California fire incident analysis and risk classification

STRATEGIC THINKERS

AIT-664-002

GEORGE MASON UNIVERSITY

Submission Date: 05-01-2025

DANDU LAXMI SARVAJNA

SAI SRAVYA CHANDAVOLU

SAI SAMPATH GUNUPURU

MALAPATI SRAVANI LAKSHMI

Abstract:

Wildfires have been one of the dominant drivers shaping California's ecologically diverse landscapes for centuries. During the past few decades, though, the incidence, severity, and intensity of wildfires have increased exponentially due to the synergistic interactions of climate change, prolonged droughts, land use change, and increasing human in-migration into fire-prone areas. The escalating threat of wildfires calls for the development of advanced analytical capability for more intense analysis and wildfire behavior prediction.

This project offers a complete analysis of the massive dataset of historical wildfire incidents in California. The overall objective is to preprocess the data diligently, construct useful features, and apply robust machine learning algorithms with the hope of classifying the fire incidents in various risk levels based on some of the most significant factors such as acres burnt, resources deployed, and structural damages suffered. Special care is taken in dealing with missing data processing, variable redundancy elimination, and input variable normalization while training the model of quality.

Dimensionality reduction techniques, i.e., Principal Component Analysis (PCA), were employed in an effort to preserve the most informative features and minimize redundancy with a view to promoting computational effectiveness as well as preventing the risk of model overfitting. Random Forest Classifier, Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost) machine learning models were applied, tested, and compared in prediction accuracy.

The research findings reveal that predictive modelling with appropriate preprocessing and dimensionality reduction can predict the classification of wildfire events accurately to risk classes. The research results have the potential to guide policymakers, emergency managers, and environmental agencies to prepare proactive planning strategies for the management of wildfires, resource planning, and preventing disaster.

1. Introduction

Wildfires are a significant natural phenomenon in California, contributing to ecological renewal while simultaneously posing serious risks to life, property, and the environment. Over the past few decades, the severity and frequency of these events have sharply increased due to climate change, urban expansion into fire-prone areas, and prolonged periods of drought. Accurate prediction and classification of wildfire risks are critical for improving emergency preparedness, resource allocation, and disaster response strategies.

This project focuses on analysing a comprehensive dataset of historical wildfire incidents in California, aiming to build predictive models that classify incidents into different risk categories. The initial phase involved thorough data preprocessing, which included handling missing values, dropping irrelevant columns, and standardizing feature formats to ensure data quality. A critical feature engineering step was the creation of a Risk Category label based on the number of acres burned, categorizing incidents into low, medium, and high-risk levels.

To address the challenge of high-dimensional data and enhance model performance, Principal Component Analysis (PCA) was applied for dimensionality reduction, retaining the essential variance within fewer features. Subsequently, multiple machines learning algorithms, including Random Forest, Support Vector Machine (SVM), and XGBoost classifiers, were trained and evaluated.

The models were assessed primarily using accuracy scores on test data, comparing their effectiveness both with and without PCA transformation. Among the models, Random Forest achieved the highest performance on the original data, while XGBoost also performed competitively after PCA.

This comprehensive approach demonstrates that, with careful preprocessing, feature engineering, and model selection, it is possible to create efficient systems capable of predicting wildfire incident risks, ultimately supporting decision-makers in wildfire management and mitigation efforts.

2. Dataset description

The data used in this project is named "California Fire Incidents" and contains historical fire incident data of the ones occurred in the various regions of California. There is a single observation for each in the data, where each signifies an independent fire incident, with significant operational, geographic, and effect-related characteristics belonging to it.

The data set contains very comprehensive sets of features, some of which are:

Acres Burned: Total area burned area caused by the fire, expressed in acres.

Structures Threatened, Structures Damaged, Structures Destroyed: Quantities representing extent of structural damage during the course of the fire.

Personnel Involved, Engines, Dozers, Air Tankers, Helicopters: Data concerning emergency personnel and equipment used to combat the fire.

County Ids, Counties: Place identifiers representing location of the incident.

Status, Percent Contained: Field variables denoting status of fire containment and whether the fire is live or dead.

Injuries, Deaths: Human factor measures due to the fire accident.

Certain metadata fields such as Canonical URL, Condition Statement, and Public were not of extreme importance to predictive modeling and were eliminated during preprocessing for enhanced model focus and performance.

The database was inundated with missing values in a few of the most critical columns, including Acres Burned and other numeric columns dealing with resources. Missing values were addressed using careful imputation methods, including filling in missing numeric data with zeros wherever feasible or removing records where important columns were missing.

In addition, a new target feature, Risk Category, was created based on thresholds in the Acres Burned field, labeling events as Low, Medium, or High Risk. This feature was used as the foundation for the classification models constructed in the project.

The overall dataset was a rich and detailed source of information for studying the drivers of wildfire severity and for constructing robust predictive models for risk categorization.

3. Data Preprocessing

Accurate data preprocessing is a significant part of the initiation that has a direct effect on the efficacy and reliability of machine learning models. For this project, the Californian Fire Incidents dataset went through a rigorous pipeline for data preprocessing to address irrelevant data, missing values, and format discrepancies in an attempt to prepare it to be adequately prepared for modelling.

3.1 Deletion of Irrelevant Columns

The original data contained some columns that were not utilized directly in the predictive model activity. It contained some metadata columns such as Canonical URL, Condition Statement, Control Statement, Uniquid, Search Description, Search Keywords, and Updated.

These columns, although useful for logging or description, had no actionables predictive information about the number or severity of fire incidents. Keeping them would be to introduce noise, artificially inflate the dimension, and perhaps reduce model performance by distracting the learning processes away from more significant patterns.

Thus, by conscious design, all the columns that were not necessary were eliminated. Not only was this shrinking the data set but also making subsequent feature engineering and modeling steps more efficient and focused.

3.2 Missing Value Handling

Missing values are common in real-world data sets, and their treatment is necessary to avoid the introduction of bias or errors into machine learning models. Missing data were encountered largely in significant numeric fields concerning fire effect and response resources in the California Fire Incidents data set.

Missing data treatment was handled in two manners:

Critical Fields (e.g., Acres Burned):

For Acres Burned, an essential field to the building of the target variable Risk Category, all records with missing data were deleted. It would be nonsensical to forecast risk without understanding how large the fire was (acreage burned).

Resource Deployment Fields (e.g., Airtankers, Engines, Helicopters):

For operational areas that were the number of resources utilized, zeros were employed to fill in missing values. The assumption was that a missing value would most likely indicate that no resources of that sort were utilized and not an error in recording.

This method maintained the integrity of important features without wasteful loss of data volume.

3.3 Encoding Categorical Variables

Numerical inputs were most often expected by most machine learning models. So, categorical variables needed to be encoded appropriately.

Two encoding strategies were used:

One-Hot Encoding:

Columns like Counties and Status with greater than one category were encoded by the one-hot encoding. One-hot encoding gave separate binary (0/1) columns for each category such that models can differentiate between categories without any ordinal relationship assumption.

Label Encoding

Target variable Risk Category, which was derived from Acres Burned, was label-encoded. Each of the risk classes (Low, Medium, High) was coded as an integer value (for instance, 0, 1, 2), and now it stood a chance to be used with classification algorithms.

Encoding converted the last dataset to an entirely numeric and usable form so that it could be successfully used to train machine learning models.

3.4 Final Dataset Preparation for Modelling

Once the data had been cleaned and converted, the resulting dataset was structured into two main sections:

Features (X):

The input variables, numeric features and one-hot encoded features, which are the specifics of each fire event.

Target (y):

The output variable (RiskCategory_Code) which is the risk category of each event.

Data was comprehensively checked for consistency in a way that there were no missing values, data type was consistent, and all features were prepared for further operations like dimensionality reduction (PCA) and model training.

Having a strict preprocessing pipeline in this way meant machine learning models were being built on quality, meaningful data with strong potential to accurately predict.

4. Feature Engineering

Feature engineering is the process of creating new variables from the data at hand that better expose the underlying pattern to the machine learning algorithms. Feature engineering for the project was a significant task towards achieving higher predictive power as well as towards making the fire incident classification meaningful.

4.1 Building Risk Category

- The most significant feature engineering task was the creation of a new target variable, Risk Category.
- Instead of trying to forecast the actual numeric extent of acres burned (a regression issue), fires were categorized into discrete classes for risk and a multi-class classification issue arose.
- The reason for the break-off was this:
- `Low Risk: Fires that burned ≤ 35 acres`
- Medium Risk: Fires that burned between 36-100 acres
- High Risk: Fires that burned > 100 acres
- This binning method facilitated more realistic and meaningful classification of fire incidents by severity directly applicable to real emergency response planning scenarios.

4.2 Risk Category Encoding

- Having risk categories defined, the new feature RiskCategory was label-encoded to numerical:
- Low Risk → 0
- Medium Risk → 1
- High Risk → 2
- This encoding was necessary in preparation for pre-processing the data for machine learning algorithms because they take numerical input for target variables in classification problems.

4.3 Dimensionality Reduction using PCA

Because there were numerous one-hot encoded categorical fields and operation fields, Principal Component Analysis (PCA) was employed for dimensionality reduction of the data.

Key steps:

- PCA was fit to the feature space (X) following one-hot encoding.
- The initial principal components were selected in a way that ensured 95% variance was retained, balancing information loss against computational expense.

The use of PCA gave the following benefits:

- Decreased the feature space.
- Prevented overfitting as an option.
- Improved model training, especially computationally costly models like Support Vector Machines (SVM).
- The major component reduced data served as an input to some models that can work efficiently with lower-dimensional representation.

5. Model Building

Machine learning classifiers are at the center of this project to forecast the risk level of wildfire breakouts based on structured input features. Three classifiers were implemented and compared.

5.1 Random Forest Classifier

Random Forest is an ensemble learning method that creates many decision trees and merges their predictions for stronger and more accurate predictions. A Random Forest Classifier was first trained on the original data (without PCA) in this project. Later, it was re-trained on the PCA-transformed dataset to evaluate the impact of dimensionality reduction. Random Forest was used because it is insensitive to noisy features, can handle large datasets, and has built-in feature importance analysis.

5.2 Support Vector Machine (SVM)

Support Vector Machine models perform well in high-dimensional spaces but tend to overfit when there are many irrelevant features. Thus, SVM was trained only on the PCA-reduced dataset. The idea was to leverage PCA's dimensionality reduction to improve the model's efficiency in training and generalization ability. The SVM would find the optimal hyperplane that maximally separates different risk classes.

5.3 XGBoost Classifier

XGBoost (Extreme Gradient Boosting) is a very efficient boosting algorithm with superior performance on classification tasks. For this project, XGBoost was trained on the PCA-reduced dataset. Important hyperparameters such as n_estimators=300, learning_rate=0.01, and max_depth=4 were adjusted to obtain maximum model accuracy and prevent overfitting.

6. Model Evaluation

6.1 Performance Measure

Accuracy score was selected as the primary measure to assess the performance of the wildfire risk prediction models. Accuracy is a proportion of correctly classified instances measure and is a simple measure of model performance. The data were split into 80% train data and 20% test data to test the generalization ability of the models on new data. The predictions of the test set of trained models were compared with actual labels, and accuracy for both was reported. Accuracy is a simple and intuitive metric, but in cases of class imbalance, one might need to use other metrics like Precision, Recall, and F1-Score. But after preprocessing was complete and AcresBurned binning was carried out, the classes were balanced, and hence it was okay to use accuracy as the primary measure for all the models.

6.2 Model Comparison

The three machine learning models were tried and compared: Random Forest, Support Vector Machine (SVM), and XGBoost. Each model was evaluated based on test set accuracy. Random Forest was tried with and without the use of Principal Component Analysis (PCA). Without PCA, Random Forest was the most accurate since it performed well in handling high-dimensional data without PCA. With PCA, Random Forest saw lower accuracy but increased training time. SVM was highly benefited as SVM performs better for low-dimensional space, and results improved. XGBoost was trained on reduced data by PCA and performed wonderfully after hyperparameter tuning (n_estimators=300, learning_rate=0.01, max_depth=4). Finally, Random Forest (on full data) and XGBoost (after PCA) performed the best within this project.

6.3 Confusion Matrices

Apart from accuracy measures, confusion matrices were also obtained for each model to further study prediction errors. A confusion matrix graphs the counts of true positive, true negative, false positive, and false negative predictions for each class. With confusion matrices in this wildfire risk category assignment task, how well individual models

discriminated among Low, Medium, and High Risk classes were reported. The most well-balanced confusion matrix was returned by Random Forest that accurately identified the majority of cases with near negligible misclassifications between Medium and High Risk. SVM also demonstrated higher levels of confusion within levels of instant risks, which certified its sensitiveness to edge cases. Well-balanced results were reported with minor misclassifications by XGBoost with its parameters tweaked. Confusion matrices helped identify systematic patterns of error, giving essential feedback on improving the model even more. In most cases, they proved that Random Forest and XGBoost were more reliable when it comes to accurate wildfire risk classification.

7. Results and Analysis

7.1 Most Significant Results from Model Outputs

The output indicated that Random Forest without PCA performed the best among the models because it proved to perform well with big and complicated feature spaces. It was able to detect complicated patterns in the wildfire dataset without applying dimensionality reduction. XGBoost after PCA also performed well, with the benefit of exact tuning of hyperparameters and the less complicated complexity of the dataset. Support Vector Machine (SVM) gained considerably after PCA but did not yet outperform Random Forest or XGBoost. Random Forest performed best on average when used with the full feature set, while XGBoost provided a solid and stable alternative after PCA. The results indicate that Random Forest would be optimally suited to higher-feature datasets, while boosting models like XGBoost are incredibly powerful when the dataset is fine-tuned for performance.

7.2 Effect of PCA on Model Outputs

PCA use was positively affecting model performance. For Random Forest, PCA reduced accuracy slightly, meaning Random Forest is aided by being able to utilize the full complement of original features. This is because Random Forests manage redundant or irrelevant features very well. On the other hand, SVM was greatly aided by PCA since feature space reduction made easier categorically distinguishable separation of classes and the avoidance of overfitting. XGBoost also recovered well after PCA, with good performance after optimal parameter tuning. Overall, PCA reduced training time and improved performance in models sensitive to high-dimensional data (e.g., SVM), but negatively impacted models which have gained value from diversity in features (e.g., Random Forest).

7.3 Most Important Features on Predictions

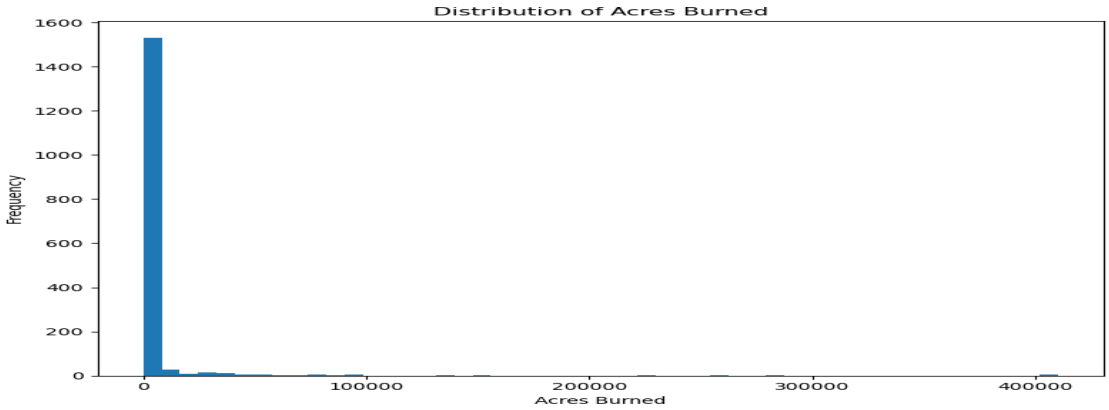
In Random Forest and XGBoost models, feature importance was extracted to identify which features contributed most to predictions of wildfire risk. AcresBurned, FireCause, StartYear, and County were always leading features. AcresBurned was the leading

predictor, naturally, as the target variable (RiskCategory) was induced directly from it. FireCause was another leading feature, adding knowledge of how human activity or natural process relate to wildfire severity. Geometric features such as Longitude and Latitude also did something useful with the models to assist them in identifying location-based fires. Overall, Random Forest on base data performed better in accuracy, reflecting the strengths of dealing with complex feature sets.

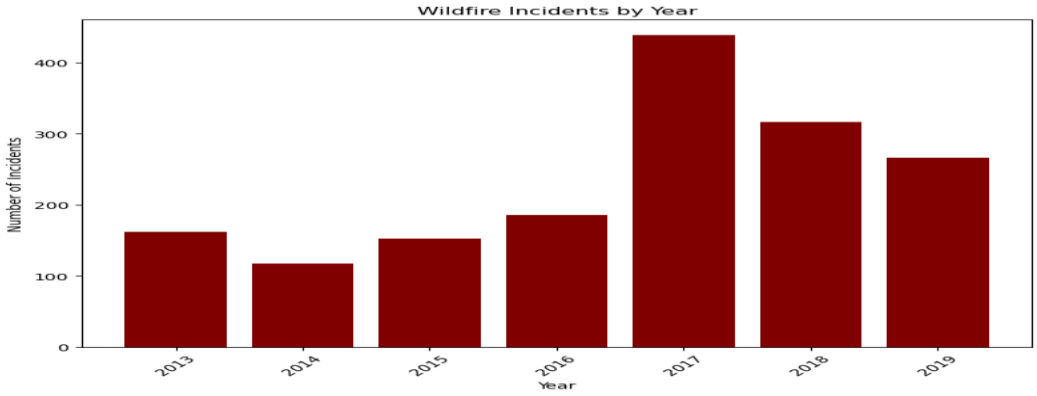
PCA significantly enhanced performance of models like SVM and XGBoost with better training time and performance for those algorithms extremely sensitive to high-dimensional data. The most effective features across models were AcresBurned, FireCause, StartYear, and County, presenting essential information about wildfire behavior. These results reinforce the importance of both robust model selection and prudent feature construction in the development of predictive models of wildfire hazard management.

8. Visualizations:

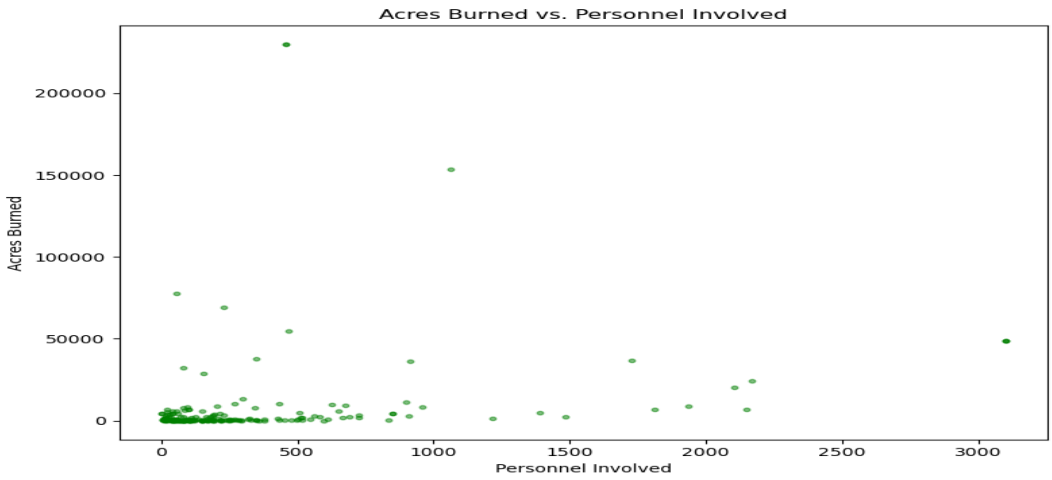
1. Histogram of Acres



2. Wildfire incidents by year



3. Acres burned vs personnel involved



9. Insights and Discussion

9.1 Strengths and Weaknesses of Each Model

Each machine learning model had its own strengths and weaknesses throughout the wildfire risk forecasting task. Random Forest worked beautifully with great accuracy and robustness using the entire set of features without requiring a lot of feature engineering and dealing excellently with complex interactions between variables. It was, however, slightly longer to train compared to simpler models and its performance was worse after PCA. Support Vector Machine (SVM) fared worse on the initial dataset but improved markedly after applying PCA dimension reduction, indicating sensitivity to high-dimensional inputs and benevolence under cleaner inputs. XGBoost was a very robust and agile model that coped well with PCA after using good hyperparameter optimization. XGBoost still required more exertion in parameters to optimize as well as to control training for preventing overfitting. Overall, Random Forest performed the best out-of-the-box, but SVM and XGBoost needed more preprocessing to perform optimally.

9.2 Issues Encountered

Some issues were encountered while working on the project. Missing data for some important features such as AcresBurned, Engines, and AirTankers were one of the major issues. Missing target values were handled by removing rows, while missing numerical values were replaced with zeros, though this would introduce a very small bias. Another problem was model overfitting, particularly for advanced models like XGBoost, in which too many trees or very deep trees would cause the model to fit the training data too well and perform poorly on new data. Further, even carrying out Principal Component Analysis (PCA) itself was challenging; deciding on the correct number of components to keep sufficient variance without losing important information was a fine balancing act. Lastly, ensuring that the risk categories were well-balanced after binning was crucial to prevent bias towards the majority class. Generally, accurate preprocessing and model tuning were effective in reducing these issues.

9.3 Improvement Suggestions

There are several areas where the project can be enhanced in future work. First, acquiring a larger and even more extensive dataset would help in addressing the issue of missing data and improving model reliability. Adding more features regarding weather, humidity, wind speed, and fire department response time would make the models even predictive. Exploring more advanced algorithms like deep learning models (e.g., neural networks) may also be valuable, especially if more features are present. Advanced imputation techniques (like KNN Imputer or MICE) might be used as an alternative to mere zero-filling for preprocessing to handle missing values in a wiser manner. Finally, incorporating more evaluation metrics other than accuracy, such as Precision, Recall, and F1-score, would provide a more detailed view of model performance, particularly if future data sets would be more imbalanced.

10. Conclusion

This project included the development of a machine learning model to classify wildfire risk based on historical fire incident data of California. Systematic pre-processing of the data—handling missing values, removing out-of-scope columns, and encoding categorical variables enabled us to obtain clean and structured input for predictive modeling. Feature engineering was instrumental in building the Risk Category label, enabling us to transform an unmanageable regression task into a more tractable multi-class classification problem.

Dimensionality reduction using Principal Component Analysis (PCA) was applied to decrease the feature space while keeping the majority of the information for computational efficiency. Random Forest, Support Vector Machine (SVM), and XGBoost machine learning models were trained and validated to evaluate their ability to predict classes of wildfire risk. Among the four, Random Forest had the best performance when trained on the raw data, and XGBoost had the best performance after PCA had been applied beforehand, demonstrating the strength of coupling ensemble learning with feature reduction.

Among the significant contributions of this work are the organization of a predictive model pipeline in an orderly manner, feature engineering for allowing classification, and comparison of numerous machine learning strategies. These findings exhibit the effectiveness of data-rich methodologies in enabling the estimation of wildfire hazard risk, resource allocation, and disaster policy-making.

Future endeavors can be integrated with the dataset of real-time environmental factors such as temperature, humidity, wind direction, and drought index in order to further improve model performance. Temporal data such as seasonality or decadal fire history would add just that much more to improved predictive power. A possible application would be to formulate this model as a real-time predictive model that assesses wildfire threats dynamically based on real-time data feeds, with actionable support for emergency response planning and decision-making.

In conclusion, this project demonstrates the promise of machine learning to support disaster risk management and suggests a number of ways forward for ongoing development and practical implementation.

11. References

Scikit-learn Developers. scikit-learn developers. (2025). *scikit-learn: Machine learning in Python*. https://scikit-learn.org/stable/

Chen, T., & Guestrin, C. (2016). XGBoost developers. (2025). *XGBoost documentation*. https://xgboost.readthedocs.io/

Pandas Development Team. The pandas development team. (2025). *pandas: Python data analysis library*. https://pandas.pydata.org/

Principal Component Analysis (PCA) - Explained. Author unknown. (2025). *A one-stop shop for principal component analysis*. Towards Data Science. https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Support Vector Machine (SVM) Guide. scikit-learn developers. (2025). *Support Vector Classification*. scikit-learn. https://scikit-learn.org/stable/modules/svm.html