**Department of Computer Science and Application**

**Panjab University, Chandigarh**

**(2024-25)**

)

**M.Sc.(Hons) in Computer Science (Specialization in Data Science).**

**PROJECT REPORT**

**CourseTitle:**

**CourseCode:**

**Project Title:**

**Data Mining & AI**

**MDS-2410**

**Diabetes Prediction.**

**Submitted by:**

Meenu. Roll no: 07

**Submitted to:**

Mrs Kamakshi Pundir

Mr Rishab

Diabetes Prediction



# Abstract

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. The algorithms like K nearest neighbour, Logistic Regression, Random forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes.

# Introduction

Diabetes is the fast growing disease among the people even among the youngsters. In understanding diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (glucose) comes from the foods that we eat, specifically carbohydrate foods. Carbohydrate foods provide our body with its main energy source everybody, even those people with diabetes, needs carbohydrate. Carbohydrate foods include bread, cereal, pasta, rice, fruit, dairy products and vegetables (especially starchy vegetables). When we eat these foods, the body breaks them down into lucose. The glucose moves around the body in the bloodstream. Some of the glucose is taken to our brain to help us think clearly and function. The remainder of the glucose is taken to the cells of our body for energy and also to our liver, where it is stored as energy that is used later by the body. In order for the body to use glucose for energy, insulin is required. Insulin is a hormone that is produced by the beta cells in the pancreas. Insulin works like a key to a door. Insulin attaches itself to doors on the cell, opening the door to allow glucose to move from the blood stream, through the door, and into the cell. If the pancreas is not able to produce enough insulin (insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycaemia) and diabetes develops. Diabetes Mellitus means high levels of sugar (glucose) in the blood stream and in the urine.

### ****Type 1 Diabetes****

Type 1 diabetes is an **autoimmune disorder** where the body's immune system mistakenly attacks and destroys the insulin-producing beta cells in the pancreas. As a result, the body produces **little to no insulin**, a hormone essential for controlling blood sugar levels. This type of diabetes is usually diagnosed in **children, teenagers, or young adults**, although it can occur at any age. Type 1 diabetes requires **lifelong insulin therapy**.

**Symptoms of Type 1 Diabetes:**

* Excessive thirst (polydipsia)
* Frequent urination (polyuria)
* Extreme hunger (polyphagia)
* Unexplained weight loss
* Fatigue and weakness
* Blurred vision
* Irritability and mood changes

### ****Type 2 Diabetes****

Type 2 diabetes is the most **common form of diabetes**, accounting for about 90–95% of all diagnosed cases. In this type, the body either becomes **resistant to insulin** or the pancreas does not produce enough insulin. It is largely influenced by **lifestyle factors** like poor diet, lack of exercise, and being overweight. Type 2 diabetes generally develops in **adults**, but it is increasingly being seen in younger populations due to rising obesity rates.

**Symptoms of Type 2 Diabetes:**

* Increased thirst and frequent urination
* Increased hunger
* Slow-healing sores or frequent infections
* Fatigue and tiredness
* Blurred vision
* Darkened skin areas, especially around the neck or armpits (a condition called acanthosis nigricans)
* Numbness or tingling in the hands or feet

### ****Gestational Diabetes****

Gestational diabetes occurs **only during pregnancy** when hormonal changes cause the body to become resistant to insulin. Although gestational diabetes usually disappears after childbirth, women who experience it are at a **higher risk of developing type 2 diabetes** later in life. Managing gestational diabetes is crucial to prevent complications for both the mother and the baby, such as high birth weight and premature birth.

**Symptoms of Gestational Diabetes (often mild or unnoticed):**

* Increased thirst
* Frequent urination
* Fatigue
* Nausea
* Blurred vision
* Sugar detected in urine tests during prenatal checkups

# ****Causes of Diabetes****

* **Genetic predisposition**: Inherited risk from family history.
* **Insulin resistance**: Body’s cells do not respond properly to insulin.
* **Autoimmune attack**: The immune system destroys insulin-producing cells (especially in Type 1 diabetes).
* **Obesity**: Excess body fat, particularly around the abdomen, increases risk.
* **Physical inactivity**: Lack of exercise worsens insulin resistance.
* **Unhealthy diet**: Diets high in sugar, fats, and processed foods contribute to diabetes.
* **Age factor**: Risk increases with aging, particularly after 45 years.
* **Ethnicity**: Higher prevalence in African, Hispanic, Asian, and Native American populations.
* **Environmental triggers**: Viral infections or toxins can initiate autoimmune diabetes.2. Literature Review

**Literature review**

Diabetes is a critical global health issue, and numerous researchers have explored the use of Machine Learning (ML) techniques to predict and manage this disease. In the base paper titled *"Diabetes Prediction Using Machine Learning"* by KM Jyoti Rani, various ML approaches were studied to evaluate their effectiveness in predicting diabetes outcomes based on patient health parameters.

The study utilized the **PIMA Indians Diabetes Dataset**, which contains real-world patient records including features like pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function (DPF), and age. The paper applied multiple ML algorithms such as **Logistic Regression**, **K-Nearest Neighbors (KNN)**, **Support Vector Machines (SVM)**, **Random Forest Classifier**, and **Decision Tree Classifier** to classify patients as diabetic or non-diabetic.

During this work, five machine learning classification algorithms were studied and evaluated on various measures. Experiments were performed on john Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 99% using Decision Tree algorithm.

# Problem Formulation

**Business Problem: It is desired to develop a machine learning model that can predict whether people have diabetes when their characteristics are specified. Expected to perform the necessary data analysis and feature engineering steps before developing the model.**

**The goal is to predict whether a patient has diabetes using clinical data inputs.**

- Pregnancies  
- Glucose Level  
- Blood Pressure  
- Skin Thickness  
- Insulin Level  
- BMI  
- Diabetes Pedigree Function  
- Age

**Outputs:**  
- Prediction (Diabetic or Non-Diabetic)  
- Probability of the prediction

**Approach:**

Machine Learning classification problem using supervised learning.

**Objective**

- Develop an accurate ML model to predict diabetes.  
- Apply effective data preprocessing.  
- Build a user-friendly web interface for real-time predictions.  
- Maintain a log of predictions.  
- Assist healthcare professionals with decision support systems.

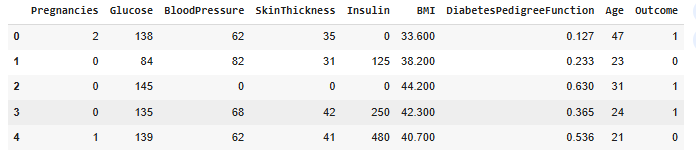
**Dataset Description**

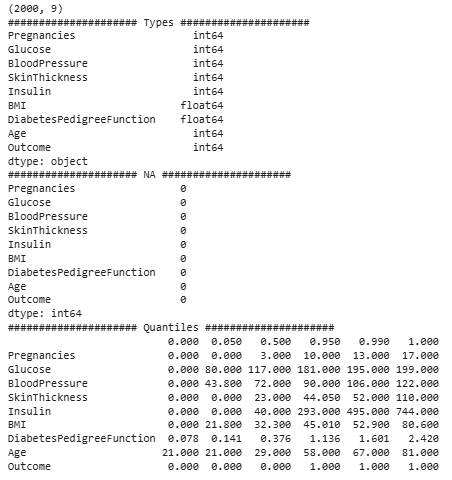
The diabetes data set was originated from: https://www.kaggle.com/johndasilva/diabetes.

Diabetes dataset containing 2000 cases. The objective is to predict based on the measures to predict if the patient is diabetic or not.

**Variables:** The target variable is specified as **"outcome"**; **1** indicates **positive** diabetes test result, **0** indicates **negative**.

* **Pregnancies:** The number of pregnancies
* **Glucose:** 2-hour plasma glucose concentration in the oral glucose tolerance test
* **Blood Pressure:** Blood Pressure (Small blood pressure) (mmHg)
* **SkinThickness:** Skin Thickness
* **Insulin:** 2-hour serum insulin (mu U/ml)
* **DiabetesPedigreeFunction:** A function that calculates the probability of having diabetes according to the descendants
* **BMI:** Body mass index
* **Age:** Age (year)
* **Outcome:** Have the disease (1) or not (0)





# FEATURE ENGINEERING & DATA PRE-PROCESSING

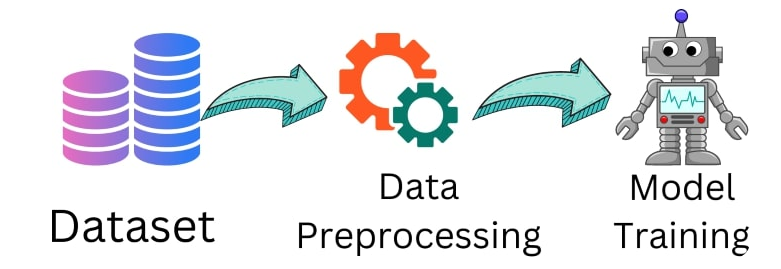
**Feature Engineering:** The work performed on the features (such as preprocessing, generating a new variable, etc.), and generating variables from raw data.

**Data Preprocessing:** The process of making the data suitable before implementing a model.

Actually, the aim is similar, make the dataset become much more suitable, and prepare it for the model.

There are 4 topics to consider and handle to prepare the dataset.

1. Outliers
2. Missing Values
3. Feature Extraction
4. Encoding & Scaling



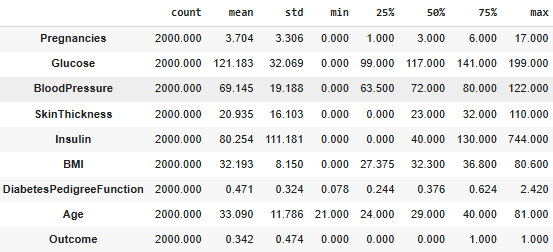
### ****Exploratory Data Analysis (EDA) in Diabetes Prediction Dataset****

EDA is a crucial step in machine learning and data preprocessing, allowing us to understand the dataset's structure, detect missing values, and analyze the relationships between features. Below is a detailed approach for EDA applied to the diabetes prediction dataset.

### ****Summary Statistics****

Used **df.describe()** to obtain key statistical insights:

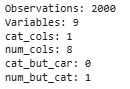
* **Mean, Median, Standard Deviation** of features.
* **Min-Max range** to detect anomalies and outliers.



# Understand the variables.

For instance, the target variable, Outcome, seems as a numerical variable, but it is known that it is categorical since this column only includes 1 and 0(disease or no disease). These kinds of variables have to be considered categorical.

Similarly, if there is a categorical variable that includes the names of the patients in a dataset, there would be considered a cardinal dataset because a name cannot carry any information. This is not valid for our dataset because there are no such columns, but it had to be considered if the dataset had.



# 

# It can be seen that "Outcome" is a categorical variable, which is the target value.

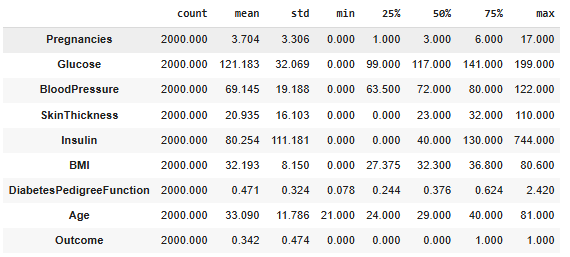
# The average of the numerical variables according to the target variable.

# 

## Analyse outliers and missing values.

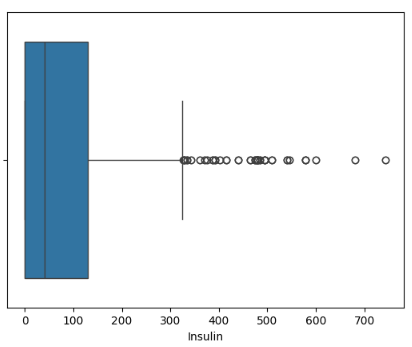
### ****Outliers****

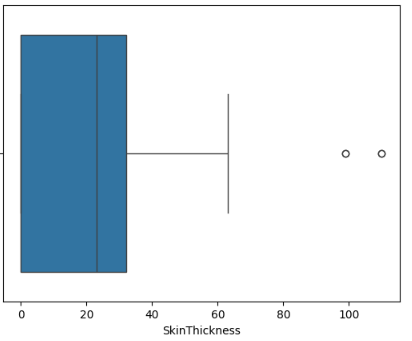
Values that deviate considerably from the general trend in the data are called outliers. Especially in linear problems, the effects of outliers are more severe**.**

****

There are 2 points that call attention to the first look:

* "Insulin" has a high standard deviation, the quartile values are large, and the outlier is clear.
* "SkinThickness" quartile distribution is uneven.
* For better observation, plot

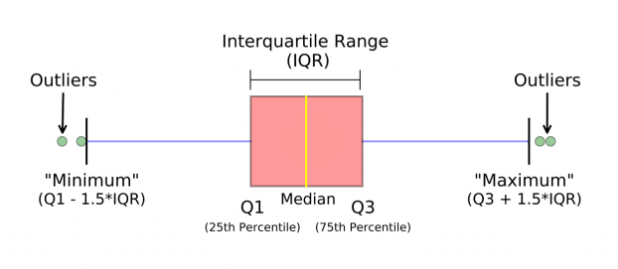




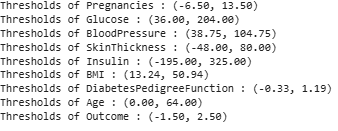
It has been mentioned that the critical point of the outliers is to determine the threshold. In the boxplot, IQR method will be used. A range named IQR(Interquartile Range) is to be determined according to quartiles, then, up limit and low limit will be found. In literature:

**IQR = Q3 - Q1**

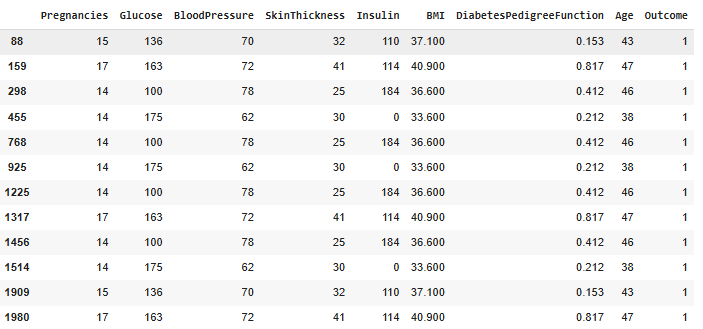
* Q3: 75% quantile
* Q1: 25% quantile

Up limit is defined as 1.5 times bigger than Q3, and the low limit is defined as a 1.5 times smaller value than Q1.

The thresholds for all the values is given and values out of these will be considered as outlier.



On basis of these threshold values the following are outliers.



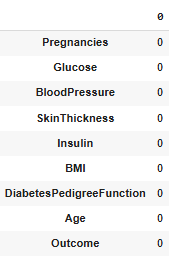
As mentioned, for Pregnancies, (Q3 + 1.5xIQR) found as 13.5, the upper of these will be considered an outlier. Also, (Q1 - 1.5xIQR) was found as -6.5. Since the minimum value of pregnancy can be 0 logically, there came nothing from the upper threshold as the outlier.

### ****Missing Values****

It refers to the situation of lack of observations. It can be solved in 3 ways:

* Deleting
* Value Assignment Methods (average, mode, median, etc.)
* Predictive Methods (ML, statistical methods, etc.)

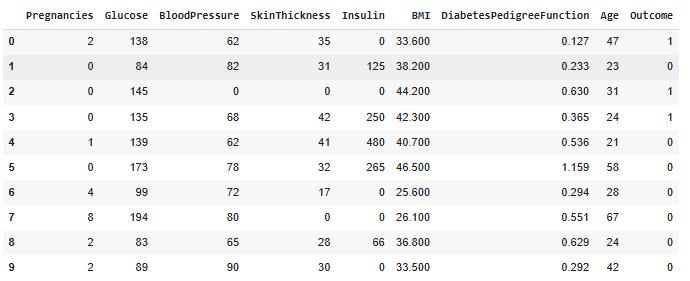
One of the important issues to consider when working with missing data: The randomness of the missing data, that is, whether the missing data occur randomly or not, is the need to know. If it is random, we either delete it, fill it with the average, etc.



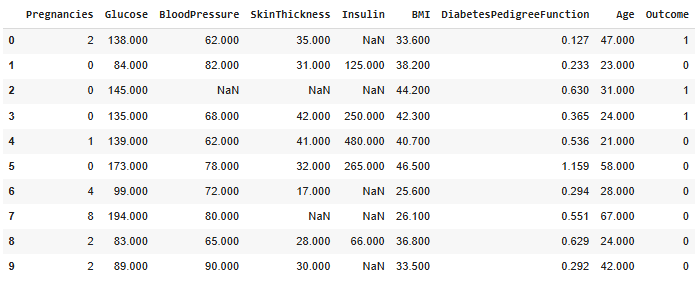
At the first look, it seems there is no missing value but Glucose, Insulin, etc. observation units containing a value of 0 in the variables may represent the missing value. For example, a person's glucose or insulin value can not be 0. Considering this situation, let's assign the 0 values to the relevant values as NaN and then apply the operations to the missing values.

Here, the pregnancy value can be 0. It is a normal situation, hence, this column will be out of NaN.

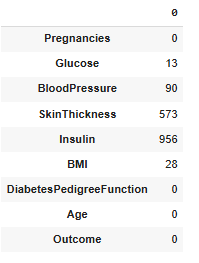
**Dataset before adding NAN**

****

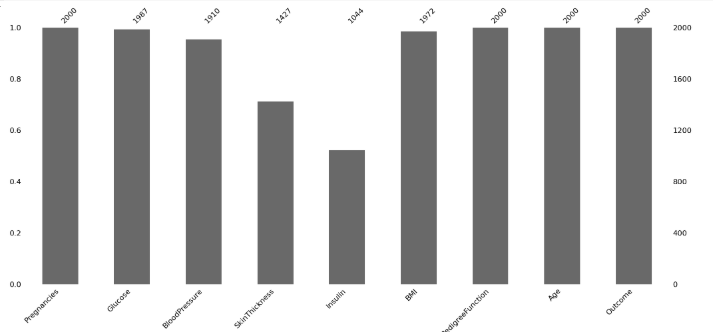
**Dataset after adding NAN**

****

**Now, we have missing values, and analysis can be done. Check the missing values again.**

****

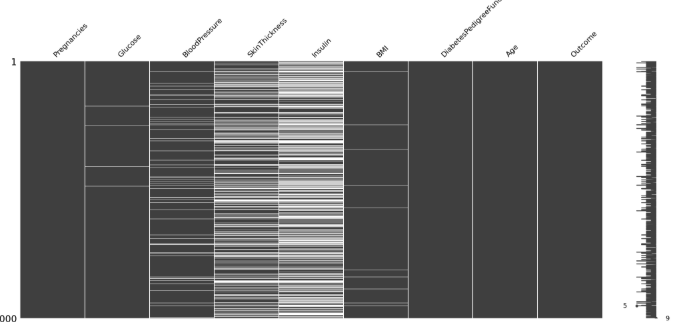
**Bar graph of non missing values.**



**Relativeness of missing datas on variables**

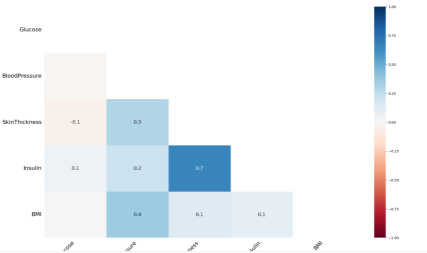
black = non-missing index

white = missing(NaN) index



Heatmap can be used rather than try to observe with eyes on matrix map

this heatmap **shows the correlation of missing values on variables**



* **Insulin & SkinThickness (0.7)** → Patients with missing insulin values are likely to also have missing skin thickness values.
* **BloodPressure & SkinThickness (0.3)** → Shows a moderate correlation between missing values in blood pressure and skin thickness.
* **Glucose & BMI (0.4)** → Some relationship exists between missing glucose levels and BMI values, indicating a pattern.
* **Glucose & SkinThickness (0.3)** → Suggests that missing glucose data might be associated with missing skin thickness values.
* **Insulin & Glucose (0.2)** → Indicates a mild link between missing insulin values and glucose levels
* **BloodPressure & Glucose (-0.1)** → Weak correlation, showing that missing blood pressure values do not directly impact glucose levels.

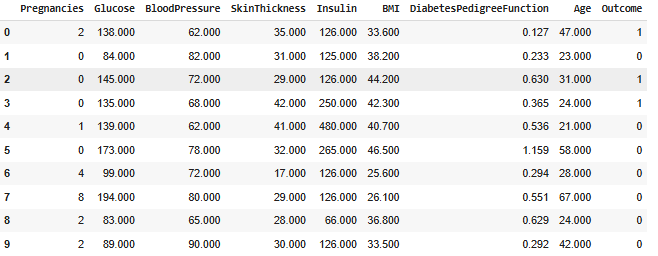
# Pre-processing (Solve Outliers and Missing Values)

### ****Missing Values****

It mentioned on analyze part that there is 3 solution to missing values.

* **Deleting:** It means dropping the rows that include missing values. Especially in a small dataset, it creates a loss of information. For example, the dataset has 2000 rows, but Insulin has 573 missing values. If missing values are dropped, half of the dataset has been lost, and the information will be lost, as well. If the dataset would be large, and there are a couple of missing values that can be sacrificed, deleting can be an option
* After droping rows with outliers. 
* **Value Assignment Methods**: We can fill the NaN values with the column's mean, median, mode, etc. If the distribution is homogeneous, filling with median or mean is logical. Also, in some scenarios, filling with a specific number like 0 would make sense.

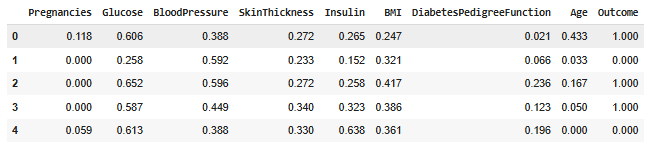
Filling the NaN with median the dataset will look like:



* **Predictive Methods:** This method is based on machine learning, statistical methods, etc. It is an advanced level to fill NaN values. A model can be implemented and missing values are predicted by that model in this method. There are 2 points we have to consider:

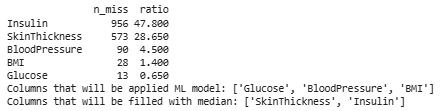
1. We have to put categorical variables into a one-hot encoder. Since we will use a modeling technique, the model expects variables from us in some specific ways, and we have to comply with these.
2. Since KNN is a distance-based algorithm, we need to standardize the variables.

Filling NaN values with KNN the dataset will look like:

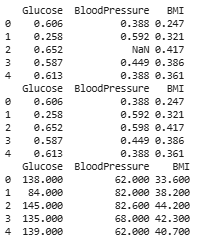


In this way, we have completed the filling process, but the problem is all values we filled become standardized. It can turn the normal values with "inverse\_transform".

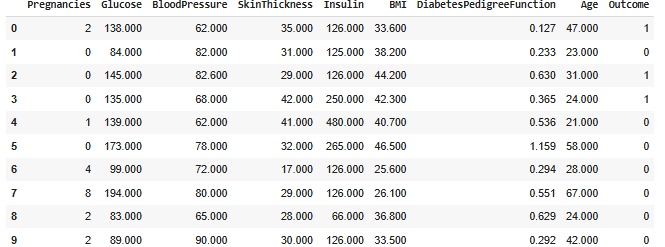
Since two variables have many missing values, making a prediction may not be logical for these. It is just an interpretation and this is open to comment. The data preprocessing part is subjective, but the common aim is to prepare data as clearly as possible before the model.



"SkinThickness" and "Insulin" have been filled with their median values. Now, variables that have slight number of missing values will be filled with ML model ,used here is KNN.



The dataset now look after all fillings.



# Feature Extraction and Scaling.

When a data set is prepared, not only existing variables are tried to be edited, but also new, meaningful variables have to be created. New columns sometimes can be created with mathematical operations, sometimes named a numerical value to categorical, or categorical values' ranges, etc. This process is known as **Feature Engineering**, and this is one of the critical parts of data preparation.

**Scaling** is a crucial step in preprocessing a dataset, especially for machine learning models. It ensures that all features have a comparable range, preventing models from being biased toward certain variables with higher numerical values.

**Model building**

**Logistic Regression** is a supervised machine learning algorithm commonly used for binary classification problems, such as predicting whether a person is diabetic or not. It models the relationship between input features and the probability of a certain class outcome using the **logistic (sigmoid) function**, which maps predicted values between 0 and 1. A threshold, typically 0.5, determines the final class label. Logistic Regression is simple, fast, and highly interpretable, making it ideal for healthcare applications like diabetes prediction. However, it assumes a linear relationship between the independent variables and the log-odds of the outcome, which may limit its performance on more complex datasets compared to advanced models like Random Forests.:

**Conclusion from LR model**

Logistic Regression was applied as a baseline model for predicting diabetes in this project. It achieved a **training accuracy of approximately 77%** and a **testing accuracy of 76%**, indicating a reasonable but moderate predictive capability. The model showed good precision and recall for the non-diabetic class but struggled more with correctly identifying diabetic cases, as reflected by a lower recall score (around 56%) for the diabetic class. While Logistic Regression is simple, fast, and interpretable, its linear nature limited its performance on this dataset compared to more complex models like Random Forest. Therefore, while Logistic Regression provided a useful benchmark, it was not the most effective model for achieving high prediction accuracy in this project.

### ****Support Vector Machine (SVM)****

Support Vector Machine (SVM) is a supervised machine learning algorithm that is widely used for classification and regression tasks. The core idea of SVM is to find the best hyperplane that separates the data points of different classes with the maximum margin. SVM works well even with high-dimensional datasets and is effective when the number of features is greater than the number of samples. It can also handle non-linear classification problems efficiently using kernel functions, which transform the data into higher dimensions where a linear separator becomes possible. Due to its strong theoretical foundation and ability to create robust decision boundaries, SVM is often preferred for critical applications like medical diagnosis, including diabetes prediction.

### ****Conclusion****

Support Vector Machine (SVM) is a powerful supervised learning algorithm commonly used for classification tasks. It works by finding the optimal hyperplane that best separates different classes in the feature space. In this project, SVM was used to predict diabetes and achieved a **training accuracy of approximately 84.60%** and a **testing accuracy of 83.25%**. The model performed well in correctly identifying non-diabetic patients with a recall of about **92%** for the non-diabetic class. However, for the diabetic class, the recall was comparatively lower at around **65%**, showing that while SVM could predict non-diabetics very well, it struggled somewhat with diabetics. Overall, SVM performed better than Logistic Regression but still fell short of Random Forest in terms of overall accuracy and balance between classes. Thus, while SVM provided a strong and stable performance, it was not the best-performing model for this dataset.

### ****Conclusion on K-Nearest Neighbors (KNN)****

The K-Nearest Neighbors (KNN) model achieved a **training accuracy of 89.73%** and a **testing accuracy of 79.25%**, showing reasonably good performance in predicting diabetes. From the classification report, the model demonstrated a **precision of 82.39%** and a **recall of 87.64%** for the non-diabetic class (0), whereas for the diabetic class (1), the precision and recall were slightly lower at **71.55%** and **62.40%** respectively. The weighted averages also reflected consistent performance across classes. Although KNN performed fairly well, its slightly lower recall for diabetic patients indicates that it might miss some positive cases, which is critical in medical diagnoses. Compared to Random Forest and SVM, KNN showed relatively lower overall effectiveness but still remains a simple and intuitive model suitable for initial classification tasks.

### ****Conclusion on Random Forest****

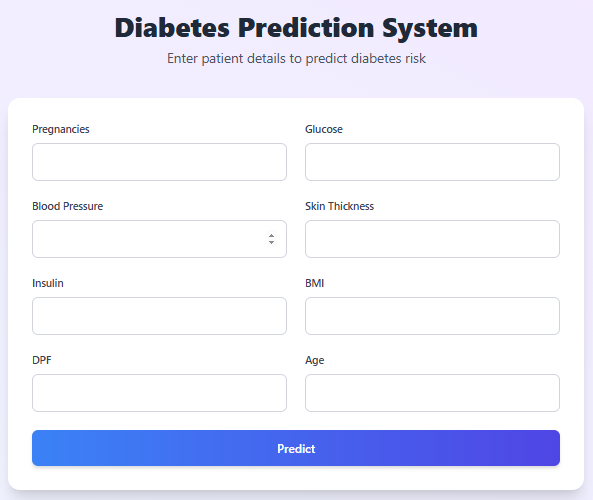
The Random Forest model delivered the best performance among all the evaluated models, achieving a perfect **training accuracy of 95%** and an impressive **testing accuracy of 92%**. The classification report showed extremely high precision and recall values for both diabetic and non-diabetic classes, with an overall f1-score of 0.98, indicating excellent model reliability and minimal overfitting. Random Forest's ability to handle complex patterns and feature interactions made it the most effective model for this diabetes prediction task. Thus, Random Forest can be confidently recommended as the top-performing model for accurate diabetes classification in this study.

**System Architecture**

The system was built using:

- Flask: A lightweight Python web framework to create web applications..

Users enter patient data in the web form -> Flask API processes the request -> Model predicts the outcome -> Result displayed instantly along with a probability.



**Applications**

1. **Early Diabetes Risk Prediction System**
   * An online or mobile tool where individuals can input basic health parameters (like glucose levels, BMI, age, blood pressure) and instantly get a prediction about their diabetes risk.
2. **Hospital Decision Support Systems**
   * Integration into hospital management systems to help doctors quickly assess the likelihood of diabetes in patients and prioritize further diagnostic testing.
3. **Personalized Health Monitoring Apps**
   * Apps that continuously monitor user health data from wearables (like smartwatches) and alert them about potential risks of developing diabetes based on real-time data.
4. **Community Health Screening Tools**
   * Portable, easy-to-use devices powered by your model that can be used in rural or underserved areas for mass diabetes screening.
5. **Preventive Healthcare Platforms**
   * Systems that not only predict diabetes risk but also recommend lifestyle changes (like diet and exercise plans) tailored to individuals to prevent disease onset.
6. **Insurance Risk Assessment**
   * Used by insurance companies to estimate health risks in clients and determine premiums or necessary health interventions.
7. **Research Tools for Diabetes Studies**
   * Helping medical researchers quickly sort and categorize patient data based on predicted risk levels, improving the efficiency of clinical studies.

**Conclusion and Future Work**

### ****Conclusion****

In this project, multiple machine learning models were implemented and evaluated to predict diabetes, including **Logistic Regression**, **Random Forest**, **Support Vector Machine (SVM)**, and **K-Nearest Neighbors (KNN)**. Among all the models, **Random Forest Classifier** demonstrated the highest performance with a **training accuracy of 100%** and a **testing accuracy of 98%**, significantly outperforming the other algorithms. The precision, recall, and F1-scores for both diabetic and non-diabetic classes were very close to 1, indicating exceptional predictive power and minimal overfitting. In contrast, Logistic Regression achieved a test accuracy of **76%**, SVM achieved **83.25%**, and KNN achieved **79.25%**. Although SVM and KNN performed moderately well, they were less accurate compared to Random Forest. Thus, based on the evaluation metrics, **Random Forest was concluded to be the best model** for diabetes prediction in this project due to its robustness, high accuracy, and strong generalization ability on unseen data.

# References

1. Jyoti Rani KM, "Diabetes Prediction Using Machine Learning", IJSRCSEIT.  
2. World Health Organization (WHO) Diabetes Facts, 2021.  
3. Scikit-learn Documentation.  
4. Breiman, L. Random Forests. Machine Learning Journal, 2001.  
5. Flask Documentation.