Diabetes Feature Extraction through machine learning approach

BY

Rafiul Islam
ID: 191-15-12871

Md Nabil Ahmed Nahid
ID: 191-15-12761

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. Md Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY,2023

APPROVAL

This Project/internship titled "**Diabetes Feature Extraction through machine learning approach**", submitted by Rafiul Islam and Md Nabil Ahmed Nahid ID No: 191-15-12871 and 191-15-12761 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on *25-01-2023*.

BOARD OF EXAMINERS

Chairman

____

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

____

Dr. Md. Monzur Morshed
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

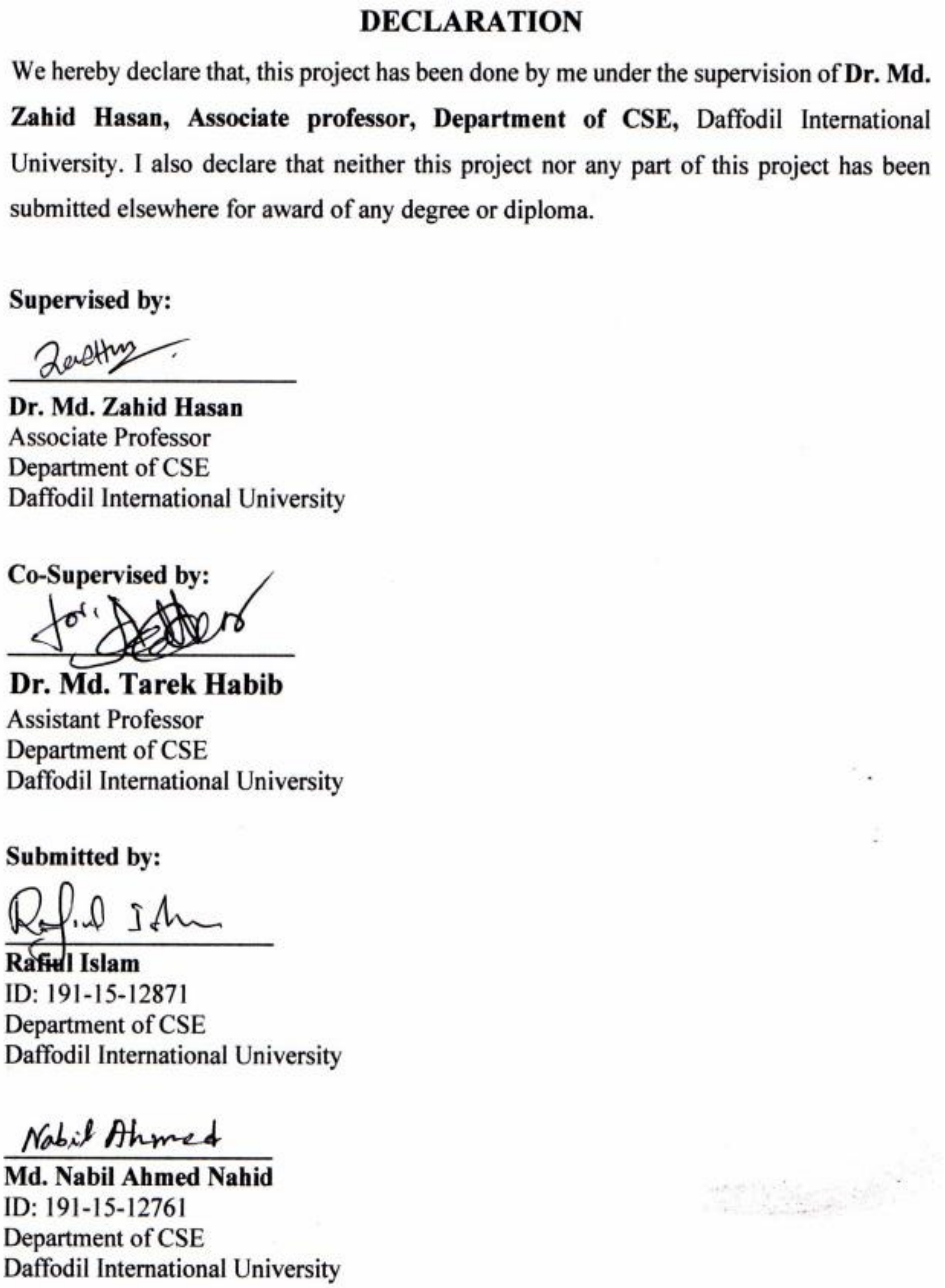
____

Dewan Mamun Raza
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

25.1.2023
____

Dr. Ahmed Wasif Reza
Associate Professor
Department of Computer Science and Engineering
East West University



DECLARATION

We hereby declare that, this project has been done by me under the supervision of **Dr. Md. Zahid Hasan, Associate professor, Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Md. Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Rafiul Islam
ID: 191-15-12871
Department of CSE
Daffodil International University

Nabil Ahmed

Md. Nabil Ahmed Nahid
ID: 191-15-12761
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to Almighty Allah for His divine blessing which makes us possible to complete the final year project/internship successfully.

We are really grateful and wish our profound indebtedness to **Md Zahid Hasan**, **Assistant professor**, Department of CSE, Daffodil International University, Dhaka, deep knowledge & keen interest of our supervisor in the field of Machine Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan**, Head**,** Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

Finally, must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

There are a number of individuals who suffer from diabetes mellitus, definitely among the most popular severe diseases. Diabetic mellitus can be caused by a variety of factors, including age, obesity, inactivity, genetics, dietary habits, blood pressure, and others. An individual's risk of developing diabetes increases chance from growing several illnesses, affect the heart, renal disease, kidneys, nerves harm, eyesight damage and so on. The various tests that are widely utilized in hospitals to diagnose diabetes are used to determine appropriate treatment, according to that diagnosis. So, in this paper, we will discover what the essential components of diabetes causes are in this essay. In areas of application where datasets containing tens or thousands of elements are available, variable and feature choice have become the focus of significant study. Our determination of whether someone is likely to develop diabetes in the future will also focus on the most crucial characteristics. We used two ML algorithms on the dataset to predict diabetes. One is KNN where the other one is K-means algorithm. We found that the model KNN works well on diabetes prediction with the accuracy of 81%.

TABLE OF CONTENTS

CONTENT	PAGE NO
CHAPTER 01: INTRODUCTION	**01-03**
1.1 Introduction	01
1.2 Motivation	02
1.3 Objectives	02
1.4 Problem Statement	02
1.5 Report Layout	03
CHAPTER 02: BACKGROUND STUDY	**04-08**
2.1 Related Work	04
2.2 Challenges	08
CHAPTER 03: RESEARCH METHODOLOGY	**09-19**
3.1 Data collection	09
3.2 Data analysis	09
3.3 Features type, data property	10
3.3.1 Pregnancies	10
3.3.2 Glucose	10
3.3.3 Blood Pressure	10
3.3.4 Skin Thickness	10
3.3.5 Insulin	10
3.3.6 BMI	10
3.3.7 Diabetes Pedigree Function	10
3.3.8 Age	10
3.4 Dataset features correlation with the target column	11
3.5 Dataset insight (Graphs with explanation & details)	12

3.6 Data Preprocessing 17

3.7 Model 17

CHAPTER 04: EXPERIMENTAL RESULT AND DISCUSSION 20-24

4.1 Result 20

4.1.1 Accuracy of model 20

4.1.2 KNN 20

4.1.3 K means clustering 22

CHAPTER 05: CONCLUSION AND FUTURE WORK 24

5.1 Conclusion 24

5.2 Future Work 24

REFERENCES 25-26

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Snap of the dataset	9
Figure 3.2: Heatmap	12
Figure 3.3: Detailed distribution of the features	13
Figure 3.4: Data Visualization	14
Figure 3.5: Histogram for each feature	15
Figure 3.6: Visualizations of train and test score	18
Figure 3.7: Center	19
Figure 4.1: Predicted label of KNN	21
Figure 4.2: ROC curve	22
Figure 4.3: Predicted label of K means clustering	23

LIST OF TABLES

TABLES	**PAGE NO**
Table 3.1: Feature name and data type	11
Table 3.2: Describe of data	16
Table 4.1: Predictors Score according to Model	20
Table 4.2: Confusion Matrix for binary classification	20
Table 4.3: Classification report for KNN	21
Table 4.4: Classification report of K means clustering	23

CHAPTER 1
INTRODUCTION

1.1 Introduction

Diabetes mellitus includes an early warning sign of excessive glucose and a few side symptoms, including constant peeing, increased thirst, longing, and weight loss. Diabetes patients, for the most part, require regular treatment; else, it has the potential to cause a host of serious issues [1]. A blood glucose level of less than 200 mg/dL during the two-hour post-stack period is considered diabetic [2], and the requirement for diabetes recognition brings forth several exams concerning diabetes recognition. There have been several prior research studies regarding diabetes machine learning recognizable evidence. There has been research on diabetes identifiable evidence using k-means, and it has produced some encouraging findings. In contrast to previous work, we conducted a more thorough assessment of numerous regular methods used to identify diabetes[3]. Our proposal is to consider the effectiveness of each of them and choose the one that is most effective. The goal of this study is to identify the optimal preprocessor for each of the classifiers we use based on a few pieces of information and standard preprocessors. We consider utilizing other classifiers when we have adjusted the parameters and various kernels to their predicted highest accuracy. At last, we analyze the relevance of each component in the configuration outcome, which would affect the information in upcoming testing. Among the subfields of artificial intelligence is machine learning, which arose from the necessity to teach PCs how to take in answers to a problem automatically. This area is called an example acknowledgment in design because the PC identifies patterns from data and makes a judgment based on the example detected[4]. It is a rich topic thoroughly and unavoidably pertaining to flag controlling, primarily by utilizing evidence based systems for learning. That have quickly garnered use in practically every industry imaginable, from healthcare to product commerce. To create our prediction model, we tested two distinct algorithms. The first technique we explored was k-means clustering, and the last one was the K-Nearest Neighbors Algorithm. Our primary goal was to develop the ideal prediction algorithm for identifying women with diabetes. We have also identified the critical features that are important in predicting diabetes.

1.2 Motivation

- Predict other diseases early as well.
- Diseases like blindness, kidney failure, heart attacks, stroke can be detected in early stage.
- Sugar level sometimes fluctuates.
- Significantly less expensive
- When a problem with blood sugar is found, doctors and patients can take steps to prevent permanent damage.
- Diabetes is not Curable. Early detection can maintain it.

1.3 Objective

- Data type analysis
- Null count analysis
- Checking the balance of the data by plotting the count of outcomes
- Data pre-processing
- Finding the best clusters 'k'
- Prediction on new data

1.4 Problem Statement

We want to detect diabetes in the early stage. To reduce the risk of complications associated with diabetes, early detection is essential. Symptoms of diabetes can be subtle and hard to recognize, so it is essential to be aware of a familial history of diabetes, being overweight, or being older than 45 are all risk factors. Using Machine learning algorithms can be used to analyze patient data and identify patterns that indicate the presence of diabetes. This can be done by looking at the patient medical history, family history, lifestyle habits, and other data points. By identifying these patterns, machine learning algorithms can accurately predict the risk of diabetes and even diagnose it in its early stages.

1.5 Report Layout

Chapter 1 Presents the research introduction

Chapter 2 Highlights a detailed review of the related literature

Chapter 3 Describes the proposed methodology with a detailed description

Chapter 4 Explains the result analysis

Chapter 5 Concludes of the present research

CHAPTER 2
LITERATURE REVIEW

2.1 Related works

Diabetes currently does not have a treatment, so early detection is crucial. This study employs neural network (NN) techniques to predict diabetes. A dataset containing data on diabetes in Pima Indians was used in their study, and the diabetes prediction model using Logistic Regression (LR) and Support Vector Machine (SVM) was effective, according to the authors. A Neural Network with two hidden layers performed 88.6% accurately when built with different epochs and hidden layers. Approximately 1.6 million people die every year from diabetes, according to the World Health Organization. One condition that develops when blood glucose/blood sugar levels are incredibly high in the body is diabetes. Diabetes, often known as diabetes type 1, is a condition that occurs when the pancreas a gland in the human body, cannot make enough insulin, and as a result of the production of insulin, the body cells can't use it [2]. Following the digestion of food, glucose is released when we eat. An insulin hormone encourages cells to ingest blood glucose and use it as fuel by traveling from the blood to the cells. Because cells cannot take up glucose when the pancreas produces insufficient insulin, the glucose stays in circulation. Therefore, blood glucose levels increase exceptionally quickly in the body. The human body experiences several symptoms of high blood sugar, including severe acute, extreme thirst, and increased urination. The typical blood sugar level concentrations in an adult are 70 to 99 micrograms /dL[5]. Diabetes is indicated if the sugar levels is more significant than 126 micrograms /dl. If a person's physical glucose level is between 100 and 125 micrograms per deciliter, they are said to have pre-diabetes. affect the heart, renal disease, kidneys, nerves harm are potential consequences of an excessively high blood sugar level in humans. Diabetes cannot be cured completely. Long-term diabetes is a leading cause of macrovascular and microvascular complications, which are health issues. Damage to the heart, brain, and legs' major blood arteries constitutes a macrovascular problem. Small blood vessels are damaged by microvascular complications affecting the kidneys, eyes, feet, and nerves. If diabetes is identified early, it can be effectively controlled. Keeping up a healthy exercise routine and food plan can assist in preventing diabetes. There has been a massive shift in the medical field in recent years toward the use of data mining and machine learning technologies. The necessary elements from the healthcare data are preprocessed and chosen using the data mining approach, and diabetes prediction is automated with machine learning. A data mining and machine learning algorithm that identifies the underlying patterns in the data will enable accurate and reliable decisions to be made using the most recent techniques. Data mining uses several approaches, such as machine learning, statistics, and database systems, to find a design in the vast quantity of datasets. The accuracy found respectively 78.8571%, 78.2857%, 77.3429%, and 88.57%. Increasing diabetes incidences are related in large part to the types of nutrition we are receiving today, as well as our irregular eating habits and schedules. A number of factors contribute to diabetes, including obesity and high blood sugar. So, in this paper, we will discover what the critical elements of the reason for diabetes are. The first sign of diabetics is excessive blood sugar, and it also includes a few side symptoms includes frequent urination, increased thirst, increased desire, and losing weight. [6]. It is important that diabetic patients receive consistent treatment because if they do not, they may suffer a variety of perilous and hazardous complications. Diabetes recognition proof has been investigated using SVMs, and some impressive results have been achieved. They present an analysis of a variety of regular diabetes identification systems that differs significantly from the previous work. They suggest considering their plan of action and choosing the most effective one. Each of the classifiers they utilize is examined with both traditional and information preprocessors, with the best preprocessor determined separately. As a result of the computations provided, they could say that a precision of about 84 percent of RF is just the best algo for forecasting diabetic. Additionally, patients must maintain a healthy level of blood sugar and adhering to just a good diet as they become older if they wish to prevent diabetic. Additionally, those with a diabetes-related family history must start taking care of themself. There are, according to the International Diabetes Federation, 382 million diabetics globally. There will be 592 million in 2035, which will be double what it is today. In diabetes mellitus, blood glucose levels are increased as a result of the disease. In addition to traditional physical and chemical testing for diabetes, a range of traditional diagnostic methods can also be used. The complex interdependence of various factors in diabetes, which affects several organs in the body, makes early detection of diabetes very challenging for medical practitioners like kidneys, eyes, heart, nerves, feet, etc [7]. One such endeavor is the use of medical data to make predictions. Combining the findings of various machine-learning approaches, the aim of this study is to create an accurate prediction of diabetes in a patient. SVM, logistic regression, and artificial neural networks all play roles in this project in order to predict diabetes. A category of metabolic illnesses known as diabetes mellitus (DM) is characterized by impaired insulin production and action as their primary cause. Hyperglycemia, which is caused by low insulin levels, impairs the metabolism of proteins, fats, and carbs. One of the most prevalent endocrine disorders, diabetes affects more than 200 million people worldwide. Increasing diabetes prevalence is expected in the years to come. Different classifications can be applied to DM. It is known that type 1 diabetes (T1D) and types 2 diabetes (T2D) are the two major clinical manifestations of this condition according to its etiopathology. Insulin resistance is one of the primary characteristics of T2D, which accounts for 90% of people with diabetes. In T1D, Langerhans islets which house pancreatic cells are thought to be destroyed by an autoimmune process, but T2D is thought to be largely caused by lifestyle, physical activity, diet, and inheritance. In the world, around ten percent of people with diabetes have type 1 diabetes, and ten percent acquire type 2 diabetes as well. Aside from these types of diabetes, there are other types such as gestational diabetes, endocrinopathies, MODY/neurodiabetes, neonatal, mitochondrial, and pregnancy diabetes. These subtypes are categorized according to the insulin secretion profile and onset. Polyuria, polydipsia, and severe weight reduction are only a few of the signs of DM. Blood glucose levels affect the diagnosis (fasting plasma glucose = 7.0 mmol/L). Using support vector machines, artificial neural networks, and logistic regression as classification techniques, the research sought to create a system that would combine all three. It is possible to transform the prediction of diabetes risk completely by using machine learning techniques in combination with a large number of epidemiological and genetic diabetes risk datasets. A large number of databases are used in the healthcare industry. A big data analytics tool is a powerful method for studying massive datasets and extracting information from them and make appropriate predictions. Every year, diabetes claims the lives of between 2 and 5 million sufferers. According to estimates, known as Insulin-Dependent Diabetes Mellitus (IDDM), Type-1 Diabetes has an estimated 629 million people worldwide in 2045. DM, or diabetes mellitus, is caused by an insufficient level of insulin in the body, which is why insulin injections are necessary [8]. The term non-insulin-dependent diabetes mellitus is also used to describe Type 2 Diabetes Mellitus. Insulin resistance is a form of diabetes in which the body cannot use insulin appropriately. When gestational diabetes is detected too late in a woman, or when it is not diagnosed until the third trimester, Type-3 Gestational Diabetes occurs. Diabetes mellitus is associated with long-term complications. It is a method of uncovering information and foretelling future events using a combination of machine learning algorithms and statistics and data mining techniques. A predictive analysis can be used to make significant predictions and decisions based on healthcare data. Predictive analytics can be accomplished using machine learning and regression techniques. Through predictive analytics, diseases can be diagnosed more accurately, patient care can be improved, resources can be optimized, and clinical outcomes can improve. Computer systems can acquire knowledge from past experiences using machine learning without having to program for every possible scenario. It is one of the most important artificial intelligence features that supports the advancement of computer systems. Machine learning is considered to be a dire need in today's situation to eliminate human efforts by supporting automation with minimum flaws. As of now, fasting blood glucose levels and oral glucose tolerance tests are the most commonly used methods for detecting diabetes. This approach takes a lot of time, though. Data mining techniques and machine learning algorithms are used in this paper to build a diabetes prediction model. A literature review is presented in Section II of the paper and discusses diabetes prediction and machine learning taxonomy. Logistic Regression gives the highest accuracy of 96%. AdaBoost classifier was found to be the most accurate with 98.8% accuracy using the pipeline. When this dataset is used in place of the existing data, the diabetes prediction model improves in accuracy and precision. This research may also be expanded to determine the likelihood that non-diabetics will develop diabetic during the few years that follow.

2.2 Challenges

Select Machine Learning Approach: To accomplish their specific tasks, researchers use different machine learning approaches. It is therefore crucial to determine the best machine learning approach for their specific problem. In order to identify diabetes at an early stage, we chose the K-nearest neighbor algorithm and clustering k means.

Accuracy Improvement: Previously k-means clustering method’s accuracy was 31%. After changing the independent attribute, then the accuracy is 69%.

CHAPTER 3
MATERIALS AND METHODS

3.1 Data collection

Our data set’s name is “Pima Indians Diabetes Database**.”** We collected the data set from Kaggle[9]. This site is an online community platform which is for data scientist ML enthusiasts. It is a very popular site from which we can find a data set like ours. As we collected this data set from Kaggle, this data set is much better than the other data sets we found. This data set helps to get much accuracy as we found fewer null values.

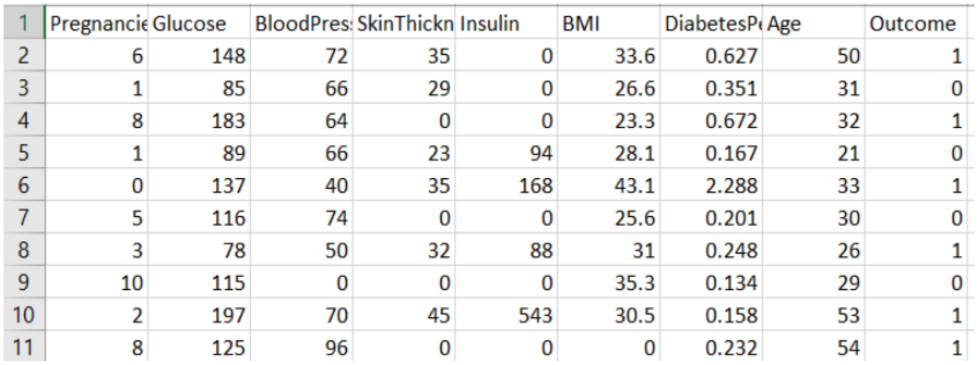
3.2 Data analysis

Dataset features

Here are 7 features in our dataset features such as Pregnancies, Glucose, BloodPressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction and age. Also, there are 769 records or instances there.

Dataset label and its properties

The outcome level is class. There are two outcomes here. The first one is 0, and the second one is 1. As it is a diabetes prediction algorithm, outcome 0 indicates a negative result which means no people with diabetes. On the other hand, outcome 1 indicates the positive side, which means people with diabetes are detected and that indicates whether the people exhibit diabetic symptoms.



1	Pregnancie	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesP	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1

Figure 3.1: snap of the dataset

3.3 Features type, data property, explain feature value in the dataset

The characteristics in a dataset is now explained in detail below:

3.3.1 *Pregnancies*

This denotes the number of occasions the lady became conceived over her lifetime.

3.3.2 *Glucose*

An oral glucose tolerance test measures plasma glucose concentration after two hours.

3.3.3 *BloodPressure*

A person's blood pressure is an effective indicator of their heart's health, and both diastolic and systolic pressures are measured. We have the diastolic pressure in this data set which is the pressure the heart experiences after contractions.

3.3.4 *SkinThickness*

Body fat is measured halfway between the olecranon of the elbow and the acromial process of the scapula on the right arm.

3.3.5 Insulin

Insulin 2 hours serum insulin rate is represented by this number.

3.3.6 BMI

The BMI measures an individual's health (mass in kilogram/height in square meters) in and is an indication of how healthy they are.

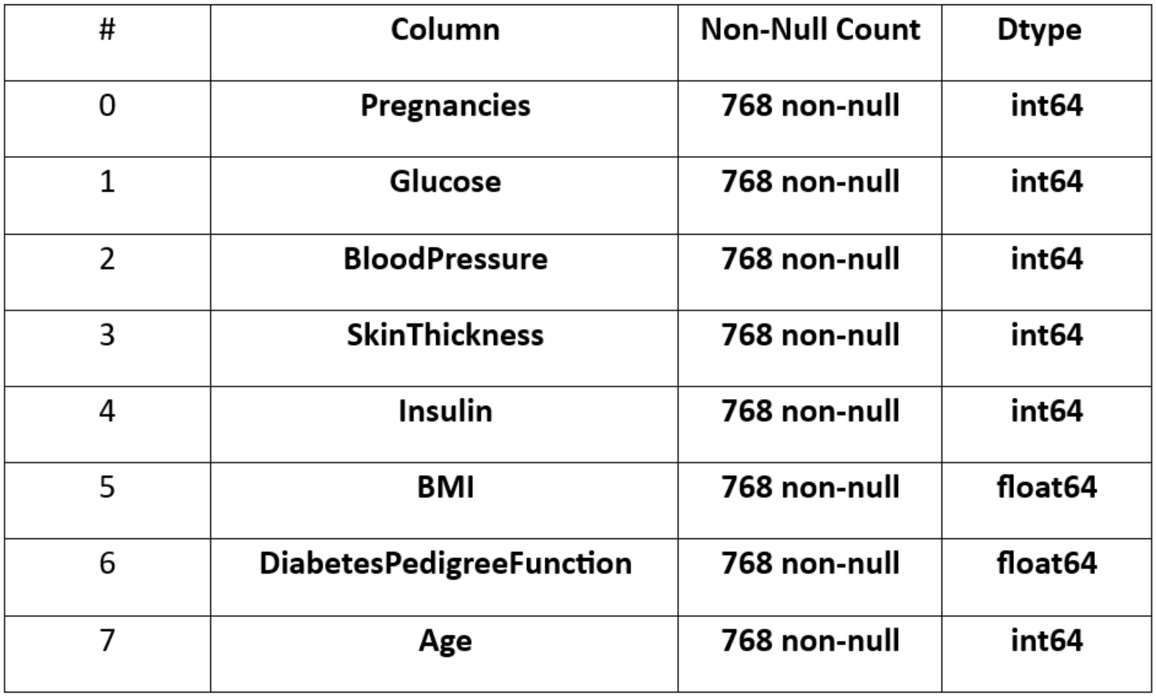
3.3.7 *DiabetesPedigreeFunction*

Having a family history of diabetes is an indication of a higher risk of developing it.

3.3.8 *Age*

Pima woman’s age in years is represented by this number.

TABLE 3.1: Feature name and data type



#	Column	Non-Null Count	Dtype
0	**Pregnancies**	**768 non-null**	**int64**
1	**Glucose**	**768 non-null**	**int64**
2	**BloodPressure**	**768 non-null**	**int64**
3	**SkinThickness**	**768 non-null**	**int64**
4	**Insulin**	**768 non-null**	**int64**
5	**BMI**	**768 non-null**	**float64**
6	**DiabetesPedigreeFunction**	**768 non-null**	**float64**
7	**Age**	**768 non-null**	**int64**

3.4 Dataset features correlation with the target column

In this segment, we will find the correlation between the dataset features and its outcomes. Here the range of the correlations is between -1 to 1, where -1 denotes a negatively correlated feature and +1 denotes the highest positively correlated feature with our target column. For *Pregnancies*, the correlation between outcomes is 0.22. For Glucose, the correlation between outcome and glucose is 0.47. For BloodPressure, the correlation between outcome and bloodpressure is 0.065. For SkinThickness, it is 0.075. For Insulin, it is 0.13. For BMI, it is 0.29. For DiabetesPedigreeFunction it is 0.17; lastly, for *age*, the correlation between outcome and dataset features is 0.24.

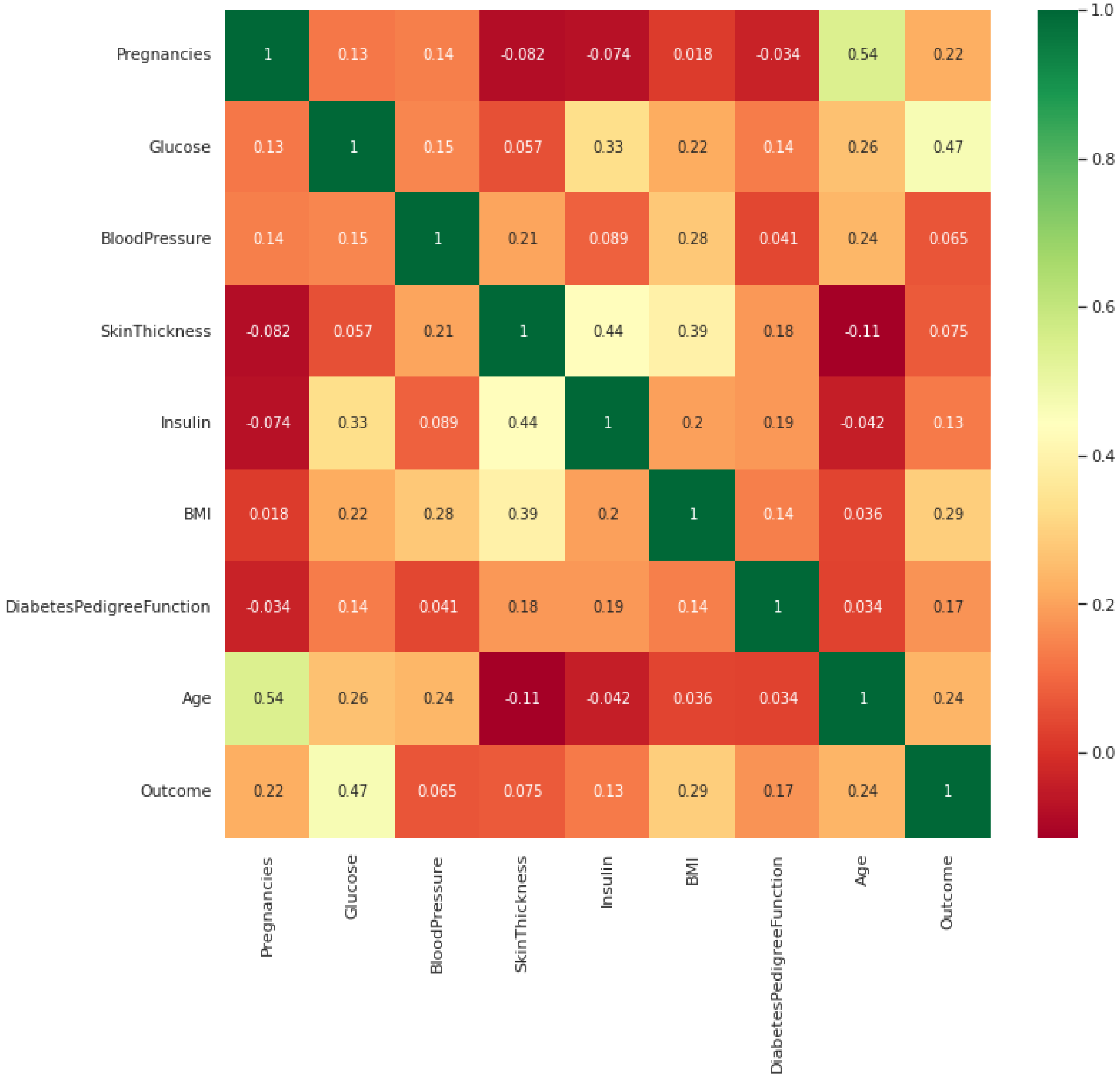


Figure 3.2: Heatmap

3.5 Dataset insight (Graphs with explanation & details)

It is a Scatterplot project with each feature and its outcome which you can observe below:

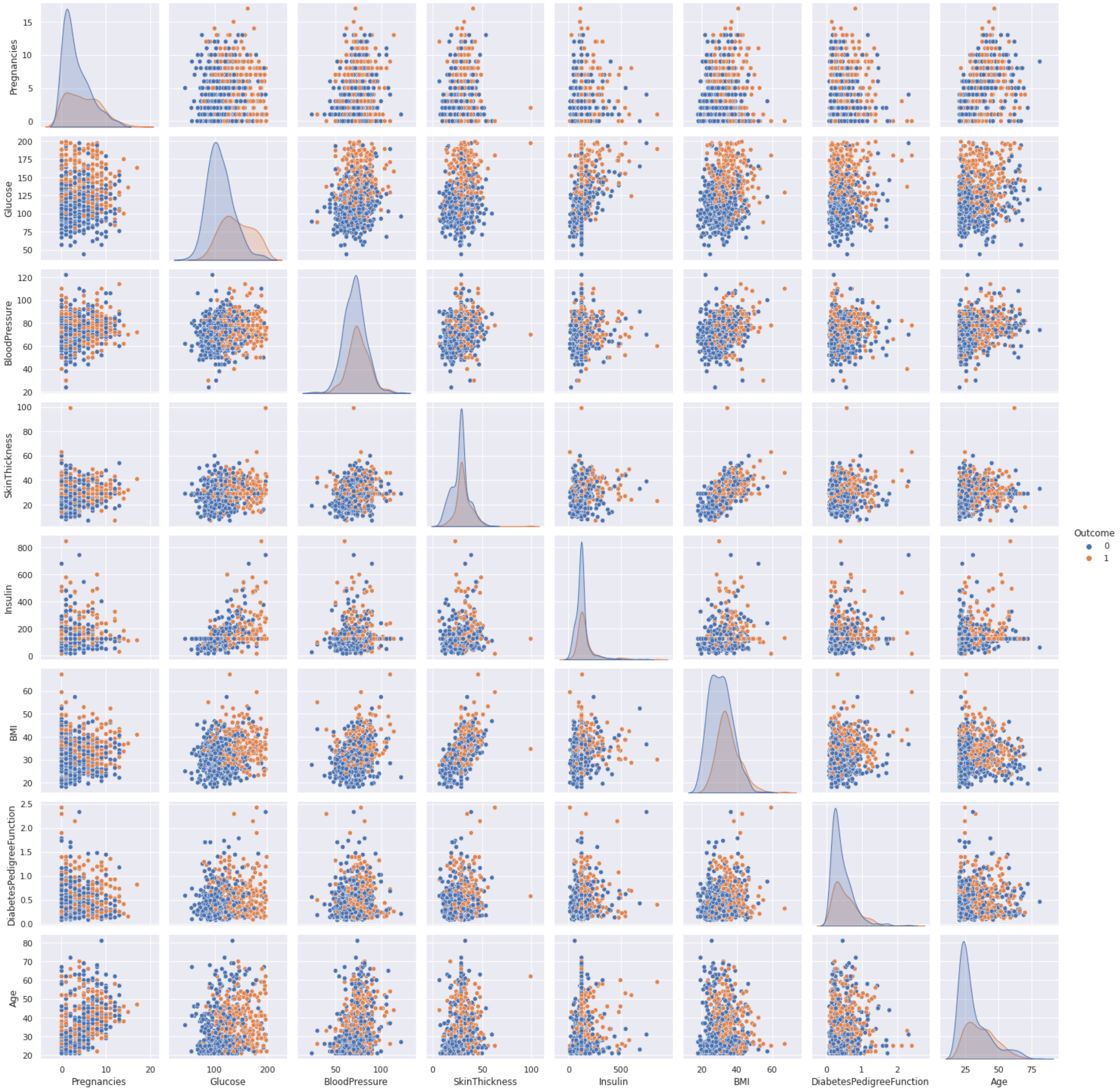


Figure 3.3: Detailed distribution of the features

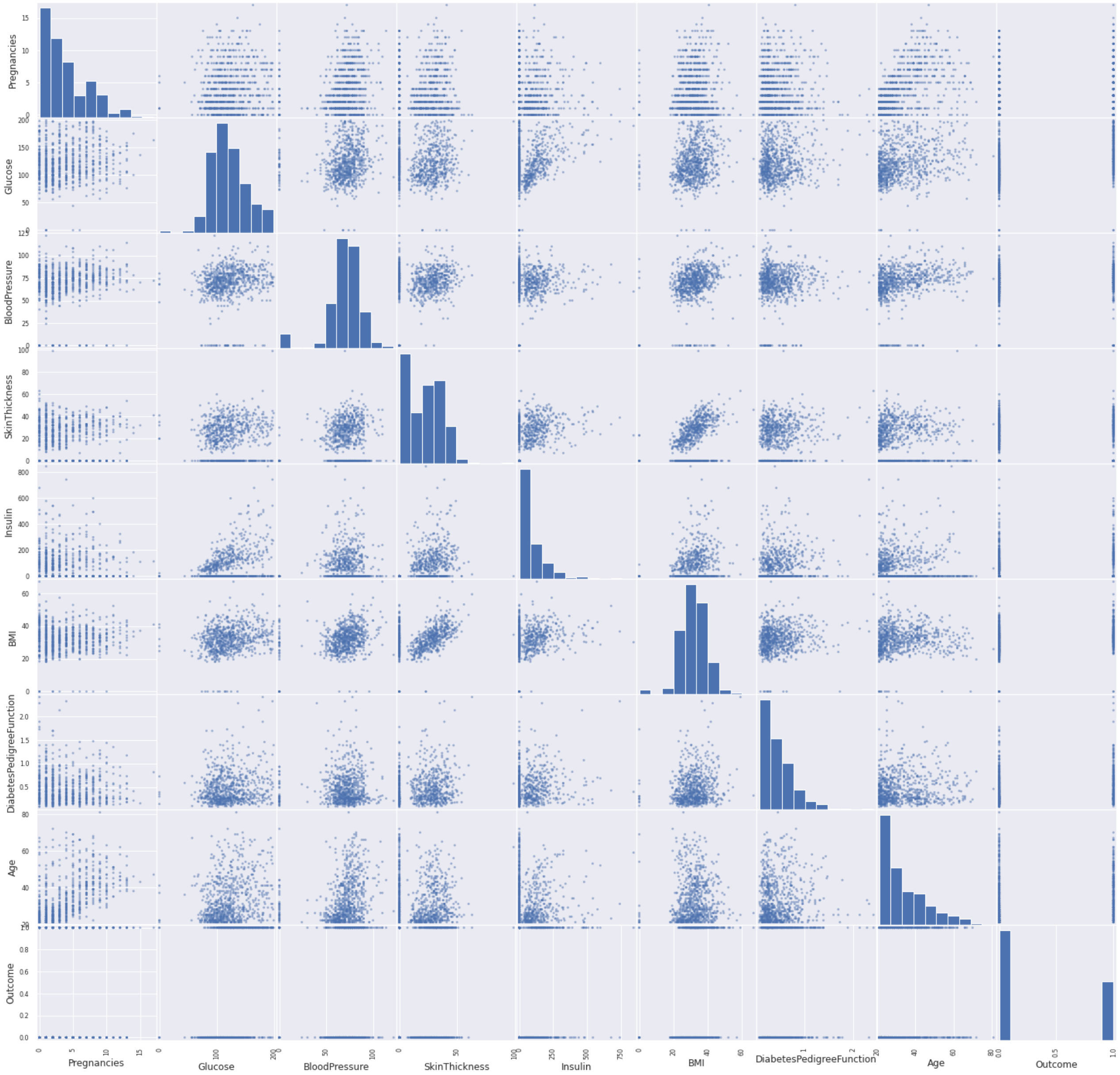


Figure 3.4: Data Visualization

The data value is shown on the x-axis, while the count is shown on the y-axis.

Pregnancies: For the range of 0-0.25 data level is higher than the other ranges. It is more than 250 data. The range 0.25-5.0 has 180 data, 5.0-7.5 has 125 data, 7.5-10.0 has 45 data, 10.0-12.5 has 90 data, and 12.5-15.0 has 50 data for Pregnancies.

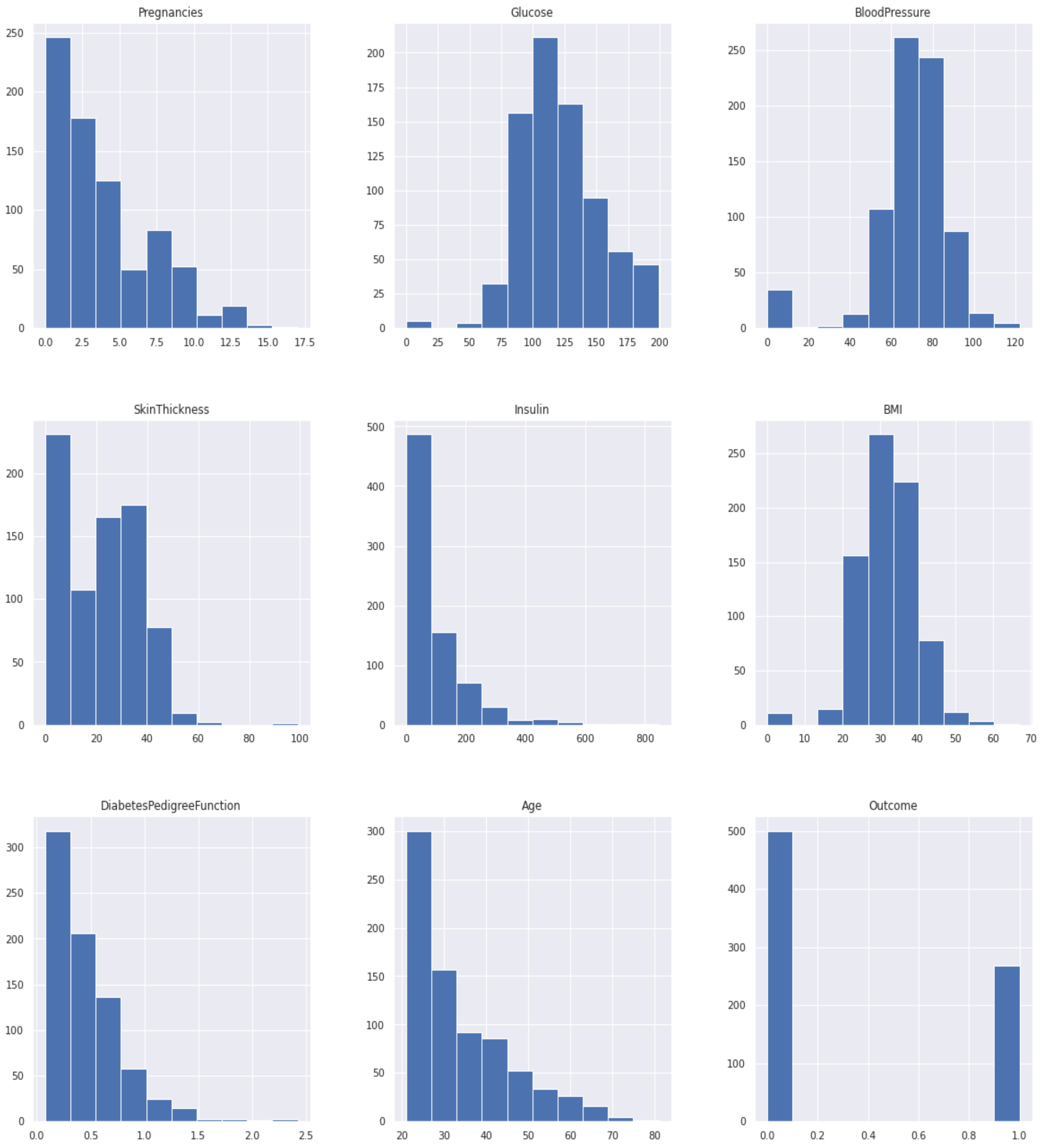


Figure 3.5: Histogram for each feature

In this table, count indicates the number of features. Mean indicates the average of the column. Std means standard deviation. How much data are spreading from the mean value is called standard deviation. Min is the particular value for features. Also, max is the particular value for features. 25%, 50%, and 75% are quartile values. Quartile value means the value of a particular column at that position. 25% quartile 1.50% is quartile2 and 75% is quartile 3.

TABLE 3.2: Describe of data

	Pregnancies	**Glucose**	**BloodPressure**	**SkinThickness**	**Insulin**	**BMI**	**DiabetesPedigreeFunction**	**Age**	**Outcome**
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	22.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.000000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

3.6 Data Preprocessing

Dataset load: We used pandas and its data frame to load the dataset.

Then we converted the data frame to an appropriate NumPy array of size n*m. Where n equals to 768 and m to 8.

While doing data processing, we had to go through a few processes. We pre-processed the data before using it. We had a few null values there. In the pre-processing stage, we used standard scaling along with minmaxscaling. This scaling is for normalizing the data.

Train tests, and splitting was also its essential parts. In dataset splitting, the train was 70%, and the test was 30%. A ratio of 70% training to 30% testing was used to divide the dataset into train and test sections.

Afterwards, we staged the dataset for training into our two distinct model which is KNN and K-means.

3.7 Model

3.7.1 KNN

In Machine Learning, K-Nearest Neighbor combines the supervised learning method with a simple algorithm. The K-NN algorithm places the most recent issue into the category with the least similarity to the general categories based on similarities between the most recent case and available cases. By evaluating the resemblance between the data and the unique informational item, the K-NN approach classifies a unique informational item. That implies that whenever there is a new set of data, an algorithm based on K-NN classification allows easy classification into a suitable suite category. Regression and classification can be accomplished using the K-NN algorithm. However, classification problems are mainly addressed by it. When updated data arrives, the K-Nearest Neighbor method sorts in the most recent data, it should be placed in a comparable category, saving the dataset during the training phase [10].

Configurations

We used KNN from the sci-kit-learn library. After that, we had two classes which were for the outcome. 1 and 0 are the results for the outcome. One indicates that diabetes has been detected, and 0 indicates a negative result. Training method: We trained the model for 15 times, and then we took the average of the best-resulting model. For train score, we found a training score of 0.85 for the position of 2nd, for the 4th position, the score was 0.83; for the 6th position, the score was 0.81; for the 8th position, the score was 0.80; for the 10th position, the score was 0.80, for the 12th position, the score was 0.79, for the 14th position, the score was 0.80. For test score, we found the test score 0.40 for each position of 2nd; for the 4th position, the score was 0.74; for the 6th position, the score was 0.25, for the 8th position, the score was 0.75, for the 10th position, the score was 0.30, for the 12th position, the score was 0.60, for the 14th position, the score was 0.58.

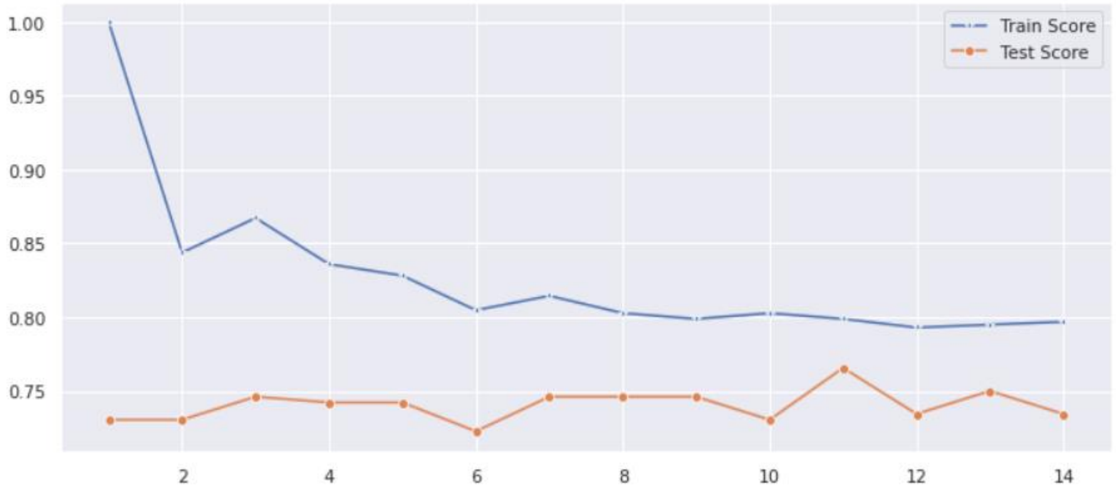
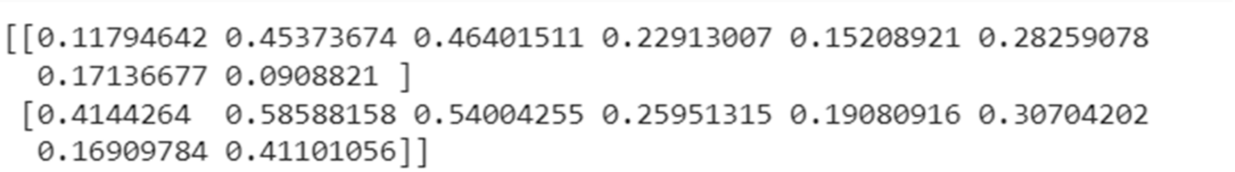


Figure 3.6: Visualizations of train and test score

3.7.2 K-means clustering

The K-Means clustering method is one of the methods for unsupervised learning. Data that has been labeled is not used in clustering in contrast to supervised learning. Objects that are similar and different from each other are divided into clusters using K-Means. The system needs to know how many clusters to create. Two clusters are represented by K = 2. There is a way of finding the optimal K for a given data set [11].

Configurations



```
[[0.11794642 0.45373674 0.46401511 0.22913007 0.15208921 0.28259078
  0.17136677 0.0908821 ]
 [0.4144264  0.58588158 0.54004255 0.25951315 0.19080916 0.30704202
  0.16909784 0.41101056]]
```

Figure 3.7: Center

After that, we had 2 classes which are for outcome. 1 and 0 are the results for outcome.1 indicates that diabetes has detected and 0 indicates the negative result.

CHAPTER 4
EXPERIMENTAL RESULTS

4.1 Results

We applied K-means and KNN algorithm to train the dataset and produce prediction independently.

4.1.1 Accuracy of Models

Analyze the outlines of our models, provided in the table below

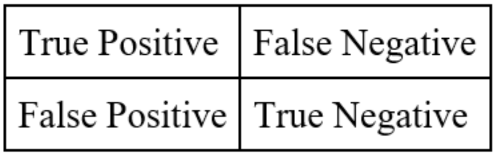
TABLE 4.1: Predictors Score according to Model

Model Name	**Accuracy (%)**
KNN	81%
K-means	69%

4.1.2 KNN

Confusion Matrix: In order to describe the effectiveness of a classification system, a confusion matrix is used. A confusion matrix displays and summarizes the output of a classification algorithm [12].

TABLE 4.2: Confusion Matrix for binary classification



True Positive	False Negative
False Positive	True Negative

For True Positive, it identifies 1 to 1 for 142 times. For this reason, it is truly positive. The false positive is 35. It means it identifies 0 as 1. But those are 0s. Similarly, False Negative is 25, and the True Negative is 54.

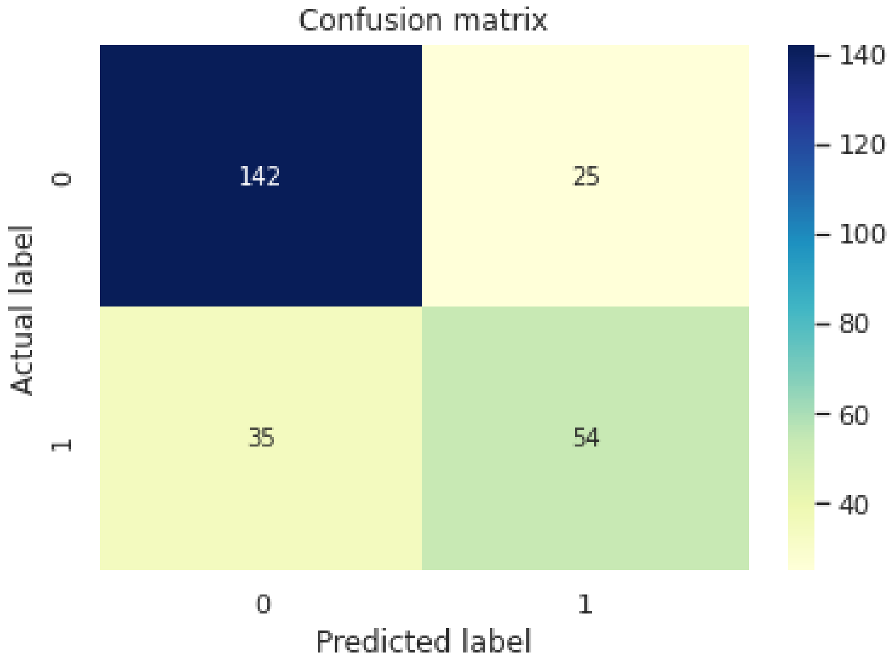


Figure 4.1: Predicted label of KNN

In the segment of precision for 0, the result is 0.83, and for 1, it is 0.80. For the segment of recall, it is 0.78 for 0 and 0.84 for one, and lastly, for the segment of f1-score, it is 0.81 for 0 and 0.82 for 1. The final accuracy is 0.81. So, we can say for KNN, it is 81%.

TABLE 4.3: Classification report for KNN

	Precision	recall	f1-score	Support
0	0.83	0.78	0.81	167
1	0.80	0.84	0.82	167
accuracy			0.81	334
macro avg	0.81	0.81	0.81	334
weighted avg	0.81	0.81	0.81	334

ROC curve: An indication of how well binary classifiers can diagnose problems is shown by a graph called a Receiver Operator Characteristic curve. Its original application was concept of spectrum sensing, but it has now been applied to a variety of fields, encompassing radiographs, medical care, machine learning, and natural disasters. I'll demonstrate how to generate a ROC curve and how to analyze one in this post[13].

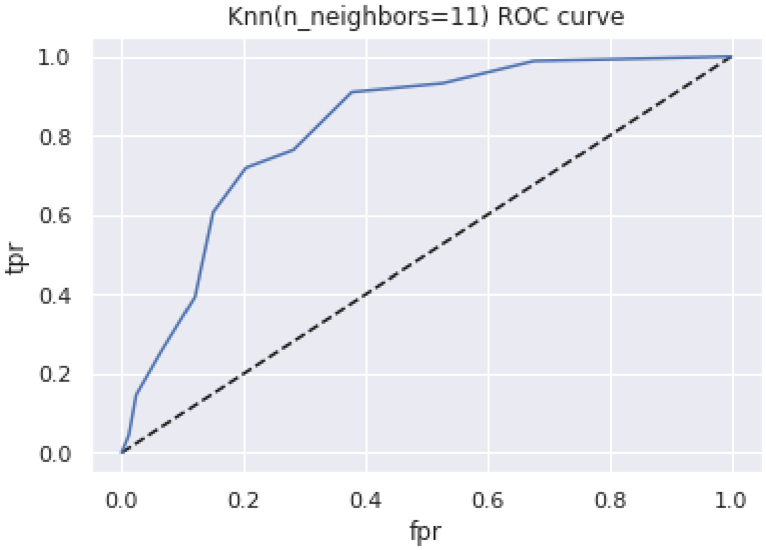


Figure 4.2: ROC curve

4.1.3 K means clustering

Classification metrics: Classification metrics are numbers that measure the effectiveness of assigning observations to classes using machine-learning methods. The binary classification method has two categories: positive and negative [14].

Confusion Matrices: A confusion matrix assists in displaying the results of a classification activity by arranging the different outcomes of the prediction and findings in a tabular format [15].

For True Positive, it identifies 1 to 1 for 77 times. For this reason, it is truly positive. False positive is 24. It means it identifies 0 as 1. But those are 0s. Similarly, False Negative is 23, and the True Negative is 30.

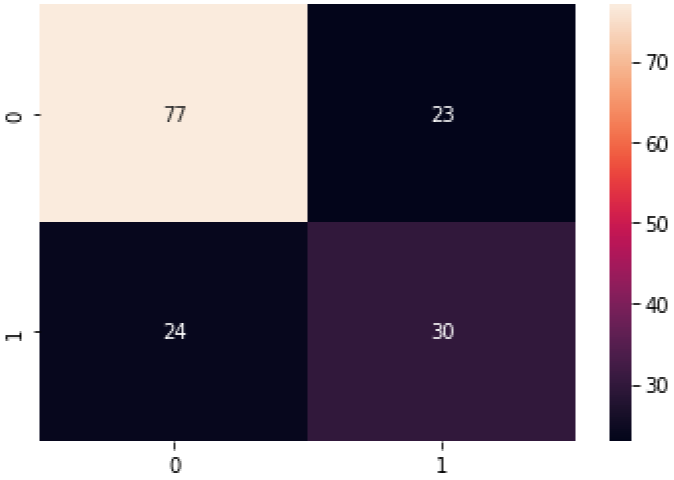


Figure 4.3: Predicted label of K means clustering

In the segment of precision for 0, the result is 0.76, and for 1, it is 0.57. For the segment of recall, it is 0.77 for 0 and 0.56 for one, and lastly, for the segment of f1-score, it is 0.77 for 0 and 0.56 for 1. The final accuracy is 0.69. So, we can say for K-means it is 69%.

Table 4.4: Classification report of K means clustering

	Precision	recall	f1-score	Support
0	0.76	0.77	0.77	100
1	0.57	0.56	0.56	54
accuracy			0.69	154
macro avg	0.66	0.66	0.66	154
weighted avg	0.69	0.69	0.69	154

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

KNN is the most effective algorithm for predicting diabetes based on the above calculations in the early stages, which gives an accuracy of around 81%. Diabetes can also be prevented by maintaining a healthy glucose level and following a healthy diet as people grow older. Furthermore, people whose families have a history of diabetes should also be careful.

5.2 Future Work

1. Create multi-disciplinary dataset
2. Enrich existing dataset
3. Increase model’s accuracy
4. Apply feature engineering

REFERENCES

[1] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, “A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques,” *1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC 2019 - Proceedings*, Nov. 2019, doi: 10.1109/UBMYK48245.2019.8965556.

[2] J. J. Khanam and S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction,” *ICT Express*, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/J.ICTE.2021.02.004.

[3] D. Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms,” *Procedia Comput Sci*, vol. 132, pp. 1578–1585, Jan. 2018, doi: 10.1016/J.PROCS.2018.05.122.

[4] W. Raghupathi and V. Raghupathi, “Analysis and Prediction of Diabetes Using Machine Learning,” *Health Inf Sci Syst*, vol. 2, no. 1, Apr. 2019, doi: 10.1186/2047-2501-2-3.

[5] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, “Classification and prediction of diabetes disease using machine learning paradigm,” *Health Inf Sci Syst*, vol. 8, no. 1, pp. 1–14, Dec. 2020, doi: 10.1007/S13755-019-0095-Z/TABLES/13.

[6] D. Dutta, D. Paul, and P. Ghosh, “Analysing Feature Importances for Diabetes Prediction using Machine Learning,” *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2018*, pp. 924–928, Jan. 2019, doi: 10.1109/IEMCON.2018.8614871.

[7] T. N. Joshi and P. M. Chawan, “Diabetes Prediction Using Machine Learning Techniques Related papers Classificat ion Of Diabet es Disease Using Support Vect or Machine Much Aziz Muslim Haemorrhage Det ect ion and Classificat ion: A Review IJERA Journal Digit al Image Processing Assessment in Mult i Slice CT Angiogram using Liner, Non-Liner and Bot h Cla… IJERA Journal Diabetes Prediction Using Machine Learning Techniques,” *Computer Engg. and Info. Tech., V.J.T.I*, vol. 8, pp. 2248–9622, 2018, doi: 10.9790/9622-0801020913.

[8] A. Mujumdar and V. Vaidehi, “Diabetes Prediction using Machine Learning Algorithms,” in *Procedia Computer Science*, 2019, vol. 165, pp. 292–299. doi: 10.1016/j.procs.2020.01.047.

[9] “Pima Indians Diabetes Database | Kaggle.” https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?fbclid=IwAR1auaaiqjpvdloT1cyxJxGze6yKDr1uoY-6PSS_vUkzcsOGIRGtQeatDGA (accessed Dec. 05, 2022).

[10] “K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint.” https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (accessed Dec. 05, 2022).

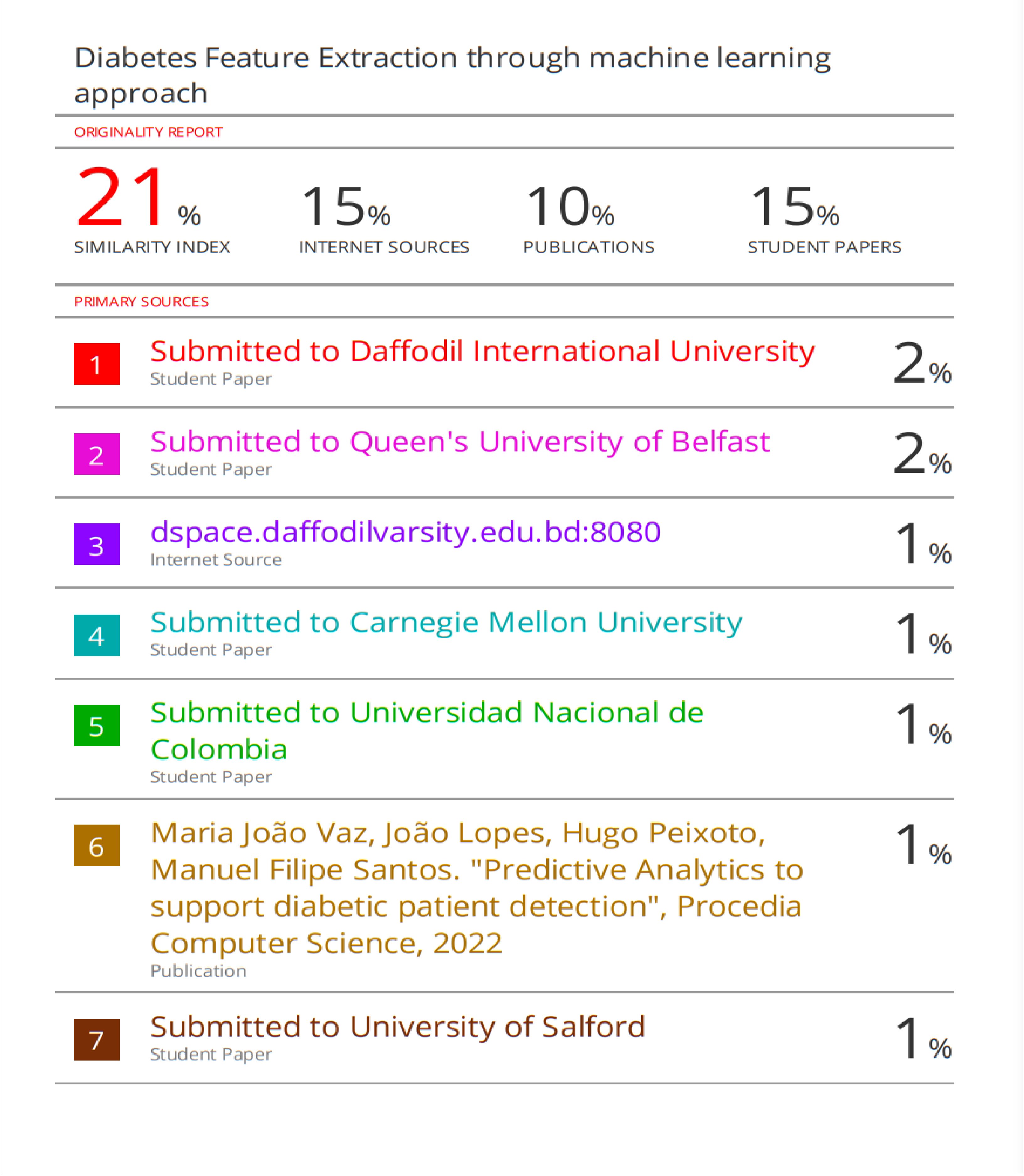
[11] “K-means Clustering Algorithm: Applications, Types, and Demos [Updated] | Simplilearn.” https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm (accessed Dec. 05, 2022).

[12] A. Kulkarni, D. Chong, and F. A. Batarseh, “Foundations of data imbalance and solutions for a data democracy,” *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*, pp. 83–106, Jan. 2020, doi: 10.1016/B978-0-12-818366-3.00005-8.

[13] “What is a ROC Curve - How to Interpret ROC Curves - Displayr.” https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/ (accessed Dec. 05, 2022).

[14] “24 Evaluation Metrics for Binary Classification (And When to Use Them) - neptune.ai.” https://neptune.ai/blog/evaluation-metrics-binary-classification (accessed Dec. 05, 2022).

[15] “What is a Confusion Matrix in Machine Learning?” https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning (accessed Dec. 05, 2022).



Diabetes Feature Extraction through machine learning approach

ORIGINALITY REPORT

21%	15%	10%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	Submitted to Queen's University of Belfast Student Paper	2%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	Submitted to Carnegie Mellon University Student Paper	1%
5	Submitted to Universidad Nacional de Colombia Student Paper	1%
6	Maria João Vaz, João Lopes, Hugo Peixoto, Manuel Filipe Santos. "Predictive Analytics to support diabetic patient detection", Procedia Computer Science, 2022 Publication	1%
7	Submitted to University of Salford Student Paper	1%