

**PREDICTION OF DIABETIC PATIENTS USING BEST MACHINE MODEL**

BY

**MEGHA KALACHIRA ANKUSH**

**P2696430**

**IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE MASTER OF SCIENCE**

IN

**DATA ANALYTICS**

**AUGUST 2023**

**SUPERVISOR: Mrs AREEJ POLINA**

**Contents**

[List of Figures 4](#_Toc143978686)

[ABSTRACT 6](#_Toc143978687)

[CHAPTER 1 – INTRODUCTION 7](#_Toc143978688)

[1.1. Project Overview 7](#_Toc143978689)

[1.2. Project Background and Motivation 8](#_Toc143978690)

[1.3. Project Aim and Objectives 13](#_Toc143978691)

[1.4. Research questions 13](#_Toc143978692)

[1.5. Scope of the Project 14](#_Toc143978693)

[1.6. Project outline 14](#_Toc143978694)

[CHAPTER 2 – LITERATURE REVIEW 15](#_Toc143978695)

[2.1. Overview 15](#_Toc143978696)

[2.2. Critical Review 20](#_Toc143978697)

[CHAPTER 3 – METHODOLOGY 21](#_Toc143978698)

[3.1 Overview 21](#_Toc143978699)

[3.2. Research Aim 21](#_Toc143978700)

[3.3. Research Framework 21](#_Toc143978701)

[3.4. Data collection methods 22](#_Toc143978702)

[3.5. Data Pre-processing 22](#_Toc143978703)

[3.6. Data analysis 23](#_Toc143978704)

[3.7. Modelling Preparation 23](#_Toc143978705)

[3.8. Modelling 23](#_Toc143978706)

[3.9. Model Evaluation 24](#_Toc143978707)

[CHAPTER 4 – ANALYSIS AND FINDINGS 25](#_Toc143978708)

[4.1. Overview 25](#_Toc143978709)

[4.2. Data Analysis and Findings 25](#_Toc143978710)

[CHAPTER 5 – CONCLUSION AND RECOMMENDATIONS 53](#_Toc143978711)

[5.1. Conclusion 53](#_Toc143978712)

[5.2. Recommendations 54](#_Toc143978713)

[References 55](#_Toc143978714)

# List of Figures

[Figure 1: Global raised fasting blood glucose, adults aged 18+ (Source: WHO) 9](#_Toc143977836)

[Figure 2: Estimates of the global prevalence of diabetes in the 20-79 year group between 2000-2021(Source: IDF) 10](#_Toc143977837)

[Figure 3: Age-wise prevalence of diabetes in high, middle and low income countries (Source: IDF) 10](#_Toc143977838)

[Figure 4: Diabetes related health expenditure in billion USD (Source: IDF) 10](#_Toc143977839)

[Figure 5: Region wise diabetes related health expenditure a) Total and b)per person (Source: IDF) 11](#_Toc143977840)

[Figure 6: Top 10 countries with diabetes and health expenditure (Source: IDF) 11](#_Toc143977841)

[Figure 7: Global percentage of diabetes deaths occurring under 70 years (Source: WHO) 12](#_Toc143977842)

[Figure 8: Research Framework 22](#_Toc143977843)

[Figure 9: Importing essential libraries and models 25](#_Toc143977844)

[Figure 10: Loading and cleaning the data sets 27](#_Toc143977845)

[Figure 11: Checking for missing values in the train DataFrame 29](#_Toc143977846)

[Figure 12:Checking for missing values in the test DataFrame 30](#_Toc143977847)

[Figure 13:Checking the data types in the train DataFrame 30](#_Toc143977848)

[Figure 14: Checking the data types in the test DataFrame 31](#_Toc143977849)

[Figure 15:Descriptive statistics of the features within the train DataFrame 31](#_Toc143977850)

[Figure 16: The features are classified into various groups based on the nature 32](#_Toc143977851)

[Figure 17: Plotting the social and economic features 32](#_Toc143977852)

[Figure 18: Social and Economic Features distribution – Subject Profile 33](#_Toc143977853)

[Figure 19: Fourteen Level age category 33](#_Toc143977854)

[Figure 20: Social and Economic Features distribution – Subject Income and Health Insurance Status 34](#_Toc143977855)

[Figure 21: Plotting Health Category features distribution 35](#_Toc143977856)

[Figure 22: Health Category features distribution 36](#_Toc143977857)

[Figure 23:Plotting Disease Category features distribution 36](#_Toc143977858)

[Figure 24:Disease Category features distribution 37](#_Toc143977859)

[Figure 25: Plotting Habits Category features distribution 38](#_Toc143977860)

[Figure 26: Habits Category features distribution 38](#_Toc143977861)

[Figure 27: Plotting histogram for BMI 39](#_Toc143977862)

[Figure 28: Plotting the relationship between features using Heatmap 40](#_Toc143977863)

[Figure 29: Heatmap 40](#_Toc143977864)

[Figure 30: Correlation of data binary to features in descending order for train DataFrame 41](#_Toc143977865)

[Figure 31: Data Preprocessing 42](#_Toc143977866)

[Figure 32: AdaBoost Classifier - Modelling and evaluation 44](#_Toc143977867)

[Figure 33:Installation and modelling of CatBoost Classifier 45](#_Toc143977868)

[Figure 34: CatBoost performance evaluation 46](#_Toc143977869)

[Figure 35: Decision Tree - Modelling and evaluation 47](#_Toc143977870)

[Figure 36:SVM – Modelling and evaluation 49](#_Toc143977871)

Figure 37:Logistic Regression-Modelling and evaluation……………………………………51

# ABSTRACT

Diabetes is one of the most common non communicable diseases in the world and is affecting population across the globe and has been increasing rapidly over the past few years. This disease can be fatal and has various complications. People with diabetes have to live with caution and be mindful at all times and based on the type of the diabetes, some have to depend on insulin injections for a lifetime. Early diabetes diagnosis allows for the implementation of preventative measures, lifestyle changes, and effective treatment plans, improving the condition's management and preventing complications. The aim of this study is to analyse and forecast diabetic patient diagnoses using the best machine modelling approaches. The goal of the project is to develop and analyse machine learning models which can precisely predict people with or without the risk of diabetes based on their habitual, clinical and demographic features to increase the accuracy, efficiency, and timeliness of diabetes diagnosis. Classic and ensemble machine learning models such as AdaBoost classifier, Logistic Regression, Decision Tree, Support Vector Machine and CatBoost Classifier were used in the prediction of diabetes among people and the Decision Tree model outperformed all the other models when the performance evaluation was conducted using different metrics such as accuracy, precision, recall and F1 score. The study also identified that various features such as general health of the subject, high blood pressure, high cholesterol, BMI, age, difficulty in walking or climbing stairs, income, physical health status, heart disorder or heart attack status, education, physical activity, stroke, cholesterol check status and heavy alcohol consumption habit showed higher correlation to diabetes and might be useful in classifying the risk of diabetes.

# CHAPTER 1 – INTRODUCTION

# 1.1. Project Overview

Diabetes is a long-term (or "chronic") and serious disorder characterized by elevated blood glucose levels caused by the body's inability to produce enough or any of the insulin hormone or to effectively utilize the insulin it does produce. Insulin is a type of hormone that controls blood glucose levels and produced by the beta cells that are found in islets of Langerhans within the Pancreas. To understand diabetes, the researcher must first study how the body functions without diabetes. Food has a variety of components such as carbohydrates, vitamins, protein, fat, and much more. Carbohydrates is mostly obtained from foods that contain carbs, which offer energy to our bodies. Carbohydrates can be found in bread, cereal, pasta, rice, fruit, dairy products, and vegetables. When a normal human being consumes such foods, their body convert this food into glucose, which is then distributed throughout their body via their bloodstream. Glucose primarily gets to the brain since it is essential for the body's thinking and functionality. The remaining glucose is delivered to the other parts of the body, including cells which is used for other metabolic activities and the liver. Insulin is a critical component that is essential for the human body to function properly. It is a hormone generated by pancreatic beta cells. It allows glucose absorption from the bloodstream to the cells in the body to be used for metabolic activities. Because the pancreas produces insulin, it requires a sufficient supply of glucose. If the pancreas does not create enough insulin, glucose accumulates, and diabetes develops in an individual.

There are different types of diabetes found in people, the most common ones being type 1, type 2, and gestational diabetes, and less common types being monogenic diabetes and cystic fibrosis-related diabetes. Type 1 diabetes, which is also known as juvenile diabetes occurs as the result of an autoimmune process in one’s body, where the body’s immune system attacks the beta cells of the pancreas that produces insulin leading to the production of little or no glucose. Even though the causes of this autoimmune reaction are not identified fully, but is believed to be caused due to genetic susceptibility or other environmental triggers. Type 1 diabetes can develop at any age and is commonly found in children and young adults and cannot be prevented. The type 1 diabetes patients will require daily insulin injections to regulate their blood glucose levels within the normal range, which otherwise might lead to various complications such as ketoacidosis, nerve damage, issues with the eyes, increased risk of skin infection, issues with the kidneys, cardiovascular disease, foot problems including numbness, high blood pressure, and stroke. According to WHO, there were 9 million people living with type 1 diabetes, globally in 2017.

Type 2 diabetes, which is the most common type diabetes accounting to more than 90 percent of global diabetes, the body fails to use the insulin produced properly in absorption of glucose from bloodstream to cells, which will use this glucose as energy for metabolic activities. This might be caused due to the condition called insulin resistance, wherein the body’s cells might have the inability to respond to the insulin produced. The result of type 2 diabetes is hyperglycaemia, also known as elevated blood glucose levels or raised blood sugar, which might eventually lead to the failure of pancreatic beta cells to produce sufficient insulin to keep up with the elevated glucose levels. This type of diabetes is preventable and the patients can live healthy life by managing their lifestyle as the major risk factors that cause type 2 diabetes include overweight and obesity, lack of exercise, age, ethnicity, and genetics.

Gestational diabetes is developed among women during their pregnancy, and it is a hyperglycaemia state where the blood glucose values rise above the normal levels, but below that of the diabetes diagnostic level. The major risks with gestational diabetes are complications during pregnancy and delivery and increased risk of type 2 diabetes both in child and mother in future.

Here, the study uses a dataset on patient demographics, medical and health information and habits, to investigate the effectiveness of the employment of machine learning algorithms in diabetes diagnosis. Diabetes binary information is present in the data set meaning subject with both diabetes positive and negative, enabling use of supervised learning techniques for prediction. Several machine learning techniques were developed and assessed, including logistic regression, decision trees, CatBoost, support vector machines, and AdaBoost. The evaluation the models' prediction efficiency, was done using the train dataset, which was used for training and validation and test dataset, was used for evaluation, subsequently and then using of performance evaluation metrics such as accuracy, precision, recall, and F1 score.

# 1.2. Project Background and Motivation

According to a WHO (World Health Organization) study from 2016, there are 422 million adults with diabetes, with an estimated 2 million deaths (WHO, 2023).

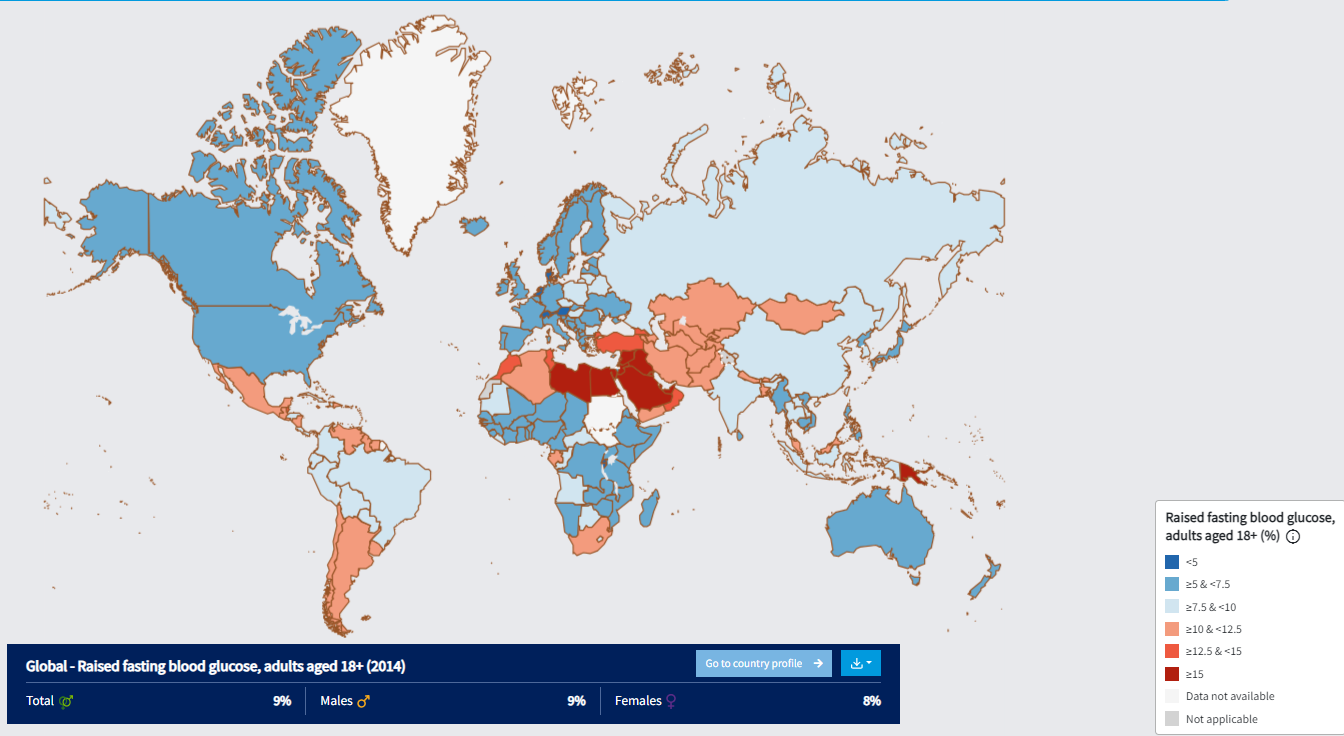


Figure 1: Global raised fasting blood glucose, adults aged 18+ (Source: WHO)

Diabetes affected 9% of persons aged 18 and above in 2014, including about 9 percent of the global male population and 8 percent of the global female population. The estimated number of people with diagnosed and undiagnosed diabetes living in the United States alone is approximately 34.2 million with a large proportion among them being aged above 18+ (Centers for Disease Control and Prevention, 2020).

Chart

Description automatically generated

Figure 2: Estimates of the global prevalence of diabetes in the 20-79 year group between 2000-2021(Source: IDF)

The studies conducted by the International Diabetes Federation shows that the number of people with diabetes in the age group of 20-79 has been increasing in an alarming rate over the past decade from 151 million in 2000 to 537 million in 2021 (International Diabetes Federation, 2021).

Chart, line chart

Description automatically generated

Figure 3: Age-wise prevalence of diabetes in high, middle and low income countries (Source: IDF)

The age wise distribution of diabetes within the high, middle and low income countries shows that the number of people with diabetes has increased with age, but is affecting the people in all age ranges. Considering the various symptoms and complications of diabetes, this will lead to increased health risks and mortality among these people regardless of age. But there is a huge potential of prevention and control of diabetes considering the percentage of the population affected based on the age groups. The financial burden the disease might bring on the working class in the given age groups is high.

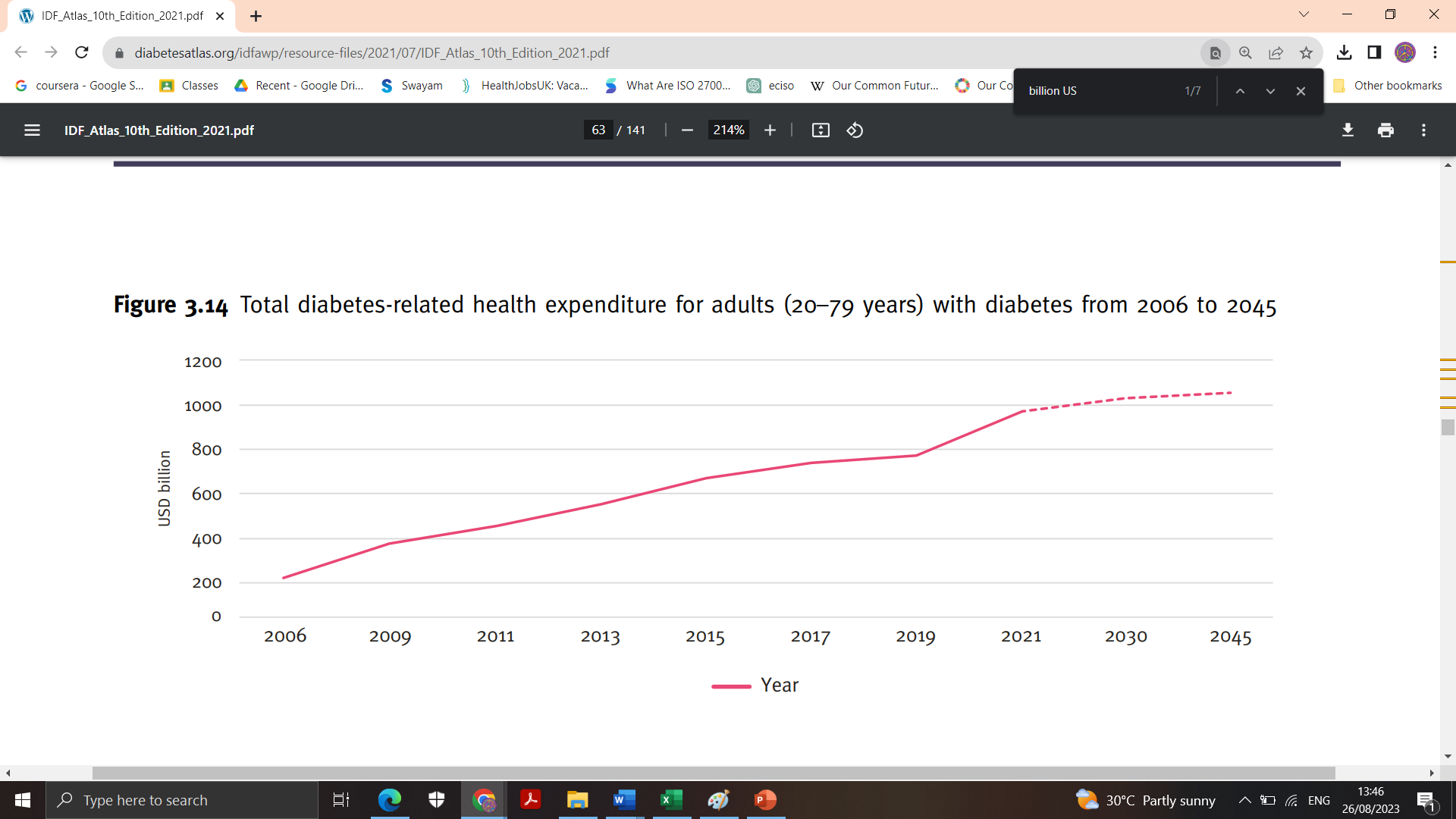


Figure 4: Diabetes related health expenditure in billion USD (Source: IDF)

The figure shows the diabetes related health expenditure in billion USD for adults in the age group of 20–79 years over a period of 2006 to 2021 and the forecasted values from after 2021 till 2045, if the trend continues. Considering the costs being approximately about 1000 billion USD in 2021, it is not wrong to observe the intensity of financial burden on the people as well as the countries are very high, affecting their financial stability and economy.

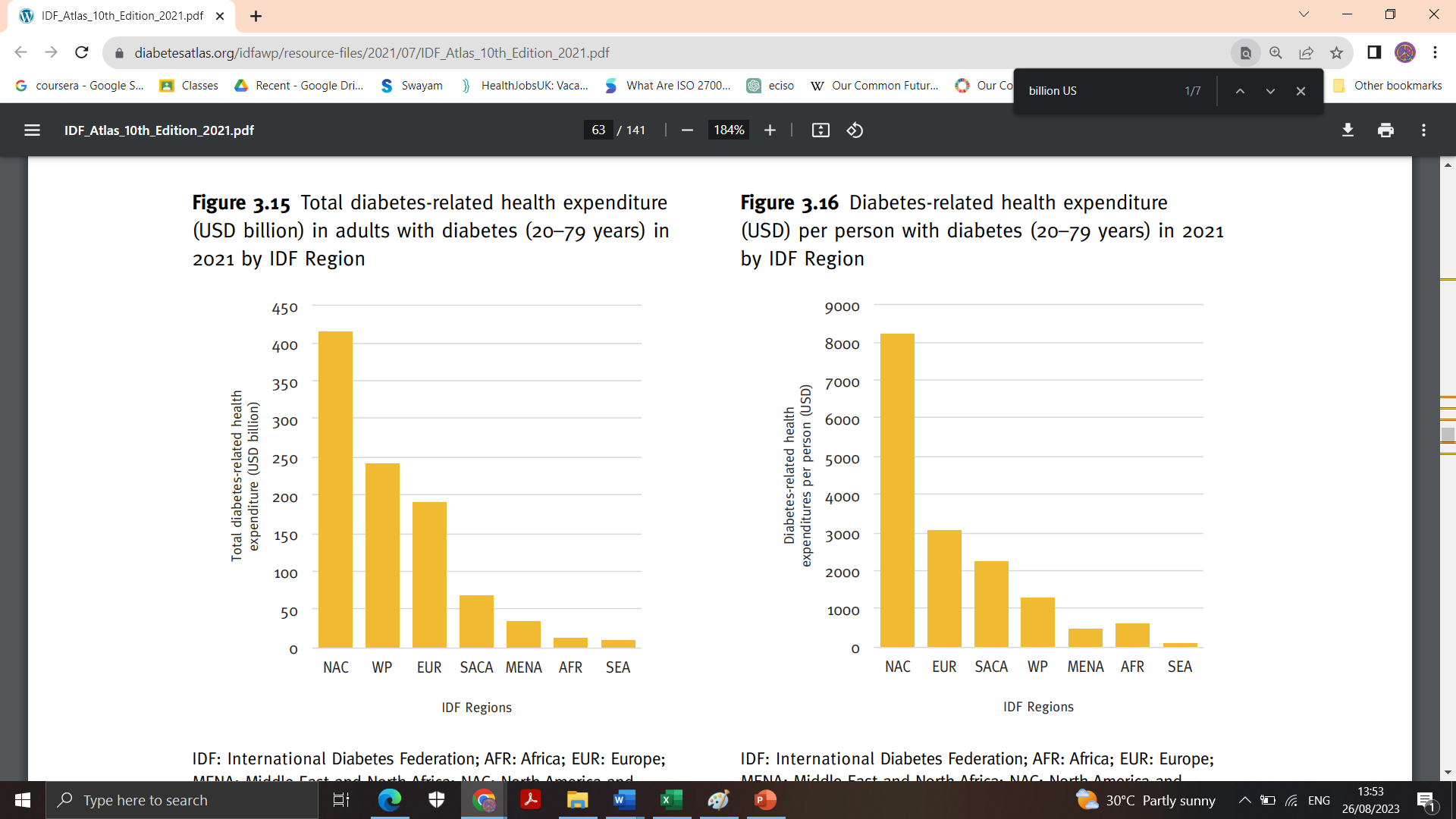


Figure 5: Region wise diabetes related health expenditure a) Total and b)per person (Source: IDF)

The report by IDF have shown that the total diabetes related health expenditure in the North America and Caribbean is the highest, followed by Western Pacific, Europe, South and Central America, Middle East and North Africa, Africa and South-East Asia, but the per person expenditure in Western Pacific and Middle East and North Africa and SEA is lesser, which might be due to the increased number of diabetes patients, while its higher in other regions, which might be due to the higher cost of healthcare and lower number of diabetes patients.

Chart, bar chart

Description automatically generated

Figure 6: Top 10 countries with diabetes and health expenditure (Source: IDF)

China being the second highest populated country in 2021, behind India, has the highest number of diabetes patients, whilst India is in second position in diabetes rankings. The cost of diabetes related healthcare though is a different story and the USA is spending approximately 5 times the number of diabetes patients as health expenditure, which shows the cost of health care in the USA. Similar trends can be observed in Brazil and Japan, where the expenditure is higher, which might be due to the increase in healthcare costs. The health care expenditure in countries like India and Pakistan are very less compared to the diabetes patients, which might be because of the cheaper healthcare or patients’ inability to use the healthcare systems due to poverty.



Figure 7: Global percentage of diabetes deaths occurring under 70 years (Source: WHO)

Diabetes was the primary cause of approximately 1.5 million fatalities in 2019, with 48% of all diabetes-related deaths occurring below the age of 70. Diabetes was responsible for an additional 460 000 kidney disease fatalities, and elevated blood glucose is responsible for about one fifth of cardiovascular mortality. Diabetes caused a three percent rise in age-standardized death rates between 2000 and 2019. Diabetes-related mortality increased by 13% in lower-middle-income nations.

Considering all these financial expenditure and burden, the complications patients have to live with and the death rate, it is necessary to prevent the increasing number of diabetes patients globally and hence this study is necessary to predict and thereby prevent diabetes in people.

# 1.3. Project Aim and Objectives

The aim of this study is to analyse and forecast diabetic patient diagnoses using the best machine modelling approaches. The goal of the project is to develop and analyse machine learning models which can precisely predict people with or without the risk of diabetes based on their habitual, clinical and demographic features to increase the accuracy, efficiency, and timeliness of diabetes diagnosis.

The objectives of this research are:

* To create prediction models that can reliably categorise patients into diabetic and non-diabetic groups or successfully identify persons at risk of diabetes by using big and diverse datasets, cutting-edge machine learning algorithms, and suitable feature selection techniques.
* Use data analysis and feature engineering techniques to gain insight on the various features in the dataset collected, to reach meaningful conclusions.
* Develop and assess different machine learning algorithms, such as Logistic Regression, Decision Trees, AdaBoost, SVM and CatBoost Classifier, for the prediction of Diabetes.
* Compare the performance of these ML models using performance evaluation metrics such as accuracy, precision, F1 score and recall.
* To give medical practitioners trustworthy, data-driven tools to aid in early diagnosis and intervention.

# 1.4. Research questions

* Can machine learning models predict the progression of diabetes or the risk of developing complications in diabetic patients?
* How do different pre-processing techniques and feature selection methods impact the performance of machine learning models in diagnosing diabetes?
* Can machine learning models integrate multiple data sources (clinical records, genetic data, lifestyle factors, etc.) to improve the accuracy of diabetes diagnosis?
* How do machine learning models handle missing or incomplete data in the context of diabetes diagnosis?
* What are the potential challenges and limitations of implementing machine learning models for diabetes diagnosis in real-world clinical settings?

# 1.5. Scope of the Project

The research in the area of diabetes prediction is very crucial considering the complications that comes with them and the financial burden that put on the overall global economy. Being a preventable disease, type 2 diabetes can be treated and controlled, it is one of the most common diabetes accountings to over 90 percent of the global diabetes patients. Considering this, there is a major scope in understanding the efficiency of using ML models in prediction and considering the number of undiagnosed diabetes, which can be majorly due to the costs for medical diagnosis, it is necessary to develop a model that will help in diabetes prediction. This can help in avoiding needless examinations, hospital stays, and treatments for conditions that are misdiagnosed or go undetected, and accurate and prompt diagnosis can lower healthcare expenditures globally. With identification of various features or attributes that are critical, it will be easy to identify at risk people and diagnosis enabling implementation of preventative measures, lifestyle changes, and effective treatment plans, improving the condition's management and preventing complications.

# 1.6. Project outline

This project has 5 chapters:

* Introduction – Contains project overview, background and motivation, aim and objectives, research questions, and project scope.
* Literature Review – Summarises existing research in the area of study and research gap.
* Methodology – Summarises the overall methodology of the research including methods, and techniques used in data collection, preprocessing, and analysis as well as models and evaluation.
* Analysis and Findings – Summarises the process and findings of data analysis and modelling.
* Conclusion and Recommendations – Summarises the results and insights gained and overall conclusion of the research by answering research questions. Also, share any recommendations for future research and implementation.

# CHAPTER 2 – LITERATURE REVIEW

# 2.1. Overview

Diabetes being one of the word’s common lifestyle diseases with a large number of death due to symptoms, or other complications, is one of the scariest. Diabetes can be either hereditary, or can be caused due to food and lifestyle habits and is affecting about 500 million people, from various age groups worldwide. Even though diabetes cannot be cured, there are treatments, but it is always better to prevent this disease by early detection and prediction.

In research paper Aguilera–Venegas et al, a type 2 diabetes mellitus prediction model using machine learning models was proposed, with an aim of predicting type 2 diabetes at most seven and half years prior to the potential disease (Aguilera-Venegas, et al., 2023). The data set was collected from the nation-wide cohort di@bet.es study in Spain and the research used four machine learning classifier models – Decision Tree, Random Forest, K- Nearest Neighbors and Neural Networks. The framework uses hyperparameters to tune the classifier models and improve their performance and the final results showed that the Random Forest model outperformed all other models with an accuracy of 92 percent. The study being nationwide, is expected to represent multiple ethnic groups, gender and age demographics.

Khaleel & Al-Bakry, proposed a machine learning framework to predict diabetes among individuals, with precision using various machine learning (ML) algorithms such as Logistic Regression algorithm, Naïve Bayes classifier, and K-nearest Neighbors classifier (Khaleel & Al-Bakry, 2023). The precision is measured by various evaluation methods such as precision, recall, and F1-measure. This research uses data set procured from PIMA Indian Diabetes Database which contains data on adult women in the United States and is used to predict diabetics risk based on various diagnostics data in the data set. The results of the study shows that the Logistic Regression classifier based framework gives maximum accuracy of about 94% compared to other models and hence can be used for future predictions.

The research paper Olisah et al., proposed a framework using machine learning models to predict and diagnose diabetes, for which they used two data sets from PIMA Indians Diabetes database and Laboratory of Medical City Hospital (LMCH) diabetes database (Olisah, et al., 2022). In addition to the machine learning models, the feature selection and missing value imputation methods were used to boost the overall performance of the classification models, to get high accuracy in prediction and diagnosis of diabetes. In this framework, Spearman correlation is adopted for feature selection and polynomial regression is adopted for missing value imputation, intending to boost the performance, followed by the use of the supervised machine learning models such as Random Forest, Support Vector Machine, and Twice-Growth Deep Neural Network (2GDNN) for classification. The results showed that when combined with feature selection and missing value imputation methods, the 2GDNN model outperformed other classifier models. The major limitations of the study are that the PIMA Indians data set contained only data on adult women from a county in US, whereas the LMCH data set contained adult Iraqi population and hence doesn’t represent the overall demographics of the diabetes patients globally.

The research Ahmed et al, proposed a model for diabetes prediction by employing a fused machine learning model (Ahmed, et al., 2022). Artificial Neural Network (ANN) models and Support Vector Machine (SVM) were used for the analysis of the dataset collected from UCI Machine Learning Repository which contains 520 data instances with 17 features. The framework has two layers, first with training data consisting of 70 percent of the dataset and second layer with 30 percent of the data set as test data. Output of this model is further used as input for the Fuzzy model, which will check if the diabetes diagnosis is positive or negative. The cloud storage system is present to store the fused models and is used for predicting the diabetic status using real time medical record of the people. This proposed Fused Model for Diabetes Prediction showed an accuracy of 94.87%, outperforming all other models.

The research paper, Lama et al, proposed a machine learning method to predict type 2 diabetes or prediabetes risk among people with abnormal glucose regulation (Lama, et al., 2021). The data used was Stockholm Diabetes Preventive Program longitudinal research data from 1992-2017 with data of more than 8000 people with normal glucose tolerance or prediabetes. The data represented only adults over the age of 34 in 1992 and doesn’t address the demographics of diabetes patients below the age 34. The model used Random Forest classifier along with SHAP Tree Explainer for prediction interpretation, XGBoost Classifier, LightGBM classifier, CatBoost classifier and Linear Regression. The study identified various important features that influences type 2 diabetes such as heredity or family history of diabetes, age, waist-hip ratio, body mass index, increased systolic blood pressure and diastolic blood pressure, lack of physical activity, gender (male gender are more at risk to diabetes), etc. The results do not share more on the performance of the machine learning models, but is used to generate individual comprehensive risk profile based on the features that influence the risk of diabetes and suggest an individualised diabetes prevention health care plans.

The research paper Hassan et al, proposed machine learning based diabetes prediction framework (Hassan, et al., 2021). The authors collected data from Khulna Diabetes Center in Bangladesh, consisting of 289 data instances and 13 features. The framework uses machine learning models like Logistic Regression, XGboost Classifier and Random Forest classifier and the results shows that Random Forest outperforms all other models with an accuracy of 86.36%. The major limitation of the research is that the data is collected only represents people from single district in Bangladesh and hence doesn’t represent the global diabetes patient demographics. Posonia et al, proposed machine learning model to predict gestational diabetes among women (Posonia, et al., 2020). The study conducted on data set with 768 data instance and 8 features and used various machine learning models and results shows that Decision Tree J48 classifier calculation gave more precise results using minimum processing time.

The research paper Hasan et al, proposed a robust diabetes prediction framework, which has multiple steps such as rejection of outliers, missing value filling, standardisation of the data, feature selection, K-fold cross validation, machine learning models for classification, and finally Multilayer Perceptron (MLP) (Hasan, et al., 2020). The classifier models used are K- Nearest Neighbors, Decision Trees, AdaBoost, Naive Bayes, Random Forest, and XGBoost, and also weighted ensemble models of the different models, by using the Area under the ROC curve of the ML model to calculate the weights. The performance of the models is then maximised using hyperparameter tuning. The data set is the PIMA Indian Diabetes Dataset, which only contains data about women from a county in the United States only. The results of the research shows that the proposed ensembling classifier model outperforms all other models and gives best results in predicting diabetes.

The research paper Tigga and Garg, proposed machine learning models to predict Type 2 diabetes with high accuracy and precision (Tigga & Garg, 2020). The primary data set collected by the researchers were compared with PIMA Indian Diabetes Dataset to have a better comparison of the models and used Machine learning models such as Logistic Regression model, Support Vector Machine, K- Nearest Neighbors, Naïve Bayes classifier, Decision Tree, Random Forest. The primary data set collected contains data of male and female patients aged above 18, and 18 attributes related to family background, health, and lifestyle of the respondents. The results showed that the Random Forest model performed with best accuracy when tested for the primary data set collected by the researchers and PIMA Indian Diabetes Dataset.

The research Malik et al, a machine learning framework is developed to predict diabetes in women, as the dataset used is an exclusive data set containing data on women procured from Frankfurt hospital in Germany (Malik, et al., 2020). In this research ten different machine learning models such as Naive Bayes, Bayes Net, Decision Tree, Random Forest, AdaBoost, Bagging, K-Nearest Neighbors, Support Vector Machine, Logistic Regression, and Multi-Layer Perceptron were used for prediction framework. The results of the accuracy evaluation showed that the frameworks with machine learning models K-Nearest Neighbors, Random Forest, and Decision Tree outperformed the other models in all the metrics of evaluation and hence the researchers claimed that they will be able to predict the results with high accuracy, helping the doctors to diagnose with ease.

Naz & Ahuja, proposed deep learning and machine learning based models to predict diabetes risk using data procured from PIMA Indians Diabetes Database (Naz & Ahuja, 2020). The model employs functional classifiers such as Artificial Neural Network (ANN), Naive Bayes (NB), Decision Tree (DT) and Deep Learning (DL) for building a novel prognostic tool for diabetes prediction which will help the health care workers as well. The results showed that Deep Learning model made prediction with maximum accuracy of 98.07%.

The research paper Warke et al, describes the need for early detection since the disease itself leads to various serious health conditions (Warke, et al., 2019). The research developed machine learning based prediction engine to analyse a diabetes dataset collected and employed various ML classifiers such as Logistic Regression, Naïve Bayes, Support Vector Machine, and K-Nearest Neighbours for prediction of diabetes. They also proposed a web application that allow users to use this prediction engine for easy prediction of diabetes. The prediction model with Naïve Bayes outperformed other models with maximum accuracy, but was computationally expensive and a static engine for disease prediction with further scope of a web application was developed. The major drawback of this study was that the data set used only had records of female patients and 9 attributes to study.

The research Sisodia & Sisodia, proposed a machine learning design model to prognosticate diabetes risk in patients with maximum precision (Sisodia & Sisodia, 2018). The research used three machine learning algorithms such as Decision Tree, Support Vector Machines and Naive Bayes are used in this experiment to detect diabetes by analysing the data collected from PIMA Indians Diabetes Database from UCI machine learning repository, which again is a collection of data about adult women only and not the best representation of world diabetes patient demographics. When measured for performance using precision, accuracy, F1 score, and recall, the Naïve Bayes outperformed all the other models and hence is the best model to get precise prediction of the diabetes risk.

The research Vijayan & Anjali, proposed a multilayer machine learning classifier model where AdaBoost algorithm is used with a base classifier for classifying the data, to predict diabetes mellitus (Vijayan & Anjali, 2016). As base classifiers, Decision Stump classifier, Support Vector Machine, Naive Bayes and Decision Tree and are tested for their accuracy. The results showed that when Decision Stump was used as base classifier, the AdaBoost Algorithm shown the best accuracy of 80.72% compared to when used with other models.

The research Perveen et al, proposed machine learning ensemble models with AdaBoost classifier and Bagging to predict diabetes mellitus among people with J48 Decision Tree used as a base learner (Perveena, et al., 2016 ). The J48 data mining technique is used for more accurate classification. The data set is collected from Canadian Primary Care Sentinel Surveillance network, which consists of data about three ordinal adults groups and henceforth does not represent different age demographics like adolescents and children. The results showed that AdaBoost ensemble model outperformed all the other models.

The research Kandhasamy & Balamurali, studied different machine learning based models such as J48 Decision Tree, K- Nearest Neighbors, Random Forest, and Support Vector Machines, to predict diabetes mellitus and the performance of these models were compared using data mining techniques (Kandhasamy & Balamurali, 2015 ). The data was collected from UCI machine learning data repository and had 20 attributes with various time stamp based data, like the eating habits, and other lifestyle activities. The study results showed that J48 Decision Tree outperformed other models with raw noisy data, whereas KNN and Random Forest models outperformed other models with pre-processed clean data. The data is more of lifestyle data and prediction is made using features of a particular day and the values may vary drastically each day in and out, and as a result the accuracy can be questionable.

The research Nai-aruna & Moungmai, proposed the machine learning model to predict the diabetes mellitus risk among people (Nai-aruna & Moungmai, 2015). The data used is collected from 26 primary health care centres in Thailand, and therefore doesn’t represent the global population as a whole, since the ethnic origin and lifestyle largely varies globally. The model uses four different ML classification models such as Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes, followed by Bagging and Boosting Techniques to improve the robustness of these ML models. The results showed that the Random Forest classifier outperformed all the other models and hence was use for creating a web-based application to predict the diabetes risk.

# 2.2. Critical Review

The major problem with the existing research were the lack of study happening on younger age demographics and also diabetes being a lifestyle disease, cannot be predicted if not studied for a dataset with diverse data from different demographics of data. Hence the research should be done to identify the models that might suit different demographics of individuals as the lifestyle changes based on the age, ethnicity, gender, etc. Also, when the results of these previous studies are considered, different models perform better with different sets of data and hence will have to identify the best among all these models and compare the performance of these models to identify which model gives maximum precision. There might not be a single model that give precise prediction for all demographics or data sets, so we will study different demographics by collecting multiple sets of data or the data with mixed demographics data.

# CHAPTER 3 – METHODOLOGY

# 3.1 Overview

The research methodology means the techniques used to build a more organised structure in the execution and evaluation of data acquired in order to achieve the research's objective and aims. This chapter on methodology of the research focuses on the different research techniques, technology, analytical instruments, machine learning models, and algorithms which can be utilized for gathering and preparing data, carry out analysis, and discuss the models that are employed to forecast diabetes among the research subjects.

# 3.2. Research Aim

The main aim of this study is to analyse and forecast diabetic patient diagnoses using the best machine modelling approaches. The overall goal of the project is to develop and analyse machine learning models which can precisely predict people with or without the risk of diabetes based on their habitual, clinical and demographic features to increase the accuracy, efficiency, and timeliness of diabetes diagnosis. This aim is further divided into simpler objectives such as create prediction models that can reliably categorise patients into diabetic and non-diabetic groups or successfully identify persons at risk of diabetes by using big and diverse datasets, cutting-edge machine learning algorithms, and suitable feature selection techniques, use data analysis and feature engineering techniques to gain insight on the various features in the dataset collected, to reach meaningful conclusions, develop and assess different machine learning algorithms, such as Logistic Regression, Decision Trees, AdaBoost, SVM and CatBoost Classifier, for the prediction of Diabetes, compare the performance of these ML models using performance evaluation metrics such as accuracy, precision, F1 score and recall and hence give medical practitioners trustworthy, data-driven tools to aid in early diagnosis and intervention.

# 3.3. Research Framework

This section includes the overall research framework or design planned for this research study. Based on the objectives and aim of the research, the overall framework is made into simple steps – Data collection, data preprocessing, data analysis, data visualisation, ML model exploration and selection, and Prediction system as in the figure below.

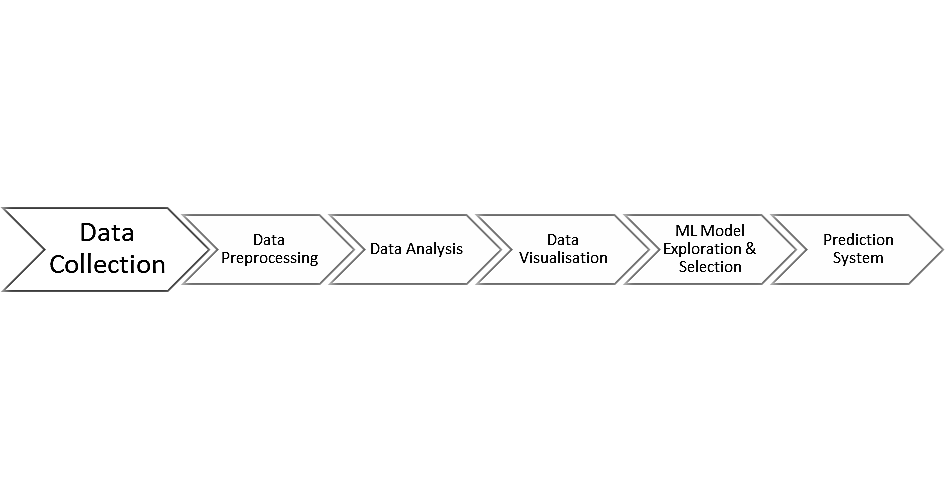


Figure 8: Research Framework

# 3.4. Data collection methods

Data collection is one of the primary and critical steps in the process of any research and there are several methods for collecting data. Only with relevant and right data, the modelling and prediction can be successfully completed or reach any conclusions. The nature of the research depends on the data collection. There are different types of data based on their source – primary data collected by the researcher themselves first hand from the research subject and secondary data which was originally collected by other researchers or organisations and then used by the researcher for further studies (Pandey & Pandey, 2015). In this research, researcher uses secondary data collected from Kaggle an open-source platform. This data will be used for analysis and modelling. We also collect pre-published literature in the subject area from various data sources such as digital library databases, university library, books, reports by organisations, etc. to get more insights in the research area and identify gaps in the literature.

# 3.5. Data Pre-processing

Data pre-processing is very critical in the modelling and analysis. The data collected from Kaggle is pre-processed as this data is raw data collected from an open-source platform and might not be suitable for immediate use. The process of data preprocessing involves prep preparing this raw data collected for analysis, by removing or altering the data which might be incorrect, missing, incomplete, irrelevant, or redundant. The need for this step is that if not used, the data which might have missed values or incorrect information will be used for analysis and modelling and similar to the presence of outliers, this may influence the results hence giving false conclusions or totally failing the process. The data pre-processing step involves finding missing values, either delete or impute these values based on some selected category, removing irrelevant data fields or columns, finding duplicate data and deleting them, correcting any errors in syntax or spelling, rename fields for better understanding of what they mean, and standardising the data set.

# 3.6. Data analysis

Data analysis is the process of analysing the data to gain meaningful insights and hence understanding the dataset, its variables and the relationships between the variables better. Here we use exploratory data analysis on the data collected from Kaggle once the pre-processing step is completed. This will give insights on any outliers, the behavioural patterns and trends within the dataset and the visualisation of the data and features are done using various plotting libraries such as Matplotlib and Seaborn plots. Graphical visualisation such as histograms, bar graphs, pie charts, heat map, etc. are used in this study for data analysis and visualisation.

# 3.7. Modelling Preparation

Once the data is visualised, and insights are gained the next step is preparing the data for modelling. For this first the dataset has to be split as test and train data and the dataset collected for this study is already divided into two datasets – train\_diabetes.csv and test\_diabetes.csv. Therefore, this step can be avoided. But the train dataset is used for training the model is divided into two – train and validation data set. The splitting is done for validation as the dataset are already separated. Here, the split is done as 20 percent for this validation test data set and balance 80 percent is used for training. This percentages can be chosen according to the size of the data set and requirements of the researcher.

# 3.8. Modelling

Machine learning has found applications in various industries and sectors and few of its major applications in healthcare sector are disease diagnosis and prognosis, medical image analysis, drug discovery, personalized treatment recommendations, etc. Machine learning models can be trained to identify trends and patterns in any dataset and use this patterns and trends to make decisions, provide recommendations, and make predictions. This has helped healthcare industry vastly and, in this research, this capability of machine learning models to identify such patterns and predict is employed. There are mainly two types of ML models - supervised and unsupervised learning models. As name suggests, supervised learning models are trained on dataset containing inputs and corresponding outputs, whereas in unsupervised, the datasets doesn’t contain output data but the model identifies the patterns and make conclusions. This research employs supervised models such as Logistic Regression, AdaBoost, Support Vector Machine, CatBoost Classifier, and Decision Trees.

# 3.9. Model Evaluation

The models perform differently with different set of data and this performance can be measured using various metrics for performance evaluation such as accuracy, recall, F1 score, precision, etc. These values are calculated based on the true values and predicted values. Accuracy measures the proportion of correct predictions to total prediction, precision measures the proportion of correct positive predictions to total positive predictions, recall measures the percentage of positives that were correctly predicted and F1 score is a harmonic mean of precision and recall.

# CHAPTER 4 – ANALYSIS AND FINDINGS

# 4.1. Overview

In this chapter, the data is analysed and classified using various machine learning models. For this Python programming language is employed by the researcher and Google Colaboratory, which is a free cloud-based platform by Google LLP is used for writing and executing the Python code in an environment of Jupyter Notebook. It is widely used among researchers as it offers free Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs), that significantly speed up the execution of tasks such as intensive mathematical functions like training and testing.

# 4.2. Data Analysis and Findings

The first step in the analysis and modelling is importing the various libraries, models and evaluation metrics essential for analysis.

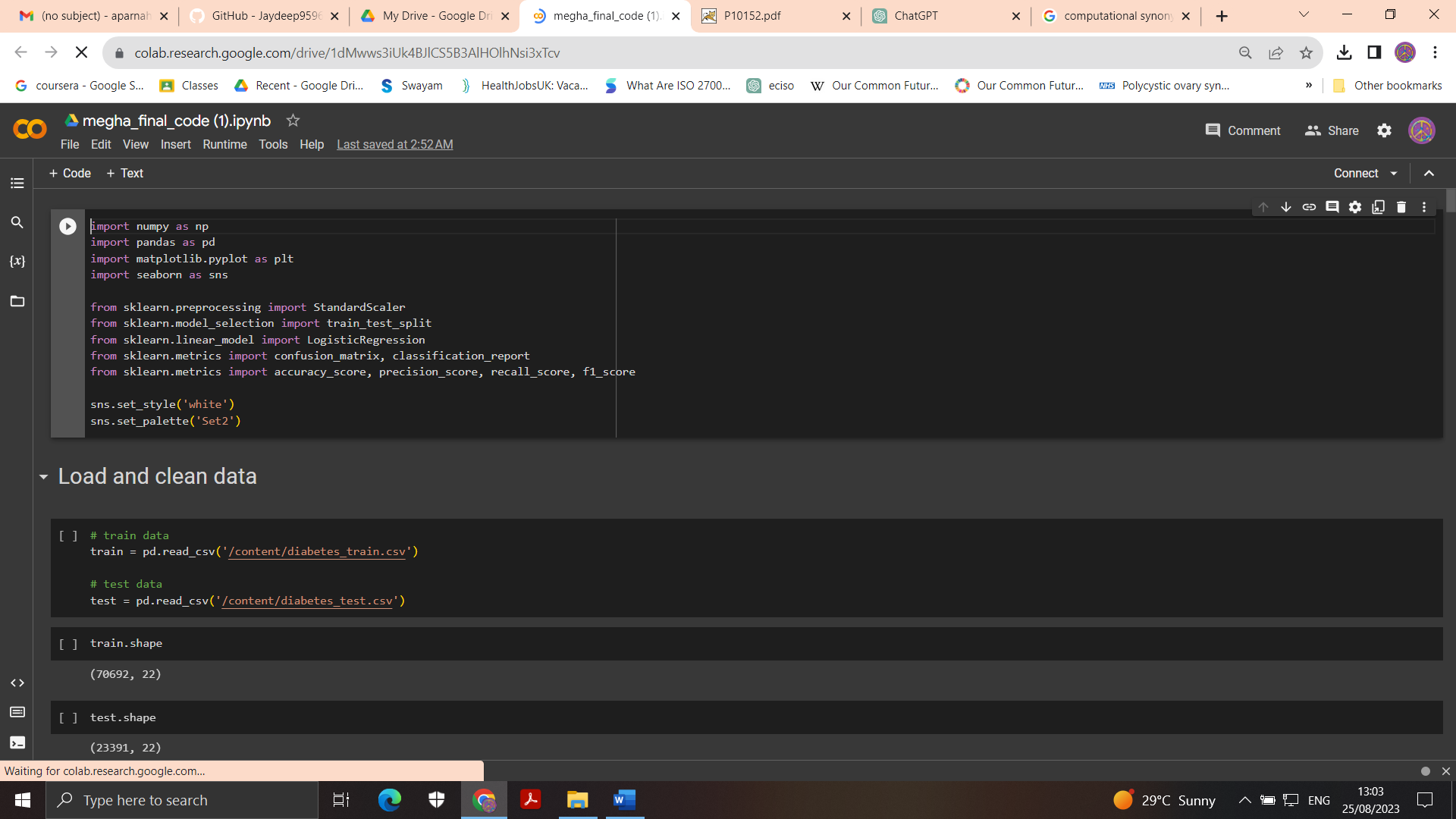


Figure 9: Importing essential libraries and models

“import numpy as np” imports the NumPy library, which is a foundational Python package for scientific computing that includes support for working with arrays and matrices, as well as a wide range of mathematical functions that may be applied to these arrays.

“import pandas as pd” is used to import Pandas, which is another essential Pandas is another essential Python library, used for data manipulation, analysis, and cleaning, and provides data structures and functions, which helps researchers to work with structured data sets, like tables, spreadsheets, and databases.

“import matplotlib.pyplot as plt” imports matplotlib.pyplot, a submodule of Python based data Matplotlib, which has a collection of functions for visualisation and plotting.

“import seaborn as sns” imports Seaborn library, which is built on top of the Matplotlib, and offers an interface for creation and presentation of more informative and more appealing statistical visualizations and graphs, especially in statistical data analysis and exploration.

“from sklearn.preprocessing import StandardScaler” imports StandardScaler, a widely used data preprocessing technique for machine learning workflows for standardisation and scaling features. This is imported from sklearn.preprocessing module within scikit-learn (also known as sklearn), a Python based machine learning library.

Similar to the previous code, “from sklearn.model\_selection import train\_test\_split” imports train\_test\_split function, which is used to split the available dataset into training and testing data sub sets from the sklearn.model\_selection module in sklearn library.

In the step, “from sklearn.linear\_model import LogisticRegression” is used for the machine learning model Logistic Regression is imported from the sklearn library. Other models which are used in the research will be

“from sklearn.metrics import confusion\_matrix, classification\_report” import performance evaluation metrics confusion matrix and classification report from the sklearn.metrics module from sklearn library. The confusion matrix is used for computing a matrix based on the true and predicted labels (True Positive, True Negative, False Positive and False Negative) whereas the classification report generates the report on the different performance evaluation metrics used such as precision, recall, F1-score, etc.

“from sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score” imports the various performance evaluation metrics from the sklearn.metrics module in sklearn library. accuracy\_score computes the accuracy, which is a widely used metric for measuring the overall performance of the ML model, which is calculated as the ratio of number of correctly predicted instances to total number of instances. precision\_score computes the precision of the model, which quantifies the percentage of actual positive predictions that were correct. recall\_score computes recall (or sensitivity/true positive rate) of the model and measures the efficiency of model in identifying the positive instances. f1\_score computes the F1-score or the harmonic mean of precision and recall, and is a more efficient metric when the classes are imbalanced.

The code “sns.set\_style('white')” is used to set the style of the Seaborn plots to an available style named 'white' to give the plots a clean and simple appearance and “sns.set\_palette('Set2')” to set the colour palette of the Seaborn plots to the predefined Set2, which define the aesthetics of the data points, lines, bars, etc. of the plots.

Once the libraries, models and performance evaluation metrics are imported, the training and testing data sets were loaded and cleaned.

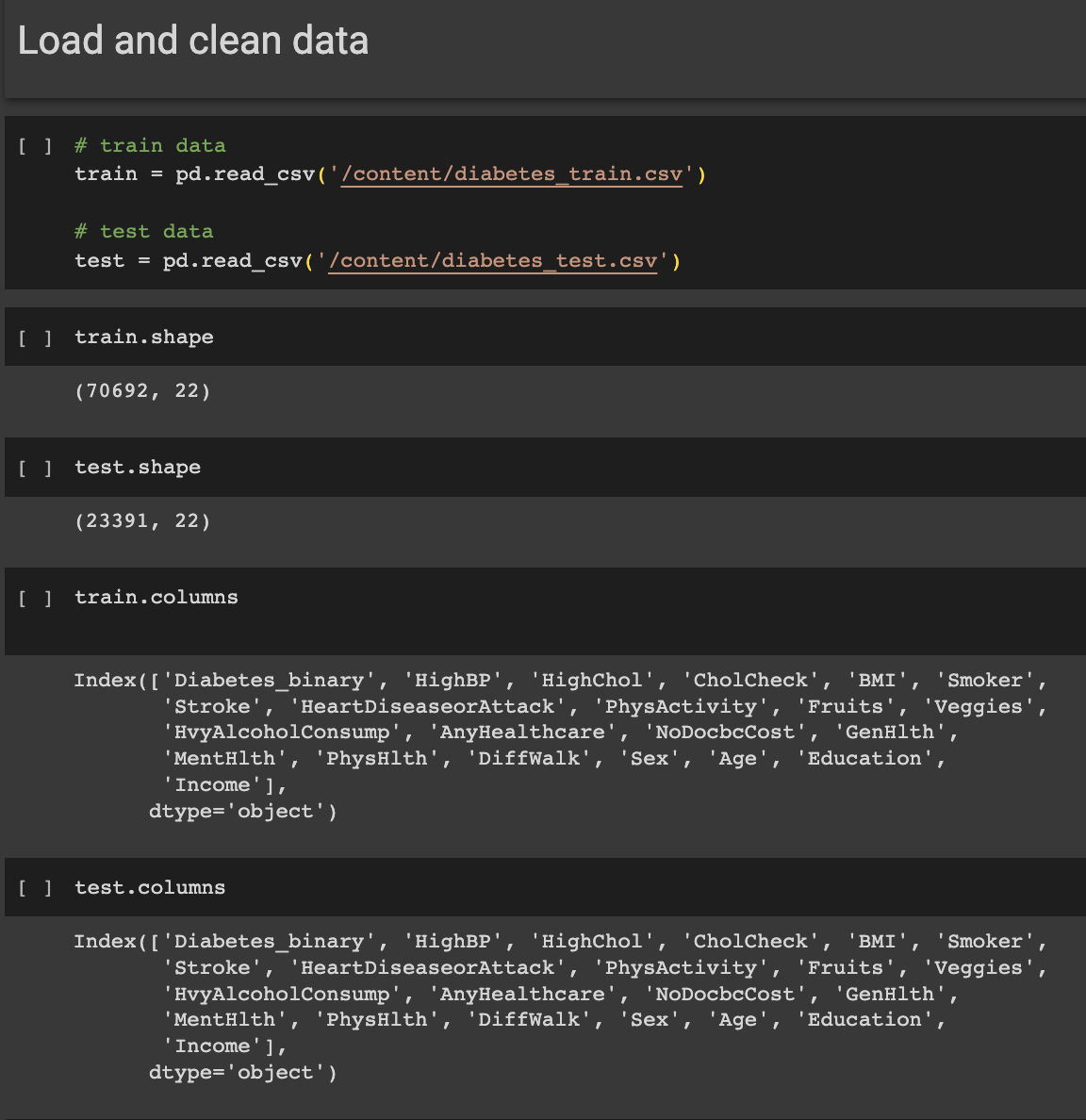


Figure 10: Loading and cleaning the data sets

The data set collected from Kaggle is already divided as two different data sets, train data and test data with 70692 instances and 23391 instances respectively. The train data set named diabetes\_train.csv and diabetes\_test.csv are loaded and stored to the DataFrame named train and test respectively. The path of these csv files is located in the content directory in the Google Colaboratory.

Using train.shape and test.shape codes, the dimensions (number of rows, number of columns) of the DataFrame train and test are returned. The values shows that there are 70692 and 23391 rows or instances in the respective DataFrame and has 22 column features or column headers, that indicates the various features used for predicting diabetes in the patients.

The code train.columns and test.columns, returns the index object from the Pandas library, that contains the column labels or headers within the DataFrame train and test respectively. This will give the insight on the various features that are used for analysis and prediction and it can be observed that the train and test DataFrame has similar features in the same order, hence doesn’t require to take any preprocessing or cleaning actions here. The column headers are:

* Diabetes\_binary
* HighBP
* HighChol
* CholCheck
* BMI
* Smoker
* Stroke
* HeartDiseaseorAttack
* PhysActivity
* Fruits
* Veggies
* HvyAlcoholConsump
* AnyHealthcare
* NoDocbcCost
* GenHlth
* MentHlth
* PhysHlth
* DiffWalk
* Sex
* Age
* Education
* Income



Figure 11: Checking for missing values in the train DataFrame

The missing values are to be identified before modelling as part of the data preprocessing and this is done using the train.isnull().sum() code and identified that there are no null values within the train DataFrame.

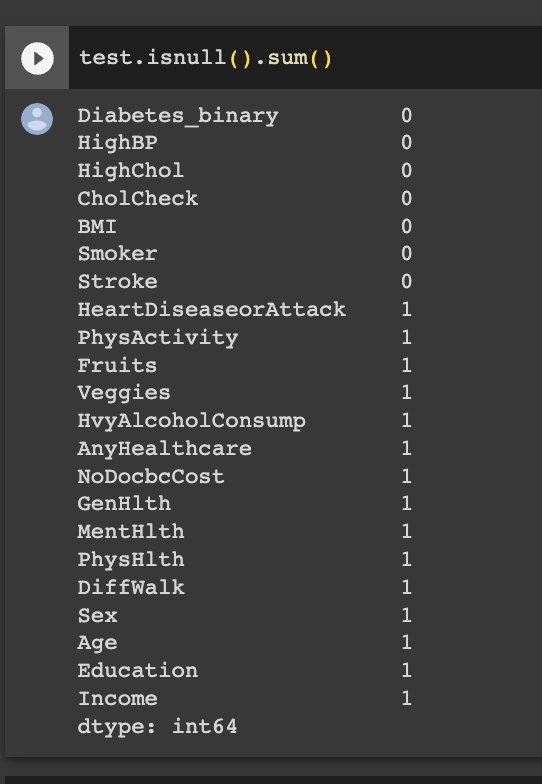


Figure 12:Checking for missing values in the test DataFrame

The missing values are to be identified before modelling as part of the data preprocessing and this is done using the test.isnull().sum() code and identified that there are few null values within the test DataFrame such as HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, and Income.



Figure 13:Checking the data types in the train DataFrame

The data types of the features are to be identified before modelling as part of the data preprocessing and this is done using the train.dtypes and is seen that the data types of all the features are 64 bit float or float64.

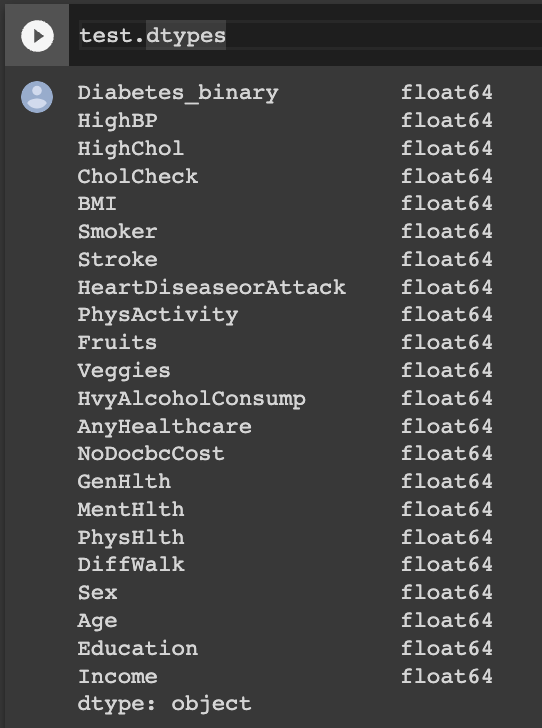


Figure 14: Checking the data types in the test DataFrame

The data types of the features are to be identified before modelling as part of the data preprocessing and this is done using the test.dtypes and is seen that the data types of all the features are 64 bit float or float64.

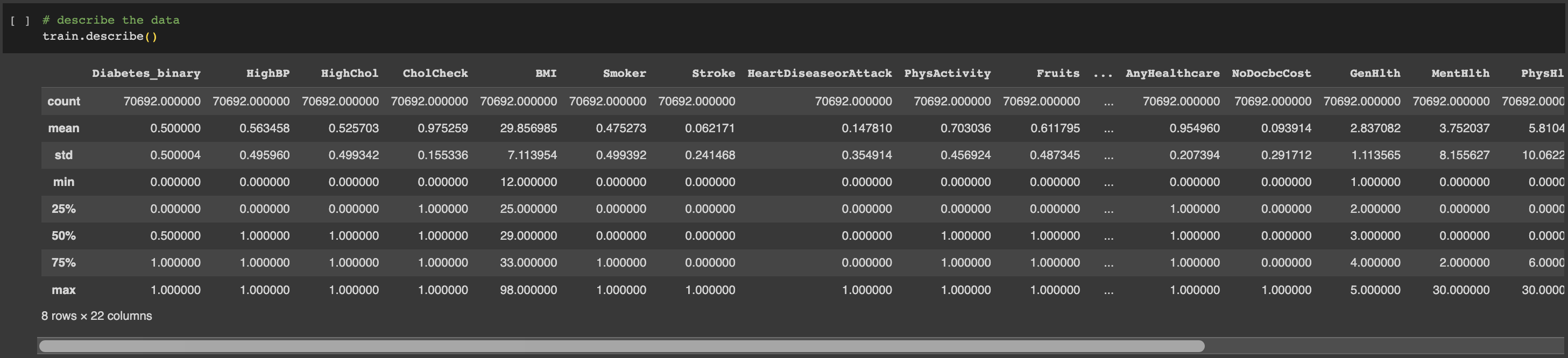


Figure 15:Descriptive statistics of the features within the train DataFrame

The code train.describe(). T is used for getting the summary statistics or descriptive statistics of the train data frame such as count, which gives the number of non-null values of the attribute or feature, mean or average value of the attribute, standard deviation of the attribute, minimum and maximum values, 25% or 25th percentile, which is the lower quartile, 75% or 75th percentile, which is the upper quartile and 50% or 50th percentile, which is the median value. These statistics helps in understanding the nature of the data in more detail. The nature of the distribution (normal or not) can be determined using the mean and median values as well.

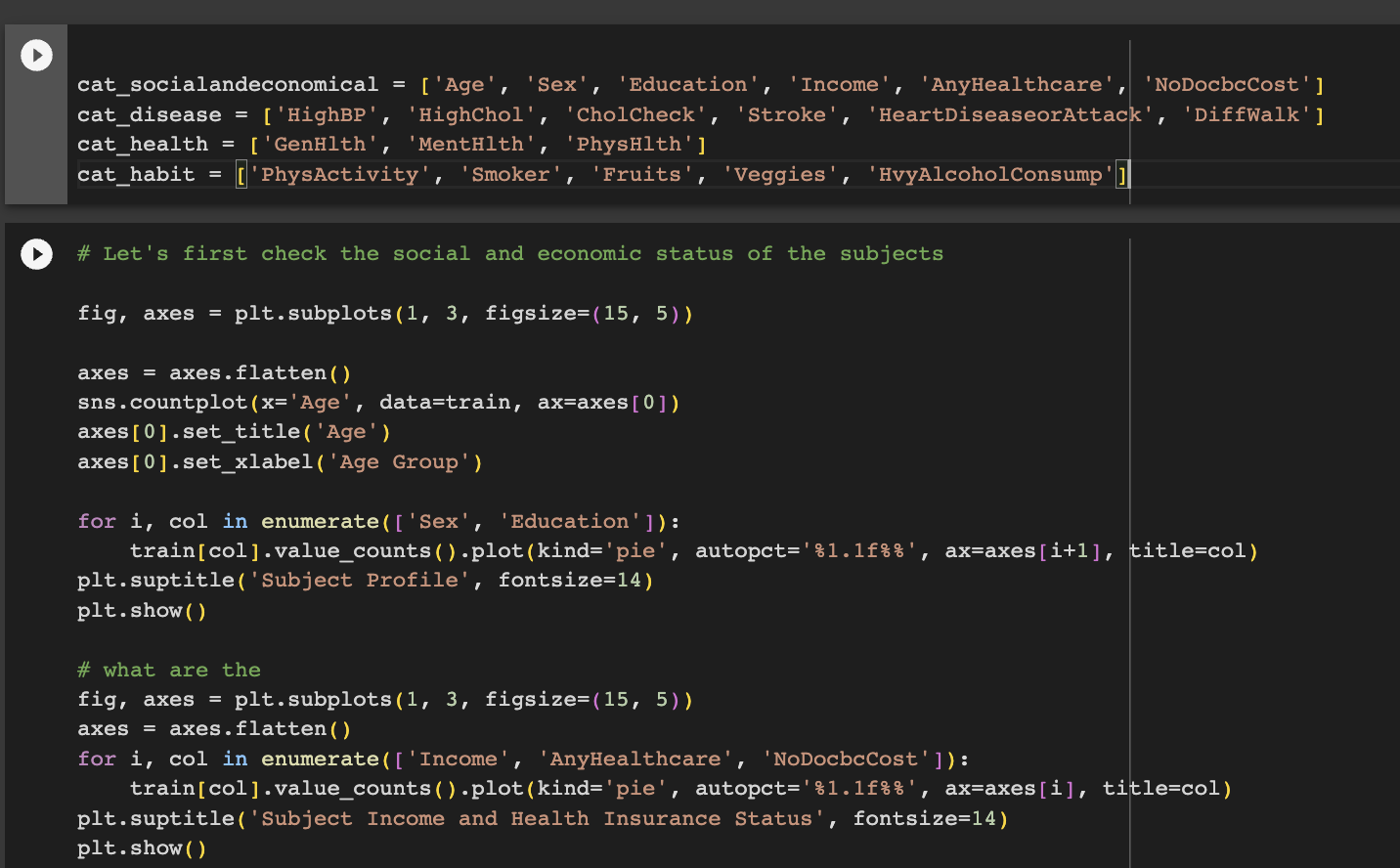


Figure 16: The features are classified into various groups based on the nature

In this section, the features are classified into various categories within the train DataFrame. The features can be related to the social and economic factors of the patient, related to disease history, related to health and habitual factors. The output shows that Age, Sex, Education, Income, AnyHealthcare, and NoDocbcCost are classified into cat\_socialandeconomical, HighBP, HighChol, CholCheck, Stroke, HeartDiseaseorAttack, DiffWalk as cat\_disease, GenHlth, MentHlth, PhysHlth into cat\_health and PhysActivity, Smoker, Fruits, Veggies, HvyAlcoholConsump into cat\_habit, which as social and economic, disease, health and habit respectively.

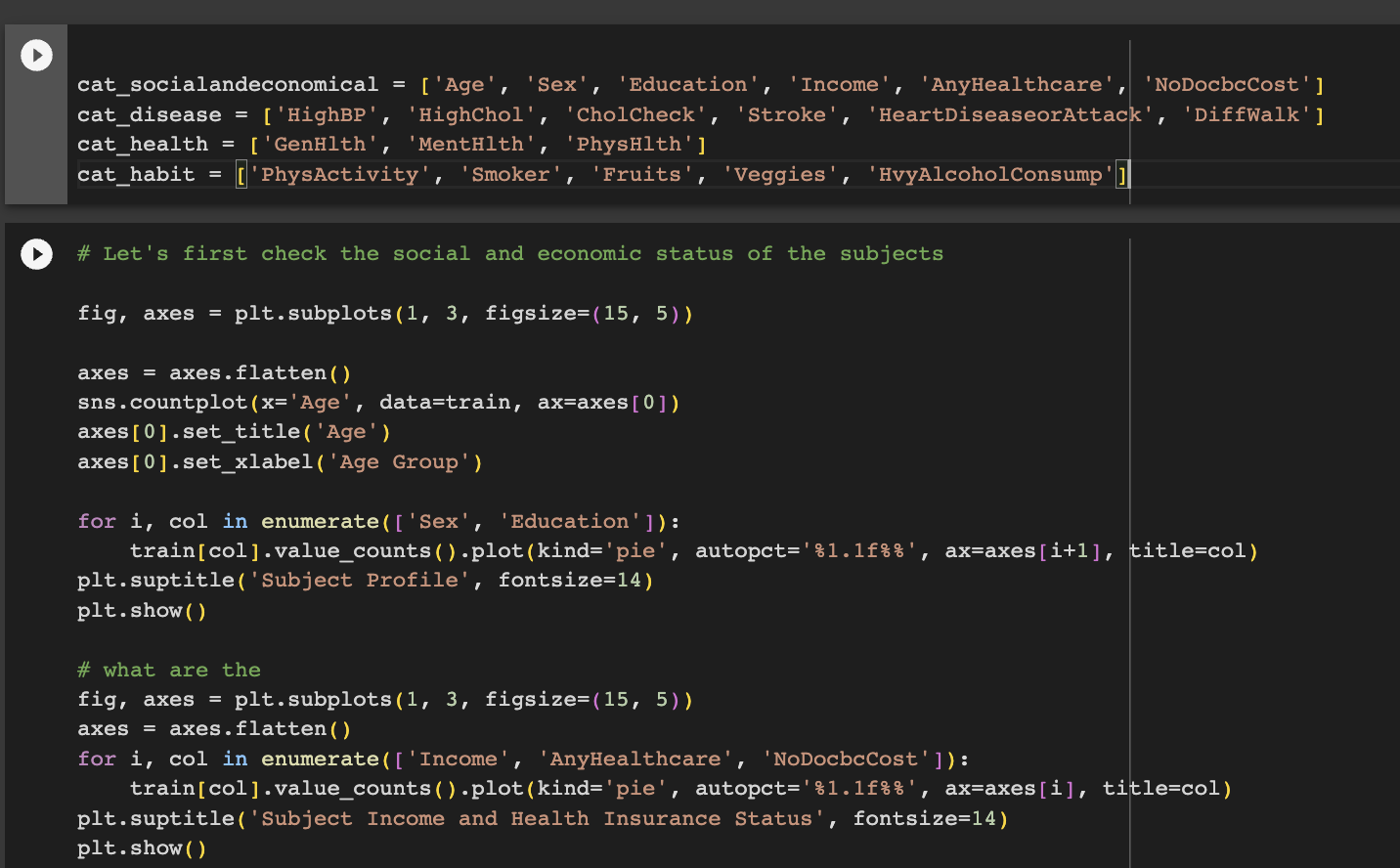


Figure 17: Plotting the social and economic features

The social and economic categorical features are plotted using the Seaborne plots in order to understand the nature of the distribution and the titles were given based on the nature of these features such as subject profile for the subject demographic and social features like age, sex and education, while using subject income and health insurance status for the economic features like income, AnyHealthcare (any health care insurance or benefits access) and NoDocbcCost (No visit to doctor because of cost).

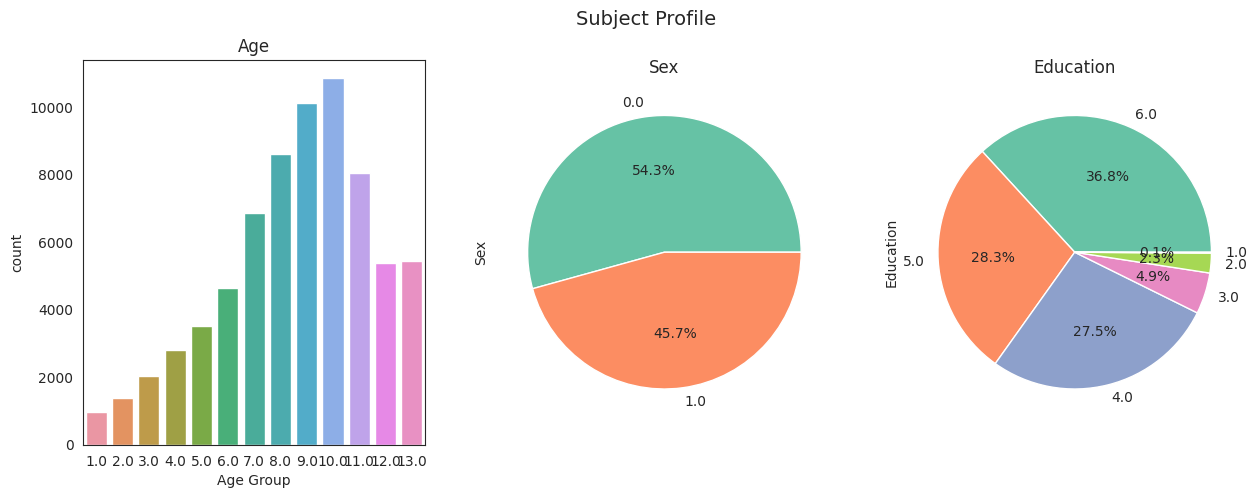


Figure 18: Social and Economic Features distribution – Subject Profile

The Subject profile is the features that define the subject such as their age, sex and education. The researchers who collected this dataset primarily have used fourteen level age categories, which is given as follows:



Figure 19: Fourteen Level age category

The bar graph for age shows that the data sets contained data of adults above the age of 18 and age of the maximum number of subjects were in category 10, i.e., the age group of 65 to 69, followed by category 9 (60 to 64) and then by 8 (55 to 59). Also, we have observed from the descriptive statistic that mean and median of age are 8.58 and 9 respectively, which means that mean values lie between 55 to 59, whereas median lies between 60 to 64. It can be observed that the standard deviation from the mean is very high, since the values are distributed unevenly varying from 18 to 80 plus years of age.

The sex or gender of the patient is presented as a pie chart and 0 represents Males and 1 represents Female subjects. The figure 10 shows that the number of male subjects in the dataset is larger than the number of female subjects.

The feature education was also categorised into 6 categories based on the levels of education and is represented using pie chart. While education levels 1,2,3 indicates set of people who did not graduate high school and discontinued education at various levels prior to high school [1 - Never attended school or only kindergarten, 2 - Grades 1 through 8 (Elementary) and 3 - Grades 9 through 11 (Some high school)], education level 4 includes people who graduated high school or GED, education level 5 includes subjects who attended college or technical school but not graduated and education level 6 includes subjects who graduated from college or Technical school. In this category, while the maximum number of subjects are having education level 6, followed by 5 and 4 respectively, only less than 10 percent of the total test subjects did not graduate high school. The median and mean values of this distribution are 5 and 4.92.

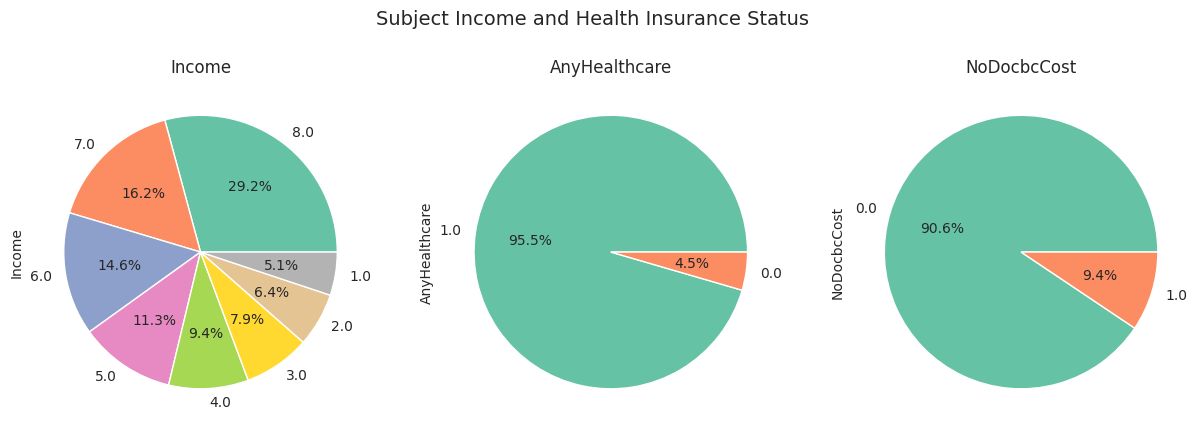


Figure 20: Social and Economic Features distribution – Subject Income and Health Insurance Status

The subject income and health insurance status shows the economic status of the subjects. Here the income levels are classified into 8 categories – 1 - less than $10,000, 2 - $10,000 to less than $15,000, 3 - $15,000 to less than $20,000, 4 - $20,000 to less than $25,000, 5 - $25,000 to less than $35,000, 6 - $35,000 to less than $50,000, 7 - $50,000 to less than $75,000, 8 - $75,000 or more. The maximum subjects were having a stable financial situation, i.e., level 8 ($75,000 or more) and the percentage of subjects with an income less than $10,000 is only 5.1 percentage. This might mean that there is probability that the financial status prevents patients from getting diagnosed as they are unable to afford the healthcare facilities. Also, the mean and median are 5.7 and 6 respectively. Hence can say that mean is somewhere between $25,000 and $50,000 and median lies between $35,000 to less than $50,000.

AnyHealthcare feature means if the subjects have any health care coverage or access and 0 indicates subjects without any health care coverage and 1 indicates subjects with health care coverage. The percentage of subjects with health care coverage is more and this indicates the strong leniency of patients without health coverage towards avoidance health checks and diagnosis.

Similarly, NoDocbcCost feature indicates the subjects that could not consult a doctor because of the cost. Here, 1 indicates subjects who couldn’t afford to visit a doctor because of cost and 0 indicates subjects who could afford to visit a doctor regardless of the cost. 9.4 percent of the subjects were not able to visit a doctor as they couldn’t afford to pay for the visit.

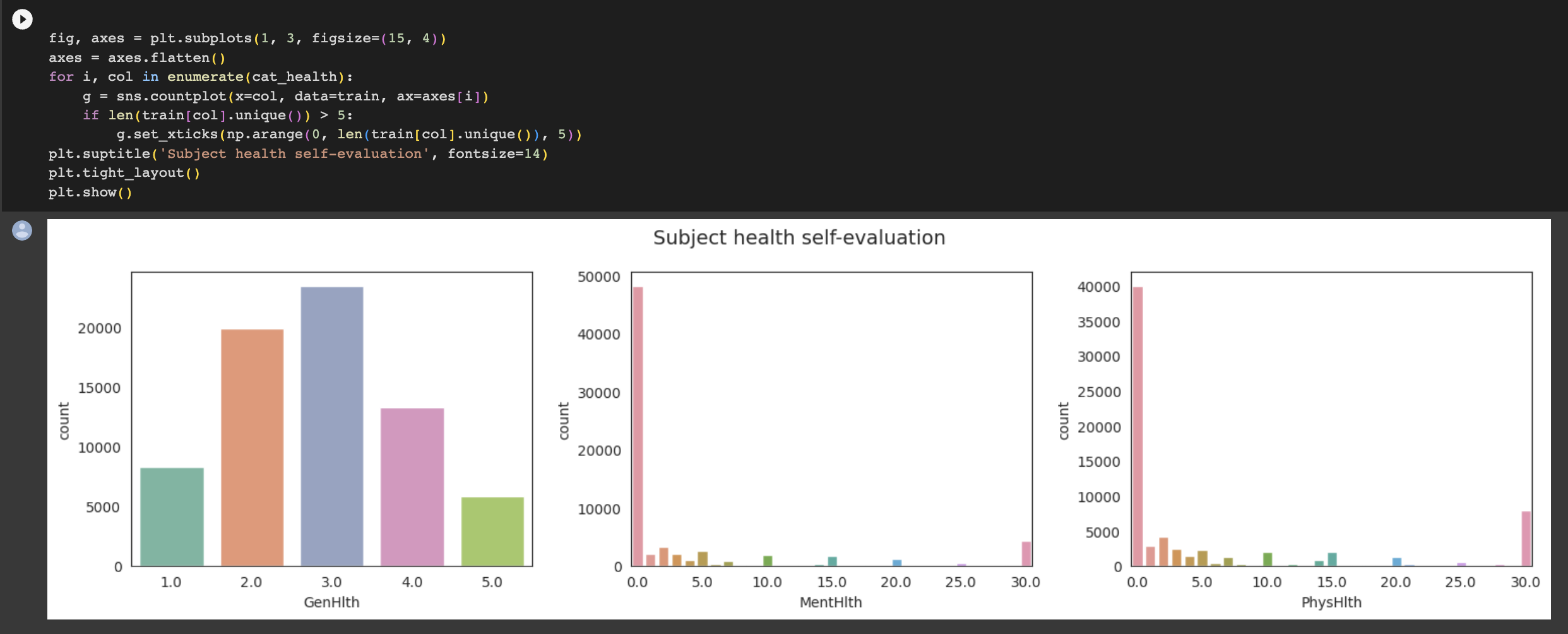


Figure 21: Plotting Health Category features distribution

The health features are plotted using the Seaborne plots in order to understand the nature of the distribution and the title was given based on the nature of these features subject health self-evaluation.

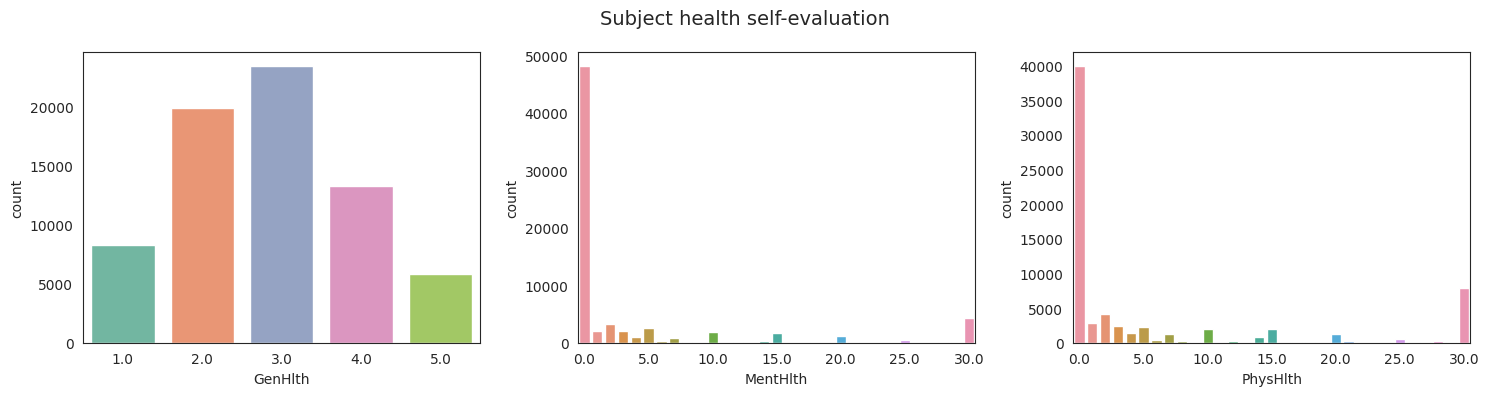


Figure 22: Health Category features distribution

The features that were self-evaluated by the subject on their health was collected and bar graphs were used for plotting them. The general health feature GenHlth was collected on a scale of 1 – 5, where the subjects were asked about their general health status (1 – Excellent, 2 – Very Good, 3 - Good, 4 - Fair, 5 - Poor). Most of the subjects responded that their health was good, followed by a good share of subjects responding that their health is very good. The mean value of this feature is 2.84, which means most of subjects thinks that their health status is either very good or good.

The mental health of the subjects (MentHlth) was collected based on their responses on the number of days when their mental health was not good due to stress, depression, and problems with emotions, during the past 30 days of data collection and it was observed that a large majority of the subjects either didn’t feel that their mental health was not good on any days or felt so for a maximum of one day.

Similarly, physical health of the subjects (PhysHlth) was collected based on their responses on the number of days when their physical health was not good due to physical illness and injury, during the past 30 days of data collection and it was observed that a large majority of the subjects either didn’t feel that their physical health was not good on any days or felt so for a maximum of one day.



Figure 23:Plotting Disease Category features distribution

The disease features are plotted using the Seaborne plots in order to understand the nature of the distribution and the title was given based on the nature of these features proportion of disease/health issue indicators as it comprised of features like high BP status, high cholesterol status, cholesterol check status, stroke status, heart disease or attack status and difficulty in walk.



Figure 24:Disease Category features distribution

The pie charts plotted for the categorical features HighBP, HighChol, CholCheck, Stroke, HeartDiseaseorAttack and DiffWalk, where 0 indicates No and 1 indicates Yes. The percentage of subjects who were diagnosed with high blood pressure and high cholesterol are high within the DataFrame and are 56.35% and 52.57 percent respectively. The number of subjects who have never checked the cholesterol are about 2.47 percent and hence there is a probability of increase in the percentage of subjects with high cholesterol. Almost 6.22 percent of the subjects had stroke, while 14.78 percent had heart diseases or had suffered from one or more heart attacks. More than one fourth of the subjects had difficulty in walking or climbing stairs, which might be used to other health condition or aging.

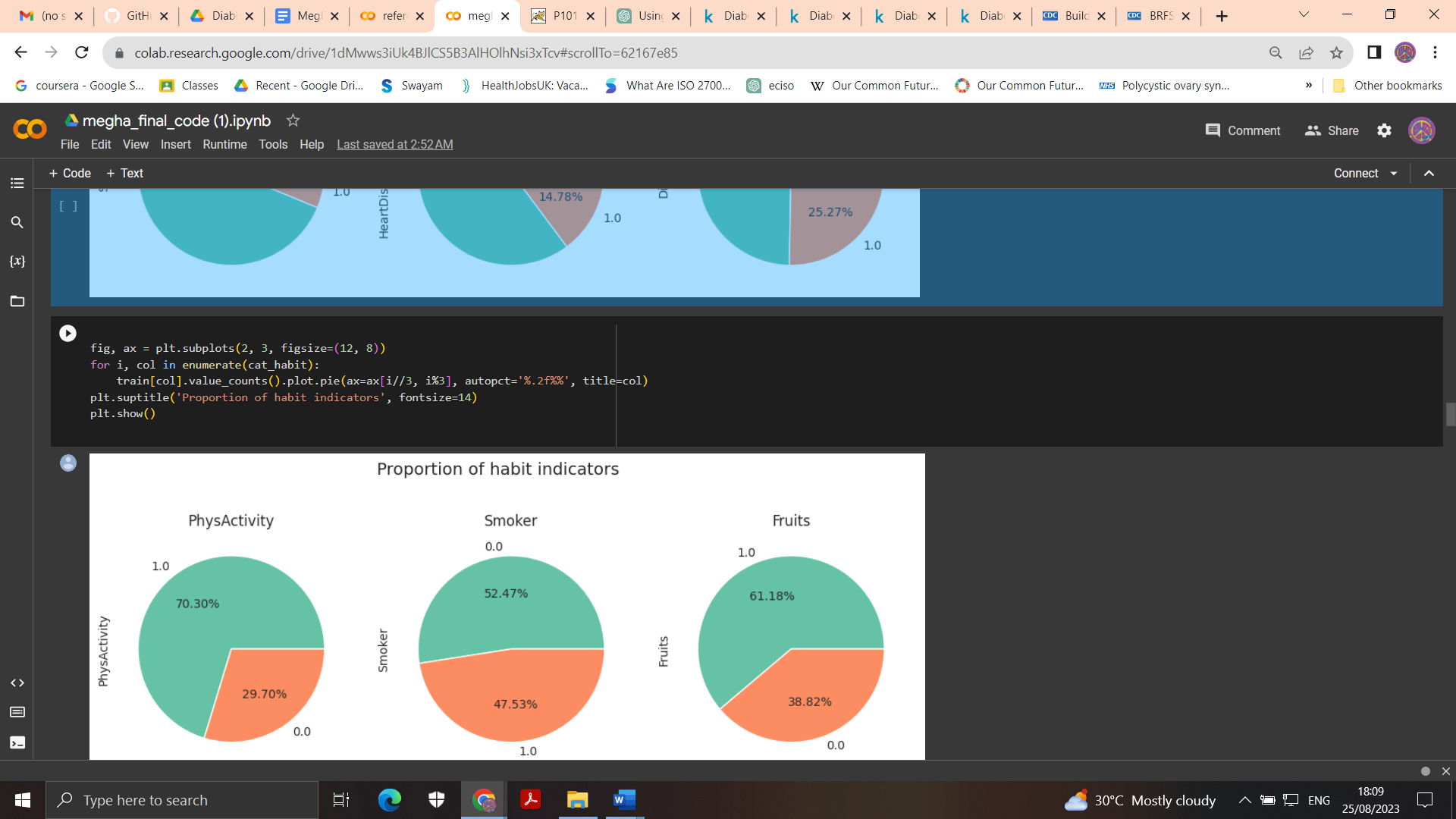


Figure 25: Plotting Habits Category features distribution

The habit features are plotted using the Seaborne plots in order to understand the nature of the distribution and the title was given based on the nature of these features as proportion of habit indicators as it comprised of features like physical activity, smoking habits, eating habits like including fruits and vegetables in the meals and heavy alcohol consumption or alcoholism.

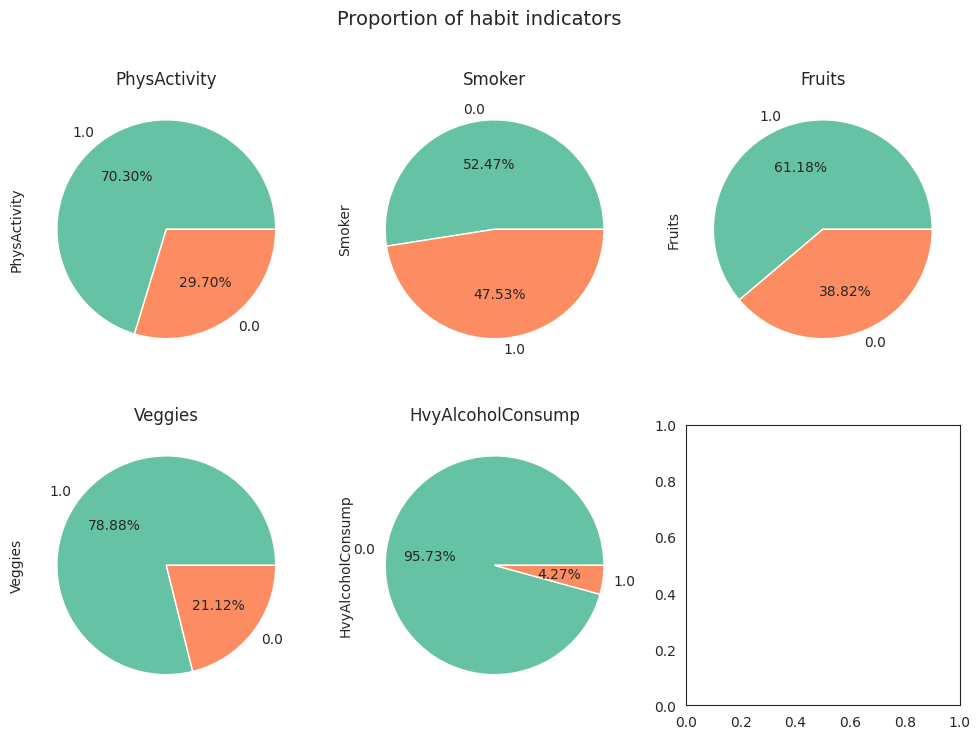


Figure 26: Habits Category features distribution

The pie charts plotted for the categorical features PhysActivity, Smoker, Fruits, Veggies, and HvyAlcoholConsumption, where 0 indicates No and 1 indicates Yes. The percentage of subjects who were participated in any physical activities or exercises such as running, calisthenics, golf, gardening, walking for exercise, etc. during the past 30 days of data collection was more than 70 percent within the DataFrame. The percentage of subjects who were smokers was 47.53 percent, but only 4.27 percent of the subject were prone to heavy alcoholism or have shown heavy alcoholic consumption habit. The fruits and Vegetables consumption pattern among the members were also comparatively high with 61.18 percent and 78.88 percentage respectively.

Further, we plot the non-categorical data BMI to understand the frequency distribution and other statistical information about it.

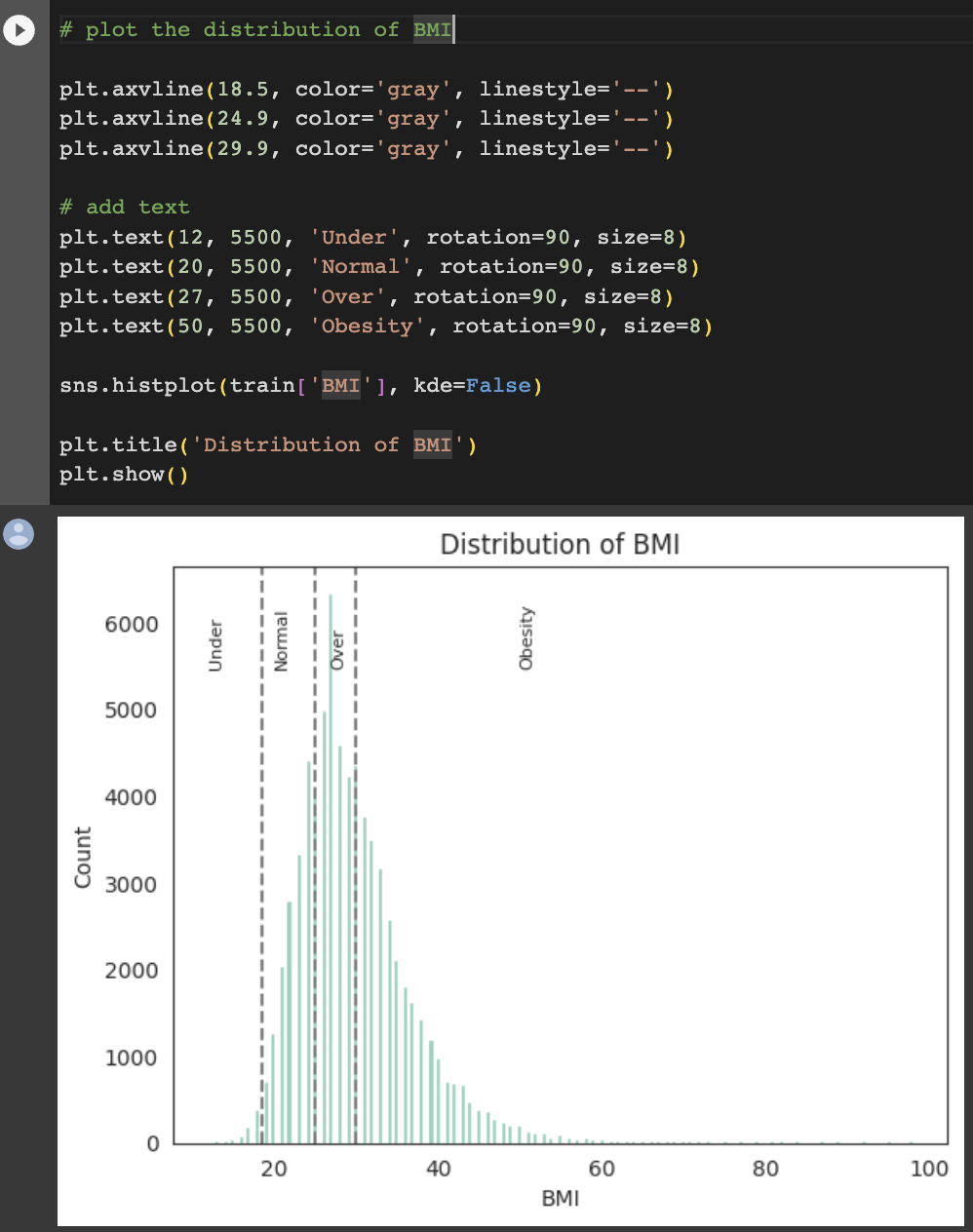


Figure 27: Plotting histogram for BMI

The BMI or body mass index of the subjects were plotted as histogram, where the BMI is categorised as below 18.5 underweight, above 18.5 to 24.9 as normal, above 24.9 to 29.9 as overweight and above 29.9 as obese. It can be observed that highest number of subjects in this data set are overweight and a large number of subjects are obese. The mean and median values from the descriptive analysis shows that mean is 29.85 whereas median is equal to 29.

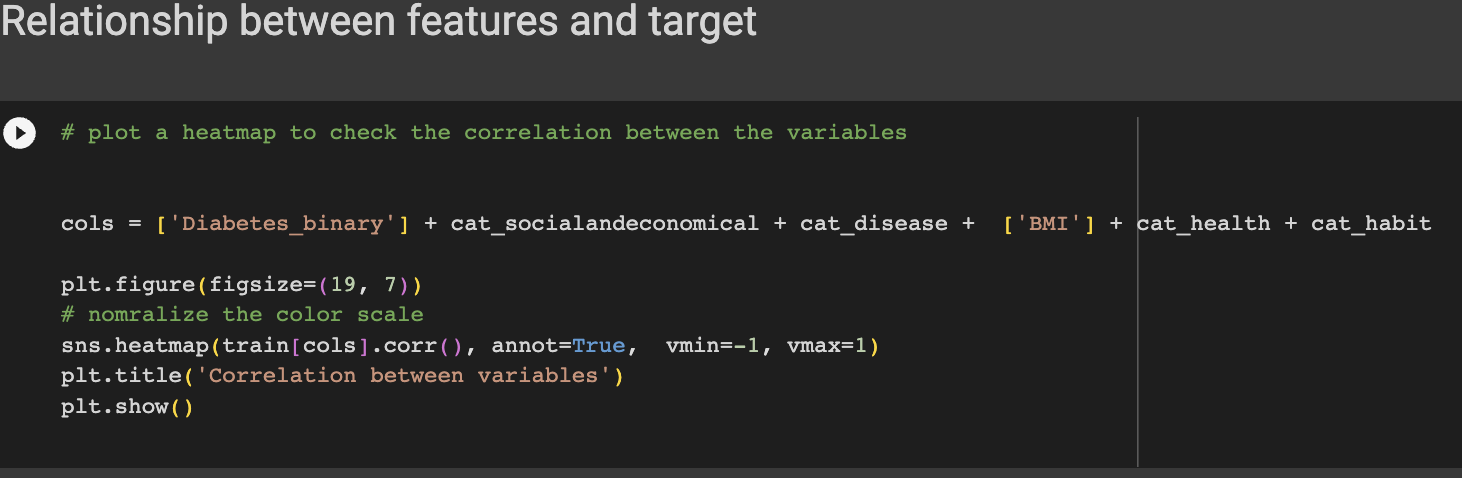


Figure 28: Plotting the relationship between features using Heatmap

Heatmap is plotted used to show the correlation between the features in the DataFrame.

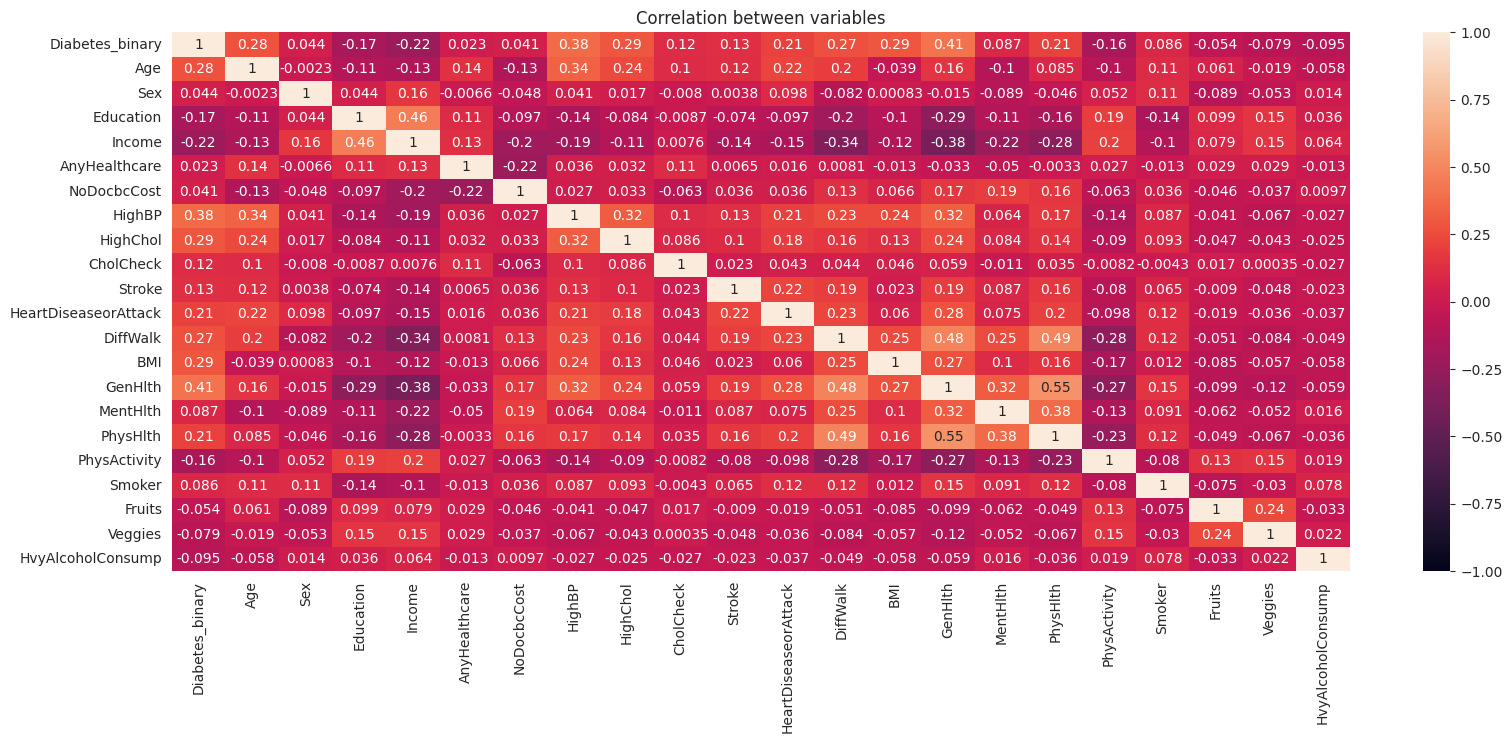


Figure 29: Heatmap

The heatmap shows that there is high positive correlation between the diabetes binary and age, high blood pressure, high cholesterol, cholesterol check, stroke, heart disease or attack, difficulty in walking, BMI, general health and physical health (increases in probability of having diabetes in proportion with increase in the features) and high negative correlation between education, income and physical activities (increase in these features decreases the probability of diabetes). If arranged in the descending order to indicate feature importance, the list will be as follows:

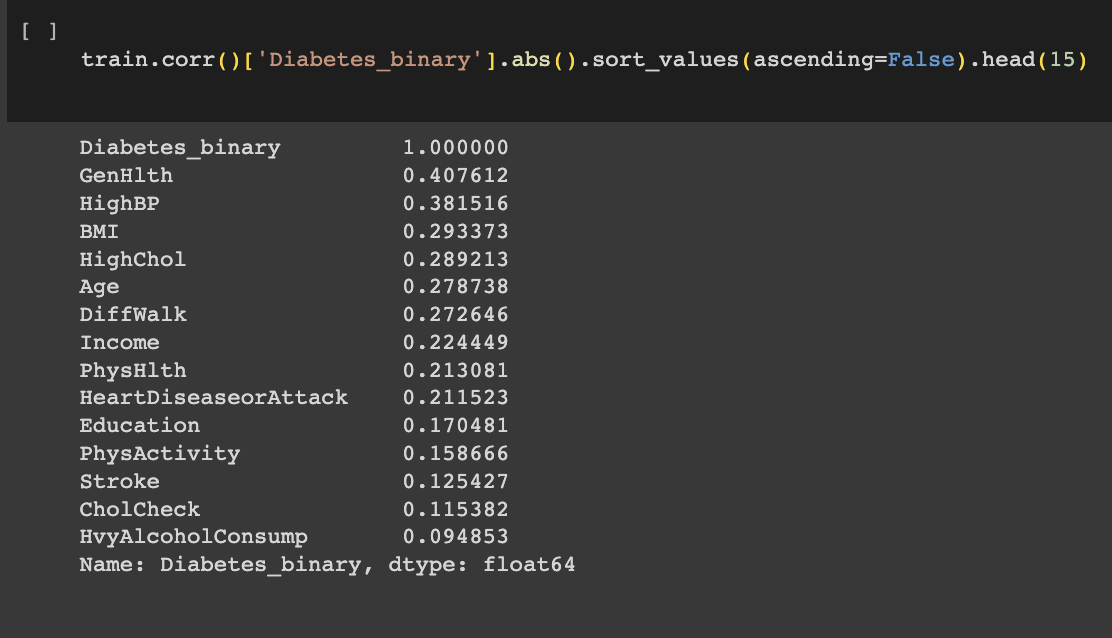


Figure 30: Correlation of data binary to features in descending order for train DataFrame

The equation gives the top 14 unique features with highest correlation with Diabetes\_binary, regardless of negative or positive correlation as GenHlth (+0.407612), HighBP (+0.381516), BMI (+0.293373), HighChol (+0.289213), Age (+0.278738), DiffWalk (+0.272646), Income (-0.224449), PhysHlth (+0.213081), HeartDiseaseorAttack (+0.211523), Education (-0.170481), PhysActivity (-0.158666), Stroke (+0.125427), CholCheck (+0.115382), and HvyAlcoholConsump (+0.094853).

The analysis of the dataframe has given better insights on the train data and feature distribution and now, the data preprocessing is done to prepare the data for modelling and prediction.

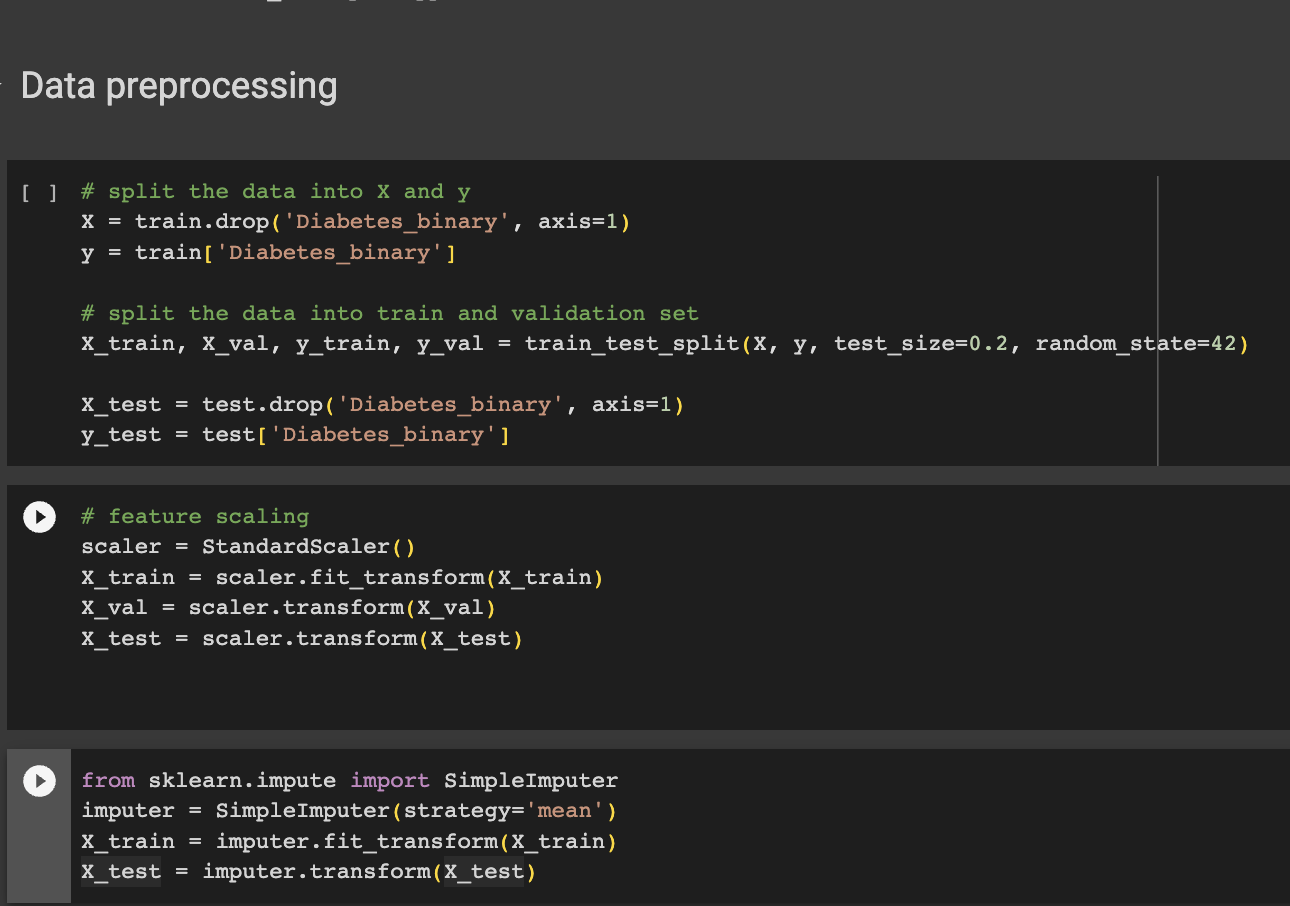


Figure 31: Data Preprocessing

The codes ‘X = train.drop('Diabetes\_binary', axis=1)’ is used to drop the column feature Diabetes\_binary from the train DataFrame and ‘y = train['Diabetes\_binary']’ is used to assign it as the target vector y which will be predicted using the models.

For splitting the data into train data and validation data subsets, ‘X\_train, X\_val, y\_train, y\_val = train\_test\_split(X, y, test\_size=0.2, random\_state=42)’ is used. Here, the size of the validation data set is 20 percent of the total data, hence using 80 percent of the data for training the models. The sampling is done at random to avoid any issues with reproducibility and also for better results.

Similarly the codes ‘X\_test = test.drop('Diabetes\_binary', axis=1)’ and ‘y\_test = test['Diabetes\_binary']’ are used to create feature matrix and target vector respectively for the test data set.

Feature scaling is the data preprocessing technique for ensuring that all the features have similar scales, so as to improve the performance of the models. First the scaler is chosen as StandardScaler, which uses the mean and other statistics for scaling. With codes ‘X\_train = scaler.fit\_transform(X\_train)’, the fit.transform method of scalar are used to scale the data. In fit.tansform, the mean and standard deviation of the data is calculated and scaling is done based on these statistics. This same mean and standard deviation values are used for scaling the validation and test data.

Once scaling is completed, the SimpleImputer is imported from the sklearn.impute module for handling the missing values in the datasets. For this the strategy the SimpleImputer is set as mean, so that the mean value of the feature can be calculated and can be used for replacing the missing data. The fit\_transform method is used for the train data (X\_train), for imputing the missing values, where the mean of each feature is calculated using the non-missing values and is used to replace the missing values. For the test data (X\_test), this calculated mean from the train data is used for replacing the missing value.

Once the preprocessing is completed, the modelling and prediction are the next steps.

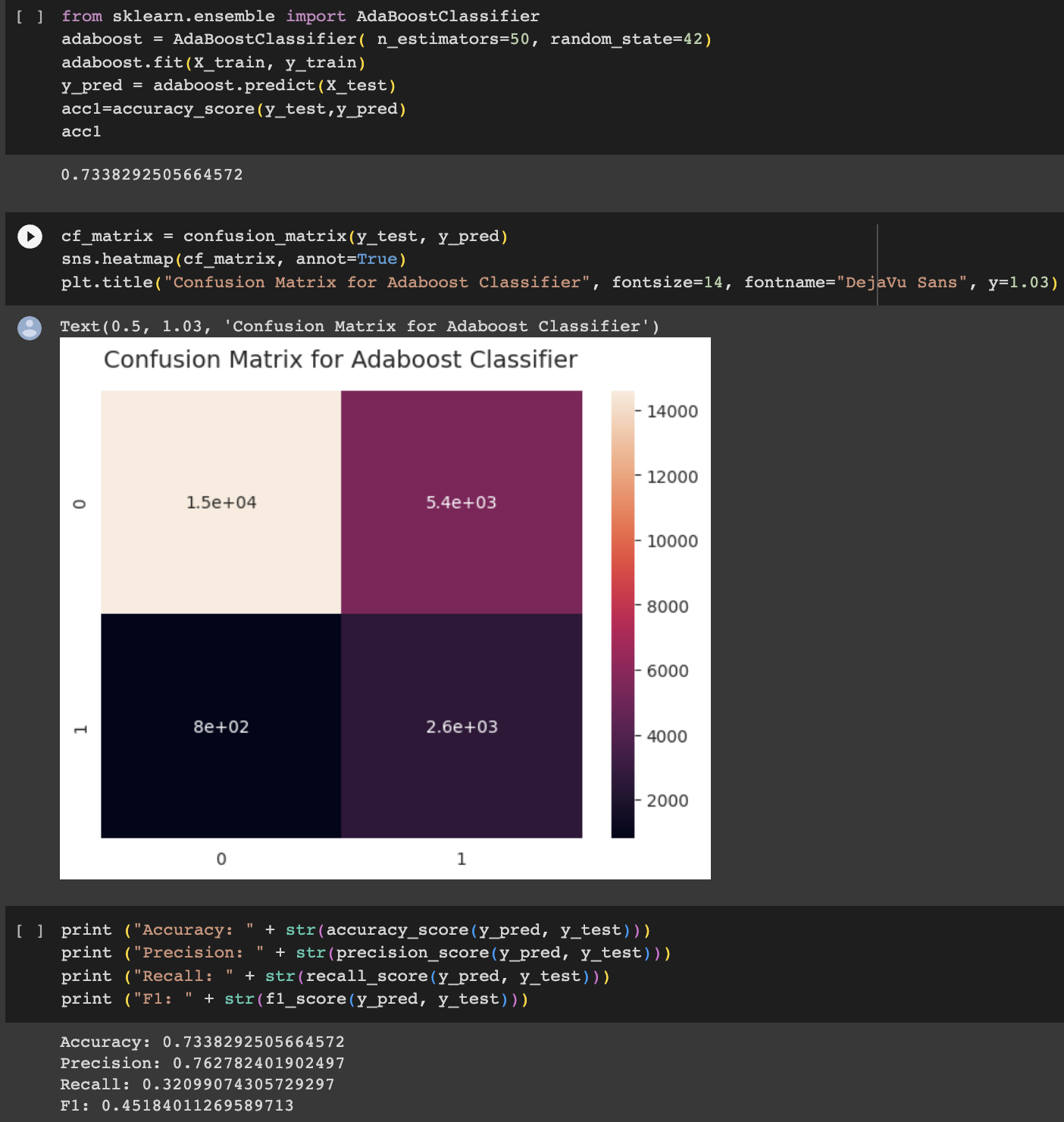


Figure 32: AdaBoost Classifier - Modelling and evaluation

The AdaBoost Classifier is imported from sklearn library and the ensemble model was created using 50 base estimators or weak learners and random\_state 42 for random number generator reproducibility. The code ‘adaboost.fit(X\_train, y\_train)’ is used to train the data using the training data and ‘y\_pred = adaboost.predict(X\_test)’ is used for making predictions. The accuracy of the model is calculated and printed. Then confusion matrix is plotted using Seaborn by comparing the predicted labels y\_pred and actual labels y\_test. The True Positives, False Positives, False Negatives and True Negatives are 1.5e+04, 5.4e+03, 8e+02 and 2.6e+03 respectively.

These values of predicted and actual labels were used to evaluate the accuracy, precision, recall and F1 score of the models and the results for the model was accuracy was 73.38 percent, precision equals 76.28, recall equals 32.10 percent and F1 score equals 45.18.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| AdaBoost | 73.38 | 76.28 | 32.10 | 45.18 |

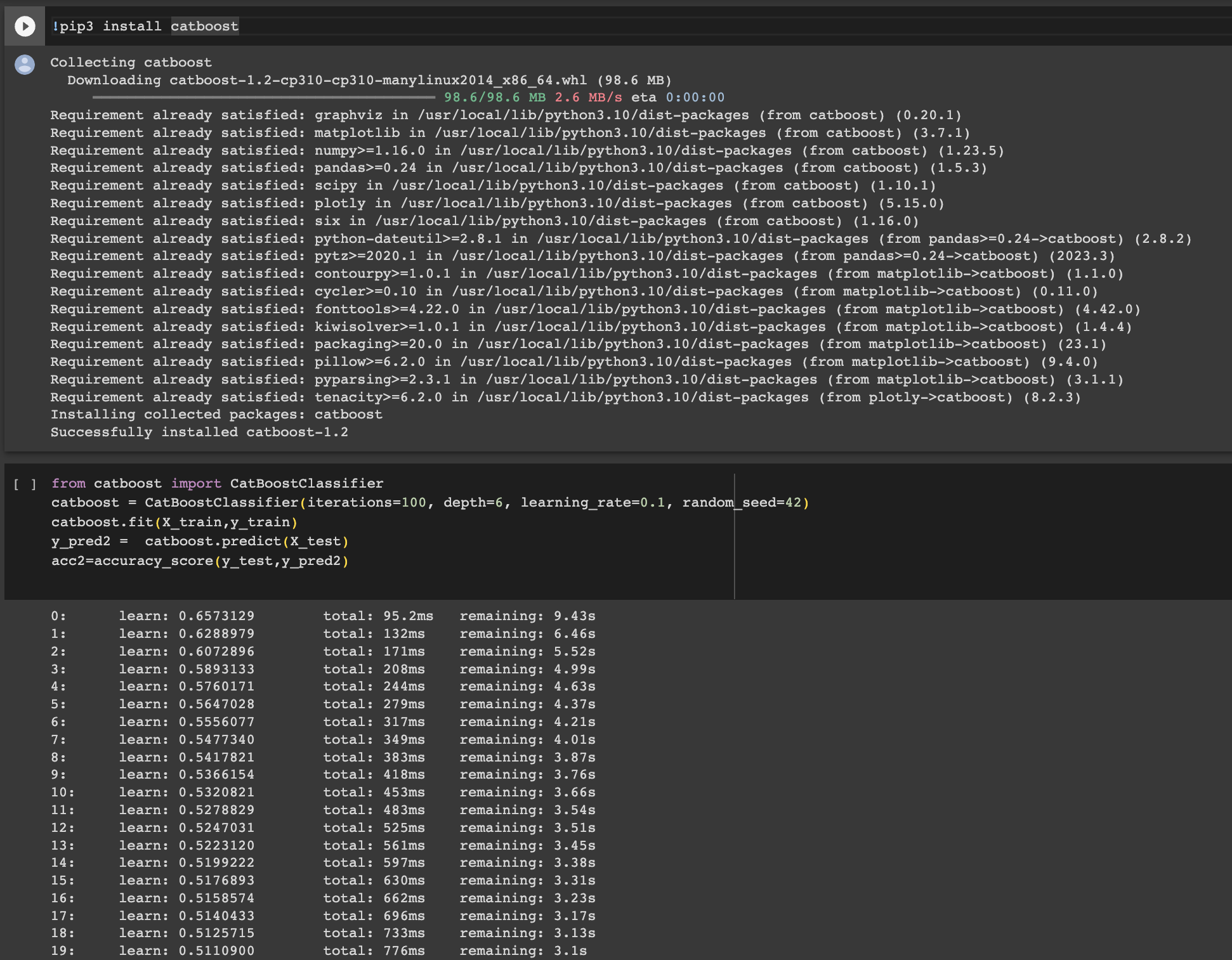


Figure 33:Installation and modelling of CatBoost Classifier

Here the CatBoost library is installed using the ‘pip3 install catboost’ code and once this is done, the CatBoost Classifier is imported from the CatBoost Library. This is an ensemble model with 100 weak learners to be trained and defined by iterations, the depth of each tree in ensemble is 6, the step size with which the boosting algorithm converges is given as the learning rate, which is equal to 0.1. The model is then trained and tested and the accuracy of the model is calculated.

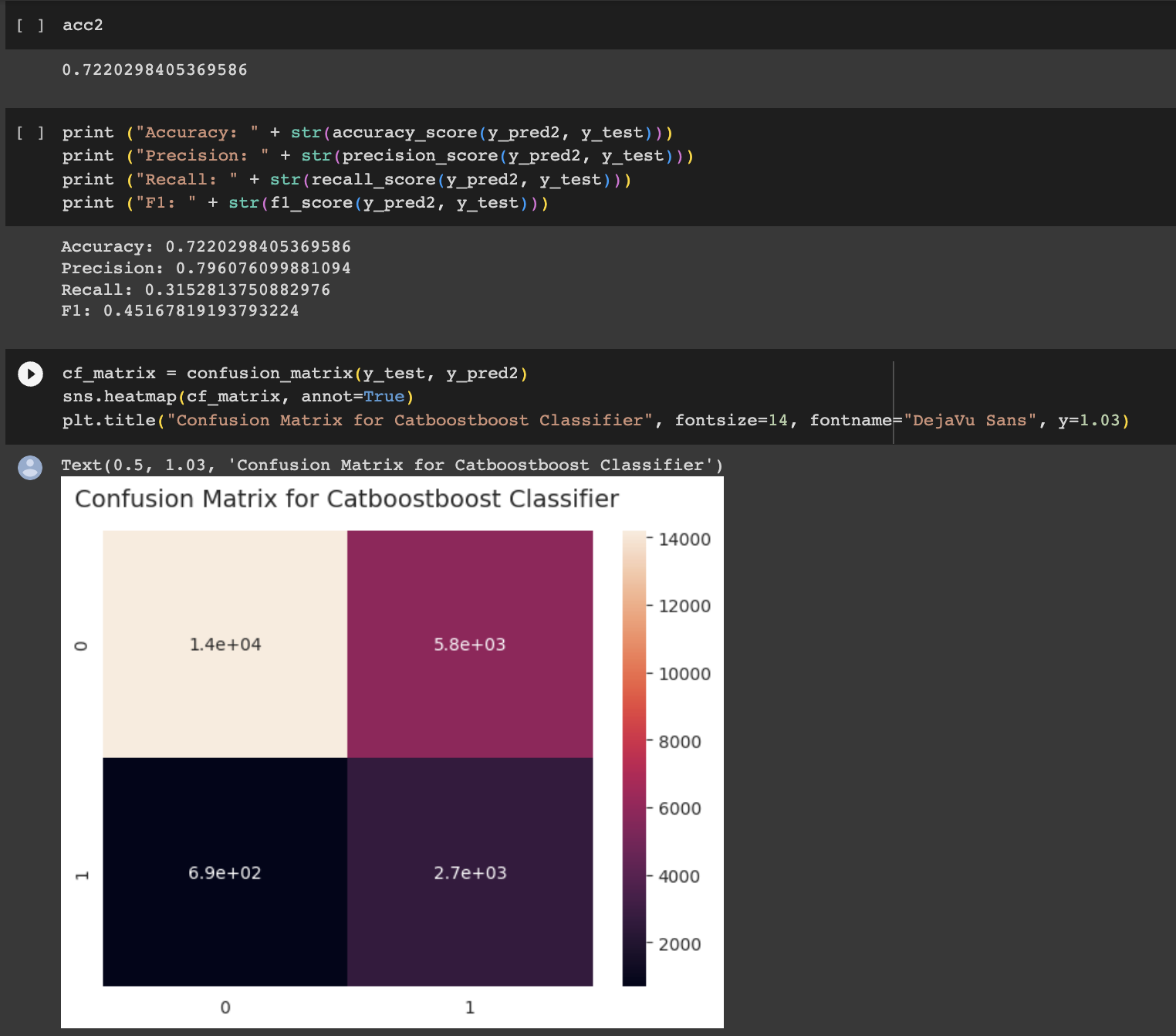


Figure 34: CatBoost performance evaluation

The accuracy of the model is calculated as 72.20 percent and then the model is further evaluated for precision, recall and F1 score, which gave values 79.61,31.53 and 45.17 percent respectively. Then confusion matrix is plotted using Seaborn by comparing the predicted labels y\_pred and actual labels y\_test. The True Positives, False Positives, False Negatives and True Negatives are 1.4e+04, 5.8e+03, 6.9e+02 and 2.7e+03 respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| CatBoost | 72.20 | 79.61 | 31.53 | 45.17 |

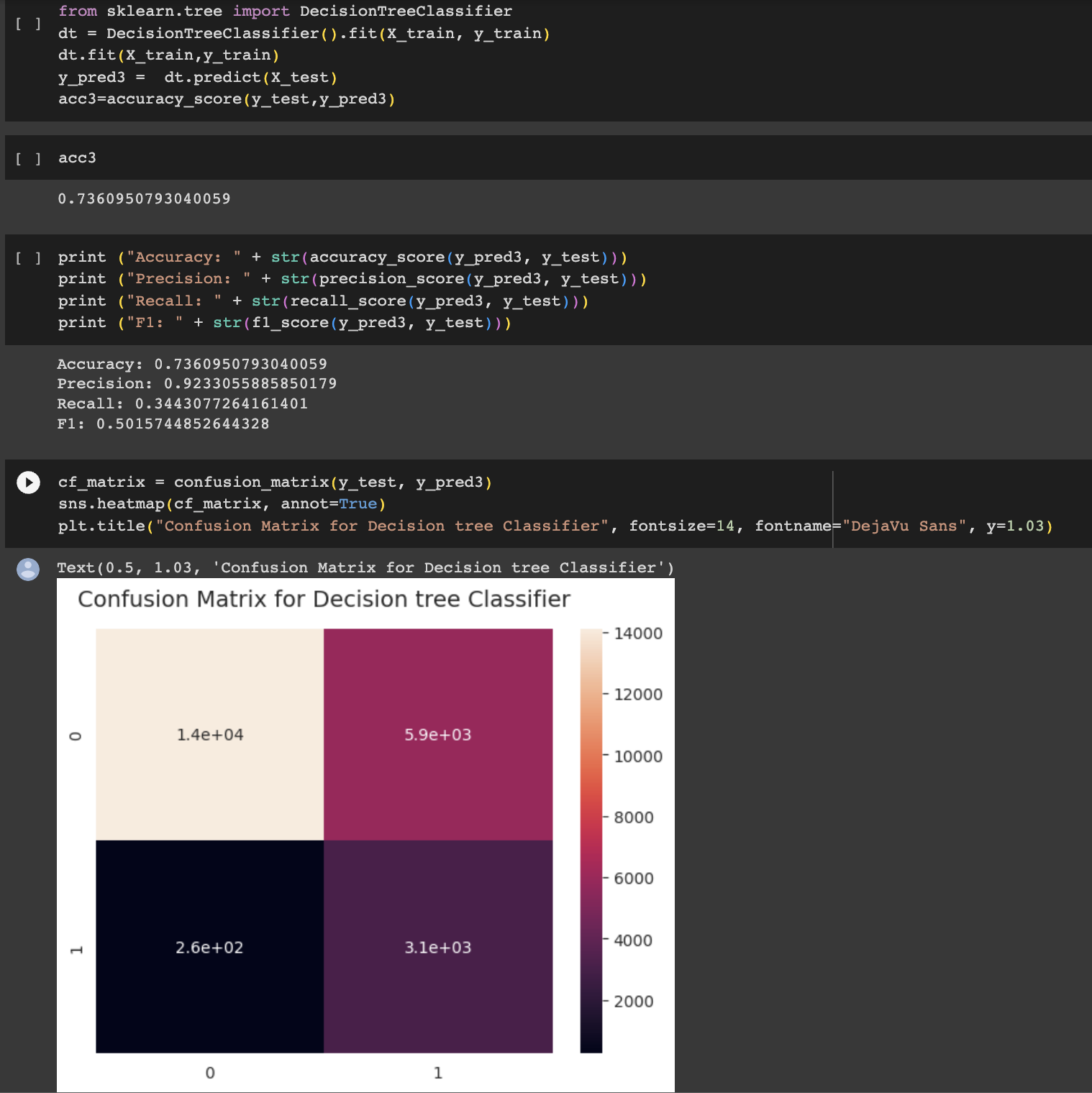


Figure 35: Decision Tree - Modelling and evaluation

The Decision Tree Classifier is imported from sklearn library and the classic model is trained using the train data and tested for predictions. The accuracy of the model is calculated and printed. Then confusion matrix is plotted using Seaborn by comparing the predicted labels y\_pred and actual labels y\_test. The True Positives, False Positives, False Negatives and True Negatives are 1.4e+04, 5.9e+03, 2.6e+02 and 3.1e+03 respectively.

The values of predicted and actual labels were used to evaluate the accuracy, precision, recall and F1 score of the models and the results for the model was accuracy was 73.60 percent, precision equals 92.33, recall equals 34.43 percent and F1 score equals 50.16.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| Decision Tree | 73.60 | 92.33 | 34.43 | 50.16 |

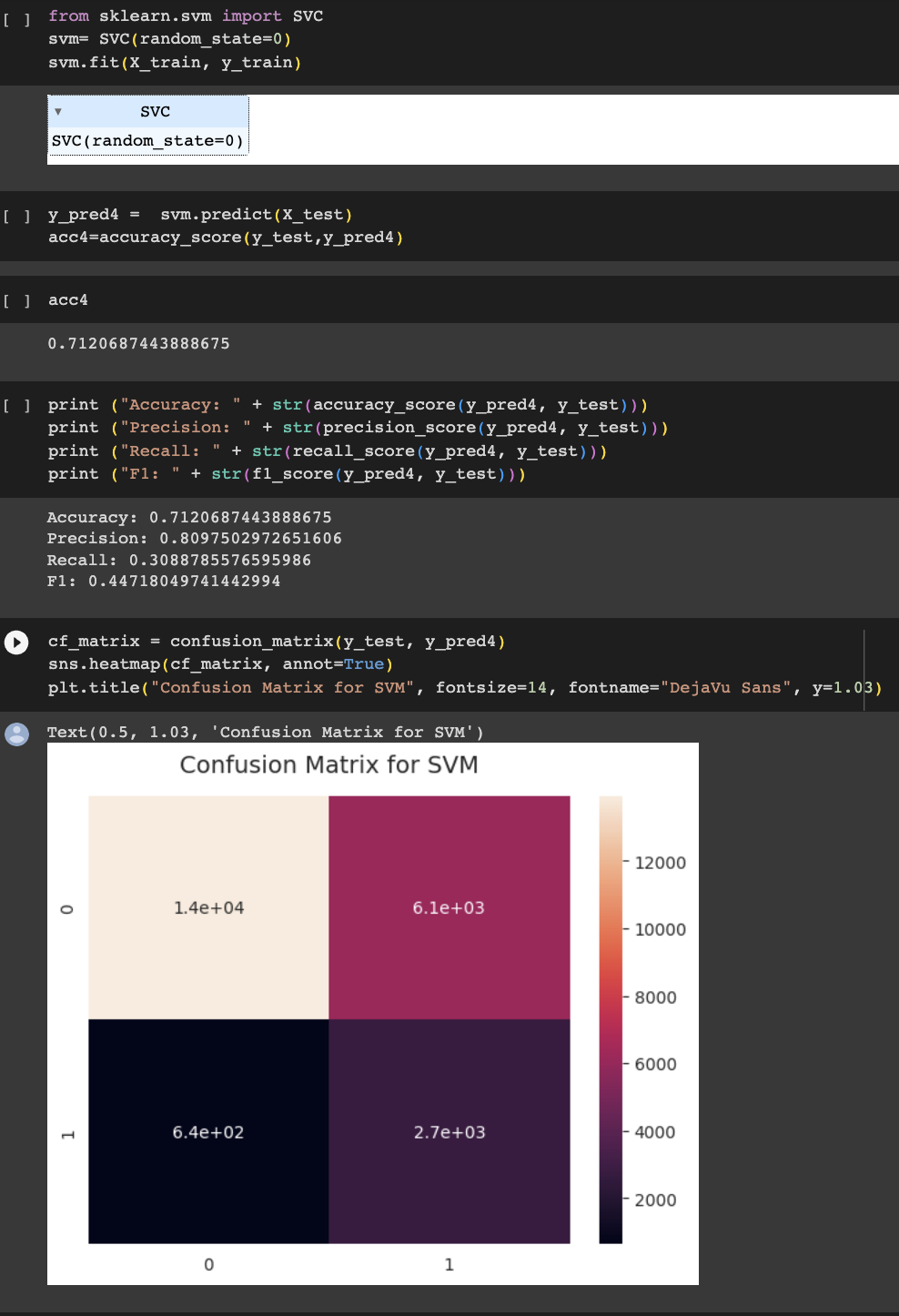
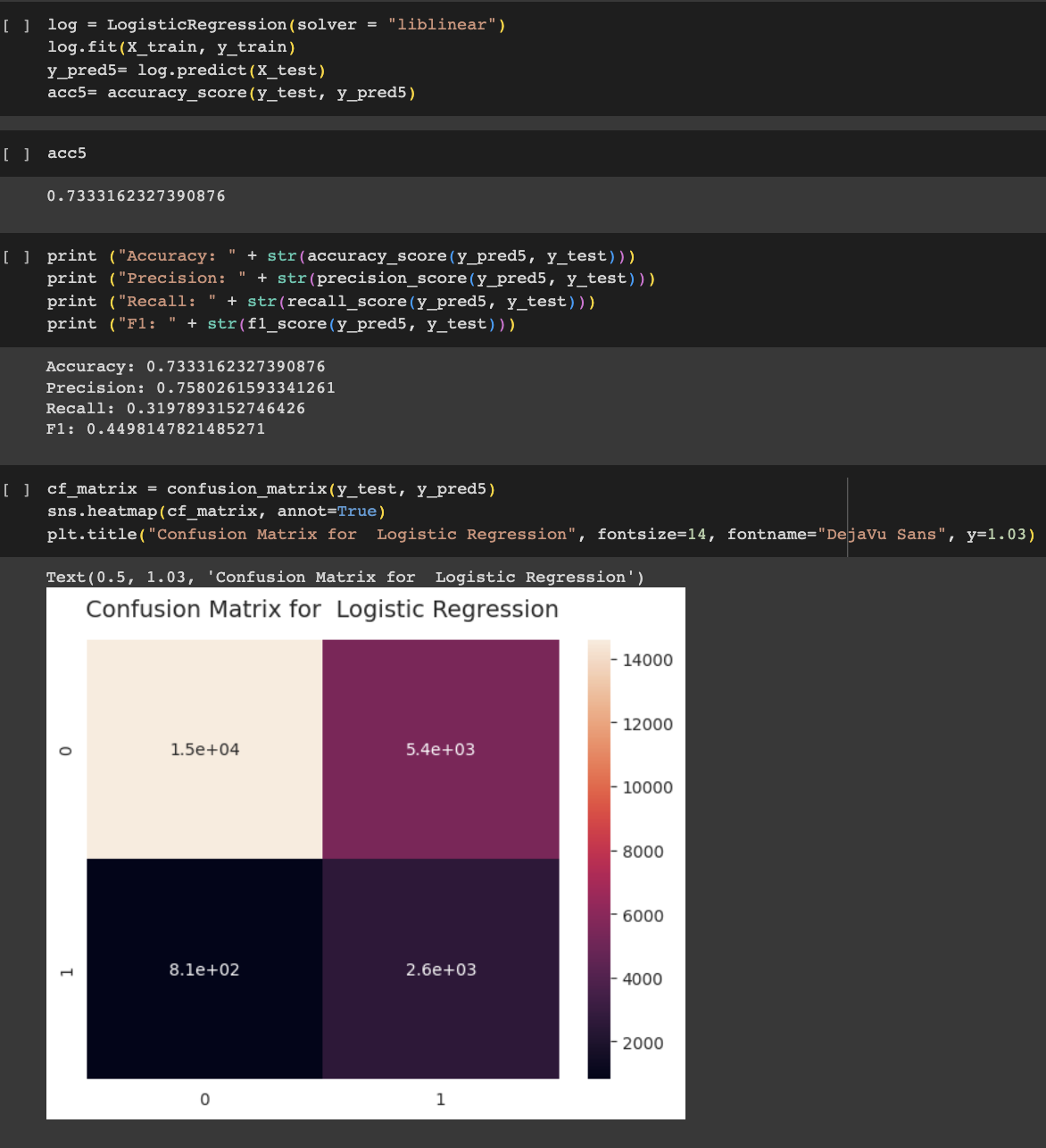


Figure 36:SVM – Modelling and evaluation

The Support Vector Machine is imported from sklearn library and the classic model is trained using the train data and tested for predictions. The accuracy of the model is calculated and printed. Then confusion matrix is plotted using Seaborn by comparing the predicted labels y\_pred and actual labels y\_test. The True Positives, False Positives, False Negatives and True Negatives are 1.4e+04, 6.1e+03, 6.4e+02 and 2.7e+03 respectively.

The values of predicted and actual labels were used to evaluate the accuracy, precision, recall and F1 score of the models and the results for the model was accuracy was 71.20 percent, precision equals 80.97, recall equals 30.89 percent and F1 score equals 44.72.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| SVM | 71.20 | 80.97 | 30.89 | 44.72 |



*Figure 37*:*Logistic Regression – Modelling and evaluation*

The Logistic Regression was already imported from sklearn library, but in this step, it is initialised and then this classic model is trained using the train data and tested for predictions. The accuracy of the model is calculated and printed. Then confusion matrix is plotted using Seaborn by comparing the predicted labels y\_pred and actual labels y\_test. The True Positives, False Positives, False Negatives and True Negatives are 1.5e+04, 5.4e+03, 8.1e+02 and 2.6e+03 respectively.

The values of predicted and actual labels were used to evaluate the accuracy, precision, recall and F1 score of the models and the results for the model was accuracy was 71.20 percent, precision equals 80.97, recall equals 30.89 percent and F1 score equals 44.72.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| Logistic Regression | 73.33 | 75.80 | 31.98 | 44.98 |

# CHAPTER 5 – CONCLUSION AND RECOMMENDATIONS

# 5.1. Conclusion

Diabetes is a very serious and chronic non communicable disease which affects the living situations of the patients and families drastically. The aim of this study was to analyse and forecast diabetic patient diagnoses using the best machine modelling approaches with an increased accuracy, efficiency, and timeliness. The research studied dataset containing data that can be collected with ease such as subject demographics, social and economic status, health status, disease status and habits and doesn’t involve clinical data. The research identified ensemble and classical machine learning models such as AdaBoost, CatBoost, LR, DT and SVM for modelling and evaluation. The data collected from the Kaggle for modelling. The raw data is pre-processed and prepared for training and testing. The models were trained, validated, tested and evaluated for diabetes prediction and the results of the performance evaluation are as given below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| AdaBoost | 73.38 | 76.28 | 32.10 | 45.18 |
| CatBoost | 72.20 | 79.61 | 31.53 | 45.17 |
| Decision Tree | 73.60 | 92.33 | 34.43 | 50.16 |
| SVM | 71.20 | 80.97 | 30.89 | 44.72 |
| Logistic Regression | 73.33 | 75.80 | 31.98 | 44.98 |

The results showed that the Decision Tree model outperformed all other models in all evaluation metrics and shown a precision of 92.33 percent. Hence, we can say that the ML models can be successfully used for prediction of diabetes. During the analysis, various important features that are correlated to diabetes were identified such as general health of the subject, high blood pressure, high cholesterol, BMI, age, difficulty in walking or climbing stairs, income, physical health status, heart disorder or heart attack status, education, physical activity, stroke, cholesterol check status and heavy alcohol consumption habit. These models and features can be therefore used for prediction systems and can be successfully integrated in the diagnosis process to reduce the number of tests conducted and hence the costs.

# 5.2. Recommendations

The study mainly used subject demographic, habits, income, health and disease status dataset for the research. This will help in reducing the number of tests conducted and hence the healthcare expenditure. But further researches can be conducted by integration of clinical data that are available as part of the patient history for better prediction and prevention of the disease. This models for prediction can also be integrated as part of the primary health services offered by government, which can offer better chances of prevention of diabetes among people.

# References

Aguilera-Venegas, G., López-Molina, A., Rojo-Martínez, G. & Galán-García, J. L., 2023. Comparing and tuning machine learning algorithms to predict type 2 diabetes mellitus. *Journal of Computational and Applied Mathematics,* Volume 427.

Ahmed, U. et al., 2022. Prediction of Diabetes Empowered With Fused Machine Learning. *IEEE Access,* Volume 10, pp. 8529-8538.

Centers for Disease Control and Prevention, 2020. *National Diabetes Statistics Report 2020. Estimates of diabetes and its burden in the United States.,* United States: Centers for Disease Control and Prevention.

Hasan, M. K. et al., 2020. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. *IEEE Access,* Volume 8, pp. 76516-76531.

Hassan, M. M. et al., 2021. Diabetes Prediction in Healthcare at Early Stage Using Machine Learning Approach. *IEEE Access,* pp. 1-5.

International Diabetes Federation, 2021. *IDF Diabetes Atlas 10th edition,* Brussels: International Diabetes Federation.

Kandhasamy, J. P. & Balamurali, S., 2015 . Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science,* Volume 47 , p. 45 – 51.

Khaleel, F. A. & Al-Bakry, A. M., 2023. Diagnosis of diabetes using machine learning algorithms. *Materials Today: Proceedings,* 80(3), pp. 3200-3203.

Lama, L. et al., 2021. Machine learning for prediction of diabetes risk in middle-aged Swedish people. *Heliyon,* Volume 7.

Malik, S., Harous, S. & El-Sayed, H., 2020. Comparative Analysis of Machine Learning Algorithms for Early Prediction of Diabetes Mellitus in Women. *Modelling and Implementation of Complex Systems,* Volume 156, pp. 95-106.

Nai-aruna, N. & Moungmai, R., 2015. Comparison of Classifiers for the Risk of Diabetes Prediction. *Procedia Computer Science,* Volume 69, p. 132 – 142.

Naz, H. & Ahuja, S., 2020. Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders,* Volume 19, p. 391–403.

Olisah, C. C., Smith, L. & Smith, M., 2022. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine,* Volume 220, pp. 1-12.

Pandey, P. & Pandey, M., 2015. *Research Methodology: Tools & Techniques.* 1 ed. s.l.:Bridge Center.

Perveena, S., Shahbaza, M., Guergachib, A. & Keshavjee, K., 2016 . Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. *Procedia Computer Science,* Volume 82 , p. 115 – 121.

Posonia, A. M., Vigneshwari, S. & Rani, D. J., 2020. Machine Learning based Diabetes Prediction using Decision Tree J48. *IEEE Access,* pp. 498-502.

Sisodia, D. & Sisodia, D. S., 2018. Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science,* Volume 132 , p. 1578–1585.

Tigga, N. P. & Garg, S., 2020. Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Computer Science,* Volume 167, p. 706–716.

Vijayan, V. V. & Anjali, C., 2016. Prediction and diagnosis of diabetes mellitus — A machine learning approach. *IEEE Xplore,* pp. 122-127.

Warke, M. et al., 2019. Diabetes Diagnosis using Machine Learning Algorithms. *International Research Journal of Engineering and Technology (IRJET),* 6(3), pp. 1470-1476.

WHO, 2023. *Diabetes.* [Online]   
Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes  
[Accessed 20 August 2023].