Diabetes Prediction

Neeharika Vasadi, Nikhitha Yarajarla, Naveenkumar Mahamkali

Department of Computer Science and Engineering, SRM University, Amaravati, 522240, Andhra Pradesh, India

Abstract

This project leverages machine learning techniques to predict diabetes diagnoses using the PIMA Indians Diabetes dataset. The dataset includes diagnostic features such as glucose levels, BMI, insulin, and age. Extensive preprocessing and exploratory data analysis (EDA) were performed to understand data patterns and relationships. Multiple machine learning algorithms, including Random Forest were implemented and rigorously evaluated. The project emphasizes the application of machine learning for healthcare analytics, offering tools to identify individuals at risk of diabetes for timely intervention.

Introduction

Diabetes is a major global health issue, and early detection is crucial for preventing complications. Using machine learning, predictive analytics offers a proactive approach to diagnosing diabetes. This project focuses on utilizing the PIMA Indians Diabetes dataset to predict diabetes outcomes.

First, the dataset undergoes preprocessing to handle missing values and standardize features. Exploratory Data Analysis (EDA) helps identify key patterns and relationships, such as the importance of glucose levels, BMI, and age in predicting diabetes. Machine learning models, specifically the Random Forest Classifier, are built to make predictions based on these features.

The model's performance is optimized using hyperparameter tuning, ensuring higher accuracy and generalizability. Cross-validation is employed to assess the model's stability and avoid overfitting. The results showcase how machine learning can significantly enhance clinical decision-making.

By accurately predicting diabetes outcomes, the model helps healthcare professionals identify at-risk individuals and plan timely interventions, ultimately improving patient care and contributing to better healthcare planning.

*Corresponding author

Email addresses: neeharika_vasadi@srmap.edu.in (Neeharika Vasadi), nikhitha_yarajarla@srmap.edu.in (Nikhitha Yarajarla),naveenkumar.m@srmap.edu.in (Naveenkumar Mahamkali)

Preprint submitted to Machine Learning Project Work *November 25, 2024*

Literature Survey

Existing research has explored machine learning models like Logistic Regression, SVM, and Neural Networks for diabetes prediction. While these methods demonstrate promise, challenges such as overfitting and inconsistent preprocessing persist.
This study improves on prior work by:

1. Utilizing robust preprocessing techniques.
2. Using Random Forest for diabetes prediction.
3. Incorporating hyperparameter tuning and cross-validation to enhance reliability.

Proposed Method

3. Methodology

3.1 Data Preprocessing

- **Missing Values**: Addressed using domain-specific imputations.
- **Standardization**: Continuous features were scaled for uniformity.
- **EDA**: Unveiled feature distributions and correlations influencing diabetes outcomes.

3.2 Model Training and Evaluation

- **Algorithms**: Random Forest Classifier.
- **Metrics**: Accuracy, precision, recall, F1-score, and confusion matrix.

3.3 Hyperparameter Tuning

- **Random Forest**: Parameters like the number of trees, max depth, and min samples split were optimized using RandomizedSearchCV.

3.4 Cross-Validation

K-fold cross-validation (k=5) was used to ensure model generalizability and minimize overfitting.

3.5 Model Deployment

The best-performing model, Random Forest Classifier, was saved for deployment using Joblib.

Results and Discussion

4.1 Model Performance: Random Forest

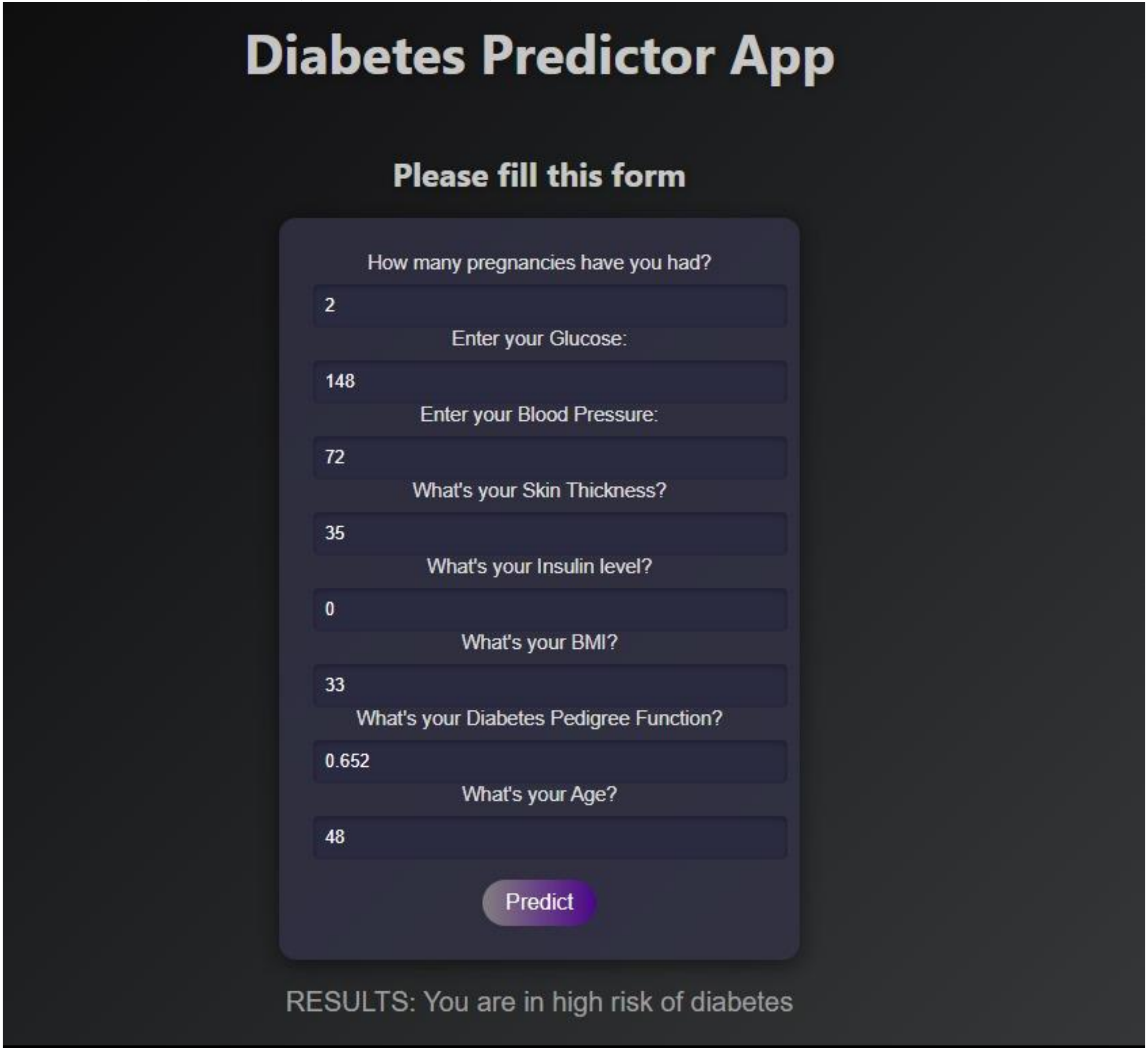
- **Accuracy**: 84%
- **Precision**: Demonstrated strong ability to correctly identify positive cases (diabetes).
- **Recall**: Balanced sensitivity in detecting true positives without overestimating false negatives.
- **F1-Score**: Balanced performance between precision and recall.

4.2 Confusion Matrix Analysis

The Random Forest model effectively classified most cases, showing a good balance between true positives and true negatives. However, some misclassifications occurred, indicating potential for further tuning or additional feature engineering.

4.3 Comparative Insights (Limited to Random Forest)

The Random Forest Classifier proved reliable due to its robustness and ability to handle interactions between features. It performed consistently across training and testing datasets, underscoring its suitability for healthcare predictions with structured numerical data.



Diabetes Predictor App

Please fill this form

How many pregnancies have you had?

2

Enter your Glucose:

148

Enter your Blood Pressure:

72

What's your Skin Thickness?

35

What's your Insulin level?

0

What's your BMI?

33

What's your Diabetes Pedigree Function?

0.652

What's your Age?

48

RESULTS: You are in high risk of diabetes

Conclusion

This study successfully demonstrated the use of machine learning for diabetes prediction, focusing on the Random Forest Classifier for robust and accurate results. The model effectively captured complex relationships between health features, such as glucose levels, BMI, and age, which are key predictors of diabetes outcomes. Rigorous data preprocessing and exploratory analysis ensured the dataset was well-prepared, contributing to reliable model performance.
The Random Forest model achieved an accuracy of 84%, highlighting its potential for real-world applications in healthcare analytics. This approach underscores the value of machine learning in identifying individuals at risk of diabetes, enabling early intervention and improved patient outcomes.
Future work could explore additional demographic and lifestyle variables, employ more advanced ensemble techniques, and develop user-friendly tools for clinicians. This project highlights the transformative role of machine learning in healthcare, paving the way for proactive, data-driven medical decision-making.

References

ttps://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

Neeharika Vasadi
AP22110011070

Nikhitha Yarajarla
AP22110011104