EXPLORATORY DATA ANLAYSIS

INTELLIGENT SMOKE DETECTION SYSTEM WITH AI SENSOR FUSION

s15089 - Sanjani Wickramasinghe

s14982 - Poornima Dissanayake

s14953 - Buddhima Senarathna

s15006 - Pasindu Sachintha

Prepared by

Group 10

ABSTRACT

The Smoke Detection System with IoT Devices and AI Model utilizes advanced sensors, including the Sensirion SPS30, and the Arduino Pro Nicla Sense ME board as the central hub to detect fire occurrences accurately.

This project's extensive dataset, comprising about 60,000 sensor readings from diverse scenarios, forms the foundation for comprehensive AI model training. This empowers the system to make informed decisions even in complex and dynamic environments. The primary challenge lies in accurately identifying the presence of fire while minimizing false alarms to enhance the system's reliability. The project places special emphasis on data analysis, including sensor data processing, cleaning, and AI sensor fusion algorithms.

These algorithms play a pivotal role in enhancing fire detection accuracy and reducing false alarms. Various environmental factors such as humidity, air pressure, and gas levels have shown correlation with fire presence. The Principal Component Analysis (PCA) helps to reduce the dimensionality of data and identify correlations among variables. To achieve a quality advanced analysis, suggestions include employing logistic regression models with consideration for highly correlated predictors and principal components.

Random Forest, an ensemble learning method, can also be utilized for its high accuracy and ease of use. Additionally, exploring separate models for identified clusters in the dataset may further improve model accuracy compared to a single model approach. The system aims to contribute to the advancement of fire detection technology, enhancing fire safety in various indoor and outdoor settings, ultimately preventing fire-related incidents and safeguarding lives and property.

TABLE OF CONTENT

.

LIST OF FIGURES

INTRODUCTION

The Smoke Detection System with IoT Devices and AI Model is a pioneering project that combines cutting-edge technology to revolutionize fire detection capabilities. By leveraging Artificial Intelligence (AI) and Internet of Things (IoT), this advanced smoke detector offers a powerful solution for early fire detection in diverse environments.

At the heart of this system lies the Arduino Pro Nicla Sense ME board, acting as the central hub for gathering data from a comprehensive array of sensors. The project harnesses the Sensirion SPS30, an advanced smoke detector, which measures particulate matter in the air with exceptional accuracy. Complementing the SPS30 are redundant sensors for Humidity/Temperature, Air Pressure, and Gas (Volatile Organic Compounds - VOC), ensuring resilience and fault tolerance in challenging conditions.

The project's extensive dataset comprises approximately 60,000 sensor readings from a diverse range of scenarios, encompassing various indoor and outdoor conditions, indoor wood and gas fires for firefighter training, and outdoor grilling scenarios. This rich dataset serves as the foundation for comprehensive AI model training, enabling the system to make informed decisions even in complex and dynamic environments.

In this project, we place a special emphasis on the data analysis part of the Smoke Detection System. We explore the intricate methodologies employed to process, clean, and interpret the vast amount of sensor data. Furthermore, we delve into the implementation of AI sensor fusion algorithms, which play a pivotal role in enhancing fire detection accuracy and reducing false alarms.

PROBLEM STATEMENT

The primary challenge lies in accurately identifying the presence of fire from the sensor data, while minimizing false alarms to enhance the system's reliability. The developed model can be integrated into the Smoke Detection System with IoT Devices and AI Model, providing enhanced fire safety in various indoor and outdoor settings. Ultimately, the project aims to contribute to the advancement of fire detection technology, safeguarding lives and property by preventing fire-related incidents.

DESCRIPTION OF THE DATASET

The dataset is a comprehensive collection of sensor readings captured from various scenarios. It consists of approximately 60,000 data points, each representing a specific timestamp and corresponding sensor measurements.

The dataset comprises the following variables:

- **Air Temperature**: Ambient temperature of the environment in degrees Celsius (°C).
- **Air Humidity**: Relative humidity of the air as a percentage (%).
- **TVOC** (Total Volatile Organic Compounds in ppb): TVOC refers to the total concentration of various volatile organic compounds present in the air.
- **eCO2** (CO2 Equivalent Concentration in ppm): The equivalent concentration of carbon dioxide (CO2) in the air.
- **Raw H2** (Raw Molecular Hydrogen): Readings of molecular hydrogen in the air which has not undergone compensation or correction for factors such as bias or temperature.
- **Raw Ethanol**: Measurement of ethanol gas in its raw or unprocessed form.
- **Air Pressure**(hPa): The atmospheric pressure in the environment.
- **PM1.0 and PM2.5** (Particulate Matter): PM1.0 represents particulate matter with a size less than 1.0 micrometer (µm), and PM2.5 represents particulate matter with a size between 1.0 µm and 2.5 µm.
- **CNT**: The sample counter, which helps to keep track of the order of sensor readings.
- **UTC**: Timestamp of each sensor reading in Coordinated Universal Time (UTC) seconds.
- **Fire Alarm** (Ground Truth): This feature is the target variable, denoting whether a fire is present or not. A value of "1" indicates the presence of a fire, while "0" indicates no fire.

MAIN RESULTS OF THE DESCRIPTIVE ANALYSIS

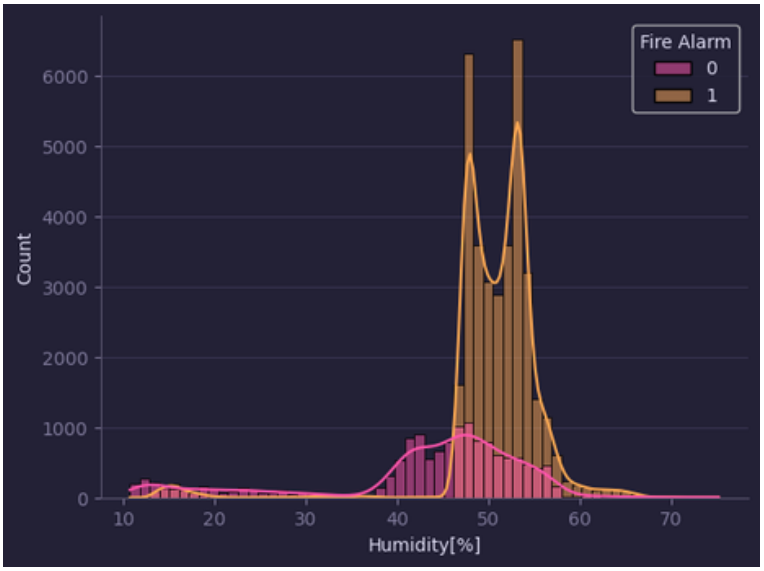

Figure 1: Histogram of humidity according to fire alarm

From this graph it's visible that the humidity has increased in case of a fire. One possible reason for this would be if the fire is contained within a confined or enclosed space, the moisture released from the burning materials, along with limited ventilation, can contribute to higher humidity levels in that specific area. The confinement of the fire can prevent the dissipation of moisture, leading to localized high humidity. Also, traditional fire alarms can cause a false alarm in high humidity levels. But it's evident that this device won't trigger an alarm just because the humidity is high because there are many cases when the humidity is high but the alarm is not triggered.

This figure shows that the air pressure has increased in the presence of a fire. Fires involve the rapid release of energy through the combustion of fuel. During combustion, gases are generated, including hot air and smoke, which can cause a temporary increase in air pressure in the immediate vicinity of the fire source. There are instances where the air pressure is high even without a fire. This might be caused by the environment the data was collected.

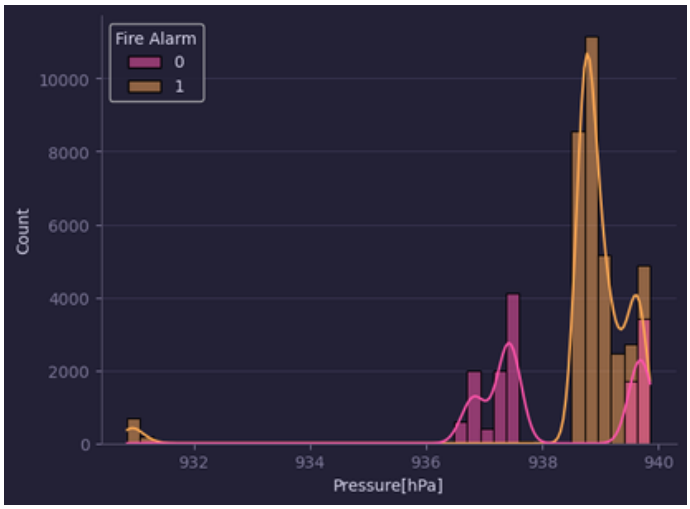

Figure 2: Histogram of pressure according to fire alarm

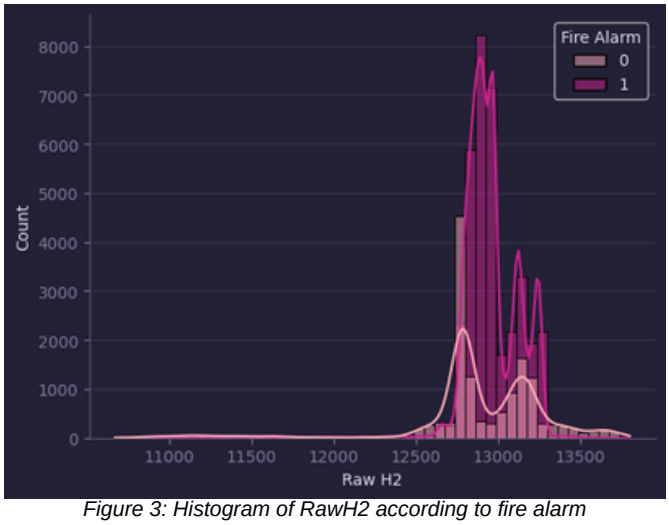

Figure 3: Histogram of RawH2 according to fire alarm

The hydrogen level shows an increment when there is a fire. In certain types of fires, hydrogen gas can be produced as a byproduct of the chemical reactions taking place. For example, when organic materials like wood, paper, or plastics burn, they release various gases, including hydrogen gas. This will result in an increment of the hydrogen gas in the air.

This figure shows the raw ethanol level distribution. If the fire involves the burning of materials that contain ethanol or ethanol-based substances, it can result in the release of ethanol gas. This can occur in scenarios where flammable liquids, alcohol-based fuels, or ethanol-containing materials are involved. This could be a possible reason for increased ethanol level. But we could see that there are high ethanol level even though a fire is not present.

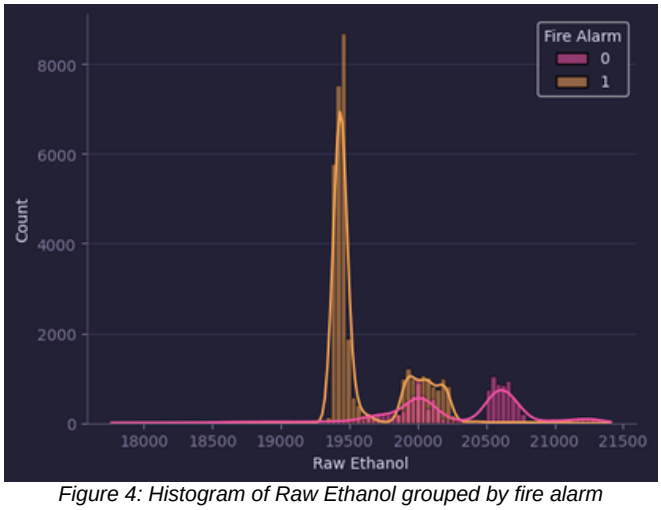

Figure 4: Histogram of Raw Ethanol grouped by fire alarm

Ethanol is commonly found in many consumer products such as cleaning solutions, hand sanitizers, perfumes, and other personal care items. If these products are used in the vicinity or if there is poor ventilation, it can result in increased ethanol levels in the air. This could be due to the testing environment.

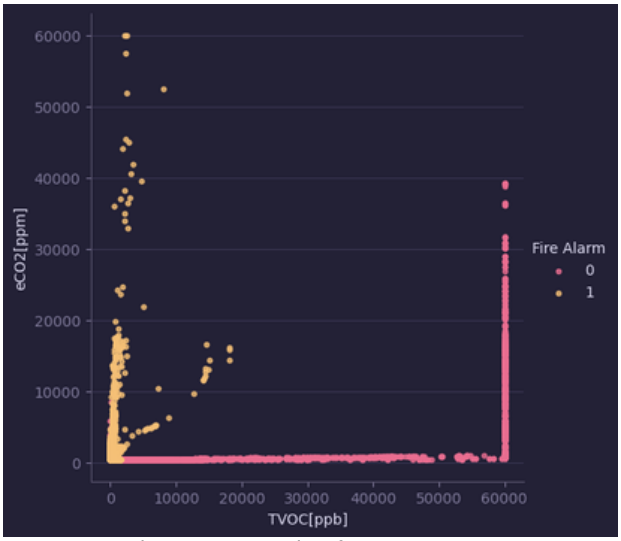

Figure 5: Scatterplot of TVOC vs eCO2

Fires often produce smoke, which contains various combustion byproducts, including carbon dioxide will result in a high eCO2 level.

A TVOC reading of 0 ppb indicates that no detectable volatile organic compounds were measured at that particular time or location. It means that the levels of volatile organic compounds, such as chemicals emitted from paints, solvents, cleaning products, or other sources, are below the detection threshold of the sensor.

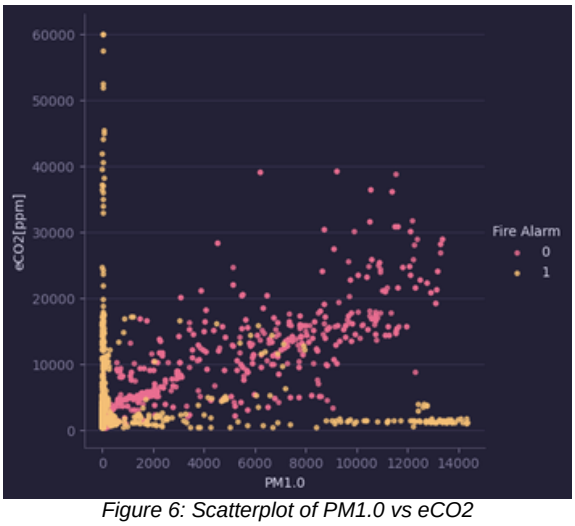

Figure 6: Scatterplot of PM1.0 vs eCO2

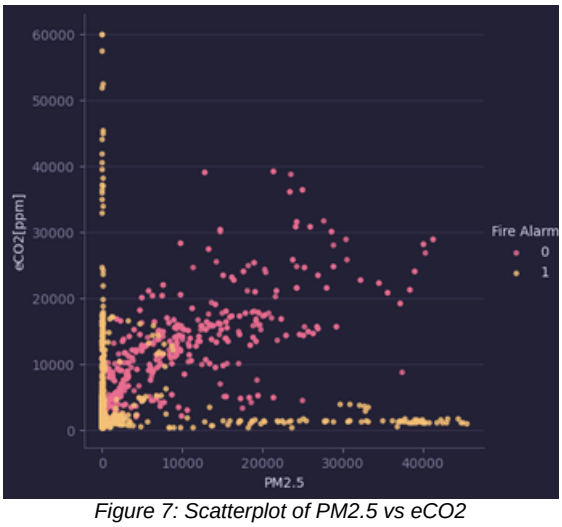

Figure 7: Scatterplot of PM2.5 vs eCO2

We could see the situation where a fire is present, eCO2 level is high but both PM1.0 and PM2.5 is very low. This may be due to the fire might be burning efficiently, resulting in minimal particulate matter emissions. In such cases, the majority of the combustion byproducts are in the form of gases, including carbon dioxide (CO2) and other volatile compounds, leading to elevated eCO2 levels. This can happen in well-controlled or controlled combustion processes, such as gas fires or clean-burning fuel sources. This is possible as there are data obtained near a gas grill.

High PM1.0 or PM2.5 but low eCO2 when a fire is present may be due to that the fire burning inefficiently or experiencing incomplete combustion. In such cases, the fire generates more particulate matter and smoke rather than producing significant amounts of carbon dioxide. This could occur with certain fuel types, combustion conditions, or ventilation limitations.

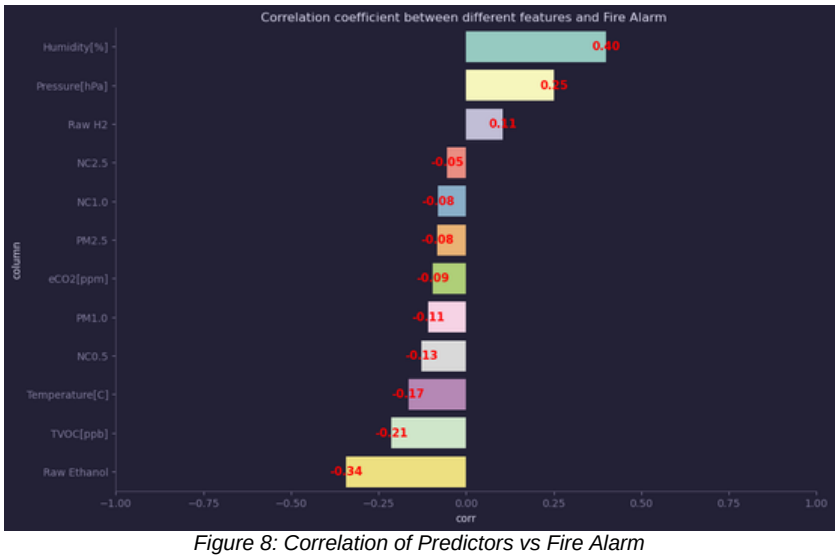

Figure 8: Correlation of Predictors vs Fire Alarm

In this figure, it shows the results of a correlation analysis that has been conducted to explore the relationships between the predictor variables and the response variable. Humidity has the highest correlation with the Fire Alarm variable, with a correlation coefficient of 0.4. This positive correlation suggests that as Humidity increases, there is a tendency for the Fire Alarm variable to increase as well.

It implies that higher humidity levels might be associated with a higher likelihood of a fire alarm being triggered.

PRINCIPAL COMPONENT ANALYSIS

By PCA we can reduce the dimensionality of our data and use less no of variables in the analysis. Also, this method is useful in dealing with multicollinearity as we identified that some variables are highly corelated with each other.

Optimal number of components?

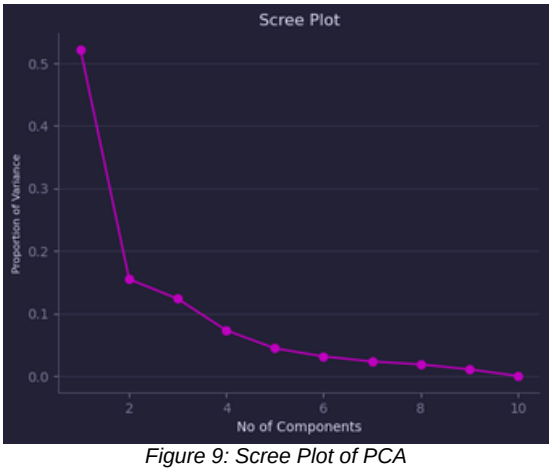

Figure 9: Scree Plot of PCA

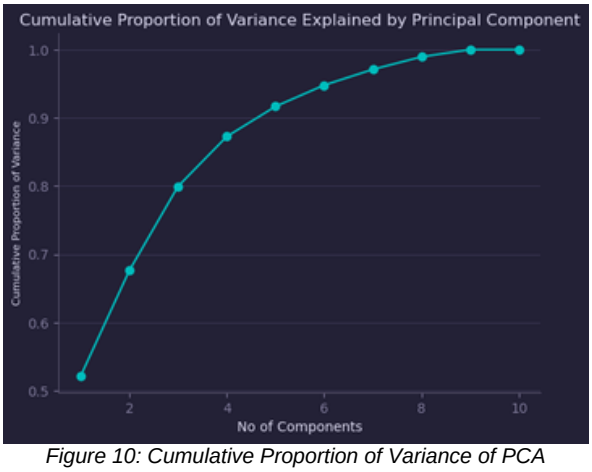

Figure 10: Cumulative Proportion of Variance of PCA

Using the scree plot and the cumulative proportion explained by the components, we can decide that the optimal number of components to be 4, which will explain approximately 87% of the variation in the data.

When the loadings plot is considered, it shows some variables are close to each other which implies that these features have similar patterns or tendencies in their variation across the data.

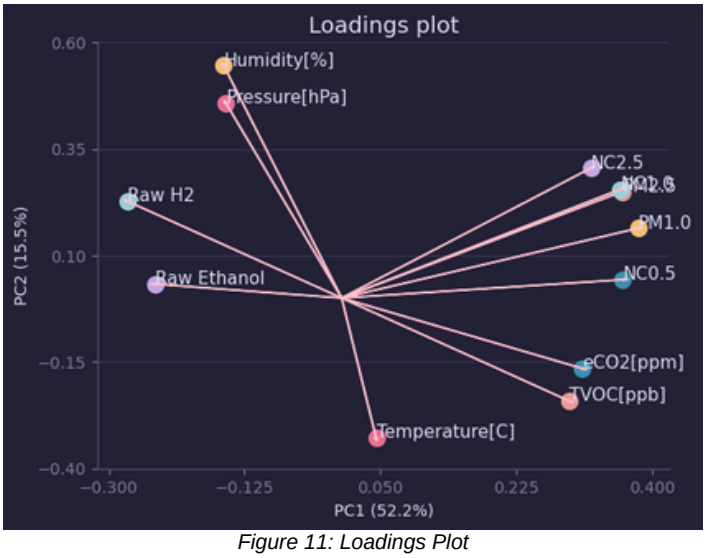

Figure 11: Loadings Plot

- **eCO2 and TVOC**

Both eCO2 and TVOC are related to air quality and can be influenced by similar environmental conditions. For example, both eCO2 and TVOC levels might increase in indoor settings with poor ventilation, high occupancy, or the presence of certain pollutants. Thus, they tend to show similar patterns of variation, leading to similar contributions in specific principal components.

- **NC2.5, NC0.5, NC1.0, PM1.0, and PM2.5**

These features represent different size fractions of particulate matter in the air, but they are all related to airborne particles and pollution levels. Therefore, they tend to exhibit similar patterns in their variations within the dataset.

- **Humidity and Pressure**

This alignment suggests that Humidity and Pressure are positively correlated or have similar patterns in their variation within the data. Humidity and pressure can be related, but their relationship is complex and can vary based on a combination of factors.

- **Raw H2 and Raw Ethanol**

Raw H2 and Raw ethanol might be influenced by similar environmental conditions, such as indoor air quality, outdoor pollution, or the presence of specific gases or chemicals in the air. For example, they both could be related to certain types of pollution, chemical emissions, or specific industrial activities.

In the loadings plot, variables are located in opposite directions which indicates that these features have different patterns of variation and are negatively correlated with each other in the dataset.

- **Humidity and Pressure are negatively related with Temperature**.

This behavior is not unexpected since these variables are often influenced by different atmospheric conditions and weather patterns. The atmospheric pressure decreases when the temperature of a place increases. Also, As temperature increases, humidity decreases. Therefore, these variables have contrary relationship.

CLUSTER ANALYSIS

There are different kinds of clustering methods and in our analysis Kmeans clustering technique is used. Initially, scaling is applied to ensure that the clustering algorithm operates on a standardized and more comparable dataset, leading to more accurate and meaningful clustering results.

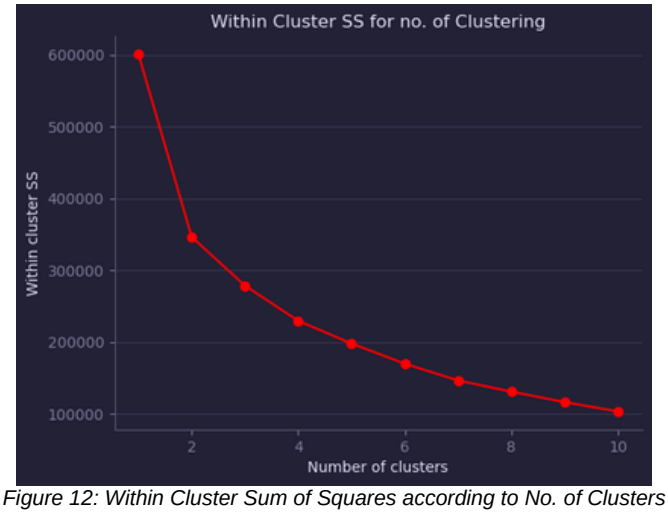

Figure 12: Within Cluster Sum of Squares according to No. of Clusters

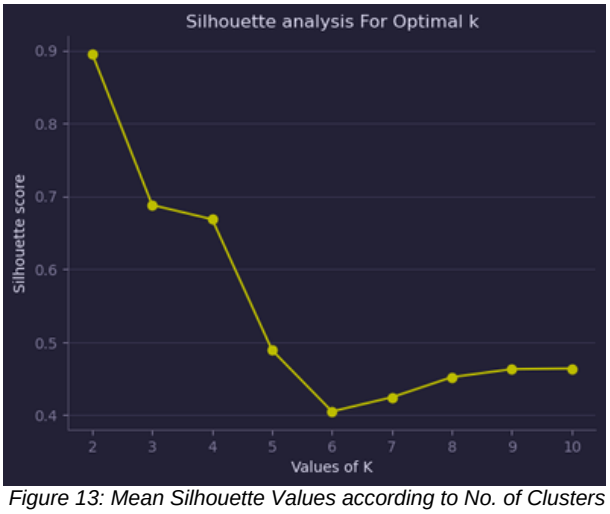

Figure 13: Mean Silhouette Values according to No. of Clusters

According to the within cluster sum of squares graph, the elbow point is located around two cluster solution and to be more precise, the mean silhouette values are also considered to choose the optimal number of clusters

Cluster 0 is dominated by observations corresponding to no fire and cluster 1 is dominated by observations where a fire is present.

SUGGESTIONS FOR A QUALITY ADVANCED ANALYSIS

- Logistic regression model will be fitted as an initial model. There are some highly correlated predictors, and we can try fitting the logistic regression model by dropping some of these highly correlated predictors. Moreover, we can also try fitting a model using the principal components we obtained.

- Random Forest can be used which is a strong ensemble learning method that can be used to solve a wide range of prediction problems, including classification. Because the method is based on an ensemble of decision trees, it offers all the benefits of decision trees, such as high accuracy, ease of use, and the absence of the need to scale data.

- Furthermore, it is evident that there are clusters in our dataset. In this case, we can try fitting separate models for each cluster and check whether the accuracy of the models increase when compared with a single model without clustering.

APPENDIX : Github link for python code and dataset