# DATA VISUALIZATION AND MODELLING FOR A SMOKE DETECTOR DATASET

By

Kathiresan A/L Kaniappan

A REPORT

SUBMITTED TO CIK ZANARIAH BINTI ZAINUDIN

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONOURS)

Faculty of Information and Communication Technology

(Kampar Campus)

JANUARY 2024

UNIVERSITI TUNKU ABDUL RAHMAN

# REPORT STATUS DECLARATION FORM

**Title**: Data Visualization and Modelling of a Smoke Detector Dataset

**Academic Session**: Jan 2024

I KATHIRESAN A/L KANIAPPAN

**(CAPITAL LETTER)**

declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,

(Author's signature)

(Supervisor's signature)

**Address**:
24, Sri Klebang C/10, Grand Retreats 2,
31200, Chemor, Perak

ZANARIAH BINTI ZAINUDIN
Supervisor's name

**Date**: 17/4/2024

**Date**: 24/4/2024

| Universiti Tunku Abdul Rahman | | | |
|---|---|---|---|
| Form Title : **Sample of Submission Sheet for FYP/Dissertation/Thesis** | | | |
| Form Number: **FM-IAD-004** | Rev No.: **0** | Effective Date: **21 JUNE 2011** | Page No.: **1 of 1** |

**FACULTY/INSTITUTE* OF INFORMATION AND COMMUNICATION TECHNOLOGY**

**UNIVERSITI TUNKU ABDUL RAHMAN**

Date: 17/4/2024

## SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS

It is hereby certified that ***KATHIRESAN A/L KANIAPPAN*** (ID No: ***20ACB03079*** ) has completed this final year project/ dissertation/ thesis* entitled "*Data Visualization and Modelling of a Smoke Detector Dataset*" under the supervision of Cik Zanariah Binti Zainudin (Supervisor) from the Department of Digital Economy Technology, Faculty/~~Institute~~* of Information and Communication Technology , and ____________________ (Co-Supervisor)* from the Department of _________________________, Faculty/Institute* of ___________________________.

I understand that University will upload softcopy of my final year project / dissertation/ thesis* in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

Kathiresan A/L Kaniappan

# DECLARATION OF ORIGINALITY

I declare that this report entitled "**DATA VISUALIZATION AND MODELLING FOR A SMOKE DETECTOR DATASET**" is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature :

Name : Kathiresan A/L Kaniappan

Date : 17/04/2024

# ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my supervisor, Zanariah binti Zainudin who has given me this bright opportunity to engage in this project. It is my first step to establish a career in the field of data science and data visualisation. A million thanks to you.

To a very special person in my life, Angeline Chung Kah Yee, for her patience, unconditional support, and love, and for standing by my side during the duration of this study and overall trying times. Finally, I must say thanks to my mother, Logammal and my family for their love, support, and continuous encouragement throughout the course and duration of study and hereafter.

# ABSTRACT

Machine learning (ML) analytics dashboard is powerful tools to monitor, analyze, and communicate the results of your data-driven projects. It can help to track key metrics, visualize trends, identify outliers, and share insights with stakeholders about the dataset. This report aims to continue the progress of the first project report by building on the pre-processing techniques decided on in that report (Flooring and Capping) to prepare a smoke detector dataset for machine learning modelling. After testing out 3 varying models and analysing the output and results, it is decided that the Multilayer Perceptron Model (MLP) has the best performance out of all the models (approx. 92%+ accuracy), also when comparing it to the benchmark model. Furthermore, the output of the model and the model itself has been imported to PowerBI. A combination of Python scripts and PowerBI visualizations has been used to visualize the data in a comprehensible and informative manner showcasing information of the dataset, model performance and key attributes.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1

# Introduction

## 1.1 Problem Statement and Motivation

The primary problem that this project seeks to solve is the lack of easily digestible and meaningful information for the public to view regarding smoke detectors, and the conditions surrounding them. The currently available information regarding the statistic of smoke detectors do not really delve into the details related to the environment regarding the smoke detectors. Thus, the public are not able to derive much meaningful relationship and inferences about smoke detectors. Furthermore, the lack of information related the details regarding the environment surrounding the smoke detector is worrying as these factors play a vital role in how smoke detectors functions. Thus, it is important to have an easily understandable and digestible source of information related to smoke detectors.

The primary motivation of this research-based project is to make information that is normally hard to understand to be more digestible and easily understood by the public to help them make informed decisions. The purpose of this information is to ensure the public are more informed about the public safety since smoke detectors play a vital role in ensuring our safety as they can help us detect fires. Having detailed as well as digestible information readily available can help people make informed decisions regarding smoke detectors (purchases, recommendation, etc.). Furthermore, another motivation is to foster discussion regarding smoke detectors. If detailed information is easily accessible to the public, it is more likely to spark discussions that may result in further advancements in this field.

Another motivation behind this research-based project is to test whether environmental factor data related to smoke detectors are sufficient to help in classifying if a fire is present via a machine learning model. There are several outcomes that may arise from this modelling procedure, these include pattern recognition of the environmental factors that affect a fire starting and risk management. Furthermore, it further supports the primary goal of data driven decisions that permeates throughout this project as the results of this model would be vital in ensuring better decisions are made by the public.

## 1.2 Research Objectives

The primary objective of this project is to produce a dashboard that has summarised the information from the dataset to showcase the various relationships between the environmental factors that affect smoke detectors. This is because most available statistical data regarding the smoke detectors don’t go into detail regarding the smoke detectors and what affects them. The purpose of this dashboard is to show end-users statistics and data that is normally unknown. These data and information can be utilised to then make more informed decisions by consumers of these products or for the purpose of public safety.

Furthermore, another goal that this paper aims to achieve is to create a machine learning model that can classify whether given based of certain data the relates to the surroundings of a smoke detector, the model is able to classify if a fire is present.

## 1.3 Project Scope and Direction

The scope of this project involves several different aspects. Firstly, the main deliverable of this project is a dashboard. This dashboard will contain several key information regarding the dataset. This dataset will be pre-processed to fit the purpose of visualization. These include key charts and graphs that provide vital information related to the various relationships between the different environmental factors that affect the smoke detectors. The charts and graphs utilised will vary to best convey the information regarding that specific statistical relationship. Moreover, the original dataset will also be pre-processed once again for the purposes of machine learning model, where the goal would be to classify the presence of fire given certain environmental factors present a smoke detector.

Furthermore, this project will entail 2 more reports down the line that will outline the progress of this research-based project. The first report, FYP 1 report, will be done around the 30-40% completion of the project. This report will most likely detail the data pre-processing and data preparations of this assignment as it is a vital and time-consuming part of the project to ensure the reliability and accuracy of the data. The second report, FYP 2 report will be done with the completion of this project. This report will discuss the dashboard of this project and will go into detail regarding the findings of the various relationships in the datasets. Furthermore, the FYP2 Report will also discuss the results of the machine learning model that has been trained for this project. The results of the model will be compared to related and contemporary research done based on similar dataset or similar purposes.

## 1.4 Contributions

The primary contribution of this research-based project is to make a detailed dashboard regarding smoke detector statistics that take into consideration the many environmental factors that affect the smoke detector and its functionalities. This research-based project will look at an extensive dataset and visualise it to produce an extensive dashboard that can be used to derive meaningful and unseen relationships. Some examples of the environmental factors include particulate matter within a certain size (in micrometers), carbon dioxide concentration, temperature, pressure, as well as number concentration of particulate matters, just to name a few. It is a novel concept to utilise this detailed data to visualise the various patterns and relationship between the environmental factors. The currently available smoke detector statistics do not go into deep details regarding the relationships between the various environmental factors. More information regarding past and current studies related to this topic will be covered in the literature review section in more detail.

Furthermore, another contribution of this research-based project would be the machine learning model that will be created upon the completion of this project. This machine learning model would further support the main goal of enforcing data-driven decision-making when it comes to smoke detectors. This is because it will use real life collected data from smoke detectors regarding environmental factors.

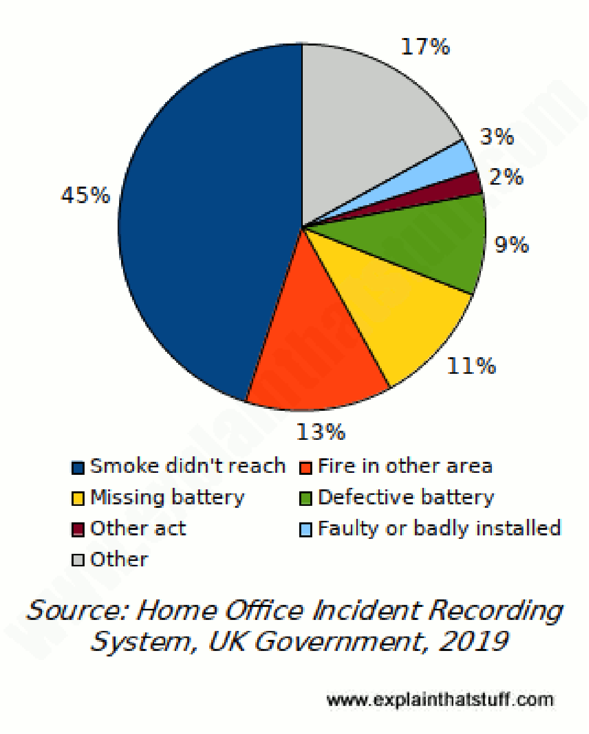
# CHAPTER 2

# Literature Reviews

For this section of the report, I will be explaining the previous statistical data that is currently available on smoke detectors, as well as some limitations and proposed solutions to these limitations.

## 2.1 Previous Works Related To Smoke Detectors

The first one that I will be reviewing is from the UK Government Home Office. As we can see from Figure 2.1, which showcases the chart that can be found from [9], the details are very limited as they only focus on a very simple statistic of why the smoke alarm did not function.



*Source: Home Office Incident Recording System, UK Government, 2019*

Figure 2.1 Chart on Why Smoke Detector did not Function

Even if the chart from [9] is lacking in detail, it still showcases some vital information to the public regarding safety. This is because it shows the possible reasons that can a smoke detector might fail to carry out its function. This information is very important as it can remind consumers to check their smoke detectors for faults.

The next material we will be reviewing is a report, [2] conducted regarding smoke alarms in US home fires. This report focuses mainly on the relation of house fires related to smoke alarms and the casualties that might have been caused as a result of the fire. The report goes into deep detail regarding the condition of the smoke detectors before the fire started, in terms of battery, its performance before the fire, the deaths caused as a result of faulty smoke detectors, effectiveness of smoke alarms in alerting residents and many more, and the reason for why it last sounded off. The report also goes in details to talk about the status of smoke alarm in reported fires, causes of smoke alarm fires, fire discovery, smoke alarms in context, and what counts as fires.

Due to the size of the report, to review the context of the report efficiently, I will be taking a look at some of the graphs and charts that are consistent with the rest of the data visualisation in the rest of the report. Listed below in Figure 2.2 and Figure 2.3 are some graphs that convey the most important information.

Experimental Results

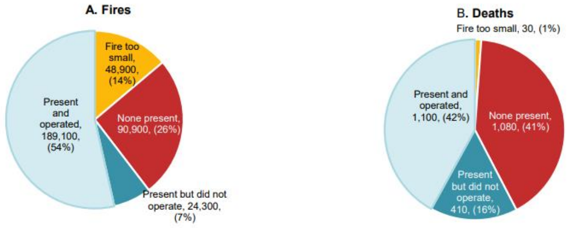


Figure 2.2 Reported Home Structure Fires and Fire Deaths by Smoke Alarm Performance 2014-2018

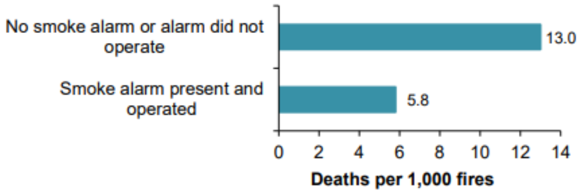


Figure 2.3 Death Rate per 1,000 Reported Home Fires by Smoke Alarm Status: 2014–2018

As we can see from Figure 2.2, we can see that the data visualised is easy to read and understand as well as visually appealing due to the colours and placing of wording and the informative labelling of pie chart components. Furthermore, we can surmise the data that is being showcased to the public is very important as it shows how smoke detector performance relates to casualties reported in home fires. This statistic is very important as it provides key information that contributes to public safety. Viewers can come up with meaningful inferences as well from the data, for example that the presence or lack of presence of smoke detectors have equal value in the pie chart that relates to the deaths in house fires, suggesting that the devices might not be very effective.

Moving on, based on Figure 2.3, like the other data visualisations in the report, we can see that is easy to understand and presents information in a digestible and informative way. Like the other data visualisation in the report as well, Figure 2.3 in [2] conveys a very strong message regarding public safety. The graph provides information using numerical measure to convey the information, allowing users to create inferences based on numbers. For example, lack of smoke alarm or lack of operating smoke alarm caused more than twice as many deaths per 1000 fires compared to functioning smoke detectors.

As we can see from the context of the report, we can see that it is very details regarding the relation of fires and smoke detectors. Much of the report focuses on the relation between how the smoke detector performs during a house fire and how it might have caused casualties due to faulty smoke detectors. This information is of high importance as users can derive many useful information regarding how a smoke detector function and how efficient they are during a real fire.

### **2.1.1 Limitations**

The data, while useful also has some negatives that hinder the potential impact it can provide to the public. For the first reviewed, as well as the other data reviewed, they are relatively low on details regarding the environment surrounding the smoke detector. However, for the first reviewed data in [11], the main limiting factor besides the lack of details regarding the surrounding environment is there is no significant inferences that can be made after viewing the data. It only serves as a reminder to test the functionality of the smoke detector. Furthermore, some of the data is slightly vague. For example, in the legend of the chart, there are two labels for Other and Other act, which are not differentiated and might cause confusion for the public.

While the information present in [2] is very important and is well delivered with plenty of informative graphs, the major drawback is that it doesn’t go into detail regarding the status of the smoke detector surroundings right as the fire goes out. This information can give consumers more insight into how the environment affects the functionality of the smoke detector in the presence of different environmental factors. This will increased information combined with the detailed insight of how the power source affects the effectiveness would have provided users with very insightful information. Another limitation noticed is that the first pie chart (Figure 2.2 A[Fires]), when the pie chart segments are added up it produces a total percentage of 101% which may confuse the viewers to the method the data is being visualised and represented.

### **2.1.2 Proposed Solutions**

The limitations of the reviewed data are unfortunate however there are some possible solutions that can be enacted to improve each of the previous works reviewed. For the data found in [11], the data can be improved by including some other statistic such as model type, or brand so that consumers can make more informed decisions. This can be done as an alternative to the surrounding environment statistics. Another step is to clarify the components of the pie chart, such as the legend so that consumers are not confused when viewing the data.

The limitation of the report contents in [2] are a hindrance towards the potential it holds towards providing the best possible information for the public and public safety.

There are some solutions that can be done to overcome its primary limitation. Which is to implement data if possible, regarding the environment of the fire and the surrounding of the smoke detector. As discussed in the limitation section for this report, it will provide much needed information that can give the public more information for them to analyse and derive inferences from. Furthermore, the pie chart should be calculated thoroughly to ensure that it stays at exactly 100% to prevent confusion.

### 2.1.3 Strengths

While the existing works reviewed have their limitations to how they can contribute to the purpose of this project, and they also have their strengths that may relate to the general objective that this project is trying to contribute to which is general health and safety for the public using smoke detections data. However, the data that is being represented in these existing works tackle the implications of the smoke detector functionality. Furthermore, the strengths of these existing may lie outside of the context of these infographics.

For example, based on Figure 2.1 and Figure 2.2, which are pie charts, the charts display their data in a way that is easily understood and graspable by the viewers. This was done by ensuring the that segments of the pie-chart are diverse in color from each other, making it easier to differentiate the colors. Furthermore, the charts display what each segment represents clearly. For Figure 2.1, a legend is used to associate each color of segment to its respective label while Figure 2.2, places the label into the pie chart segments directly, Due to a lower number of segments, this method is viable as it doesn't over-crowd the pie chart. Moreover, in the Figure 2.1

and Figure 2.2 (Pie Chart B[Deaths]) we can see the total segments of the pie charts added up to become 100% which follows the good design practices of data visualizations for pie charts.

Based on Figure 2.3, the information is displayed is concise and informative while using the necessary scales for the graph. Secondly, the bar charts are given with labels indicating what the data represents as well as the value it represents. This prevents any confusion on the side of the viewers.

## 2.2 Previous Works on Data Visualization and PowerBI

Based on the importance of data visualization for decision making in [11], the article expresses the implications and positive effects that data visualization has especially when it comes to aid in decision making for business. These data-driven decisions may improve the quality of the entire business. The article [11], mentions that data visualization improves the understanding of data the public has by reducing its complexity as well enhancing its readability and perception. Data visualizations aid various individuals from the public who might be analyzing stock market forecasts which are visualized to marketing leads who may be studying the sales data. It enhances data perception by revealing hidden trends that support prediction and analysis. This increased perception can be used to model the data in a way that future data may be predicted approximately, and this ability is crucial for business to make decisions based on this data. Often, stakeholders of companies who make the important decisions in a company may not understand the data just by looking at it in its raw format. Thus, data visualizations help in ensuring that even individuals who are not directly involved with the raw data will understand the implications.

Based on [14], there are also some disadvantages to data visualizations that should be avoided in this project. Some of these disadvantages are improper visualization, incorrect conclusions, and inexact representation.

Improper visualization can be briefly summarized as misrepresenting the data visually by either using the wrong techniques or using wrong scales for the data and visualization. These can be overcome by ensuring the individuals in charge of visualization, understand the data as well as have the required training in PowerBI to carry out the task.

Incorrect conclusions can be defined as viewers arriving to the wrong conclusion based of the visualization provided as visualized data can cause confusion if not explained properly. To tackle this issue, it is important to explain the aspects of the visualization clearly and ensure that the information displayed is relevant to the goals of the visualization.

Based on [14], it also states that data visualization may cause inexact representation which is the misrepresentation of numerical data. To combat this, it is vital to ensured that the visualization is labelled clearly to provide detailed information. Furthermore, on a technical level, it should be ensured that the numerical data is pre-processed effectively before visualizing it.

## 2.3 Previous Works on Smoke Detector Data for Machine Learning Model

Due to the nature of this project, there are not many sources that have researched this dataset that this project is heavily relies on. Thus, to serve as a benchmark to this study, especially when it comes to the data modelling aspect, the closest paper available will be used as a point of reference for the study as well as the results of the completed model.

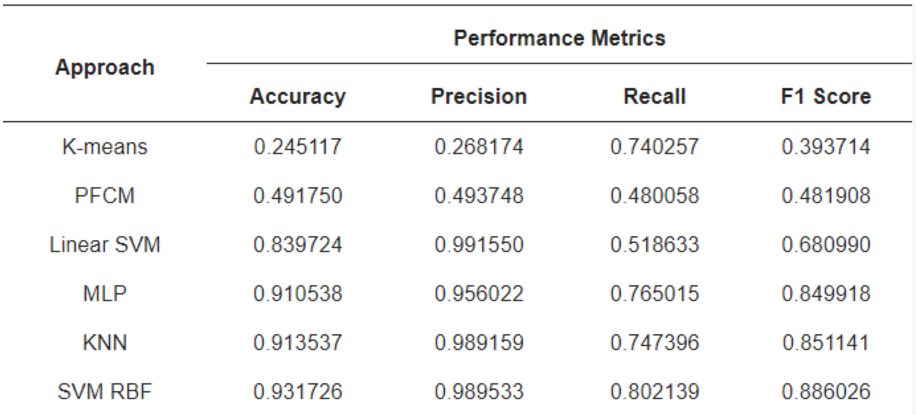
The research paper that fits these criteria the best is the paper titled “Detection of Smoking in Indoor Environment Using Machine Learning” by Jae Hyuk Cho. This paper is identified as the best candidate for a point of reference due to the similarities of the purpose of the modelling as well as the similarities found in the dataset. While the dataset used is not the exact same many of the attributes used are very similar in the data used by the paper, as well as this project. The similar attributes are that of TVOC, CO2, PM2.5, Temperature and Humidity data as well. However, it should be noted that the data this project utilizes, also has PM data in other scales such as PM1.0

The primary purpose of [6] is to identify cigarette smoke emissions by detecting the typical gases (CO2, total volatile organic compounds, etc.) with the use of machine learning approach to classify the presence of cigarette smoke. It should be noted that the data used for the analysis and machine learning was collected via IoT sensors that collected the desired data. The data was collected by placing the relevant sensors (ICT-Particulate Matter Sensor (SPS 30), ICT-TVOC sensor (SVM30), etc.) into a smoking machine (BORGWALDT’s Smoking Machine RM200A2). The mentioned sensors would record the data from the vaping chamber where to smoke is released.

When it comes to the machine learning aspect of [6], the paper specified that they have used Min-Max Scaler to pre-process the data. The justification for this decision is due to the possibility of abnormal values in the data due to the low reliability and high sensitivity of the IoT sensors. The paper suggests that these values should be eliminated and treated as an outlier.

As for the models used in [6], they have opted for 6 models using both supervised and unsupervised learning. These models are linear SVM, K-means clustering, PFCM (Possibilistic Fuzzy C-means), MLP (Multiplayer Perceptron) , KNN ( K-Nearest Neighbors), SVM RBF (Support Vector Machine Radial Basis Function). The paper also suggest that they have used Rectified Linear Unit (ReLu) as the activation function.

The table below shows the results and performance of the models used by the research in [6]



| Approach | Performance Metrics | | | |
|---|---|---|---|---|
| | Accuracy | Precision | Recall | F1 Score |
| K-means | 0.245117 | 0.268174 | 0.740257 | 0.393714 |
| PFCM | 0.491750 | 0.493748 | 0.480058 | 0.481908 |
| Linear SVM | 0.839724 | 0.991550 | 0.518633 | 0.680990 |
| MLP | 0.910538 | 0.956022 | 0.765015 | 0.849918 |
| KNN | 0.913537 | 0.989159 | 0.747396 | 0.851141 |
| SVM RBF | 0.931726 | 0.989533 | 0.802139 | 0.886026 |

Figure 2.4 Performance Metrics of Model from [6]

Experimental Results

The table below shows the comparison of the performance of the models used by the research in [6].

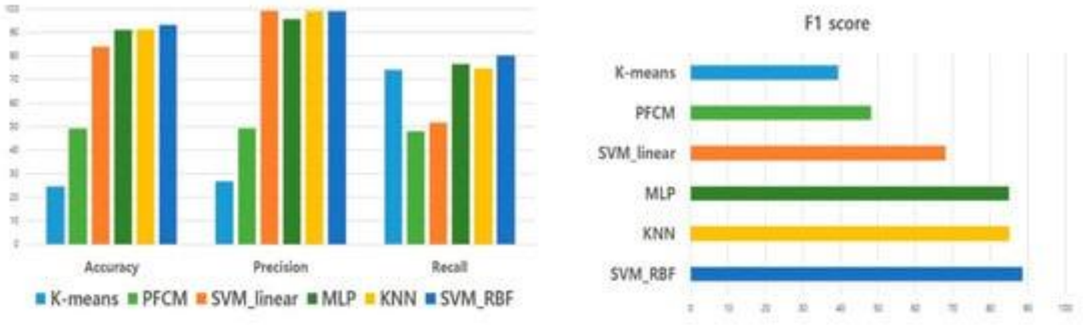


Figure 2.5 Comparison of Performance Metrics of Model from [6]

The graph below shows the ROC Curve of the MLP, Linear SVM and SVM RBF models from [6].

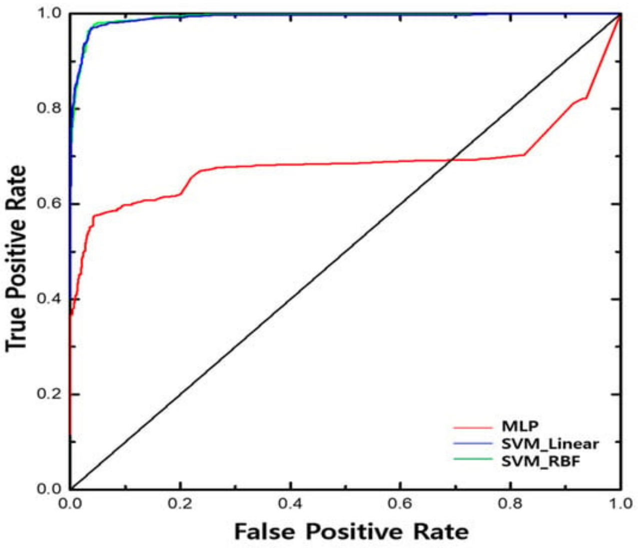


Figure 2.6 ROC Curve of several models from [6]

Based on the results from the models, the SVM RBF showed the most promising results out all of those that was tested.

As to how the research conducted in [6], relates to this research paper, there can be several vital similarities between these 2 studies that may aid us in the completion of this project. Firstly, while the dataset used by the research in [6] is not available for view, all the attributes listed in [6], are also part of the dataset of this project (as available to see in Chapter 4). The attributes mentioned in [6] are TVOC, CO2, Temperature, Humidity, PM1.0 and PM2.5. Thus, in terms of attributes 100% of the attributes found in [6], are part of the attributes in this project. As for the other attributes found in our dataset, that are not present in the data from [6], they will either be removed during pre-processing as they do not further the visualization or modelling purposes of this project, or their significance will be measured later on during pre-processing.

Furthermore, beyond similar attributes and scales, the data from [6] also faces a similar issue in their research when it comes to outlier data collected by the IoT sensor due to the low reliability and high sensitivity of the sensors. Thus, we may experiment on similar approaches used in [6] to treat its outliers and view its effects on our project. Moreover, the data used in [6] and this research project both are collected from IoT sensors, further proving similarities between both projects.

Finally, an important similarity between both projects is the nature of the model which is a binary classification model. Where in [6], the purpose is to detect the presence of cigarette smoke based on the IoT data of the environment surrounding the smoke, in this research paper the purpose is to model and visualize the model that can binarily classify the presence of smoke from IoT smoke detector data. Both purposes while different in some areas, are quite similar in concept, especially when we consider the similarity of the dataset and method of obtaining the data. Thus this is why this paper, was chosen as a point of reference and benchmark for this study.

# CHAPTER 3

# System Model

The processes of the project were categorized into different phases in the development, which were project pre-development, data exploration, data pre-processing, data analysis and visualization.

## 3.1 Overview

Methodology is one of the most vital processes of conducting research as it serves as a guideline to research process. This chapter outlines the methodology I have conducted throughout this research.

## 3.2 Research Framework for  Data Visualization and Modelling

### 3.2.1 Research Framework Flowchart

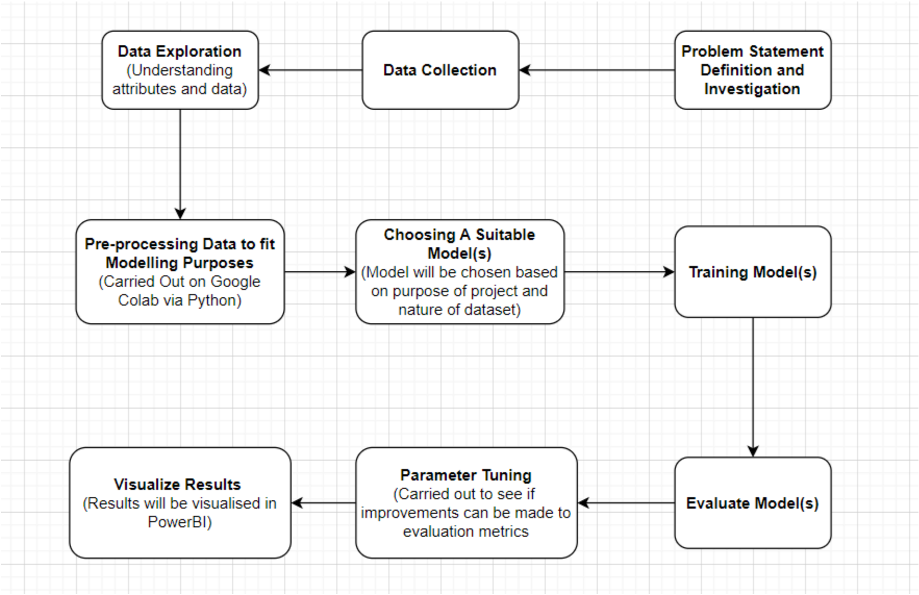

Problem Statement Definition and Investigation
Data Collection
Data Exploration (Understanding attributes and data)
Pre-processing Data to fit Modelling Purposes (Carried Out on Google Colab via Python)
Choosing A Suitable Model(s) (Model will be chosen based on purpose of project and nature of dataset)
Training Model(s)
Evaluate Model(s)
Parameter Tuning (Carried out to see if improvements can be made to evaluation metrics
Visualize Results (Results will be visualised in PowerBI)

Figure 3.1 Research Framework

## 3.3 System Model for Data Visualization and Modelling

### 3.3.1 Problem Statement and Definition

The primary step is to identify a problem statement and motivations that this project would like to attempt to tackle. In this case, this paper aims to solve 2 primary problems, which are to visualize data that is normally inaccessible so that it is easier to understand, and this project also aims to test whether environmental factor data related to smoke detectors are sufficient to help in classifying if a fire is present via a machine learning model(s). The success of the model will be evaluated in later sections.

### 3.3.2 Data Collection

The data collection process is important as it involves getting the data that would be at the center of the project. The data collected should have its initial purpose clearly defined, as well as abiding by laws of copyright and intellectual property. The source of the data should also be reputable and be able to be tracked backed to its source. In the case of this project, the data was sourced from a Kaggle page that explains the content of the data and its purpose, which is to serve as a dataset for a machine learning project for binary classification and that the data was collected using IoT sensors (smoke detectors).

### 3.3.3 Data Exploration

Data Exploration is carried out to understand the attributes of the dataset and the various interactions that are present in the dataset. From this, it is possible to know which attributes to prioritize when it comes to training the desired model. For this project, Point-

Biserial Correlation is used to check the correlation between the attributes in the dataset and the target variable.  According to [10], Point-Biserial Correlation is appropriate for data where the data is mostly numerical continuous data while the target variable is a binary variable. This mirrors the situation of this project; thus, it was used to find out the correlation.

### 3.3.4 Data Pre-processing

Before analysis and visualization can be done on the smoke detector dataset, an important step is pre-processing the dataset. Based on the article in [5], pre-processing is an important step as it improves accuracy and reliability as well as the consistency of the dataset and consequently whatever the dataset is used for. Since by the end of this project, visualizations will be made for the general public's use, these qualities are vital. The data pre-processing includes cleaning missing values as mentioned above as well as data transformation such as smoothing to remove noise from the dataset. Aggregation, discretization, and normalization techniques will be employed to make the data easier to work with for example in making specific columns of the dataset and combining them into bins to make them easier to read. The data pre-processing method will be carried out mainly via Python language in Google Colab to carry out statistical analysis on the data.

Based on [5], it also states that pre-processing data for a machine learning project is important, especially in the early stages to ensure the data that is being used to train and subsequently test the model is reliable and accurate. In the case of this project, this will mostly entail dealing with the outliers, and the approach used will be heavily influenced by [6], due to the similarities of the dataset.

### 3.3.5 Choosing Suitable Model(s)

Choosing the appropriate model for a machine learning model is a vital step as it must fit the purpose of the project, while also providing with acceptable results. Based on Figure 6 (The machine learning algorithms cheat sheet to find the appropriate algorithms) from [6]. Some of the models that are appropriate for this project (classification algorithm) are SVM, KNN, Naïve Bayes, Neural Network, and many other models. However, due to similar datasets and purposes of projects between this project and [6], the models used for this project will be like those used in [6].

### 3.3.6 Training Model(s)

Training the model(s) chosen is the primary step of any machine learning model. The pre-processed data will be split into a training and test set, each consisting of the input variables and the target variables. Training will be carried using the training data in various models and the performance of the models will be compared and discussed in later sections of this report.

### 3.3.7 Evaluate Model(s)

After training the models for the project, the test data is used to evaluate the performance. Several evaluation metrics will be used including Confusion Matrix, ROC Curve, Precision, Recall, F1 Score, FPR and others. The best performing model will be decided based on the most desirable results and justification will be given based on the purposes of this project in later sections.

### 3.3.8 Parameter Tuning

Based on [7], there are various important reasons why parameter tuning is a vital part of the machine learning process. This is because hyperparameters have a noticeable impact on the performance, speed, and requirements of a model. Tweaking and adjusting the parameters such as batch sizes and learning rate may improve the overall performance of a given model. Thus, it is an essential part of this project to ensure the model is performing as good as possible.

### 3.3.9 Data Visualization

Data Visualization is a key part of the project as it is one of the final deliverables for this project in the form of a dashboard that is currently aimed to have at least 2 important visualizations that contain key information that may be useful the to the public. For the visualization process, the software that will be heavily used will be PowerBI, this is due to its built in Python script editor that can help with the visualization process.

# CHAPTER 4

# System Design

## 4.1 Overview

This section will go through the model choosing, training, and testing process of this project. Models chosen will be discussed along with the parameters and justifications behind carry out the specifications of the modelling. The results of the models will be discussed and analyzed in further detail in upcoming chapters.

## 4.2 Model(s) Selection

Three primary models will be chosen for this classification project, these models are Support Vector Machine model using the Radial Basis Function kernel, K-Nearest Neighbors model, and MLP model. These models are chosen due to their success in the experiment in [6]. The reason we have chosen the models used in this paper are due to the similarities of objectives and data used in the research in [6] and this paper as well.

### 4.2.1 SVM RBF Model

Based on [12], SVM with RBF kernel is suitable in classification tasks due to it being a powerful machine learning algorithm that utilizes non-linear and high dimensional data as well. This model works by mapping the input data into a high dimensional space where the classes can be separated by the hyperplane while RBF kernel function is used to measure the similarity between the data into previously mentioned high dimensional feature space. The RBF kernel is defined as:

$$K\ (x, x') = exp\ (- gamma\ ||\ x - x'\ ||^2)$$

The gamma variable is a hyperparameter that determines the width of the kernel while the values in the ||.|| are used to calculate the Euclidian Distance between the points. The x and x' are the data points.

### 4.2.2 KNN Model

This model was chosen due to its impressive performance in the study in [6], being the second-best performing model in the study. Based on the study, they have employed KNN model using the input attributes as a vector and have set the value of k = 3, utilizing Euclidian Distance. However, one thing to note is that KNN model is incredibly sensitive to outliers and data imbalances [1], and since the data used in the project has persistent issues of large ranges between the data, even after pre-processing to cap and floor the outliers, this may cause issues in the model's training and testing results.

### 4.2.3 MLP Model

Based on [6], their third best performing model is an Multilayer Perceptron (MLP). The model had performed respectably given their dataset and with the similarities between the datasets between this paper and [6], this model has been selected for the purpose of this project. As an introduction, MLP can be defined as a Feed-Forward Neural Network that is a generalized form of a Linear Model that takes several various steps before making decisions. Normally, datasets with average size and complexity would require only a single hidden layer. However, due to the high complexity and size of the dataset used by this project, we would have to opt for more layers and neurons.

## 4.3 Model(s) Training

### 4.3.1 SVM RBF Model

The SVM model uses the SVC function that is available from the sklearn.svm library. There are many parameters that contribute to the effectiveness and accuracy of the model, but in this scenario, only 2 parameters will be adjusted. These hyperparameters are the kernel value which will be set to 'rbf' as that is the kernel function that is desired, and the C value which functions as the regularization value will be set to 10.0 as it produces the better results when compared to other values.

### 4.3.2 KNN Model

The KNN model was trained using the pre-processed data from the previous phase of the model and an alternate dataframe that used the pre-processed data that was scaled suing StandardScaler function. This is due to the KNN model being sensitive to outliers and data imbalance [1], which this data has. However, the 2 different versions of data used did not produce varying results (will be discussed in upcoming chapters). For the parameters of the model, the only change that has been made is the number of neighbors. The other parameters such as weights, p (power) value, metrics value and others were tested but there were no changes to the results of the model. The only parameter that had changes to results was the number of neighbors which was set to the square root of the number of records in the data [8], which in this case the closest value is 250. However, one thing to note is that the results of this model did not change much when n_neighbours values change from 3 to 250+. The only changes that can be noticed is when the n_neighbours value was set very high (e.g 2500). Thus, it can be determined that there is an issue faced by this model when trying to train using the data, the possible reasonings and justifications will be studied and provided in the analysis of results section.

### 4.3.3 MLP Model

The MLP model used for this project had several versions, and the final version will used the model that has better overall performance. The previous versions of the MLP model had some changes such as having larger number of neurons in each layer, having more layers, having larger/lower number of epochs for training, having larger/smaller values for learning rate and different batch sizes and shuffling training data and carrying out regularization techniques using Dropout() functions.

The final version of the MLP model had 2 hidden layers with the first layer having 64 neurons and the second layer having 32 neurons. The activation function used for the 2 hidden layers are ReLU and the activation function used for output layer was sigmoid function. ReLU was used as it helps the model to learn complex data better by introducing non-linearity. Sigmoid function was used due to its ability to translate the results into a probability value that is stored into the list y_pred. If the predicted value has a value greater than 0.5, it is listed as 1 and anything less than that is listed as 0 in the final predicted values list of y_pred_binary. The “Adam” (Adaptive Moment Estimation) function was used due to its ability to update the networks weights iteratively based on the training data and its effectiveness in handling sparse gradients [1]. The learning rate was set to 0.001 due to experimentation with higher and lower learning rates providing worse results. The loss perimeter as set to ‘binary_crossentropy’ because of the purpose of this model being binary classification. Moreover, accuracy was used as the metrics as it is a simple and efficient way to view the correctness of a model. Furthermore, for the model fitting, 20 epoch was used as it provided better results than 10 epochs and larger number of epochs caused the accuracy to decrease and stay stagnant without any changes. Finally, the batch size used was 32.

# CHAPTER 5

# Experiment Results

## 5.1 Results of SVM RBF Model

The table below shows the evaluation metrics of the SVM Model that has been trained and tested using the pre-processed data as well as the parameters from the previous chapter.

| Evaluation Metrics | Values |
|---|---|
| Accuracy | 0.8349 |
| Precision | 0.8544 |
| Recall | 0.8349 |
| F1 Score | 0.8134 |

Table 5.1 Tabulation of Results of SVM RBF Model

Based on the results of the model we can see that the SVM RBF Model has performed respectably well when we consider that the dataset used by this model is quite large and that SVM is not preferred for large datasets due to the high computational costs and worse performance [4] associated with training. The increase in computational costs can be seen with the training time for this model being approximately 1 minute + compared to the sub 1 minute time of the other models tested.

Based on Table 5.1, we can see that the SVM RBF model has an accuracy of 83.49% which can be interpreted as having relatively high correctness is prediction of all classes. The 85.44% precision score of the model suggests that the model can accurately identify the positive classes correctly based of all positively classified results. Next, the recall score of 83.49% suggests that the model can effectively capture a big portion of the positive instances of the dataset. Recall score is a vital score in instances where it is important to identify positive cases accurately. Finally, the F1 Score of 81.34 % indicates there is a good balance between the Precision and Recall metrics. Since the data used has an imbalance between the positive and negative target values, F1 Score is a vital score since it considers the negative and positive instances of the results.

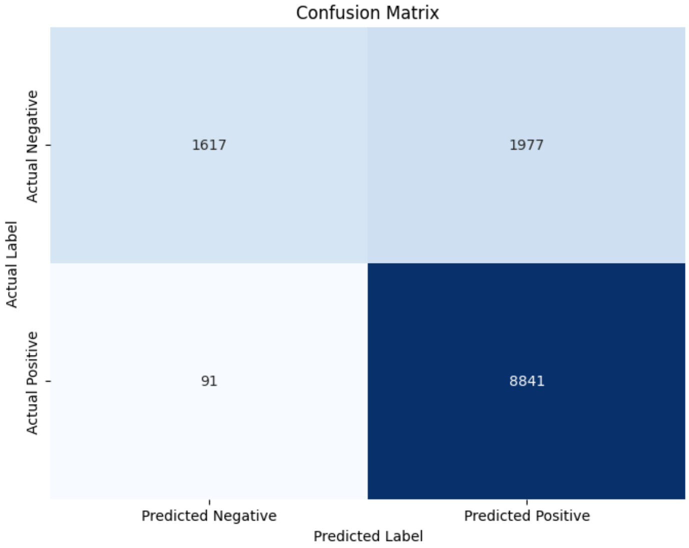


Figure 5.1 Confusion Matrix of SVM RBF Model

We can see that based on Figure 5.1, that the SVM RBF Model has classified the actual positive values accurately, but it has many false positives (more than the True Negative Values predicted), suggesting that due to the inherit data imbalance of the data, the model is choosing the majority target value(positive) instead of predicting accurately. This may affect the integrity of the correctly predicted values in the Predicted Positive – Actual Positive Values as most of it may be classified as so due to the data imbalance. When it comes to the other values, the model has performed quite well as it boasts a high TPR (Recall) of 83.49% which is moderately good for minimizing false negatives which is important as this would mean less cases where the model predicts no fire when there is actually a fire. The SVM RBF model has a FPR of 55% which is quite high and is not desirable.

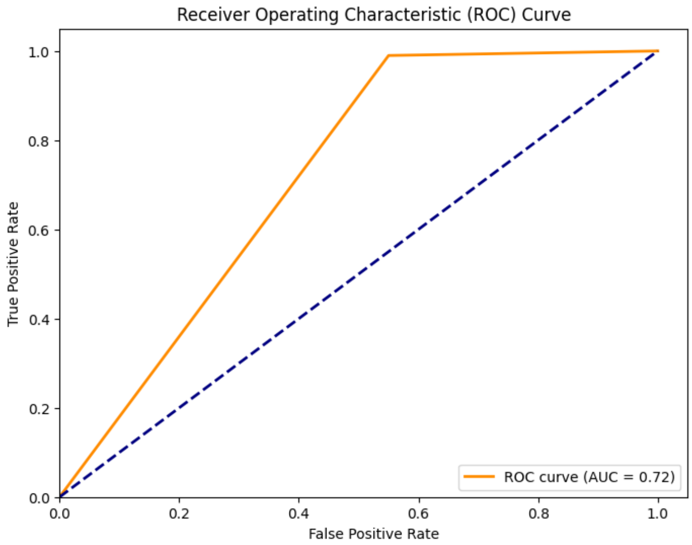


Figure 5.2 ROC Curve of the SVM RBF Model

The figure above shows the ROC Curve of the SVM RBF Model which is a visualization for tradeoff between the true positive rate and the false positive rate. Generally, a ROC Curve that lies closer to the top left and an AUC score that is high is desirable. The SVM RBF model shows a decent AUC Score of 0.72, however the curve displayed by the model, while good, still has a lot of room for improvement.

The AUC Score and ROC Curve should be compared to the other models to choose the best possible model for the purpose of this project.

## 5.2 Results of KNN Model

| Evaluation Metrics | Values |
|---|---|
| Accuracy | 0.9973 |
| Precision | 0.9973 |
| Recall | 0.9973 |
| F1 Score | 0.9973 |

Table 5.2 Tabulation of Results of KNN Model

In a previous section (4.2.2), it is known that this model is sensitive to cases where the dataset has certain issues, in particular, data imbalances and outliers/large ranges in the data. Thus, it is expected that the KNN model will have some difficulties in the training phase of the model. Once again, it should be noted that the performance of the results did not change much even after changing the number of neighbours value by a large amount during different iterations of training as well as experimenting with different proportions of training and testing splits for the data.

Based on table 5.2, we can notice that there is an abnormality in the results on the KNN Model. It produces the exact same results for all evaluation metrics of the model after testing the model with the testing data. Some may attribute these results as an overwhelming success in the terms of the model's ability to classify the target variable. However, due to the inherit problems in the dataset caused by the high sensitivity of the equipment used to collect the data, the data contains a lot of noise, outliers, and imbalances that may affect the training effectiveness of the KNN model which is already sensitive to these issues. This pattern of extremely high scores permeates throughout the other evaluation metrics such as the Confusion Matrix and the ROC Curve. Thus, it would be safe to assume this model is not a suitable fit for the classification purpose of this project.

It should be noted that when performing k-cross validation the scores from the validation process also repeat the similar pattern of extremely high scores across all the folds, on multiple ranges of number of folds. The mean of the fold scores stay consistently at 99% with a low standard deviation of 0.0011 suggesting high consistency between all the folds. A point of note is that the values provided are for a split of 5 in the k-cross validation, but the scores and standard deviation do not change drastically with increased or decreased number of folds.

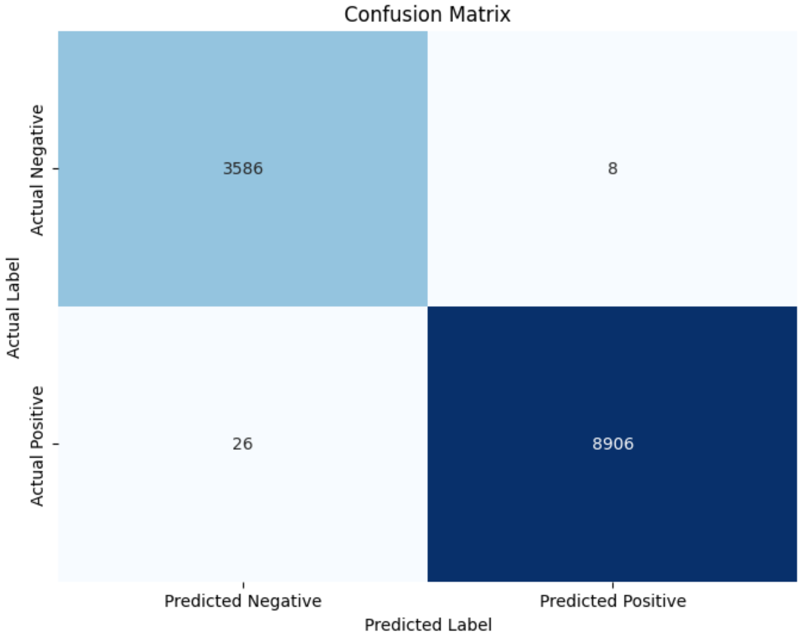


Figure 5.3 Confusion Matrix of KNN Model

Based of the confusion matrix in figure 5.3, it is apparent again that the model is performing too well as the model has performed quite well as it boasts a high TPR (Recall) of 99.73% which is quite good for minimizing false negatives classifications by the model and an even better FPR of 0.22% which is too good.

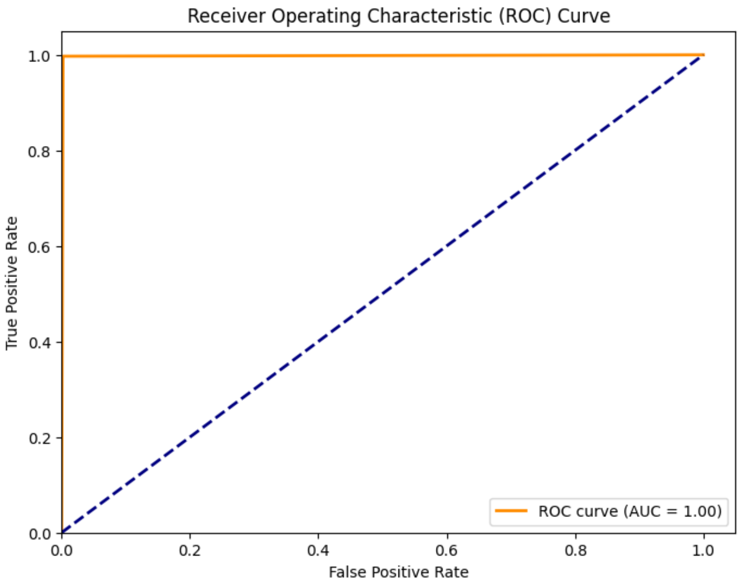


Figure 5.4 ROC Curve of KNN Model

Based on the Figure 5.4, we can see that the KNN Model tested has produced a perfect ROC Curve with a perfect AUC Score of 1.00. An AUC score of 1.00 indicates that the model can accurately classify all the test data provided to the corresponding target variable. However, due to the model's sensitivity to outliers, imbalances, and noise, it would be safe to assume that this AUC Score is not reliable and that other models should be considered.

## 5.3 Results of MLP Model

| Evaluation Metrics | Values |
|---|---|
| Accuracy at 20th EPOCH | 0.9420 |
| Precision | 0.9377 |
| Recall | 0.9841 |
| F1 Score | 0.9603 |

Table 5.3 Tabulation of Results of KNN Model

Based on the table above we can see the results of the MLP (Multilayer Perceptron) model is quite good, especially with the inherit issues with the dataset as mentioned in other sections. However, one thing to note is that there is a degree of variety in the results of the model. The results tabulated are the mean results of the model performance over 5 separate and isolated training and testing session The range of the results is around 90.4% ~ 95.2%.

Based on the above table 5.3, we can see that the results obtained the evaluation metrics gained from MLP (Multilayer Perceptron) Model are quite high, with an accuracy of 94.2% at 20th Epoch of training, with the accuracy increasing overtime as each epoch passes with occasional dips and rises. After each epoch, the loss value decreases steadily. Furthermore, the precision score for the MLP Model is 93.77% indicates that the model is effective in ensuring less false positive errors which is useful for the purpose for this project as it would mean less cases where the fire is detected when there is no fire, which may cause unnecessary issues in real life scenarios. The MLP Model has a recall score of 98.41 % which is quite high and suggests that the model can accurately capture a large majority of the positive instances in the dataset. Next, the model has a F1 Score of 96.03%, indicating a good balance between precision and recall

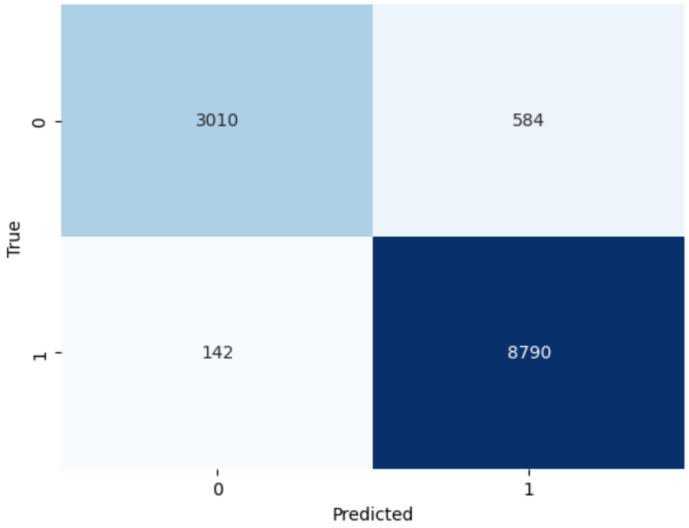


Figure 5.5 Confusion Matrix for MLP Model

Conclusion

Based of the Figure 5.5, we can see that the precision and recall scores viewed in the Table 5.3 are reflected in the Confusion Matrix by the low number of false negatives and false positives. For the purpose of this project, the number of false negatives should be as low as possible within a well performing model as it would indicate less cases where a fire is classified as no fire based on the smoke detector dataset. In that regard, the MLP model has performed quite well as it boasts a high TPR (Recall) of 98.41% which is quite good for minimizing false negatives. The MLP Model provides a FPR of 16.25% which while could be improved upon is still generally acceptable. Overall, the model performs its classification tasks quite well and still has some room for improvement in terms of reducing FPR even further.

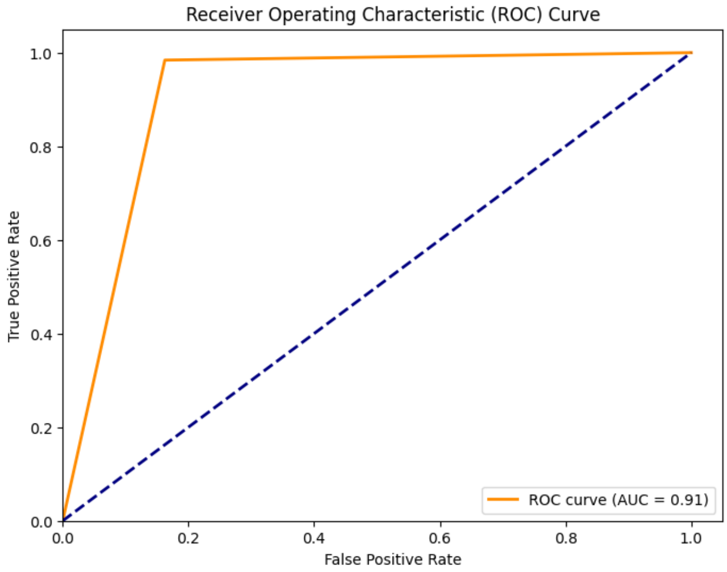


Figure 5.6 ROC Curve of MLP Model

Based of the ROC Curve in Figure 5.6, it is apparent the model has performed quite well with an AUC Score of 0.91 indicating it can effectively distinguish between positive and negative instances. Furthermore, based of the curve that sits quite close to the top left corner of the plot, it indicates the model's performance is quite good.

## 5.4 Comparison of Model Results

Based on the tables 5.1, 5.2 and 5.3, we can conclude several interesting findings regarding the model performance of the SVM RBF model, KNN model and the MLP model. Firstly, we can deduce based on the results of the KNN model that the model is unfit for the purpose of this project due to the unreliable results and the model's performance that is too good to be true, even with performing k-cross validation to prevent overfitting. A possible cause of this phenomenon is data leakage which refers to a scenario where the training set unintentionally contains information of the test set, causing the model to be trained on the validation set and the model memorising the outcome instead of truly learning from the data provided [13]. This issue can be overcome by finding new test data for this KNN model, that is not a part of the original dataset. Thus, at the current point of this project, it is recommended not to use the KNN Model.

The next deduction that can be made from the results of the models are the MLP model performs better than the SVM RBF model. The results of the SVM RBF model are acceptable however the MLP model outperforms it in almost all ways as we can see based on the evaluation metrics and confusion matrixes, especially when it comes to the MLP model's higher TPR and lower FPR scores. The SVM RBF model has a slightly lesser number of false negative classifications, which is good, however it is a marginally larger number of false positives as can be seen by its higher FPR. Furthermore, the MLP model also boasts a better AUC Score (0.91) than the SVM RBF model's AUC Score (0.72), indicating a model that performs a reasonable amount better. These findings that out the models tested for this project that the MLP is the best performing model and relates back unto the research objectives set out for this project, being to train a model effectively using the environmental data surrounding the smoke detector.

| | Project Model | Benchmark Model |
|---|---|---|
| Accuracy | 0.8349 | 0.9317 |
| Precision | 0.8544 | 0.9895 |
| Recall | 0.8349 | 0.8021 |
| F1 Score | 0.8134 | 0.8860 |

Table 5.4 Comparison of Results of SVM RBF Model between Project and Benchmark

| | Project Model | Benchmark Model |
|---|---|---|
| Accuracy | 0.9973 | 0.9135 |
| Precision | 0.9973 | 0.9891 |
| Recall | 0.9973 | 0.7474 |
| F1 Score | 0.9973 | 0.8511 |

Table 5.5 Comparison of Results of KNN Model between Project and Benchmark

| | Project Model | Benchmark Model |
|---|---|---|
| Accuracy | 0.9420 | 0.9105 |
| Precision | 0.9377 | 0.9560 |
| Recall | 0.9841 | 0.7650 |
| F1 Score | 0.9603 | 0.8499 |

Table 5.6 Comparison of Results of MLP Model between Project and Benchmark

| 🟥 | 🟨 | 🟩 |
|---|---|---|
| Inferior Performance | Flawed/Issue Performance | Better performance |

Table 5.7 Color Coded Meanings

When we relate back to [6], which serves as a point of reference to this project due to its similar dataset, purpose, data collection method, and models used, it can be observed that the MLP model trained in this project surpasses the performance of the MLP model trained in that research. This can be see clearly based on Tables 5.4, 5.5 and 5.6 which show the clear comparison between the model performance in the benchmark model and this project. We can see that our models perform better (except for KNN model which has an issue that has been discussed) in most or at least some regards than the benchmark model as can be seen with the color coded comparisons.Furthermore, the MLP model also outperforms all the other models trained in [6]. The high scores of the MLP model are quite exemplary considering the characteristics of the dataset, that would make training a model difficult.

In conclusion, based of all the results from the model training, as well as the results from [6], it can be concluded that the MLP model trained has performed very well in classifying the presence of fire based of the smoke detector dataset.

## 5.5 Feature Importance for Visualization

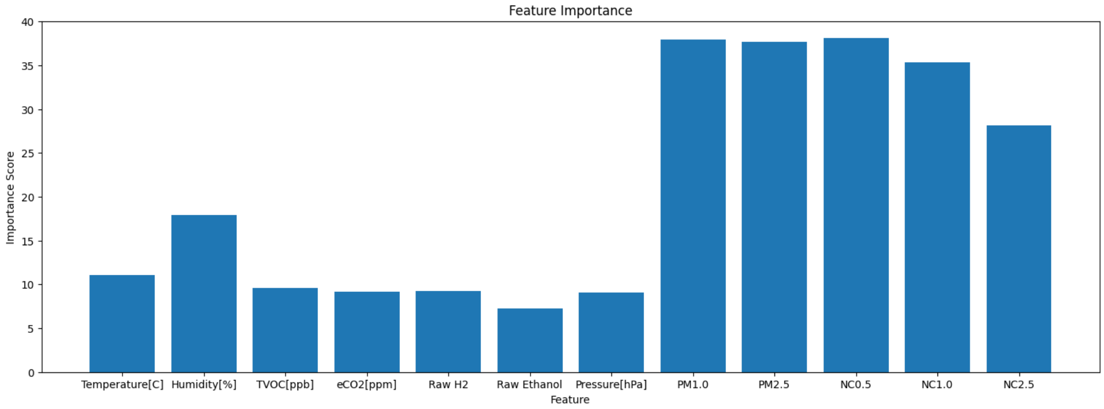


Figure 5.7 Feature Importance of MLP Model

The graph above showcases the Feature Importance of each attribute of the MLP model that has been chosen for the purposes of this project, These attributes will be used as a basis for the upcoming visualizations of the model and the dataset in PowerBI. Due to the PM and NC attributes using similar data just in different scales, the paper will be using the common smallest scale for both attributes which are PM1.0 AND NC1.0. Furthermore, we will also be using the Humidity [%], Temperature [C] and TVOC [ppb] attributes for the upcoming visualizations.

## 5.6 Visualization

The PowerBI visualization based of the model consists of 3 distinct parts that are visualized in the PowerBI file. These parts are the Dataset Summary, Distribution of Key Attributes, and the Model Information. The dataset summary page consists of several visualizations that highlight the nature of the datasets and some information that highlights the fundamentals of the dataset. These are a count of each output of the model, a correlation heatmap showcasing the correlation of important variables as outlined in

5.5, a table that functions as a header of the dataset, and a diagram showcasing each of the attributes that make up the dataset.

The next part, Distribution of Key Attributes, aims to show the viewers the distribution in the form of boxplots for the important attributes. This is due to the initial heavy presence of outliers in the original dataset. This visualization would show the new distribution of the dataset.

The next part, which is Model Information, serves to showcase information regarding the model performance. This is done via key visualizations such as ROC Curve, Precision-Recall Curve, a table that shows the evaluation metrics of the model, a pair plot of important variables and the predicted outputs as well as a confusion matrix. Something that should be noted is that due to limitations of PowerBI, it is unable to plot a confusion matrix in the traditional python way, thus to achieve this goal, the use of PowerBI cards were used with some advanced filtering to get the desired results.

# CHAPTER 6

# Conclusion

## 6.1 Summary of Project

In conclusion to this project, it can be determined that the pre-processed data from Project I, and the insights from the EDA conducted during the duration of the project has been a vital step in ensuring that the data (after undergoing refinements in pre-processing) is suitable for the machine learning aspect of the research paper. The models chosen for this project are based of the 3 best performing models from the benchmark paper, have proven to have interesting results in the outcomes of the experiments with this project's MLP Model outperforming the benchmark's MLP model after finetuning the hyperparameters of the model to ensure the best possible results. Furthermore, the data from the modelling phase has been showcased in the visualization aspect of the project in the PowerBI file, that summarizes the Dataset Summary, Distribution of Key Attributes, and the Model Performance.

## 6.2 Future Work

The project has achieved good results overall, but improvements can be made to further improve the project in the future. Firstly, instead of using a pre-made dataset, it may be more useful to use a new dataset that is created specifically for the purpose of this project with the use of good quality sensors, equipment, and environment to ensure the data does not have issues such as outliers or imbalances, like this data used currently.

Furthermore, another improvement that can be made with more experimentation and time is to arrive at the MLP model that has a lesser number of False Positives as while it is still a good performing model, the high number of False Positives remains its only flaw as of now. A False positive in this case would mean that the model predicts there’s a fire when there is none. In a practical scenario this would cause owners of a smoke detector inconvenience as it would falsely identify presence of fire.

Finally, the biggest enhancement that can be made with the project is to implement the model and its capabilities into small devices that may carry out its functions in predicting presence of fire based on environmental factors. Then, it would visualize its results into an app or dashboard to enable viewers to understand the data more easier which would strengthen all of our goals and objectives we have set out of this project

# REFERENCES

[1] “Adam,” *DeepAI*, May 17, 2019. https://deepai.org/machine-learning-glossary-and-terms/adam-machine-learning#:~:text=Adam%2C%20short%20for%20Adaptive%20Moment (accessed Mar. 12, 2024).

[2] Ahrens, M. (2021). Smoke Alarms in US Home Fires (NFPA ® ). https://www.nfpa.org//-/media/Files/News-and-Research/Fire-statistics-and-reports/Detection-and-signaling/ossmokealarms.pdf

[3] C. L. Karr, J. M. Sanil, and R. E. Barlow, "Visualizations in Bayesian analysis," in Proceedings of the 21st Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, 2001, pp. 143-154.

[4] I. Ahmad, M. Basheri, M. J. Iqbal and A. Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," in IEEE Access, vol. 6, pp. 33789-33795, 2018, doi: 10.1109/ACCESS.2018.2841987. keywords: {Intrusion detection;Support vector machines;Radio frequency;Training;Forestry;Kernel;Detection rate;extreme learning machine;false alarms;NSL–KDD;random forest;support vector machine}

[5] Indeed Editorial Team, What is data preprocessing? (with importance and examples), https://ca.indeed.com/career-advice/career-development/data-preprocessing (accessed Jul. 16, 2023).

[6] J. H. Cho, “Detection of Smoking in Indoor Environment Using Machine Learning,” *Applied Sciences*, vol. 10, no. 24, p. 8912, Dec. 2020, doi: https://doi.org/10.3390/app10248912.

[7] N. Kasture, “Why Hyper parameter tuning is important for your model ?,” *Analytics Vidhya*, Nov. 16, 2020. https://medium.com/analytics-vidhya/why-hyper-parameter-tuning-is-important-for-your-model-1ff4c8f145d3

[8] R. Rahim and A. S. Ahmar, “Cross-validation and validation set methods for choosing K in KNN algorithm for Healthcare Case Study,” *JINAV: Journal of Information and Visualization*, vol. 3, no. 1, pp. 57–61, Jul. 2022. doi:10.35877/454ri.jinav1557

[9] UK Government Home Office: Data sheet Fire0704: Percentage of smoke alarms that did not operate in primary dwelling fires and fires resulting in casualties in dwellings, by type of alarm and reason for failure, September 2

[10] “Point-Biserial Correlation in SPSS Statistics - Procedure, assumptions, and output using a relevant example.,” *statistics.laerd.com*. https://statistics.laerd.com/spss-tutorials/point-biserial-correlation-using-spss-statistics.php

[11] S. Sharma, “Data Visualization and its impact in decision making in business,” ResearchGate, Mar. 2023, doi: 10.13140/RG.2.2.21906.12483.

[12] S. Eskandar, “Introduction to RBF SVM: A Powerful Machine Learning Algorithm for Non-Linear Data,” *Medium*, Mar. 26, 2023. https://medium.com/@eskandar.sahel/introduction-to-rbf-svm-a-powerful-machine-learning-algorithm-for-non-linear-data-1d1cfb55a1a#:~:text=RBF%20SVM%20works%20by%20mapping

[13] Swetha, “Data Leakage and its effect on Machine Learning models,” *Medium*, Aug. 15, 2023. https://medium.com/@swethac42/data-leakage-and-its-effect-on-machine-learning-models-67c8edc588d4 (accessed Mar. 22, 2024).

[14] “What are the advantages and disadvantages of data visualization?,” Tableau. https://www.tableau.com/data-insights/data-visualization/advantages-disadvantages

APPENDIX

# POSTER

FACULTY OF INFORMATION TECHNOLOGY
FINAL YEAR PROJECT ii

SUPERVISOR : CIK ZANARIAH BINTI ZAINUDDIN

STUDENT: KATHIRESAN A/L KANIAPPAN

## DATA VISUALIZATION AND MODELLING OF A SMOKE DETECTOR DATASET

BACHELOR OF COMPUTER SCIENCE (HONS)

### MOTIVATIONS

- SOLVE LACK OF EASILY UNDERSTANDABLE VISUAL DATA RELATED TO SMOKE DETECTOR
- TEST IF SMOKE DETECTOR DATA AVAILABEL IS SUFFICIENT FOR TRAINING MACHINE LEARNING MODELS

### OBJECTIVES

- CREATE DASHBOARD THAT CAN VISUALIZE KEY DATA
- CREATE MACHINE LEARNING MODEL TO CLASSIFY PRESENCE OF FIRE BASED ON DATA

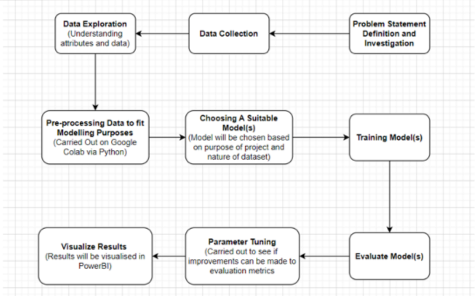
### SCOPE

- DASHBOARD IN POWERBI
- MACHINE LEARNING MODEL

### CONTRIBUTIONS

- ENFORCING DATA-DRIVEN DECISION MAKING BASED OF THE DASHBOARD AND MACHINE LEARNING MODEL

### RESEARCH FRAMEWORK


Problem Statement Definition and Investigation
Data Collection
Data Exploration (Understanding attributes and data)
Pre-processing Data to fit Modelling Purposes (Carried Out on Google Colab via Python)
Choosing A Suitable Model(s) (Model will be chosen based on purpose of project and nature of dataset)
Training Model(s)
Evaluate Model(s)
Parameter Tuning (Carried out to see if improvements can be made to evaluation metrics
Visualize Results (Results will be visualised in PowerBI)

### SYSTEM DESIGN

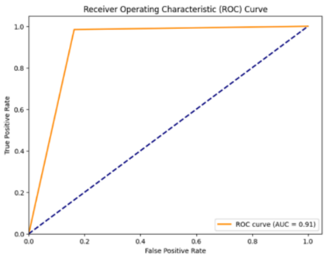
3 MODELS FOR TRAINING

- **SVM RBF MODEL** : BEST PERFORMING MODEL IN BENCHMARK AND POWERFUL CLASSFICATION MODEL
- **KNN MODEL** : HIGH PERFORMANCE IN BENCHMARK, SENSITIVE TO OUTLIERS AND DATA IMBALANCE
- **MLP MODEL**: THIRD BEST PERFROMING MODEL IN BENCHMARK, ARCHITECTURE BASED ON COMPLEXITY OF DATA

### EXPERIMENT RESULTS

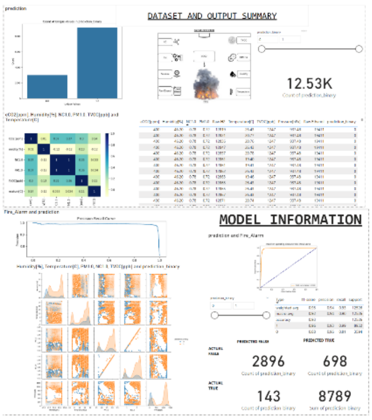
BEST PERFORMING MODEL: MLP, OUTPERFORMING ALL BENCHMARK MODELS

| Evaluation Metrics | Values |
|---|---|
| Accuracy at 20th EPOCH | 0.9420 |
| Precision | 0.9377 |
| Recall | 0.9841 |
| F1 Score | 0.9603 |



### VISUALIZATION

DONE IN POWERBI BY LOADING MODEL INTERNALLY



AND MORE (DISTRIBUTION OF KEY ATTRIBUTES)

### CONCLUSION

- PREPROCESSING USED: FLOORING AND CAPPING
- MODEL: MLP
- VISUALIZED IN POWERBI

# FINAL YEAR PROJECT WEEKLY REPORT

*(Project II)*

| | |
|---|---|
| **Trimester, Year: 3,3** | **Study week no.: 2** |
| **Student Name & ID: Kathiresan A/L Kaniappan, 2003079** | |
| **Supervisor: Cik Zanariah Binti Zainuddin** | |
| **Project Title: DATA VISUALIZATION AND MODELLING FOR A SMOKE DETECTOR DATASET** | |

| |
|---|
| **1. WORK DONE**<br>-Discussed project 2 goals and lessons learn from moderator feedback from project 1 |
| **2. WORK TO BE DONE**<br>-perform further pre-processing if needed for the new purpose of modeling as well as the original goal of visualization |
| **3. PROBLEMS ENCOUNTERED**<br>- distribution of previously pre-processed data may not be fit for modelling |

**4. SELF EVALUATION OF THE PROGRESS**

-The next few weeks will be busy with the pre-processing data and training, so it is better to complete the previous sections of the report earlier, especially since they have to be revised for the addition of the modelling purpose.

____________________ ____________________

Supervisor’s signature Student’s signature

# FINAL YEAR PROJECT WEEKLY REPORT

*(Project II)*

| | |
|---|---|
| **Trimester, Year: 3,3** | **Study week no.: 4** |
| **Student Name & ID: Katbiresan A/L Kaniappan, 2003079** | |
| **Supervisor: Cik Zanariah Binti Zainuddin** | |
| **Project Title: DATA VISUALIZATION AND MODELLING FOR A SMOKE DETECTOR DATASET** | |

| |
|---|
| **1. WORK DONE**<br>-Discussed with supervisor regarding the dataset new pre-processed dataset and the technique used for it. |
| **2. WORK TO BE DONE**<br>**-** study the relevance of the attributes for the data in terms of machine learning. Point Biserial Correlation can be used as it fits the criteria of the modelling purpose and dataset. |
| **3. PROBLEMS ENCOUNTERED**<br>- Encountered issues with pre-processing data as it took multiple techniques and tries to find an appropriate technique |

**4. SELF EVALUATION OF THE PROGRESS**

-Fostered thinking outside the box to solve the issue resolved at this time

Supervisor’s signature

Student’s signature

# FINAL YEAR PROJECT WEEKLY REPORT

*(Project II)*

| | |
|---|---|
| **Trimester, Year: 3,3** | **Study week no.: 8** |
| **Student Name & ID: Katbiresan A/L Kaniappan, 2003079** | |
| **Supervisor: Cik Zanariah Binti Zainuddin** | |
| **Project Title: DATA VISUALIZATION AND MODELLING FOR A SMOKE DETECTOR DATASET** | |

**1. WORK DONE**

-Completed the revised literature review and research methodology sections that will be used throughout this project

-After discussion with supervisor, 3 models have been decided for this project based on the benchmark paper, which are SVM RBF, KNN and MLP

-Have started training models

**2. WORK TO BE DONE**

-Finish training models

-Evaluate model performance

**3. PROBLEMS ENCOUNTERED**

-KNN Model has proved slight difficult to train due its sensitivities to outliers are data imbalances

**4. SELF EVALUATION OF THE PROGRESS**

- Spent a lot of time experimenting and problem solving to solve issues with training the models for SVM and KNN, managed to solve the issue for SVM however, the KNN model seems difficult based on the dataset issues.

Supervisor’s signature

Student’s signature

# FINAL YEAR PROJECT WEEKLY REPORT

*(Project II)*

| | |
|---|---|
| **Trimester, Year: 3,3** | **Study week no.: 10** |
| **Student Name & ID: Kathiresan A/L Kaniappan, 2003079** | |
| **Supervisor: Cik Zanariah Binti Zainuddin** | |
| **Project Title: DATA VISUALIZATION AND MODELLING FOR A SMOKE DETECTOR DATASET** | |

**1. WORK DONE**

-After training and evaluating results, the best performing model, MLP has been chosen as the appropriate model and has been chosen for the visualization

-Exported model to PowerBI

**2. WORK TO BE DONE**

-Visualize results in PowerBI

**3. PROBLEMS ENCOUNTERED**

- Issues finding appropriate file type for the model to be exported in PowerBI

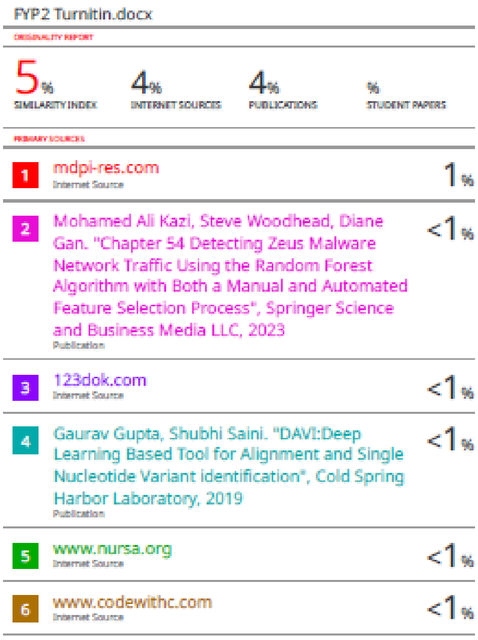
**4. SELF EVALUATION OF THE PROGRESS**

- Visualization requires creative thinking as well as design ideas that will help users get the important information. Carrying out this will require me to learn and apply new ideas and knowledge, especially when it comes to this unfamiliar software.

Supervisor’s signature

Student’s signature

# PLAGIARISM CHECK RESULT



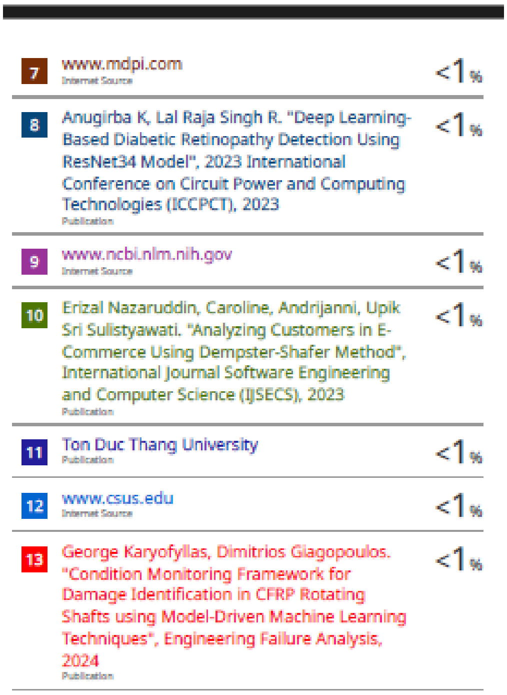
FYP2 Turnitin.docx

ORIGINALITY REPORT

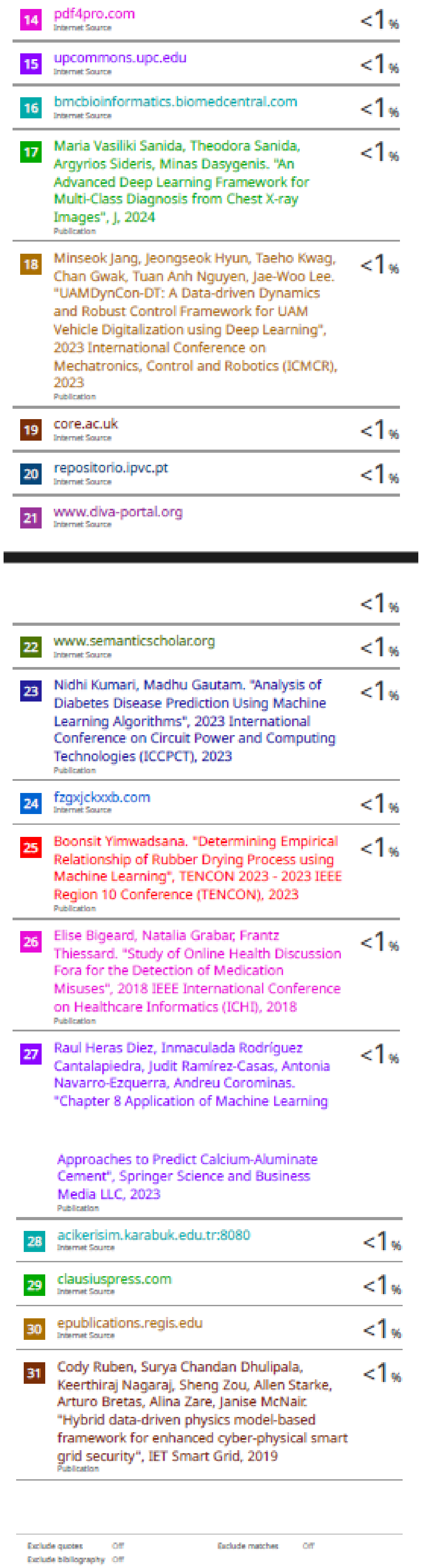
| 5% | 4% | 4% | % |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | mdpi-res.com<br>Internet Source | 1% |
| 2 | Mohamed Ali Kazi, Steve Woodhead, Diane Gan. "Chapter 54 Detecting Zeus Malware Network Traffic Using the Random Forest Algorithm with Both a Manual and Automated Feature Selection Process", Springer Science and Business Media LLC, 2023<br>Publication | <1% |
| 3 | 123dok.com<br>Internet Source | <1% |
| 4 | Gaurav Gupta, Shubhi Saini. "DAVI:Deep Learning Based Tool for Alignment and Single Nucleotide Variant identification", Cold Spring Harbor Laboratory, 2019<br>Publication | <1% |
| 5 | www.nursa.org<br>Internet Source | <1% |
| 6 | www.codewithc.com<br>Internet Source | <1% |



| | | |
|---|---|---|
| 7 | www.mdpi.com<br>Internet Source | <1% |
| 8 | Anugirba K, Lal Raja Singh R. "Deep Learning-Based Diabetic Retinopathy Detection Using ResNet34 Model", 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), 2023<br>Publication | <1% |
| 9 | www.ncbi.nlm.nih.gov<br>Internet Source | <1% |
| 10 | Erizal Nazaruddin, Caroline, Andrijanni, Upik Sri Sulistyawati. "Analyzing Customers in E-Commerce Using Dempster-Shafer Method", International Journal Software Engineering and Computer Science (IJSECS), 2023<br>Publication | <1% |
| 11 | Ton Duc Thang University<br>Publication | <1% |
| 12 | www.csus.edu<br>Internet Source | <1% |
| 13 | George Karyofyllas, Dimitrios Giagopoulos. "Condition Monitoring Framework for Damage Identification in CFRP Rotating Shafts using Model-Driven Machine Learning Techniques", Engineering Failure Analysis, 2024<br>Publication | <1% |



| | | |
|---|---|---|
| 14 | pdf4pro.com<br>Internet Source | <1 % |
| 15 | upcommons.upc.edu<br>Internet Source | <1 % |
| 16 | bmcbioinformatics.biomedcentral.com<br>Internet Source | <1 % |
| 17 | Maria Vasiliki Sanida, Theodora Sanida, Argyrios Sideris, Minas Dasygenis. "An Advanced Deep Learning Framework for Multi-Class Diagnosis from Chest X-ray Images", J, 2024<br>Publication | <1 % |
| 18 | Minseok Jang, Jeongseok Hyun, Taeho Kwag, Chan Gwak, Tuan Anh Nguyen, Jae-Woo Lee. "UAMDynCon-DT: A Data-driven Dynamics and Robust Control Framework for UAM Vehicle Digitalization using Deep Learning", 2023 International Conference on Mechatronics, Control and Robotics (ICMCR), 2023<br>Publication | <1 % |
| 19 | core.ac.uk<br>Internet Source | <1 % |
| 20 | repositorio.ipvc.pt<br>Internet Source | <1 % |
| 21 | www.diva-portal.org<br>Internet Source | <1 % |
| 22 | www.semanticscholar.org<br>Internet Source | <1 % |
| 23 | Nidhi Kumari, Madhu Gautam. "Analysis of Diabetes Disease Prediction Using Machine Learning Algorithms", 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), 2023<br>Publication | <1 % |
| 24 | fzgxjckxxb.com<br>Internet Source | <1 % |
| 25 | Boonsit Yimwadsana. "Determining Empirical Relationship of Rubber Drying Process using Machine Learning", TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON), 2023<br>Publication | <1 % |
| 26 | Elise Bigeard, Natalia Grabar, Frantz Thiessard. "Study of Online Health Discussion Fora for the Detection of Medication Misuses", 2018 IEEE International Conference on Healthcare Informatics (ICHI), 2018<br>Publication | <1 % |
| 27 | Raul Heras Diez, Inmaculada Rodríguez Cantalapiedra, Judit Ramírez-Casas, Antonia Navarro-Ezquerra, Andreu Corominas. "Chapter 8 Application of Machine Learning Approaches to Predict Calcium-Aluminate Cement", Springer Science and Business Media LLC, 2023<br>Publication | <1 % |
| 28 | acikerisim.karabuk.edu.tr:8080<br>Internet Source | <1 % |
| 29 | clausiuspress.com<br>Internet Source | <1 % |
| 30 | epublications.regis.edu<br>Internet Source | <1 % |
| 31 | Cody Ruben, Surya Chandan Dhulipala, Keerthiraj Nagaraj, Sheng Zou, Allen Starke, Arturo Bretas, Alina Zare, Janise McNair. "Hybrid data-driven physics model-based framework for enhanced cyber-physical smart grid security", IET Smart Grid, 2019<br>Publication | <1 % |

| | | | |
|---|---|---|---|
| Exclude quotes | Off | Exclude matches | Off |
| Exclude bibliography | Off | | |

PLAGIARISM CHECK RESULT

| | | | |
|---|---|---|---|
| **Form Title: Supervisor's Comments on Originality Report Generated by Turnitin** | | | |
| Form Number: FM-IAD-005 | Rev No.: 0 | Effective Date: 01/10/2013 | Page No.: 1of 1 |

**FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY**

| | |
|---|---|
| **Full Name(s) of Candidate(s)** | Kathiresan A/L Kaniappan |
| **ID Number(s)** | 2003079 |
| **Programme / Course** | Computer Science |
| **Title of Final Year Project** | DATA VISUALIZATION AND MODELLING FOR A SMOKE DETECTOR DATASET |

| **Similarity** | **Supervisor's Comments (Compulsory if parameters of originality exceed the limits approved by UTAR)** |
|---|---|
| **Overall similarity index:__5___ % Similarity by source**<br>Internet Sources: __4___%<br>Publications: __4__%<br>Student Papers: __0___% | |
| **Number of individual sources listed** of more than 3% similarity: _0_ | |
| **Parameters of originality required, and limits approved by UTAR are as Follows:**<br>**(i) Overall similarity index is 20% and below, and**<br>**(ii) Matching of individual sources listed must be less than 3% each, and**<br>**(iii) Matching texts in continuous block must not exceed 8 words**<br>***Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.*** | |

Note: Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

***Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.***

| | |
|---|---|
| Signature of Supervisor | Signature of Co-Supervisor |
| Name: ZANARIAH BINTI ZAINUDIN | Name: __________________________ |
| Date: 24/4/2024 | Date: ___________________________ |

# UNIVERSITI TUNKU ABDUL RAHMAN

## FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY (KAMPAR CAMPUS)

## CHECKLIST FOR FYP2 THESIS SUBMISSION

| | |
|---|---|
| Student Id | 20ACB03079 |
| Student Name | Kathiresan A/L Kaniappan |
| Supervisor Name | Cik Zanariah Binti Zainudin |

| **TICK (√)** | **DOCUMENT ITEMS**<br>Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item. |
|---|---|
| √ | Title Page |
| √ | Signed Report Status Declaration Form |
| √ | Signed FYP Thesis Submission Form |
| √ | Signed form of the Declaration of Originality |
| √ | Acknowledgement |
| √ | Abstract |
| √ | Table of Contents |
| √ | List of Figures (if applicable) |
| √ | List of Tables (if applicable) |
| | List of Symbols (if applicable) |
| | List of Abbreviations (if applicable) |
| √ | Chapters / Content |
| √ | Bibliography (or References) |
| √ | All references in bibliography are cited in the thesis, especially in the chapter of literature review |
| √ | Appendices (if applicable) |
| √ | Weekly Log |
| √ | Poster |
| √ | Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005) |
| √ | I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report. |

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.

(Signature of Student)
Date: 24/04/2024