Minor Project Final Report on

“Diabetes Prediction Using ML”

****

**[SUB CODE :CMP 390]**

in partial fulfillment for the award of the degree of

Bachelor of Information Technology Engineering

**SUBMITTED BY**

**Dina Khadka [20120018]**

**Dipesh Kr. Chaudhary [20120019]**

**Nabeena Adhikari [20120036]**

**Saheena Khadka [20120048]**

**OCTOBER,2023**

**SUBMITTED TO**

**Department of Computer and IT Engineering**

**Everest Engineering College**

**Sanepa, Lalitpur**

**COPYRIGHT**

The author has agreed that the library, Everest Engineering Collage, may make this report freely available for inspection. Moreover, the author has agreed that permission for extensive copying of this project report for scholarly purpose may be granted by the lecturers, who supervised the project works recorded here in or, in their absence, by the Head of Department wherein the project report was done. It is understood that the recognition will be given to the author of the report and to the Department of Computer and Electronics, EEC in any use of the material of this project report. Copying or publication or other use of this report for financial gain without approval of the Department and author’s written permission is prohibited. Request for permission to copy or to make any other use of the material in this report in whole or in part should be addressed to:

Head

Department of IT, Computer and Electronics Engineering

Everest Engineering College

Sanepa, Kathmandu

Nepal

**CERTIFICATE**

The undersigned certify that they have read and recommended to the Department of Electronics and Computer Engineering for acceptance, a project report entitled “Diabetese Prediction Using ML”, submitted by Dipesh kr. Chaudhary, Dina Khadka, Nabeena Adhikari and Saheena Khadka in partial fulfillment of the requirement for the Bachelor’s degree in Information Technology.

**Supervisor:**

...................................

Shailesh Pandey

Lecturer

**External Examiner:**

...................................

[Name of External Examiner]

[Designation]

**ACCEPTANCE**

The project report entitled “Diabetese Prediction Using ML”, submitted by Dipesh kr. Chaudhary, Dina Khadka, Nabeena Adhikari and Saheena Khadka in partial fulfillment of the requirement for the Bachelor’s degree in Information Technology has been accepted as a bona fide record of work independently carried out by the group in the department.

……………………………..

Er. Birodh Rijal

**Principal**

Everest Engineering and Management College,

Kathmandu, Nepal.

……………………………..

Anuj Ghimire

**Head of the Department**

Department of Electronics and Computer Engineering, Everest Engineering and Management College, Kathmandu, Nepal.

# ACKNOWLEDGEMENT

We would like to express our gratitude and appreciations to all those who gave us the possibility to complete this report. Special thanks are due, to our supervisor Mr. Sailesh Pandey sir, for providing us with sound counsel and assistance at critical times and for leading us in the correct direction.

We’d also like to use this opportunity to thank everyone in the Department of Engineering of Everest Engineering College for letting and helping us throughout the completion of the project. We’d want to express our gratitude to our friends and family for their support and encouragement during our work

# ABSTRACT

The following report presents a proposal for a research project on "Diabetes Prediction Using Machine Learning." The main objective of this paper is to develop prediction modelling of the given medical data of patients with and without diabetes. Through this project, we aim to create hybrid models or types that can be easily used by doctors to treat patients with diabetes. In this method, we will use algorithms such as Naïve Bayes, k-NN to predict the chances of a random person with having diabetes, keeping his health conditions in mind. This process thus, enables doctors to easily group, classify and categorize the disease type and accordingly, treatment can be given to them.

*Keywords: Classification, Diabetes, Machine Learning, Naïve Bayes, KNN Prediction.*

Contents

[ACKNOWLEDGEMENT v](#_Toc147497993)

[ABSTRACT i](#_Toc147497994)

[LIST OF FIGURES iv](#_Toc147497995)

[LIST OF TABLES v](#_Toc147497996)

[Chapter 1: INTRODUCTION 1](#_Toc147497997)

[1.1 Background 1](#_Toc147497998)

[1.2 Statement Of Problem 2](#_Toc147497999)

[1.3 Objectives 2](#_Toc147498000)

[1.4 Scope 2](#_Toc147498001)

[1.5 Applications 3](#_Toc147498002)

[1.6 Hardware and Software Requirements 4](#_Toc147498003)

[Chapter 2 : LITERATURE REVIEW 5](#_Toc147498004)

[Chapter 3 : METHODOLOGY 12](#_Toc147498005)

[3.1 Introduction 12](#_Toc147498006)

[3.2 Block Diagrams 12](#_Toc147498007)

[3.1.2 WORK FLOW 13](#_Toc147498008)

[3.1.3 USE CASE DIAGRAM 13](#_Toc147498009)

[3.3 ALGORITHMS 14](#_Toc147498010)

[3.3.1 K-NEAREST NEIGHBOUR 14](#_Toc147498011)

[3.3.2 Naïve Baye’s 14](#_Toc147498012)

[3.4 IMPLEMENTATION PLAN 15](#_Toc147498013)

[3.4.1 Data Exploration And Cleaing 15](#_Toc147498014)

[3.4.2 Data Pre-processing And Analysis: 15](#_Toc147498015)

[3.4.2 Model Selection and Evaluation : 16](#_Toc147498016)

[3.4.4 Model Building : 16](#_Toc147498017)

[3.4.5 Model Deployment And Prediction : 16](#_Toc147498018)

[3.4.6 Documentation and Reporting 16](#_Toc147498019)

[Chapter 4 : RESULTS AND ANALYSIS 18](#_Toc147498020)

[4.1 Result and analysis 18](#_Toc147498021)

[4.1.1 Data collection/exploration and cleaning 18](#_Toc147498022)

[4.1.2 Preprocessing and EDA 18](#_Toc147498023)

[4.1.3 Model Evaluation And Selection 20](#_Toc147498024)

[4.1.4 Model Building From Scratch 21](#_Toc147498025)

[4.1.5 Model Deployment 21](#_Toc147498026)

[4.2 Limitations 21](#_Toc147498027)

[4.3 Work Schedule 22](#_Toc147498028)

[Chapter 5 : Conclusion and Future Enhancement 23](#_Toc147498029)

[5.1 Conclusion 23](#_Toc147498030)

[5.2 Future Enhancement 23](#_Toc147498031)

[REFERENCES 24](#_Toc147498032)

[APPENDICES 25](#_Toc147498033)

[Data collection/exploration and cleaning 25](#_Toc147498034)

[Selected relevant features for diabetic prediction: 25](#_Toc147498035)

[Data Preprocessing and EDA 29](#_Toc147498036)

[Dataframe information and Handling Duplicates: 29](#_Toc147498037)

[Checking Outliers: 30](#_Toc147498038)

[Correlation Between features: 31](#_Toc147498039)

[Relation Between Features and Diabetes Binary 31](#_Toc147498040)

[Relation between different invidual features and Diabetese Binary 32](#_Toc147498041)

[Model Evaluation and Selection 35](#_Toc147498042)

[Model Evaluation with cross validation and without scaled testing inputs i.e. unscaled validation and testing sets : 35](#_Toc147498043)

[Model Evaluation with cross validation and without scaled testing inputs i.e. unscaled validation and testing sets : 35](#_Toc147498044)

[Relation between feature’s different values v/s correct prediction and incorrect prediction: 37](#_Toc147498045)

[Model Building from scratch 38](#_Toc147498046)

[Deployed Model 38](#_Toc147498047)

# LIST OF FIGURES

[Figure 2.1:Performance of Predictive Models , on BRFSS, 2014 [Source : [2] ] 7](#_Toc147498048)

[Figure 2.2: Association Between Covariates,BRFSS2014[Source : [2]] 9](#_Toc147498049)

[Figure 2.3: AUC-ROC behavior: inductive learning.[Source: [4]] 10](#_Toc147498050)

[Figure 3.1: Proposed system architecture 12](#_Toc147498051)

[Figure 3.2 : Work flow of the model[Source: [7]] 13](#_Toc147498052)

[Figure 3.3: Use case diagram 14](#_Toc147498053)

[Figure 0.1: Dataframe information 29](#_Toc147498054)

[Figure 0.2: Handling duplicates 30](#_Toc147498055)

[Figure 0.3:Checking outliers 30](#_Toc147498056)

[Figure 0.4: Correlation between features 31](#_Toc147498057)

[Figure 0.5: Relation between Features and Diabetes Binary 31](#_Toc147498058)

[Figure 0.6: Age vs Diabetese Binary 32](#_Toc147498059)

[Figure 0.7: RoutineCheckupSpan vs DiabetesBinary 32](#_Toc147498060)

[Figure 0.8 : HeartDiseaseAttack vs DiabetesBinary 32](#_Toc147498061)

[Figure 0.9: BMI vs DiabeteseBinary 33](#_Toc147498062)

[Figure 0.10 : Genhlth vs DiabeteseBinary 33](#_Toc147498063)

[Figure 0.11: Physhlth vs Diabetes Binary 33](#_Toc147498064)

[Figure 0.12: PhysActivity vs DiabeteseBinary 34](#_Toc147498065)

[Figure 0.13: HvyAlcoholConsump vs DiabetesBinary 34](#_Toc147498066)

[Figure 0.14: Income vs Diabetes Binary 34](#_Toc147498067)

[Figure 0.15:Model Evaluation without scaled inputs 35](#_Toc147498068)

[Figure 0.16:Model Evaluation with scaled inputs 36](#_Toc147498069)

[Figure 0.17: Value vs Correct vs Incorrect prediction 37](#_Toc147498070)

[Figure 0.18: Accuracy of built model 38](#_Toc147498071)

[Figure 0.19:User Interface 39](#_Toc147498072)

# LIST OF TABLES

[Table 4.1:Work Schedule 22](#_Toc147498073)

[Table 0.1: Selected relevant features for diabetic prediction 29](#_Toc147498074)

#### ABBREVIATIONS

ML :MACHINE LEARNING

k-NN :K NEAREST NEIGHBOUR

DM : Diabetes Milletius

SVM : Support Vector Machine

GNB -> GAUSSIAN NAIVE BAYE’S

CDC -> CENTRES FOR DISEASE CONTROL AND PREVENTION

BRFSS -> BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM

# Chapter 1: INTRODUCTION

## 1.1 Background

Diabetes is a chronic metabolic disorder characterized by high blood sugar levels, resulting from insufficient insulin production or impaired insulin function.

**Types of Diabetes :-**

**Type 1** diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

**Type 2** diabetes means that the cells produce a low quantity of insulin or the body can’t use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.

Since, of those 3, type 2 diabetes is the most prevalent and accounts for 90% to 95% of all cases and unlike others it is a predictable and preventable disease because it usually develops later in life (age >30) as a result of lifestyle (eg, low physical activity, obesity status) and other (eg, age, sex, race, family history) risk factors , our main focus will be in type 2 diabetes.

Early detection and timely intervention are crucial for effective diabetes management and prevention of complications

Machine learning, a branch of artificial intelligence, has shown promise in analyzing large datasets and identifying patterns to predict diabetes risk in individuals.

## 1.2 Statement Of Problem

According to an article of WHO [1] , about 422 million people worldwide have diabetes, the majority living in low-and middle-income countries, and 1.5 million deaths are directly attributed to diabetes each year. Both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades

To identify the risk factors for a variety of human diseases(mainly chronical diseases), in 1984 CDC initiated the state-wide Behavioral Risk Factor Surveillance System (BRFSS), an ongoing, state-based, random-digit–dialed telephone survey of noninstitutionalized US adults aged 18 years or older.

Since, original dataset contains behavioral risk factors of all chronical disease rather of diabetic only and many unwanted datas, assistance of ‘Relevant Research Paper using BRFSS for Diabetes ML’ [2] was taken to extract relevant features causing diabetes. The goal of our study was to build predictive models for type 2 diabetes using 2022 BRFSS data by applying machine learning techniques, including K Nearest Nesighbour (k-NN) and Gaussian Naive Bayes classifiers.

## 1.3 Objectives

The major work we will be doing under this project is listed below:

* prediction of diabetes susceptibility

The early diagnosis and prognosis of a diabetes make it more likely to respond to effective treatment and can result in a greater probability of surviving, less morbidity, and less expensive treatment. Significant improvements can be made in the lives of diabetic patients by detecting diabetes early and avoiding delays in care.

## 1.4 Scope

The scope of this project “DIABETES PREDICTION USING ML” encompasses following things :

* **Data Sources:** It leverages diverse sections of datas that are risk factors for diabetes such as: Chronic Health Conditions, Pre-diabetes, Physical /mental /economical conditions, Demographics(age,sex,education,etc.)
* **Geographic Applicability:** This project’s output can be used for global use and can adapt to different healthcare settings, irrespective of geographic location.
* **Healthcare Domains:** This project can be used as an reference for various healthcare domains, including diagnostics and researches.
* **Limitations :** This project is based on BRFSS 2022 data by CDC, which is basically for collecting datas of risk factors of chronical diseases like diabetes ,heart disease of US. So, it’s not necessary that it had all the risk factors for diabetes included ,Though this project had tried it’s best to extract most relevant risk factors among available in the datasets.
* **Future Expansion:** This project is scalable, as it can be make more efficient as the emergence of new technologies or new datasets.

## 1.5 Applications

The applications of machine learning in diabetes prediction include:

a. **Clinical Decision Support:** Machine learning algorithms can assist healthcare professionals in making more informed decisions by providing risk assessments and predictions to aid in the diagnosis and treatment planning process.

b. **Remote Monitoring and Telemedicine:** Implementing machine learning algorithms in mobile apps or online platforms enables individuals to conveniently assess their diabetes risk remotely. This can promote proactive health management and encourage timely medical interventions.

c. **Public Health Initiatives:** Population-level screening using machine learning algorithms can help public health organizations identify high-risk groups and implement targeted interventions, such as lifestyle modification programs and awareness campaigns.

d. **Research and Development:** Machine learning techniques can be applied to large-scale health datasets to identify novel risk factors, uncover hidden patterns, and contribute to the ongoing research and understanding of diabetes etiology and prevention.

## 1.6 Hardware and Software Requirements

Certainly, below is a table listing the hardware and software required to build this project.

|  |  |  |
| --- | --- | --- |
| **Hardware Requirement** | **Software Requirement** | **Version** |
| Personal Computer or Laptop | Python | 3.10 |
|  | Google Colab (for cloud-based development) | Latest version |
|  | Anaconda (for managing Python environments) | Latest version |
|  | Spyder (IDE for Python development) | Latest version |
|  | Streamlit (for web application development) | Latest version |
|  | Libraries and Frameworks | Various (e.g., Scikit-Learn, Mtplotlib,Numpy, Pandas) |
| Monitor, Keyboard, and Mouse | Operating System (e.g., Windows, macOS, Linux) | Latest supported versions |

# : LITERATURE REVIEW

1. This report [3] of CDC is a national diabetes statistics report of between 2015-2018 for analysing risk factors for diabetes-related complications and their frequencies. They are as:-

* Smoking:
* 19.8% were tobacco users based on self-report or levels of serum cotinine.
* 13.8% reported current cigarette smoking.
* 37.1% had quit smoking but had a history of smoking at least 100 cigarettes in their lifetime.
* Overweight and Obesity :
* 89.8% were overweight or had obesity, defined as a body mass index (BMI) of 25 kg/m2 or higher. Specifically:
  + 27.7% were overweight (BMI of 25.0 to 29.9 kg/m2).
  + 45.8% had obesity (BMI of 30.0 to 39.9 kg/m2).
  + 16.2% had extreme obesity (BMI of 40.0 kg/m2 or higher).
* Physical Inactivity :
* 34.3% were physically inactive, defined as getting less than 10 minutes a week of moderate or vigorous activity in each physical activity category of work, leisure time, and transportation.
* A1C :
* 49.4% had an A1C value of 7.0% or higher. Specifically:
  + 24.9% had an A1C value of 7.0% to 7.9%.
  + 11.4% had an A1C value of 8.0% to 9.0%.
  + 13.2% had an A1C value higher than 9.0%.
* 10.4% of adults aged 18–44 years had A1C levels of 10% or higher, compared to 9.4% of those aged 45–64 years and 2.6% of those aged 65 years or older
* High Blood Pressure :
* 69.0% had a systolic blood pressure of 140 mmHg or higher or diastolic blood pressure of 90 mmHg or higher or were on prescription medication for their high blood pressure.
* High Cholesterol:
* 44.3% had a non-HDL level of 130 mg/dL or higher. Specifically:
  + 22.1% had a non-HDL level of 130 to 159 mg/dL.
  + 12.5% had a non-HDL level of 160 to 189 mg/dL.
  + 9.7% had a non-HDL level of 190 mg/dL or higher.

1. The authors of this journal [2] analyzed cross-sectional data on 138,146 participants, including 20,467 with type 2 diabetes, from the 2014 Behavioral Risk Factor Surveillance System.They built several machine learning models for predicting type 2 diabetes, including support vector machine, decision tree, logistic regression, random forest, neural network, and Gaussian Naive Bayes classifiers. They used univariable and multivariable weighted logistic regression models to investigate the associations of potential risk factors with type 2 diabetes..

**Predictive models for type 2 diabetes :**

All classifiers had a high test accuracy (74.3%–82.4%) and high AUC values (0.7182–0.7949) *Figure 2.1*. Although the neural network model gave the highest accuracy (82.4%), specificity (90.2%), and AUC (0.7949) values, its sensitivity (37.8%) was the lowest. In contrast, although the decision tree model had the lowest accuracy (74.3%), specificity (78.2%), and AUC (0.7182) values, its sensitivity (51.6%) was the highest. Other classifiers gave intermediate and reasonable accuracy, sensitivity, specificity, and AUC values. Overall, the predictive models for type 2 diabetes had similar and good prediction performance with only slight differences.

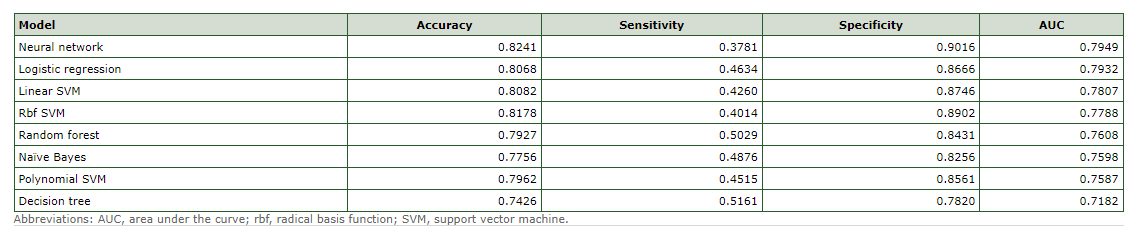


Figure 2.1:Performance of Predictive Models , on BRFSS, 2014 [Source : [2] ]

Of the 8 predictive models, the neural network model gave the best model performance with the highest AUC value; however, the decision tree model is preferred for initial screening for type 2 diabetes because it had the highest sensitivity and, therefore, detection rate. They confirmed previously reported risk factors and also identified sleeping time and frequency of checkup as 2 new potential risk factors related to type 2 diabetes.

**Risk factors affecting type 2 diabetes :**

Unadjusted and adjusted ORs from univariable and multivariable weighted logistic regression models are summarized in *Figure 2.2*. Compared with women, men had a significantly higher risk of type 2 diabetes (aOR = 1.38; 95% CI, 1.29–1.48). Risk of developing type 2 diabetes increased as age and BMI increased, and it decreased as income increased. Compared with married respondents, all other groups had similar risk of developing type 2 diabetes, with the exception of divorced respondents for whom risk of type 2 diabetes was lower. Respondents of all races/ethnicities except Native Hawaiian/other Pacific Islanders had a significantly higher risk of type 2 diabetes than white respondents, and Asians had the highest risk after adjusting for other variables.

Respondents who slept 6 or fewer, or 9 or more hours per day had a higher unadjusted odds of type 2 diabetes than respondents who slept from 7 to 8 hours per day, but the significance remained only for those who slept 9 or more hours per day after adjusting for all other variables (aOR = 1.13; 95% CI, 1.03–1.25). Compared with respondents whose last checkup was within the last 3 to 5 years, those whose last checkup was less than 1 year ago or who had never had a checkup had higher unadjusted odds of developing type 2 diabetes. After adjusting for all other variables, odds of developing type 2 diabetes were significantly higher only for those whose last checkup was less than 1 year ago (aOR = 2.31; 95% CI, 1.86–2.85). Although the adjusted odds of never having had a checkup were 1.43, they were not significant.

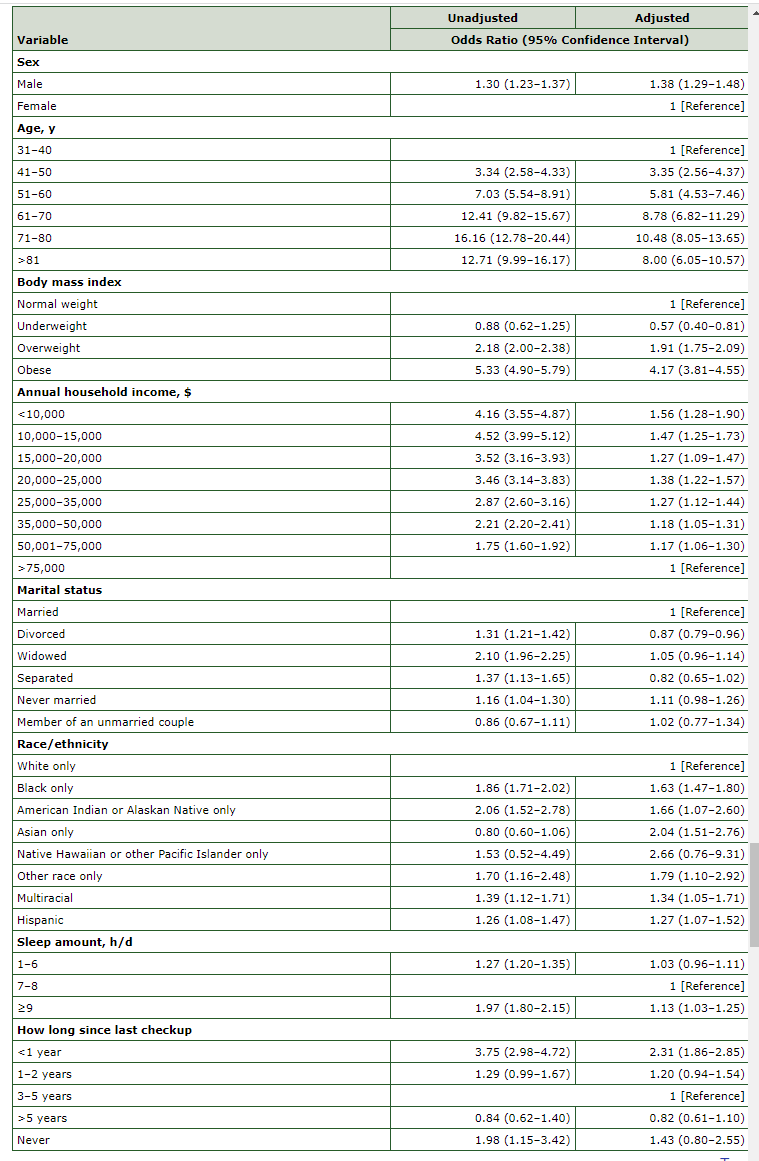


Figure 2.2: Association Between Covariates,BRFSS2014[Source : [2]]

1. The authors of this paper [4] proposed a worker-centric, IoT-enabled framework for monitoring the health, well-being, and functional ability of individuals using AI tools. The focus is on predicting the risk of type 2 diabetes by applying different components of the Knowledge Discovery in Database (KDD) process, including dataset creation, feature selection, and classification using supervised machine learning models. The proposed ensemble model, called WeightedVotingLRRFs, is used to improve the prediction of diabetes risk.

The research compares the performance of the WeightedVotingLRRFs model with other models and risk scoring systems such as the Finnish Diabetes Risk Score (FINDRISC) and Leicester risk score. The evaluation is conducted using data extracted from the English Longitudinal Study of Ageing (ELSA) database.

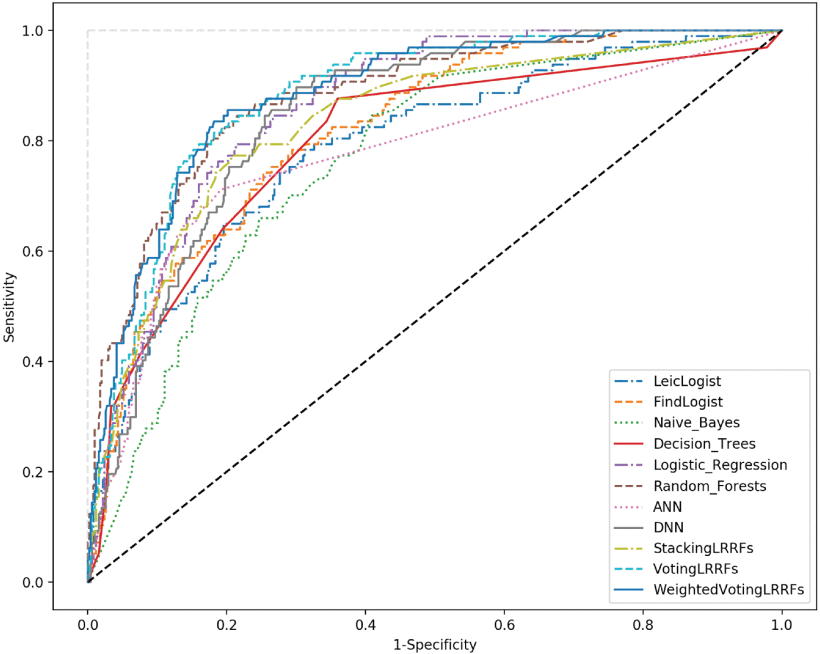


Figure 2.3: AUC-ROC behavior: inductive learning.[Source: [4]]

Figure 3 summarizes the performance metrics values for the diabetes prediction according to the adopted ML models.

1. A survey for detecting and predicting of diabetes using machine learning techniques [5]. This paper focuses on machine learning promising the improving accuracy of perception and diagnosis of the diseases. The various machine learning techniques that are used to classify the data sets include supervised, unsupervised, reinforcement, semi-supervised, and deep learning, evolutionary learning algorithms. It also shows the comparison of the two methods namely, Naïve Bayes and Artificial Neural Networks (ANN). The Bayesian Network applies the Naïve Bayes theorem which firmly assumes that the presence of any attribute in a class is not related to the presence of any other attribute, making it much more advantageous, efficient and independent.
2. Yasodhaet al. [6]uses the classification on diverse types of datasets that can be accomplished to decide if a person is diabetic or not. The diabetic patient’s data set is established by gathering data from hospital warehouse which contains two hundred instances with nine attributes. These instances of this dataset are referring to two groups i.e. blood tests and urine tests. In this study the implementation can be done by using WEKA to classify the data and the data is assessed by means of 10-fold cross validation approach, as it performs very well on small datasets, and the outcomes are compared. The naïve Bayes, J48, REP Tree and Random Tree are used. It was concluded that J48 works best showing an accuracy of 60.2% among others

# : METHODOLOGY

## 3.1 Introduction

This chapter explains the details about the method that is being used while developing our project. The project method is just a fancy way of describing a system used to do something. It helps to organize investigation in a scientific way to overcome problems, extract the truth about the subject to achieve the goal of the project.

## 3.2 Block Diagrams

#### 3.1.1 System Architecture

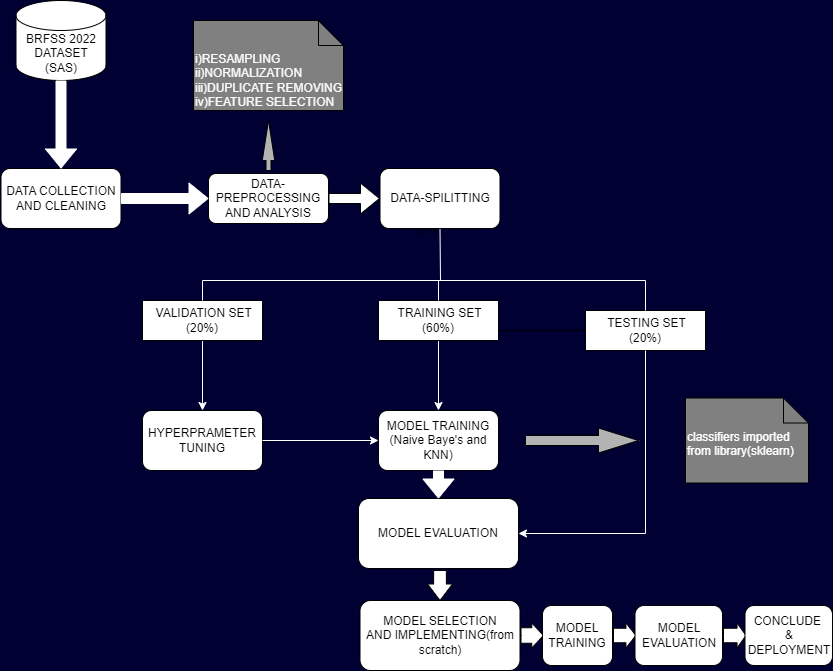


Figure 3.1: Proposed system architecture

### 3.1.2 WORK FLOW

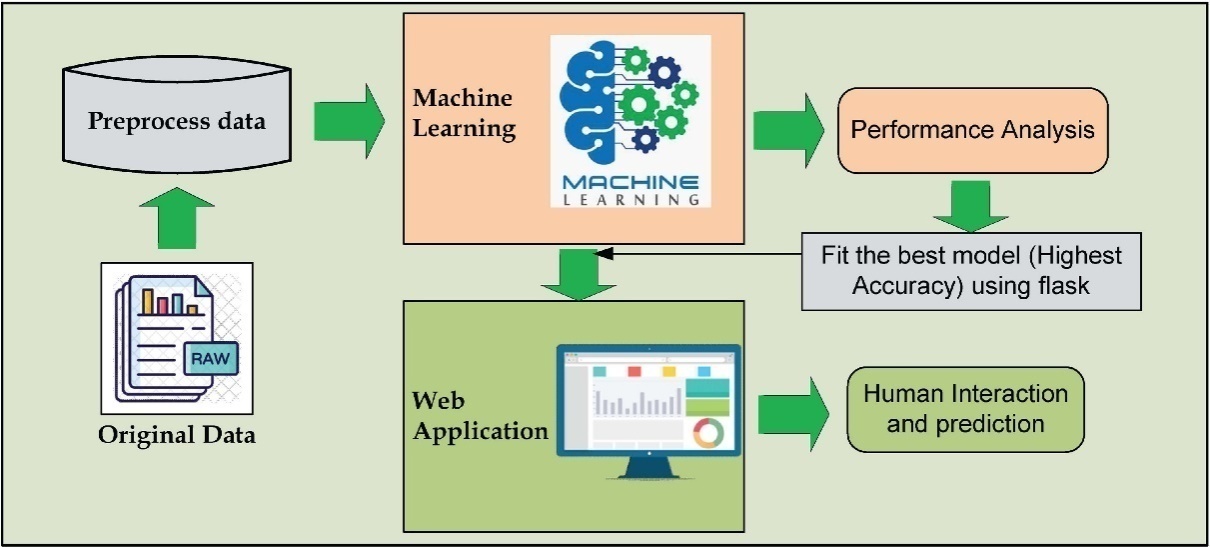


Figure 3.2 : Work flow of the model[Source: [7]]

### 3.1.3 USE CASE DIAGRAM

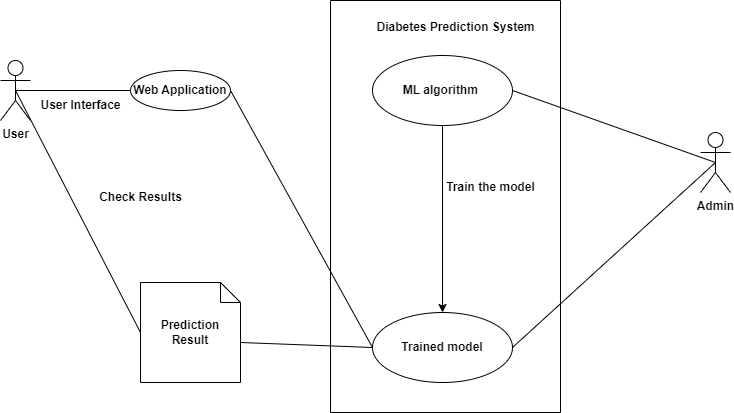


Figure 3.3: Use case diagram

## 3.3 ALGORITHMS

### 3.3.1 K-NEAREST NEIGHBOUR

* KNN is a simple and intuitive algorithm used for both classification and regression tasks.
* It works on the principle of similarity, where the class of a data point is determined by the classes of its k nearest neighbors.
* KNN does not explicitly learn a model from the training data but instead stores the entire dataset in memory.
* To predict the class of a new data point, KNN calculates the distances (e.g., Euclidean distance) between the new point and all the training data points.
* It then selects the k nearest neighbors and assigns the majority class among them as the predicted class for the new point.
* KNN requires defining the value of k, which determines the number of neighbors considered.
* Implementation: In Python, one can use the scikit-learn library to implement KNN for diabetes prediction. The KNeighborsClassifier class can be used, and one can specify the number of neighbors (k) as a parameter.

### 3.3.2 Naïve Baye’s

* Naive Bayes is a probabilistic algorithm based on Bayes' theorem and assumes that the features are conditionally independent given the class variable.
* It is fast, simple, and performs well in many real-world applications, including text classification and medical diagnosis.
* Naive Bayes calculates the probability of each class given the feature values and selects the class with the highest probability as the prediction.
* It estimates the probabilities using the training data's class distributions and feature probabilities.
* There are different variants of Naive Bayes, including Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes, which are suitable for different types of data.
* Gaussian Naive Bayes assumes that the continuous input features follow a Gaussian distribution, while Multinomial and Bernoulli Naive Bayes are designed for discrete or binary features.
* Implementation: The scikit-learn library provides classes such as GaussianNB, MultinomialNB, and BernoulliNB for implementing different variants of Naive Bayes for diabetes prediction. You can use these classes based on the type of features in your dataset.

Each of these classifier algorithms has its strengths, weaknesses, and assumptions. The choice of algorithm depends on factors such as the dataset size, complexity of the problem, interpretability requirements, and available computational resources. Depending on accuracy we will decide best algorithm among both and **will implement from scratch**.

## 3.4 IMPLEMENTATION PLAN

### 3.4.1 Data Exploration And Cleaing

* Explore CDC’s BRFSS annual dataset
* Clean the dataset in ML usable format

### 3.4.2 Data Pre-processing And Analysis:

* Perform data preprocessing steps such as removing duplicates, removing rows with empty cells, resampling, normalization
* Analyse relation between features themselves and, between featares and target variable
* Select features on the basis of relations or using chi square
* Check outlier

### 3.4.2 Model Selection and Evaluation :

* Split data into training ,validation and testing sets
* Use validation set for hyperparameter tuning of both the models (gaussian naïve baye’s and k-NN classifiers)
* Evaluate gaussian naïve baye’s and k-NN tuned classifiers on the basis of accuracy score ,on the testing data
* Evaluate whether testing sets requires scaling or not
* Select the Model with better performance on the dataset

### 3.4.4 Model Building :

* .Learn about the selected model, using their respective sickit learn documentation
* Build an model from scratch using that documentation
* Train the model
* Evaluate the model by using testing set
* Save the trained model for further use

### 3.4.5 Model Deployment And Prediction :

* Develop a user-friendly interface or API using streamlit in spyder
* API should be able to input the necessary health data and provide predictions for diabetes status.
* Prediction is done using that saved model

### 3.4.6 Documentation and Reporting

* Document the entire implementation process, including data preprocessing steps, feature engineering, algorithm selection, and model training.
* Prepare a detailed report summarizing the project, including the problem statement, methodology, results, and conclusions.
* Include visualizations, code snippets, and any other relevant information to support the documentation.

.

# : RESULTS AND ANALYSIS

## 4.1 Result and analysis

In this chapter, we present the results of our project, including the data collection and cleaning process, preprocessing and exploratory data analysis (EDA), model evaluation and selection, and the implementation of a custom machine learning model. We analyze the outcomes and draw insights from each of these phases.

### 4.1.1 Data collection/exploration and cleaning

#### 4.1.1.1 Problems Encountered

* BRFSS,2022 Datasets [8]downloaded from CDC was in ASCII code
* Original dataset contains so many irrelevant features for diabetes prediction
* Dataset contains features name and values that needed to be modified
* Dataset wasn’t in the ML usable format

#### 4.1.1.2 Solution of problem encountered

* Dataset was converted into SAS using Format.sas7bcat code and finally into CSV format using SAS studio for academics
* Out of those 326 features , 20 features [*Table 0.1*] was selected as the risk factors for diabetes using research paper [3]
* Codebook [9]provided by the CDC was taken as help to study about the features.
* Dataset’s Feature variables were renamed in understanding way[*Table 0.1*]
* Data’s values were modified in order to be in ML usable format [*Table 0.1*]

### 4.1.2 Preprocessing and EDA

#### 4.1.2.1 Problems Encountered

* Values were in both float and integers [*Figure 0.1*]
* Duplicates of records were present [*Figure 0.2*]
* BMI,Genhlth,MenHlth,Physhlth had some outliers[*Figure 0.3*]

#### 4.1.2.2 Solutions for problems encountered

* Float values were changed to integer values
* Duplicates records were removed [*Figure 0.2*]
* No need for handling outliers, as they were not because of measurement errors

#### 4.1.2.3 Conclusions derived after EDA

#### 4.1.2.3.1 Correlation between features [*Figure 0.4*]

* GenHlth and PhysHlth are highly correleted with each other.(positive relation) with value 0.49
* GenHlth and Income are highly correleted with each other .(negetive relation) with value -0.3

#### 4.1.2.3.2 About Variables

* Some of them are continuous and Some of them are discrete
* Some of them are categorical and some of them are real numbers

#### 4.1.2.3.3 Relation between Features and Diabetes Binary[*Figure 0.5*]

* Smoker, AnyHealthcare,SleepTime , NoDocbccost,MentHlth and sex are least correlated with Diabetes\_binary.
* BMI , stroke , HeartDiseaseorAttack ,RoutineCheckupSpan, PhysActivity , HvyAlcoholconsump , GenHlth , PhysHlth , Age , Education , Income and DiffWalk have a significant correlation with Diabetes\_binary.

#### 4.1.2.3.4 Relation between individual major features and Diabetes Binary

* When age increases ,risk of diabetes also increases. And the median of the age of diabetic people is greater than that of non-diabetic people.[*Figure 0.6*]
* almost all diabetic patients had last visited a doctor for a routine checkup within a past year.[*Figure 0.7*]
* large amount of diabetic patients were not reported heart disease[*Figure 0.8*]
* people range between 23-38 BMI have more likely to have Diabetic.[*Figure 0.9*]
* percentage of diabetic pateint on a particualar General Health condition is high of poor and fair[*Figure 0.10*]
* We can say that PhysHlth Group 0-5 have impact on Diabetic[*Figure 0.11*]
* Percentage of diabetic patients is higher in "No physical activity class"[*Figure 0.12*]
* People who don't have heavy alcohol consumption are more likely be healthier.But , since there is not enough information here about HvyAlcoholConsump so, we can not predict the effect of HvyAlcoholConsump on Diabetes.[*Figure 0.13*]
* We can conclude that in the high level of income, the healthy people is more than others[*Figure 0.14*]

### 4.1.3 Model Evaluation And Selection

In the model evaluation and selection phase, we compared different machine learning algorithms (k-NN nad Gaussina Naïve Baye’s) to determine the best-performing model for our project. We conducted the following analyses:

#### 4.1.3.4 k-NN vs. Gaussian Naive Bayes

We evaluated the performance of two classification models, k-Nearest Neighbors (k-NN) and Gaussian Naive Bayes (GNB), and selected the best-performing model based on validation accuracy. • [Figure 0.16][Figure 0.15]

* We assessed model performance both in scaled inputs and unscaled inputs
* Model comparison: We compared the accuracy of k-NN and GNB on the validation set.
* Hyperparameter tuning: Hyperparameter tuning was performed to optimize the selected model. Gridseach cv was used for this.
* Test set evaluation: We assessed the final model's performance on the test set and generated a classification report.
* We assessed the impact of values of each features on orrect and incorrect predictions [*Figure 0.17*]

**Results :**

* kNN performed better on our dataset
* Accuracy of model on scaled validation and test set found to be higher than that of unscaled
* We concluded that some particular values of features have equal or greater error than correct prediction [*Figure 0.17*]

**FEATURES**---------------------------------------->**VALUES**

* BMI ---------------------------------------------> 15 to 19
* Stroke -------------------------------------------> 1
* HeartDiseaseorAttack-------------------------> 1
* RoutineCheckupSpan-------------------------> 4
* GenHlth-----------------------------------------> 4 and 5
* PhysHlth----------------------------------------> 10,14,15,20,30
* DiffWalk----------------------------------------> 1
* Age----------------------------------------------> 10 to 13
* Income------------------------------------------> 1 to 4

### 4.1.4 Model Building From Scratch

We implemented a custom k-Nearest Neighbors (k-NN) classifier from scratch, allowing us to gain a deeper understanding of the algorithm.

* Model implementation: We provided an overview of the custom k-NN classifier and its functioning.
* We used same hyperparameter that was selected as best in Model selection.
* Testing and accuracy: The custom k-NN model was tested on the test data, and accuracy was calculated.[*Figure 0.18*]

### 4.1.5 Model Deployment

* An user interface [*Figure 0.19*] was built by using streamlit on spyder ide
* An saved trained model of the ,odel built from scratch was used for prediction
* Streamlit was used deployment

## 4.2 Limitations

“Diabetese prediction using ML” project has the following limitations:

* Users or patients can check whether they are diagnosed with type-2 diabetese or not, just by entering some features values within the range, without any suggestion of medical consultancy.
* It needs the data for different features within range, only then the prediction can be done.
* Users or patients need to be familiarized with the computerized system as prediction is done on computers.
* Users or patients could enter any data within the range without any medical consultant.
* Features are most relevant for predicting daiabetese only among dataset we used. There might be more impactful features which were not present in our dataset.
* Only one patient can be predicted at a time.

## 4.3 Work Schedule

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sn | Tasks | Duration | 2023   |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 1 6 | 7  6 | 14  6 | 23 6 | 30 6 | 7  7 | 14  7 | 23  7 | 30  7 | 7  8 | 14  8 | 30  8 | 8  9 | 16  9 | 24  9 | 30  9 | 3  10 | |
| 1 | Requirement and Gathering Analysis | 4w 5d | |  | | --- | |  | |
| 2 | Research | 6w 4d | |  | | --- | |  | |
| 3 | Design | 5w 2d | |  | | --- | |  | |
| 4 | coding | 9w | |  | | --- | |  | |
| 5 | Teting | 13w 2d | |  | | --- | |  | |
| 6 | Implementation and documantation | 14w 4d | |  | | --- | |  | |

Table 4.1:Work Schedule

# : Conclusion and Future Enhancement

## 5.1 Conclusion

Type-2 diabetes, if detected early, can save the lives of thousands of people. This project aids real-world patients and doctors in gathering as much as information as possible.

The research papers aided us in gathering knowledge on Diabetes in order to analyze the project we proposed. Machine learning algorithms helps to classify and predicts whether the users or patients have Diabetes or not. Machine Learning algorithms can be utilized in medical research because they advance the system, reduce human errors, and manual errors.

## 5.2 Future Enhancement

“Diabetes Prediction Using ML” project has the great scope in medical field in near future in helping more patients to recover from the type-2 diabetes because of the early precautions they would get after the early detection of the diabetese. For the future, we will be featuring the ‘Multiple’ segment, where records of many patients can be passed at once as input by health care professionals making diagnosing them all at once possible.

# REFERENCES

|  |  |
| --- | --- |
| [1] | P. /. A. Loke/WHO, "DIABETES," WHO, [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab\_3. [Accessed 15 06 2023]. |
| [2] | N. O. L. J. L. D. Xie Z, "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques," *Prev Chronic Dis,* p. 16, 2019. |
| [3] | CDC, "Risk Factors for Diabetes-Related Complications," U.S. Department of Health & Human Services, US, 2022. |
| [4] | O. K. E. D. S. A. N. F. a. K. M. N. Fazakis, "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction," 29 july 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9491154. [Accessed 17 06 2023]. |
| [5] | M. Y. R. Mrs. Priyanka Indoria, "INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)," 24 03 2018. [Online]. Available: https://www.ijert.org/a-survey-detection-and-prediction-of-diabetes-using-machine-learning-techniques. [Accessed 10 06 2023]. |
| [6] | M. G. A. M. K. S. Abdullah A. Aljumah, "Application of data mining: Diabetes health care in young and old patients," 2 july 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157812000390#aep-abstract-id17. [Accessed 18 06 2023]. |
| [7] | Nazin Ahmed, "science direct," 7 12 2021. [Online]. Available: www.sciencedirect.com/science/article/pii/S2666307421000279. [Accessed 9 6 2023]. |
| [8] | A. TEBOUL, "Diabetes Health Indicators Dataset from BRFSS 2015," 04 02 2021. [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes\_binary\_health\_indicators\_BRFSS2015.csv. [Accessed 15 06 2023]. |

# 

# APPENDICES

## Data collection/exploration and cleaning

### Selected relevant features for diabetic prediction:

|  |  |  |  |
| --- | --- | --- | --- |
| **‘Variable name in original dataset’**  **: ‘Changed name’** | **Description** | **Original Values** | **Modified values**  **(Original:new)** |
| 'DIABETE4':'DiabetesBinary' | (Ever told) you have diabetes? | 1 :Yes ,  2 :Yes, but female told only during pregnancy  3: No  4: No, pre-diabetes or borderline diabetes  7: Don’t know/Not Sure  9: Refused—Go to Section 08.01 AGE | 2:0, 3:0, 1:1, 4:0  Remove all 7 (dont knows)  Remove all 9 (refused) |
| ,  '\_BMI5':'BMI' | Body Mass Index (BMI) | 1 - 9999 1 or greater Notes: WTKG3/(HTM4\*HTM4) (Has 2 implied decimal places) | no changes, just note that these are BMI \* 100. So for example a BMI of 4018 is really 40.18) |
| 'SMOKE100':'Smoker' | Have you smoked at least 100 cigarettes in your entire life? | 1 : Yes  2 : No  7 : Don’t know/Not Sure  9 : Refused | Change 2 to 0 because it is No  Remove all 7 (dont knows)  Remove all 9 (refused) |
| 'ADDEPEV3':'Depression' | (Ever told) (you had) a depressive disorder (including depression, major depression, dysthymia, or minor depression)? | 1 : Yes  2 : No  7 : Don’t know/Not sure  9 : Refused | change 2 to 0 ,means never told/had depressive order  Remove all 7 (dont knows)  Remove all 9 (refused) |
| 'CVDSTRK3':'Stroke' | (Ever told) (you had) a stroke | 1 : Yes  2 : No  7 : Don’t know/Not sure  9 : Refused | change 2 to 0 ,means never told/had stroke  Remove all 7 (dont knows)  Remove all 9 (refused) |
| '\_MICHD':'HeartDiseaseorAttack' | Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) | 1 : Reported having MI or CHD  2 : Did not report having MI or CHD | change 2 to 0  Remove all 7 (dont knows)  Remove all 9 (refused) |
| 'CHECKUP1':'RoutineCheckupSpan' | About how long has it been since you last visited a doctor for a routine checkup? | 1: Within past year  2: Within past 2 years  3: Within past 5 years  4: 5 or more years ago  7: Don’t know/Not sure  8: Never  9: Refused | not changed  remove all 7,8,9(don't know/not sure,never,refused) |
| 'SLEPTIM1':'SleepTime' | On average, how many hours of sleep do you get in a 24-hour period? | 1 – 24: Number of hours [1-24]    77: don’t know/Not Sure  99 : Refused | 1: 1–6 hours, 2: 7–8 hours, 3: ≥9 hours  Remove all 7 (dont knows)  Remove all 9 (refused) |
| '\_RFDRHV8':'HvyAlcoholConsump' | Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week | 1: No  2 : Yes  9:Don’t know / Refused / Missing | change 1 to 0 for not heavy drinkers and 2 to 1 for heavy drinkers  remove all 9(don't know/refused/missing) |
| '\_HLTHPLN':'AnyHealthcare', | Adults who had some form of health insurance | 1 : Have some form of insurance  2 : Do not have some form of health insurance  9 : Don´t know, refused or missing insurance response | change 2 to 0 for not having any form of health insurance  remove all 9(don't know/refused/missing) |
| 'MEDCOST1':'NoDocbcCost' | Was there a time in the past 12 months when you needed to see a doctor but could not because you could not afford it? | 1 : Yes  2 : No  7 : Don’t know/Not sure  9 : Refused | change 1 to 0 for could not afford medical cost and 2 to 1 for able to afford  remove all 7,9(don't know/not sure, refused) |
| 'GENHLTH':'GenHlth' | Adults with good or better health | 1 : Good or Better Health  2 : Fair or Poor Health  9 : Don’t know/Not Sure Or Refused/Missing | NO change range of(excellent to poor ) --> (1 to 5)  remove all 7,9(don't know,refused) |
| '\_TOTINDA':'PhysActivity' | Adults who reported doing physical activity or exercise during the past 30 days other than their regular job | regular job 1: Had physical activity or exercise, 2: No physical activity in past 30 days | 1 for physical activity  change 2 to 0 for no physical activity  Remove all 9 (don't know/refused) |
| 'SEXVAR':'Sex' | Indicate sex of respondent | 1: Male, 2: Female | change 2 to 0 (female as 0). Male is 1 |
| 'MENTHLTH':'MentHlth', | Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? | 1 – 30: Number of days  88: None  77: Don’t know/Not sure  99: Refused | change 88 to 0 because it means none (no bad mental health days)  remove 77 and 99 for don't know not sure and refused |
| 'PHYSHLTH':'PhysHlth' | for how many days during the past 30 days was your physical health not good? | 1 – 30: Number of days  88: None  77: Don’t know/Not sure  99: Refused | change 88 to 0 because it means none (no bad physical health days)  remove 77 and 99 for don't know not sure and refused |
| 'DIFFWALK':'DiffWalk' | Do you have serious difficulty walking or climbing stairs? | 1:Yes, 2 :No, 7: Don’t know/Not Sure, 9: Refused | change 2 to 0 for no. 1 is already yes  remove 7 and 9 for don't know not sure and refused |
| '\_AGEG5YR':'Age' | Fourteen-level age category | 1 is 18-24 all the way up to 13 is 80 and older. 5 year increments.  14: missing or don’t know | remove 14 because it is don't know or missing |
| 'EDUCA':'Education' | Level of education completed | 1 : Never Attended School'  2: Grades 1 through 8  3 : Grades 9 through 11  4 : Grade 12 or GED'  5 : College 1 year to 3 years  6 : College 4 yr | Remove 9 for refused: |
| 'INCOME3':'Income' | What is your annual household income from all sources? | 1: <$10 K, 2: $10–$15 K, 3: $15–$20 K, 4: $20–$25 K, 5: $25–$35 K, 6: $35–$50 K, 7: $50–$75 K, 8: $75 K-$100 K, 9: $100–$150 K, 10: $150–$200 K, 11: > $200 K, | Remove 77 and 99 for don't know and refused |

Table 0.1: Selected relevant features for diabetic prediction

## Data Preprocessing and EDA

### Dataframe information and Handling Duplicates:

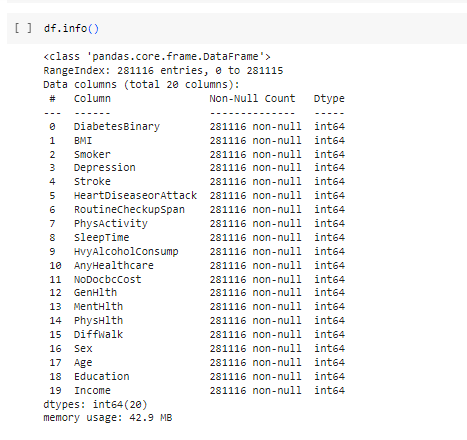


Figure 0.1: Dataframe information

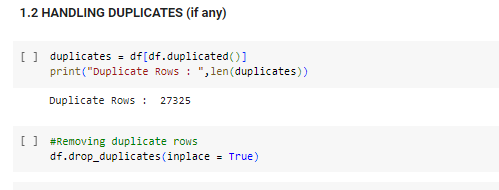


Figure 0.2: Handling duplicates

### Checking Outliers:

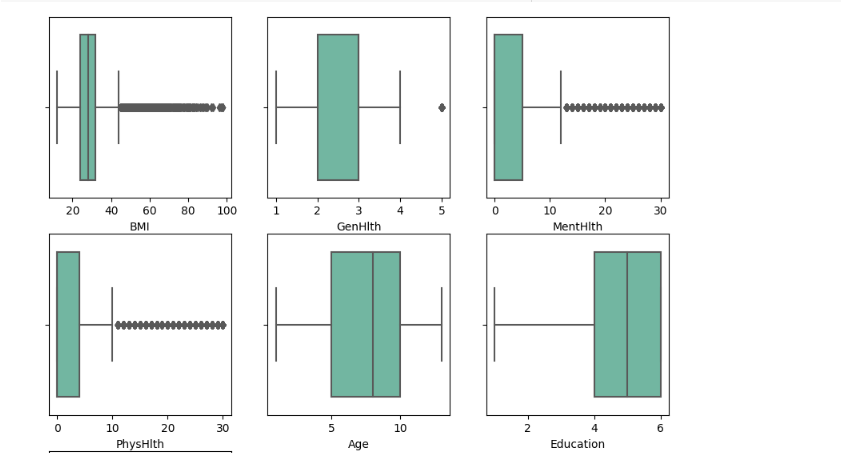


Figure 0.3:Checking outliers

### Correlation Between features:

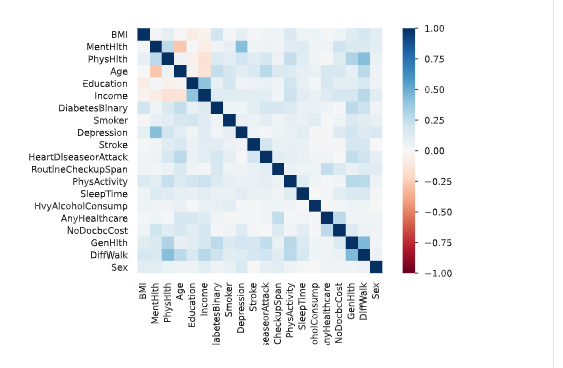


Figure 0.4: Correlation between features

### Relation Between Features and Diabetes Binary

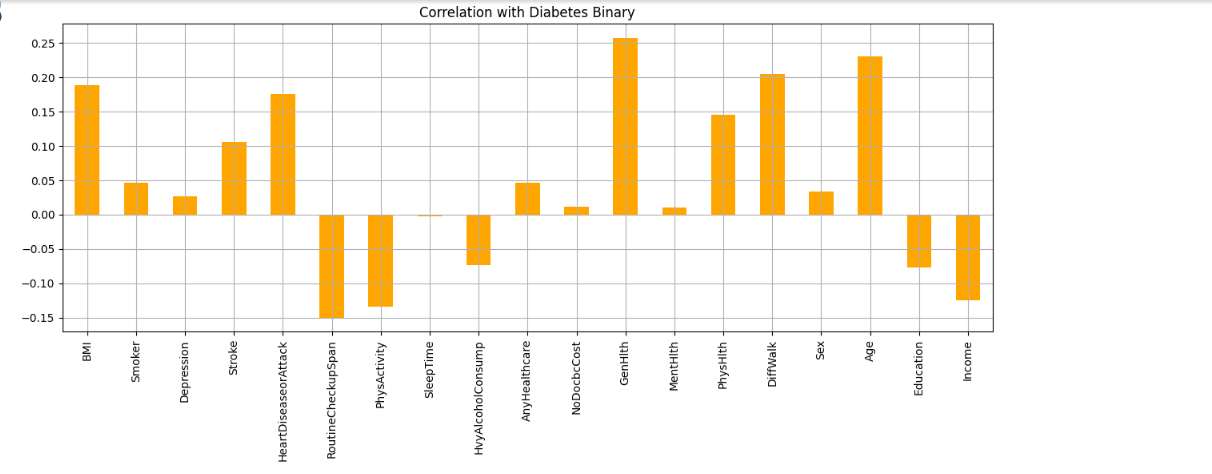


Figure 0.5: Relation between Features and Diabetes Binary

### Relation between different invidual features and Diabetese Binary

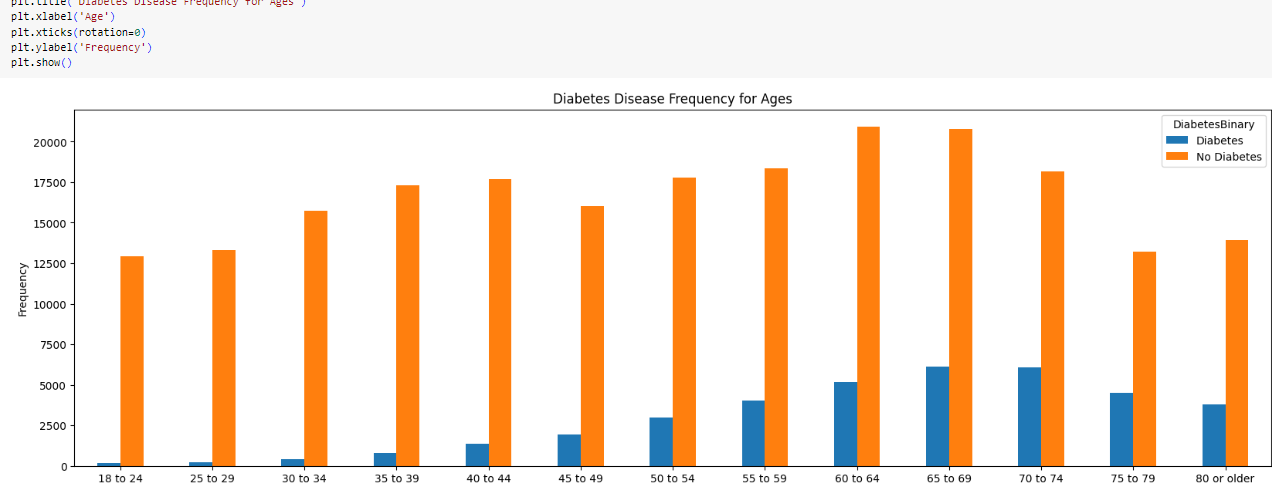


Figure 0.6: Age vs Diabetese Binary

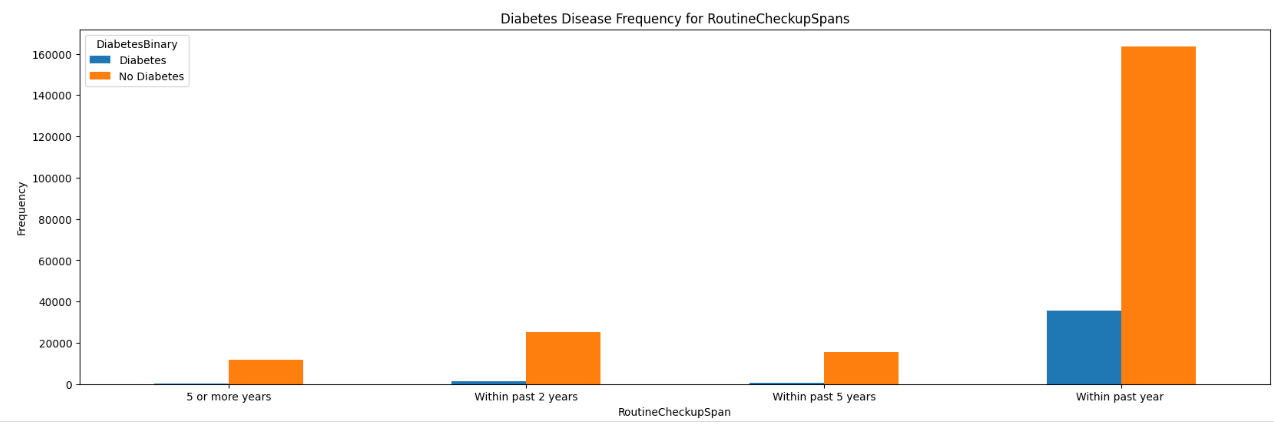


Figure 0.7: RoutineCheckupSpan vs DiabetesBinary

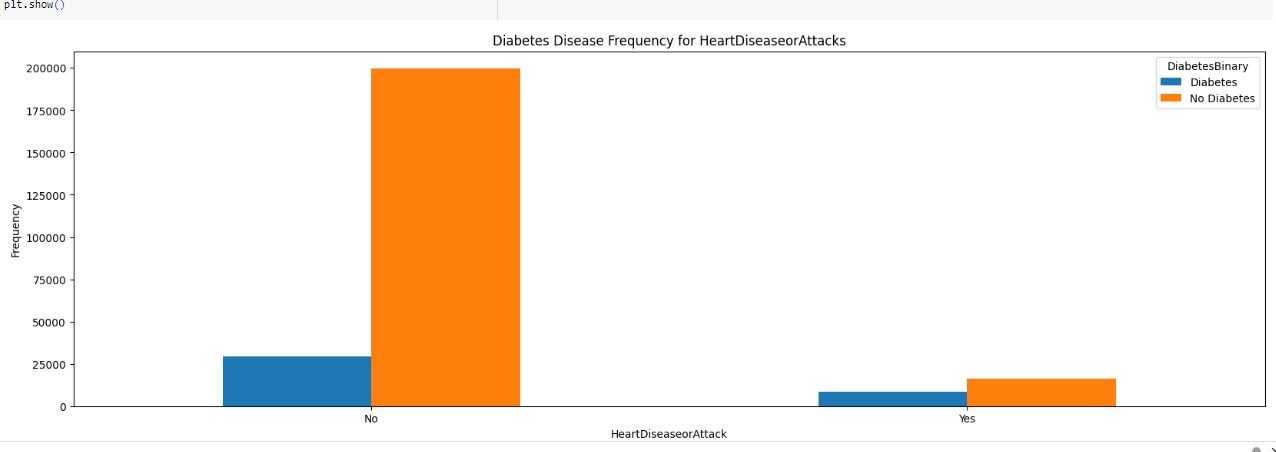


Figure 0.8 : HeartDiseaseAttack vs DiabetesBinary

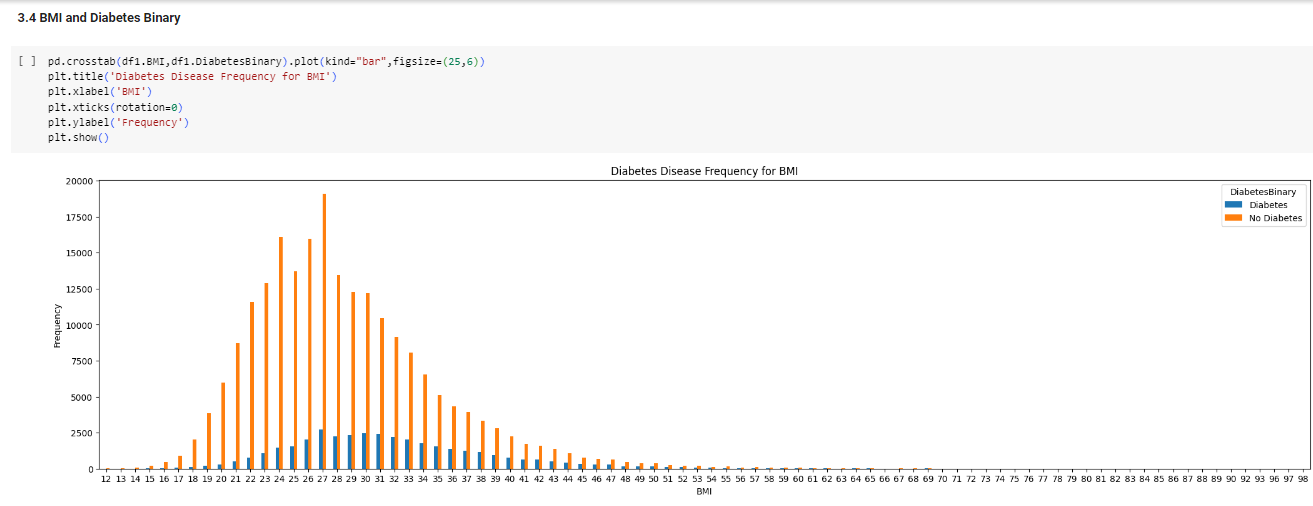


Figure 0.9: BMI vs DiabeteseBinary

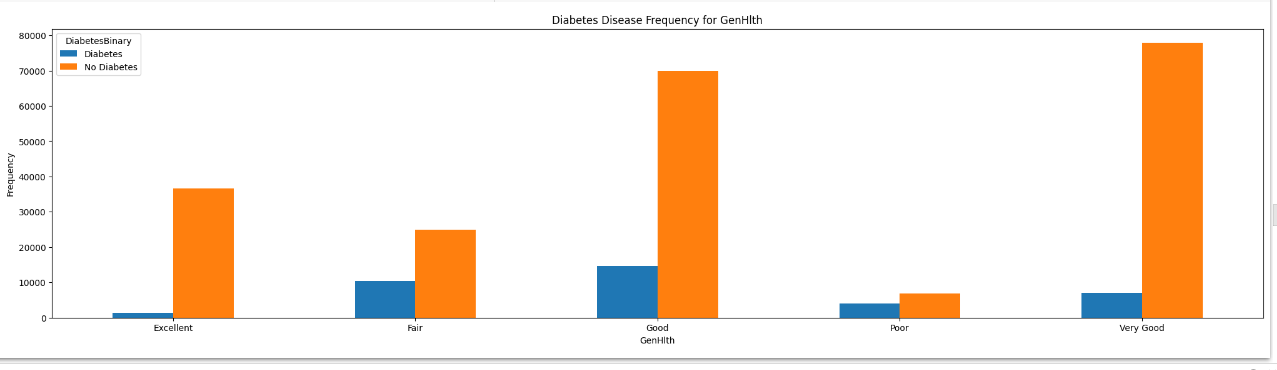


Figure 0.10 : Genhlth vs DiabeteseBinary

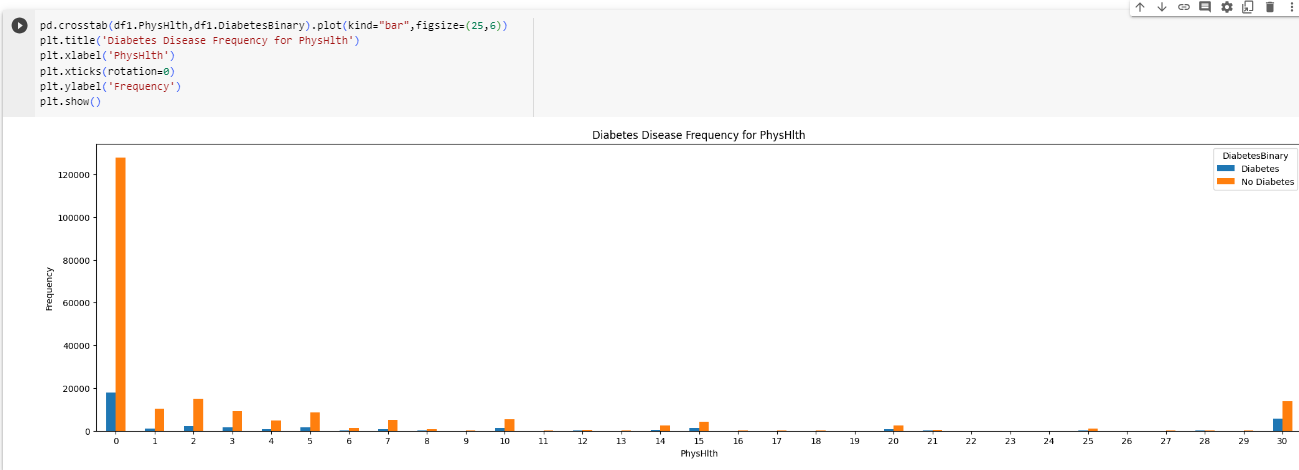


Figure 0.11: Physhlth vs Diabetes Binary

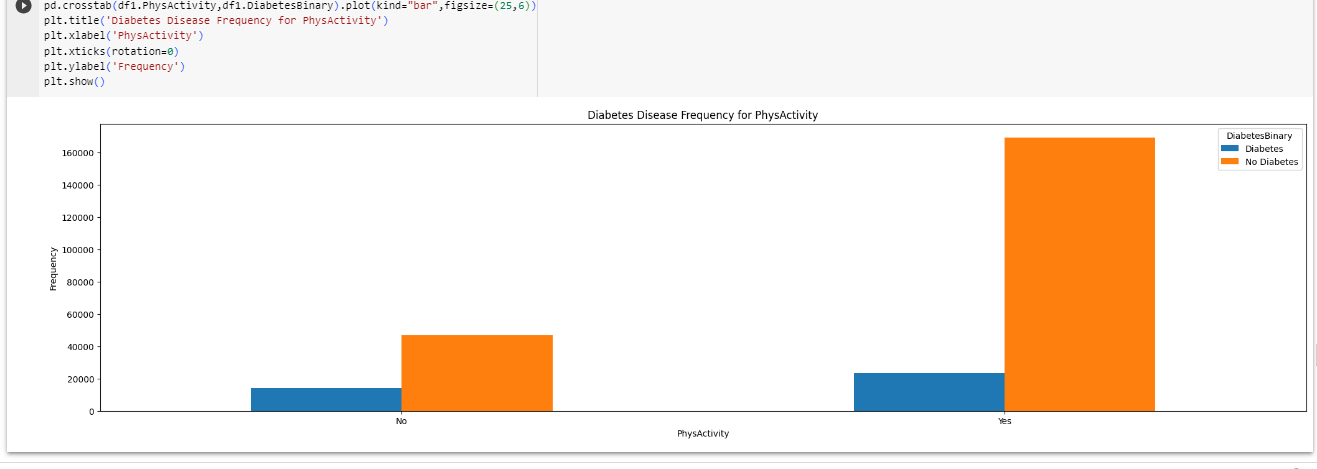


Figure 0.12: PhysActivity vs DiabeteseBinary

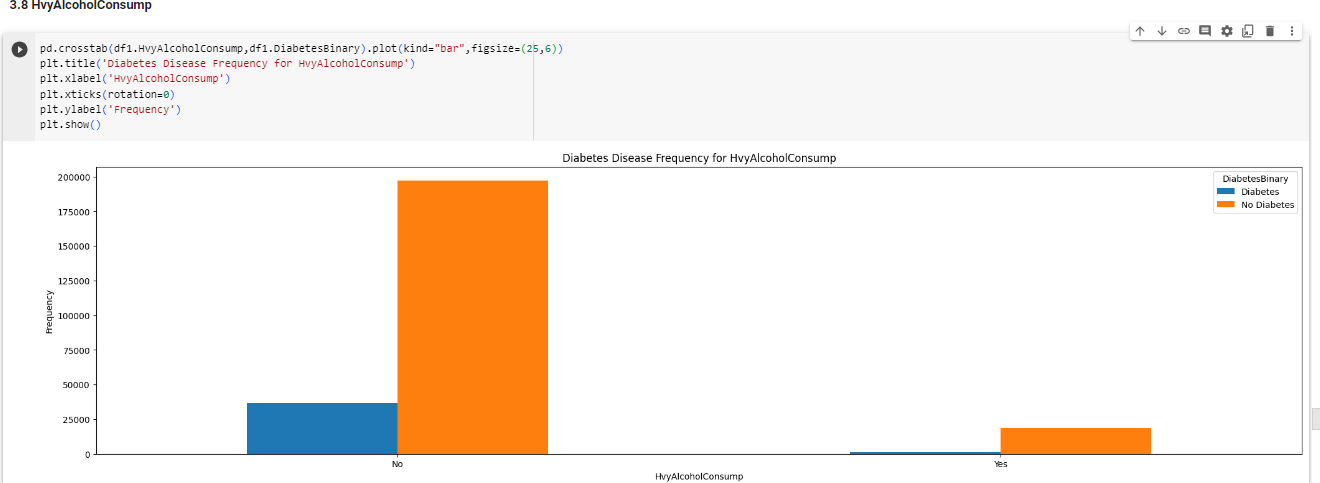


Figure 0.13: HvyAlcoholConsump vs DiabetesBinary

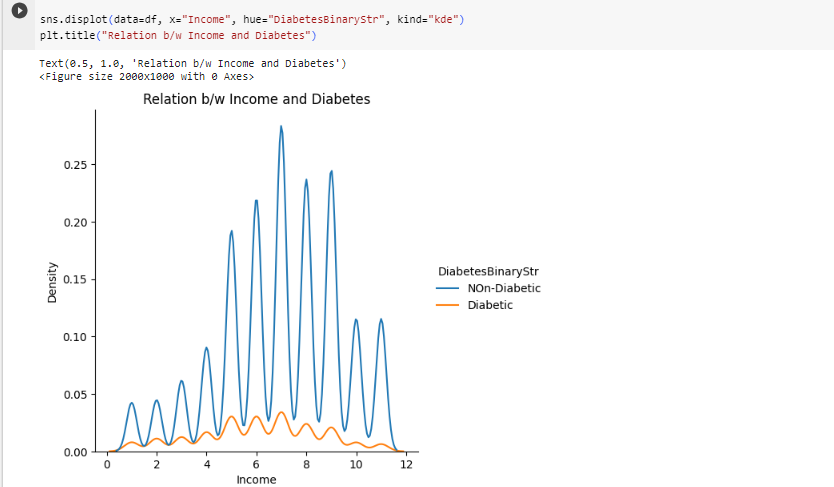


Figure 0.14: Income vs Diabetes Binary

## Model Evaluation and Selection

### Model Evaluation with cross validation and without scaled testing inputs i.e. unscaled validation and testing sets :

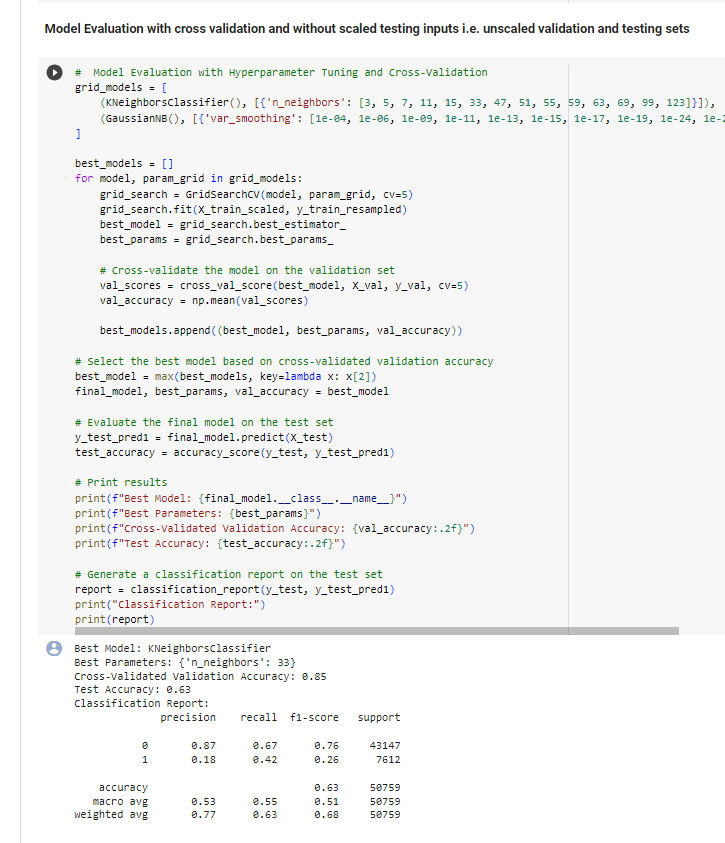


Figure 0.15:Model Evaluation without scaled inputs

### Model Evaluation with cross validation and without scaled testing inputs i.e. unscaled validation and testing sets :

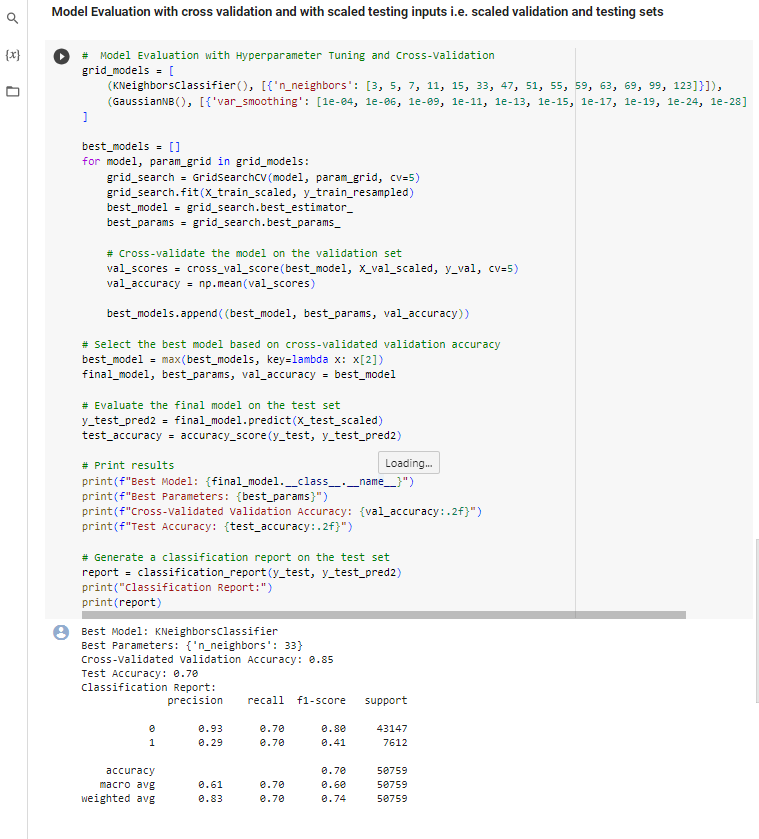


Figure 0.16:Model Evaluation with scaled inputs

### Relation between feature’s different values v/s correct prediction and incorrect prediction:

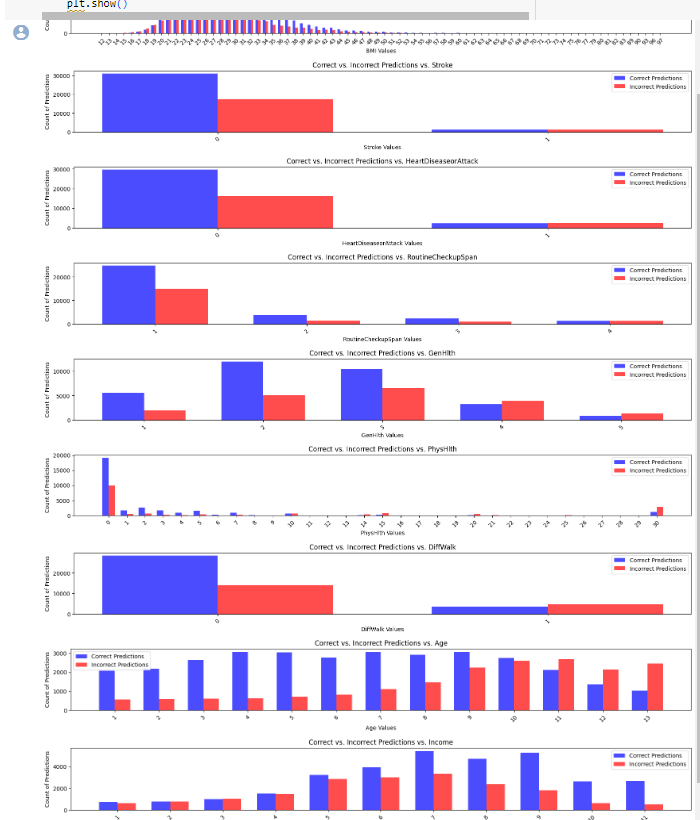


Figure 0.17: Value vs Correct vs Incorrect prediction

## Model Building from scratch

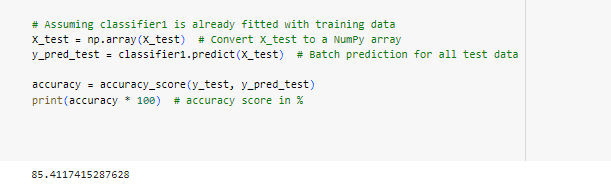


Figure 0.18: Accuracy of built model

## Deployed Model

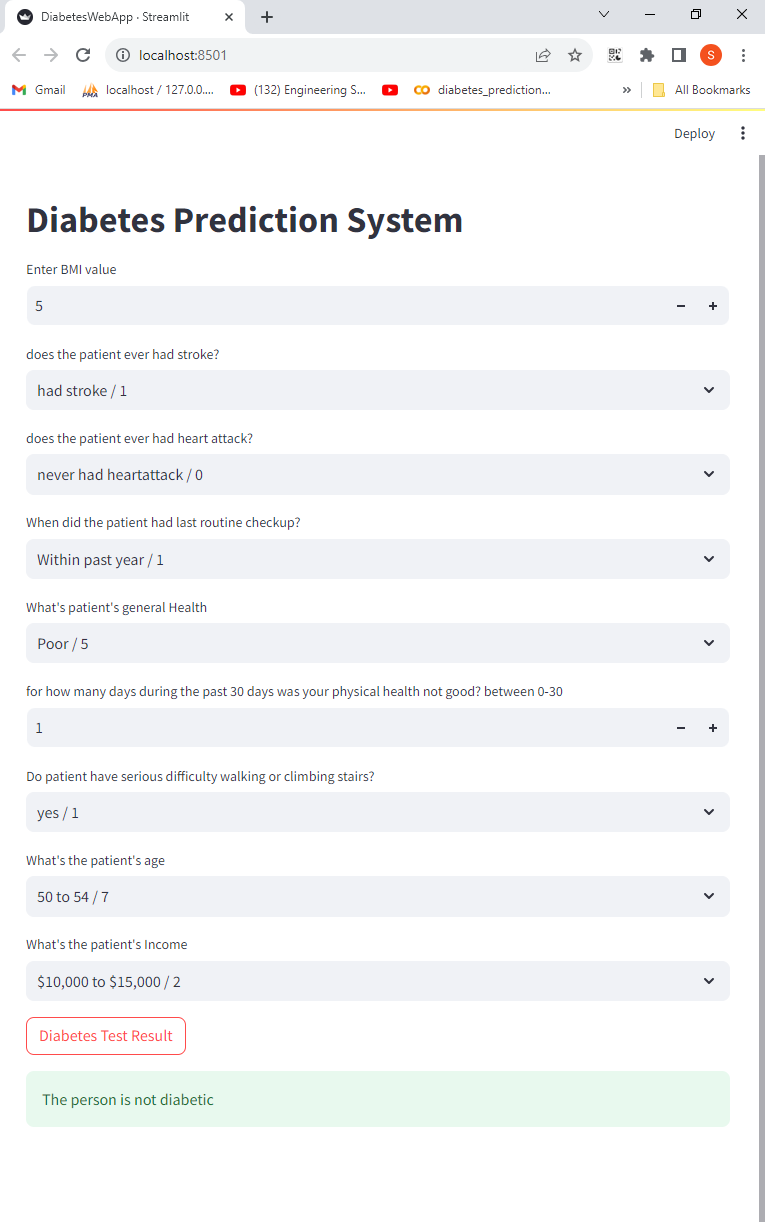


Figure 0.19:User Interface