Cairo University
Faculty of Engineering
Systems & Biomedical Department

# Diabetic Prediction Using ML

## Artificial Intelligence in Medicine (SBME3021)

**Done by :**

| | | |
|---|---|---|
| **St : Yassin Essam Mohamed** | **Sec :2** | **BN:49** |
| **St : Abdelrahman Khaled** | **Sec :1** | **BN:50** |
| **St : Mustapha Megahed** | **Sec :2** | **BN:31** |
| **St : Ahmed Hossam Eldin** | **Sec :1** | **BN:6** |

1. Introduction
2. Review of Literature
3. Data Preparation
4. Data visualization
5. Data Cleaning
6. Results and discussion
7. Conclusion

# 1. Introduction

Diabetes is a very dangerous disease and does not able to cure. If this disease affect once, it will maintain in your life time. At the same time, your blood having too much of glucose can cause health issues. Like kidney disease, heart disease, stroke ,eye problems, dental disease, foot problems ,nerve damage.so you can take step to oversee your diabetes and avert these complications.
The familiar types of diabetes:

- Type 1 diabetes
- Type 2 diabetes
- Gestational diabetes

**Type1 Diabetes** Body does not able to produce insulin. Its affect children and young adults. Also it can affect at any age. Peoples affected by this type of diabetes to take insulin every day.

**Type2 Diabetes** Body does not able to produce or use insulin. This type of diabetes mostly affected on middle-aged and up in years.

**Gestational Diabetes** Women's are mostly affected by this type of diabetes. This type of diabetes develops during pregnancy. Gestational diabetes causes high blood sugar that can affect your pregnancy and you baby's health.

# 2. Review of Literature

| No. | PAPER TITLE | DATASET | ALGORITHM | OUTCOME & ACCURACY |
|---|---|---|---|---|
| 1. | Prediction Of Diabetes Using Classification Algorithm. | PPID. | Decision tree, SVM and Naive Bayesian. | Diabetes detection At an early stage. Accuracy 76%. |
| 2 | Comparison of three data mining models for predicting diabetes of prediabetes by rick factors. | To test 735 patients, they are came from two communities in Guangzhou, china. | Logistic algorithm, ANN and Decision tree. | The highest accuracy Is 77.87%. |
| 3 | Decision support system for a chronic disease-Diabetes. | PPID. | Naive Bayes , logistic regression , Extreme Gradient Boosting. | The accuracy for Extreme Gradient Boosting algorithm is 81%. |
| 4 | Prediction Of Diabetes Diagnosis Using Classification Based Data Mining Techniques. | Multi-dimensional Healthcare dataset. | Logistic -Regression, Multilayer -perception and KNN. | The accuracy level in Binary Logistic Regression is 0.69, Multilayer Perception is 0.71 and K-Nearest neighbor is 0.80. |
| 5 | Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach | PPID. | KNN, Naive Bayes, Random Forest and J48 | To get the best result in ensemble approach, when combining individual techniques. Also called the hybrid model. This provides the best performance and accuracy than the single one |
| 6 | Data Mining Models Comparison for Diabetes Prediction | PPID | Decision tree, Naive Bayes and KNN. | The result of this paper is that decision tree is the best prediction algorithm. It gives an accuracy level of 75.65%. |
| 7 | Decision Support Systems for Predicting Diabetes Mellitus –A Review | PPID | SVM, Naive Bayes and Decision Tree. | The accuracy for Decision Tree with pre-processing technique was 78.17%. |
| 8 | Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women | PIMA | SVM , Decision Tree. | SVM gives 82% accuracy. |

## 3. Data Preparation

We aren't going to create our own data set, instead, we will be using an existing data set called the "Pima Indians Diabetes Database" provided by the UCI Machine Learning Repository (famous repository for machine learning data sets). We will be performing the machine learning workflow with the Diabetes Data set provided above

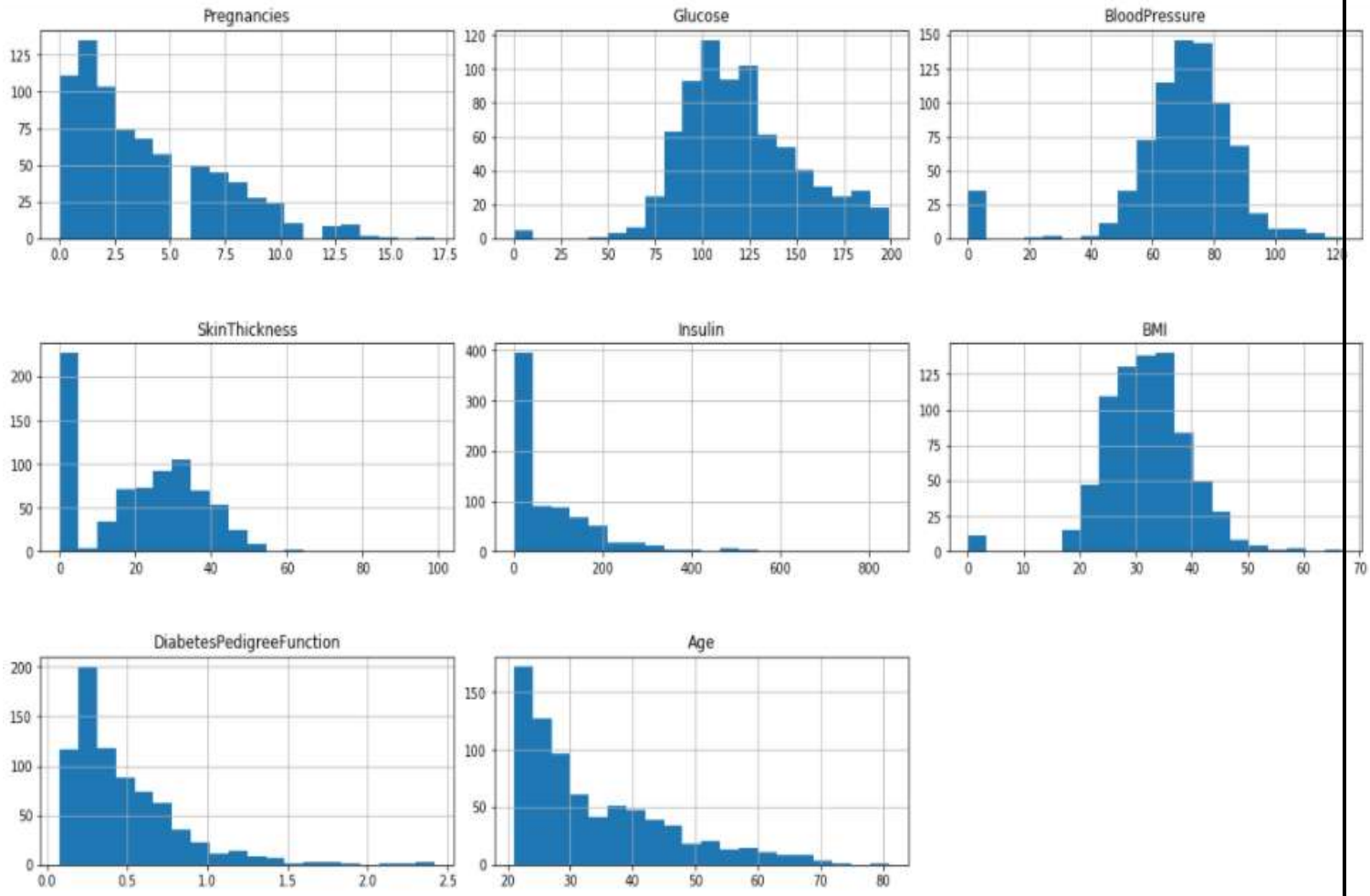
**The dataset contains 768 records (row) each has 9 columns as following:**

| Column | Description |
|---|---|
| Pregnancies | Number of times pregnant |
| Glucose | Plasma glucose concentration a 2 hours in an oral glucose tolerance test |
| Blood Pressure | Diastolic blood pressure (mm Hg) |
| Skin Thickness | Triceps skin fold thickness (mm) |
| Insulin | 2-Hour serum insulin (mu U/ml) |
| BMI | Body mass index (weight in kg/(height in m)^2) |
| Diabetes Pedigree Function | Diabetes pedigree function |
| Age | Age (years) |
| Outcome | Class variable (0 or 1) 268 of 768 are 1, the others are 0 |

# 4. Data visualization

When encountered with a data set, first we should analyze and "get to know" the data set. This step is necessary to familiarize with the data, to gain some understanding of the potential features and to see if data cleaning is needed

## Histogram Visualization




we used histogram to help us visualize the data distribution of the features and if there are any outliers.

# 5. Data Cleaning

There are several factors to consider in the data cleaning process.

1. Duplicate or irrelevant observations.
2. Bad labeling of data, same category occurring multiple times.
3. Missing or null data points.
4. Unexpected outliers.

| | |
|---|---|
| Pregnancies | 0 |
| Glucose | 5 |
| Blood Pressure | 35 |
| Skin Thickness | 227 |
| Insulin | 374 |
| BMI | 11 |
| Diabetes Pedigree Function | 0 |
| Age | 0 |
| Outcome | 0 |

**We have replaced each NAN value with mean values of the feature.**

## 6. Results and discussion

| Model | Accuracy |
|---|---|
| K Nearest neighbors | 78.57 % |
| Random Forest | 75.97 % |
| Support Vector Classifier | 73.38 % |
| Naive Bayes | 71.43 % |
| Logistic Regression | 71.43 % |
| Decision tree | 68.18 % |

K Nearest neighbors has the highest accuracy (78.57%), since we have properly labeled data; the data is almost noise-free and it's relatively small dataset.

## 7. Conclusion

Eventually, we can conclude that Glucose is the most important factor in determining the onset of diabetes followed by BMI and Age.

Other factors such Pregnancies, Blood Pressure, Skin Thickness, and Insulin also contributes to the prediction.

As we can see, the results derived makes sense as one of the first things that is actually monitored in high-risk patients is the Glucose level. An increased BMI might also indicate a risk of developing Type II Diabetes. Normally, especially in case of Type II Diabetes, there is a high risk of developing as the age of a person increases (given other factors).