Diabetes Prediction Using XGBoost

A machine learning project that predicts whether a patient is Non-Diabetic (0), Pre-Diabetic (1), or Diabetic (2) using clinical laboratory data. The final model achieves 98.5% accuracy and identifies the top medical indicators contributing to diabetes risk.

PROJECT OVERVIEW

This project explores clinical health data to build a predictive model capable of classifying diabetes status. The workflow includes:

- Data cleaning & preprocessing
- Feature engineering
- Exploratory data analysis (EDA)
- Categorical encoding
- Scaling
- Model building with multiple algorithms
- Cross-validation
- Final model selection
- Model interpretation (feature importance)
- Saving the trained model for deployment
- Optional Streamlit app for real-time prediction

The XGBoost Classifier emerged as the best-performing model with outstanding accuracy and generalization.

KEY FEATURES

- Multi-class diabetes prediction (0 = Non-diabetic, 1 = Pre-diabetic, 2 = Diabetic)
- XGBoost model achieving 98.5% accuracy
- Automatic feature engineering & one-hot encoding

- Proper train-test split with stratification

- Evaluation using classification report & confusion matrix

- Feature importance visualization

- Model exported as .pkl for deployment

- Optional Streamlit app for interactive predictions

TECHNOLOGIES USED

Programming: Python

Libraries:

- pandas

- numpy

- matplotlib

- seaborn

- scikit-learn

- xgboost

- joblib

- streamlit (for deployment)

PROJECT STRUCTURE

diabetes-prediction-xgboost/

- diabetes_prediction.ipynb

- app.py

- diabetes_xgboost_model.pkl

- requirements.txt

- README.md

- images/

MODEL PERFORMANCE

Best Model: XGBoost Classifier

- Accuracy: 98.5%

- Macro Avg F1-Score: 0.96

- Weighted Avg F1-Score: 0.99

Confusion Matrix Highlights:

- 95% of Non-Diabetic cases correctly predicted

- 86% of Pre-Diabetic cases correctly predicted

- 99% of Diabetic cases correctly predicted

TOP FEATURES (XGBoost Feature Importance)

1. HbA1c

2. BMI

3. Cholesterol

4. Triglycerides (TG)

5. Creatinine (Cr)

6. HDL

7. Urea

8. VLDL

9. LDL

10. Age Range (40-50)

HOW TO RUN THE STREAMLIT APP

pip install -r requirements.txt

streamlit run app.py

SAVING THE MODEL

import joblib

joblib.dump(xgb_model, "diabetes_xgboost_model.pkl")

CONCLUSION

This project demonstrates how structured medical data can be leveraged to build a highly accurate, interpretable diabetes classification model. The final XGBoost model is robust, generalizes well, and can be integrated into clinical decision-support systems or patient-facing wellness applications.

AUTHOR

Nikechukwu Ndukwe

Data Analyst & ML Practitioner