DIABETES RISK PREDICTION USING MACHINE LEARNING

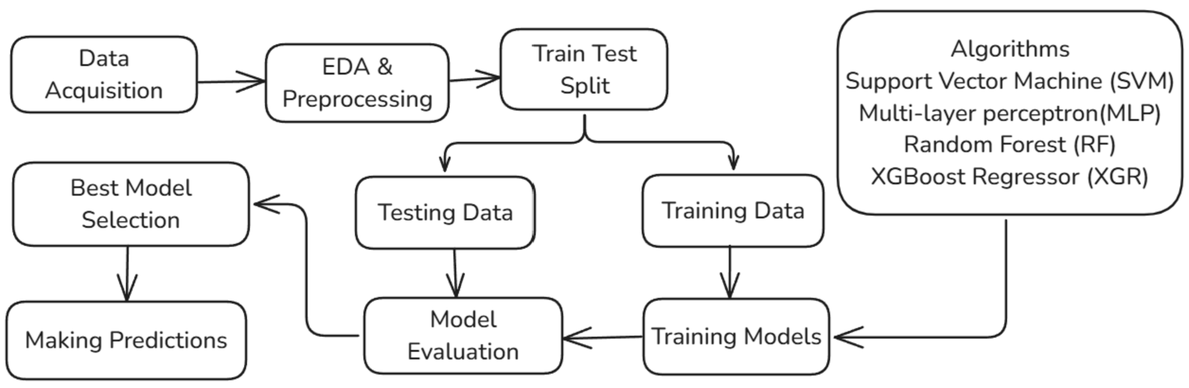
Aviral Vashistha and Bhavya Bharadwaj

Department of Computer Science & Engineering
SRM Institute Of Science & Technology, Delhi-NCR Campus, Modinagar, Ghaziabad

Foundation Day 2025

ABSTRACT

Early detection of diabetes is crucial for effective management. This study evaluates four machine learning models—Random Forest, XGBoost, SVM, and MLP—on a dataset of 100,000 patient records. To enhance interpretability, we integrate Explainable AI using LIME to analyze key features influencing predictions. Model performance is assessed using accuracy, precision, recall, and F1-score to identify the most reliable approach for diabetes classification.



ML Model Developement Flow Chart

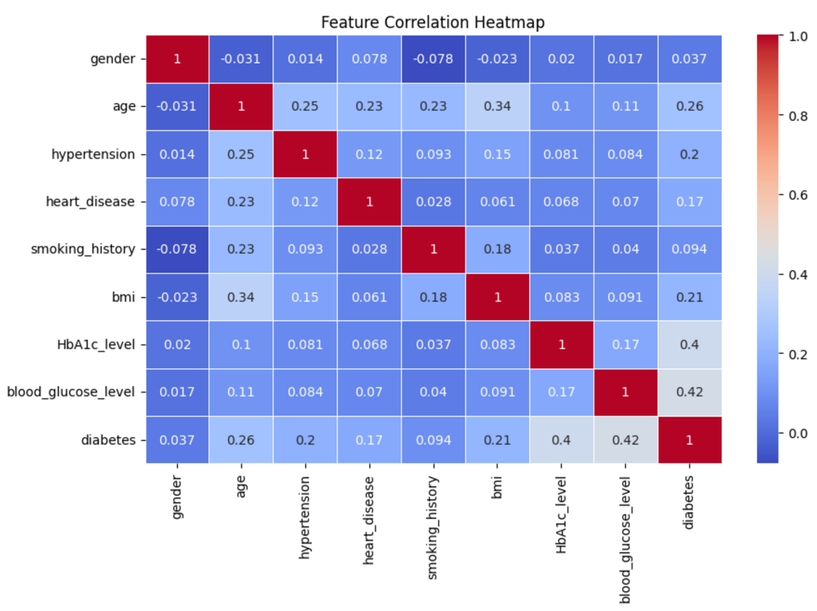
Error Evaluation Table

Models	RMSE	MAE	MBE
XGBoost	0.178326	0.0318	0.0188
RandomForest	0.166283	0.02765	0.02735
SVM Classifier	0.18775	0.03525	0.03405
Neural Networks	0.170734	0.02915	0.02915

- Random Forest has the lowest RMSE (0.166283) and MAE (0.02765), indicating it has the best overall performance among the models in terms of error minimization.
- XGBoost has slightly higher RMSE and MAE than Random Forest but lower MBE, meaning it has less bias compared to other models.
- SVM Classifier has the highest RMSE (0.18775) and MAE (0.03525), indicating the poorest performance in terms of prediction accuracy.
- Neural Networks perform slightly worse than RandomForest but better than SVM Classifier, showing a balance between accuracy and error.

Correlation Matrix

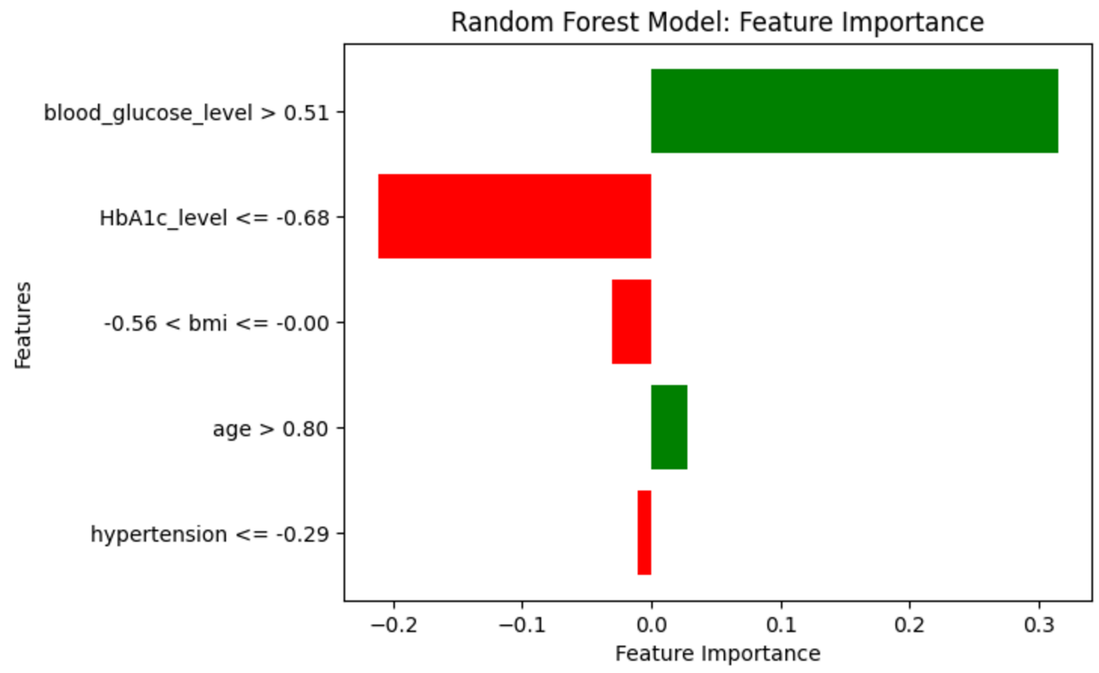
- Diabetes is most strongly linked with HbA1c level and blood glucose level.
- Age plays a role in several health conditions, including diabetes, hypertension, and BMI.
- Heart disease has a weak correlation with most factors, indicating other underlying influences.
- Smoking history does not show strong links with diabetes or heart disease, which may require further investigation.



BACKGROUND OF THE STUDY

- Diabetes as a Global Health Challenge – A prevalent chronic disease causing severe complications, necessitating improved early detection methods.
- Importance of Early Diagnosis – Early identification helps in effective disease management, reducing complications and improving patient outcomes.
- Role of Machine Learning in Prediction – ML algorithms can analyze complex health data to uncover patterns that aid in more accurate predictions.
- Model Selection for Effective Prediction – Algorithms such as Random Forest (RF), XGBoost, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) are widely used for medical diagnostics.
- Enhancing Interpretability with LIME – Local Interpretable Model-Agnostic Explanations (LIME) provides insights into model decisions, increasing trust and usability in clinical settings.
- Performance Evaluation: Accuracy, precision, recall, and F1-score determine the best model.

Explainable AI's feature importance



This visualization illustrates feature importance using LIME for one of the observation in the diabetes prediction model.

Key Insights:

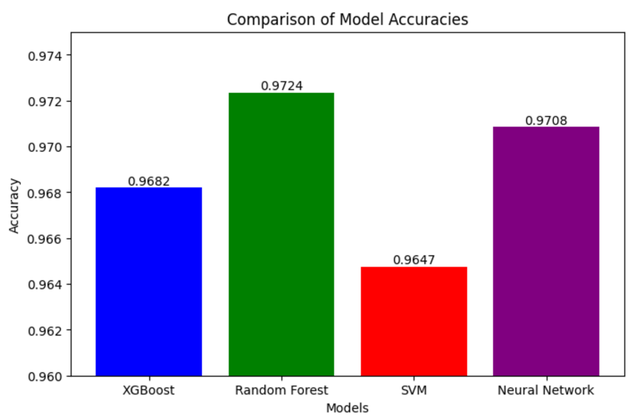
- Blood Glucose Level is the most influential factor, strongly increasing diabetes risk.
- HbA1c Level has a negative impact, lowering the likelihood of diabetes for this observation.
- BMI and Age have minor contributions, with less influence than Blood Glucose and HbA1c levels.
- Hypertension has a small negative effect, making little difference in the prediction.
- Green bars indicate risk-increasing features, while red bars represent risk-reducing factors.

This enhances model transparency, aiding in medical decision-making.

CONCLUSIONS

- Feature importance analysis identified HbA1c Level and Blood glucose level as the most influential in predicting diabetes, highlighting its critical role in the prediction process.
- Among the models tested, Random Forest achieved the highest accuracy in diabetes prediction.
- Future scope is to minimize the value of False Negative in the model using hyper-parameter tuning.

Accuracies of Models ➡



MAJOR REFERENCES

- Smith, J. W., et al. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Kaggle Link.
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
- Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.