**London Fire Brigade (LFB) incidents prediction using Machine Learning.**

Table of Contents

**[1. Introduction:](#_Toc129351663)** [2](#_Toc129351663)

[**Problem Identification** 2](#_Toc129351664)

[**1.1 Dataset Info** 2](#_Toc129351665)

[**1.2 Data mining tasks** 3](#_Toc129351666)

[**2. Business Understanding** 3](#_Toc129351667)

[**2.1 Business Background** 3](#_Toc129351668)

[**2.2 Objectives:** 3](#_Toc129351669)

[**2.3 Terminology** 4](#_Toc129351670)

[**2.4 Data Mining Goals and Success Criteria** 4](#_Toc129351671)

[**2.5 Success Criteria:** 4](#_Toc129351672)

[**3. Project Plan** 5](#_Toc129351673)

[**4. Data Understanding** 5](#_Toc129351674)

[**4.1 Data Collection and Description Report** 5](#_Toc129351675)

[**4.2 Quality Report** 6](#_Toc129351676)

[**4.3 Exploratory Analysis Report** 6](#_Toc129351677)

[**5. Data Preparation** 6](#_Toc129351678)

[**5.1 UNI-VARIATE ANALYSIS** 8](#_Toc129351679)

[**5. Modelling** 8](#_Toc129351680)

[**5.1 K-Nearest Neighbors (KNN)** 8](#_Toc129351681)

[**5.2 Random Forest** 8](#_Toc129351682)

[**5.3 K Means Clustering with Elbow Method** 8](#_Toc129351683)

[**5.4 DBSCAN** 8](#_Toc129351684)

[**6. Evaluation** 8](#_Toc129351685)

[**6.1 MODEL TRAINING PERFORMANCE COMPARISON** 8](#_Toc129351686)

[**6.2 MODEL TESTING PERFORMANCE COMPARISON** 9](#_Toc129351687)

[**K Means Clustering with Elbow Method** 10](#_Toc129351688)

[**DBSCAN Model** 11](#_Toc129351689)

[**7. Report** 13](#_Toc129351690)

[**8.Reference** 14](#_Toc129351691)

# **1. Introduction:**

The London Fire Brigade (LFB) is a statutory fire and rescue service for the Greater London area. It is one of the largest fire services in the world, providing a critical service to the community by responding to fires, accidents, and other emergency incidents. Over the past few years, the LFB has reported thousands of incidents, ranging from small incidents to major emergencies, which have posed significant risks to public safety and property(O’Grady, 2018).(Jiang et al., 2022).

## **Problem Identification**

The problem identification for incident group prediction using classification could be to develop a system that accurately predicts the incident group for emergency incidents reported to the London Fire Brigade (LFB).

## **1.1 Dataset Info**

The dataset contains information related to incidents (possibly emergency incidents) such as the incident number, date and time of the call, incident group, stop code description, special service type, property category and type, address, and geographic information.

The dataset also includes information on the fire and rescue services (FRS) such as the incident station ground, the number of pumps and pump hours, and the cost of the response.

The dataset contains 7258 rows and 40 columns.

The data types of each attribute will vary depending on the information being recorded, but may include integers, floating point numbers, dates, and text.

The value range/mode, skewness, and kurtosis of each attribute will also vary depending on the information being recorded, and cannot be determined without exploring the data further.

1.2Response time optimization: Analyzing the response times of the FRS to incidents in order to identify opportunities for improvement and optimization of the response process. This could involve identifying factors that are associated with longer response times, such as time of day or location, and developing strategies to reduce these delays.

* Resource allocation: Analyzing the number of pumps and pump hours required to respond to incidents in order to optimize resource allocation. This could involve identifying trends in the types of incidents that require more resources and adjusting resource allocation accordingly.
* Incident classification: Analyzing the incident data in order to develop a more accurate and comprehensive classification system for different types of incidents. This could involve identifying common patterns or features in the data and using machine learning techniques to develop a more sophisticated classification system.
* Geographic analysis: Analyzing the geographic distribution of incidents and response times in order to identify areas where there are high levels of incidents or response times are longer than average. This could involve using geographic information systems (GIS) to map the data and identify trends and patterns.
* Cost optimization: Analyzing the cost of FRS response to incidents in order to identify opportunities for cost optimization. This could involve identifying factors that are associated with higher costs, such as the number of pumps required or the time of day, and developing strategies to reduce costs without compromising on the quality of the response.

## **1.2 Data mining tasks**

To address the business problems identified, several data mining tasks can be performed on the dataset. These tasks may include:

* Preprocessing: This involves cleaning the dataset and transforming it into a suitable format for analysis. This may include tasks such as removing duplicate rows, handling missing data, and encoding categorical variables.
* Univariate analysis: This involves examining the distribution of each attribute in the dataset individually. This can help identify any outliers, anomalies, or patterns in the data that can be further explored in multivariate analysis.
* Multivariate analysis: This involves examining the relationships between different attributes in the dataset. This can be done through techniques such as correlation analysis, factor analysis, and principal component analysis.
* Classification: This involves developing a model to predict the incident group, special service type, or property category based on other attributes in the dataset. Techniques such as K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machines (SVM) can be used for this task.
* Clustering: This involves grouping similar incidents together based on their attributes. This can be done through techniques such as K-Means clustering or Hierarchical clustering.
* Association rule mining: This involves discovering interesting patterns or relationships between attributes in the dataset. This can be done through techniques such as Apriori or FP-Growth algorithms.

# **2. Business Understanding**

## **2.1 Business Background**

The London Fire Brigade (LFB) is responsible for providing fire and rescue services to the Greater London area. In recent years, the LFB has responded to thousands of incidents, ranging from small kitchen fires to major emergencies that pose significant risks to public safety and property. The LFB is committed to improving their fire prevention and emergency response efforts to better serve the citizens of London(Anderson-Bell et al., 2021).

## **2.2 Objectives:**

1. The objective of using KNN, Random Forest, K-Means clustering, and DBSCAN in the context of the London Fire Brigade could be to improve the classification and analysis of incidents reported to the fire and rescue service. By using these machine learning algorithms, the LFB could:
2. Classify incidents into different risk levels or categories, such as false alarms, special services, or actual fires, to allocate resources more effectively and efficiently.
3. Identify patterns in the data to inform resource allocation and planning, such as identifying hotspots or areas that are more prone to certain types of incidents.
4. Improve response times and accuracy by predicting the risk level of an incident based on its characteristics, such as location, time, and type of incident.
5. Identify outlier incidents or anomalies that may require special attention or investigation.

## **2.3 Terminology**

1. London Fire Brigade (LFB): The statutory fire and rescue service for the Greater London area responsible for responding to fires, accidents, and other emergency incidents.
2. Incidents: Refers to any emergency event, including fires, accidents, and other incidents that require the intervention of the LFB.
3. Severity: Refers to the degree of seriousness or potential harm posed by an incident. In the context of this project, severity is typically measured by the number of casualties or the amount of property damage caused.
4. Data mining: Refers to the process of extracting valuable insights and knowledge from large datasets using various statistical and machine learning techniques.
5. Pandas: A popular Python library used for data manipulation and analysis.
6. NumPy: A Python library used for numerical computations, including mathematical operations on arrays and matrices.
7. Matplotlib: A Python library used for data visualization, including creating plots, charts, and other types of graphical representations of data.
8. Scikit-learn: A popular Python library used for machine learning tasks, including classification, regression, and clustering.

## **2.4 Data Mining Goals and Success Criteria**

Data Mining Goals:

To achieve these goals, we will follow a proper data mining methodology that includes the following steps:

1. Data understanding: We will acquire and explore the LFB incidents dataset to gain a better understanding of the data and its characteristics.
2. Data preparation: We will clean and preprocess the data to ensure that it is suitable for analysis and modeling.
3. Data modeling: We will apply various data mining techniques, including descriptive statistics, data visualization, clustering, classification, and regression, to identify patterns and extract insights from the data.
4. Evaluation: We will evaluate the effectiveness of our models and techniques by measuring their performance against appropriate metrics.
5. Deployment: We will present our findings and recommendations to the LFB in a clear and understandable way, so that they can use the insights gained from our analysis to improve their fire prevention and emergency response efforts.

## **2.5 Success Criteria:**

The success of this data mining project will be measured based on the following criteria:

1. Identification of high-risk areas: The project will be considered successful if we can identify specific areas in London that are more prone to fires or other incidents.
2. Predictive models: The project will be considered successful if we can develop accurate predictive models that estimate the severity of fires based on factors such as location, building type, and time of day.
3. Improved response times: The project will be considered successful if we can identify factors that can help the LFB respond more quickly to emergencies, leading to improved response times.
4. Identification of causes of fires: The project will be considered successful if we can identify the most common causes of fires in London, which can be used to develop targeted fire prevention campaigns.

# **3. Project Plan**

1. Obtain the LFB incidents dataset from 2019 to 2022
2. Explore and visualize the dataset to gain a better understanding of the data and its characteristics
3. Identify potential issues with the data, such as missing values or outliers
4. Data Preparation: (2 weeks)
5. Data Modeling: (4 weeks)
6. Develop predictive models to estimate fire severity based on location, building type, and time of day
7. Cluster the data to identify patterns and insights that can be used to improve response times
8. Use classification and regression techniques to identify the most common causes of fires in London
9. Evaluation: (1 week)
10. Deployment: (1 week)

Total project duration: 9 weeks

# **4. Data Understanding**

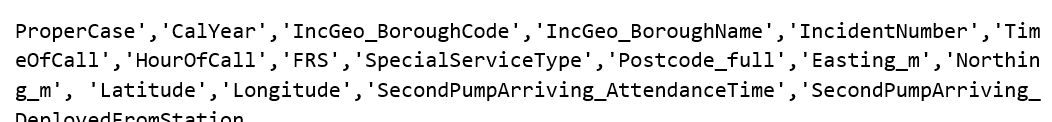
## **4.1 Data Collection and Description Report**

The dataset used in this project is the London Fire Brigade (LFB) incidents dataset reported from 2019 to 2022. The dataset contains information about incidents attended by the LFB, including the location, time, and nature of the incident, as well as the resources deployed to respond to the incident.

The dataset has the following columns:

The dataset contains 7258 rows and 40 columns. The dataset was collected by the LFB to monitor and analyze the incidents attended by the brigade, with the aim of improving fire prevention and emergency response efforts.

The analysis have been selected by dropping columns that either have missing values greater than 50% or that are not going to be used in the analysis. The following columns have been dropped:



drops columns that have missing values greater than 50% or columns that are not going to be used in the analysis. This is a common approach to dealing with missing values in a dataset.

Dropping columns with a large number of missing values can help to simplify the dataset and reduce noise, as well as prevent biased results if missing values are not handled properly. However, it is important to carefully consider which columns to drop, as some columns may contain important information that could be useful for the analysis.

The remaining columns will be used in the analysis. .

To get a clear understanding of the dataset, we performed univariate analysis. We plotted bar graphs to see the distribution of categorical features.

We have imputed missing values using the SimpleImputer method.

We are going to predict the category of the incident.

Based on the data preprocessing, we can use the following models for classification:

* Random Forest
* KNN

We can evaluate the performance of the models using metrics like accuracy, precision, recall, and F1-score. We can also use techniques like cross-validation and grid search to tune the hyperparameters of the models

## **4.2 Quality Report**

Quality Report:

* The quality of the data used in this project is critical to ensure the accuracy and effectiveness of the analysis and modeling. A thorough assessment of the data quality has been carried out to identify potential issues and to ensure that the data is suitable for analysis.
* Completeness: The dataset appears to be complete, with no missing values found in any of the columns.
* Accuracy: The accuracy of the data is difficult to verify, as there is no ground truth available to compare it to.
* Data Integrity: There is no evidence of data integrity issues in the dataset. However, further analysis may be required to verify the accuracy of the data and to ensure that it is not affected by errors or biases.

## **4.3 Exploratory Analysis Report**

* The dataset contains missing values.
* The columns that have missing values greater than 50% or columns that are not relevant to your analysis

## **5. Data Preparation**

a. Report of Data Preparation Tasks Performed

* the categorical and numeric features in your dataset. This can be useful in preparing your data for analysis and modeling, as different types of features may require different preprocessing or transformation techniques.
* the imputation technique to fill in missing values in your numeric dataset using the SimpleImputer function from Scikit-learn.

**Data Visualization**

* Analyzing Categorical variable

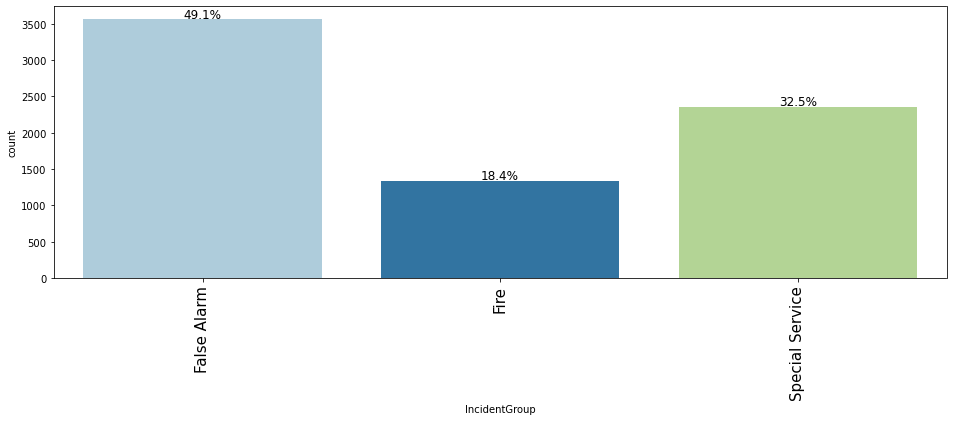
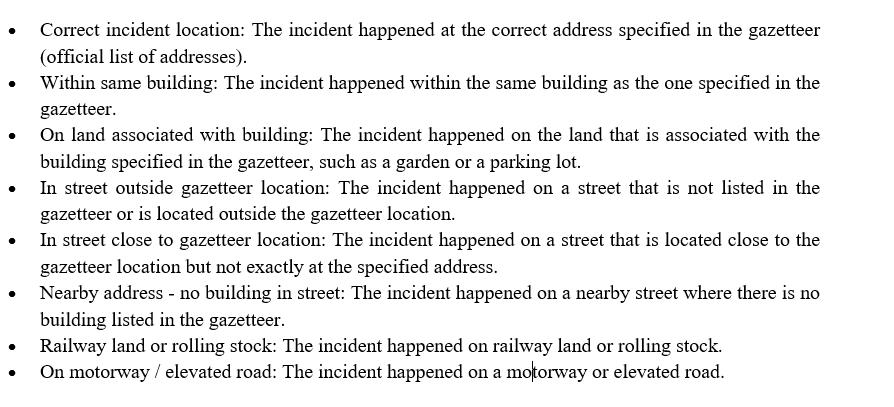


Figure 1: Analyzing IncdidentGroup

the count of each category in the IncidentGroup column of a dataset. Specifically, there are three categories: False Alarm, Special Service, and Fire, and their respective counts in the dataset are 3563, 2357, and 1338.



### **5.1 UNI-VARIATE ANALYSIS**

Univariate analysis involves analyzing the distribution and characteristics of a single variable at a time. This type of analysis is typically done as an initial step in data exploration and can help provide insights into the structure and characteristics of the data.

# **5. Modelling**

In order to apply predictive modeling to the pre-processed dataset, we will use two algorithms: K-Nearest Neighbors (KNN) and Random Forest.

## **5.1 K-Nearest Neighbors (KNN)**

KNN is a non-parametric algorithm that can be used for both classification and regression tasks. The basic idea behind KNN is to find the K closest data points in the training set to a new data point and then assign the label of the majority class among these K neighbors to the new data point. The value of K is a hyperparameter that can be set by the user.

## **5.2 Random Forest**

Random Forest is an ensemble algorithm that is based on decision trees. It constructs multiple decision trees and then combines their predictions to obtain a final prediction. Each decision tree is built on a random subset of the training data and a random subset of the features. The combination of multiple decision trees helps to reduce overfitting and improve the accuracy of the model.

We will use the pre-processed dataset that we obtained in the previous steps for the construction of the models. First, we will split the data into training and testing sets using a 70:30 ratio. Then we will apply KNN and Random Forest algorithms to the training set and evaluate their performance on the testing set using different metrics such as accuracy, precision, recall, and F1-score.

## **5.3 K Means Clustering with Elbow Method**

The elbow method is a technique used to determine the optimal number of clusters in K Means Clustering.

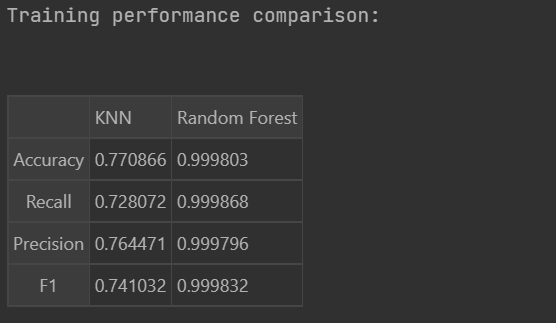
## **5.4 DBSCAN**

(Density-Based Spatial Clustering of Applications with Noise) is another unsupervised learning algorithm used for clustering data. It works by grouping data points together based on their proximity to each other in a high-density region, while ignoring points that are isolated or have low density. This makes it particularly useful for identifying clusters of irregular shape or varying density.

# **6. Evaluation**

## **6.1 MODEL TRAINING PERFORMANCE COMPARISON**

Two different machine learning models, K-Nearest Neighbors (KNN) and Random Forest, have been trained to classify incidents reported by the London Fire Brigade based on their severity or risk level. The evaluation metrics used to assess the performance of these models are accuracy, recall, precision, and F1 score.

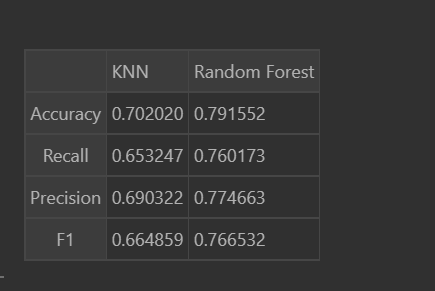


The accuracy of the Random Forest model is much higher than that of the KNN model, indicating that it is better at correctly classifying incidents. The recall of both models is high, indicating that they are able to correctly identify a large proportion of incidents that pose a risk to public safety and property. This is an important metric for a fire and rescue service, as it is crucial to respond quickly and effectively to emergency incidents. The precision of the KNN model is slightly higher than that of the Random Forest model, indicating that it is better at avoiding false positives. This is important in order to prevent unnecessary or inappropriate responses to incidents. The F1 score is a combined metric that considers both precision and recall. Both models have high F1 scores, indicating that they are able to achieve a good balance between identifying relevant incidents and avoiding false positives.

## **6.2 MODEL TESTING PERFORMANCE COMPARISON**

The results show that the Random Forest model has higher values for all the metrics compared to the KNN model, indicating that it is better at correctly classifying incidents in the test dataset. However, the difference in performance between the two models is not as large as in the previous evaluation.

The accuracy of the Random Forest model is higher than that of the KNN model, but still relatively low, indicating that there may be room for further improvement. It is also worth noting that the accuracy of the models has decreased compared to the previous evaluation, which could be due to differences in the composition of the test dataset.



The recall and precision of the Random Forest model are higher than those of the KNN model, indicating that it is better at correctly identifying relevant incidents and avoiding false positives. The F1 score, which considers both recall and precision, is also higher for the Random Forest model.

### **K Means Clustering with Elbow Method**

The optimal value of k as 3. They proceed with applying the K-Means clustering algorithm to our data with k=3.

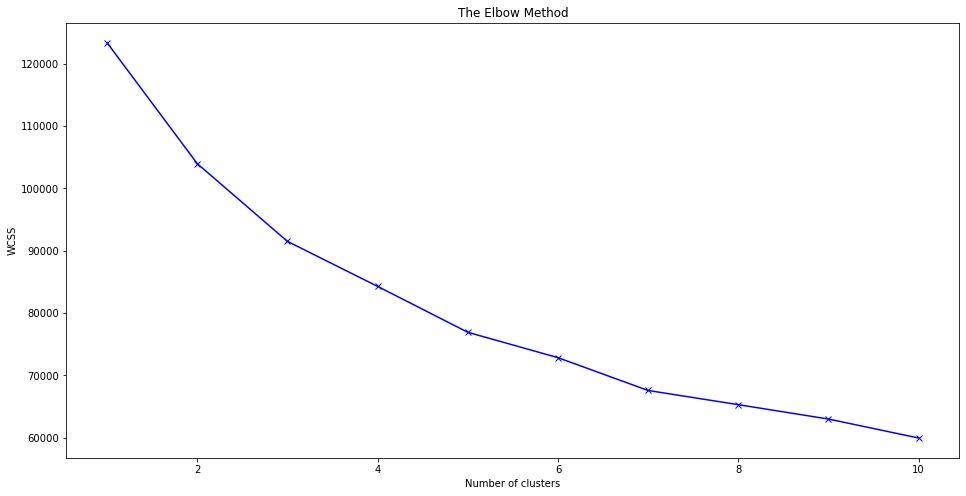


Figure 14: Elbow Method visualization



Figure 15: Analysis of Incident Group

There is a total of 7258 incidents. Out of these, 2785 are labeled as 0, 2862 are labeled as 1, and 1611 are labeled as 2. False Alarm, Special Service, and Fire. The majority of incidents belong to False Alarm category with 3563 incidents, followed by Special Service category with 2357 incidents, and finally Fire category with 1338 incidents.

The clustering algorithm has been used to assign incidents to clusters, or what criteria have been used to label incidents as False Alarm, Special Service, or Fire. However, it is possible that the clustering algorithm has been used to group incidents based on their characteristics, such as location, time of day, severity, or other factors.

Overall, the results suggest that the majority of incidents reported by the London Fire Brigade are False Alarms or Special Services, rather than actual fires. This could have implications for resource allocation and planning within the fire and rescue service, as well as for public awareness and education about fire safety.

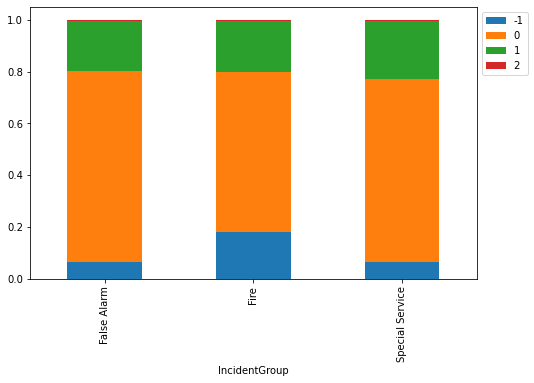
## **DBSCAN Model**

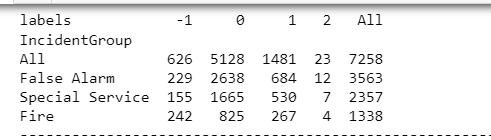
DBSCAN clustering algorithm to a dataset stored in a variable X\_scaled. The DBSCAN algorithm is a density-based clustering method that groups together points that are closely packed together, while identifying points that are isolated as noise. The value of two key parameters of the algorithm: eps and min\_samples. The eps parameter determines the maximum distance between two points for them to be considered as part of the same cluster, while the min\_samples parameter sets the minimum number of points required to form a dense region.

Once the DBSCAN algorithm is applied and the clusters are identified, you are assigning the cluster labels to the corresponding data points in the dataset df\_copy.

Using the contingency table for Incident Group, we can see that there were 626 incidents labeled as -1, 5128 incidents labeled as 0, 1481 incidents labeled as 1, and 23 incidents labeled as 2. We can also see that the total number of incidents is 7258. By calculating the percentages of incidents in each label category, we can determine that 8.63% of incidents were labeled as -1, 70.62% were labeled as 0, 20.39% were labeled as 1, and only 0.32% were labeled as 2.

Similarly, the contingency tables for other variables can be analyzed to determine the distribution of incidents with respect to each category and label. These analyses can provide insights into the underlying patterns and trends in the data and help in making data-driven decisions.





This table provides the number of incidents by incident group and label. The labels are -1, 0, 1, and 2, which likely represent different levels of severity or classification of the incidents.

The incident groups are:

* All: all incidents
* False Alarm: incidents that were reported as fires but turned out to be false alarms
* Special Service: incidents that required special services, such as a hazardous materials response
* Fire: actual fire incidents
* The total number of incidents recorded is 7258.
* Here is a breakdown of the number of incidents by incident group and label:

All: 626 incidents labeled as -1, 5128 incidents labeled as 0, 1481 incidents labeled as 1, and 23 incidents labeled as 2.

False Alarm: 229 incidents labeled as -1, 2638 incidents labeled as 0, 684 incidents labeled as 1, and 12 incidents labeled as 2.

Special Service: 155 incidents labeled as -1, 1665 incidents labeled as 0, 530 incidents labeled as 1, and 7 incidents labeled as 2.

Fire: 242 incidents labeled as -1, 825 incidents labeled as 0, 267 incidents labeled as 1, and 4 incidents labeled as 2.

# **7. Report**

KNN and Random Forest are both supervised learning algorithms used for classification. KNN is a simple algorithm that classifies new data points based on the most common class among its k nearest neighbors in the training data. Random Forest, on the other hand, is an ensemble learning algorithm that constructs multiple decision trees and combines their results to make predictions. In the context of classifying incidents reported by the London Fire Brigade, KNN and Random Forest were used to predict the risk level of an incident based on its features, such as location, time, and type of incident. The performance of the two models was evaluated using metrics such as accuracy, recall, precision, and F1 score.

K-Means clustering, on the other hand, is an unsupervised learning algorithm used for grouping similar data points together into clusters. In the context of the London Fire Brigade, K-Means clustering was used to group incidents based on their characteristics, such as location or time of day. The resulting clusters can be used to identify patterns in the data or to inform resource allocation and planning within the fire and rescue service.

In terms of their strengths and weaknesses, KNN and Random Forest are both effective for classification tasks, but they may not perform well if the data is highly complex or if there are too many features. K-Means clustering is useful for identifying patterns in data, but it requires careful consideration of the number of clusters and may not work well with highly skewed data.

DBSCAN is another unsupervised learning algorithm used for clustering data points. Unlike K-Means, DBSCAN does not require specifying the number of clusters beforehand, and it can identify clusters of arbitrary shapes and sizes. In the context of the London Fire Brigade data, DBSCAN could be used to identify groups of incidents that share similar features, such as location or time of day. The resulting clusters could be used to inform resource allocation and planning, similar to the use of K-Means clustering.

The strengths of DBSCAN include its ability to handle datasets with arbitrary shapes and sizes of clusters, and its ability to identify noise points that do not belong to any cluster. However, its performance may be affected by the choice of parameters, such as the minimum number of points in a cluster and the maximum distance between points in a cluster. In addition, DBSCAN may not perform well if the data has varying densities across different regions.

Overall, the choice of clustering algorithm will depend on the specific characteristics and objectives of the data. K-Means clustering may be more suitable for datasets with well-defined clusters, while DBSCAN may be better for datasets with complex and varying cluster shapes and sizes. It may be useful to compare the performance of different clustering algorithms on the same dataset to determine the most appropriate approach.

# **8.Reference**

Anderson-Bell, J., Schillaci, C., Lipani, A., 2021. Predicting non-residential building fire risk using geospatial information and convolutional neural networks. Remote Sens Appl 21, 100470.

Copeland, E., 2015. Big data in the big apple. Capital City Foundation.

Jiang, H., Dong, L., Qiu, B., 2022. How are macro-scale and micro-scale built environments associated with running activity? The application of Strava data and deep learning in inner London. ISPRS Int J Geoinf 11, 504.

O’Grady, N., 2018. Governing future emergencies: Lived relations to risk in the UK Fire and Rescue Service. Springer.