Seth Harris, Joaquin Ramirez, Collin Real, Leonel Salazar

Professor Wang

STA 6543-01T

10 August 2024

Predictive Modeling Analysis for Early Detection of Diabetes in Pima Indians

**Introduction and Background:**

Diabetes is a chronic disease that poses a significant health threat globally, particularly among specific populations like the Pima Indians, who have a notably high incidence rate of diabetes. Early detection of diabetes is crucial for managing and potentially mitigating the disease’s impact. This project aims to apply predictive modeling techniques to medical data from the Pima Indians to predict the onset of diabetes. The goal is to identify the most effective machine learning model for early detection, which could be instrumental in improving healthcare management and developing preventive strategies.

This study utilizes a dataset containing various medical details, including the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, age, and diabetes outcome. By employing different machine learning models, we seek to determine which model most accurately predicts diabetes, providing valuable insights for healthcare professionals.

**Data Structure:**

The dataset used in this analysis is the Pima Indians Diabetes Database, sourced from Kaggle. It consists of 768 records and 9 attributes, where 8 are input features, and 1 is the target variable (outcome), indicating whether the patient has diabetes. The input features include:

• **Pregnancies:** Number of times the patient has been pregnant.

• **Glucose:** Plasma glucose concentration two hours after an oral glucose tolerance test.

• **Blood Pressure:** Diastolic blood pressure (mm Hg).

• **Skin Thickness:** Triceps skinfold 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 that represents the genetic predisposition of diabetes.

• **Age:** Age of the patient in years.

The dataset has been thoroughly preprocessed to handle missing values and normalize the data, ensuring consistency and comparability across different models.

**Statistical Learning Methods:**

We employed several predictive modeling techniques to analyze the dataset, including:

• **Logistic Regression:** A basic linear model used for binary classification problems like this one. It predicts the probability of a binary outcome based on one or more predictor variables.

• **Decision Trees:** A non-linear model that splits the data into subsets based on the value of input features, resulting in a tree structure that can be used for classification.

• **Random Forests:** An ensemble learning method that builds multiple decision trees and merges them to get a more accurate and stable prediction.

• **Support Vector Machine (SVM):** A linear classifier that finds the hyperplane that best separates the data into classes.

• **k-Nearest Neighbors (k-NN):** A simple, instance-based learning method that classifies new cases based on the majority vote of the k-nearest neighbors.

• **Neural Networks:** A complex model that mimics the human brain, consisting of interconnected neurons that process information and learn patterns in the data.

Each model was trained and evaluated based on accuracy, precision, recall, F1-score, and AUC-ROC to determine their effectiveness in predicting diabetes.

**Exploratory Analysis:**

Histogram plots were used to visualize the distribution of the diabetes dataset. Histograms are useful for understanding the shape of the data, and identifying whether data is skewed to the right or left. Histograms are also useful for identifying outliers and how data is spread across various intervals. Predictors, such as Glucose, illustrate normal distribution in the dataset:

A graph of glucose

Description automatically generated

However, predictors, such as DiabetesPedigreeFunction, illustrate non-normality distribution:

A graph of a patient's function

Description automatically generated

Despite a few of the diabetes’ predictors showing non-normal distribution, we chose not to transform the data with log, square root, or cubed root.

**Analysis Results:**

The results of the predictive modeling analysis are as follows:

• **Random Forest Model:** This model outperformed the others with the highest Area Under the Curve (AUC) score of 87%, indicating its superior ability to distinguish between diabetic and non-diabetic cases. It also had the highest accuracy (79%) and precision (83%).

A graph of different colored rectangular bars

Description automatically generatedA graph of different colored rectangular shapes

Description automatically generated

A graph of different colored rectangular shapes

Description automatically generated

A graph of different colored rectangular shapes

Description automatically generated

A graph of different colored squares

Description automatically generated

• **Logistic Regression and SVM:** Both models also performed well, with high AUC, accuracy, and precision. However, they did not surpass the performance of the Random Forest model.

• **Key Predictors:** The Random Forest model highlighted that glucose levels, insulin concentration, and age are the most influential predictors of diabetes risk. High glucose levels and insulin concentrations are strongly associated with diabetes, while age also plays a crucial role in its development and management.

A graph with different colored squares

Description automatically generated

**Conclusions and Discussions:**

The findings from this study suggest that the Random Forest model is the most effective predictive model for early detection of diabetes in Pima Indians. By focusing on key indicators such as glucose levels, insulin concentration, and age, healthcare providers can develop tailored wellness plans to minimize the risk of diabetes. Regular monitoring and screening, along with lifestyle changes, can significantly improve health outcomes for individuals at risk.

The implications of this study extend beyond the Pima Indian population, as the methodology and insights gained can be applied to other populations at risk of diabetes. Future work could involve expanding the dataset, incorporating additional variables, and testing other advanced machine learning models to further enhance the predictive power of early diabetes detection.

Works Cited

Kuhn, Max, and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013.

“Kaggle: Pima Indians Diabetes Database.” *Kaggle*, n.d., www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.