Jurnal Aplikasi Sains Data

Journal homepage: https://jasid.upnjatim.ac.id/index.php/jasid

Application of XGBoost for Risk Level Classification of Fires in Surabaya City in 2024 and Interactive Spatial Visualization Based on Streamlit

Sarah Aprilia Hasibuan[1], Divia Prisillia Prisca[2], Annita Fadhilah Aprilia[3], Dwi Arman Prasetya[4], Prismahardi Aji Riyantoko[5]

[1,2,3,4]Data Science Study Program Universitas Pembangunan Nasional "Veteran" Jawa Timur,
[5]Department of Information and Communication Okayama University Japan

[*]arman.prasetya.sada@upnjatim.ac.id

ABSTRACT

Abstract— Fire in urban areas such as Surabaya City is a non-natural disaster that can have a significant impact on public safety, economic stability, and the environment. This study aims to develop a fire risk level classification model using Extreme Gradient Boosting (XGBoost) algorithm based on selected predictor variables, namely response time, fire subtype, and number of victims affected. The dataset consists of 859 fire events throughout 2024, enriched with spatial and demographic attributes. The research methodology involved data preprocessing (including label coding and normalization), class imbalance handling with Synthetic Minority Over-sampling Technique (SMOTE), model training with XGBoost, and evaluation using metrics such as accuracy, precision, recall, and f1-score. The classification model achieved excellent performance, with an overall accuracy of 1.00% and perfect precision, recall, and f1-score of 1.00 across all risk categories (low, medium, and high). Confusion matrix and ROC curve analysis confirmed the high predictive ability of this model. In addition, the results were visualized using a Streamlit-based interactive dashboard to enhance the usability of the model for decision-making. These findings highlight the potential of XGBoost as a powerful tool for fire risk classification and emphasize its relevance in supporting early warning systems and evidence-based disaster mitigation policies in urban environments.

Keywords: fire, risk classification, XGBoost, SMOTE, response time, interactive dashboard

I. INTRODUCTION (10PT)

Fires in urban areas such as Surabaya are one of the non-natural disasters that have a significant impact on human safety, economic losses and environmental damage. With the characteristics of a densely populated city and high intensity of economic activity, Surabaya is vulnerable to fire incidents, whether triggered by human negligence, technical disturbances, or environmental factors.

Data from the Regional Disaster Management Agency (BPBD) shows that the main causes of fires generally involve electrical short circuits, gas leaks, and unsafe behavior from the community. To minimize the impact and improve preparedness, a data-driven approach is needed that can accurately predict potential risks.

In this study, the Extreme Gradient Boosting (XGBoost) method is used, an ensemble-based classification algorithm proven to excel in handling complex and imbalanced data. XGBoost is able to form accurate predictive models by utilizing spatial and demographic variables, such as incident

* *Corresponding author.*
E-mail address: arman.prasetya.sa
Doi: 10.33005/jasid.v1i2.24

location, response time, gender, age, and incident type [1]. The use of XGBoost has also been proven effective in the classification of various environmental disasters such as forest fires [2] and the prediction of geological phenomena such as earthquakes [3].

Different from traditional approaches such as logistic regression, XGBoost is able to capture not only linear relationships between variables, but also complex non-linear patterns. This makes XGBoost a relevant choice of method in the analysis of factors affecting fire risk levels and officer response speed [4].

Apart from modeling, the presentation of analysis results also plays an important role in supporting decision making. Therefore, in this research, an interactive dashboard was built using Streamlit, which allows real-time visualization of data and prediction results. One of the main features of this dashboard is an interactive map that displays the distribution of fire occurrence locations based on the model classification results. This map makes it easy for users, such as BPBD or policy makers, to identify vulnerable areas, see spatial patterns, and compare risk categories between sub-districts or villages.

Through a dynamic and user-friendly interface, the dashboard also provides filters and simulation options to change prediction parameters directly. This provides flexibility in evaluating various risk scenarios and response times, and supports the development of evidence-based emergency response and mitigation strategies.

The integration between XGBoost, spatial data, and Streamlit's interactive dashboard provides a comprehensive and applicable analytic approach. Thus, the results of this research are expected to contribute to the improvement of early warning systems, optimization of resource allocation, and the development of more adaptive and data-driven fire risk mitigation policies [5].

II. RESEARCH METHODOLOGIES

This research methodology is structured in several systematic stages to build a fire risk prediction model in urban areas, especially Surabaya, and present an interactive spatial visualization based on Streamlit. These stages include data collection, pre-processing, data exploration, modeling using XG Boost, spatial data, and Streamlit interactive dashboard.

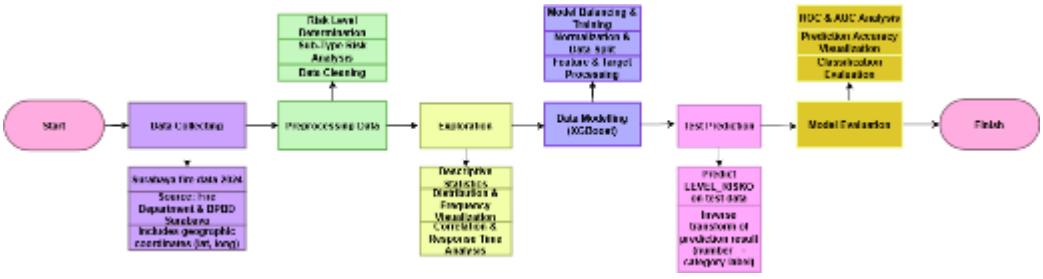

Figure 1. Research Flow

A. Data Collection

The data used in this study is data on fire incidents in Surabaya City throughout 2024 which was obtained directly from the Surabaya City Regional Disaster Management Agency (BPBD). The data used in this study is data on fire incidents in the Surabaya City area in 2024. This data was obtained from the Fire and Rescue Department of Surabaya City and BPBD Surabaya City, and consists of important attributes that reflect the characteristics of fire incidents and their locations. The data is also equipped with geographic information (latitude and longitude) to support spatial analysis. The variables contained in the dataset are as follows:

a. KECAMATAN : The sub-district level administrative area where the fire occurred. This variable is important in spatial and cluster analysis of fire-prone areas.
b. KELURAHAN: Administrative areas are more specific than sub-districts, i.e. kelurahan. This helps improve resolution in risk mapping.
c. WILAYAH : A general description of the area, which can be a classification (such as densely populated, industrial, or residential).
d. LOKASI : A specific address or more detailed point of fire occurrence, useful for geo-referencing and accurate mapping.
e. OPEN TICKET : The time or date of entry of the fire report into the fire service system.

f. WAKTU TIBA: The time when officers arrive at the scene. The difference between this time and the report time is used to calculate the response time.
g. WAKTU RESPON : The time taken from the report being received until the first response is made, usually in minutes. This can be used as an indicator of the efficiency of emergency response.
h. KORBAN TERDAMPAK : The number of people directly affected by the fire incident, both physically and materially.
i. KORBAN LUKA : Totally of people who suffered minor or major injuries as a result of the fire.
j. KORBAN MENINGGAL : The number of fatalities in a fire incident.
k. SUB JENIS : The specific type of fire incident, e.g. residential fire, vacant lot fire, or industrial fire.
l. latitude dan longitude : The geographical coordinates of the fire scene. These variables are very important in the interactive map-based spatial visualization process, as they determine the precise position of the incident point on the digital map.

B. Preprocessing

Data pre-processing is an important stage to prepare the dataset to be suitable and ready for use in time series analysis. The following steps were taken :

a. Removes duplicate values and invalid or null values.
   Response Time Conversion, Response time columns that were originally text were converted to numeric format in seconds. This allows the data to be used in the analysis and classification of risk levels based on the speed of officers in responding to fire incidents.
b. Calculation of Average Casualties per Sub-Type of Event
   Data is grouped by sub-type of fire, such as residential, vacant land or industrial fires, and then the average number of casualties affected in each sub-type is calculated. This average value represents the severity of each type of fire.
c. Risk Grouping Based on Event Subtypes
   Based on the calculated average casualty values, the event subtypes were classified into three risk categories: low, medium and high. This grouping was done statistically by dividing the data into three parts based on quartiles.
d. Merging Risk Information into the Main Dataset
   The results of the risk classification per sub-type were then merged into the main dataset, so that each row of fire event data had the additional attribute of a risk category based on the type of event.
e. Risk Level Classification of Each Event
   To determine the risk level of each fire event, information from several aspects is combined, namely the number of victims affected, sub-type risk categories, and officer response time. Using a certain classification logic, each event is categorized into low, medium or high risk levels.

C. Exploration

In the data exploration stage, the aim is to understand the distribution, patterns and relationships between variables in the fire dataset. This process is very important to gain initial insight into the characteristics of the data and help in determining the right modeling strategy. The explorative steps taken include:

a. Descriptive Statistics of Dataset : The first step is to look at the descriptive statistics of all numeric columns in the dataset, such as the maximum, minimum, mean, and standard deviation values. This is useful to determine the range of values and the distribution of data, as well as to identify possible outliers or extreme values.
b. Officer Response Time Distribution : A histogram visualization with a density curve is used to see the distribution pattern of response time (in seconds). From this graph it can be seen whether the response time tends to be normal, skewed to one side, or has a wide spread. This gives an idea of the extent of variation in the response speed of firefighters in responding to incidents.
c. Number of Incidents by Subdistrict : A visualization of a horizontal bar chart was conducted to display the number of fire cases by subdistrict. This helps identify which administrative areas experience the most fires, which can then be the focus of risk maps

or policy interventions.

d. Number of Incidents by Fire Subtype : This visualization shows the frequency of incidents by fire type, such as residential, warehouse or vacant land fires. This information is important to understand the most frequent types of fires and allows further analysis of their characteristics.
e. Average Response Time per Risk Level : Bar charts are used to illustrate the relationship between response time and risk level (Low, Medium, High). This visualization aims to see if slower response times correspond to higher risk levels, thus providing additional insight in the validation of the risk classification model.
f. Correlation Map between Numerical Variables : Heatmaps are used to illustrate the correlation between numerical variables in a dataset. The colors and numbers in the heatmap indicate the strength and direction of the relationship between variables, for example between the number of casualties, response time, and other variables. This helps in determining which features are relevant for use in modeling.

D. Data Modeling

Modeling Model evaluation includes data preparation, feature processing, and predictive model training using the XGBoost algorithm. XGBoost works by forming an ensemble model consisting of a number of decision trees, where each tree is incrementally added to reduce the prediction error generated by the previous tree. This process follows the gradient boosting method, which gradually minimizes the error through the contribution of each tree.

a. XGBoost Model Training: The XGBoost model was trained using the resampled training data. This algorithm was chosen because it is known to provide high accuracy and is efficient in handling numerical and unbalanced data. XGBoost or Extreme Gradient Boosting is a highly efficient and accurate decision tree-based boosting method. Basically, XGBoost builds the model incrementally by adding new trees that aim to correct the prediction error of the previous model. Each iteration of XGBoost estimates a new predicted value using the first and second derivatives of the loss function, which makes it highly optimized for both classification and regression tasks. Mathematically, XGBoost minimizes the following objective function $L(\phi) = \Sigma_{\mathrm{i}=1}^{\mathrm{n}} l\left(y_i, \hat{y}_i^{(t)}\right) + \Sigma_{k=1}^{t} \Omega(\mathrm{f_k})$.
b. Encoding categorical variables: The categorical variables SUB JENIS and LEVEL_RISKO were converted into numerical representations using the Label Encoding technique. This is so that the variables can be processed in a machine learning model, which can only handle numerical data.
c. Feature selection and Targets: The three main features used in the modeling are; response time duration in seconds, fire type in numerical form, and number of victims affected. The prediction target is LEVEL_RISIKO_ENC, which is the encoded fire risk level.
d. Feature normalization: Feature data is normalized using the StandardScaler method, so that all features have a mean of zero and a standard deviation of one. Normalization is important to avoid the dominance of one feature in the model.
e. Training and Test Data Split: The data was divided into 80% for training and 20% for testing using the stratified split technique, in order to keep the class distribution balanced in the training and test data.
f. Data Balancing with SMOTE: Since the dataset showed class imbalance in the target variable, the SMOTE (Synthetic Minority Over-sampling Technique) technique was used to oversample the minority class only in the training data. This is so that the model is not biased towards the majority class.

F. Test Prediction

After the model training process is complete, the next step is to make predictions on the test data to assess the model's ability to cluster areas based on disaster risk levels. The model produces predictions in the form of numerical labels for the LEVEL_RISIKO variable, indicating categories such as “low”, ‘medium’, or “high” depending on the clustering results. To facilitate interpretation, these numerical labels are then converted back to their original categorical form using the inverse transform function of the encoder used in the preprocessing stage. This transformation is important so that the prediction results can be understood more clearly and can be compared directly with the original labels on the test data, making the evaluation of model performance more accurate and informative.

III. RESULTS AND DISCUSSIONS

Researchers analyzed fire data for Surabaya City in 2024 using the pandas and numpy libraries, along with visualizations created using matplotlib and seaborn. This analysis aims to classify the risk level of fire events based on response time, fire sub-types, and the number of victims. The fire dataset used consists of 859 rows and 15 columns, which contain information about the time of the incident, location, type of fire, number of victims, and response time. The columns used include OPEN TICKET, WAKTU TIBA, and WAKTU RESPON to calculate the duration of the officer's response, the columns of KORBAN TERDAMPAK, KORBAN LUKA, and KORBAN MENINGGAL to measure the impact of the incident, and SUB JENIS to categorize the type of object that caught fire. In addition, the data is also equipped with latitude and longitude coordinate information used for spatial analysis and visualization purposes on the Streamlit-based interactive dashboard.

	TANGGAL	JENIS KEJADIAN	KECAMATAN	KELURAHAN	WILAYAH	LOKASI	OPEN TICKET	WAKTU TIBA	WAKTU RESPON	KORBAN TERDAMPAK	KORBAN LUKA	KORBAN MENINGGAL	SUB JENIS	latitude	longitude
0	1-Jan-24	Kebakaran	Bubutan	Jepara	Surabaya Pusat	(Asrama AD) Dupak Masigit XII,	6:48:10	6:54:12	0:06:02	90.0	0.0	0.0	Bangunan Selain Hunian	-7.249170	112.750830
1	3-Jan-24	Kebakaran	Kenjeran	Tanah Kali Kedinding	Surabaya Utara	Jl. Pogot 105	8:35:54	8:41:50	0:05:56	2.0	1.0	0.0	Bangunan Selain Hunian	-7.232647	112.770340
2	4-Jan-24	Kebakaran	Sukolilo	Klampis Ngasem	Surabaya Timur	Jl. Klampis Ngasem	4:27:04	4:33:00	0:05:56	2.0	0.0	0.0	Hunian	-7.288672	112.777086
3	4-Jan-24	Kebakaran	Tenggilis Mejoyo	Kendangsari	Surabaya Timur	Jl. Tenggilis Barat No. 18	11:25:19	11:29:18	0:03:59	2.0	0.0	0.0	Bangunan Selain Hunian	-7.321486	112.750010
4	6-Jan-24	Kebakaran	Wonocolo	Siwalankerto	Surabaya Selatan	Jl. Kutisari Utara VII No 42	1:05:36	1:12:00	0:06:24	0.0	0.0	0.0	Jaringan Kabel	-7.333983	112.742276
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
854	26-Dec-24	Kebakaran	Tambaksari	Gading	Surabaya Timur	Jl. Kenjeran No.422	13:49:10	13:55:00	0:05:50	1.0	0.0	0.0	Bangunan Selain Hunian	-7.248090	112.778315
855	26-Dec-24	Kebakaran	Wonokromo	Wonokromo	Surabaya Selatan	Stasiun wonokromo	18:04:00	18:10:00	0:06:00	0.0	0.0	0.0	Kendaraan	-7.301928	112.739095
856	29-Dec-24	Kebakaran	Mulyorejo	Kejawan Putih Tambak	Surabaya Timur	Jl. Raya Laguna KJW Putih Tambak caffe objek c...	6:49:56	6:54:11	0:04:15	0.0	0.0	0.0	Elektronik	-7.271760	112.816047
857	30-Dec-24	Kebakaran	Pabean Cantian	Krembangan Utara	Surabaya Utara	Muteran Gg. IV,	6:00:26	6:07:09	0:06:43	3.0	0.0	0.0	Hunian	-7.249170	112.750830
858	31-Dec-24	Kebakaran	Sukomanunggal	Simomulyo	Surabaya Barat	Jl. Simo Pomahan Baru Barat IV	14:14:03	14:21:00	0:06:57	2.0	0.0	0.0	Elektronik	-7.259961	112.705109

Figure 2. Fire Data of Surabaya City 2024

After the data is ready for use, the researcher performs a pre-processing stage by using the pandas library to convert the duration of the response time into a numeric format in seconds using the 'pd.to_timedelta' function, then the results are stored in a new column called RESPON_DETIK. Next, data was grouped based on SUB JENIS of fire to calculate the average number of victims affected. Based on the average value, the risk level category was determined using the discretization technique with 'pd.qcut' into three risk levels, namely Low, Medium and High. These risk categories were then merged back into the main data through the merge process.

The next step is to create a new column LEVEL_RISIKO to store the classification results of the fire risk level. This classification process considers three main aspects, namely the number of victims affected, sub-type risk categories, and response time in seconds. The determination of the risk label is done with a conditional logic function, where the final risk value is assigned as High, Medium, or Low based on the combination of the three variables. This pre-processing results in final data that is ready to be used for further analysis and visualization.

	TANGGAL	JENIS KEJADIAN	KECAMATAN	KELURAHAN	WILAYAH	LOKASI	OPEN TICKET	WAKTU TIBA	WAKTU RESPON	KORBAN TERDAMPAK	KORBAN LUKA	KORBAN MENINGGAL	SUB JENIS	latitude	longitude	RESPON_DETIK	RISIKO_SUB_JENIS	LEVEL_RISIKO
0	1-Jan-24	Kebakaran	Bubutan	Jepara	Surabaya Pusat	(Asrama AD) Dupak Masigit XII,	6:48:10	6:54:12	0:06:02	90.0	0.0	0.0	Bangunan Selain Hunian	-7.249170	112.750830	362.0	Tinggi	Tinggi
1	3-Jan-24	Kebakaran	Kenjeran	Tanah Kali Kedinding	Surabaya Utara	Jl. Pogot 105	8:35:54	8:41:50	0:05:56	2.0	1.0	0.0	Bangunan Selain Hunian	-7.232647	112.770340	356.0	Tinggi	Sedang
2	4-Jan-24	Kebakaran	Sukolilo	Klampis Ngasem	Surabaya Timur	Jl. Klampis Ngasem	4:27:04	4:33:00	0:05:56	2.0	0.0	0.0	Hunian	-7.288672	112.777086	356.0	Tinggi	Sedang
3	4-Jan-24	Kebakaran	Tenggilis Mejoyo	Kendangsari	Surabaya Timur	Jl. Tenggilis Barat No. 18	11:25:19	11:29:18	0:03:59	2.0	0.0	0.0	Bangunan Selain Hunian	-7.321486	112.750010	239.0	Tinggi	Sedang
4	6-Jan-24	Kebakaran	Wonocolo	Siwalankerto	Surabaya Selatan	Jl. Kutisari Utara VII No 42	1:05:36	1:12:00	0:06:24	0.0	0.0	0.0	Jaringan Kabel	-7.333983	112.742276	384.0	Sedang	Rendah
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
854	26-Dec-24	Kebakaran	Tambaksari	Gading	Surabaya Timur	Jl. Kenjeran No.422	13:49:10	13:55:00	0:05:50	1.0	0.0	0.0	Bangunan Selain Hunian	-7.248090	112.778315	350.0	Tinggi	Sedang
855	26-Dec-24	Kebakaran	Wonokromo	Wonokromo	Surabaya Selatan	Stasiun wonokromo	18:04:00	18:10:00	0:06:00	0.0	0.0	0.0	Kendaraan	-7.301928	112.739095	360.0	Sedang	Rendah
856	29-Dec-24	Kebakaran	Mulyorejo	Kejawan Putih Tambak	Surabaya Timur	Jl. Raya Laguna KJW Putih Tambak caffe objek c...	6:49:56	6:54:11	0:04:15	0.0	0.0	0.0	Elektronik	-7.271760	112.816047	255.0	Sedang	Rendah
857	30-Dec-24	Kebakaran	Pabean Cantian	Krembangan Utara	Surabaya Utara	Muteran Gg. IV,	6:00:26	6:07:09	0:06:43	3.0	0.0	0.0	Hunian	-7.249170	112.750830	403.0	Tinggi	Sedang
858	31-Dec-24	Kebakaran	Sukomanunggal	Simomulyo	Surabaya Barat	Jl. Simo Pomahan Baru Barat IV	14:14:03	14:21:00	0:06:57	2.0	0.0	0.0	Elektronik	-7.259961	112.705109	417.0	Sedang	Sedang

Figure 3. Risk Level Classification Result

After pre-processing and classification of risk levels based on the number of victims, fire sub-types, and response times, the data is ready for further analysis through visual data exploration (EDA). This stage aims to understand the characteristics of the distribution of events, response time patterns, and the distribution of fire locations throughout 2024. One of the initial visualizations presented is the

distribution of response times as shown in the following figure.

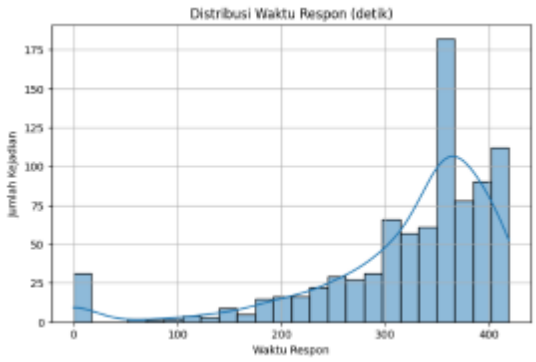

Figure 4. *Response Time Distribution*

The figure represents the distribution of response times to fire incidents in Surabaya for the year 2024. The majority of incidents had response times between 300 and 420 seconds (approximately 5 to 7 minutes), with the peak of the distribution being around 360 seconds. This suggests that most incidents are dealt with relatively quickly, although there are also a number of incidents with very short or slow responses. This distribution gives an initial idea of the efficiency of response times and is an important indicator in the risk classification of incidents.

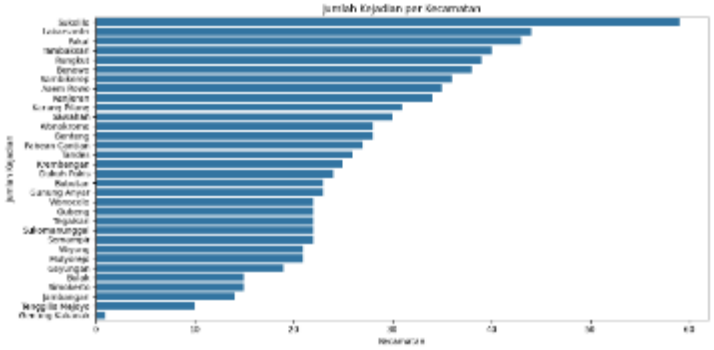

Figure 5. *Number of Occurrences per Sub-district*

Based on the visualization results above, it is known that Sukolilo sub-district recorded the highest number of fire incidents during 2024, with 59 incidents. This is followed by Lakarsantri and Pakal sub-districts with 44 and 43 incidents respectively. On the other hand, Genting Kalianak sub-district has the lowest number of incidents, with only 1 incident. This uneven fire distribution pattern shows the urgency of a location-based approach in prioritizing and formulating fire mitigation policies.

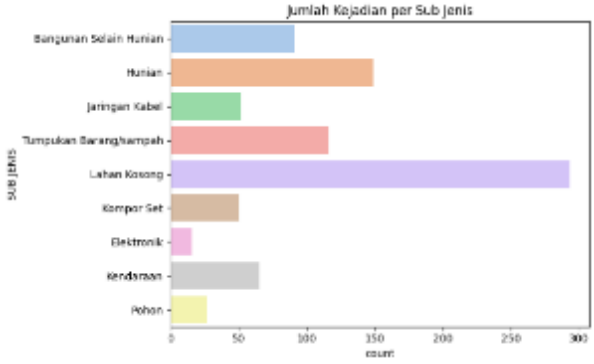

Figure 6. *Number of Occurrences per Sub type*

Based on the visualization of the number of fire incidents by sub-type, it is known that Empty Land

is the category with the highest frequency, with 294 incidents throughout 2024. Other sub-types that also dominate are Occupancy with 147 incidents and Piles of Goods/Waste with 115 incidents. Meanwhile, the category with the lowest number of occurrences is Electronics with only 18 occurrences. This pattern shows that open areas such as vacant land and residential neighborhoods are the main hotspots that require more attention in fire prevention strategies.

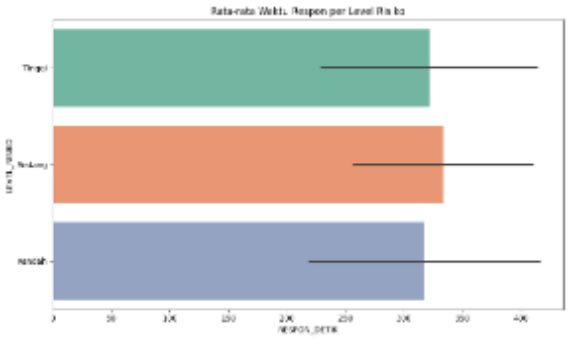

Figure 7. *Average Response Time Per Risk Level*

Based on the visualization of the average response time per risk level, it is observed that the Medium risk category exhibits the highest average response time, followed by High and Low which have relatively close average response time values. All three risk levels show an average response time above 300 seconds, with considerable variation within each category. This pattern indicates that response time is not only influenced by the risk level, but also possibly other factors such as the location of the incident and accessibility, so optimizing response at all risk levels remains an important challenge in fire management efforts.

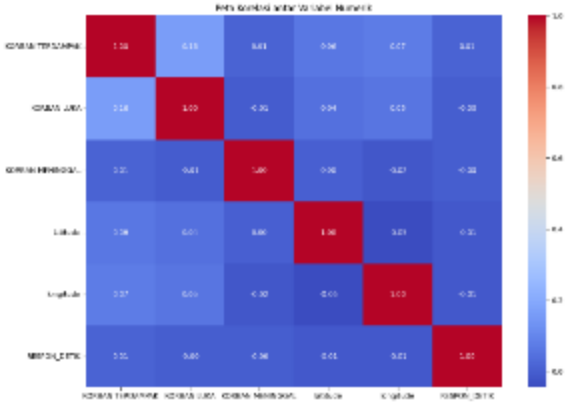

Figure 8. *Correlation Map*

Based on the correlation map between numerical variables, it can be seen that the relationship between the observed variables tends to be very weak. The highest correlation exists between the KORBAN TERDAMPAK and the ORBAN LUKA with a value of 0.16, while the correlation between the other variables is close to zero on average. These low correlation values indicate that there is no strong linear relationship between the number of victims, the location of the incident (latitude and longitude), or the response time (RESPON_DETIK). However, since the risk level classification was predetermined in the preprocessing stage through a categorical approach, this correlation analysis serves more to strengthen the understanding of the characteristics of each numerical variable.

Therefore, to capture non-linear patterns that are not detected by simple correlation analysis, this study utilizes the Extreme Gradient Boosting (XGBoost) algorithm as a more sophisticated classification approach capable of identifying complex relationships between variables. XGBoost is an ensemble-based classification method capable of efficiently identifying non-linear and complex patterns between variables. The algorithm combines a number of decision trees incrementally and features regularization to prevent overfitting and handle imbalanced data.

LEVEL_RISIKO

```
Rendah   505
Sedang   280
 Tinggi    74
```

In this study, XGBoost is used to predict fire LEVEL_RISKO based on numerical and spatial variables, such as number of casualties, response time, and incident coordinates. Next, data pre-processing was performed to prepare the features and targets to be used in model training. First, label encoding is performed on the SUB JENIS and LEVEL_RISKO columns using LabelEncoder, because machine learning algorithms such as XGBoost can only process numeric data. The encoding results are stored in the new columns SUB_JENIS_ENCODED and LEVEL_RISIKO_ENC. Then, three variables are determined as predictor features (RESPON_DETIC, SUB_JENISTS_ENCODED, and KORBAN TERDAMPAK) which are stored in the X variable. Meanwhile, LEVEL_RISIKO_ENC is used as the classification target or label, and is stored in the y variable. This process is important as the first step in building a classification model to predict the risk level of fire events.

```
              precision    recall  f1-score   support

      Rendah       0.99      1.00      1.00       101
      Sedang       1.00      0.99      1.00       101
      Tinggi       1.00      1.00      1.00       101

    accuracy                           1.00       303
   macro avg       1.00      1.00      1.00       303
Weighted avg       1.00      1.00      1.00       303
```

Next, data normalization is performed using StandardScaler to equalize the scale between features, so that the model can work more optimally. Next, the SMOTE (Synthetic Minority Over-sampling Technique) technique was applied to balance the amount of data in each risk level category (LEVEL_RISIKO) class (Low, Medium, and High). SMOTE adds synthetic data to classes that have less data, so that the model is not biased towards the majority class. After the data was balanced, the data was divided into training and test data with a proportion of 80:20 using stratification to keep the class distribution balanced. The classification model is then built using the XGBoost algorithm (XGBClassifier), which is trained on the normalized and SMOTE data. The model evaluation results are shown in the classification report, which shows that the model has a very high performance, with precision, recall, and f1-score values of 1.00 for all classes (Low, Medium, and High). This shows that the model is able to classify the test data very accurately.

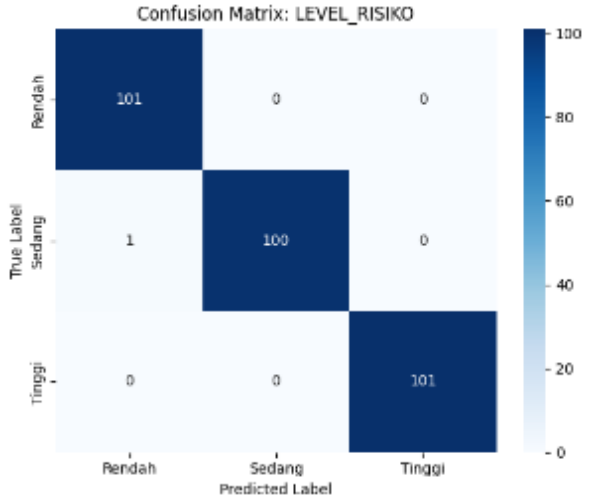

Figure 9. *Risk Level Heat Map*

The confusion matrix in Figure 4 is visualized using heatmaps from the seaborn library to show the results of the model evaluation against the risk level category (LEVEL_RISIKO) which consists of

three classes: Low, Medium, and High. The predicted label (y_pred) is returned to its original form using inverse_transform, then compared with the actual label (y_test) using confusion_matrix. As a result, the model successfully classified 302 out of 303 test data correctly, with only one error in the Medium class predicted as Low. This visualization in Figure 4 shows that the model has very high accuracy and is able to distinguish each class very well.

Calculating and displaying the ROC Curve in the case of multi-class classification. First, the original label (y_test) is binarized with label_binarize as ROC requires a binary format. Then, the model calculates the prediction probability (predict_proba) for each class.

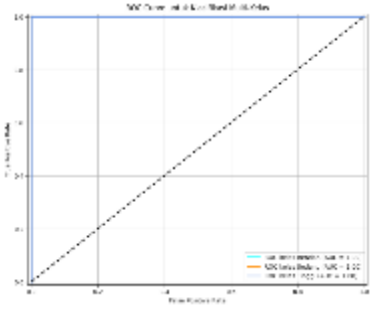

Figure 10. *ROC Curve*

Next, ROC curve and AUC (Area Under Curve) values were calculated for each class using roc_curve and auc. The results are visualized in the ROC graph shown in figure 5 which displays the lines for each class in different colors. This ROC curve shows the model's ability to distinguish between classes-the higher the AUC (closer to 1), the better the model performs in classifying that class.

We created the Surabaya Fire Dashboard 2024 which is an interactive visualization system designed to display the results of the analysis of fire incidents in Surabaya City throughout 2024. This dashboard presents important information such as total incidents, number of victims affected, victims injured, and victims died. In addition, users can filter the data by time range, fire subtype, and subdistrict to obtain more specific information. An interactive map is also provided to display the geographical distribution of incident locations, making it easier to identify fire-prone areas.

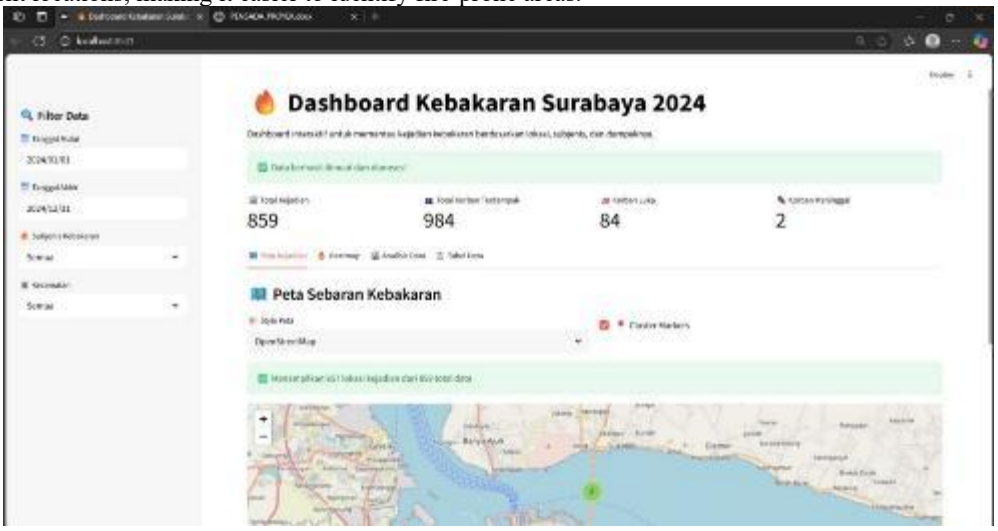

Figure 11. *Dashboard*

a. Dashboard Title, The Surabaya Fire Dashboard 2024 is an interactive display designed to monitor fire incidents in Surabaya City for the year 2024. This dashboard presents information based on location, fire subtype, and impact on the community.
b. Data Status, The system displays a notification that the data has been successfully loaded and processed, indicating that the dashboard is ready to be used for data exploration and analysis.
c. Summary Statistics, The dashboard presents summary data as follows :

- Total Events: 859
- Total Victims Affected: 984 people
- Injured Victims: 84 people
- Death Victims: 2 people

This information provides an overview of the frequency of fire incidents and their impact on fatalities.

d. Data Filters, The panel on the left allows users to filter data by :
   - Start Date and End Date (default: January 1 - December 31, 2024)
   - Fire Subtype (dropdown option to show all or specific type)
   - District (dropdown option to show all or specific district)

e. Fire Distribution Map, The main section of the dashboard displays an interactive map showing
   the location of the fire incident :
   - Users can choose the map display style (e.g. OpenStreetMap)
   - The Cluster Markers option is enabled to group locations for easier reading.
   - It was noted that 857 event locations were successfully visualized out of a total of 859 data points.

f. Additional Navigation, The dashboard provides several tabs for advanced exploration :
   ❖ Incident Map

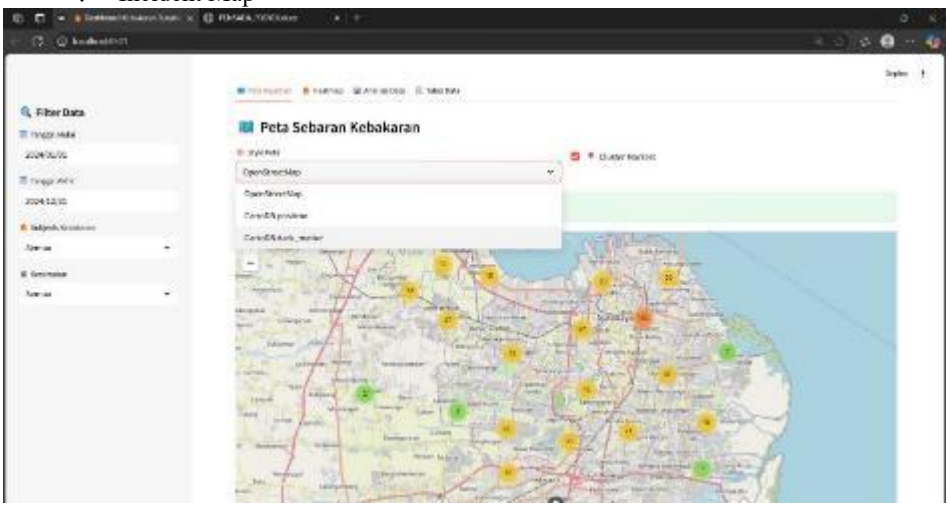

Figure 12. Incident Map

The Map Style section allows the user to select the background appearance of the map according to visual preference or analysis needs. This feature is provided in the form of a dropdown menu that can be selected interactively. Available map style options include :

a. OpenStreetMap, A default map style that displays geographic details such as roads, buildings, and boundaries with a familiar, standardized look.
b. CartoDB positron, Displays a map with a cleaner and minimalist look, dominated by light colors, suitable for data visualization that wants to stand out from the map background.
c. CartoDB dark_matter, Provides a dark mode map view, which is ideal for displaying data with contrasting colors, especially in low lighting conditions or when highlighting specific visual patterns.

❖ **Heatmap**

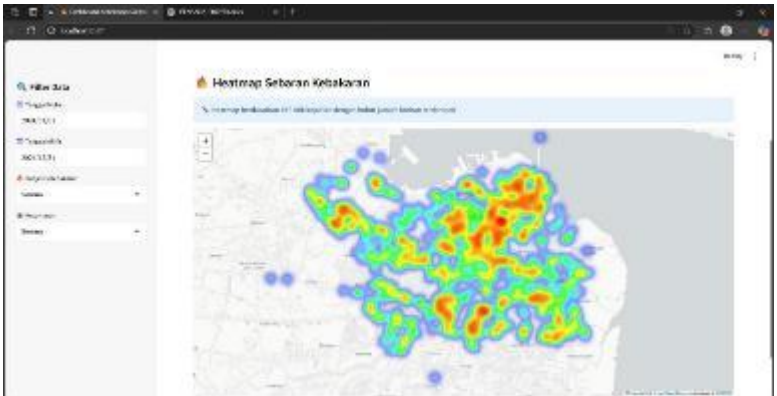

Figure 13. Dashboard of Heatmap(Heatmap Sebaran Kebakaran)

The "Heatmap of Fire Distribution(Heatmap Sebaran Kebakaran)'' page on this dashboard presents a heatmap visualization depicting the spatial distribution of 857 fire hotspots in Surabaya City throughout 2024. This heatmap is created by considering the weight of the number of victims affected, so that areas with greater impact are displayed in more intense colors, such as red for the highest density, orange and yellow for medium density, and green and blue for low density. The map is interactive and comes with a zoom feature as well as filters by date, fire subtype and subdistrict, allowing users to focus the analysis on specific areas or categories. This visualization is very helpful identifying fire-prone areas and can be used as a basis for planning more targeted mitigation and disaster management strategies.

❖ **Data Analysis**

Figure 14. Dashboard of Data Analysis

The Data Analysis(Analisis Daata) page on this dashboard is the result of an Exploratory Data Analysis (EDA) process on fire data in Surabaya City for 2024, which aims to understand the characteristics of the data prior to further modeling. One of the visualizations displayed is the response time distribution graph (in seconds), which shows the frequency distribution of the number of fire incidents based on the length of officer response time. This helps identify patterns or anomalies in fire response performance, such as whether there are frequent delays or most incidents are handled quickly. With this analysis, stakeholders can evaluate response efficiency and design strategies to improve emergency services in the future.

❖ **Data Table**

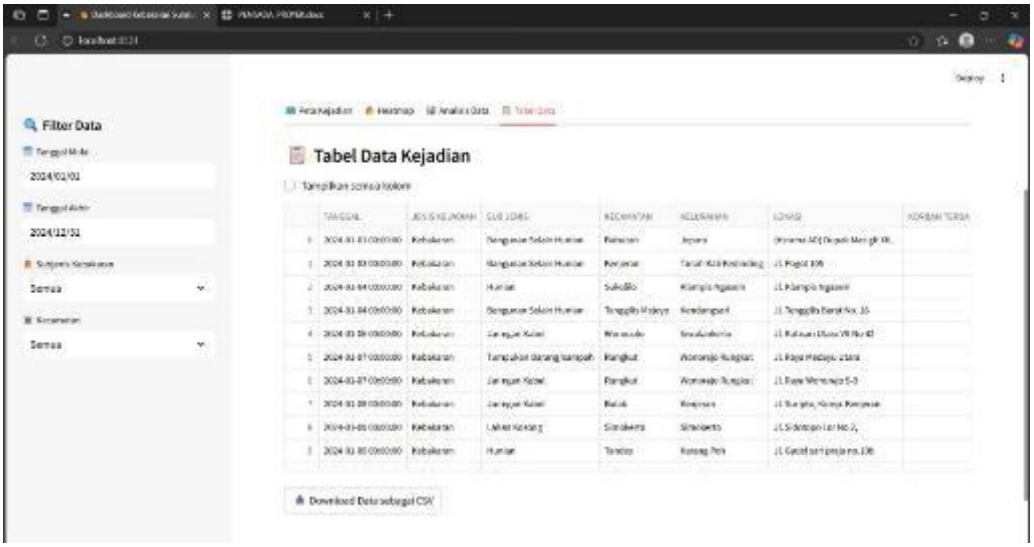

Figure 15. Dashboard of Data Table

The Data of Table(Data Tabel) page of this fire dashboard provides detailed information on all fire incidents recorded in Surabaya City during the period January 1 to December 31, 2024. The table includes several important columns, including: date of incident, type of incident (e.g. "Fire"), sub-type (e.g. "Non-Residential Building", "Cable Network", "Vacant Land"), sub-district and village where the incident took place, address of the incident location, and number of victims affected. The data is displayed in an interactive tabular form that allows users to filter by date, fire subtype, and specific subdistrict. In addition, there is a checkbox feature to display all additional columns if desired, as well as a Download Data as CSV button to allow users to download and further analyze the data independently. This feature is very useful for researchers, policy makers, and other related parties to understand the pattern of events and establish more effective prevention measures.

CONCLUSION

Based on the results of the implementation and evaluation of the classification model using the Extreme Gradient Boosting (XGBoost) algorithm on Surabaya City fire data in 2024, it was found that the model was able to classify the level of fire risk (LEVEL_RISIKO) very accurately. The model was trained using numerical and spatial variables such as response time, number of victims affected, and the encoding results of the event subtype. Preprocessing included label encoding, data normalization with StandardScaler, and data balancing using the SMOTE technique to address class imbalance.

The model showed very high performance with precision, recall, and f1-score values of 1.00 across all classes (Low, Medium, and High), and an overall accuracy of 1.00%. The confusion matrix visualization shows that only one data was misclassified out of a total of 303 test data. In addition, the multiclass ROC graph shows AUC values close to 1 for all classes, indicating that the model has excellent discriminative ability.

These results show that XGBoost is a highly effective method for fire risk classification in urban areas. This study recommends the application of this method in early warning systems and data-driven decision-making for more targeted fire risk mitigation policies in Surabaya City and other areas.

REFERENCES

[1] Ghaitsa Amany Mursianto, dkk. (2021). *Perbandingan Metode Klasifikasi Random Forest dan XGBoost Serta Implementasi Teknik SMOTE pada Kasus Prediksi Hujan*. SENAMIKA.

[2] Ichwanul Muslim Karo Karo. (2020). *Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan*. Journal of Software Engineering, Information and Communication Technology.

[3] Adam Kharis Pratama, dkk. (2023). *Klasifikasi Data Gempa Bumi di Pulau Jawa Menggunakan Algoritma Extreme Gradient Boosting*. JATI (Jurnal Mahasiswa Teknik Informatika), Vol. 7 No. 4.

[4] Rearizth Muhammad Daffaa, dkk. (2025). *Perbandingan XGBoost dan Logistic Regression dalam Memprediksi Credit Card Customer Churn*. Jupiter, Vol. 3 No. 3.

[5] Ichwanul Muslim Karo Karo. (2020). *Ibid*, hlm. 14-15.