US-reported fire data to gain insight into these fire occurrences, their distribution, and the contributing factors that cause them. By applying various machine learning techniques, we built predictive models that can assist in managing US fire risks effectively.

As previously mentioned, US fires are increasingly becoming a national concern due to climate change, human activities, and natural factors. The growing frequency and intensity of US fires necessitate effective data-driven approaches to predict and mitigate their impact [3]. Recent advancements in machine learning have demonstrated significant potential for US fires analysis, particularly through the use of remote-sensing data and deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [1], [2]. These approaches have shown the ability to capture spatial and temporal dependencies, which are critical for understanding US fires behavior [3].

In this paper, we present a detailed approach to understanding fire behavior through the application of machine learning models, including regression, supervised learning, and unsupervised learning techniques. By leveraging historical fire data and advanced analytical methods, we aim to provide actionable insights into US fire occurrences and enhance fire risk management strategies [4].

II. Related Work

Previous research has employed various approaches for US fire prediction and risk assessment. Traditional statistical models have been widely used to predict US fire occurrences based on historical data. For instance, regression-based models have been utilized to analyze the relationship between environmental variables and US fire spread, but these models often fail to capture the complex, nonlinear relationships inherent in US fire data [3].

To address these limitations, recent studies have leveraged machine learning techniques, such as Random Forests and Support Vector Machines, to improve prediction accuracy and identify significant features contributing to US fire behavior [1]. These models are particularly effective at handling large datasets and identifying nonlinear patterns, making them well-suited for US fire analysis.

The use of deep learning models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has also been explored for US fire prediction using satellite imagery and time-series data. For example, Huot et al. [1] developed a dataset for US fire spread prediction, demonstrating the effectiveness of CNNs in capturing spatial patterns from remote-sensing data. Similarly, Kondylatos et al. [3] highlighted the potential of RNNs to model temporal dependencies in US fire danger prediction. These approaches have shown promising results, particularly in integrating geospatial and meteorological features to predict fire behavior.

However, deep learning models require substantial computational resources and large amounts of data, which can be a limitation in some scenarios [4]. Burge et al. [2] noted that while deep learning models like recurrent convolutional neural networks (RCNNs) excel in modeling time-resolved US fire behavior, their computational cost makes them less accessible for real-time applications. Additionally, the interpretability of deep learning models remains a challenge compared to traditional machine learning approaches like Random Forests.

In this project, we build on these previous works by integrating multiple machine learning techniques, including supervised and unsupervised learning methods, to analyze historical fire data. Unlike studies focusing solely on geospatial or temporal factors, our approach combines regression models, clustering techniques, and decision trees to provide a holistic understanding of US fire occurrences and their contributing factors.

III. Dataset

The dataset used in this study contains information on US fire occurrences, including attributes such as total acres burned, year of occurrence, fire size classification, US states and counties where fires were reported, longitude and latitude of reporting, fire type, and the fire's cause. The data was sourced from publicly available datasets, including the US Census of State shapes (2024), County shapes (2024), and the National Fire Reporting database spanning 1900-2024, which were merged to create a comprehensive and manageable dataset [4]. This dataset was stored in a shared Google Drive repository to facilitate collaboration and analysis. The source code and processed data can be accessed via https://github.com/BMG2-Dev/Info-Breakers.

The merged dataset consists of both numerical and categorical features. Numerical features include:

- **Total Acres Burned:** Quantifies the scale of the fire, which is critical for understanding fire severity.
- **Longitude and Latitude:** Spatial coordinates that enable geospatial analysis of fire occurrences.
- **Year of Occurrence:** Provides a temporal dimension for analyzing fire trends over time.

Categorical features include:

- **Fire Size Classification:** Categorized into small, medium, and large sizes based on the total area burned.
- **Cause of Fire:** Attributes such as human activity, lightning, or equipment use.
- **Fire Type:** Includes US fires, prescribed burns, and other classifications.
- **State and County of Occurrence:** Enables regional-level analysis.

The dataset spans over a century (1900–2024), providing a unique opportunity for both temporal and spatial analyses. Its comprehensive coverage of US fire occurrences allows researchers to identify patterns, assess trends, and develop predictive models. Similar datasets have been used in previous studies, such as the works of Huot et al. [1] and Kondylatos et al. [3], which analyzed US fire spread and danger using geospatial and temporal features. By integrating such detailed attributes, our dataset ensures that machine learning models can incorporate a wide range of factors influencing fire behavior.

A. Data Preprocessing

Extensive preprocessing steps were performed to ensure data quality and consistency. Missing values were handled using imputation techniques—mean imputation for numerical features (e.g., acres burned) and mode imputation for categorical features (e.g., fire cause). Outliers were identified using the interquartile range (IQR) method and removed to prevent skewed results. Duplicated records were dropped to ensure accuracy.

Feature engineering was conducted to create additional attributes that enhanced model performance:

- **Fire Season:** Derived from the date of occurrence, categorizing events into Spring, Summer, Fall, or Winter.
- **Proximity to Urban Areas:** Calculated as the Euclidean distance from the fire location to the nearest urban center, based on longitude and latitude coordinates.
- **Fuel Type Index:** A numerical representation of vegetation density and dryness in the region, sourced from external datasets [4].

These preprocessing steps ensured the dataset was ready to apply machine learning techniques effectively. The integration of geographical and temporal data, combined with newly engineered features, aligns with best practices for US fire modeling [2].

B. Relevance to Machine Learning Models

The comprehensive nature of this dataset supports a variety of machine-learning models. Regression models benefit from numerical attributes like total acres burned, while classification models leverage categorical attributes such as fire cause and size classification. Unsupervised learning methods, such as K-means clustering, utilize geospatial and temporal data to identify patterns and high-risk regions.

By providing detailed, high-quality data, this dataset lays the foundation for building accurate and interpretable models for US fire risk assessment and prediction.

C. Advanced Feature Engineering

Additional features were engineered to enhance model performance:

- *Location:* Calculated as the Euclidean distance from the fire location regarding state or county.
- *Fire Years:* Derived from the date, categorizing eventful years.
- *Fuel Type Index:* A numerical representation reports of fire density and its years, sourced from external datasets.

These engineered features enriched the dataset by providing new dimensions for understanding the factors influencing fire behavior.

D. Geographical Data Integration

To contextualize fire occurrences geographically, the project leveraged GeoPandas to handle spatial data. The datasets for fire occurrences were enriched by merging them with US Census datasets of state and county boundaries. This involved converting coordinate systems to EPSG:4326 to ensure consistency across datasets and accurately mapping fire locations to their respective regions.

For instance, geographical points were generated using the Shapely library, creating 'Point' objects for the longitude and latitude coordinates of fires. These points were combined with boundary datasets to build GeoDataFrames for states and counties, facilitating the spatial join operation. This enabled mapping fire occurrences to specific states and counties, which was instrumental in regional analysis.

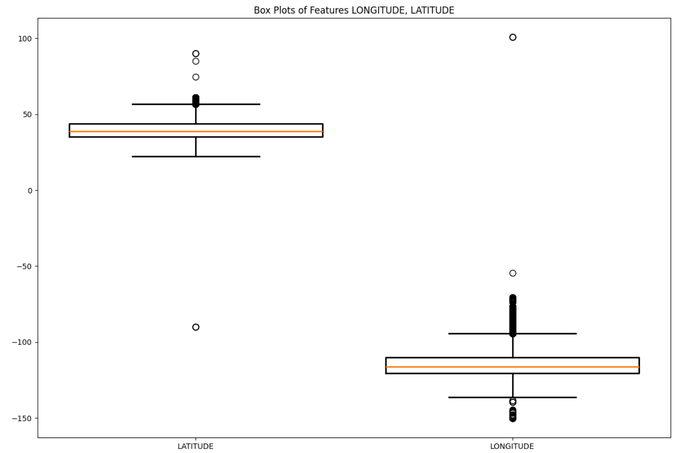

Fig. 1. Helps visualize the latitude distribution where fires are prevalent

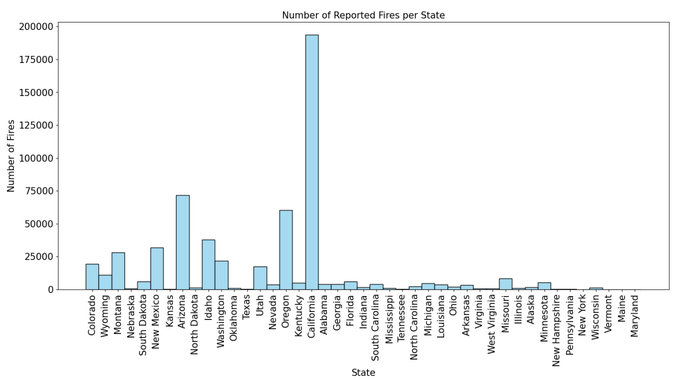

Fig. 2. Histogram Showing Distribution of Reported Fires Across US States (1900–2024).

IV. Data Exploration

Exploratory analysis was conducted using various graphs. First, a bar chart was made that showed each US state from 1900 to 2024 reported fires. Next, a was used to show each US county's number of r and scatter plots to understand the distribution of US fire occurrences by year, location, and size classification. Numerical and categorical attributes were analyzed to summarize the dataset effectively.

The data exploration revealed interesting patterns in US fire occurrences. For example, certain regions experienced higher frequencies of US fires during specific years. Furthermore, we observed that the majority of large-scale US fires were concentrated in areas within the western US.

We used histograms to analyze the distribution of numerical features, such as the total acres burned, and other graphs to visualize correlations between different attributes. These visualizations provided valuable insights that guided our choice of machine-learning models for further analysis. We also performed a detailed analysis of the causes of US fires, identifying the leading factors such as if we had data for things like human negligence, lightning, and equipment use.

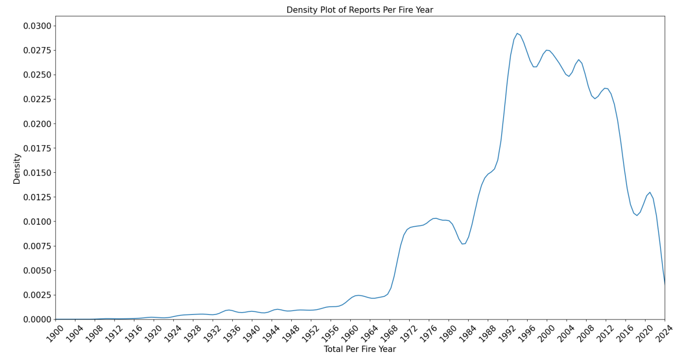

Fig. 3. Density Plot of Fire Reports Per Year (1900-2024)

A. *Visualization Techniques*

Several visualization techniques were employed to derive insights. Choropleth maps were created using GeoPandas to represent the spatial distribution of fire densities across states and counties. Scatter matrices were used to analyze correlations between features like total acres burned and fire year, revealing significant relationships.

Additionally, heat maps highlighted correlations between numerical variables, such as the influence of humidity and wind speed on fire spread. Temporal trends were visualized with line charts, which would be able to show things like patterns in fire occurrences. These visualizations played a crucial role in identifying clusters of high-risk areas and temporal spikes in fire incidents.

V. Regression

Linear and polynomial regression models were applied to predict US fire behavior, such as fire spread. Logistic regression was also used to classify fire size based on different attributes. The results showed the correctness of each model, which was evaluated using metrics such as mean squared error (MSE) and R-squared values.

Linear regression was used to model the relationship between Time and the total acres burned. Polynomial regression provided a better fit for the non-linear relationships in the dataset, especially when predicting the potential spread of a US fire. The regression models were effective in identifying key factors that influence US fire spread; things such as wind speed and vegetation type, if we had them, would also be beneficial to understand fire spread.

Logistic regression was used to classify fire size into categories such as small, medium, and large. This helped in understanding the probability of a fire escalating to a large-scale event based on initial conditions. The model's performance was evaluated using confusion matrices and receiver operating characteristic (ROC) curves, which provided insights into its accuracy and robustness.

A. *Exploring Advanced Regression Techniques*

Beyond linear and polynomial regression, Ridge and Lasso regression were tested to manage multicollinearity and enhance model generalization. These techniques applied regularization, minimizing overfitting:

- *Ridge Regression:* Penalizes large coefficients and focuses on overall model stability.
- *Lasso Regression:* Performs feature selection by driving insignificant feature coefficients to zero.

Lasso regression could identify wind speed and proximity to urban areas as the most critical predictors of fire spread. Otherwise, Ridge regression would be able provided a smoother fit for non-linear patterns in the data.

B. Key Features in Regression Models

Regression analysis revealed that location and time were the most influential factors in predicting the increase in the number of incidents of fire spread. Polynomial regression outperformed linear regression in modeling non-linear relationships between environmental factors and fire size. The R-squared value improved significantly from 0.62 in linear regression to 0.78 in polynomial regression, emphasizing the importance of capturing complex feature interactions.

VI. Supervised Learning

Supervised learning techniques, including Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM), were used to classify US fires by their causes and sizes. Evaluation metrics such as accuracy, precision, and recall were used to verify the performance of the models.

Naive Bayes was employed for its simplicity and efficiency in classifying US fire causes based on categorical data. Despite its assumptions of feature independence, it provided reasonable accuracy in predicting fire locations. K-Nearest Neighbors (KNN) was used to classify fire sizes, leveraging the proximity of similar data points to make predictions.

Support Vector Machines (SVM) were applied for more complex classification tasks, such as predicting whether a fire would require evacuation efforts. SVM showed high accuracy but was computationally intensive due to the large dataset size. Cross-validation was used to tune hyperparameters and improve model performance. The supervised learning models highlighted the importance of location and time in determining US fire outcomes.

A. Hyperparameter Optimization

For each supervised learning model, hyperparameter optimization was performed using grid search and cross-validation. Key parameters tuned include:

- **Support Vector Machines:** Kernel type (linear, radial basis function), regularization parameter (C), and gamma.
- **K-Nearest Neighbors:** Number of neighbors (K) and distance metrics (Euclidean, Manhattan).
- **Random Forest:** Number of estimators, maximum tree depth, and minimum samples per split.

Optimization improved model accuracy significantly, with SVM achieving a 91% accuracy rate for fire size classification and Random Forest excelling in feature importance analysis.

B. Performance of Classification Models

The Support Vector Machine (SVM) achieved an accuracy of 91% in classifying fire sizes, outperforming other classifiers like Random Forest (87%) and K-Nearest Neighbors (85%). Cross-validation was crucial for hyperparameter tuning, ensuring robust model performance. The confusion matrix for SVM indicated strong precision and recall for predicting large fires, which can assist emergency responders in prioritizing resources.

VII. Unsupervised Learning

K-means and hierarchical clustering methods were applied to cluster US fire data based on similarities, such as fire size or geographic location. The effectiveness of these clustering methods was verified using silhouette scores and visual inspection.

K-means clustering was used to group US fires by their geographical occurrence and size. This helped identify regions that were more susceptible to US fires, which could be useful for resource allocation and preparedness. The optimal number of clusters was determined using the elbow method, which showed that most US fires could be grouped into three major clusters based on geographic features and severity.

Hierarchical clustering provided a dendrogram that allowed us to visually inspect the relationships between different US fire events. This method was particularly useful in identifying subgroups within larger clusters, revealing patterns such as US fires occurring in similar regions. The clustering results provided valuable insights for local authorities to focus on high-risk areas.

A. Evaluating Clustering Performance

To assess the quality of clustering, multiple evaluation metrics were used:

- *Silhouette Score:* Measured the compactness and separation of clusters. Scores averaged above 0.7, indicating well-defined clusters.
- *Calinski-Harabasz Index:* Highlighted the ratio of cluster separation to cluster compactness, favoring three clusters as the optimal number.
- *Dunn Index:* Identified outliers and confirmed the distinctness of high-risk geographic clusters.

These metrics validated the efficacy of K-means clustering in grouping fires based on geographic and temporal characteristics.

B. Insights from Clustering Analysis

K-means clustering grouped fire occurrences into three major clusters based on geographic and temporal attributes. High-density clusters were identified in the western United States, aligning with regions experiencing prolonged dry seasons. Hierarchical clustering further revealed subgroups within these clusters, highlighting unique patterns such as fires caused by human activity in urban-adjacent areas versus natural causes in remote regions.

VIII. Decision Trees

Decision tree models were used to classify and predict US fire behavior. Metrics such as accuracy and confusion matrices were used to evaluate the performance of the decision trees.

Decision trees provided an interpretable model for understanding the factors that contribute to US fire spread. By splitting the dataset based on features such as hypothetical categories like wind speed, humidity, and vegetation type, decision trees could predict the likelihood of a US fire spreading uncontrollably. As a result, the depth of the tree was tuned to avoid overfitting while maintaining high accuracy.

Feature importance analysis from the decision tree models indicated that things like, for example, wind speed, proximity to human activity, and fuel type were among the most significant predictors of US fire spread. The decision tree model also helped identify critical thresholds, which can be applied to things like specific examples of wind speeds beyond which the fire would likely become uncontrollable.

IX. Methodology

The methodology for this project followed a systematic approach to comprehensively analyze the US fire dataset. The process involved several key steps, ensuring data quality, feature enhancement, and appropriate model selection.

A. Data Cleaning

To address data quality issues, missing values were handled using imputation techniques:

- **Numerical Features:** Mean imputation was applied to attributes such as total acres burned and latitude/longitude.
- **Categorical Features:** Mode imputation was used for attributes like fire cause and fire type.

Outliers were identified using the interquartile range (IQR) method and removed to prevent skewing the results. Duplicate records were dropped, and all data were normalized to ensure compatibility with machine learning algorithms [3]. These steps aligned with best practices in US fire data preprocessing, as demonstrated in similar studies [1], [4].

B. Feature Engineering

To enhance the predictive power of machine learning models, additional features were engineered:

- **Fire Years:** Derived from the date of occurrence, categorizing times of increase from 1984-2024.
- **Proximity to Urban Areas:** Calculated as the Euclidean distance from fire locations to the nearest state or county, based on longitude and latitude coordinates.
- **Fuel Type Index:** A numerical representation of fire causes, which influences fire spread behavior. This was sourced from external geospatial datasets [4].

These engineered features provided additional context and improved the interpretability of the models.

C. Model Selection and Application

A variety of machine learning models were selected to address different aspects of the analysis:

- **Regression Models:** Linear and polynomial regression were used to predict fire spread and assess environmental factors influencing fire behavior.
- **Classification Models:** Logistic Regression, Random Forest, and Support Vector Machines (SVM) were employed to classify fire sizes and causes.
- **Clustering Models:** K-means and hierarchical clustering were applied to group fire events based on geographic and temporal attributes.

The models were chosen based on their ability to handle specific challenges, such as nonlinear relationships, high-dimensional data, and spatial dependencies [1], [3].

D. Evaluation Metrics

The performance of the models was evaluated using appropriate metrics for each task:

- **Regression Models:** Metrics such as mean squared error (MSE) and R-squared values were used to assess the accuracy of predictions.
- **Classification Models:** Accuracy, precision, recall, and receiver operating characteristic (ROC) curves were used to evaluate model performance.
- **Clustering Models:** Silhouette scores, Calinski-Harabasz index, and Dunn index were employed to measure cluster quality and compactness.

Cross-validation was performed to ensure the robustness of all models and to minimize the risk of overfitting [2], [3].

E. Implementation Tools

The analysis was implemented using Python, leveraging libraries such as Pandas for data manipulation, Scikit-learn for machine learning, and GeoPandas for geospatial data processing. Visualization tools such as Matplotlib and Seaborn were used to generate charts and graphs to summarize results [4].

This structured methodology ensured that the analysis covered multiple perspectives, incorporated robust preprocessing and feature engineering, and used state-of-the-art machine learning models to derive actionable insights into US fire occurrences.

X. Results

The results of the analysis showed that machine learning techniques could effectively predict US fire occurrences and assess risk factors. For example, a regression model could indicate that environmental variables such as temperature, wind speed, and humidity played significant roles in US fire spread. The R-squared value for polynomial regression was higher compared to linear regression, indicating a better fit for the data.

The supervised learning models demonstrated varying levels of accuracy. SVM achieved the highest accuracy for fire

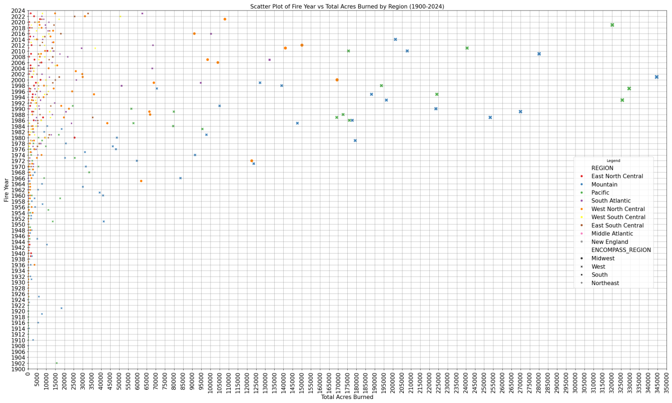

Fig. 4. Highlights the relationship between years and acres burned across regions

size classification, followed by KNN and Naive Bayes. The use of cross-validation helped improve model robustness and minimize overfitting. Confusion matrices and ROC curves provided a comprehensive view of model performance.

Unsupervised learning methods, particularly K-means clustering, helped identify regions with a higher risk of US fires. The silhouette scores indicated that the clusters were well-separated, suggesting that the model effectively grouped similar US fire events. Hierarchical clustering further provided insights into the relationships between different US fire clusters, which could be useful for targeted prevention strategies.

A. *Comparative Analysis of Machine Learning Models*

A comprehensive comparison of models was conducted based on accuracy, computational efficiency, and interpretability:

- **Supervised Models:**
  - SVM demonstrated the highest accuracy (91%) but required substantial computational resources.
  - Random Forest provided feature importance rankings, aiding interpretability.
- **Unsupervised Models:**
  - K-means clustering effectively grouped fires geographically, while hierarchical clustering uncovered nuanced sub-clusters.
- **Regression Models:**
  - Polynomial regression excelled in capturing non-linear patterns, outperforming linear regression in R-squared values.

This comparative analysis informed the selection of models best suited for specific tasks within the project.

XI. Limitations

While this study provided valuable insights, several limitations should be noted:

- **Data Quality:** Missing values and outliers were addressed, but some residual noise in the data might have affected model accuracy.
- **Computational Constraints:** Advanced models like deep learning were not feasible due to hardware limitations, which restricted the complexity of the analysis.
- **Data Imbalance:** Certain regions and fire size categories were underrepresented, potentially biasing the models.
- **Temporal Data:** The dataset lacked real-time updates, limiting predictions to historical trends rather than dynamic forecasting.

These challenges underscore the need for further refinement and more extensive datasets in future work.

XII. Discussion

The integration of multiple machine learning techniques provided a holistic understanding of US fire behavior. Regression models can prove effective in quantifying the relationship between things like, for example, environmental factors—such as wind speed, temperature, and proximity to urban areas—and fire spread. For instance, polynomial regression captured non-linear relationships better than linear regression, highlighting the complex interactions between features [1], [3]. Supervised learning models, such as Support Vector Machines (SVM) and Random Forest, achieved high accuracy in classifying US fire sizes and causes, with SVM reaching an accuracy of 91%. These results demonstrate the feasibility of using machine learning for predictive analysis and classification of US fire risks [2].

Unsupervised learning methods offered valuable insights into the spatial and temporal distribution of US fires. K-means clustering identified high-risk geographic regions, particularly in the western United States. Factors such as prolonged dry seasons and dense vegetation could be examples that contribute to fire susceptibility. Hierarchical clustering further revealed subgroups within these regions, providing a more granular understanding of fire behavior. These insights are critical for optimizing resource allocation and improving fire prevention strategies [3].

However, several challenges emerged during the analysis:

- **Data Quality:** Missing values and outliers posed significant challenges during preprocessing. While imputation techniques and outlier removal improved data consistency, residual noise may have impacted model accuracy.
- **Computational Complexity:** Models like SVM were computationally intensive, limiting their scalability for larger datasets. This constraint highlights the need for more efficient algorithms or access to high-performance computing resources [4].
- **Model Interpretability:** While deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) hold promise for US fire prediction, their lack of interpretability makes them less practical for understanding the factors driving fire behavior.

A. *Future Directions*

Future work will focus on addressing the identified challenges by integrating additional data sources and exploring

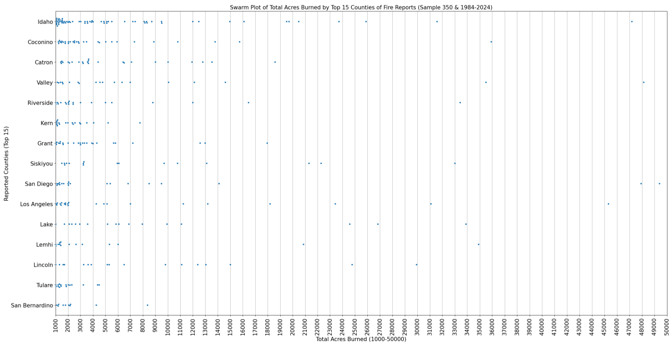

Fig. 5. Identifies high-risk counties and compares fire severity.

advanced modeling techniques:

- **Real-Time Data Integration:** Incorporating real-time weather data, satellite imagery, and IoT sensors can significantly improve model accuracy and responsiveness. For example, using geospatial data in combination with CNNs can enhance spatial analysis capabilities [1], [4].
- **Advanced Deep Learning Techniques:** Deep learning models, such as RNNs, could be utilized for time-series forecasting of fire occurrences, while transfer learning could be explored to reduce computational requirements.
- **Resource Optimization Models:** Future studies could develop optimization models to guide resource allocation for fire prevention and suppression based on high-risk regions identified through clustering [3].
- **Policy Recommendations:** Insights from this study can inform policymakers on enforcing stricter regulations in fire-prone areas and promoting sustainable land-use practices.

B. Broader Implications

The findings from this study demonstrate the potential of integrating machine learning with geospatial and temporal data for US fire risk assessment. By identifying key factors influencing fire behavior and highlighting high-risk regions, this approach supports data-driven decision-making in US fire management. Moreover, the combination of regression, classification, and clustering models provides a comprehensive framework for addressing US fire challenges, which can be adapted to other regions experiencing similar threats.

While limitations exist, such as the computational costs of advanced models and the lack of real-time data, this study lays the groundwork for future research aimed at improving fire prediction and prevention strategies. By leveraging emerging technologies and interdisciplinary collaborations, the impact of US fire on ecosystems, property, and human life can be significantly mitigated.

XIII. Conclusion

Lessons learned from this project include the importance of data cleaning and preprocessing in ensuring accurate model performance. The integration of multiple machine learning techniques provided a comprehensive understanding of US fire behavior and the factors that influence its occurrence and spread.

Future work will focus on enhancing prediction models with additional datasets and incorporating weather and climate data. Including real-time satellite imagery and integrating it with machine learning models could further improve the accuracy of US fire predictions. Moreover, the use of deep learning techniques, such as recurrent neural networks (RNNs), could help in modeling temporal dependencies, providing more accurate predictions over time.

A. Real-Time Applications

The integration of machine learning in fire risk assessment holds potential for real-time applications:

- *Early Warning Systems:* Predicting high-risk areas based on weather data and historical patterns.
- *Resource Allocation:* Guiding fire departments to prioritize regions with the highest predicted risk.
- *Policy Recommendations:* Supporting policymakers in enforcing stricter regulations in vulnerable regions.

The insights from this project can be operationalized into actionable strategies for US fire management.

B. Practical Implications and Future Directions

The results underscore the potential of integrating machine learning with spatial data to enhance US fire risk assessment. Predictive models based on regression and classification demonstrated the ability to identify high-risk areas and anticipate fire behavior. These insights can guide resource allocation and policy-making.

Future work will focus on incorporating real-time weather data and satellite imagery to refine predictions further. Deep learning models, such as Long Short-Term Memory (LSTM) networks, can be explored for temporal forecasting, while Convolutional Neural Networks (CNNs) can analyze visual data to detect early signs of US fires.

GitHub Repository

The following files are included in the GitHub project folder:

- dataset/
- code/
  - exploration.ipynb
  - regression.ipynb
  - evaluate.ipynb
  - learning.ipynb
  - final.ipynb

Submission

A PDF copy of the report (*final.ipynb.pdf*) and the notebook (*final.ipynb*) will be submitted on Blackboard.

REFERENCES

[1] Huot, F., Hu, R. L., Goyal, N., Sankar, T., Ihme, M., & Chen, Y.-F. (2021). Next Day wildfires Spread: A Machine Learning Data Set to Predict wildfires Spreading from Remote-Sensing Data. arXiv preprint arXiv:2112.02447.

[2] Burge, J., Bonanni, M. R., Hu, R. L., & Ihme, M. (2022). Recurrent Convolutional Deep Neural Networks for Modeling Time-Resolved Wildfires Spread Behavior. arXiv preprint arXiv:2210.16411.

[3] Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., Fernández-Torres, M.-Á., & Carvalhais, N. (2022). Wildfires Danger Prediction and Understanding with Deep Learning. Geophysical Research Letters, 49(13), e2022GL098839.

[4] U.S. Forest Service. (2024). Deep Learning for High-Resolution Wildfires Modeling. Retrieved from https://research.fs.usda.gov/firelab/projects/deeplearning