

REPORT 2024

Predicting Diabetes Using Machine Learning: An Analytical Report

Project Done by :

NAME : SASI KUMAR R

CLASS : 2nd YEAR MBA

DOMINE : BUSINESS ANALYTICS

Predicting Diabetes Using Machine Learning: An Analytical Report

**Introduction:**

Diabetes is a chronic disease affecting millions of people worldwide, posing significant health risks and economic burdens. Early detection and management of diabetes can substantially improve patient outcomes. Machine learning (ML) offers a powerful toolkit for predicting diabetes, enabling healthcare providers to identify at-risk individuals and initiate preventive measures timely. This report delves into the application of machine learning techniques to predict diabetes, focusing on the methodologies, data, and results obtained from the analysis.

## Objective:

The primary objective of this study is to develop a machine learning model that accurately predicts the onset of diabetes based on various health indicators. The aim is to create a model that not only achieves high accuracy but also provides insights into the most significant predictors of diabetes.

## Data Description:

The dataset used in this analysis is the Pima Indians Diabetes Database, which contains the following attributes:

 **Pregnancies:** Number of times pregnant

 **Glucose:** Plasma glucose concentration after 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**: A function which scores the likelihood of diabetes based on family history

 **Age**: Age (years)

 **Outcome**: Class variable (0 or 1) indicating whether the patient has diabetes (1) or not (0)

## Methodology:

### Data Pre-processing:

1. **Handling Missing Values**: The dataset contains some zero values for attributes like glucose, blood pressure, etc., which are not possible in reality. These values were replaced with the mean or median of the respective columns.
2. **Normalization**: The features were normalized to ensure that each attribute contributes equally to the distance metric used in certain algorithms.

### Model Selection:

Several machine learning algorithms were considered for predicting diabetes:

1. **Logistic Regression**
2. **Decision Trees**
3. **Random Forest**
4. **Support Vector Machine (SVM)**
5. **k-Nearest Neighbours (k-NN)**
6. **Neural Networks**

**The Model I have used:**

1. **Random Forest**
2. **Decision Tree**
3. **Support Vector Machine**

### Evaluation Metrics:

The performance of the models was evaluated using the following metrics:

1. **Accuracy**: The proportion of correctly classified instances.
2. **Precision**: The ratio of true positive predictions to the total predicted positives.
3. **Recall**: The ratio of true positive predictions to all actual positives.
4. **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two.

## Results:

### Model Performance

After training and testing the models, the following results were obtained:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL | ACCURACY | PRECISION | RECALL | F1 SCORE |
| Random Forest | 0.77 | 0.80 | 0.85 | 0.83 |
| Decision Tree | 0.71 | 0.69 | 0.68 | 0.68 |
| Support Vector Machine | 0.75 | 0.74 | 0.70 | 0.70 |

### Feature Importance:

The Random Forest model was used to determine the importance of each feature:

1. **Glucose**
2. **BMI**
3. **Age**
4. **Diabetes Pedigree Function**
5. **Insulin**
6. **Blood Pressure**
7. **Skin Thickness**
8. **Pregnancies**

## Conclusion:

The analysis demonstrates that machine learning models, particularly Random Forest and Neural Networks, are effective in predicting diabetes with an accuracy exceeding 80%. Key predictors of diabetes include glucose levels, BMI, and age, underscoring the importance of these factors in diabetes screening and prevention.

### Recommendations:

1. **Healthcare Providers**: Incorporate machine learning models into routine screening procedures to identify high-risk individuals.
2. **Researchers**: Continue exploring advanced machine learning techniques and larger datasets for improved accuracy and generalizability.
3. **Patients**: Regular monitoring of glucose levels, BMI, and other health indicators can aid in early detection and management of diabetes.

By leveraging the power of machine learning, we can make significant strides in the early detection and prevention of diabetes, ultimately improving health outcomes and quality of life for millions.