Final Report
BAIT 509 Business Applications of Machine Learning

Team Name: Allen Yu, Athena Xu, Gwen Fang, Mia Ling, Wendy Liu

Student ID: 30979249, 53905113, 28843514, 62106489, 301385516

Date: 2025 Feb 14

4.1 Background, Motivation, and Business Question

4.1.1 Background

Diabetes is a growing global health concern, with over 537 million individuals currently living with the condition. Early diagnosis and intervention are critical for preventing complications, reducing healthcare costs, and improving patient outcomes. Healthcare providers play a pivotal role in this process, as they can leverage predictive analytics to identify high-risk individuals early, allowing for timely interventions and personalized care plans.

4.1.2 Motivation

Healthcare systems often struggle with resource allocation, especially when it comes to chronic disease management. By integrating machine learning models into routine screening processes, healthcare providers can:

- Improve Screening Efficiency: Identify individuals at high risk without relying heavily on invasive lab tests.
- Enhance Preventive Care: Provide early, data-driven recommendations to mitigate the development of diabetes.
- Optimize Resource Allocation: Focus diagnostic and treatment resources on patients who are most likely to benefit from intervention.

4.1.3 Business Question

How can predictive analytics assist healthcare providers in identifying individuals at high risk for diabetes, enabling early interventions and better resource allocation?

4.2 Data and Statistical Questions

4.2.1 Data Description

The dataset used in this analysis is sourced from the Kaggle Diabetes Prediction Dataset (link).
This dataset includes demographic, behavioral, and physiological attributes that contribute to diabetes risk prediction.

Features include:

- **Demographic Features:** Gender, Age
- **Health Indicators:** Hypertension, Heart Disease, BMI, HbA1c Level, Blood Glucose Level

- **Categorical Features**: BMI Category (Normal, Overweight, Obese), Age Group (Middle-Aged, Senior)
- **Behavioral Features:** Smoking History (Current, Ever, Former, Never, Not Current)
- **Derived Features:** BMI Squared, HbA1c Level Squared, Interaction Terms (Hypertension & Heart Disease)
- **Outlier Indicators:** Flags for Age, BMI, HbA1c Level, Blood Glucose Level

The target variable is Diabetes (0 = No, 1 = Yes), making this a binary classification problem.

4.2.2 Statistical Question

To address the business problem of "How can predictive analytics help healthcare providers identify high-risk diabetes patients?", the statistical question we focus on is:

"Which features have the highest predictive power for determining diabetes risk in patients?"

This question helps to identify the most influential factors contributing to diabetes risk, aiding in targeted interventions, policy decisions, and personalized healthcare strategies.

4.2.3 Why is this Statistical Question Useful?

Improving Predictive Accuracy

- Identifying key predictors (e.g., **Blood Glucose Level, BMI, Smoking History**) refines our model, improving early detection of diabetes.
- Focusing on the most significant features helps simplify the model while maintaining accuracy, making predictions more reliable.

Supporting Preventive Healthcare

- Recognizing modifiable risk factors (e.g., BMI, Smoking History) enables early lifestyle interventions.
- Patients at risk can take preventive actions like dietary changes and regular glucose monitoring before reaching a critical health stage.

Cost-Effective Screening

- Using non-invasive indicators (such as BMI and smoking status) allows for lower-cost screening methods, reducing reliance on expensive tests.
- Healthcare providers can focus resources on high-risk individuals, improving efficiency.

Business & Policy Applications

- **Insurance Providers:** Can adjust health coverage plans based on risk factors.
- **Healthcare Providers:** Can implement targeted screening programs for early diabetes detection.
- **Public Health Initiatives:** Can focus on reducing preventable risk factors, improving community health outcomes.

4.2.4 Limitations of This Statistical Question

- **Correlation vs. Causation:** The analysis identifies correlations but does not confirm causation.
- **Feature Overlap:** Some engineered features (e.g., HbA1c Level Squared) may introduce redundancy.
- **Unaccounted External Factors:** Variables like diet, exercise, genetic predisposition, and socioeconomic factors are missing.
- **Potential Model Bias:** If the dataset is not diverse, the model may **not generalize well** to other populations.

4.2.5 Feature Engineering to Address the Question

- **Handling Missing Values**: Imputation for missing demographic and health indicators.
- **Removing Duplicates**: Ensuring data integrity by eliminating redundant entries.
- **Categorization:** BMI and Age were transformed into categorical variables.
- **Feature Engineering Enhancements**:
  - **Polynomial Features**: Squaring continuous variables (e.g., BMI^2, $HbA1c^2$) to capture nonlinear relationships.
  - **Interaction Terms**: Combining hypertension and heart disease to examine compounding effects.
  - **Outlier Detection**: Creating flags to identify extreme values in key health indicators.

4.3 Exploratory Data Analysis (EDA)

4.3.1 Dataset Overview

The dataset used in this study is sourced from the Pima Indians Diabetes Dataset (UCI) and contains 100,000 records with 9 features. The goal is to predict diabetes occurrence based on demographic, behavioral, and physiological attributes.

4.3.2 Feature Breakdown

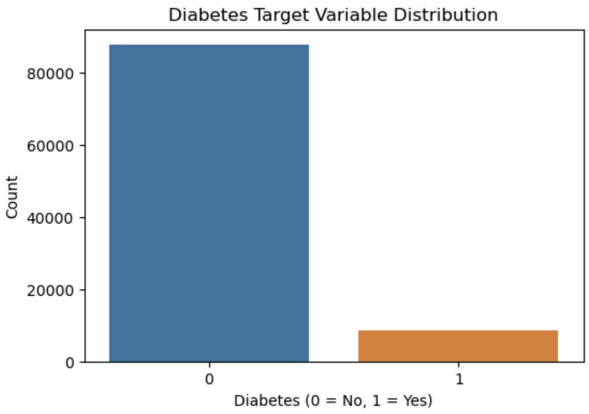
1. **Categorical Features:**

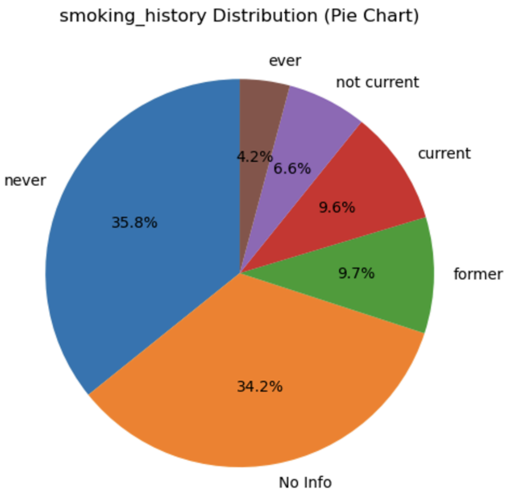
    - Gender: Male, Female
    - Smoking History: Never, Current, Former, etc.
2. **Numerical Features:**
    - Age: Continuous variable representing patient age.
    - BMI: Body Mass Index, indicating body weight relative to height.
    - HbA1c Level: Long-term blood sugar indicator.
    - Blood Glucose Level: Instantaneous blood glucose measurement.
3. **Binary Features:**
    - Hypertension: 0 = No, 1 = Yes
    - Heart Disease: 0 = No, 1 = Yes
    - Diabetes (Target Variable): 0 = No Diabetes, 1 = Diabetes

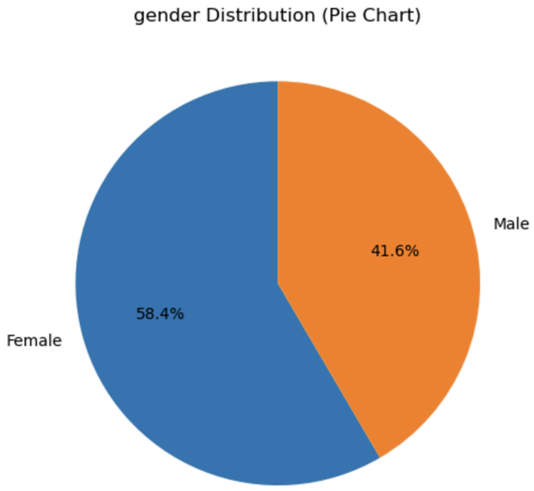
4.3.3 Data Distributions

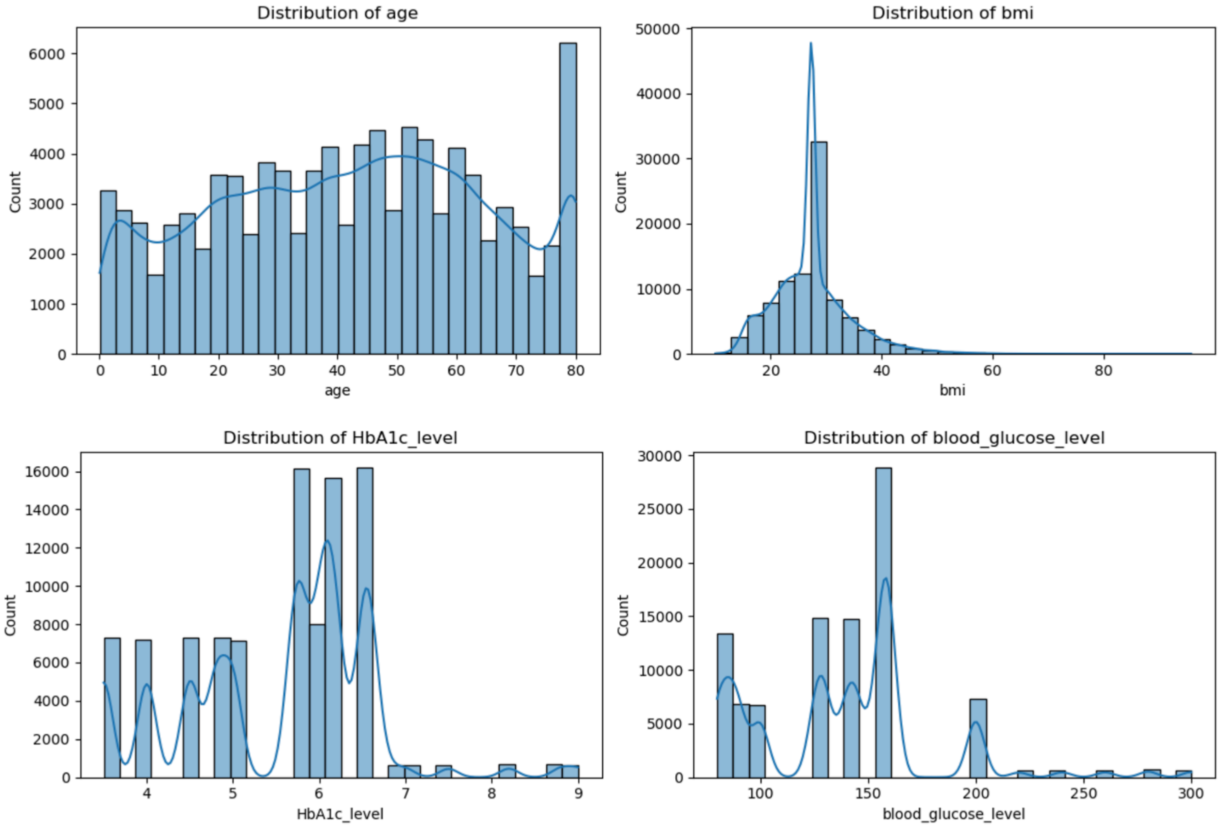
Visualizations:

Below are key visualizations that illustrate the distribution of important features:









Key Observations from Feature Distributions:

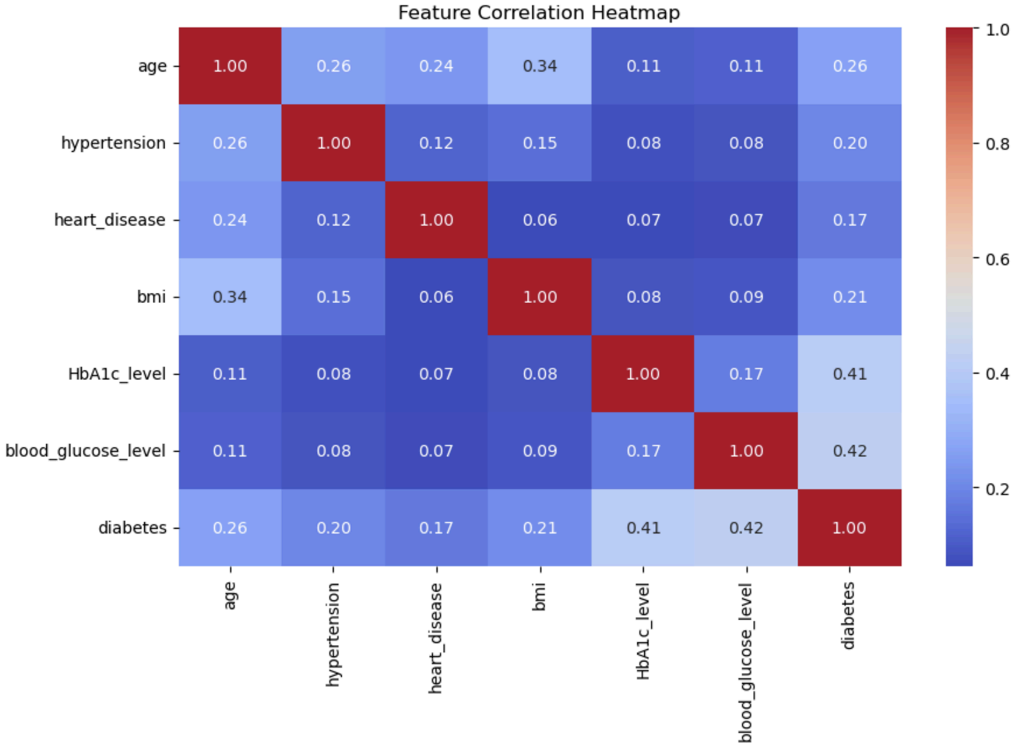
- **Age Distribution:** Skewed towards middle-aged and older individuals, peaking around 80 years, with a great proportion of younger individuals as well.
- **BMI Distribution:** Most values fall between 20-40, with a concentration around 25-30, but some outliers exceed 80.
- **HbA1c & Blood Glucose Levels:** Show multiple peaks, indicating potential subgroups in the data and suggesting different risk profiles.
- **Smoking History:** The dataset shows a relatively balanced distribution across smoking status categories, but there is a notable group with missing or unknown smoking history.

Takeaways:

- The skewness in Age and HbA1c levels suggest possible transformations later on.
- There is the presence of outliers in BMI and glucose levels.

4.3.4 Data Visualization

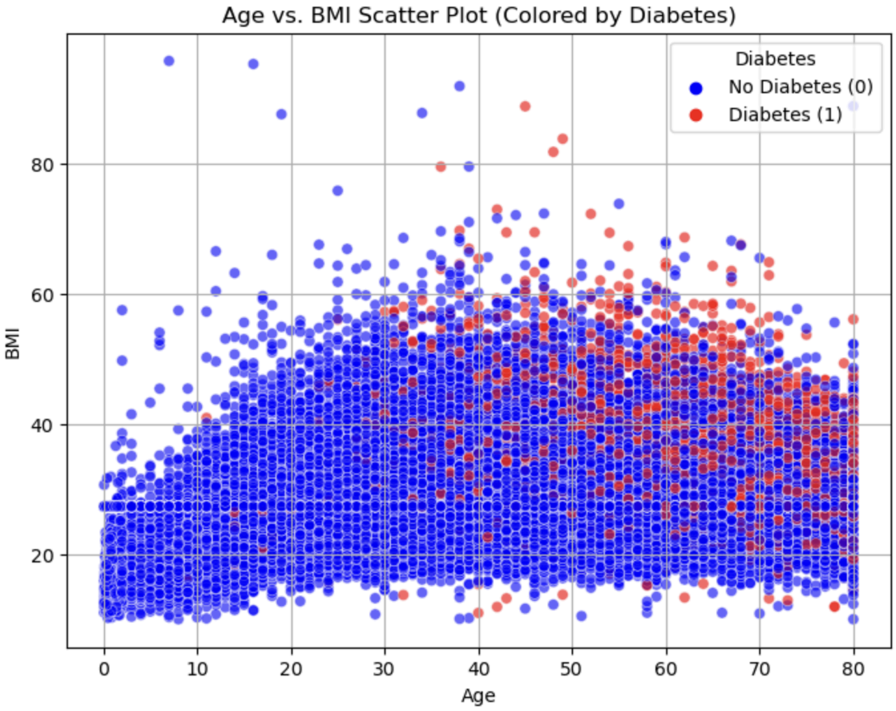
Correlation Analysis & Exploration of the relationship between age, BMI, and diabetes



The correlation heatmap reveals key relationships between features:

- Strongest Predictors of Diabetes:
  - Blood Glucose Level (0.42 correlation with diabetes)
  - HbA1c Level (0.41 correlation with diabetes)
  - Age (0.26 correlation)
  - BMI (0.21 correlation)
  - Hypertension (0.20 correlation)
- Feature Independence: No extreme multicollinearity detected, ensuring diverse predictive variables.
- The correlation heatmap shows that blood glucose and HbA1c levels have the strongest correlation with diabetes, indicating that blood glucose and HbA1c levels could be the most relevant predictors.

Out of curiosity, we also created a scatter plot to explore the relationship between age, BMI, and diabetes. We can see that individuals with higher BMI, particularly those above 30, tend to show higher diabetes prevalence (red dots) as shown below.



4.3.5 Key Considerations for Model Building

- Class Imbalance: Since diabetic cases are underrepresented, balancing techniques could be necessary.
- Nonlinear Relationships: HbA1c and Blood Glucose show non-linear trends, suggesting models like Random Forest may perform well.
- Feature Engineering Potential: Interaction terms (e.g., Hypertension + Heart Disease) and polynomial transformations (e.g., BMI^2, $HbA1c^2$) could enhance predictive power.

4.4 Method & Results

To build our machine learning models, we followed a structured workflow consisting of feature preprocessing, feature selection and engineering, performance metrics selection, model selection, hyperparameter tuning, and model performance comparison.

4.4.1 Feature Preprocessing & Engineering

We processed and refined features by handling missing values, encoding categorical variables, performing feature selection and engineering, scaling numerical features, and downsampling before splitting the dataset into training and testing sets.

1. **Handling Missing Values:** Although our dataset was already cleaned and contained no missing values, we applied *SimpleImputer* with a median strategy to handle potential missing values in future datasets.
2. **Feature Engineering:** To enhance model performance, we created the following new variables:
    a. **BMI Categories:** We categorized raw BMI values into groups based on standard guidelines: underweight (<18.5), normal (18.5–24.9), overweight (25–29.9), and obese (≥30). This categorization helps the model capture distinct diabetes patterns based on weight ranges rather than dispersed BMI values.
    b. **Age Groups:** We grouped raw age into three categories: youth (18–29 years), adults (30–59 years), and seniors (60+ years). This allows the model to capture relationships between diabetes risk and age patterns across different life stages.
    c. **Interaction Features:** We created an interaction feature combining hypertension and heart disease to capture their combined effect on diabetes risk. This reflects the reality where the co-occurrence of these conditions may amplify risks.
    d. **Polynomial Features:** Based on exploratory data analysis (EDA), we replaced some important features with their polynomial terms. Given the strong correlation between HbA1c levels, glucose levels, and BMI with the target variable, we incorporated these nonlinear relationships to improve prediction accuracy. Iterative testing confirmed that including these features enhanced model performance.
3. **Encoding Categorical Variables:** We used *LabelEncoder* to convert categorical variables into numeric form. Encoded variables included:
    a. **Gender**: 0 for female, 1 for male.
    b. **Smoking Status**: 0 for non-smoker ("never"), 1 for former smoker ("not current," "former," "ever"), 2 for current smoker ("current"), and -1 for no information.
4. **Scaling:** Before splitting the data into training and testing sets, we standardized numerical features using *StandardScaler* to ensure consistent scaling. Scaled variables included BMI, age, HbA1c levels, and glucose levels.
5. **Downsampling**: To address class imbalance (90% non-diabetes vs. 10% diabetes), we downsampled the majority class to 15,000 samples, resulting in a more balanced dataset.

4.4.2 Performance Metrics Selection

To address the business questions, we selected performance metrics suitable for a healthcare setting. Our primary objective was to identify as many potential diabetic patients as possible while minimizing false negatives. Our secondary objective was to ensure the accuracy of these predictions to reduce unnecessary consultations and care for false positives. Consequently, we prioritized the F1-score, which balances precision and recall.

1. **Primary Metric**: F1-score, prioritizing the balance between recall and precision.

2. **Secondary Metrics**:
    a. Recall: Prioritizing the reduction of false negatives (missed cases).
    b. Precision: Focusing on reducing false positives (unnecessary interventions).
3. **Other Metrics for Reference**: Accuracy and ROC-AUC, which provide insights into the overall balance and performance of the models.

4.4.3 Model Selection

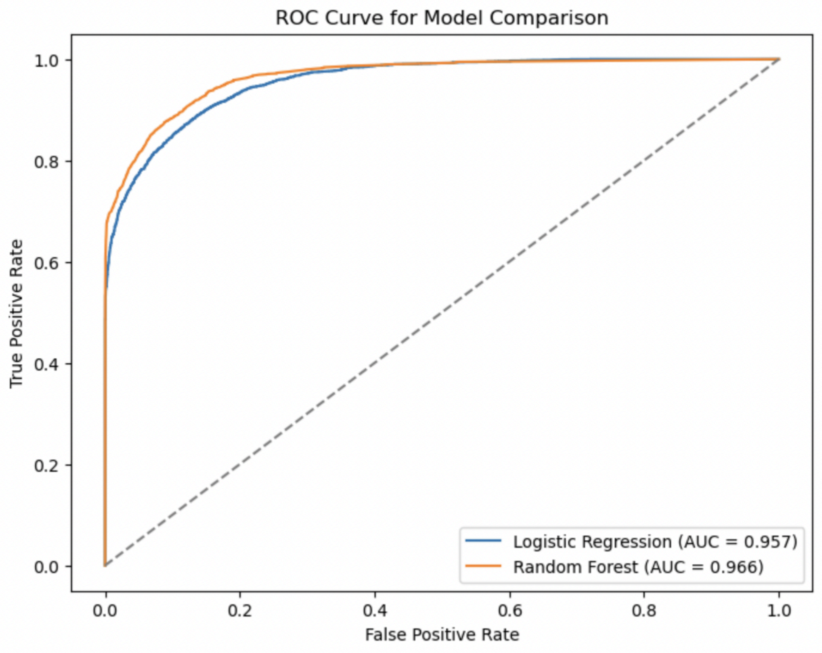
We selected two popular classification models for our initial investigation: Random Forest and Logistic Regression.

- **Random Forest:**
    - Capable of capturing complex relationships that linear models like logistic regression cannot.
- **Logistic Regression:**
    - Easier to interpret and explain to stakeholders, as it provides insights into the impact of individual features on predictions.
    - Outputs a probability score, which can be interpreted as the risk of having diabetes, making it clinically intuitive.

We built both models and evaluated their performance without hyperparameter tuning to establish a baseline understanding. The results are summarized below:

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	0.8866	0.8719	0.8043	0.8359	0.9568
Logistic Reg.	0.8997	0.8712	0.8478	0.8579	0.9658

We further validated the models' performance by plotting ROC curves, which confirmed similar performance trends.

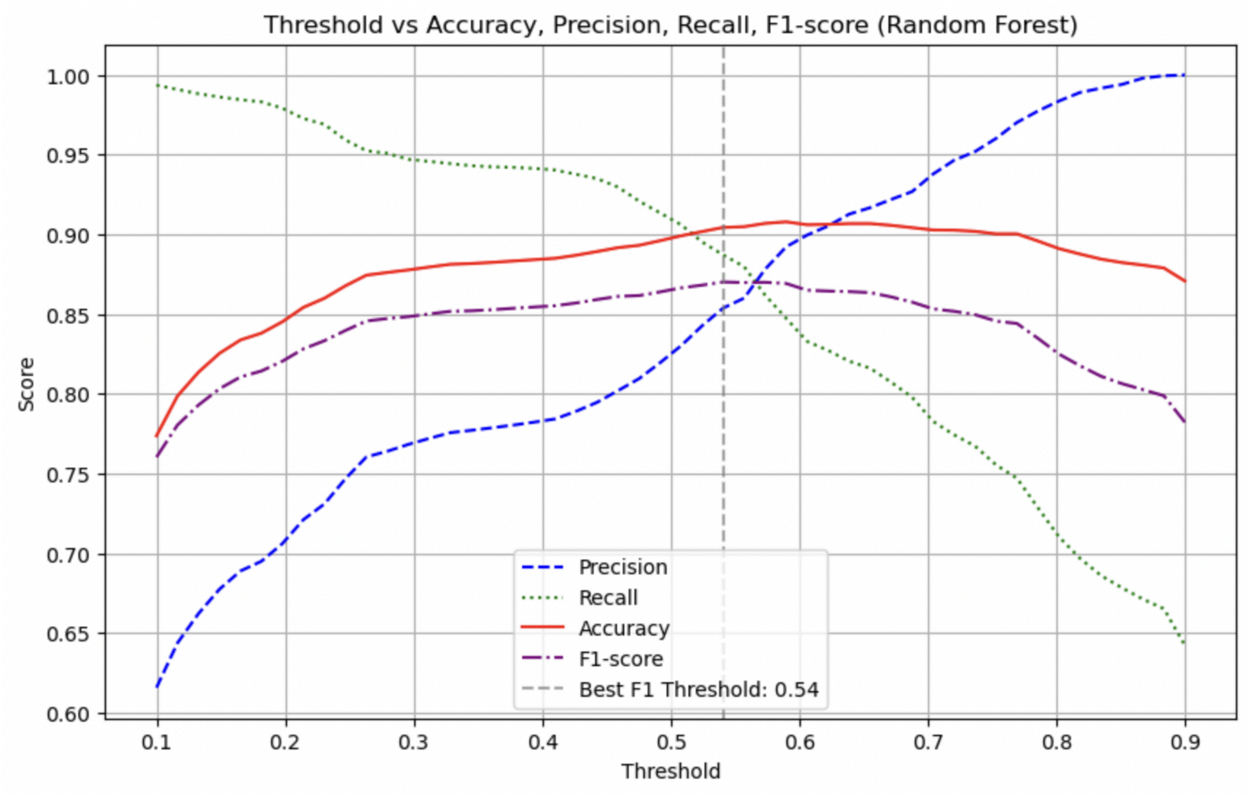


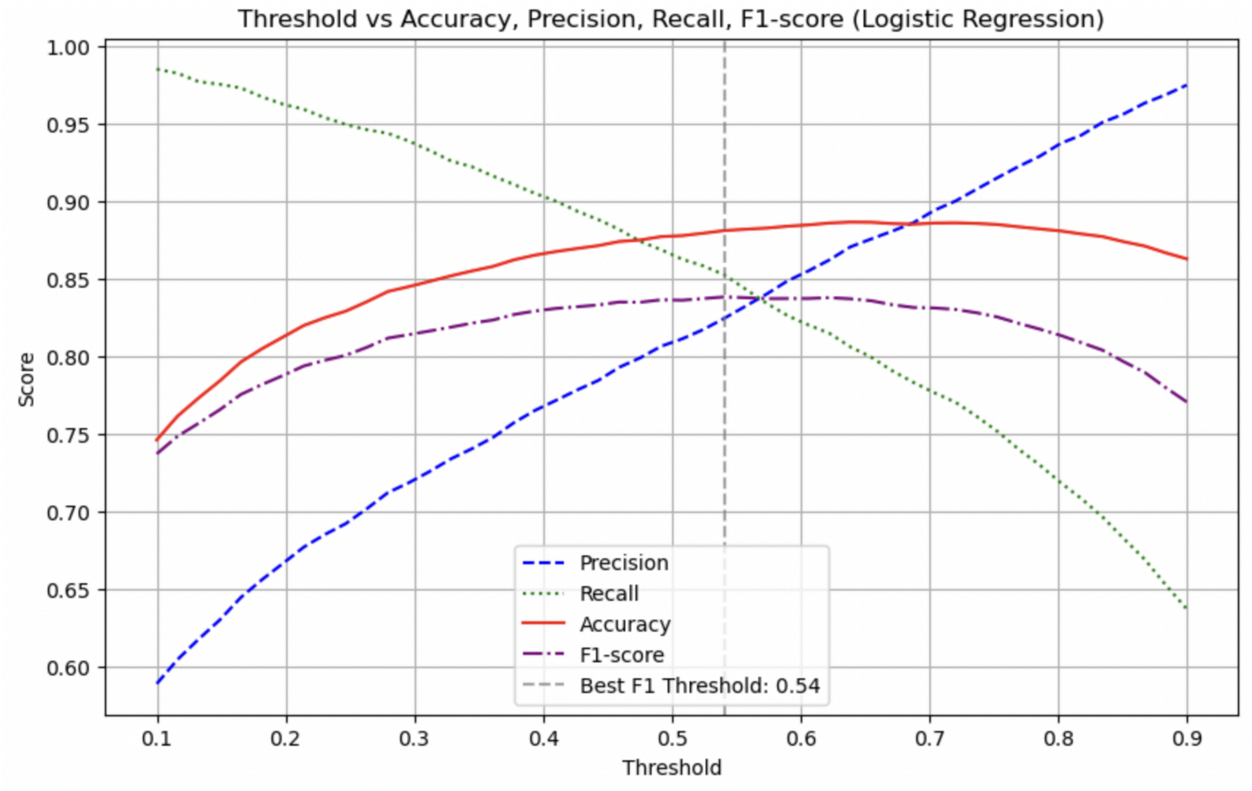
4.4.4 Hyperparameter Tuning & Prediction Threshold Tuning

We performed hyperparameter tuning using *RandomizedSearchCV* with 7-fold cross-validation to identify optimal hyperparameters for both models.

- **Random Forest**: The parameters adjusted included n_estimators, min_samples_split, min_samples_leaf, max_features, and max_depth.
- **Logistic Regression**: The parameters adjusted included solver, penalty, max_iter, and C.

Additionally, we tuned the decision threshold to optimize performance metrics. By plotting the relationship between metrics and thresholds, we selected a threshold of **0.54**, which maximized the F1-score.





4.4.5 Results Comparison After Tuning

After tuning, the models' performance improved as follows:

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Tuned Random Forest	0.9039	0.8527	0.8872	0.8696	0.9724
Tuned Logistic Reg.	0.8811	0.824	0.8531	0.8383	0.9568

4.4.6 Conclusion

The **Tuned Random Forest** was selected as the final model due to its higher **F1-score (0.8696)** compared to Logistic Regression (0.8383). Additionally, Random Forest outperformed Logistic Regression across all key metrics (precision, recall, accuracy, and ROC-AUC), making it the most suitable model for deployment. Continuous monitoring and retraining are recommended as new data becomes available.

However, we acknowledge that the performance gap between the two models was not significant. With additional feature engineering or data manipulation, Logistic Regression could be further improved. Its simplicity and interpretability remain valuable, particularly for providing insights into patient risk probabilities, which can inform clinical decision-making. Future research should explore enhancing Logistic Regression for deployment alongside Random Forest, ensuring a balance between predictive performance and interpretability.

4.5 Communication of Results and Advice to a Non-expert

This section explains how our machine learning model can be used in real-world healthcare settings to assist medical professionals in identifying diabetes risk. Our approach aims to enhance early detection, improve resource allocation, and reduce reliance on manual assessments. The proposed system automates the diabetes screening process by integrating machine learning into clinical workflows.

4.5.1 Two-Step Risk Screening Approach for Healthcare Providers

Our approach consists of two main steps: a preliminary risk estimation using logistic regression and a final diagnosis using a more complex model once lab test data is available.

Step 1: Initial Screening with Logistic Regression (No Lab Tests Required)

- The first step provides a fast and accessible risk assessment without requiring blood tests.
- This model estimates a patient's probability of diabetes based on:
  - Age
  - BMI (calculated from height & weight)
  - Hypertension history
  - Heart disease history
  - Smoking status
  - Gender
  - Interaction between hypertension and heart disease
- These details can be collected through patient questionnaires or during routine checkups.
- The logistic regression model calculates a risk probability score, which is used to decide whether further testing is necessary.

Step 2: Lab Testing & Advanced Prediction with Random Forest

- If a patient's predicted diabetes risk exceeds a threshold (e.g., 50%), they are recommended for further lab testing.
- The lab tests include:
  - HbA1c levels (long-term blood sugar levels)
  - Blood glucose levels (current sugar levels)

- These new test results are then fed into the Random Forest model, which makes a final diabetes prediction.

This two-step approach is designed to reduce unnecessary lab testing while ensuring high-risk patients receive further evaluation.

4.5.2 Model Performance: With vs. Without Lab Test Data

We tested two versions of our model:

1. Using only easily available patient data (excluding HbA1c & glucose)
2. Including lab test data (HbA1c & glucose levels)

Performance Without Lab Tests

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.7593	0.6799	0.6308	0.8269
Logistic Reg.	0.7182	0.6133	0.5955	0.7773

- The logistic regression model performs better than the random forest model when lab test data is not included.
- The AUC of 0.827 suggests that the model is effective even without blood test data, making it suitable for early-stage risk screening.

Performance with Lab Test Data

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.9039	0.8527	0.8872	0.8696
Logistic Reg.	0.8811	0.824	0.8531	0.8383

- Including HbA1c and blood glucose levels significantly improves performance.
- The random forest model achieves an F1-score of 0.8696, meaning it effectively detects diabetes cases while minimizing false positives.
- This confirms that lab tests remain crucial for highly accurate diabetes prediction.

4.5.3 Practical Applications in Healthcare Settings

This system is highly adaptable and can be integrated into various medical workflows, including primary care clinics, telemedicine platforms, and community health programs.

Use Case 1: Primary Care Clinics

- A patient visits their primary care doctor for a routine checkup.
- The doctor inputs basic health information into the logistic regression model.
- If the predicted diabetes risk exceeds 50%, the patient is recommended for blood tests.
- Once test results are available, they are entered into the random forest model for final risk assessment.
- This optimizes lab testing resources and ensures high-risk patients receive early intervention.

Use Case 2: Telemedicine & Remote Healthcare

- A healthcare provider offers online consultations for patients in rural or underserved areas.
- Patients submit their age, BMI, smoking status, and medical history through an online form.
- The logistic regression model provides a preliminary diabetes risk score.
- High-risk patients are referred to local clinics for lab testing.
- This expands access to early diabetes screening without requiring immediate in-person visits.

Use Case 3: Workplace & Community Health Programs

- A company introduces an employee health screening program.
- Employees fill out a short health questionnaire.
- The logistic regression model flags individuals at high risk.
- The company provides subsidized lab tests for those flagged, ensuring early detection.
- This helps corporate wellness programs reduce long-term healthcare costs and support employee well-being.

4.5.4 Advantages of Automating the Screening Process

- Reduces reliance on doctors for early screening. Traditionally, experienced doctors manually evaluate risk factors. With this model, screening is automated, allowing nurses or non-specialists to conduct preliminary assessments.
- Optimizes healthcare resources. By only referring high-risk patients for lab tests, the system reduces unnecessary testing costs and improves efficiency.

- Provides faster and more accessible screening. The first screening step requires no lab tests, making it usable in clinics, pharmacies, and telemedicine services. This allows wider coverage and helps detect diabetes earlier, particularly in resource-limited settings.

4.5.5 Limitations & Future Improvements

- Data imbalance and model generalization. The dataset contains fewer diabetes-positive cases, which may limit its ability to generalize. Future work could collect more diverse patient data to improve robustness.
- Reliance on lab tests for final decision. The most accurate predictions require HbA1c and blood glucose tests. Future research could explore non-invasive alternatives, such as wearable glucose monitors or dietary tracking.
- Customizing decision thresholds. The 50% threshold for lab test referrals can be adjusted based on medical guidelines. Future improvements could include personalized thresholds based on age, BMI, and pre-existing conditions.

4.5.6 Next Steps & Deployment Considerations

- Deploy the logistic regression model in clinics for initial screening.
- Develop an automated system to flag high-risk patients for lab tests.
- Refine risk thresholds based on real-world healthcare data.
- Expand the dataset to improve model generalization before full deployment.

Reference

Mustafa, T. Z. (n.d.). *Diabetes Prediction Dataset*. Kaggle. Retrieved February 14, 2025, from https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data

OpenAI. (2024). *ChatGPT* (Feb 2024 version) [Large language model]. Retrieved February 14, 2025, from https://chat.openai.com