Medical Data Analysis for Diabetes Risk Prediction and Health Management

  RUI WANG RAGHUNATH MALE NIRANJAN REDDY DUMPA

1. Introduction

Objective of the Study  
This project demonstrates the use of predictive modeling for diagnosing diabetes based on clinical and lifestyle attributes. The CRISP-DM methodology provides a structured approach for data analysis, ensuring clear objectives and rigorous evaluation.

Relevance of Study  
Diabetes affects millions globally. Predictive models that integrate patient data offer potential to improve diagnostic efficiency, reduce costs, and enable early interventions, particularly in resource-constrained settings.

1. CRISP-DM Methodology

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a six-phase framework for building data-driven solutions:

|  |  |
| --- | --- |
| Phase | Description |
| Business Understanding | Identifying business objectives, such as diagnosing diabetes. |
| Data Understanding | Exploring the dataset, identifying trends, and understanding distributions. |
| Data Preparation | Cleaning, transforming, and organizing data for modeling. |
| Modeling | Selecting and training machine learning algorithms to predict diabetes outcomes. |
| Evaluation | Assessing model performance using metrics like accuracy, precision, recall, and F1-score. |
| Deployment | Preparing the model for real-world applications, such as in healthcare systems. |

1. Dataset Attributes

* The dataset used in the study is the Pima Indians Diabetes Dataset, available from the UCI Machine Learning Repository.
* This dataset is widely used for evaluating machine learning models in healthcare.

The dataset comprises 768 observations with 8 predictor variables and 1 target variable (Outcome).

|  |  |  |
| --- | --- | --- |
| Attribute | Description | Type |
| Pregnancies | Number of times pregnant | Numeric |
| Glucose | Plasma glucose concentration | Numeric |
| Blood Pressure | Diastolic blood pressure (mm Hg) | Numeric |
| Skin Thickness | Triceps skinfold thickness (mm) | Numeric |
| Insulin | 2-hour serum insulin (mu U/ml) | Numeric |
| BMI | Body mass index (weight in kg/height in m²) | Numeric |
| DiabetesPedigreeFunction | Genetic predisposition | Numeric |
| Age | Age of the patient | Numeric |
| Outcome | Diabetes diagnosis (1 = Diabetes, 0 = No Diabetes) | Binary |

1. Methods

* **Missing Values:** Attributes like Skin Thickness and Insulin had missing values. These were input using median values to retain dataset integrity.
* **Outliers:** Statistical techniques such as Interquartile Range (IQR) were used to detect and remove outliers.
* **Correlation Analysis:** Identified significant relationships between variables like Glucose and BMI with the target variable (Outcome).
* **Scaling:** Standardization was applied to normalize features, enhancing model performance.

The following machine learning algorithms were implemented and compared:

1. Logistic Regression

A simple, interpretable model for binary classification.

Advantages: Fast, interpretable, and effective with linearly separable data.

Limitations: Limited when relationships are non-linear.

1. Decision Tree

A rule-based model offering high interpretability.

Advantages: Easy to visualize and understand.

Limitations: Prone to overfitting with small datasets.

1. Random Forest

An ensemble of decision trees that aggregates outputs for robust predictions.

Advantages: Handles variable interactions well and reduces overfitting.

Limitations: Less interpretable compared to single-tree models.

Following are the evaluation metrics:

* Accuracy: Measures the percentage of correctly predicted outcomes.
* Precision: Evaluates how many predicted positive cases are truly positive.
* Recall (Sensitivity): Assesses the model's ability to identify all true positive cases.
* F1-Score: Combines precision and recall into a single metric, providing a balanced view of performance.

1. Results

* The models were evaluated on their ability to correctly predict diabetes status based on the attributes provided.
* Summary of Results:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Recall | F1-Score |
| Logistic Regression | 78% | 74% | 75% |
| Decision Tree | 73% | 71% | 70% |
| Random Forest | 87% | 86% | 85% |

Best Model: Random Forest emerged as the top performer, achieving an accuracy of 87% and a strong F1-score of 85%.

* **Feature Importance**
  + Glucose and BMI were the most predictive attributes.
  + Secondary predictors included Age and Diabetes Pedigree Function.

1. Implications

* The use of machine learning in healthcare offers significant advantages in improving diagnostic accuracy and efficiency.
* Such models can aid clinicians in early identification of high-risk individuals, potentially saving lives and reducing healthcare costs.
* Diverse Datasets: Incorporating larger and more diverse patient data for generalizability.
* Advanced Algorithms: Exploring deep learning techniques for feature extraction and model accuracy improvement.
* Integration with Systems: Embedding the model into clinical decision support systems for real-time diagnosis.

1. Conclusion

This study highlights the power of predictive modeling in healthcare, specifically in diagnosing diabetes. By leveraging the structured CRISP-DM methodology, the project demonstrates a systematic approach to data analysis and machine learning. With Random Forest as the standout performer, the findings underscore the importance of robust evaluation and the potential for machine learning to revolutionize medical diagnostics.