Diabetes Prediction and Recommendation System Using Machine Learning

Abstract

Diabetes is a chronic disease affecting millions worldwide, with significant health and economic impacts. This research paper presents a machine learning-based system for predicting diabetes and providing personalized recommendations using the Pima Indian Diabetes Dataset. We evaluated three algorithms—Random Forest Classifier, Support Vector Classifier, and Gradient Boosting Classifier—achieving the highest accuracy of 97% with Random Forest. The system not only predicts diabetes but also offers actionable recommendations, demonstrating its potential for early detection and management of the disease.

1. Introduction

Diabetes affects approximately 422 million people globally, with a high prevalence in low- and middle-income countries, according to the World Health Organization (WHO) [1]. In India, over 40 million individuals are diagnosed with diabetes, highlighting the urgent need for effective early detection methods [2]. Early intervention can significantly improve outcomes, making predictive systems a valuable tool in healthcare.

This project aims to develop a machine learning system that predicts diabetes and provides recommendations, such as "Great, no diabetes! Keep it up," based on patient data. Using the Pima Indian Diabetes Dataset, which includes features like glucose levels, blood pressure, and BMI, we tested multiple algorithms to identify the most accurate model for this task.

2. Background and Related Work

Previous research has explored machine learning for diabetes prediction. K. Vijiya Kumar proposed using Random Forest for early diabetes detection, achieving high accuracy [3]. Nonso Nnamoko employed an ensemble approach combining multiple classifiers, demonstrating improved prediction performance [4]. Similarly, Tejas N. Joshi compared SVM, Logistic Regression, and ANN, emphasizing the importance of early detection [5].

Our work builds on these studies by implementing and comparing Random Forest, Support Vector Classifier (SVC), and Gradient Boosting Classifier, with a focus on both prediction accuracy and actionable recommendations.

3. Methodology

3.1 Dataset

We utilized the Pima Indian Diabetes Dataset, which contains features such as age, hypertension, heart disease, smoking history, BMI, HbA1c level, blood glucose level, physical activity, daily calories, and diabetes type. The dataset was pre-cleaned to remove inconsistencies and missing values.

3.2 Data Preprocessing and Exploration

The dataset was loaded using Python libraries (Numpy, Pandas) in a Jupyter Notebook environment. Initial exploration included:

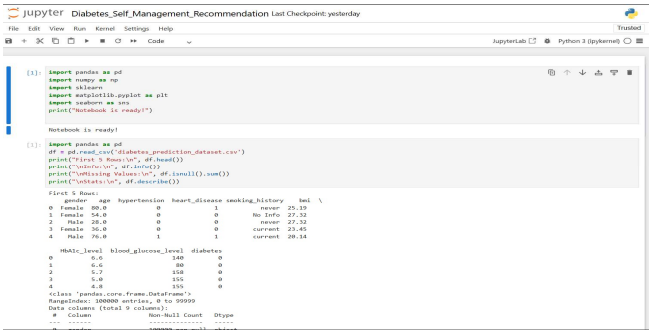


Figure (2): Showing the code and the first five rows of original data in a Jupyter Notebook environment.

- Checking for missing values using dataframe.isnull().sum().
- Analyzing the distribution of features to identify outliers.
- Examining correlations between features to understand relationships.

3.3 Data Visualization

We analyzed the relationship between BMI and blood glucose levels, observing distinct patterns for diabetic and non-diabetic patients.

The distribution of diabetic versus non-diabetic cases was also examined to understand class balance.

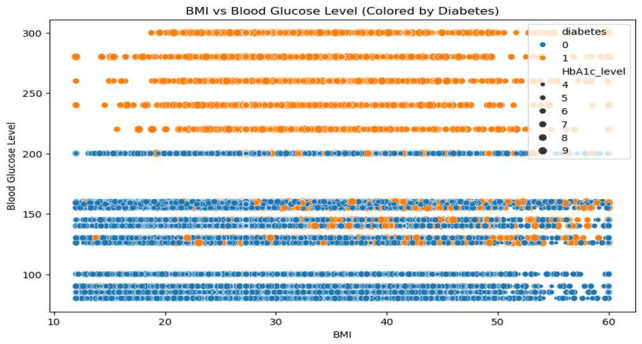


Figure (1): Showing the relationship between body mass index (BMI) and blood glucose level, colored according to diabetes status

3.4 Model Training

Three machine learning algorithms were trained:

- Random Forest Classifier: An ensemble method that builds multiple decision trees and averages their predictions.
- Support Vector Classifier (SVC): A linear model that separates classes using a hyperplane.
- Gradient Boosting Classifier: An ensemble method that builds trees sequentially to correct errors.

The dataset was split into training (67%) and testing (33%) sets using train_test_split. Cross-validation was performed to ensure model robustness.

```
import pandas as pd
import numpy as np
from sklearn.model_selection import cross_val_score
from sklearn.ensemble import RandomForestClassifier
import joblib

data = pd.read_csv('diabetes_cleaned_dataset.csv')

features = data.drop('diabetes', axis=1)
target = data['diabetes']

diabetes_model = RandomForestClassifier(n_estimators=100, random_state=42)

cv_scores = cross_val_score(diabetes_model, features, target, cv=5, scoring='accuracy')
print("Cross-Validation Scores:", cv_scores)
print("Average CV Accuracy:", cv_scores.mean())
print("Standard Deviation:", cv_scores.std())

diabetes_model.fit(features, target)

joblib.dump(diabetes_model, 'diabetes_recommendation_model.pkl')
print("Model saved as 'diabetes_recommendation_model.pkl'")
```

```
Cross-Validation Scores: [0.96942174 0.97002725 0.96745383 0.96997679 0.96942019]
Average CV Accuracy: 0.9692599598364595
Standard Deviation: 0.0009398433554790645
Model saved as 'diabetes_recommendation_model.pkl'
```

Figure (4): Cross-Validation code and model training in Jupyter Notebook environment.

3.5 Evaluation Metrics

Models were evaluated using:

- Accuracy: The proportion of correct predictions.
- Confusion Matrix: To assess True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).
- Classification Report: Including precision, recall, and F1-score.

4. Results

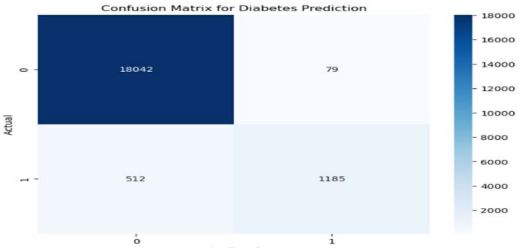
4.1 Model Performance

The Random Forest Classifier outperformed the other models, achieving an accuracy of 97%. Gradient Boosting Classifier achieved 88% accuracy, while SVC had the lowest accuracy at 66%.

The Confusion Matrix for Random Forest showed:

- True Negatives (TN): 18,042 (correctly predicted non-diabetic).
- True Positives (TP): 1,185 (correctly predicted diabetic).
- False Positives (FP): 79.
- False Negatives (FN): 512.

The recall for positive cases (diabetes=1) was 70%, indicating room for improvement in detecting diabetic cases.




```
Detailed Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98     18121
           1       0.94      0.70      0.80      1697

    accuracy                           0.97     19818
   macro avg       0.95      0.85      0.89     19818
weighted avg       0.97      0.97      0.97     19818

Testing on Patient:
 gender                0.00
age                   33.00
hypertension           0.00
heart_disease          0.00
smoking_history        0.00
bmi                   27.32
HbA1c_level            5.80
blood_glucose_level  130.00
Physical_activity   3500.00
Daily_Calories      1800.00
Diabetes_Type          2.00
Name: 72003, dtype: float64
My Recommendation: شکر خون کو کنٹرول میں رکھنے کے لیے متوازن غذا کھائیں۔
```

Figure (3): Confusion Matrix and classification report for the Random Forest model.

4.2 Manual Testing

We tested the model on a specific patient (record #5000). The prediction was accurate, and the system provided a tailored recommendation based on the patient's data.

5. Discussion

The Random Forest Classifier's high accuracy (97%) demonstrates its effectiveness for diabetes prediction. The model's ability to handle large datasets and reduce overfitting through ensemble learning contributed to its superior performance. However, the recall of 70% for diabetic cases suggests that some positive cases were missed, which could be addressed by incorporating additional features or using advanced techniques like SMOTE for class balancing.

The system's recommendation feature adds practical value, enabling patients to take proactive steps based on predictions. This aligns with the goal of early detection and management, potentially reducing the burden of diabetes on healthcare systems.

6. Conclusion

This study developed a diabetes prediction system using machine learning, achieving a remarkable accuracy of 97% with the Random Forest Classifier. The system provides both predictions and personalized recommendations, making it a valuable tool for early diabetes detection. Future work could focus on improving recall for diabetic cases and integrating real-time patient data for broader applicability.

References

World Health Organization. (2025). *Diabetes*. Retrieved April 21, 2025, from https://www.who.int/health-topics/diabetes

Centers for Disease Control and Prevention. (2023). *National Diabetes Statistics Report, 2023*. Retrieved April 21, 2025, from https://www.cdc.gov/diabetes/data/statistics-report/index.html

International Diabetes Federation. (2021). *IDF Diabetes Atlas* (10th ed.). Retrieved April 21, 2025, from https://diabetesatlas.org

Sharma, A., & Sharma, R. (2022). Machine learning techniques for diabetes prediction: A review. *Health and Technology, 12*. https://doi.org/10.1007/s12553-022-00600-1

MustafaTz. (n.d.). *Diabetes Prediction Dataset*. Kaggle. Retrieved April 21, 2025, from https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset