Diabetes Prediction with Data Science

Presented by: Mohamed Abudabra

Supervisor: Nachiketa Chakraborty

University of Reading - BSc Computer Science

Date: May 8, 2025

Project Overview

Focus: Predicting Type 2 Diabetes using ML

Dataset: Pima Indian Diabetes Dataset

Methods: Data preprocessing, SMOTE, PCA, SHAP

Final Model: Gradient Boosting (GBM-DRU)

Accuracy: 87%

Aims and Objectives

Build interpretable, robust ML model

Handle class imbalance (SMOTE)

Feature selection with RFE and PCA

Explainability via SHAP, Permutation

Ethical and calibration evaluation

Data Preprocessing

- Outlier removal using IQR
- Handling missing/implausible values
- Feature scaling (StandardScaler)
- Balanced data using SMOTE

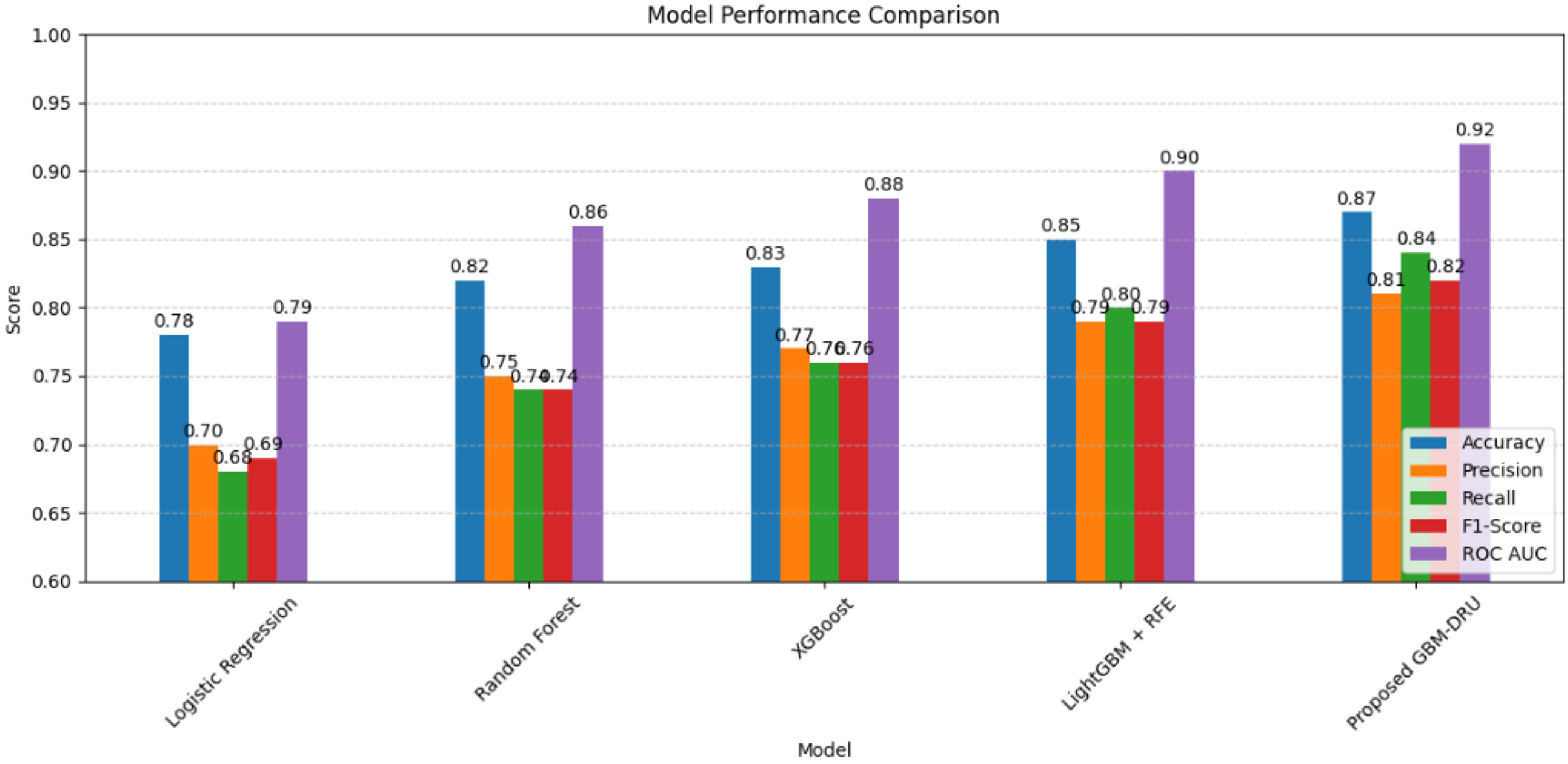
Model Development

Evaluated models: Logistic Regression, RF, XGBoost, LGBM

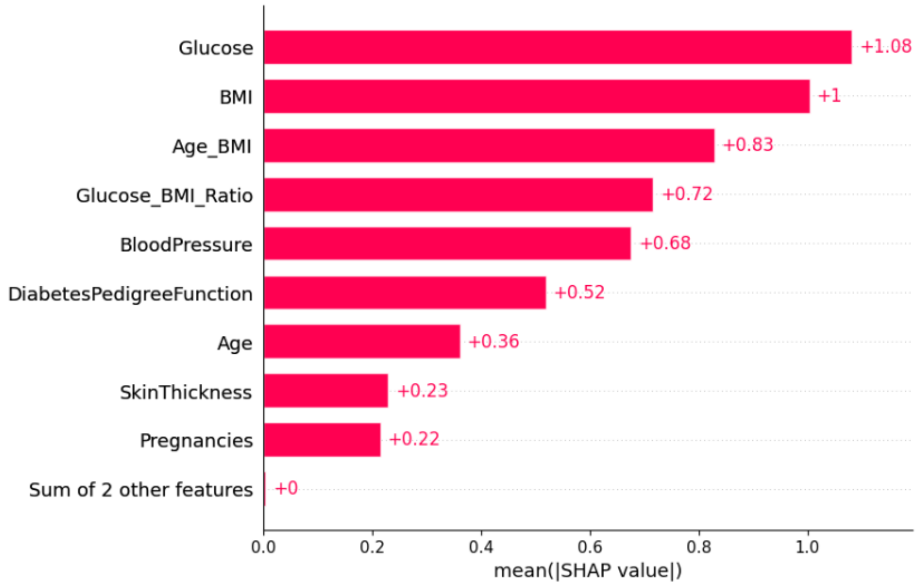
Final model: GBM-DRU = LGBM + PCA + SMOTE

Hyperparameter tuning with CV

Performance Metrics



Explainability



Tools: SHAP and Permutation Importance

Key predictors: Glucose, BMI, Age, Insulin

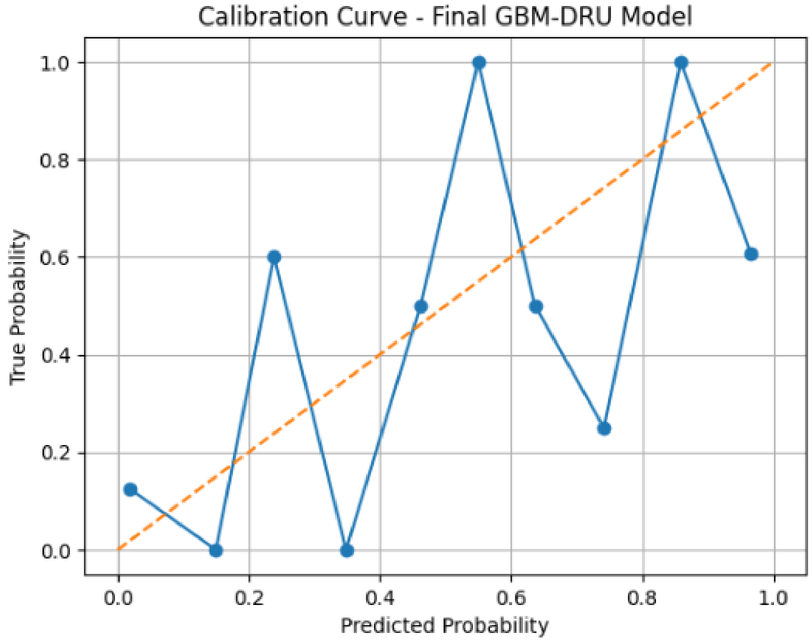
Supports clinical trust and transparency

Model Calibration

Calibration curves validate probability output

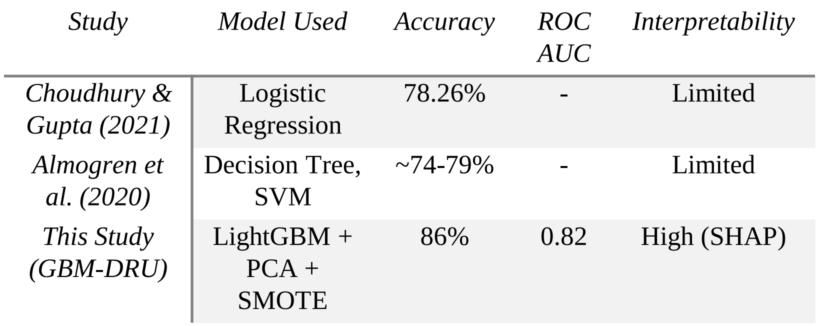
Slight underestimation at high confidence

Consider isotonic regression or Platt scaling



Comparison to Prior Research

- Choudhury & Gupta (2021): 78.26% (LogReg)
- GBM-DRU: 86% Accuracy
- Significant improvements in explainability and calibration



Study	*Model Used*	*Accuracy*	*ROC AUC*	*Interpretability*
Choudhury & Gupta (2021)	Logistic Regression	78.26%	-	Limited
Almogren et al. (2020)	Decision Tree, SVM	~74-79%	-	Limited
This Study (GBM-DRU)	LightGBM + PCA + SMOTE	86%	0.82	High (SHAP)

Conclusion

GBM-DRU outperforms traditional models

Methodologically rigorous and ethically sound

Ready for integration in clinical support systems

References

- Abid, A., Adelani, D., & Awoyemi, J. (2021). Explainable AI for medical diagnosis: A survey on datasets, methods, evaluation and challenges. International Journal of Environmental Research and Public Health, 18(13), 7346. https://doi.org/10.3390/ijerph18137346
- Almogren, A., Almazroi, A. A., Aljaffan, N., & Ali, M. (2020). Diabetes prediction using machine learning: Comparative study. Procedia Computer Science, 170, 376–381. https://doi.org/10.1016/j.procs.2020.03.065
- Choudhury, T., & Gupta, D. (2021). A machine learning approach to diabetes prediction and classification. International Journal of Environmental Research and Public Health, 18(13), 7346. https://doi.org/10.3390/ijerph18137346
- Kaggle. (n.d.). Pima Indians Diabetes Database. Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
- Kottwitz, S. (2021). LaTeX beginner's guide: Create visually appealing texts, articles, and books for business and science using LaTeX. Packt Publishing. ISBN: 9781801072588.
- Lamport, L. (1994). LATEX: A document preparation system: User's guide and reference manual. Addison-Wesley.
- University of Reading. (2023a). Avoiding unintentional plagiarism: Guidance on citing references for students at the University of Reading. https://libguides.reading.ac.uk/citing-references/avoidingplagiarism
- University of Reading. (2023b). Different styles & systems of referencing: Guidance on citing references for students at the University of Reading.

Thank You!

Questions?