4th International Conference on Inventive Research in Computing Applications
ICIRCA 2022

# Diabetes Prognosis using Machine Learning

**Manuscript ID:** ICIRCA-584

**Authors:**

Aravinda Raman J, Rakshan Kotian

Manipal Institute of Technology, Manipal, India

# Introduction

- Diabetes Mellitus (‘Mellitus’ means ‘Sweet’ in Latin) is a disease that occurs when the blood glucose, also called blood sugar, is too high in a person’s body [1]. Blood glucose is the main source of energy and comes from the food one eats.

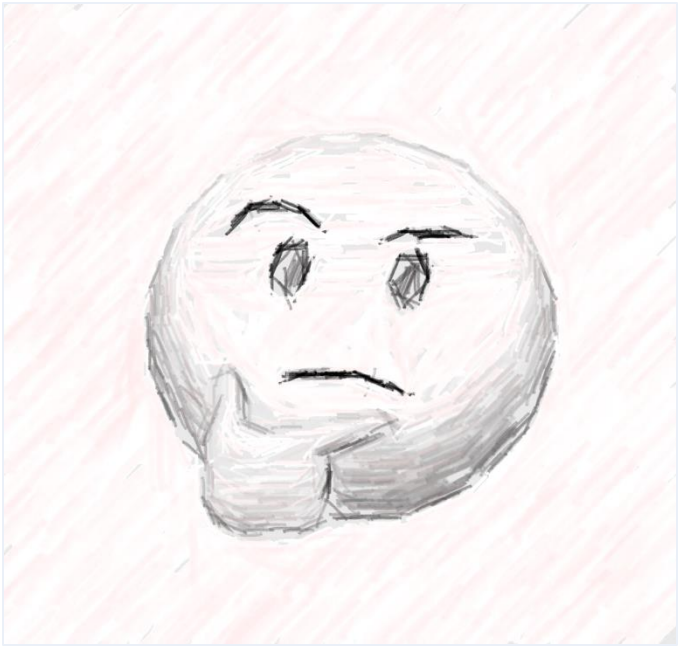
- Insulin, a hormone made by the pancreas, helps glucose from food get into the cells to be used for energy. Sometimes our body doesn’t make enough—or any—insulin or doesn’t use insulin well. Glucose then stays in the blood and doesn’t reach the cells [2].

- Cause of diabetes vary depending on the genetic makeup, family history, ethnicity, health etc..

- Over time, having too much glucose in one’s blood can cause health problems. Although diabetes has no cure as of now, one can take steps to manage it and stay healthy.

- Diabetes is also creator of different kinds of diseases like heart attack, blindness etc. [3]. The normal identifying process is that the patients need to visit a diagnostic center, consult a doctor and sit tight for a day or more to get their reports.

# Understanding of the Project



## Question

To Identify whether the patient is having diabetes or not.

## Goal

Goal of this project is to identify the probability of diabetes in patients using machine learning.

## Usefulness

The rules derived will be helpful for doctors to identify patients suffering from diabetes. Furthermore, predicting the disease early leads to treating the patient before it becomes critical.

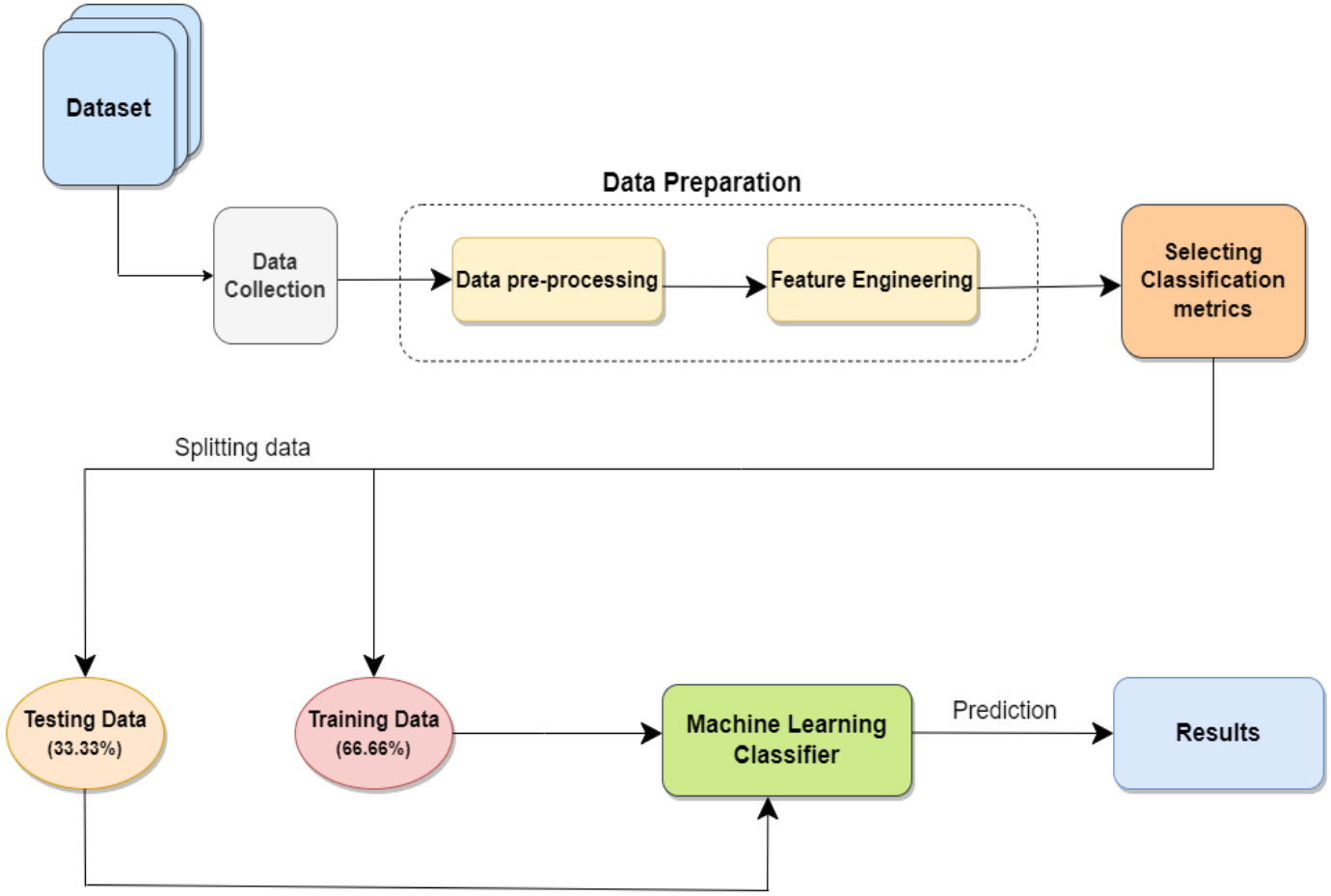
# Related works

| Authors | Work | Description |
|---|---|---|
| S. K. Reddy, T. Krishnaveni, G. Nikitha and E. Vijaykanth | Diabetes Prediction Using Different Machine Learning Algorithms [4] | Proposed the diabetes prediction model using two different classifiers, namely K Nearest Neighbor (KNN) and Random Forest (RF). They have split the data into 80% training and 20% test sets for this purpose. In this study, appropriate data preprocessing has not been performed, which has impacted the accuracy of the results. |
| P. Sonar and K. JayaMalini | Diabetes Prediction Using Different Machine Learning Approaches [5] | Provided a method to diagnose diabetes early on using various models, namely, SVM, KNN, Random Forest, Naive Bayes, and Artificial Neural Network (ANN). A training and test set ratio of 75% and 25% were used. The highest accuracy achieved with their experiment was 82% using ANN and SVM. |

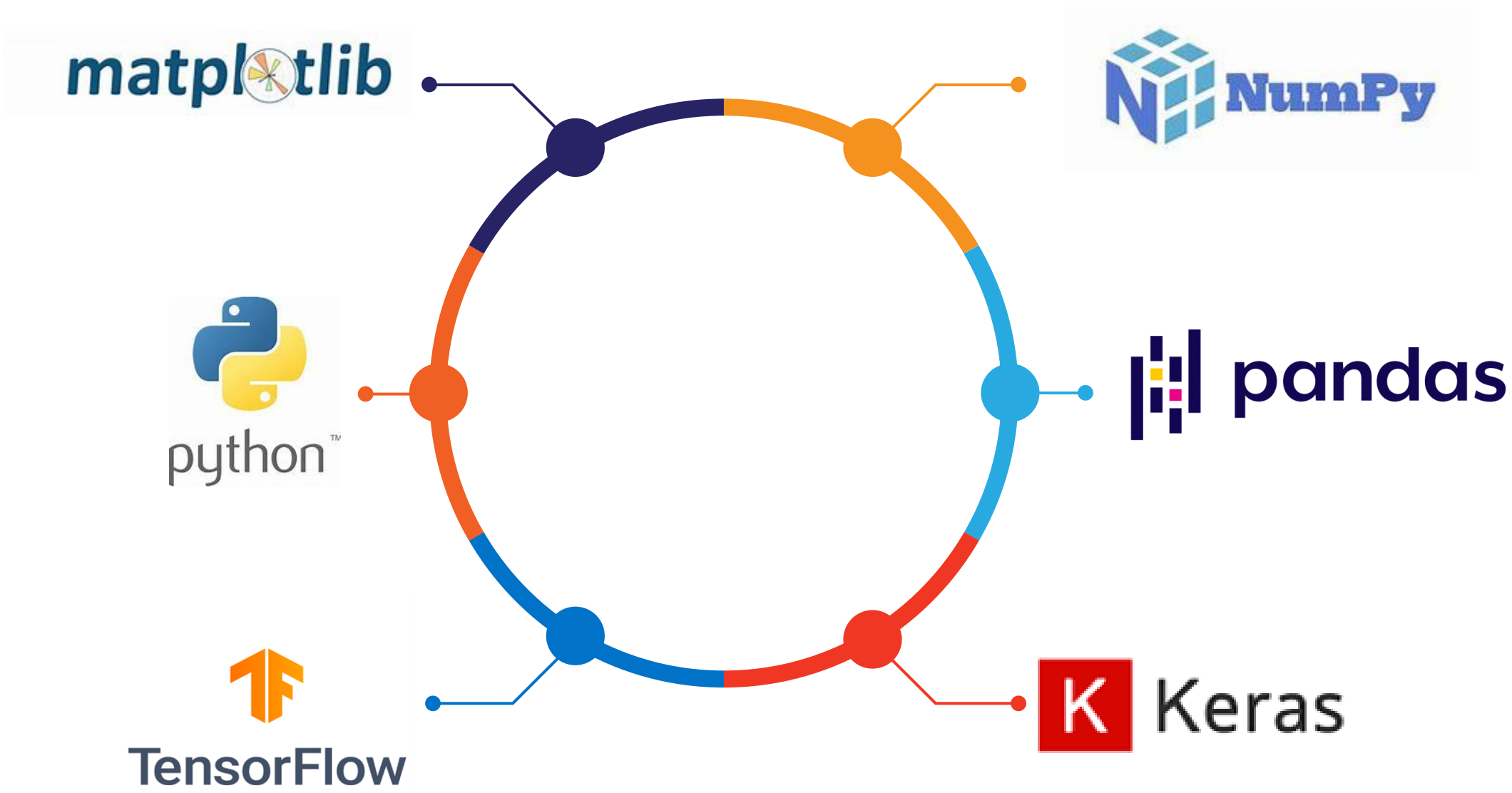
# Related works

| Authors | Work | Description |
| --- | --- | --- |
| A. C. Lyngdoh, N. A. Choudhury and S. Moulik | Diabetes Disease Prediction Using Machine Learning Algorithms [6] | Examined model overfitting and underfitting to explain why specific machine learning classifiers produce superior results. K-Fold Cross Validation was also incorporated for better accuracy results. Finally, optimum results in terms of time and an accuracy of 76% using the KNN classifier were obtained. |
| M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah | Prediction of Diabetes Using Machine Learning Algorithms in Healthcare [7] | Examined the accuracy using machine learning classifiers: Naive Bayes, KNN, SVM, Logistic Regression, Decision Tree, and Random Forest. They divided the data into training and testing data, with 70% and 30%, respectively. Both SVM and KNN models yielded an accuracy of 77%, which was the highest in that experiment. |

# Proposed Diabetes Prediction Method



# Tools and Technologies used



# Data Understanding

**Data Source:**

- The dataset used in this paper was obtained from the Kaggle “Pima Indians Diabetes Database.”
- Originally, it is from the National Institute of Diabetes and Digestive and Kidney Diseases.
- This dataset has 768 samples of diabetic and healthy individuals.
- In particular, all patients here are **females** of at least 21 years of age.

# Data Understanding

## Sample Data

| Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | 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 |
| 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 3 | 78 | 50 | 32 | 88 | 31 | 0.248 | 26 | 1 |
| 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |
| 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 |
| 8 | 125 | 96 | 0 | 0 | 0 | 0.232 | 54 | 1 |
| 4 | 110 | 92 | 0 | 0 | 37.6 | 0.191 | 30 | 0 |
| 10 | 168 | 74 | 0 | 0 | 38 | 0.537 | 34 | 1 |
| 10 | 139 | 80 | 0 | 0 | 27.1 | 1.441 | 57 | 0 |
| 1 | 189 | 60 | 23 | 846 | 30.1 | 0.398 | 59 | 1 |
| 5 | 166 | 72 | 19 | 175 | 25.8 | 0.587 | 51 | 1 |
| 7 | 100 | 0 | 0 | 0 | 30 | 0.484 | 32 | 1 |

# Attributes in the Dataset

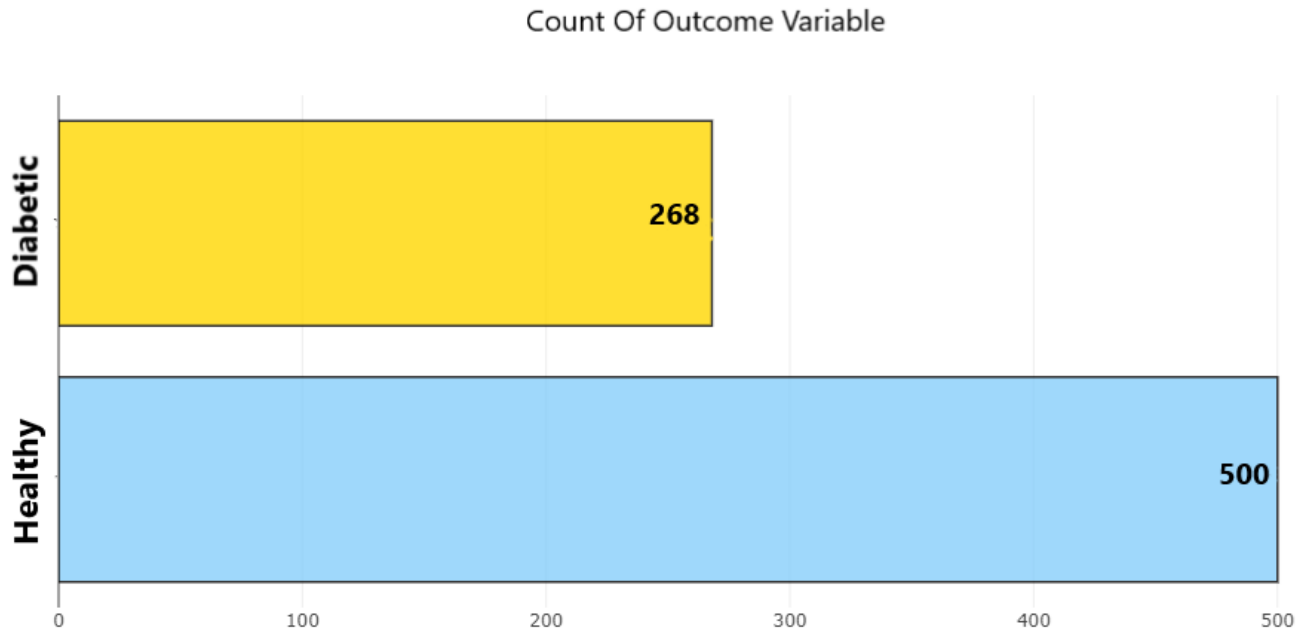
- The dataset consists of **9** attributes:

| Features | Description |
|---|---|
| Pregnancies | Number of pregnancies a patient has experienced. |
| Glucose | Glucose level in patients (mIU/L). |
| Blood Pressure | Diastolic Blood Pressure that is recorded at a particular time (mmHg). |
| Skin Thickness | Triceps skin fold thickness (mm). |
| Insulin | Amount of insulin present in the body (mIU/L). |
| Body Mass Index | BMI or Body Mass Index of an individual. |
| Diabetes Pedigree Function | Family history of diabetes disease. |
| Age | Age of an individual (years). |

- These **8** mentioned in the table above are independent attributes.

# Attributes in the Dataset

- ‘Outcome’ is the dependent variable which contains 268 ‘Diabetic’ and 500 ‘Healthy’ patient records.



# Data Pre-Processing

- The first step includes handling null values in the dataset.
- **Handling Zero/Null Values:** Missing values can be replaced with that particular attribute's mean, median, mode, or random variable.
- In this study, the missing values in the dataset were replaced with the median of their respective attributes as it proved to be more conducive for the model's prediction.

# Outlier Analysis

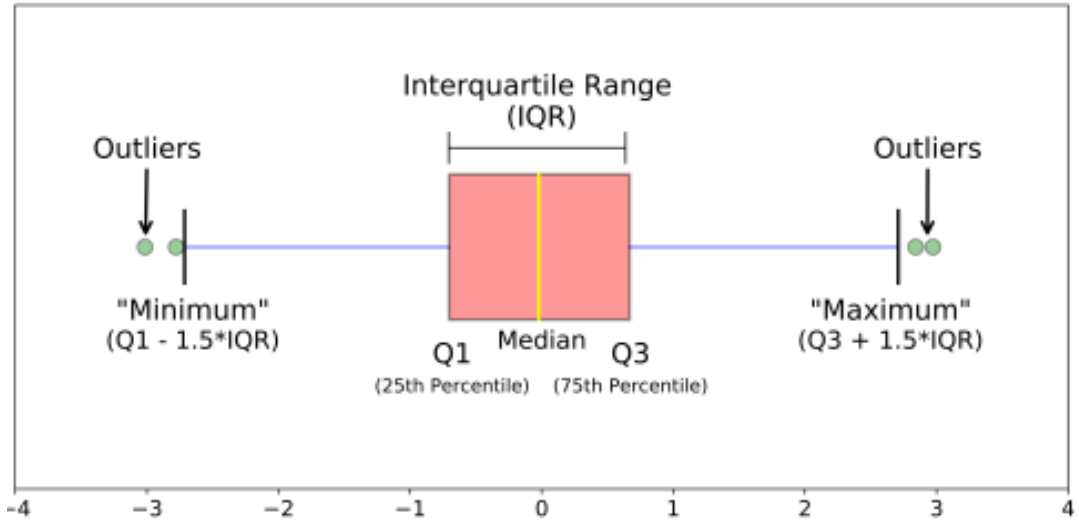
•In this step, outlier analysis and treatment has been performed for all the variables in the dataset.

• This step is crucial in order to avoid any bias in the prediction and so that it does not have a detrimental effect on the final prediction.

•Interquartile range (IQR) of each attribute was calculated individually, followed by finding the maximum and minimum limits.

•Any data value beyond maximum or below minimum were eliminated.

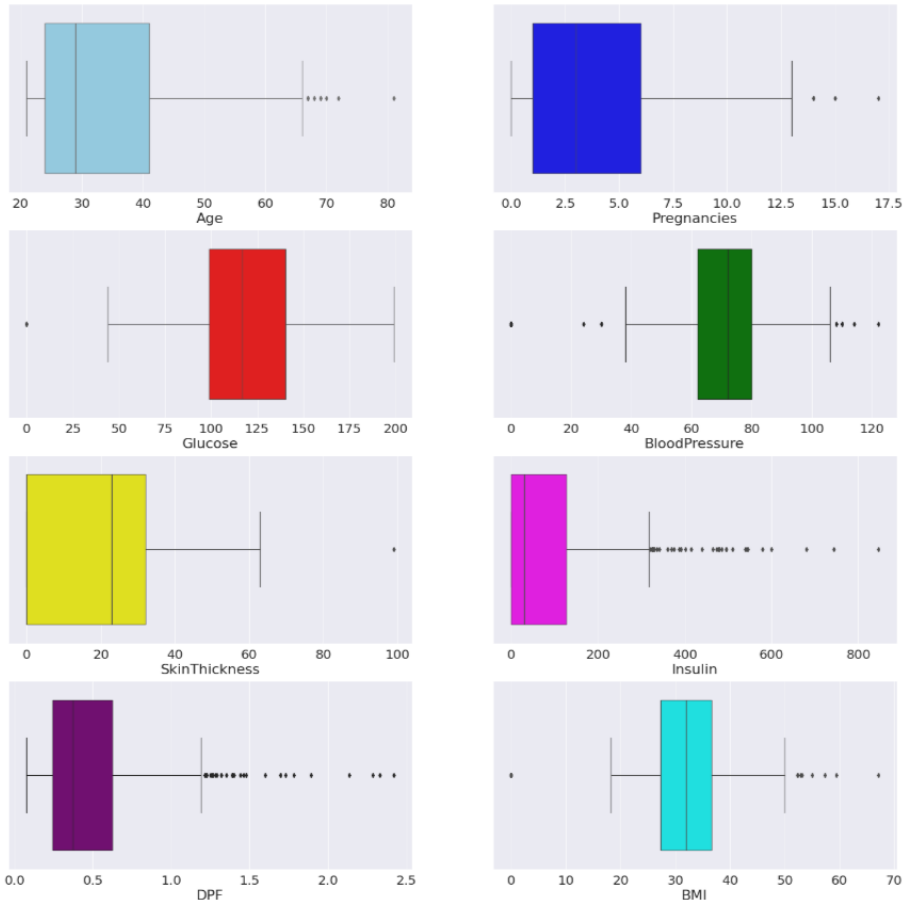
• A total of 732 values were remaining in the dataset after the removal of all the outliers.



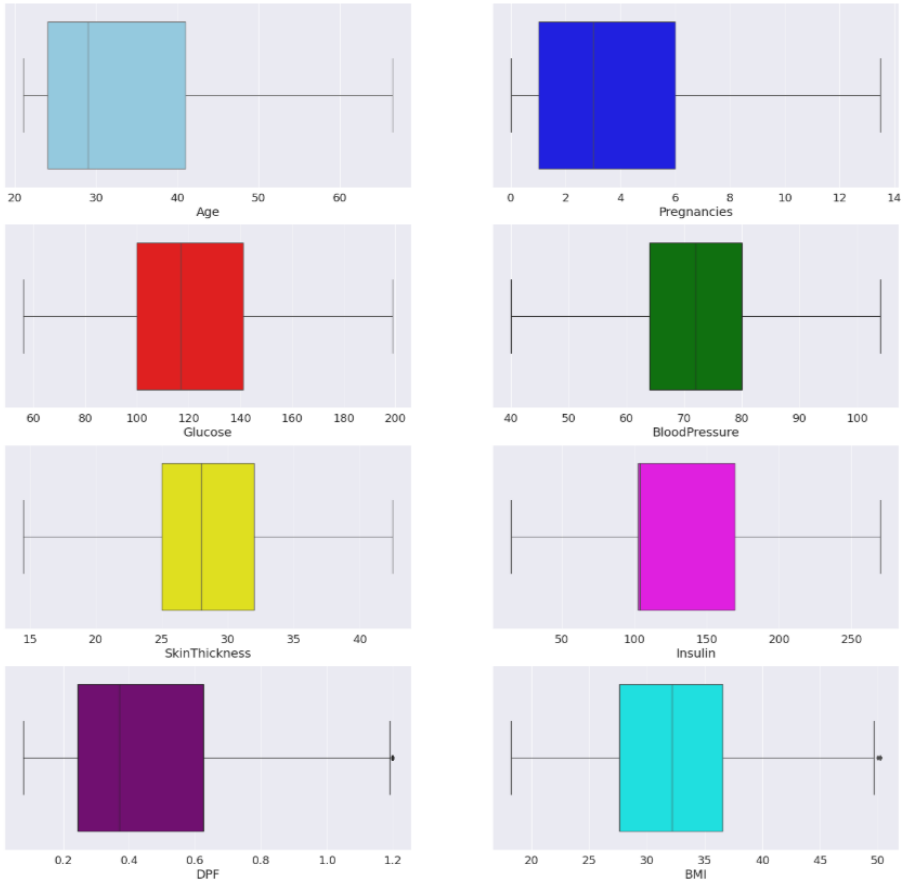

**Different Parts of a Boxplot [8]**

# Boxplots of all Variables

## Before removing the outliers




## After removing the outliers



# Categorizing Data

- Different ranges were found out for each continuous variable in the data set. Based upon these ranges, categorization was done.

**Glucose:** Glucose level in patients.

| Low | Normal | Prediabetes | Diabetes |
|---|---|---|---|
| ≤ 70 | > 70 and ≤ 99 | > 99 and ≤ 126 | > 126 |

**Blood Pressure (Diastolic):** Blood Pressure that is recorded at a particular time.

| Low | Normal | High |
|---|---|---|
| Below 60 | 60-90 | 90 or more |

**Skin Thickness:** Triceps skin fold thickness.

| Low | Normal | High |
|---|---|---|
| <23 | 23 | >23 |

# Categorizing Data

**Insulin:** Amount of insulin present in the body.

| Low | Normal | High |
|---|---|---|
| <16 mIU/L | 16-166 mIU/L | >166 mIU/L |

**BMI(weight in kg/ height in $m^2$):** BMI or Body Mass Index of an individual

| Under-weight | Normal | Over-weight | Obesity 1 | Obesity 2 | Obesity 3 |
|---|---|---|---|---|---|
| <18.5 | 18.5-24.9 | 25-29.9 | 30-24.9 | 35-39.9 | >39.9 |

**Diabetes Pedigree Function:** Family history of diabetes disease

| Low | Medium | High |
|---|---|---|
| 0-0.78 | 0.79-1.561 | >1.57 |

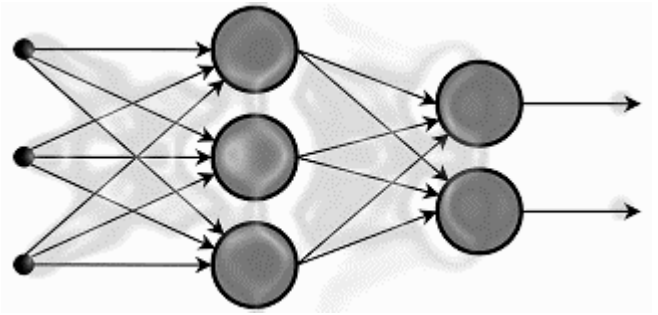
**Age:** Age of an individual in years

| Young | Adult | Old |
|---|---|---|
| 20-44 | 45-74 | 75-100 |

**No. of Pregnancies:** Number of pregnancies a patient has experienced.

| Normal | Above Normal | Highest |
|---|---|---|
| <6 | 6-12 | >12 |

# Modeling



- This phase includes application of appropriate model to the data.
- ‘Robust Scaler’ was used for scaling the data.
- Data was divided into 66.67% training and 33.33% testing data.
- The machine learning algorithms used in this study are:
  - **K Nearest Neighbor (KNN)**
  - **Random Forest Classifier**
  - **Artificial Neural Network (ANN)**

# Evaluation of the Results

## Confusion Matrix

| | Predicted Yes | Predicted No |
|---|---|---|
| **Actual Yes** | **Tp** (True positive) | **Fp** (False positive) |
| **Actual No** | **Fn** (False negative) | **Tn** (True negative) |

- **True positive (Tp):** These are cases in which the model correctly classified a patient as diabetic.
- **True negative (Tn):** Model correctly predicted patient as non-diabetic and they don't have the disease.
- **False positive (Fp):** Model predicted patient as diabetic, but they actually do not have the disease. (Also known as a "Type I error.")
- **False negative (Fn):** Model predicted patient as non- diabetic, when they actually are. (Also known as a "Type II error.")

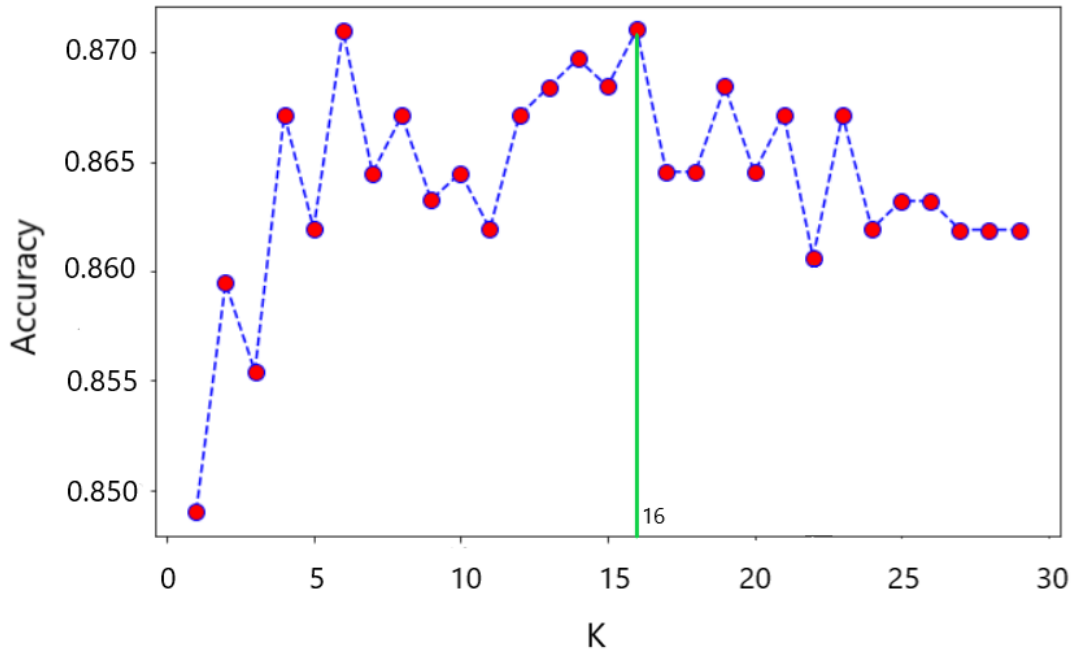
**Accuracy** = Tp + Tn / Tp + Tn + Fp + Fn

**Precision (P)** = Tp / Tp + Fp

**Recall (R)** = Tp / Tp + Fn

**F1 Score** = 2 P * R / P + R

# K Nearest Neighbor (KNN)



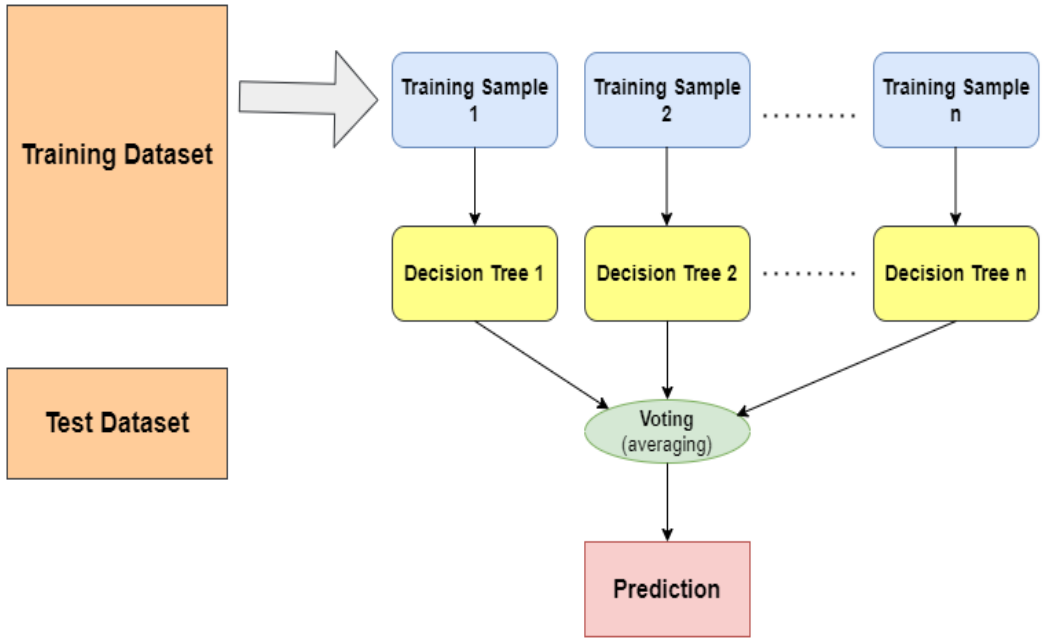

Selecting the appropriate K value for Training

| | Predicted Diabetes | Predicted No Diabetes |
|---|---|---|
| Actual Diabetes | 155 | 20 |
| Actual No Diabetes | 24 | 57 |

Confusion matrix for KNN

The Accuracy for KNN is 82.81%

# Random Forest Classifier (RF)



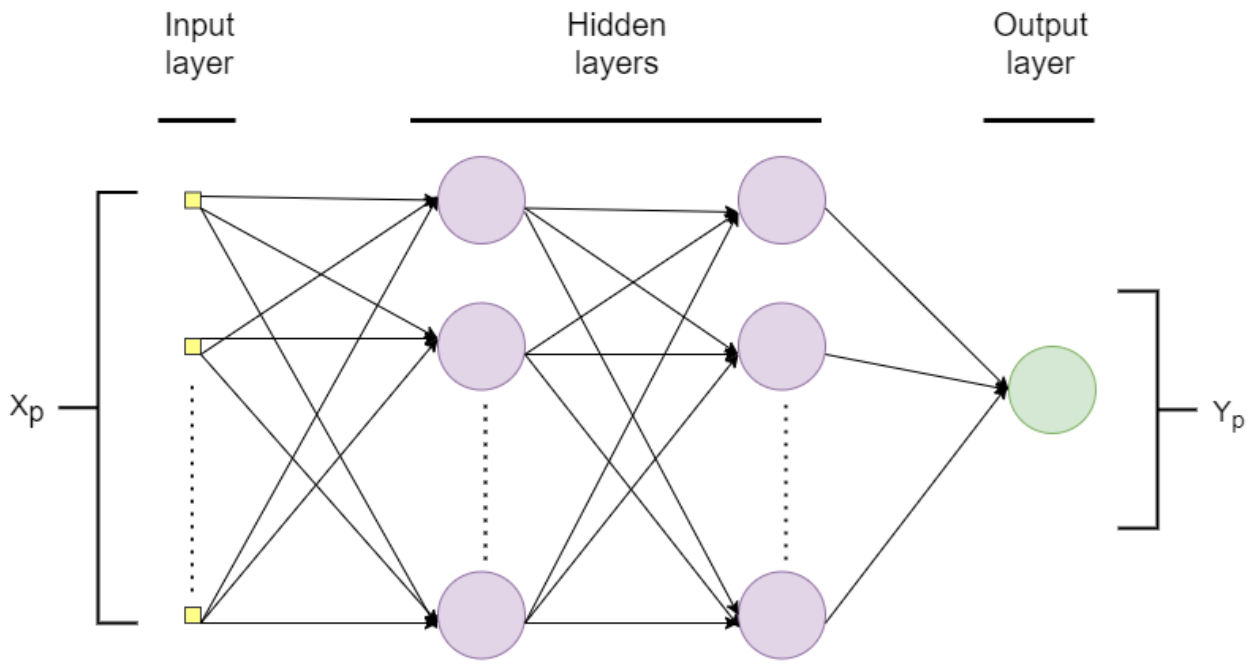

RF Model

| | Predicted Diabetes | Predicted No Diabetes |
|---|---|---|
| Actual Diabetes | 163 | 12 |
| Actual No Diabetes | 19 | 62 |

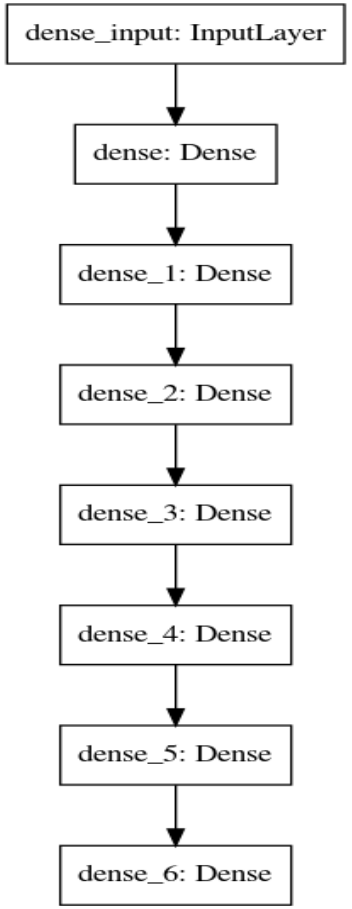
Confusion matrix for RF

The Accuracy for RF is 87.89%

# Artificial Neural Network (ANN)

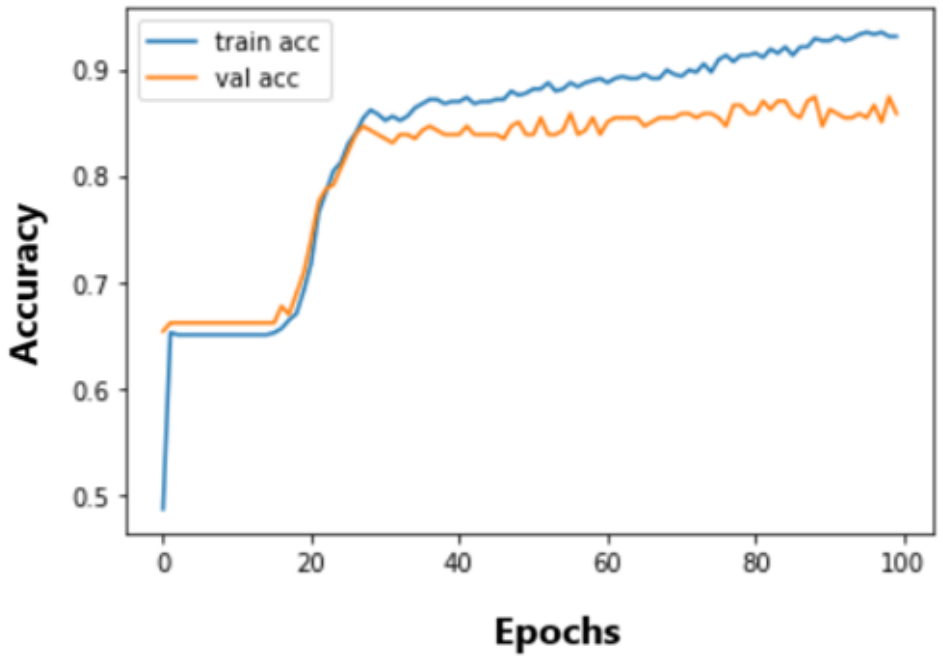



**Neural Network Model**




**ANN Architecture Used**

# Artificial Neural Network (ANN)




Accuracy vs. Epochs for ANN

| | Predicted Diabetes | Predicted No Diabetes |
|---|---|---|
| Actual Diabetes | 160 | 15 |
| Actual No Diabetes | 20 | 61 |

Confusion matrix for ANN

The Accuracy for ANN is 86.32%

# Evaluation of the Results

- We tried alternate algorithms to compare the different classification metrics and time for computation of the model.
- In our experiment, Random Forest outperformed KNN and ANN for prediction.
- Also, the Random Forest algorithm is better in terms of precision, recall, and f1-score, although it takes a slightly higher computational time compared to KNN.

| Model | Precision | Recall | F1-Score | Accuracy | Computation Time (Training + Testing) in sec |
|---|---|---|---|---|---|
| **KNN** | 87% | 89% | 88% | 82.81% | 1.43 |
| **RF** | **90%** | **93%** | **91%** | **87.89%** | **1.61** |
| **ANN** | 89% | 91% | 90% | 86.32% | 1.64 |

# Comparison with Previous Works

- The accuracy of the models obtained in this experiment exceeds the accuracies of previous works.
- As shown in the table, our Random Forest Classifier model, whose accuracy is 87.89%, exceeds the accuracy of the model presented in [9].
- Furthermore, our KNN and ANN models outperform prior efforts [4], [10] in accuracy.

| Model | Best Accuracy (Previous Works) | Accuracy (Our Approach) |
|---|---|---|
| **KNN** | 80.8% | **82.81%** |
| **RF** | 84.1% | **87.89%** |
| **ANN** | 85.09% | **86.32%** |

# Conclusion and Future Scope

- We have attained the highest accuracy of **87.89%** using the **Random Forest Classifier**.
- Our proposed system eliminates the need to visit the healthcare center physically for diabetes diagnosis.
- It has significant potential in medical science for the detection of various other medical data accurately.
- The neural network model may be enhanced by training for additional epochs, employing more hidden layers, and making a few more hyperparameter tunings.
- In the future, ensemble models may be employed to improve the outcomes.
- Increasing the feature importance of some of the features and forming new standardized datasets can help in improving the prediction.
- Moreover, adding more appropriate features and data values into the dataset can help in significantly improving the models' prediction.

# References


[1] Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, Di Angelantonio E, Ingelsson E, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet. 2010 Jun 26;375(9733):2215-22. doi:10.1016/S0140-6736(10)60484-9.

[2] Diabetes. (2021, November 10). World Health Organization. Retrieved June 26, 2022, from https://www.who.int/news-room/factsheets/detail/diabetes.

[3] Diabetes UK. Differences between type 1 and type 2 diabetes. Retrieved June 26, 2022, from https://www.diabetes.org.uk/diabetes-thebasics/differences-between-type-1-and-type-2-diabetes.

[4] S. K. Reddy, T. Krishnaveni, G. Nikitha and E. Vijaykanth, ”Diabetes Prediction Using Different Machine Learning Algorithms,” 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 1261-1265, doi: 10.1109/ICIRCA51532.2021.9544593

[5] P. Sonar and K. JayaMalini, ”Diabetes Prediction Using Different Machine Learning Approaches,” 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 367-371, doi: 10.1109/ICCMC.2019.8819841.

# References

[6] A. C. Lyngdoh, N. A. Choudhury and S. Moulik, ”Diabetes Disease Prediction Using Machine Learning Algorithms,” 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2021, pp. 517-521, doi: 10.1109/IECBES48179.2021.9398759.

[7] M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah, ”Prediction of Diabetes Using Machine Learning Algorithms in Healthcare,” 2018 24th International Conference on Automation and Computing (ICAC), 2018, pp. 1-6, doi: 10.23919/IConAC.2018.8748992.

[8] Galarnyk, Michael. “Https://Towardsdatascience.Com/UnderstandingBoxplots-5e2df7bcbd51.” Towards Data Science, https://towardsdatascience.com/. Accessed 12 July 2022

[9] Srinivasaiah, Raghavendra & Jankatti, Santosh. (2020). Performance evaluation of random forest with feature selection methods in prediction of diabetes. International Journal of Electrical and Computer Engineering (IJECE). 10. 353. 10.11591/ijece.v10i1.pp353-359

[10] Pradhan, Nitesh & Rani, Geeta & Dhaka, Vijaypal & Poonia, Ramesh. (2020). Diabetes prediction using artificial neural network. 10.1016/B978-0-12-819061-6.00014-8

# THANK YOU

# APPENDIX

# Heatmap (using seaborn) explaining how each attribute is related to the other