GlucoSense: AI-Powered Diabetes Detection for Early Intervention

Name of Presenter: Aastha Singh

🔗 Project Repository

Introduction

Problem Statement

- Diabetes is a growing global health issue, with cases rising significantly over the past 15 years due to lifestyle factors.
- Early detection can prevent complications and improve health outcomes.

Objective

- To develop an AI-powered model that predicts a person's diabetes status (Healthy, Pre-Diabetic, or Diabetic) using healthcare and lifestyle data.

Significance

- Supports early intervention, reducing healthcare costs and improving patient well-being.
- Empowers healthcare providers to make data-driven decisions for patient care.

Methodology

Data Preparation:

- Clean data (handle missing values, duplicates, outliers) and transform features (encode categorical, normalize numerical).

Feature Selection:

- Use Correlation Analysis & Feature Importance (Random Forest, Decision Trees) to identify key features.

Model Development:

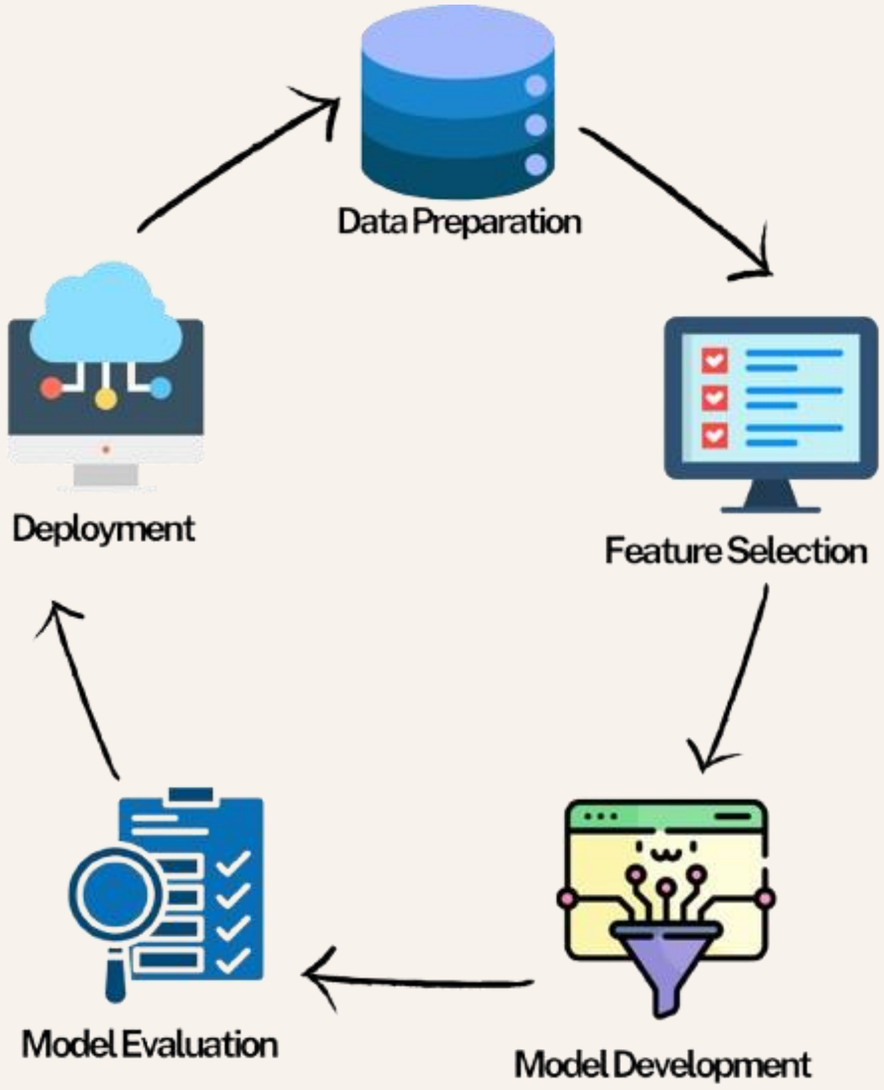
- Train models (Logistic Regression, Random Forest, Decision Trees), and optimize using Cross-Validation & Hyperparameter Tuning.

Model Evaluation:

- Evaluate with Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Select the best model.

Deployment & Presentation:

- Deploy the best model and document key findings and recommendations.



Data Exploration

Dataset Overview:

- Total Records: 520 rows
- Total Features: 17 columns (16 input, 1 output)
- Purpose: Predict diabetes based on healthcare and lifestyle factors.

Key Features:

- Age (Numerical): Age of the individual
- Gender (Categorical): Male/Female
- Symptoms (Binary): Polyuria, Polydipsia, etc.
- Lifestyle Factors (Binary): Obesity, Alopecia, Muscle Stiffness, etc.

Data Insights:

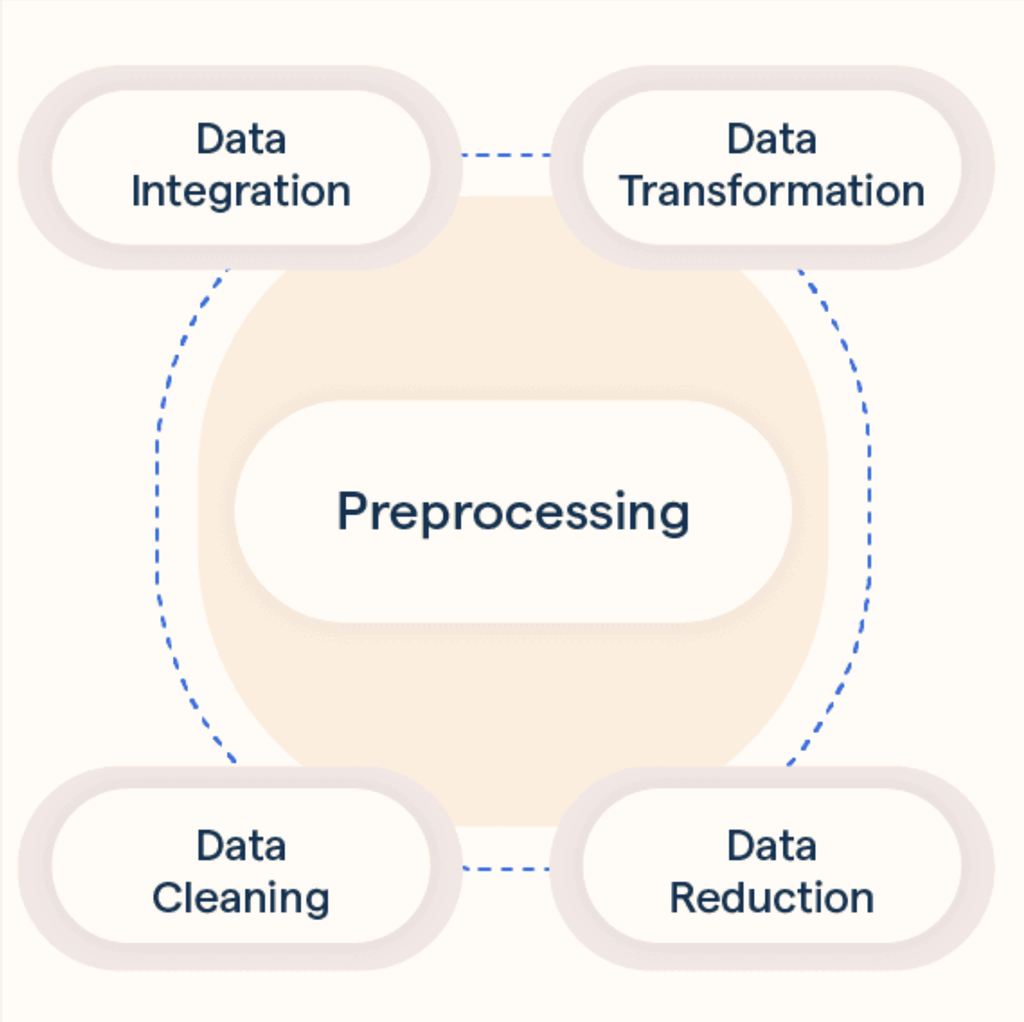
- Class Distribution: Likely imbalance (more non-diabetic cases).
- Data Quality: No missing values; binary features (0/1) except Age.
- Preprocessing: Encode Gender (Male = 1, Female = 0) and normalize Age.

Data Preprocessing

01. Handling Missing Data: Impute missing values using mean/median or predictive models.
No Missing values

02. Removing Duplicates: Identify and remove any duplicate records to ensure data quality.
Number of duplicate rows: 269

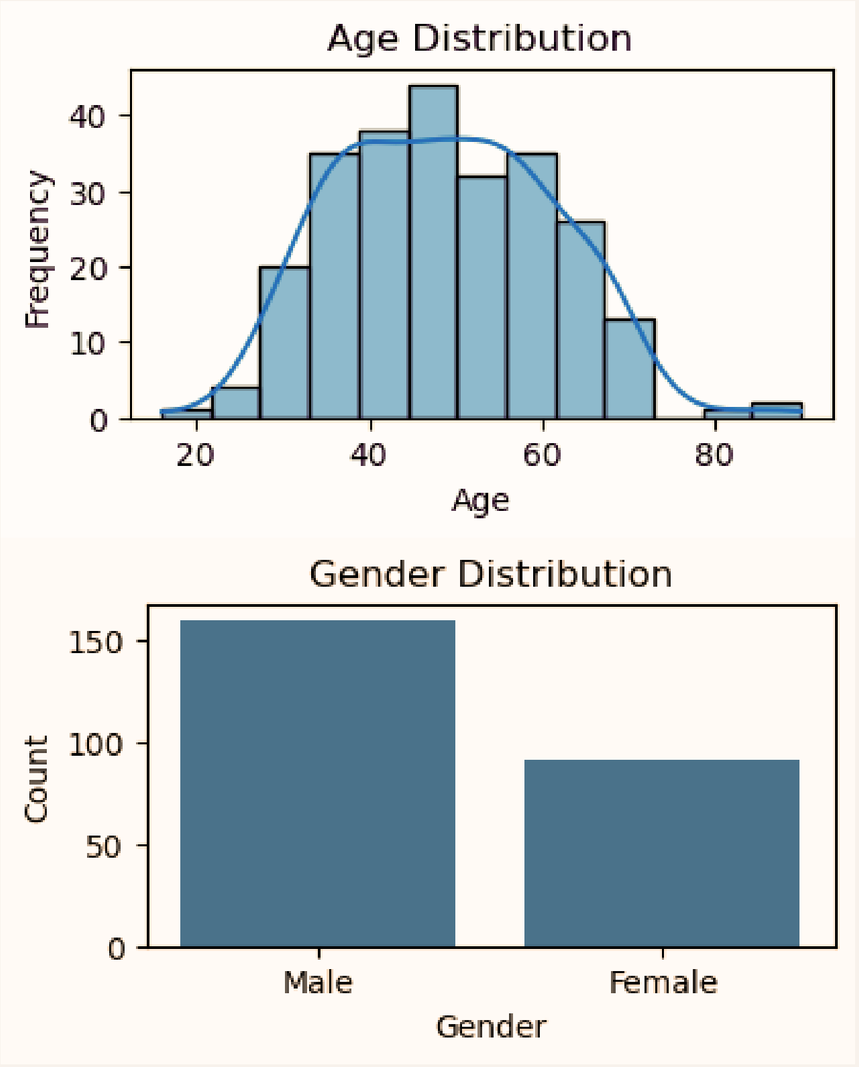
03. Encoding Categorical Data:
Convert Gender to binary: Male = 1, Female = 0.
Encode other binary features: Polyuria, Polydipsia.



Exploratory Data Analysis

Purpose: Highlight the patterns, trends, and distributions in the data.

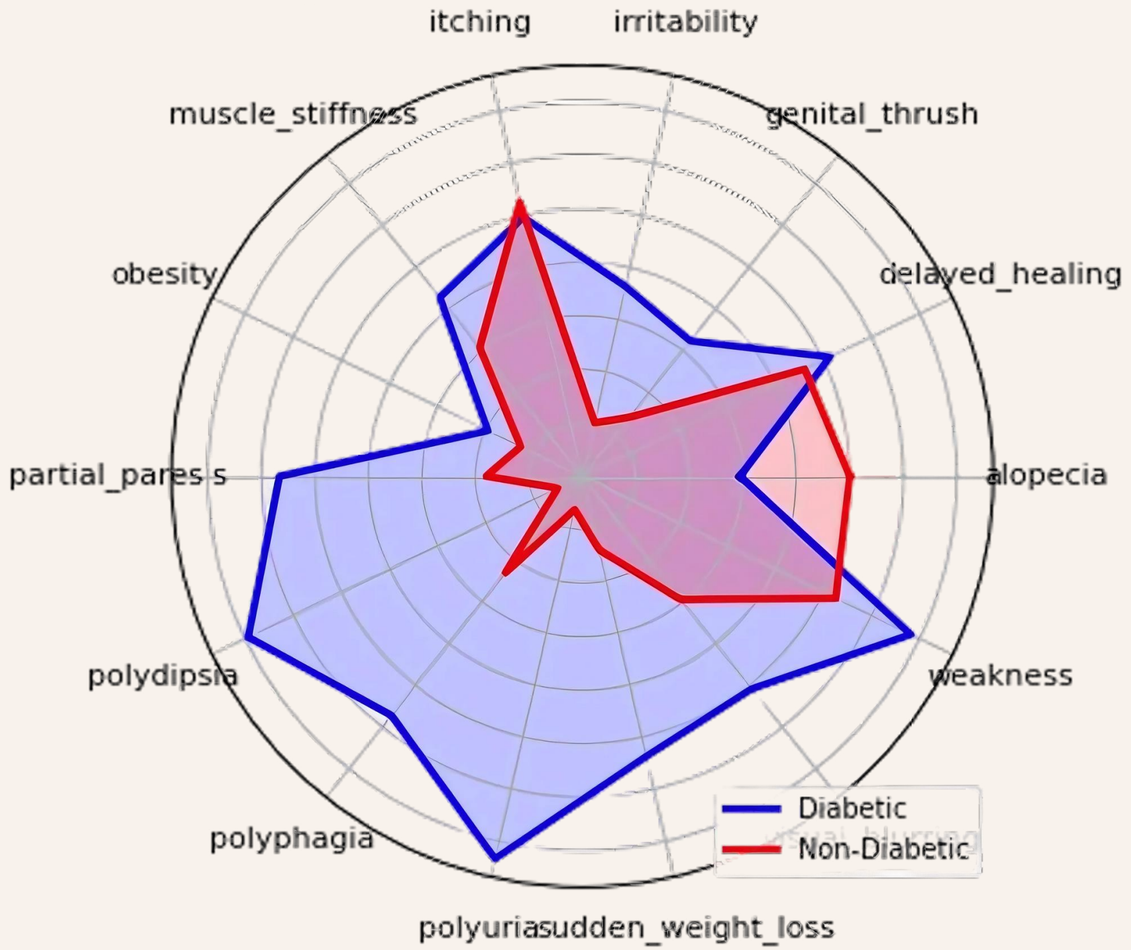
- Univariate Analysis: Distribution of Age, Gender, and other features using histograms.
- Bivariate Analysis: Correlation between features (e.g., Age vs. Class) using scatter plots or heatmaps.
- Outliers: Visuals showing box plots for features like Age.
- Class Distribution: Bar chart showing the ratio of diabetic vs. non-diabetic samples.
- Visuals:
- Heatmap for correlation analysis.
- Boxplots for visualizing outliers in features like Age.
- Histograms to show the distribution of key features like Age.



Feature Selection

Key Observations from the Correlation Matrix:

- Strong Correlation with Class (Target):
  - Polyuria (0.62) and Polydipsia (0.59) are strong indicators of diabetes.
- Moderate Correlation Among Symptoms:
  - Polyuria and Polydipsia (0.52) often occur together.
  - Weakness and Sudden Weight Loss (0.36) show a moderate relationship.
- Low or No Correlation:
  - Alopecia, Genital Thrush, and Muscle Stiffness have low predictive power for diabetes.
- Age:
  - Age has very low correlation with most symptoms and may not strongly predict diabetes in this dataset.



Model Development

Model Development Summary

- Data Preparation:
  - Feature selection focused on symptoms.
  - Data split: 70% training, 30% testing.
- Models & Key Hyperparameters:
  - Random Forest: n_estimators=100, max_depth=None.
  - Gradient Boosting: learning_rate=0.1, n_estimators=100.
  - SVM: kernel='rbf', C=1.0.
  - MLP: hidden_layer_sizes=(100,), activation='relu'.
  - Naive Bayes: Assumes predictor independence.
  - k-NN: n_neighbors=5.
  - Decision Tree: criterion='gini'.
  - Logistic Regression: penalty='l2', C=1.0.
- Key Insights:
  - Hyperparameters optimized model accuracy and generalization.

Libraries Used:

pandas	Data manipulation and preprocessing
numpy	Numerical computations
scikit-learn	Machine learning models and evaluation
matplotlib	Data visualization
seaborn	Advanced visualization
xgboost	Gradient boosting implementation
sklearn.decomposition (PCA)	Dimensionality reduction
sklearn.preprocessing	Feature scaling and preprocessing
sklearn.svm	Support Vector Classification
sklearn.model_selection	Hyperparameter tuning
sklearn.ensemble	Ensemble models
sklearn.tree	Decision tree modeling
sklearn.neighbors	k-Nearest Neighbors modeling
sklearn.neural_network	Neural networks
sklearn.naive_bayes	Naive Bayes models

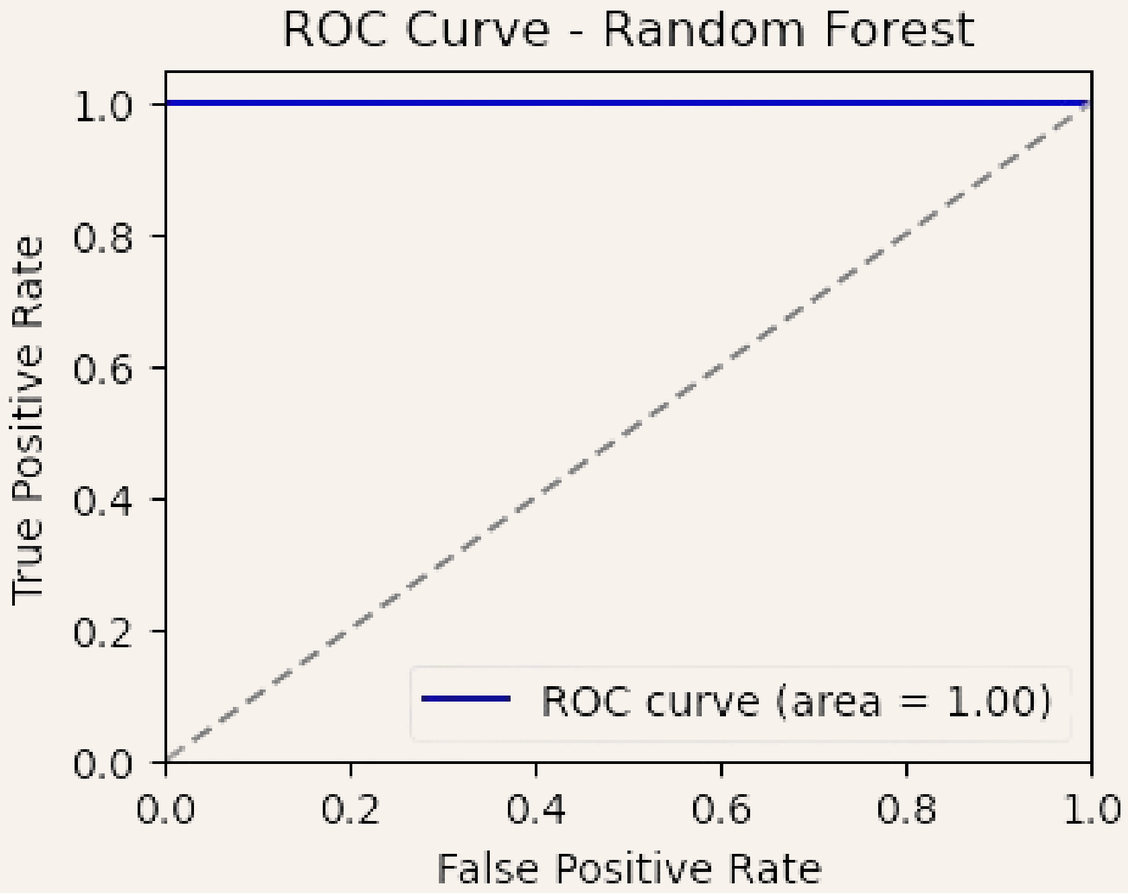
Model Evaluation

Model	Accuracy	ROC-AUC	Log Loss	MCC	Specificity	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-Score (Class 0)	F1-Score (Class 1)
Logistic Regression	0.93	0.99	0.17	0.87	0.97	0.87	0.98	0.97	0.91	0.92	0.94
SVM	0.97	1.00	0.04	0.94	1.00	0.93	1.00	1.00	0.95	0.96	0.98
Naive Bayes	0.94	0.99	0.20	0.88	0.95	0.90	0.97	0.95	0.94	0.93	0.95
k-NN	0.94	0.98	0.76	0.88	0.97	0.89	0.98	0.97	0.92	0.93	0.95
Decision Tree	0.94	0.99	0.14	0.88	0.97	0.89	0.98	0.97	0.92	0.93	0.95
Random Forest	0.99	1.00	0.08	0.98	1.00	0.98	1.00	1.00	0.98	0.99	0.99
MLP	0.97	1.00	0.05	0.94	1.00	0.93	1.00	1.00	0.95	0.96	0.98
Gradient Boosting	0.99	0.99	0.07	0.98	1.00	0.98	1.00	1.00	0.98	0.99	0.99

Results and Analysis

Random Forest is the best model due to its high performance across key metrics. It achieves excellent accuracy, a perfect ROC-AUC, and perfect precision, recall, and specificity. The model effectively distinguishes between classes and handles both positive and negative cases well, making it the most reliable choice for this task.

- Accuracy (0.99): Best performance.
- ROC-AUC (1.00): Perfect class separation.
- Log Loss (0.08): Confident predictions.
- MCC (0.98): Strong correlation.
- Specificity (1.00): Accurately identifies negatives.
- Precision and Recall (1.00): No false positives or negatives.
- F1-Score (0.99): Balanced precision and recall.



Thank you very much!

Aastha Singh
2004aasthasingh@gmail.com

Mr. Ravi
Mentor – Infosys Internship Program
springboardmentor431n@gmail.com