Term Project Deliverable 2

Dilara Albayrak
Graduate School of Building Science
Middle East Technical University
Ankara, Turkey
dilara.kececi@metu.edu.tr

***Abstract*—This study investigates two key questions related to fire management using the 2023 Yılı Yangın Müdahale İstatistiği dataset. First, it examines whether fire department arrival times differ significantly across seasons, regions, and fire causes using Shapiro-Wilk and Kruskal-Wallis tests. Second, it develops and compares machine learning models to predict fire outcomes and requires extinguishing resources based on event characteristics. After data cleaning and feature engineering, various models, including Random Forest, XGBoost, and ANN, were trained and tuned. Model performance was evaluated against baseline models using standard metrics and statistical tests. The findings support improved fire risk assessment and emergency resource planning.**

Keywords—arrival time analysis, fire-related predictions, decision support systems

I. Introduction

Fires occur at different levels of intensity in many different parts of the world every year due to different reasons, causing serious loss of life and property. Therefore, detailed examination and analysis of fire datasets, together with the lessons learned from these fires, has the power to contribute to the prevention of future fires, or to the significant reduction of loss of life and property, number of injuries, and damage caused by fires by increasing the effectiveness of fire control. Within the scope of this study, studies on fire management at various scales, from global to specific regions in the literature were examined [1][5]. Particularly decision support mechanisms appear to be quite effective in literature. In a study aiming to estimate the time periods when fire risk is high in Saskatoon, the highest risk areas, seasons, days and hours were determined by developing two functions, namely survival and hazard functions. In the study, Kaplan-Meier survival probability estimator and Cox hazard model were used [2].

Another study conducted with data from fires in Oregon between 2012 and 2023 aimed to conduct risk analysis with machine learning applications. Decision tree classifiers and a Bayesian regularized neural network were used in the study. When the results of the study were evaluated, it was determined that the age of the victims, the response time of the fire department, and the presence of working smoke or fire detectors were the most important data in predicting fire outcomes [3].

In a global study on fire severity, seven forecasting models were used in determining seasonal fire severity, including Naive forecasting, Autoregressive Integrated Moving Average models (ARIMA), Exponential smoothing (ETS), Short-term load forecasting (STLF), TBATS, Generalized linear model for count time series (TSGLM), Artificial neural networks (ANN) and Prophet [1]. In another study focusing on the Upper Colorado River basin as a more regional study, a comprehensive data set was used including various data such as weather conditions, and soil types. Features such as air temperature and vapor pressure deficit were determined to be effective predictors using Random Forest and ANN models [4].

When the studies in literature are examined, it is seen that not only the fire intensity, risky areas and time intervals but also the potential behavior of the fire are the subjects of research. In a study using Brazilian governmental open data, which is a comprehensive dataset including satellite data, machine learning models such as AdaBoost, Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Machine (SVM) were used to predict potential fire behavior [5].

In this type of comprehensive research, inferences have been made that can positively affect both resource allocation and fire response, including the estimation of the number of sources that should be used [6]. In this context, it has been concluded that studies on the time of fire response, the time intervals when fires are most likely to occur, and the prediction of fire results depending on different factors have important contributions to fire response.

II. Problem Definition

Fire incidents are critical for security, emergency resource planning, and public health at different scales, whether global or regional. Timely response and effective management and use of firefighting resources are critical to minimize fire-related damages and loss of life. However, variations in event characteristics such as location, fire cause, and fire time can lead to uncertainties that can negatively affect rapid decision-making. This study aims to investigate the temporal, spatial, and causal factors affecting fire response and outcomes using a dataset of fires that occurred in Izmir in 2023. Two research questions were determined within the scope of the study. While the first research question aims to evaluate the operational efficiency of fire response teams, the second research question focuses on fire outcomes and resource planning. Determining possible delays, regional differences, or seasonal inefficiencies in different fire departments is one of the important steps within the scope of this study. For this purpose, the main research questions are divided into sub-questions based on each specific target.

Q1: Does the average arrival time change significantly in different seasonal periods of the year, different address areas, and different fire causes?

- Q1.A: Does the average arrival time change significantly in different seasonal periods of the year?
- Q1.B: Does the average arrival time change significantly in different address areas?
- Q1.C: Does the average arrival time change significantly in different fire causes?

Q2: Is it possible to predict the outcome and resource requirements of a fire based on its cause, location, and arrival time?

- Q2.A: Is it possible to predict the outcome of a fire based on its cause, location, and arrival times?
- Q2.B: Is it possible to predict the water resource requirements of a fire based on its cause, location, and arrival time?

It is thought that being able to accurately predict the severity of a fire and its resource requirements before arrival can help speed up the necessary preparation for intervention, direct fire crews more accurately, and increase the effectiveness of resource use. Considering these research questions, it is aimed to reach meaningful results with statistical analyses and machine learning implementations with given fire dataset.

III. Data Preparation And Preprocessing

The research questions explained in the problem definition section were created based on the "2023 Fire Response Statistics" dataset. The dataset consisting of 12,986 fire incident records includes 23 features, both categorical and numeric features.

First, in order to facilitate the next stages of the study, column names were edited, and Turkish characters were replaced with English ones and changed to lower case. Then, the data types of date data, team departure time data and arrival time data were changed to appropriate ones.

In the next step, missing data for each feature were evaluated to control data quality. At this stage, since approximately 76% of the structure shape (yapi_sekli) feature was missing, it was eliminated from the dataset. Because it was thought that estimating such missing data for a categorical feature could increase bias in subsequent analyses and machine learning implementations.

For the arrival time feature (varis_suresi) with 618 missing values, median imputer was used because no significant correlation was observed when its relationship with other numerical features was examined. The highest Pearson correlation coefficient was observed as 0.023. Before this stage, outliers in the arrival time data were determined and removed from the data using the Interquartile Range (IQR) method to prevent distortion of measures of central tendency and ensure statistical processing. For the amount of water used feature (kullanilan_su_miktari_(m³)) with 3 missing values, KNN Imputer was used because a strong positive correlation was detected with other numerical feature kullanilan_kopuk_miktari_(kg) as 0.35. Thus, the missing data handling stage was completed.

In the next step, new variables were created to support temporal and severity-based analysis. First, month, and day_of_week, and were extracted from the date feature to capture temporal trends. To facilitate subsequent analysis, a season column was created by referencing examples from the literature. Then, the fire outcome (yangin_sonucu) column was recoded into an ordinal severity label with categories "low", "moderate", and "high". Aggregate features such as total_casualty and total_animal_loss were computed to quantify incident impact.

Since some features' data heavily skewed such as amount of water used (m³), amount of foam used (kg), amount of dry chemical powder used (kg), and total animal loss, logarithmic scale transformation was applied to transform data into more normally distributed way. In this point, since these data seem possible to be real, not like the arrival time outliers when compared, logarithmic scale transformation was selected instead of removing extreme values from dataset.

Before the feature selection stage, the relationships of the organized data in the dataset were evaluated with a correlation heatmap for numerical ones and the distribution of numerical variables in the dataset was examined. Then, various plots were created for categorical data. Also, a time series plot was created for historical data.

As a result of the examination of visual representations and descriptive analysis, it was seen that there is a strong positive correlation between the amount of water and foam used (0.35). The highest frequency for arrival time is 4 minutes. It was seen that the most common cause of fire is cigarettes and/or matches with a value of 5556. Fires occurred most frequently in the summer, then in the fall, and least frequently in the spring. The distribution of fire causes and fire occurrences is shown on Fig.1.

For the first feature selection, both correlation maps (Pearson Correlation Values) and a feature selection algorithm were used. To capture non-linear and interaction effects based on how often a feature splits the data, a tree-based algorithm was applied with Random Forest Classifier. In this step, to prevent data leakage, dataset divided into test (0.2) and training (0.8) data. Label Encoder was used because there was no hierarchical relationship between all categorical data and almost all features were found to be important when OneHot Encoder was used. Since the values of the data in the dataset were very different from each other, StandardScaler was used for normalization. Thus, the important features are determined. According to the Random Forest Feature Importance results, the most important features were found as fire type (0.19), amount of water used (0.11), district (0.1), fire cause (0.09), arrival time (0.08), and season (0.03). The aim is to use the selected features for statistical analysis and machine learning; another feature selection strategy was applied in Model Interpretability section.

IV. Hypothesis Testing for Arrival Time Distributions

Based on the first research question, the null hypotheses were defined as:

- There is no statistically significant difference in arrival times across season.
- There is no statistically significant difference in arrival times across address regions.
- There is no statistically significant difference in arrival times across fire causes.

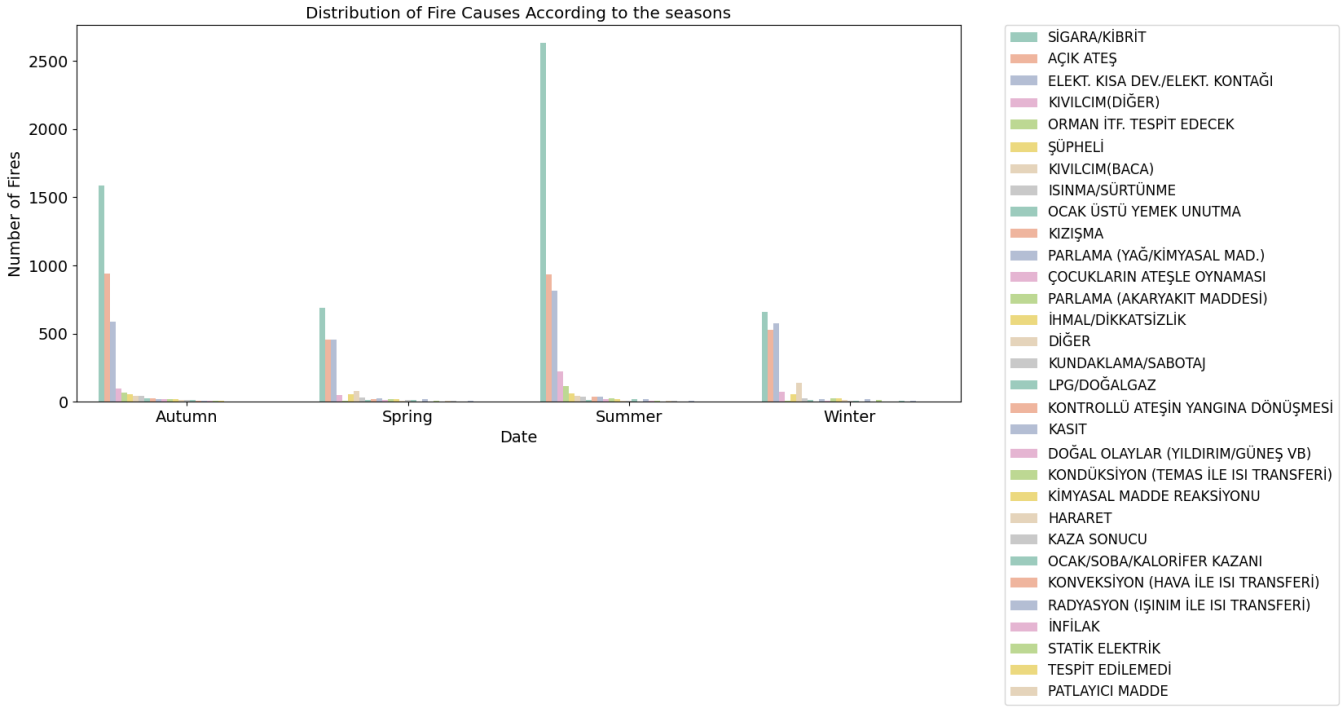

Fig. 1. Distribution of Fire Causes

To answer the first research question based on the sub questions, distribution of the variables was checked with Shapiro-Wilk test at first. According to the Shapiro-Wilk test results, none of the main groups (season, address region, fire cause) have normally distributed arrival time values except for a few with very small sample sizes.

All major groups have p-values < 0.05, and null hypothesis of normality rejected. Therefore, Kruskal-Wallis H-Test as a non-parametric alternative was used.

According to the result of Kruskal-Wallis H test, since p-values are extremely small, the null hypothesis can be easily rejected. Therefore, it can be said that there are statistically significant differences in arrival times across seasons, address regions, and fire causes.

V. DEVELOPMENT OF MACHINE LEARNING MODELS

To answer the second research question, based on the sub-questions, both classification and regression tasks were defined.

A. Classification

For the classification tasks, Random Forest Classifier, Logistic Regression, Naive Bayes, XGBoost, Dummy Classifier, k-NN and ANN models were developed. With this approach, comprehensive comparison across model families was aimed at tree-based ensemble (Random Forest, XGBoost), instance-based (k-NN), neural-based (ANN), and baseline (Dummy). Performances of models were compared in Table I.

B. Regression

A diverse set of regression models was selected for this task. Linear Regression as a simple benchmark for linear relationships, Dummy Regressor as a naive baseline, Random Forest and Gradient Boosting for capturing non-linear patterns with feature importance, XGBoost for its regularized, high-performance boosting capabilities, ANN for modeling hidden nonlinear interactions, Support Vector Regression for its robustness to outliers and its suitability for noisy real-world datasets, and k-NN for learning based on local similarity. Performances of models were compared in Table II.

TABLE I. PERFORMANCES OF CLASSIFICATION MODELS

Model	Accuracy	F1 Macro	Precision Macro	Recall Macro
Random Forest	0.8984	0.3973	0.4408	0.3835
k-NN (k=5)	0.8995	0.3748	0.4470	0.3671
XGBoost	0.9053	0.3640	0.4330	0.3574
Naive Bayes	0.9149	0.3185	0.3050	0.3333
Logistic Regression	0.9149	0.3185	0.3050	0.3333
Dummy (Baseline)	0.9149	0.3185	0.3050	0.3333
ANN	0.9149	0.3185	0.3050	0.3333

TABLE II. PERFORMANCES OF REGRESSION MODELS

Model	MAE	RMSE	R^2 Score
Gradient Boosting	0.5256	0.7567	0.1815
ANN Regressor	0.5456	0.7785	0.1336
XGBoost	0.5396	0.7851	0.1188
Linear Regression	0.5555	0.7887	0.1107
SVR	0.5216	0.7925	0.1021
Random Forest	0.5580	0.8121	0.0572
K-Nearest Neighbors	0.5650	0.8178	0.0440
Dummy Regressor	0.6127	0.8364	-0.0000

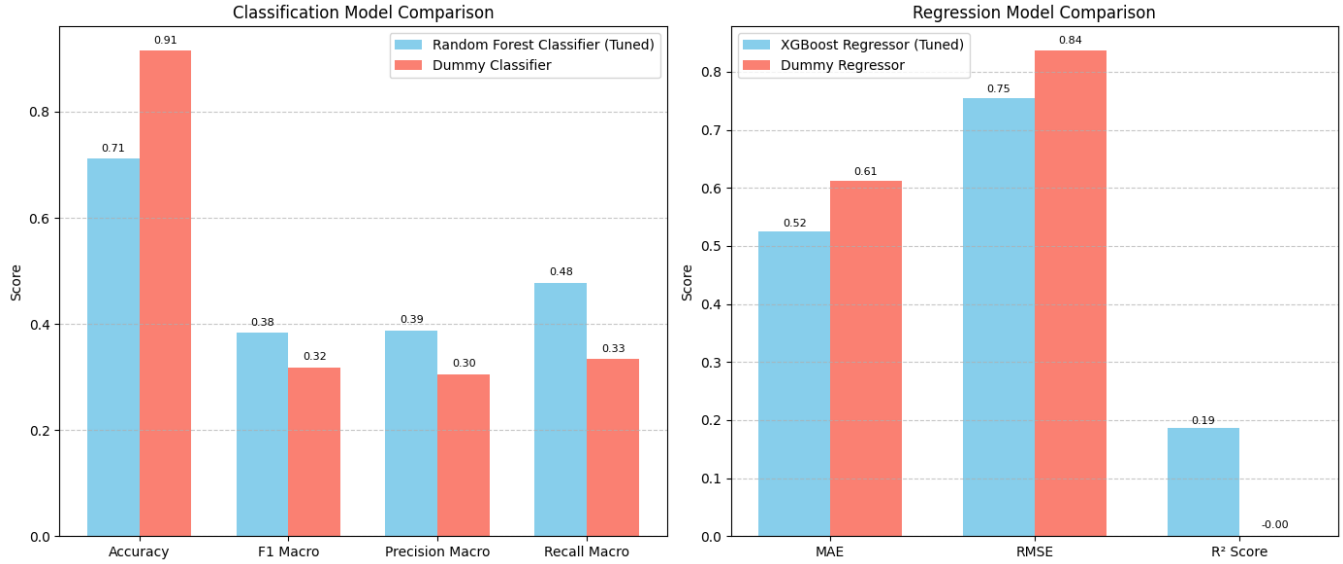

Fig. 2. Comparison of Tuned Model Performances with Baseline Model Performances

When the performances of classification models evaluated, it can be clearly seen that accuracy, F1 macro, precision macro and recall macro values are the same for Naïve Bayes, Logistic Regression, Dummy Classifier and ANN models. Their identical performances can be explained as all models predicted only the majority class (low). It that the strong class imbalance in the dataset prevented the models from learning to distinguish between the other classes.

VI. Model Optimization via Hyperparameter Tuning

In the hyperparameter tuning phase, GridSearchCV was used for regression models (Random Forest, XGBoost, and Gradient Boosting) because the parameter space was relatively small and allowed for exhaustive search. Parameters like n_estimators (number of trees) were tested with values [100, 200, 300] to evaluate how increasing the number of trees affected performance without causing excessive training time. max_depth, min_samples_split, and min_samples_leaf were selected to control tree complexity and prevent overfitting. For boosting models, learning_rate, subsample, and colsample_bytree were included to tune learning dynamics and tree diversity. For classification tasks, RandomizedSearchCV was used to cover a larger space more efficiently. For Logistic Regression, only lbfgs and liblinear solvers were tested since they support the chosen penalty type (l2) and are suitable for small to medium-sized datasets; the third option, sag, was excluded due to its sensitivity to feature scaling and slower convergence in some cases. For k-NN, values for n_neighbors and distance metrics were selected to see how local versus global decision boundaries performed. ANN tuning included different layer sizes, activation functions (tanh, relu), and learning strategies (constant, adaptive) to balance learning capacity and convergence. All models were evaluated using 3-fold cross-validation to assess generalization and reduce variance due to data splits. Three-fold cross-validation was chosen as a balance between computational efficiency and reliable performance estimation during hyperparameter tuning across multiple models. Thus, it was aimed to prevent overfitting and predict the generalization ability of the models. Cross-validation acted as the validation approach for hyperparameter tuning, internally. The hyperparameter values were chosen based on prior empirical studies and practical constraints such as runtime and model interpretability. Based on the results, the best tuned regression model was XGBoost Regressor, and the best tuned classification model was Random Forest Classifier as seen in Fig.2.

VII. Model Interpretability

To interpret how input features influence model predictions, both feature importance and permutation importance were applied. Since feature importance may introduce bias while offering fast results, permutation importance was used as a robust alternative that evaluates the impact of each feature by measuring the performance drop when it is shuffled. According to the results, the classification model primarily relies on fire cause to predict outcome severity. District and arrival time follow, while address region has the lowest contribution. When permutation importance was evaluated, fire causes still have the highest impact on F1 macro score (0.06). The other features show close to zero or even slightly negative impact, suggesting they might not improve prediction significantly beyond noise.

The regression model on the other hand, primarily relies on address region, contributing over 70% of the model's learned importance. Fire causes (0.12) and arrival time (0.11) seem moderately important. District (0.05) is the least used feature in tree splits. When permutation importance was evaluated, fire cause has the largest actual impact on predictive performance (0.1). Address regions follow (0.08), which aligns with the tree-based view but not as dominantly. District (0.03) and arrival time (0.02) have a mild but positive impact. While XGBoost's tree structure prioritized address region, actual model performance (via permutation) suggests fire cause is more critical for accurate generalization. This highlights the difference between model-internal decision paths.

As mentioned in the Introduction section, in the literature, predicting fire outcome as a research question is quite common. When the outputs were evaluated in the context of feature importance, results of this study supports the arrival time, seasonal changes and address region (risky areas or the

distance based) are important to predict fire outcomes. But their importance can be change across different data.

VIII. Performance Evaluation Against Baseline Models

When tuned Random Forest Classifier and XGBoost Regressor models were compared with baseline models, the dummy (baseline) model shows higher accuracy (0.91) then the Random Forest Classifier (0.71) as seen in Fig. 2. However, this can be misleading due to class imbalance since the dummy model always guesses the majority of class.

Based on F1 Macro, Precision Macro, and Recall Macro scores, the Random Forest Classifier significantly outperforms the dummy model. It can be interpreted as the Random Forest model is clearly superior in capturing all classes fairly while the dummy classifier is only good at the majority class.

When performances of regression models are evaluated, XGBoost model significantly reduces both absolute and squared errors, and indicates it captures some of the variance in the target. In contrast, the dummy model seems not explanatory ($R^2 \approx 0$).

To check the tuned models' superiority to the baseline models, for the regression task, Mann-Whitney U Test was conducted since it is appropriate ordinal or continuous data and does not require the normality in data**Error! Reference source not found.**. Based on the p-value (0.001), null hypothesis was rejected. Thus, the superiority of tuned models to the baseline models (dummy) was statistically proved.

IX. Conclusion and Future Work

In the scope of this study, the dataset was systematically cleaned, preprocessed, and analyzed to prepare for subsequent statistical and machine learning applications. Missing values were handled using appropriate imputation strategies, according to the variable distributions and correlations. Outliers in critical variables such as arrival time were removed using the IQR method to ensure data integrity. New features were created to capture temporal, geographic, and severity-related dimensions of fire incidents. For this purpose, date, and fire outcome features were used. Preliminary exploratory analysis and feature selection methods provided insights into the most informative variables for modeling tasks. These foundational steps have created a structured dataset, ready for statistical analysis and predictive modeling in the next phase.

To answer the first research question, two statistical analyses were conducted, as Shapiro-Wilk for normality check and Kruskal–Wallis H for testing group differences. Results indicated statistically significant differences in arrival times across seasons, address regions, and fire causes.

The second research question was divided into classification and regression tasks. For the classification task, the best-performing model was the Random Forest Classifier, achieving an F1-macro score of 0.3973. For regression, the Gradient Boosting Regressor yielded the best results, with an MAE of 0.5256, RMSE of 0.7567, and R^2 score of 0.1815.

Hyperparameter tuning was applied to enhance model performance. RandomizedSearchCV was used for classification models (Random Forest, XGBoost, Logistic Regression, k-NN, and ANN), while GridSearchCV was used for regression models (Random Forest, XGBoost, and Gradient Boosting), employing predefined parameter grids. A 3-fold cross-validation strategy was adopted to ensure robust evaluation. The objective was to minimize prediction errors in regression tasks and improve F1-macro scores in classification tasks. Results confirmed that the tuned models outperformed their baseline counterparts, validating the effectiveness of hyperparameter tuning. The best-tuned regression model was the XGBoost Regressor, and the best-tuned classification model remained the Random Forest Classifier.

Interpretability was assessed using feature importance and permutation importance methods. The classification model heavily relied on the fire cause variable, while the regression model primarily depended on the address region feature, which accounted for over 70% of the model's predictive importance.

Finally, a comparison between baseline and tuned models showed that the tuned models significantly outperformed baselines in most metrics, except for accuracy in the classification task, likely due to class imbalance in the dataset.

The results of the study provide meaningful answers to both research questions. For the first question, statistical tests showed that arrival times significantly vary across seasons, address regions, and fire causes. This means that when and where a fire happens, as well as what caused it, can affect how quickly teams arrive. For the second question, classification models like Random Forest were able to predict fire outcomes with moderate success using cause, location, and arrival time. Regression models for predicting water use had weaker results, although Gradient Boosting performed best among them. These findings suggest that machine learning can help estimate fire outcomes and resource needs, but better results may be possible with more detailed data or additional features.

Based on the analysis of the 2023 Fire Response Statistics dataset, this study provides valuable insights for data-driven decision-making in fire management. By examining variables such as fire cause, location, seasonal distribution, and arrival time, the study contributes to three critical areas: fire prevention, risk assessment, and resource planning.

From a fire prevention perspective, identifying seasonal and regional patterns in fire causes allows local governments to implement targeted awareness campaigns and adjust control strategies accordingly.

From a risk assessment perspective, the integration of classification models to predict fire outcomes based on location, cause, and arrival time helps prioritize incidents that are more likely to have serious consequences. For example, if both datasets show that fires in densely populated urban areas tend to escalate more rapidly, municipal planning can focus on response station proximity in these areas. Finally, regression analysis predicting water usage enables more efficient resource allocation.

Overall, consistent variables and model performance across the dataset support the generalizability of the findings and suggest that the models can be extended to future datasets for ongoing monitoring and adaptive planning.

REFERENCES

[1] L. N. Ferreira, D. A. Vega-Oliveros, L. Zhao, M. F. Cardoso, and E. E. N. Macau, "Global fire season severity analysis and forecasting," *Computers & Geosciences*, vol. 134, p. 104339, Jan. 2020, doi: https://doi.org/10.1016/j.cageo.2019.104339.

[2] G. Yeboah and P. Y. Park, "Using survival analysis to improve pre-emptive fire engine allocation for emergency response," *Fire Safety Journal*, vol. 97, pp. 76–84, Mar. 2018, doi: 10.1016/j.firesaf.2018.02.005.

[3] A. Schmidt, E. Gemmil, and R. Hoskins, "Machine Learning Based Risk Analysis and predictive Modeling of Structure Fire related Casualties," *Machine Learning With Applications*, p. 100645, Mar. 2025, doi: 10.1016/j.mlwa.2025.100645.

[4] H. Han, T. A. Abitew, H. Bazrkar, S. Park, and J. Jeong, "Integrating machine learning for enhanced wildfire severity prediction: A study in the Upper Colorado River basin," *The Science of the Total Environment*, vol. 952, p. 175914, Aug. 2024, doi: 10.1016/j.scitotenv.2024.175914.

[5] J. N. S. Rubí, P. H. P. De Carvalho, and P. R. L. Gondim, "Application of machine learning models in the behavioral study of forest fires in the Brazilian Federal District region," *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105649, Dec. 2022, doi: 10.1016/j.engappai.2022.105649.

[6] H. Liz-López, J. Huertas-Tato, J. Pérez-Aracil, C. Casanova-Mateo, J. Sanz-Justo, and D. Camacho, "Spain on fire: A novel wildfire risk assessment model based on image satellite processing and atmospheric information," *Knowledge-Based Systems*, vol. 283, p. 111198, Nov. 2023, doi: 10.1016/j.knosys.2023.111198

[7] Du Prel, B. Röhrig, G. Hommel, and M. Blettner, "Choosing statistical tests: part 12 of a series on evaluation of scientific publications," *Deutsches Ärzteblatt International*, vol. 107, no. 19, pp. 343–348, 2010.