Predictive Modeling for Diabetes Diagnosis: An Application of CRISP-DM Methodology

Omkar Nagarkar

1 Abstract

This research paper explores the application of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology for developing a predictive model to diagnose diabetes. Utilizing a dataset of patients with various medical predictor variables and one target variable, Outcome, six classification models were trained, evaluated, and compared to identify the most accurate and reliable model for predicting diabetes.

2 Introduction

Diabetes is a chronic health condition that affects millions of people worldwide. Early and accurate diagnosis is crucial for managing the disease and preventing complications. This study aims to apply data science techniques to develop a predictive model for diabetes diagnosis, utilizing the CRISP-DM methodology.

3 Methodology

The CRISP-DM methodology, comprising six phases—Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment—was employed to guide the research process.

4 Data Understanding andPreparation

The dataset comprises several medical predictor variables such as Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Pregnancies, along with the target variable, Outcome. The data was thoroughly analyzed, visualized, cleaned, and pre-processed, including handling missing values and feature scaling.

5 Modeling

Six classification models—Logistic Regression, Decision Tree Classifier, Random Forest Classifier, K- Nearest Neighbors, Support Vector Machine, and Naive Bayes—were trained and evaluated based on metrics such as accuracy, precision, recall, F1 score,and AUC-ROC score.

6 Evaluation

The performance of the six classification models was evaluated using various metrics such as accuracy, precision, recall, F1 score, and AUC-ROC score.

Table 1: Model Evaluation Metrics

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC Score
Random Forest	77.92%	72.73%	59.26%	65.31%	73.63%
K-Nearest Neighbors	75.32%	66.00%	61.11%	63.46%	72.06%
Support Vector Machine	73.38%	64.44%	53.70%	58.59%	68.85%
Naive Bayes	70.13%	56.67%	62.96%	59.65%	68.48%
Logistic Regression	70.78%	60.00%	50.00%	54.55%	66.00%
Decision Tree	68.18%	55.32%	48.15%	51.49%	63.57%

Based on these results, the Random Forest Classifier emerged as the best-performing model, achieving the highest scores across all evaluation metrics. Feature importance analysis further revealed the significance of features like Glucose, BMI, DiabetesPedigreeFunction, and Age in making accurate predictions.

7 Deployment

The final Random Forest model was retrained on the entire dataset and prepared for deployment. Continuous monitoring and maintenance are recommended post-deployment to ensure sustained model performance.

8 Conclusion

This study successfully applied the CRISP-DM methodology to develop a predictive model for diabetes diagnosis. The Random Forest Classifier emerged as the best-performing model, achieving the highest scores across all evaluation metrics. Feature importance analysis further revealed the significance of features like Glucose, BMI, DiabetesPedigreeFunction, and Age in making accurate predictions.

9 References

1. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.
2. American Diabetes Association. (2014). Diagnosis and classification of diabetes mellitus. Diabetes care, 37(Supplement 1), S81-S90.
3. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.