**WHITE PAPER/REPORT**

**Predicting Diabetes Risk Using Machine Learning on Patient Health Indicators**

**By Serge Nane**

**DSC680 Applied Data Science**

**Fall 2025**

**Table of Contents**

1. Introduction .................................................... 3

2. Dataset and Methodology .......................................... 3

3. Results .......................................................... 3

4. Discussion ....................................................... 6

5. Ethical Considerations ........................................... 6

6. Conclusion and Future Work ....................................... 6

7. References ....................................................... 7

8. Appendix ......................................................... 7

9. 10 Anticipated Audience Questions ................................ 8

**1. Introduction**

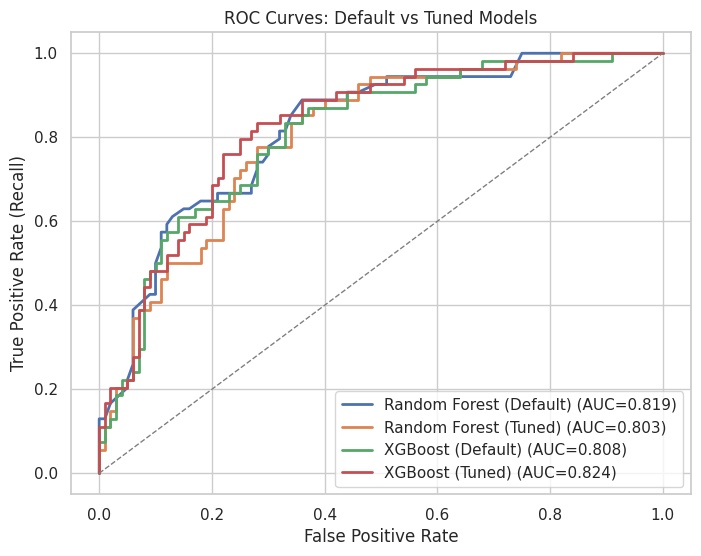
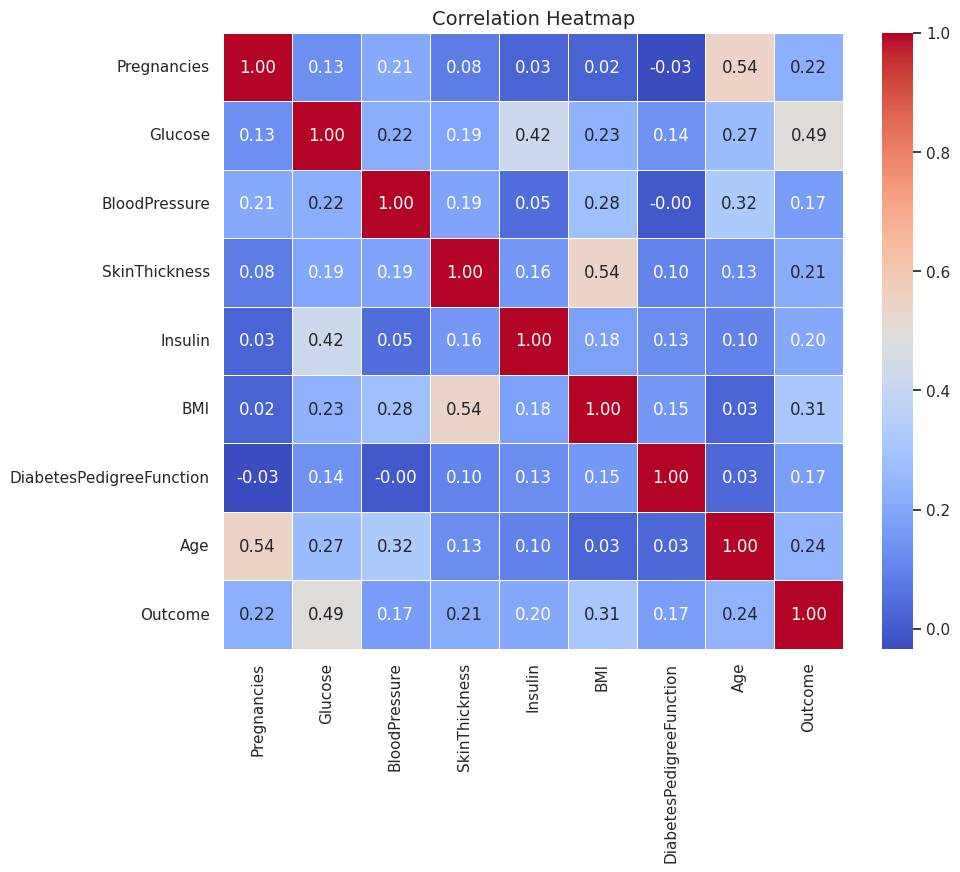
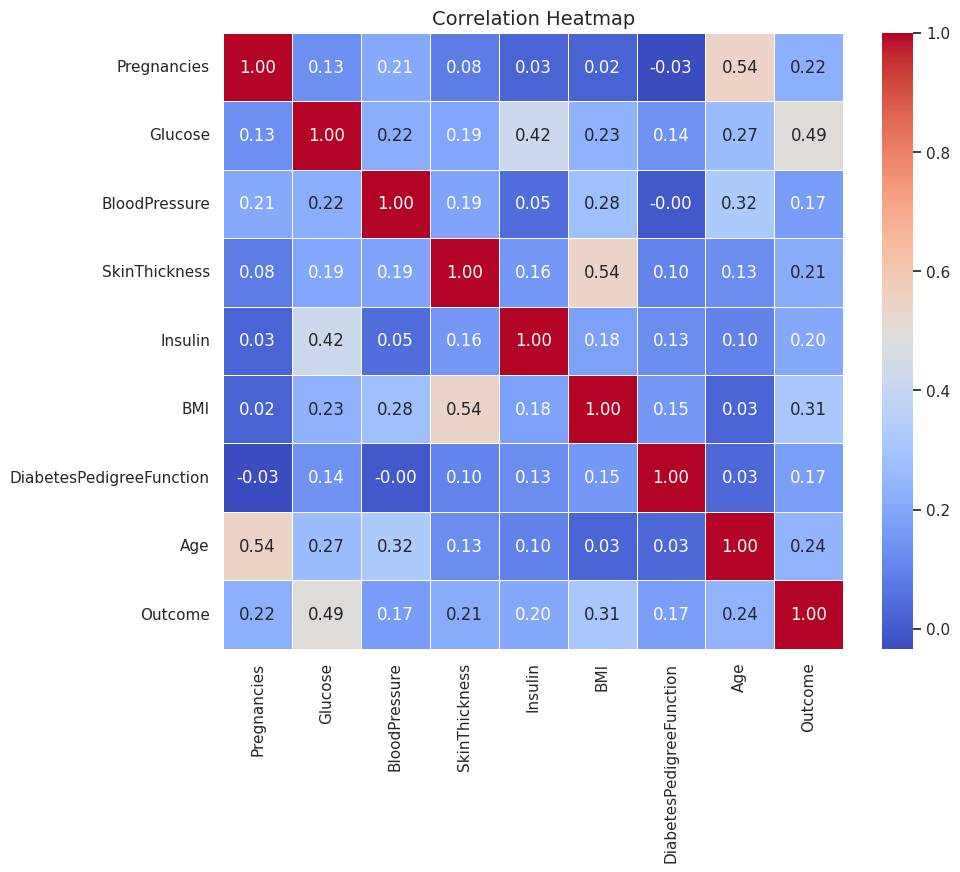
Diabetes is one of the most prevalent chronic diseases worldwide, leading to severe complications like cardiovascular disease, kidney failure, and blindness. Early identification of at-risk individuals is critical to reduce complications and healthcare costs. However, current diagnosis usually happens after symptoms appear, limiting prevention opportunities. This project applies machine learning on patient health indicators to predict diabetes onset. We evaluated Logistic Regression, Decision Tree, Random Forest, and XGBoost models to determine the most accurate approach and identify which features (like glucose, BMI, and age) are most influential in prediction.

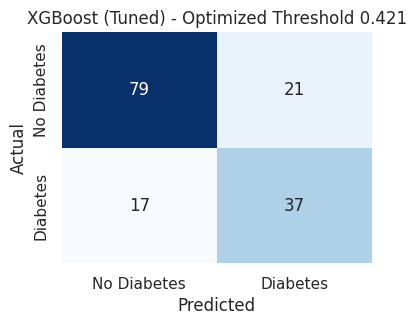
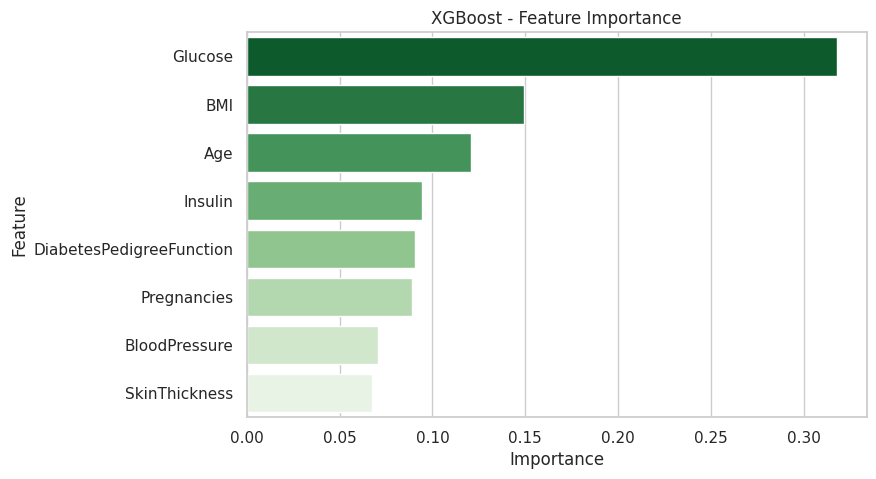
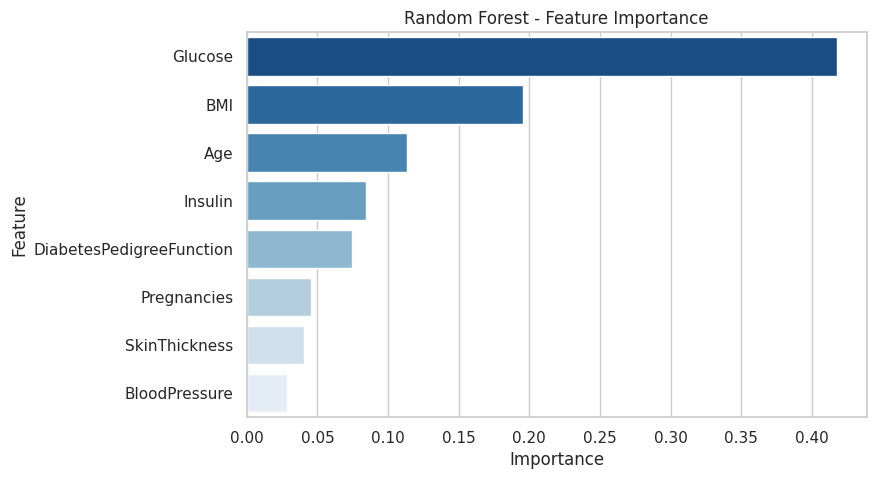
**2. Dataset and Methodology**

The Pima Indians Diabetes Database from UCI and Kaggle was used, consisting of 768 female patients with 8 predictors and 1 target variable. Preprocessing included handling missing values (replacing invalid zeros with medians), scaling features, and applying SMOTE oversampling to balance classes. Exploratory data analysis assessed feature distributions and correlations. Models (Logistic Regression, Decision Tree, Random Forest, XGBoost) were trained with an 80/20 split and evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC. Hyperparameters were tuned using Randomized Search, and thresholds were optimized to improve recall.

**3. Results**

**XGBoost** achieved the strongest performance (**Accuracy 0.75, ROC-AUC 0.824**), followed by Random Forest (**Accuracy 0.73, ROC-AUC 0.803**). Glucose, BMI, and Age were the most predictive variables. Correlation analysis showed that Glucose was strongly associated with the diabetes outcome. ROC curves confirmed XGBoost outperformed other models, and a confusion matrix showed improved positive class detection after threshold tuning. Feature importance plots showed Random Forest relied heavily on Glucose while XGBoost distributed weights more evenly, improving generalization.





**4. Discussion**

These results show that machine learning can support early diabetes risk detection. XGBoost outperformed others due to balanced feature weighting and resistance to overfitting. While promising, the small dataset limits generalizability, and model interpretability should be improved before clinical deployment.

**5. Ethical Considerations**

* **Data Privacy**: The Dataset is anonymized, and handling sensitive health data must follow privacy regulations such as HIPAA.
* **Bias and Fairness:** The model is trained on a specific demographic (Pima Indian women). Performance may not generalize equitably to other populations without retraining on diverse data.
* **Interpretability**: While feature importance provides some insight, the "black-box" nature of complex models like XGBoost can hinder trust. Techniques like SHAP analysis should be used to provide clearer explanations for predictions.
* **Responsible Use**: This model must be framed as a screening tool, not a diagnostic tool. Its purpose is to flag at-risk individuals for further confirmatory testing by a medical professional, never to replace clinical judgment.

**6. Conclusion and Future Work**

XGBoost showed the highest performance and balanced use of features, making it a strong decision-support tool for predicting diabetes risk. Future work should include larger datasets, additional risk factors, and explainability tools like SHAP to build clinician trust.

**7. References**

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care, 261–265.

UCI Machine Learning Repository. (n.d.). Pima Indians Diabetes Database. Retrieved from <https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes>.

Kaggle. (n.d.). Pima Indians Diabetes Database. Retrieved from <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>.

**8. Appendix**

* Full code available in the attached Jupyter Notebook ().
* The detailed model performance metrics table for all models has been moved here from the Results section:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | 0.71 | 0.58 | 0.67 | 0.62 | 0.811 |
| **Decision Tree** | 0.70 | 0.57 | 0.61 | 0.59 | 0.681 |
| **Random Forest** | **0.75** | 0.63 | **0.72** | 0.67 | 0.814 |
| **XGBoost** | 0.73 | 0.60 | 0.65 | 0.62 | 0.808 |

**9. 10 Anticipated Audience Questions**

1. Why was the Pima dataset chosen?

2. How was class imbalance addressed?

3. Why were these specific algorithms selected?

4. How was overfitting prevented?

5. How were missing values handled?

6. Which feature had the highest impact on predictions?

7. How could this model be validated in real clinical settings?

8. What are the ethical concerns regarding prediction errors?

9. Could this approach be adapted to other chronic diseases?

10. What steps are needed before deploying such a model in hospitals?