Text, logo

Description automatically generated

BSc (Hons) Computing

**SOUTHAMPTON SOLENT UNIVERSITY**

Data Analysis Report

Author : 10152420

Course title : Data Analysis, Tools, and Application

**Tutor’s name : Muhammad Jugon**

***Github: https://github.com/sarkersh/Data-Analysis-Tools-and-applications.git***

**Table of Contents**

[**1** **Introduction** 1](#_Toc124739406)

[**1.1** **Project Motivations/Problems** 1](#_Toc124739407)

[**1.2** **Project Aim** 1](#_Toc124739408)

[**1.3** **Project Objectives** 1](#_Toc124739409)

[**2** **Methodology** 2](#_Toc124739410)

[**2.1** **Data Collection** 2](#_Toc124739411)

[**2.2** **Data Preparation** 3](#_Toc124739412)

[**2.2.1** **Basic Data Analysis on Excel** 3](#_Toc124739413)

[**2.3** **Data Analysis** 14](#_Toc124739414)

[**2.3.1** **Logistic Regression** 14](#_Toc124739415)

[**2.3.2** **Decision Tree** 15](#_Toc124739416)

[**2.3.3** **Data Visualization on Power BI** 17](#_Toc124739417)

[**2.4** **Results** 20](#_Toc124739418)

[**2.5** **Conclusion** 21](#_Toc124739419)

[**3** **Project Management** 21](#_Toc124739420)

[**3.1** **Project Plan** 21](#_Toc124739421)

[**3.2** **Project Collaboration** 21](#_Toc124739422)

[**3.3** **Learning Reflection** 22](#_Toc124739423)

[**References** 23](#_Toc124739424)

[**Appendix** 25](#_Toc124739425)

# **Introduction**

## **Project Motivations/Problems**

By now it is clear to all that COVID-19 was deadly and an infectious disease. The infected people experienced respiratory illness, which varied from mild to moderate, and such individuals recovered without any need of special treatment. However, the older people, including the patients of chronic respiratory disease, diabetes, cardiovascular disease, and cancer suffered a lot and their illness increased.

Keeping all these aspects in mind, here the project intends to know if the patient has high risk or low risk based on the provided current symptom, status, and medical history of the COVID-9 patients. And knowing the risks can help to provide extra healthcare to the patients and make the necessary medical resources available.

But in such a tough time, the problem encountered by the healthcare providers is the difficulty of knowing the required medical resources needed by the positive tested patients. There was also shortage of medical resources, and a lack of proper plan for distributing the resources correctly to all. Knowing the details of the positive tested patients can be highly beneficial for the medical authorities as it helps in procuring and making the arrangement of required resources for saving patients’ lives.

## **Project Aim**

The aim of this project is to build the knowledge and gain understanding of real time analysis.

## **Project Objectives**

This project’s key objective is about building a best machine learning model, which can effectively predict if the patient is in high risk or not. This is determined based on the provided current symptoms of the COVID-19 patients, their medical history and their respective status.

# **Methodology**

## **Data Collection**

It is a process about collecting and measuring the information based on the interest, systematic fashion, which helps in answering the test hypothesis or the asked questions, and finally examining the outcomes. However, it is stressed on the collection of honest and accurate collection (*Data collection,* 2023).

The covid 19 dataset is collected from a popular website called Kaggle, and the respective dataset link is (Nizri, 2022):-

<https://www.kaggle.com/datasets/meirnizri/covid19-dataset>

This data is collected from the Mexican government, and its dataset has a huge amount of information associated to the anonymized patients, which also contains their pre-conditions. And this is one of the important points to be noted in this dataset, as it will help in determining the decisions. There are totally 21 features in this raw dataset, and comprises of the information of totally 1,048,576 patients, who are unique. The point to be noted in this dataset is that the Boolean feature represents 1 as Yes, and 2 and No.

1. **sex:** 1 and 2, female and male, respectively.
2. **age:** Patient’s age.
3. **classification:** Findings of the COVID test. The values 1to 3 denotes that COVID-19 was diagnosed in the patient in different degrees, whereas 4 or above denotes inconclusive test or the patient is not infected with COVID-19.
4. **patient type:** type of care the patient received in the unit. 1 for returned home and 2 for hospitalization.
5. **pneumonia:** whether the patient already have air sacs inflammation or not.
6. **pregnancy:** whether the patient is pregnant or not.
7. **diabetes:** whether the patient has diabetes or not.
8. **copd:** Indicates whether the patient has Chronic obstructive pulmonary disease or not.
9. **asthma:** whether the patient has asthma or not.
10. **inmsupr:** whether the patient is immunosuppressed or not.
11. **hypertension:** whether the patient has hypertension or not.
12. **cardiovascular:** whether the patient has heart or blood vessels related disease.
13. **renal chronic:** whether the patient has chronic renal disease or not.
14. **other disease:** whether the patient has other disease or not.
15. **obesity:** whether the patient is obese or not.
16. **tobacco:** whether the patient is a tobacco user.
17. **usmr:** Indicates whether the patient treated medical units of the first, second or third level.
18. **medical unit:** type of institution of the National Health System that provided the care.
19. **intubed:** whether the patient was connected to the ventilator.
20. **icu:** Indicates whether the patient had been admitted to an Intensive Care Unit.
21. **date died**: If the patient died indicate the date of death, and 9999-99-99 otherwise.

## **Data Preparation**

### **Basic Data Analysis on Excel**

Now, we are going to do some data analysis on excel that will help user to understand the data such as mean, standard error, median, mode, standard deviation, kurtosis, skewness, minimum, maximum and range, which as tabulated below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| USMER |  | MEDICAL\_UNIT | | SEX |  |
|  |  |  |  |  |  |
| Mean | 1.632194 | Mean | 8.980565 | Mean | 1.499259 |
| Standard Error | 0.000471 | Standard Error | 0.003636 | Standard Error | 0.000488 |
| Median | 2 | Median | 12 | Median | 1 |
| Mode | 2 | Mode | 12 | Mode | 1 |
| Standard Deviation | 0.482208 | Standard Deviation | 3.723278 | Standard Deviation | 0.5 |
| Sample Variance | 0.232525 | Sample Variance | 13.8628 | Sample Variance | 0.25 |
| Kurtosis | -1.69938 | Kurtosis | -1.63762 | Kurtosis | -2 |
| Skewness | -0.54829 | Skewness | -0.51569 | Skewness | 0.002962 |
| Range | 1 | Range | 12 | Range | 1 |
| Minimum | 1 | Minimum | 1 | Minimum | 1 |
| Maximum | 2 | Maximum | 13 | Maximum | 2 |
| Sum | 1711478 | Sum | 9416796 | Sum | 1572086 |
| Count | 1048575 | Count | 1048575 | Count | 1048575 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PATIENT\_TYPE | | INTUBED |  | PNEUMONIA | |
|  |  |  |  |  |  |
| Mean | 1.190765 | Mean | 79.52288 | Mean | 3.346831 |
| Standard Error | 0.000384 | Standard Error | 0.036005 | Standard Error | 0.011634 |
| Median | 1 | Median | 97 | Median | 2 |
| Mode | 1 | Mode | 97 | Mode | 2 |
| Standard Deviation | 0.392904 | Standard Deviation | 36.86889 | Standard Deviation | 11.91288 |
| Sample Variance | 0.154374 | Sample Variance | 1359.315 | Sample Variance | 141.9167 |
| Kurtosis | 0.477805 | Kurtosis | 0.666747 | Kurtosis | 60.43622 |
| Skewness | 1.574104 | Skewness | -1.63293 | Skewness | 7.898181 |
| Range | 1 | Range | 98 | Range | 98 |
| Minimum | 1 | Minimum | 1 | Minimum | 1 |
| Maximum | 2 | Maximum | 99 | Maximum | 99 |
| Sum | 1248606 | Sum | 83385699 | Sum | 3509403 |
| Count | 1048575 | Count | 1048575 | Count | 1048575 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| AGE |  | PREGNANT | | DIABETES |  |
|  |  |  |  |  |  |
| Mean | 41.7941 | Mean | 49.76558 | Mean | 2.186404 |
| Standard Error | 0.016511 | Standard Error | 0.046397 | Standard Error | 0.005297 |
| Median | 40 | Median | 97 | Median | 2 |
| Mode | 30 | Mode | 97 | Mode | 2 |
| Standard Deviation | 16.90739 | Standard Deviation | 47.51073 | Standard Deviation | 5.424242 |
| Sample Variance | 285.8598 | Sample Variance | 2257.27 | Sample Variance | 29.4224 |
| Kurtosis | 0.064148 | Kurtosis | -1.99985 | Kurtosis | 306.9149 |
| Skewness | 0.28356 | Skewness | -0.01136 | Skewness | 17.54368 |
| Range | 121 | Range | 97 | Range | 97 |
| Minimum | 0 | Minimum | 1 | Minimum | 1 |
| Maximum | 121 | Maximum | 98 | Maximum | 98 |
| Sum | 43824251 | Sum | 52182948 | Sum | 2292609 |
| Count | 1048575 | Count | 1048575 | Count | 1048575 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| COPD |  | ASTHMA |  | INMSUPR |  |
|  |  |  |  |  |  |
| Mean | 2.260569 | Mean | 2.242626 | Mean | 2.298132 |
| Standard Error | 0.005012 | Standard Error | 0.004994 | Standard Error | 0.005335 |
| Median | 2 | Median | 2 | Median | 2 |
| Mode | 2 | Mode | 2 | Mode | 2 |
| Standard Deviation | 5.132258 | Standard Deviation | 5.114089 | Standard Deviation | 5.462843 |
| Sample Variance | 26.34007 | Sample Variance | 26.15391 | Sample Variance | 29.84266 |
| Kurtosis | 343.8073 | Kurtosis | 346.2124 | Kurtosis | 302.7734 |
| Skewness | 18.59072 | Skewness | 18.64969 | Skewness | 17.45371 |
| Range | 97 | Range | 97 | Range | 97 |
| Minimum | 1 | Minimum | 1 | Minimum | 1 |
| Maximum | 98 | Maximum | 98 | Maximum | 98 |
| Sum | 2370376 | Sum | 2351562 | Sum | 2409764 |
| Count | 1048575 | Count | 1048575 | Count | 1048575 |

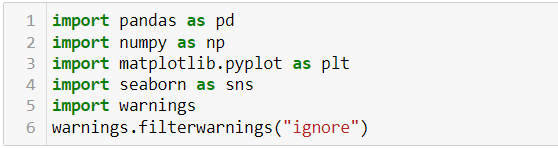
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| HIPERTENSION | | OTHER\_DISEASE | | CARDIOVASCULAR | |
|  |  |  |  |  |  |
| Mean | 2.128989 | Mean | 2.435143 | Mean | 2.26181 |
| Standard Error | 0.005114 | Standard Error | 0.006491 | Standard Error | 0.005073 |
| Median | 2 | Median | 2 | Median | 2 |
| Mode | 2 | Mode | 2 | Mode | 2 |
| Standard Deviation | 5.236397 | Standard Deviation | 6.646676 | Standard Deviation | 5.19485 |
| Sample Variance | 27.41985 | Sample Variance | 44.1783 | Sample Variance | 26.98647 |
| Kurtosis | 329.617 | Kurtosis | 202.6078 | Kurtosis | 335.4063 |
| Skewness | 18.16562 | Skewness | 14.29972 | Skewness | 18.36179 |
| Range | 97 | Range | 97 | Range | 97 |
| Minimum | 1 | Minimum | 1 | Minimum | 1 |
| Maximum | 98 | Maximum | 98 | Maximum | 98 |
| Sum | 2232405 | Sum | 2553430 | Sum | 2371677 |
| Count | 1048575 | Count | 1048575 | Count | 1048575 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| OBESITY |  | RENAL\_CHRONIC | | TOBACCO |  |
|  |  |  |  |  |  |
| Mean | 2.125176 | Mean | 2.25718 | Mean | 2.214333 |
| Standard Error | 0.005054 | Standard Error | 0.005015 | Standard Error | 0.005198 |
| Median | 2 | Median | 2 | Median | 2 |
| Mode | 2 | Mode | 2 | Mode | 2 |
| Standard Deviation | 5.175445 | Standard Deviation | 5.135354 | Standard Deviation | 5.323097 |
| Sample Variance | 26.78523 | Sample Variance | 26.37186 | Sample Variance | 28.33536 |
| Kurtosis | 337.534 | Kurtosis | 343.3665 | Kurtosis | 318.9616 |
| Skewness | 18.38079 | Skewness | 18.57758 | Skewness | 17.89128 |
| Range | 97 | Range | 97 | Range | 97 |
| Minimum | 1 | Minimum | 1 | Minimum | 1 |
| Maximum | 98 | Maximum | 98 | Maximum | 98 |
| Sum | 2228406 | Sum | 2366822 | Sum | 2321894 |
| Count | 1048575 | Count | 1048575 | Count | 1048575 |

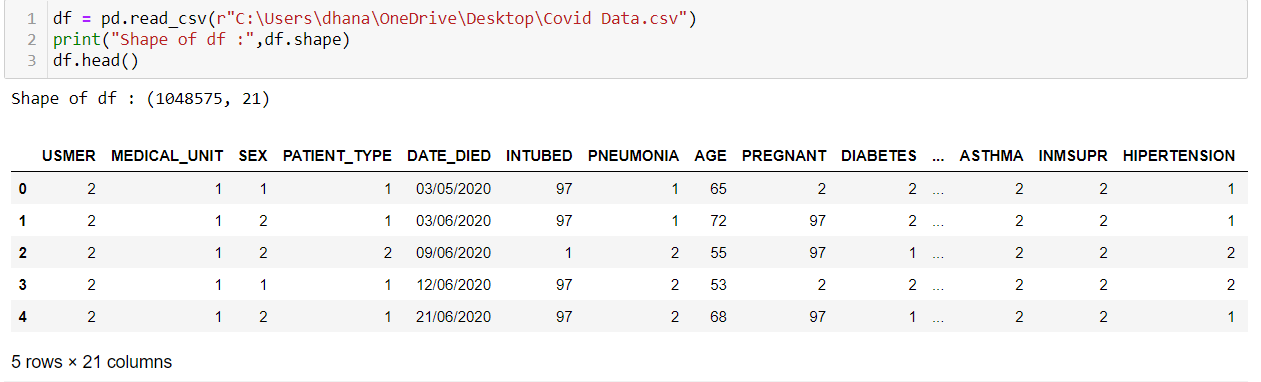
|  |  |  |  |
| --- | --- | --- | --- |
| CLASIFFICATION\_FINAL | | ICU |  |
|  |  |  |  |
| Mean | 5.305653 | Mean | 79.55397 |
| Standard Error | 0.001837 | Standard Error | 0.03596 |
| Median | 6 | Median | 97 |
| Mode | 7 | Mode | 97 |
| Standard Deviation | 1.881165 | Standard Deviation | 36.82307 |
| Sample Variance | 3.538783 | Sample Variance | 1355.939 |
| Kurtosis | -1.62064 | Kurtosis | 0.670949 |
| Skewness | -0.42492 | Skewness | -1.63424 |
| Range | 6 | Range | 98 |
| Minimum | 1 | Minimum | 1 |
| Maximum | 7 | Maximum | 99 |
| Sum | 5563375 | Sum | 83418308 |
| Count | 1048575 | Count | 1048575 |

Always, it is essential to prepare the data for the further analysis. And the steps are represented as follows:

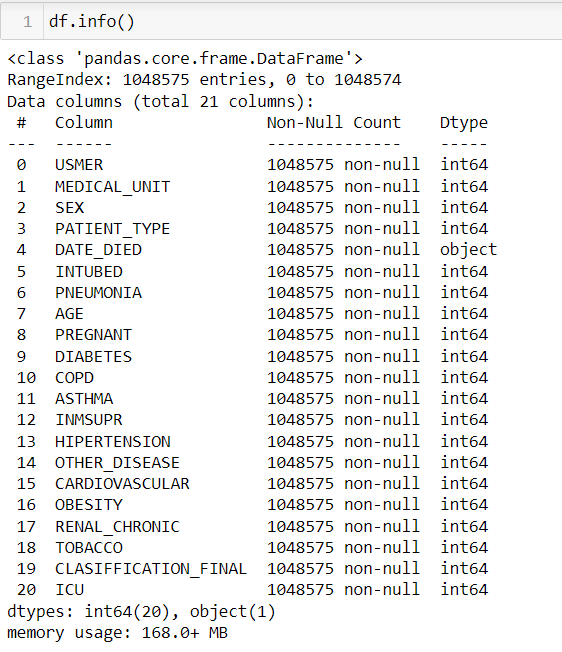
**STEP 1:** Import the needed libraries.



**STEP 2:** Import the COVID-19 dataset as presented below.

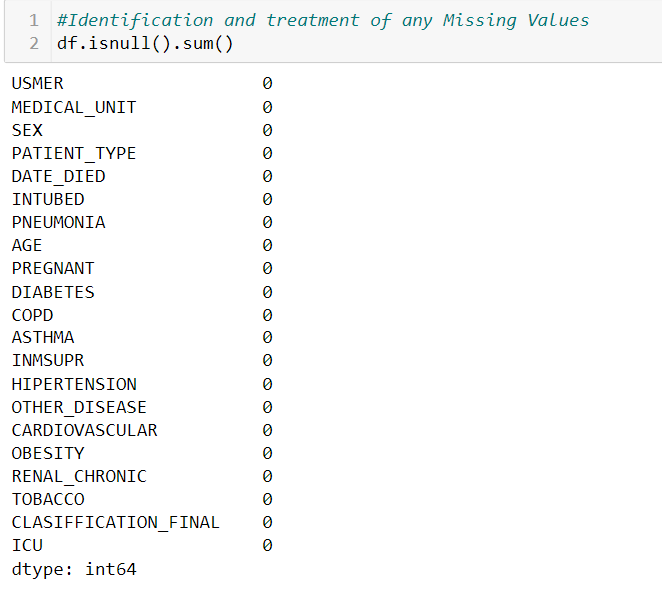


The dataset information is represented as follows:



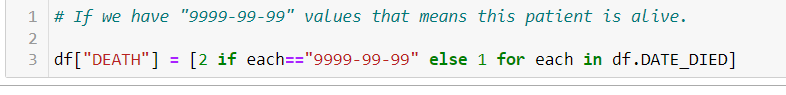
According to the result, the given dataset has 20 variables.

**STEP 3:** Next, check the values that are missing in our dataset, as demonstrated below.

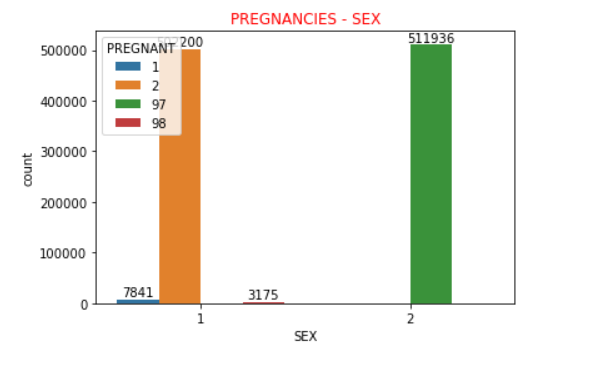


According to the result, the given dataset does not have any null values.

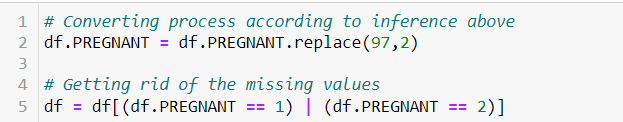
**STEP 4:** Afterwards, move on with the preparation of the Date died Column as presented below.



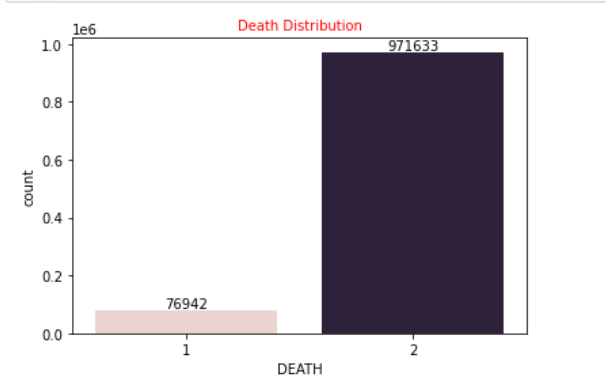
**STEP 5:** Then, it is required to create visualization for Pregnant vs sex with the help of a bar chart as displayed below.

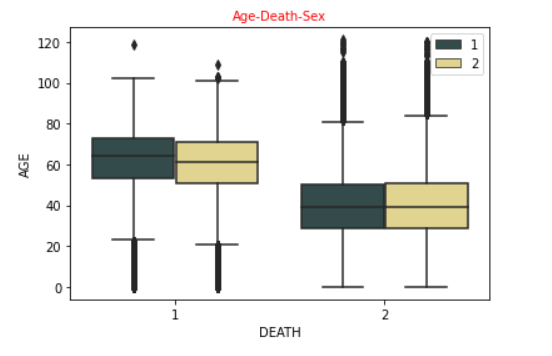


As per the above plot, it is observed that all "97" values are for males, and males cannot be pregnant. Thus, it is necessary to convert from 97 to 2, which is demonstrated below.



**STEP 6:** Creation of the visualization for Death Distribution, Age based Death distribution and Sex based on death distribution, as follows:

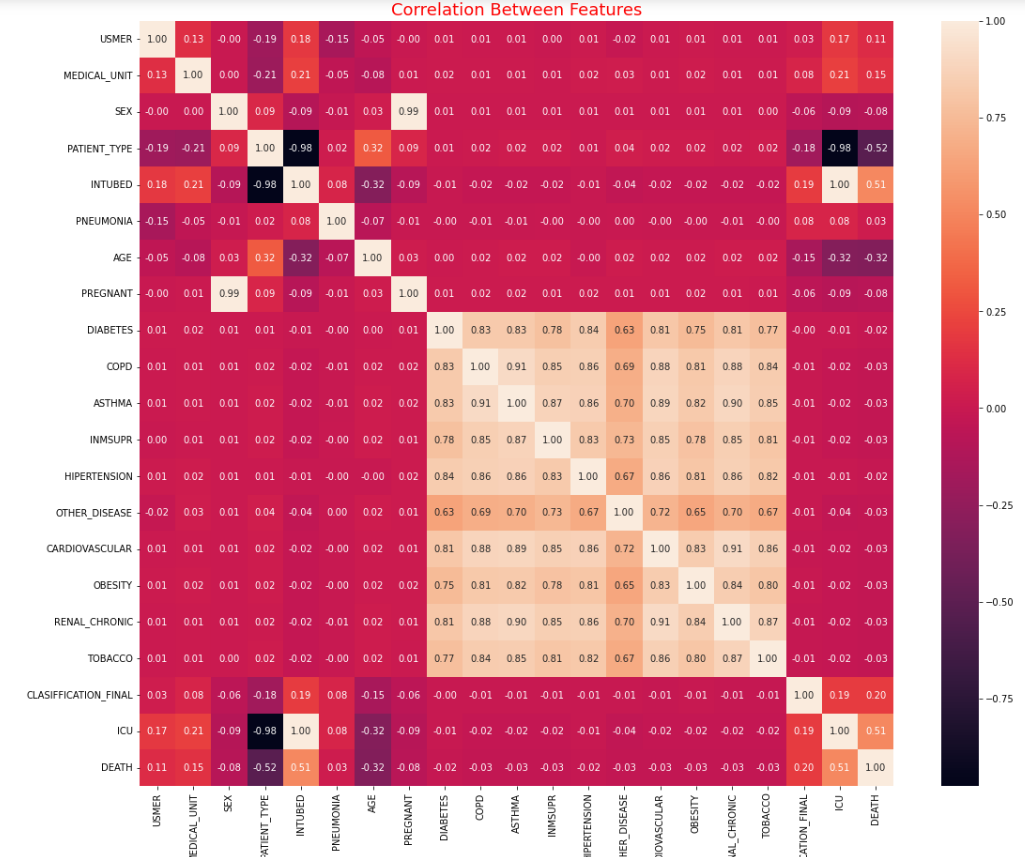




From the above depicted chart, it is evident that the older patients have high chances of dying when compared to the younger patients. On the other hand, no serious difference is found between the female and male based on the average rate of the patients.

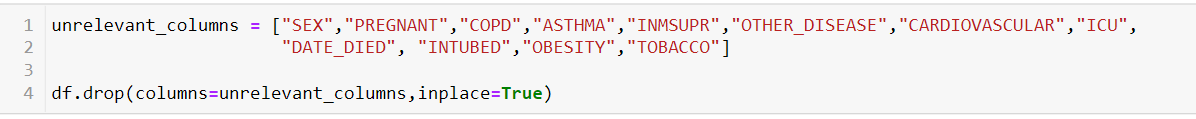


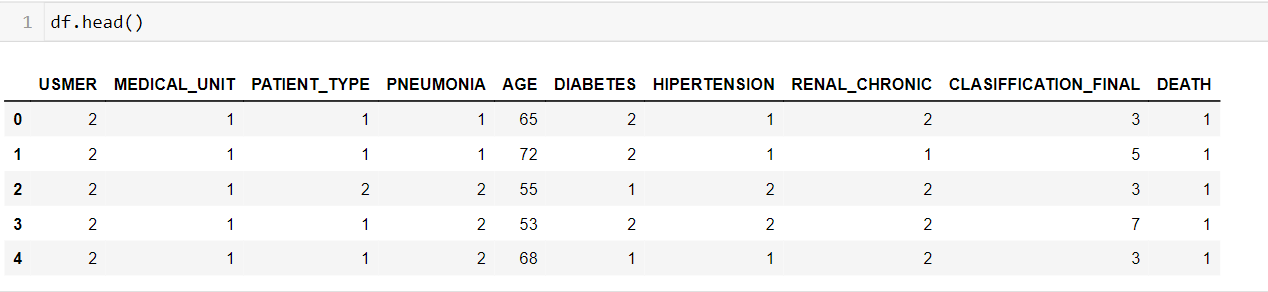
As per the result, in contrast to the female patients, the male patients are more likely to die due to COVID-19.

**STEP 7:** Check the correlation between the features, as represented below:

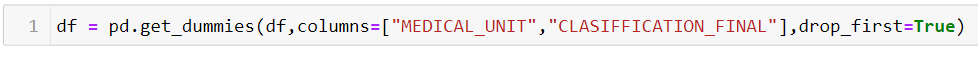
According to the correlation result, we will drop the features that have low correlation with "DEATH" feature as determined below.

**STEP 8:** Now, drop the low correction with the death Feature, as demonstrated below.

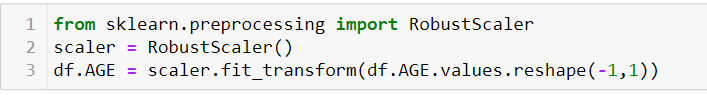




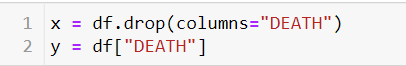
**STEP 9:** Now, ensure preparation of the categorial features for our data analysis.



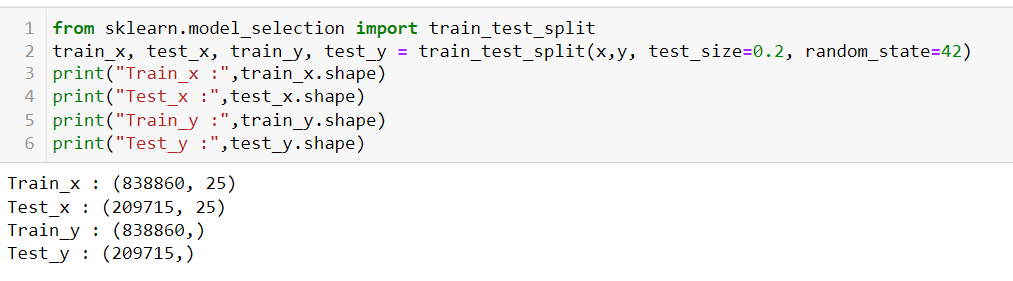
**STEP 10:** Feature Scaling is done, as follows.



**STEP 11:** For our analysis, ensure setting the Target and the response features.



**STEP 12:** Split the data into Train data (80%) and Test data (20%), just as shown below.



## **Data Analysis**

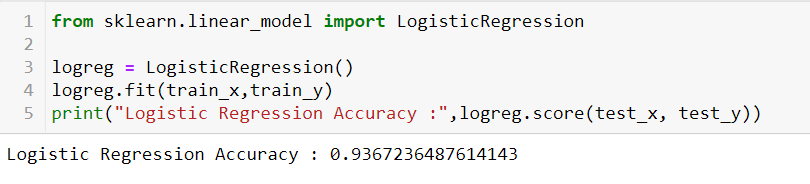
Data analysis can be referred as one of the important parts of this report, which takes the raw data, and then converts it into useful information that benefits the users to make decisions. It is highly essential to collect and analyze the data, as it support to answer the questions of the studied topic (Johnson, 2022).

In this report the purpose of data analysis is to take a look at how the situation is at present and acknowledge the mistakes, followed by planning to have reduced risk in patients due to COVID-19 base on their medical history. And, with respective to this purpose, machine learning models will be created to predict whether the patient is in high risk or not. Here, the Python programming Language is used, and the selected machine learning models are explained and represented in the following sections:

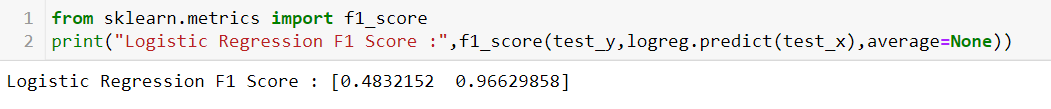
### **Logistic Regression**

The very first machine learning model that is created is the logistic regression model, which is for predicting the COVID-19 deaths based on the patients’ medical history. For this the step and the code is represented as follows (Banoula, 2022) (*Logistic regression for machine learning,* 2021):

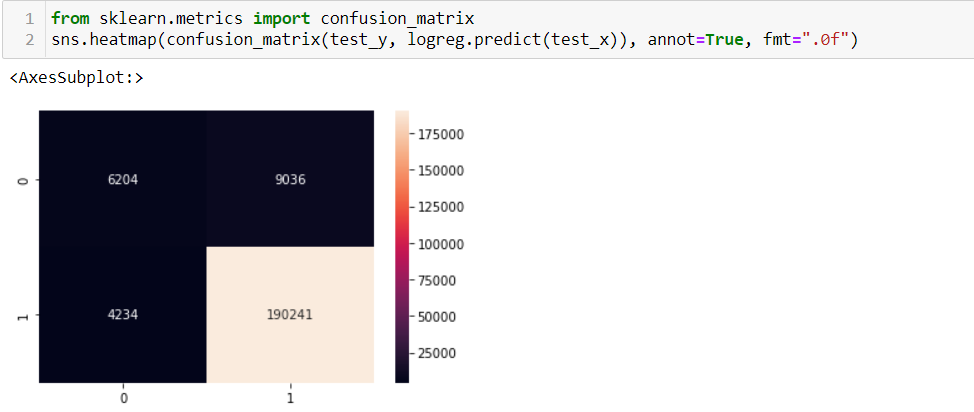
**Model Created**



And also, evaluate the model by determining the f1 score as demonstrated below (*Machine learning - logistic regression,* 2023).

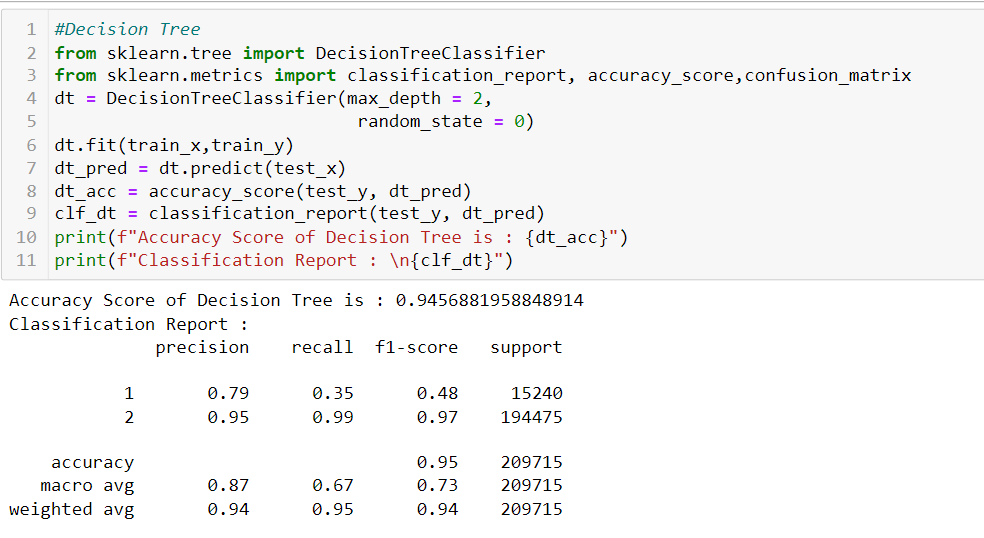


The confusion matrix for the logistic regression is demonstrated below (*What is logistic regression?* 2023).

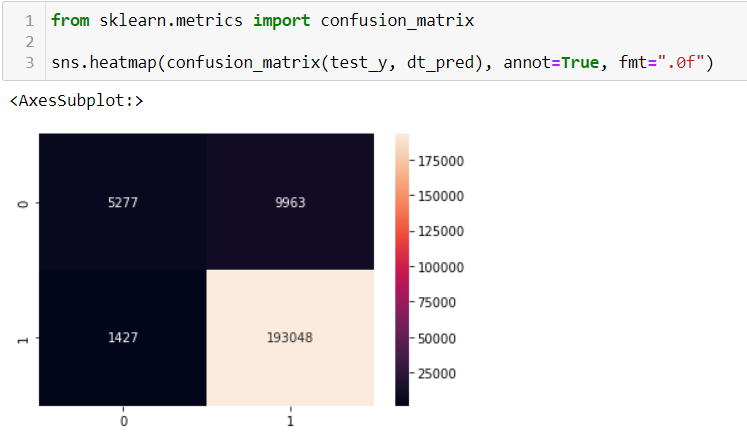


### **Decision Tree**

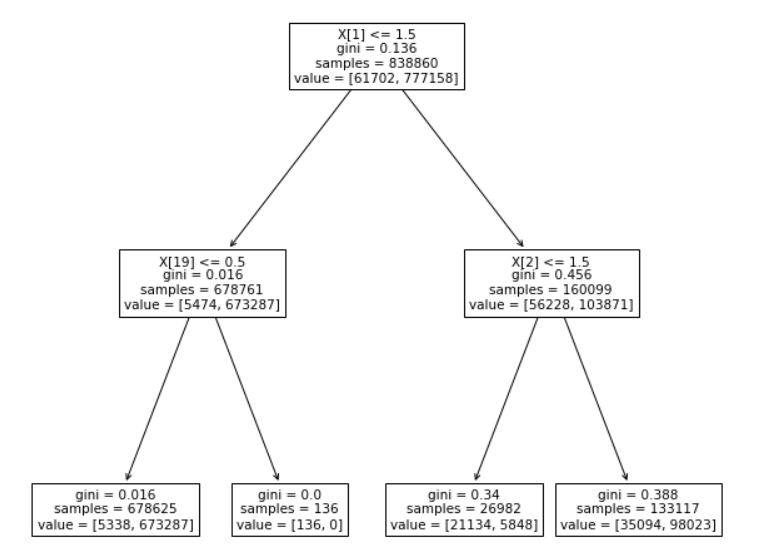
The next machine learning model includes the decision tree, and the respective code is represented as follows (*Decision tree,* 2022) (Yadav, 2019):



The confusion matrix for decision tree model as demonstrated below.

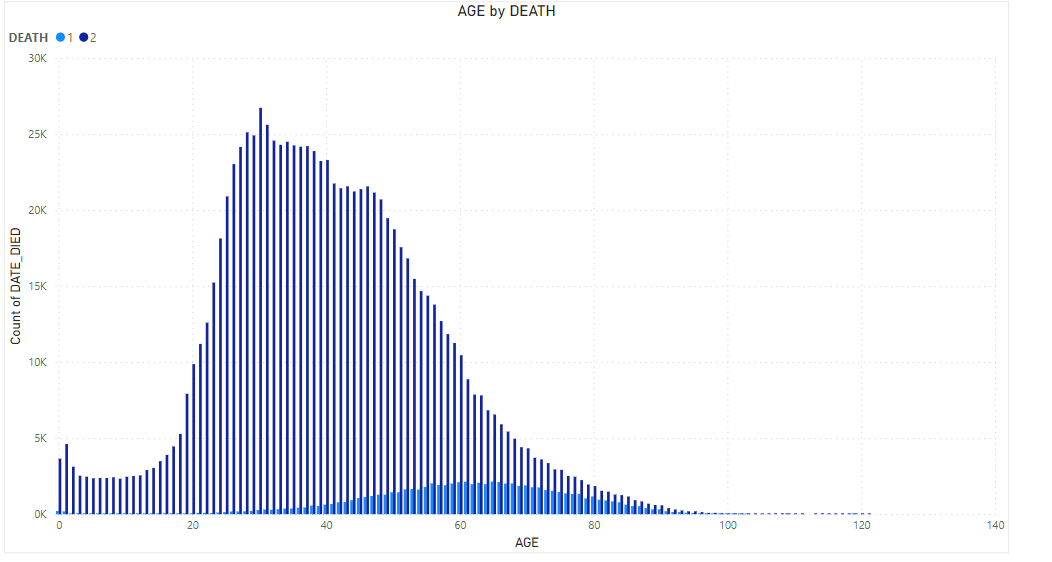


The tree visualization of decision tree model as demonstrated below.

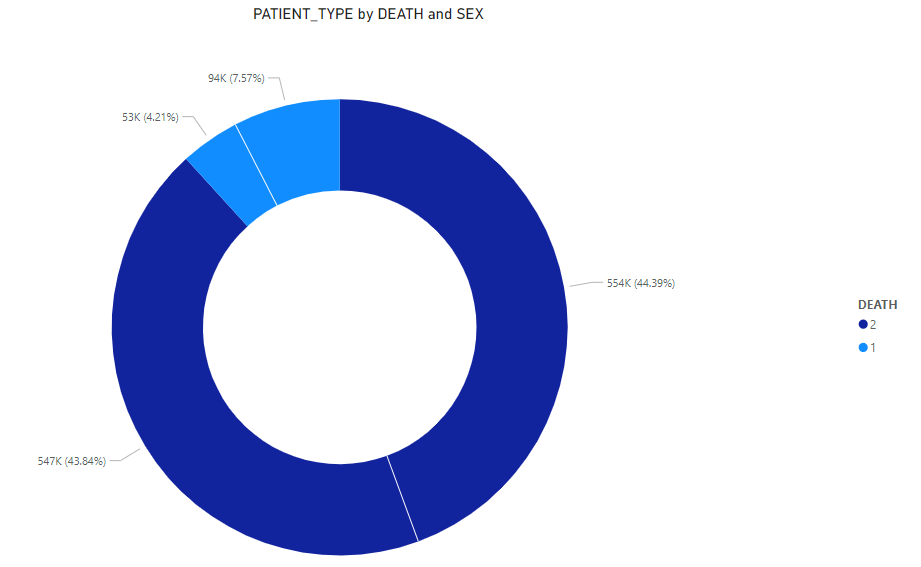


### **Data Visualization on Power BI**

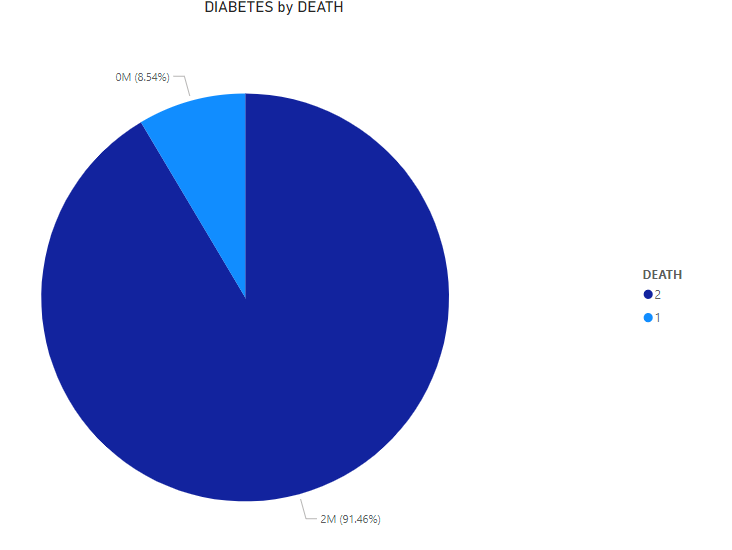
In this section, data visualization will be created with the help of Power BI tool. The below visualization is used to display the Age based on the Death with the help of a bar chart, as demonstrated below (*What is data visualization? Microsoft power bi*, 2023).



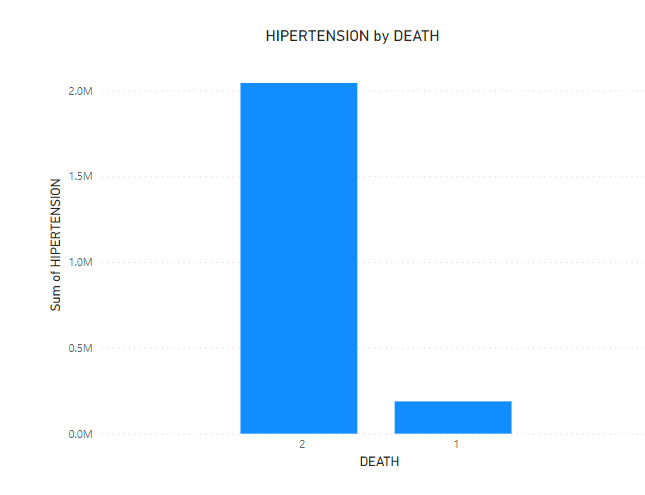
According to the above visualization, age between 30 to 50 has higher death rate compared with below 30 and higher 60 age. The next visualization is used to shows the death based on the gender with the help of a donut chart, as displayed below (Pandey, 2019).



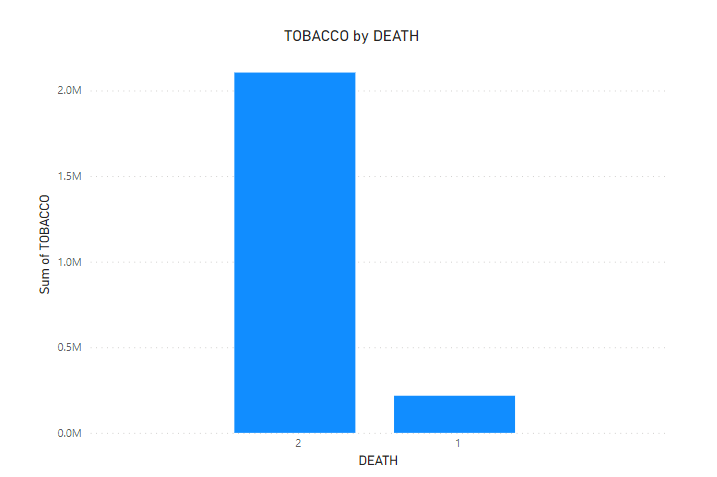
As per the above visualization, the male gender has higher death rate when compared to female. Next, we are going to visualize diabetes patients by death with the help of a pie chart as demonstrated below.



According to the visualization, high diabetes leads to increase the deaths of COVID-19. Next, the hyper tension is visualized based on diabetes with the help of death using a bar chart, as demonstrated below.



As per the result, high hyper tension leads to increase the covid 19 deaths. Afterwards, we are going to visualize the tobacco based on death with the help of bar chart as demonstrated below.



As shown in the above visualization, the patient who had higher tobacco will lead to increase the deaths via COVID-19.

The final dashboard for COVID19 cases as demonstrated below:

Chart, histogram

Description automatically generated

## **Results**

Based on the interpreted result of the created machine learning models, the result of logistic regression states that this model has 93.6% of accuracy to predict the deaths due to COVID-19. For evaluation, the f1 score is 96% and confusion matrix predicted that 10438 patients live on COVID-19, and 199,277 patients are alive. Based on the decision tree, 94.5% of accuracy is observed for predicting the deaths due to COVID-19. For evaluation, the precision is 87%, recall is 67%, f1 score is 73%, and confusion matrix predicted that, 15240 patients live on COVID-19 and 194475 patients are alive.

From the comparison of the two models, the decision tree model is found to have the highest accuracy than the logistic regression model. Therefore, decision tree can be determined as the best model for predicting the deaths via COVID-19. And, it is concluded that, totally 15240 patients live with COVID-19 and 194475 patients survived.

## **Conclusion**

This project has successful determined the best machine learning model to effectively predict if the patients have high risk of COVID-19 or not, and the model is decision tree model. This model is acknowledged for its accuracy and will help the medical authorities to know the patients and provide the required medical resources on time, which will in turn help in decreasing the deaths.

# **Project Management**

## **Project Plan**

The following table exhibits the work plan.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **WBS** | **Tasks** | **Start Date** | **Finish Date** | **Duration** |
| 1 | **Introduction** | **01/13/2023** | **01/16/2023** | 4 days |
| 1.1 | Project Motivations/Problems | **01/14/2023** | **01/14/2023** | 1 day |
| 1.2 | Project Aim | **01/15/2023** | **01/15/2023** | 1 day |
| 1.3 | Project Objectives | **01/16/2023** | **01/16/2023** | 1 day |
| 2 | **Methodology** | **01/17/2023** | **01/24/2023** | 8 days |
| 2.1 | Data Collection | **01/18/2023** | **01/18/2023** | 1 day |
| 2.2 | Data Preparation | **01/19/2023** | **01/19/2023** | 1 day |
| 2.3 | Data Analysis | **01/20/2023** | **01/20/2023** | 1 day |
| 2.3.1 | Logistic Regression | **01/21/2023** | **01/21/2023** | 1 day |
| 2.3.2 | Decision Tree | **01/22/2023** | **01/22/2023** | 1 day |
| 2.4 | Results | **01/23/2023** | **01/23/2023** | 1 day |
| 2.5 | Conclusion | **01/24/2023** | **01/24/2023** | 1 day |
| 3 | **Project Management** | **01/25/2023** | **01/26/2023** | 2 days |
| 3.1 | Project Plan | **01/25/2023** | **01/25/2023** | 1 day |
| 3.2 | Project collaboration | **01/26/2023** | **01/26/2023** | 1 day |
|  |  |  |  |  |

## **Project Collaboration**

As a group, we first met and decided the plan of how to work on this project. Then one of the team members divided the roles based on the agreement from the respective team member. Throughout the process, the Google Meet, where the doubts and the completed work was discussed.

## **Learning Reflection**

This analysis has helped to learn the real time data analysis working. I have learned that the decision tree model can provide highly accurate predictions to support the COVID-19 data, to predict the risks and control the deaths by providing the required medical resource to the needed patients.

No issues were encountered in this analysis; However, it was tricky to correctly choose the variables to get the correct and desired data visualizations.

In future, I would like to work on improved data visualization representation.

# **References**

1. Banoula, M. (2022) *An introduction to logistic regression in python*, *Simplilearn.com*. Simplilearn. Available at: https://www.simplilearn.com/tutorials/machine-learning-tutorial/logistic-regression-in-python#:~:text=algorithm%3A%20logistic%20regression.-,What%20is%20Logistic%20Regression%3F,one%20or%20more%20independent%20variables. (Accessed: January 13, 2023).
2. *Data collection* (2023) *ori.hhs.gov* . Available at: https://ori.hhs.gov/education/products/n\_illinois\_u/datamanagement/dctopic.html (Accessed: January 13, 2023).
3. *Decision tree* (2022) *CORP-MIDS1 (MDS)*. Available at: https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/ (Accessed: January 13, 2023).
4. Johnson, D. (2022) *What is data analysis? research, types & example*, *Guru99*. Available at: https://www.guru99.com/what-is-data-analysis.html (Accessed: January 13, 2023).
5. *Logistic regression for machine learning* (2021) *Capital One*. Available at: https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/ (Accessed: January 13, 2023).
6. *Machine learning - logistic regression* (2023) *Tutorials Point*. Available at: https://www.tutorialspoint.com/machine\_learning\_with\_python/machine\_learning\_with\_python\_classification\_algorithms\_logistic\_regression.htm (Accessed: January 13, 2023).
7. Nizri, M. (2022) *Covid-19 dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/meirnizri/covid19-dataset (Accessed: January 13, 2023).
8. *What is data visualization?: Microsoft power bi* (2023) *Microsoft Power BI*. Microsoft. Available at: https://powerbi.microsoft.com/en-us/data-visualization/ (Accessed: January 13, 2023).
9. *What is logistic regression?* (2023) *IBM*. Available at: https://www.ibm.com/in-en/topics/logistic-regression (Accessed: January 13, 2023).
10. Yadav, P. (2019) *Decision tree in machine learning*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96 (Accessed: January 13, 2023).
11. Pandey, P. (2019) *Data visualization with Power Bi*, *DataCamp*. DataCamp. Available at: https://www.datacamp.com/tutorial/data-visualisation-powerbi (Accessed: January 13, 2023).

# **Appendix**









Graphical user interface, text, application, email

Description automatically generated



