

**School of Computer Science and Engineering**

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

**Final Review Report**

**Programme: B.Tech**

**Course:** CSE2004 Database Management Systems

**Slot:** D2

**Faculty:** Dr. M. Premalatha

**Component:** J

# Title: Prediction and Diagnosis of Diabetes using Classification Mining Techniques.

**Team Member(s): [Min 1, Max-3 in a team]**

Harsh Gupta (20BCE1310)

Aryan Raj Wadhwa (20BCE1328)

Siddharth S (20BCE1353)

**Abstract**

**[What, Why, How]**

Prediction and Diagnosis of Diabetes using Classification Mining Techniques

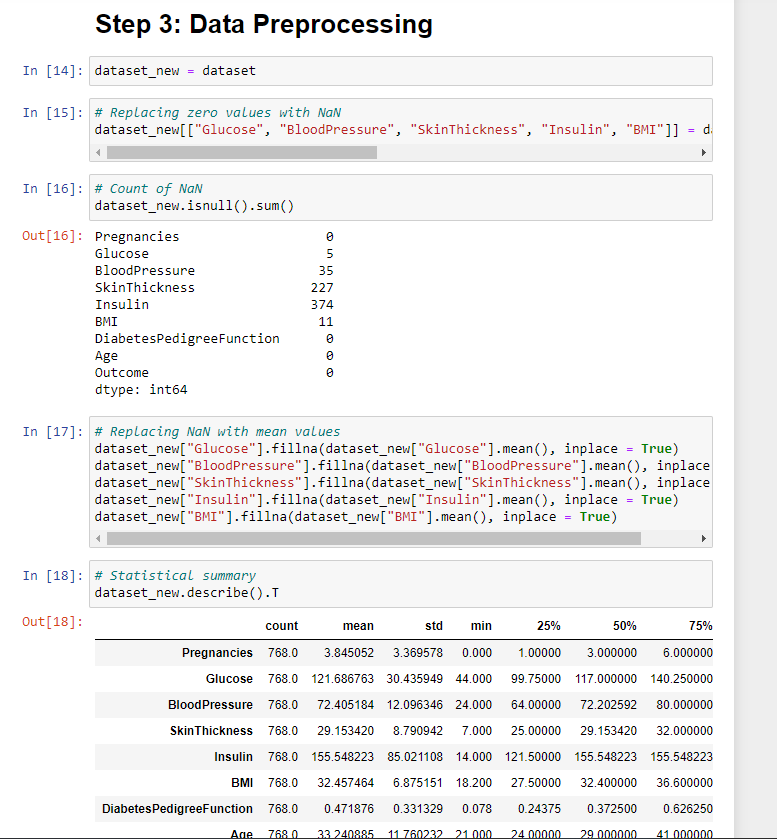
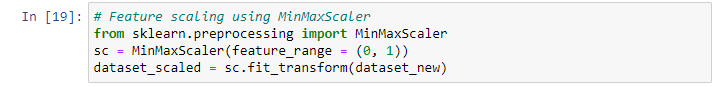
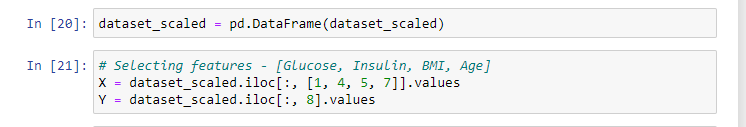
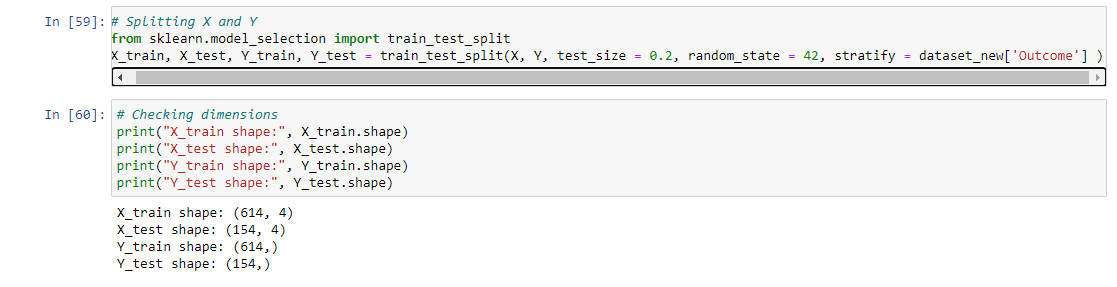
Diabetes is a condition in which the body does not properly process food for use as energy. Most of the food we eat is turned into glucose, or sugar, for our bodies to use for energy. The pancreas, an organ that lies near the stomach, makes a hormone called insulin to help glucose get into the cells of our bodies. When you have diabetes, your body either doesn't make enough insulin or can't use its own insulin as well as it should. This causes sugars to build up in your blood. This is why many people refer to diabetes as “sugar.”

Diabetes has affected over 246 million people worldwide with a majority of them being women. According to the WHO report, by 2025 this number is expected to rise to over 380 million. The disease has been named the fifth deadliest disease in the United States with no imminent cure in sight. With the rise of information technology and its continued advent into the medical and healthcare sector, the cases of diabetes as well as their symptoms are well documented.

This paper aims at finding solutions to diagnose the disease by analyzing the patterns found in the data through classification analysis by employing Classification methodology. The research hopes to propose a quicker and more efficient technique of diagnosing the disease, leading to timely treatment of the patients

**Introduction**

**[Brief explanation about the project]**

1. **Data Set Description**: [Explain the dataset in detail with the number of columns and rows identified with pre-processing steps. Also specify URL from where the dataset is retrieved]  
   The dataset chosen for this project is titled “Pima Indians Diabetes Database” and is obtained from Kaggle.com and can be found under the URL  
   <https://www.kaggle.com/uciml/pima-indians-diabetes-database>  
     
   The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset contains a total of 9 columns with 768 total records.  
   The columns are listed as follows:  
     
   **Pregnancies:** The number of pregnancies that had happened to the subject of a record.  
   **Glucose:** Plasma glucose concentration after 2 hours in an oral glucose tolerance test.  
   **Blood Pressure:** Diastolic blood pressure (measured in mm Hg)  
   **Skin thickness:** Triceps skin fold thickness (measured in mm)  
   **Insulin:** 2-Hour serum insulin  
   **BMI:** Body Mass Index (weight in kg / (height in m)2)  
   **Diabetes Pedigree Function:** It utilizes information from a person’s family history to predict how diabetes will affect that individual.   
   **Age:** Age of the individual in years  
   **Outcome:** The outcome of the test (1 if diabetic, 0 if non diabetic)  
     
   In the pre-processing stage, we take attributes that shouldn’t have empty values (such as BMI, BloodPressure, SkinThickness, Insulin and Glucose) and convert all 0 values to NaN. The mean of the remaining values is then calculated, and filled in.   
     
   Using MinMaxScaler from the sklearn package, we scale all these attributes to the range(0,1). This is normalisation and it is the most important step.  
     
   We then identify only the relevant attributes (based on the correlation strength to outcome), which is Glucose, BMI, Insulin and Age.   
     
   Finally, the training and test datasets are split, with a test to training ratio of 2:8. Our dataset is now ready to be processed by the classification algorithms.   
   
2. Consider the schema alone and normalize it till BCBF using schema decomposition  
   **(Our records contain 768 rows but we have used only 5 for a sample )**

**Normalization of the table**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pregnancies** | **Glucose** | **Blood**  **Pressure** | **Skin**  **Thickness** | **Insulin** | **BMI** | **Diabetes**  **Pedigree** | **Age** | **Outcome** |
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

**Diabetes database**

**Diabetes(Pregnancies , Glucose , BloodPressure , SkinThickness , Insulin , BMI , DiabetesPedigree , Age , Outcome)**

1NF – Identify Key, FDs

**Glucose -> Insulin , BMI , DiabetesPedigree,Skin Thickness**

**Outcome**

**Pregnancies -> Age**

**Blood Pressure**

As **BMI** increases, insulin resistance also increases which results in increased blood **glucose** level in body. Since body weight is associated **with BMI**, it may be expected that **BMI** should correlate **with** blood **glucose** levels.

**Insulin** helps the cells absorb **glucose**, reducing **blood sugar** and providing the cells with **glucose** for energy.

**Skin thickness** is primarily determined by collagen content and is increased in insulin-dependent diabetes mellitus (IDDM).

**After 1NF**

**Table 1:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Glucose (pk)** | **Insulin** | **BMI** | **Diabetes**  **Pedigree** | **Skin**  **Thickness** |
| 148 | 0 | 33.6 | 0.627 | 35 |
| 85 | 0 | 26.6 | 0.351 | 29 |
| 183 | 0 | 23.3 | 0.672 | 0 |
| 89 | 94 | 28.1 | 0.167 | 23 |
| 137 | 168 | 43.1 | 2.288 | 35 |

**Table2:**

|  |
| --- |
| **Outcome(pk)** |
| 1 |
| 0 |

**Table 3:**

|  |  |
| --- | --- |
| **Pregnancies** | **Age** |
| 6 | 50 |
| 1 | 31 |
| 8 | 32 |
| 1 | 21 |
| 0 | 33 |

**Table 4:**

|  |
| --- |
| **Blood(pk)**  **Pressure** |
| 72 |
| 66 |
| 64 |
| 66 |
| 40 |

**Table 5:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pregnancies** | **Glucose**  **(fk)** | **Blood**  **Pressure**  **(fk)** | **Skin**  **Thickness** | **Insulin** | **BMI** | **Diabetes**  **Pedigree** | **Age** | **Outcome**  **(fk)** |
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

2NF – No Partial Dependency – Decompose – Non key depends on part of key

attribute(s)

Table 1 violates 2NF as Insulin also determines Skin thickness

Therefore,

**Glucose -> Insulin, BMI, DiabetesPedigree**

**Outcome**

**Pregnancies ->Age**

**Blood Pressure**

**Insulin -> Skin Thickness**

**After 2NF:**

**Table 1:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Glucose (pk)** | **Insulin(fk)** | **BMI** | **Diabetes**  **Pedigree** |
| 148 | 0 | 33.6 | 0.627 |
| 85 | 0 | 26.6 | 0.351 |
| 183 | 0 | 23.3 | 0.672 |
| 89 | 94 | 28.1 | 0.167 |
| 137 | 168 | 43.1 | 2.288 |

**Table 2:**

|  |  |
| --- | --- |
| **Insulin (pk)** | **Skin Thickness** |
| 0 | 0 |
| 94 | 23 |
| 168 | 35 |

**Table 3:**

|  |
| --- |
| **Outcome(pk)** |
| 1 |
| 0 |

**Table 4:**

|  |  |
| --- | --- |
| **Pregnancies** | **Age** |
| 6 | 50 |
| 1 | 31 |
| 8 | 32 |
| 1 | 21 |
| 0 | 33 |

**Table 5:**

|  |
| --- |
| **Blood(pk)**  **Pressure** |
| 72 |
| 66 |
| 64 |
| 66 |
| 40 |

**Table 6:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pregnancies** | **Glucose**  **(fk)** | **Blood**  **Pressure**  **(fk)** | **Skin**  **Thickness** | **Insulin**  **(fk)** | **BMI** | **Diabetes**  **Pedigree** | **Age** | **Outcome**  **(fk)** |
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

3NF – No Transitive Dependency – No Non key to Non key dependency

Table 4 violates 3NF as Pregnancies is a non key and it determines Age

Therefore,

**Glucose -> Insulin , BMI , DiabetesPedigree**

**Outcome**

**Pregnancies**

**Age**

**Blood Pressure**

**Insulin -> Skin Thickness**

**After 3NF:**

**Table 1:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Glucose (pk)** | **Insulin(fk)** | **BMI** | **Diabetes**  **Pedigree** |
| 148 | 0 | 33.6 | 0.627 |
| 85 | 0 | 26.6 | 0.351 |
| 183 | 0 | 23.3 | 0.672 |
| 89 | 94 | 28.1 | 0.167 |
| 137 | 168 | 43.1 | 2.288 |

**Table 2:**

|  |  |
| --- | --- |
| **Insulin (pk)** | **Skin Thickness** |
| 0 | 0 |
| 94 | 23 |
| 168 | 35 |

**Table 3:**

|  |
| --- |
| **Outcome(pk)** |
| 1 |
| 0 |

**Table 4:**

|  |
| --- |
| **Pregnancies (pk)** |
| 6 |
| 1 |
| 8 |
| 0 |

**Table 5:**

|  |
| --- |
| **Age(pk)** |
| 50 |
| 31 |
| 32 |
| 21 |
| 33 |

**Table 6:**

|  |
| --- |
| **Blood(pk)**  **Pressure** |
| 72 |
| 66 |
| 64 |
| 66 |
| 40 |

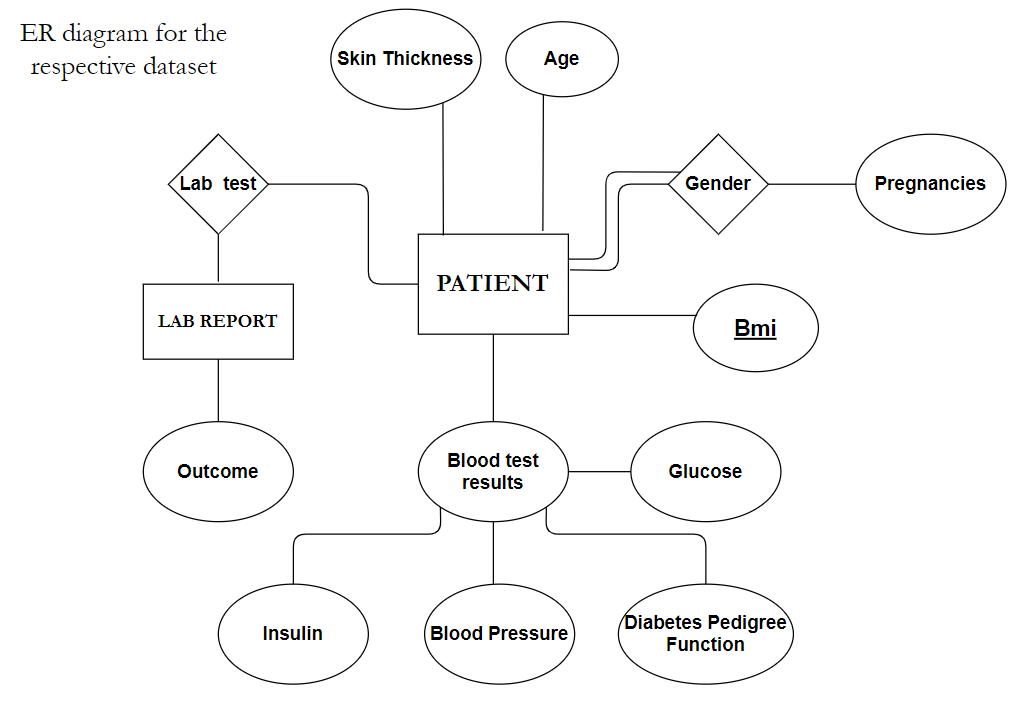
**Table 7:**

**Table 6:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pregnancies**  **(fk)** | **Glucose**  **(fk)** | **Blood**  **Pressure**  **(fk)** | **Skin**  **Thickness** | **Insulin**  **(fk)** | **BMI** | **Diabetes**  **Pedigree** | **Age**  **(fk)** | **Outcome**  **(fk)** |
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

**Therefore, the Diabetes Database is Normalized.**

1. Draw the ER diagram for the final decomposed schema stating the key attributes, mapping cardinalities, participation constraint, and so on

  
KEY ATTRIBUTES

In the following ER diagram/table, key attributes are the constraints that help in identifying the table uniquely from others.

1. Primary Key - BMI is the primary key attribute here, as it should be different for each person

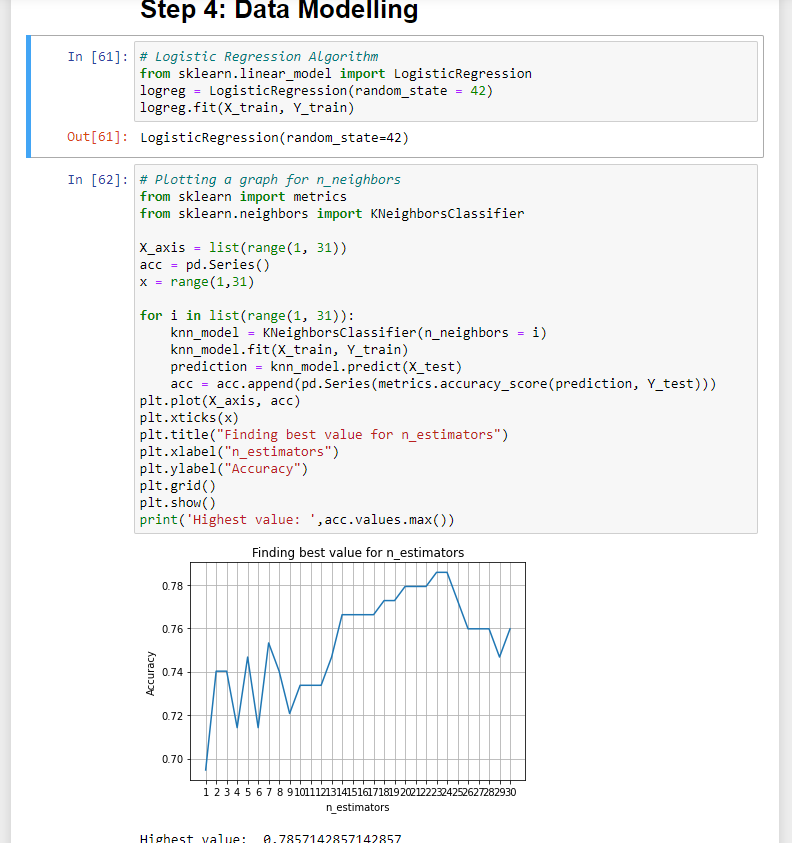
2. Candidate Keys - Insulin, Blood Pressure, Glucose are all candidate keys, as they ensure uniqueness when all taken together.

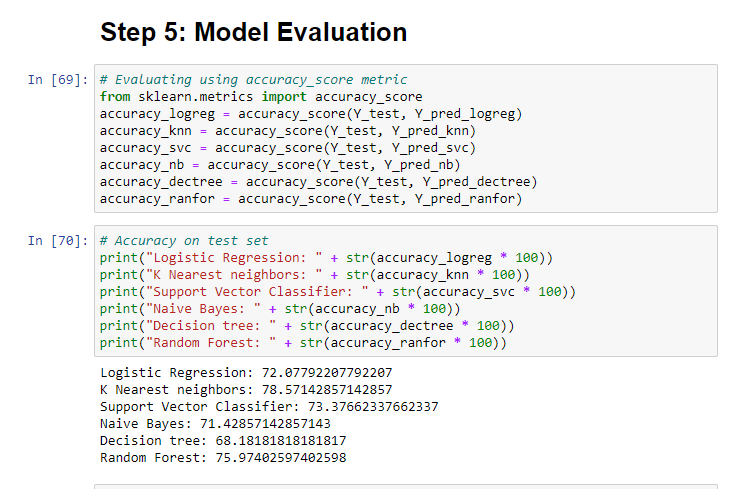
3. Alternate Key - Skin Thickness can be considered an alternate key, as two or more people rarely have the same depth of skin layer.

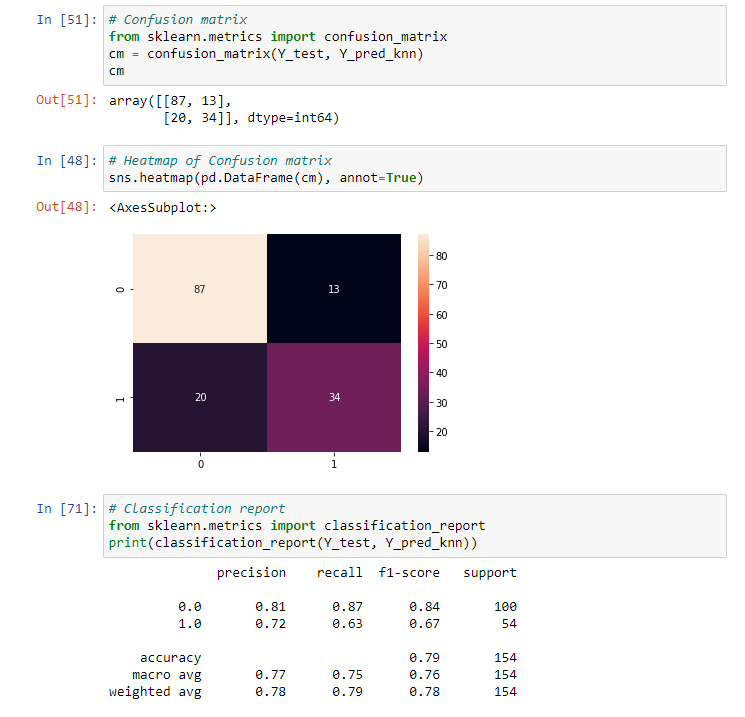
4. Foreign Key - Outcome is a foreign key as it is an attribute of another table LAB REPORT which is linked to the PATIENT table through the dependency of the test happening for diabetes.

There are two mapping cardinalities present here "Lab Test" and "gender", both of many to many kind, meaning entities are related to each other in any number of ways.

1. **Methodology and Algorithm used:** [Explain the suitable methodology u have identified with justification for choosing the same. Also specify the tentative algorithms [Min 2] to be used]  
     
   For the purpose of this project, we have employed the use of the **k-nearest neighbours** algorithm, as it produces the best accuracy with the given dataset.   
   **K-nearest neighbours** is a supervised machine learning algorithm that can be used in both classification and regression tasks. It works on the assumption that when all the datapoints are mapped onto a plane, similar datapoints will be grouped together. The distance between such datapoints is found, and it groups n such neighbours that are closest to each other together.   
     
   Other algorithms that are used include:  
   1. **Naïve Bayes algorithm**, which is a classification technique based on Baye’s Theorem with an assumption of independence among predictors. The classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.   
   2. **Support Vector Classifier**, which creates the best line of decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.   
   3. **Decision tree algorithms**: A decision tree is a structure that includes a root node, branches and leaf nodes. Each internal node denotes a test on an attribute. Each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.  
   4. **Random forest classifier**: A bunch of decision trees with the root nodes randomized.   
   5. **Logistics regression** is a statistical analysis method used to predict a data value based on prior observations of a data set. A logistic regression model predicts a dependant data variable by analysing the relationship between one or more existing independent variables.
2. **Implementation**Our first step is to find out at what n value is the KNN algorithm most accurate. For this, we run a for loop and plot the accuracies against n values and we find that accuracy is at its maximum value at n = 24.

  
With this knowledge, we train our primary KNN algorithm with n = 24, using our training set. Similarly, we train all the other algorithms used. All these algorithms are found in the sklearn package.   
Once the algorithms are trained with the training set, we then use the predict function for each of the algorithms and find the accuracy score. Here, we do indeed find that KNN has the highest accuracy at 78.57%. Hence, we continue to the results while using the KNN algorithm.   

1. **Results and Discussion**As mentioned in the previous section, using the KNN algorithm, we find an overall accuracy of 78.57%.   
   A classification report generated using functions from the sklearn package reveal that the precision of [not having diabetes] is 81% and the precision of [having diabetes] is 72%. Overall, the f1 scores for the two are .84 and .67, which are favourable results.  
     
   The confusion matrix visualises the precision results.   
   The accuracy of the algorithm depends on the datasets and random biases that can be introduced when splitting the dataset into the training and testing parts. Additionally, since this is a real-life example, there is a level of unpredictability that is a result of human nature, hence resulting in a less than 100% precision. However, for the purposes of a fairly accurate prediction, the objective of the project has been achieved.
2. **Conclusions**The automatic diagnosis of diabetes is an important real-world medical problem. Detection of diabetes in its early stages is the key for treatment. This paper shows how Classification Algorithms are used to model actual diagnosis of diabetes for local and systematic treatment, along with presenting related work in the field. Experimental results show the effectiveness of the proposed model. The performance of the techniques was investigated for the diabetes diagnosis problem. Experimental results demonstrate the adequacy of the proposed model.

In future it is planned to gather the information from different locales over the world and make a more precise and general prescient model for diabetes conclusion. Future study will likewise focus on gathering information from a later time period and discover new potential prognostic elements to be incorporated. The work can be extended and improved for the automation of diabetes analysis.