MULTI-CLASS CLASSIFICATION OF DIABETES USING MACHINE LEARNING TECHNIQUES

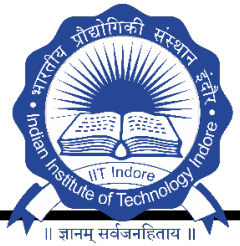
CS 403/603: Machine Learning-Project

Presented by

Avinash Bharti (PhD 2501101010)
Maha Janani M (MS 2504102003)
Rajesh Kanna (MS 2504101010)
Chirki Sivaji (PhD 2501131008)
Anurag Raghuvanshi (PhD 2501105007)

13 NOVEMBER 2025

CONTENTS



PROBLEM MOTIVATION & OBJECTIVE

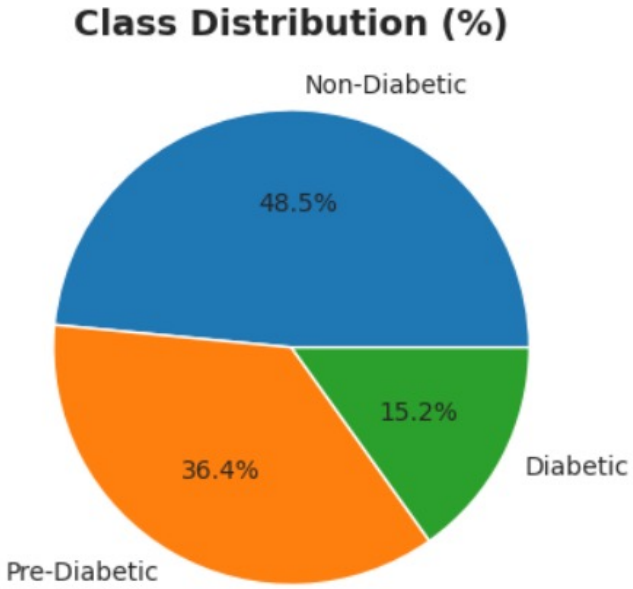
- Diabetes cases are increasing rapidly worldwide. Therefore, making an early detection critical for prevention.
- Traditional diagnosis often misses the prediabetes stage, delaying intervention.
- Multiple clinical biomarkers contain important signals but they are difficult to interpret manually.
- Objective: The objective of proposed work is to help in finding an easier, faster and better way of diagnosing the diabetes disease into multiple classes using machine learning approach.
- The machine learning approach of diagnosis of diabetes can automate most of the process and as a result helping the doctors in saving a lot many more patients.

DATASET DESCRIPTION

TOTAL SAMPLES 264

TARGET CLASS

- Diabetic
- Non-Diabetic
- Pre-Diabetic



FEATURES

- Gender
- Age
- Urea
- Creatinine
- HbA1c
- Cholesterol
- TG
- HDL
- LDL
- VLDL
- BMI

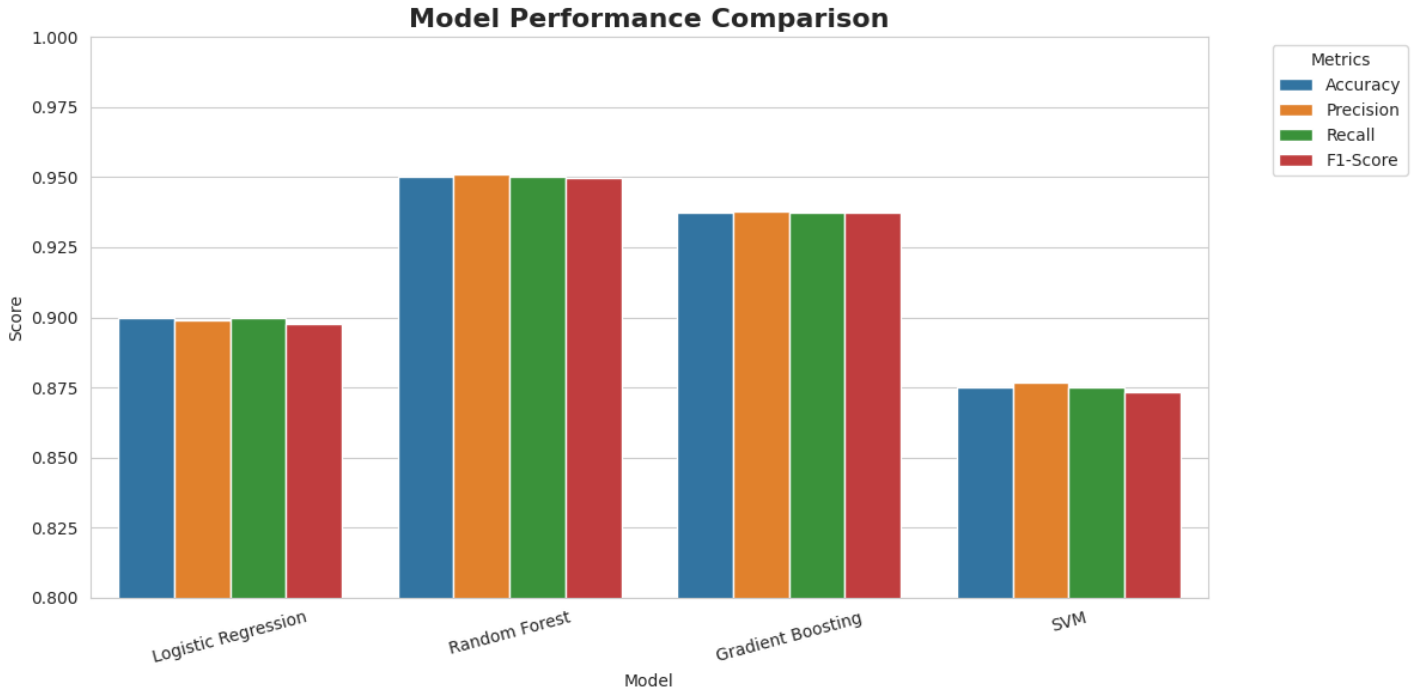
DATA PRE-PROCESSING

- Data split → 70% training & 30% testing
  - Training samples: 184
  - Testing samples: 80
- Standard Scaler applied:
  - fit_transform() on training data
  - transform() on test data
- Scaling ensures all features are on a similar range for ML model performance.

MODAL TRAINING PROCESS

- Four machine learning models initialized are,
  - Logistic Regression,
  - Random Forest,
  - Gradient Boosting &
  - SVM.
- All models are trained on scaled training data.
- Metrics calculated for each model are,
  - Accuracy,
  - Precision,
  - Recall &
  - F1-score.
- 5-fold Cross-Validation used to measure model stability and generalization.

MODAL PERFORMANCE COMPARISON



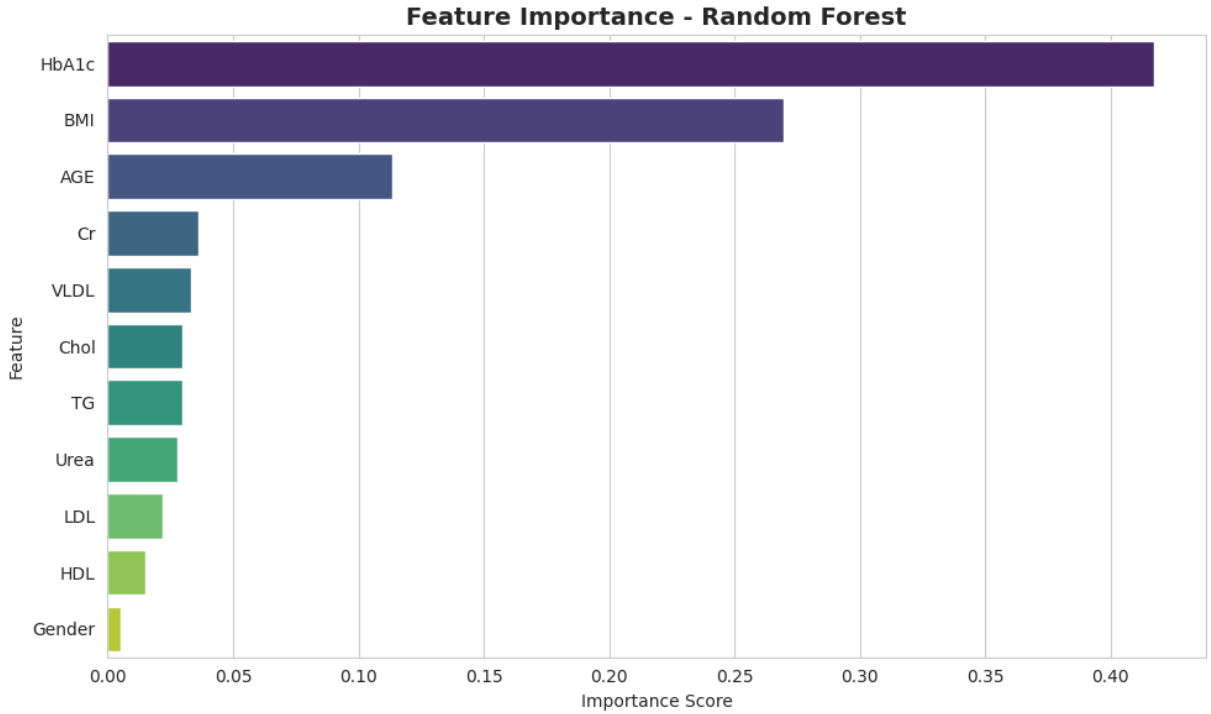
Random Forest(Best Model)
Accuracy: 0.95 | F1: 0.95 | CV: 0.9836

Logistic Regression
Accuracy: 0.90 | F1: 0.90 | CV: 0.8913

Gradient Boosting
Accuracy: 0.94 | F1: 0.94 | CV: 0.9730

SVM (RBF Kernel)
Accuracy: 0.88 | F1: 0.88 | CV: 0.8315

FEATURE IMPORTANCE



- Moderate impact features

  Cholesterol, Urea, VLDL, Creatinine, TG, LDL.

- Least impactful:

  HDL.

  Gender – 0.0053 (minimal influence)

- **HbA1c – 0.4171**
  (strongest predictor of diabetes stage)

- **BMI – 0.2696**
  (second most important metabolic indicator)

- **Age – 0.1137**
  (age increases diabetes likelihood)

CONCLUSION

- A multi class prediction can help doctors in approaching the treatment of diabetes disease for a better prognosis.
- The output values show that the Random Forest is having better accuracy value than the other three models.
- The overall accuracy of the Random forest is 95.00%, which is due to the of availability of HbA1c (the main factor in detection of diabetes).
- A lot of suffering can be avoided by the detection and prevention of diabetes at an early stage.

Thank You