Infosys Springboard Internship 5.0 Batch 2

GLUCOSENCE- AI POWERED DIABETES DETECTION FOR EARLY INTERVENTION

PRESENTED BY-

ADITYA KUMAR

MENTOR -RAVI

PROBLEM STATEMENT:

- Diabetes is becoming increasingly prevalent worldwide, largely due to changes in lifestyle habits.
- Early detection plays a vital role in preventing severe complications associated with the disease.
- Many individuals hesitate or delay testing, which often results in late diagnoses and worsened health outcomes.
- Developing an AI-driven model can help identify diabetes in its early stages.
- Early detection through AI tools can encourage timely medical intervention and improve overall disease management.

OBJECTIVES:

- **Goal:** Develop a model that predicts whether an individual is healthy, prediabetic, or diabetic.
- **Importance:** Early detection enables individuals to take preventive measures, reducing the likelihood of serious complications.
- **Focus:** Investigate the impact of lifestyle habits, such as diet, physical activity, and sleep patterns, on the progression of diabetes.
- **Result:** Provide valuable insights to doctors and healthcare professionals, aiding in diabetes prevention and patient care.

DATA OVERVIEW:

For the GlucoSense: *AI-Powered Diabetes Detection for Early Intervention project, the dataset was obtained from Kaggle, specifically the "diabetes_risk_prediction_dataset.csv. " This dataset includes key healthcare and lifestyle details crucial for predicting diabetes risk.*

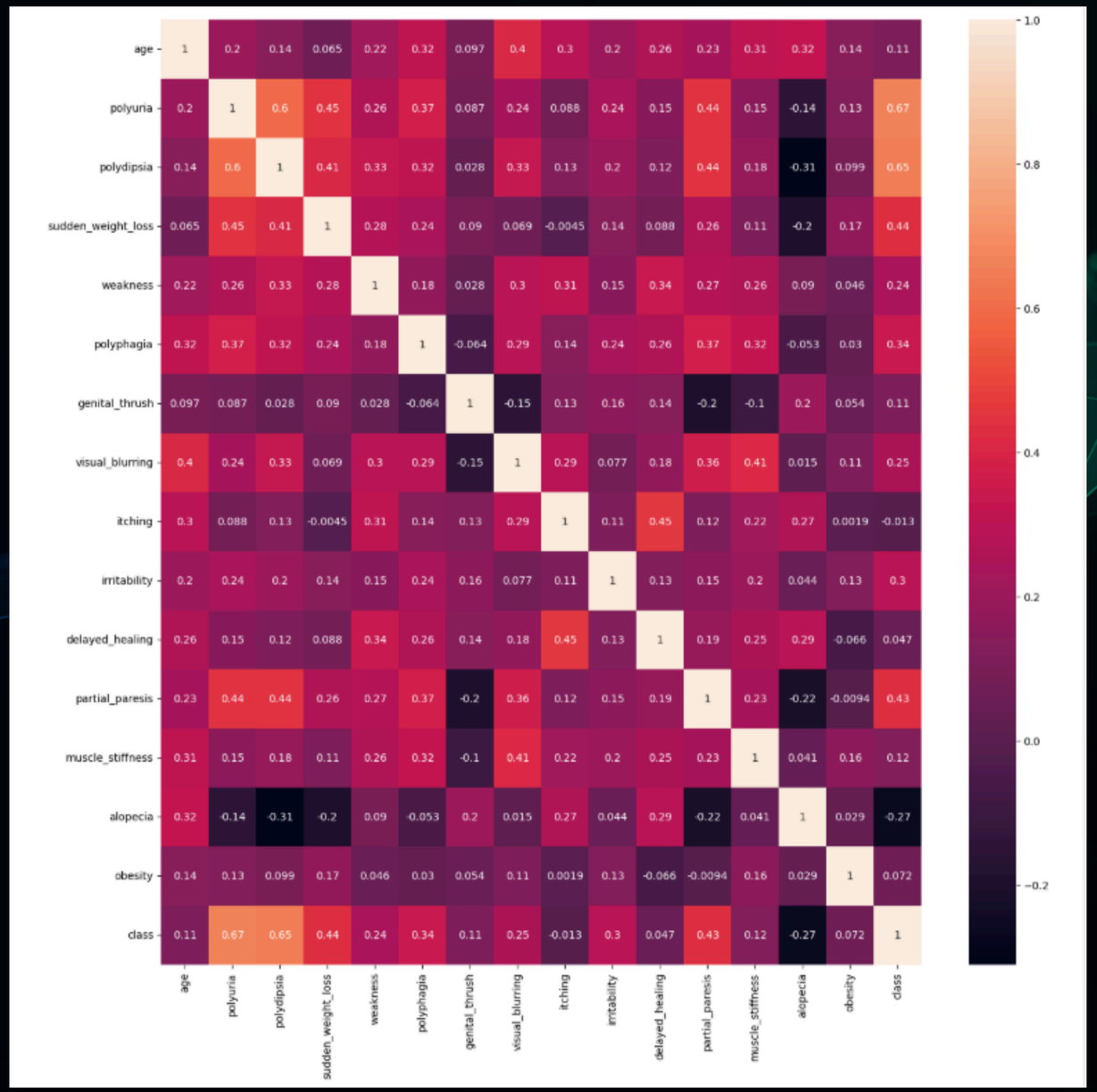
Features in the Dataset:

- *Polyuria, Polydipsia, Sudden Weight Loss, Weakness*
- *Polyphagia, Genital Thrush, Visual Blurring, Itching Irritability,*
- *Delayed Healing, Partial Paresis Muscle Stiffness, Alopecia, Obesity, and Class (target label)*

DATA EXPLORATION (EDA) AND DATA PREPROCESSING:

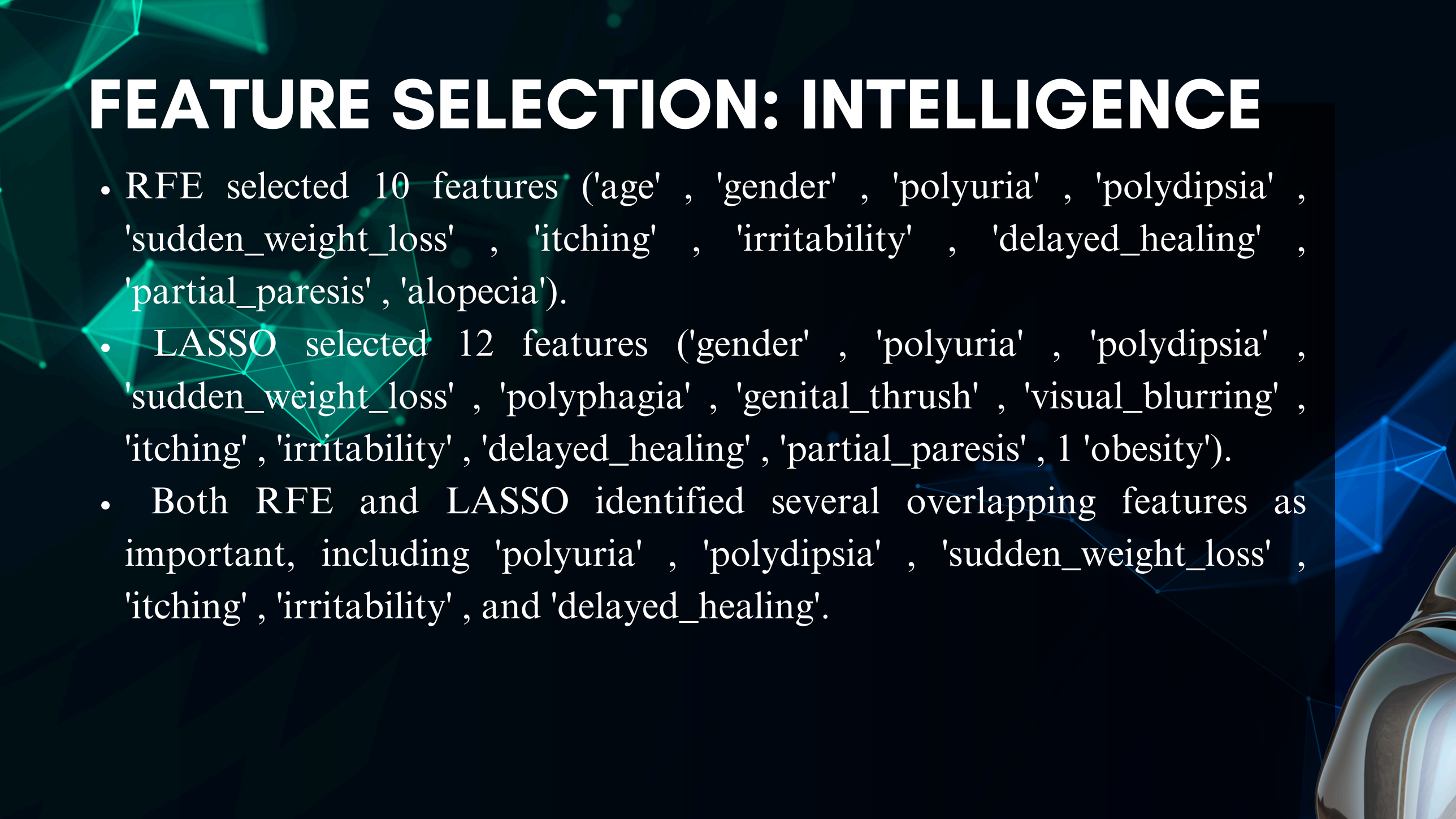
- Checked dataset structure (520 rows, 17 columns) and ensured no missing values.
- Removed 269 duplicate rows for unbiased analysis.
- Identified key features like Polyuria and Polydipsia as strong predictors through statistical analysis and visualizations.
- Used a correlation matrix to explore relationships between variables.
- Cleaned data by removing duplicates and outliers (IQR method).

CORELATION MATRIX



FEATURE SELECTION: INTELLIGENCE

- RFE selected 10 features ('age' , 'gender' , 'polyuria' , 'polydipsia' , 'sudden_weight_loss' , 'itching' , 'irritability' , 'delayed_healing' , 'partial_paresis' , 'alopecia').
- LASSO selected 12 features ('gender' , 'polyuria' , 'polydipsia' , 'sudden_weight_loss' , 'polyphagia' , 'genital_thrush' , 'visual_blurring' , 'itching' , 'irritability' , 'delayed_healing' , 'partial_paresis' , 1 'obesity').
- Both RFE and LASSO identified several overlapping features as important, including 'polyuria' , 'polydipsia' , 'sudden_weight_loss' , 'itching' , 'irritability' , and 'delayed_healing'.



DIMENSIONALITY REDUCTION:

Principal Component Analysis (PCA):

- PCA was employed to transform the dataset into a lower-dimensional space.
- Analysis revealed that 14 out of 16 features captured 95% of the data variability, indicating that dimensionality reduction was not necessary.

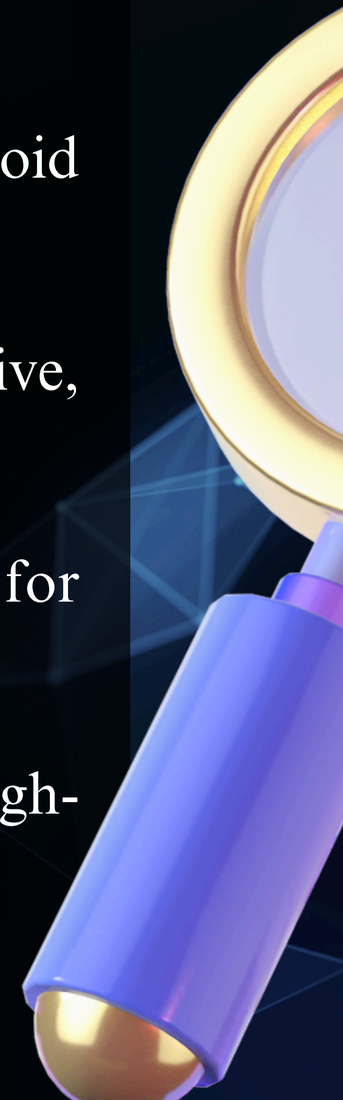
Impact:

- Simplifies data, reduces computational load, and lessens the risk of overfitting in high-dimensional datasets.
- In this instance, the dataset was already concise, so dimensionality reduction was not required for effective modeling.

CLASSIFICATION MODEL:

Classification models were employed to forecast diabetes status (Diabetic, Pre-Diabetic, or Healthy) based on key attributes.

- **Logistic Regression:** Estimates probabilities for binary outcomes using a sigmoid function. Simple and interpretable for predicting categorical results.
- **Decision Tree:** Splits data into subsets via feature values, forming an intuitive, visual decision-making flowchart.
- **Random Forest:** Combines multiple decision trees trained on random subsets for better accuracy and less overfitting.
- **SVM:** Finds an optimal hyperplane to separate classes, excelling in high-dimensional, non-linear data.



HYPERPARAMETER TUNING:

Hyperparameter tuning is the critical process of refining a model's parameters to attain optimal performance. For this project, key techniques were implemented to fine-tune the classification models:

Methods Used:

- Grid Search: Employed a methodical approach to evaluate diverse combinations of hyperparameters, systematically examining the performance of each configuration to identify the optimal values.
- Random Search: Explored a randomly selected subset of the hyperparameter space, providing a more expedient approach for tuning models with a substantial number of parameters.

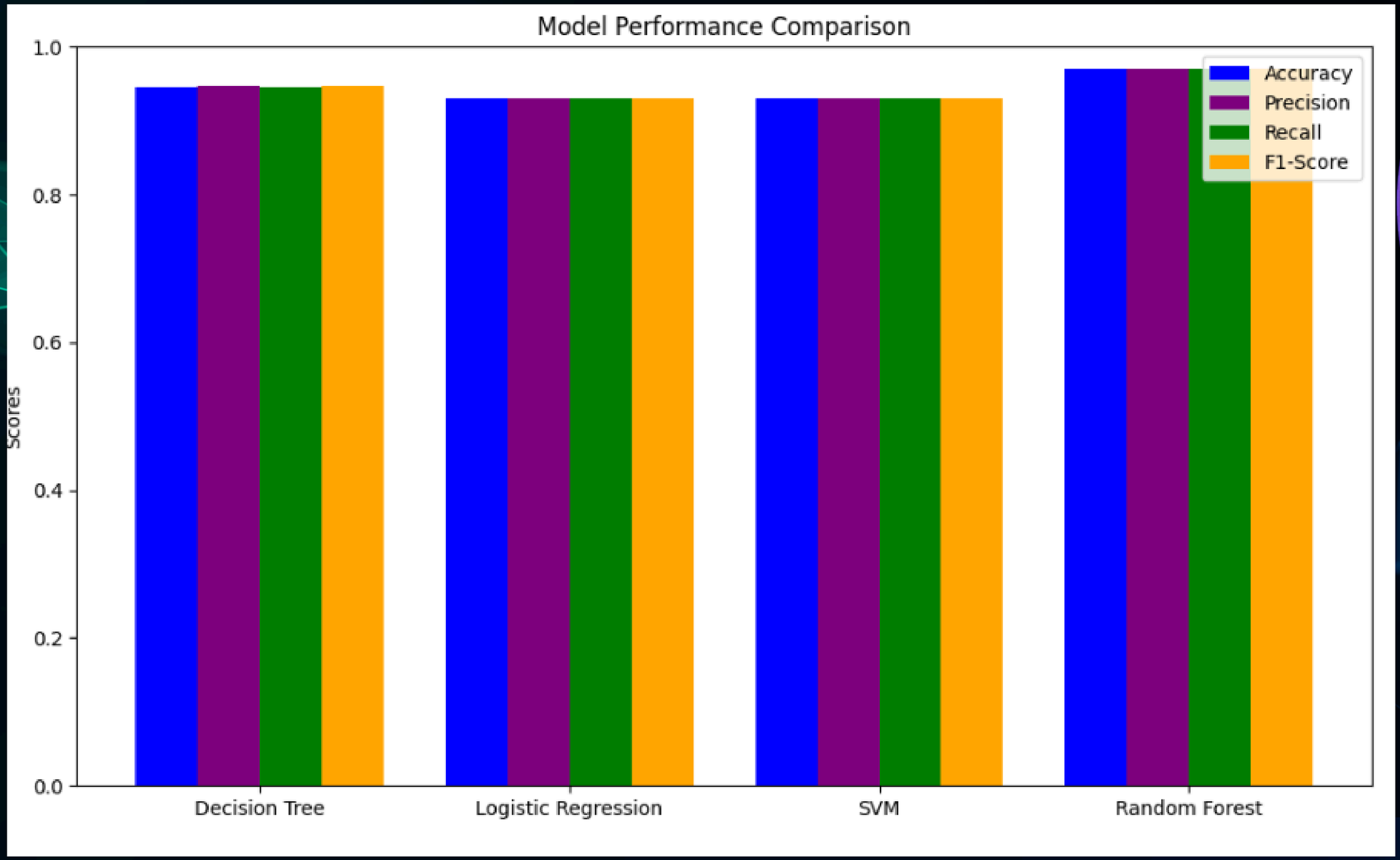
- **EVALUATION METRICS BEFORE HYPERPARAMETER TUNING:**

	Model	Accuracy	Precision	Recall	F1 Score	AUC
0	Logistic Regression	0.823529	0.825000	0.942857	0.880000	0.944643
1	Decision Tree	0.882353	0.891892	0.942857	0.916667	0.846429
2	Support Vector Machine	0.901961	0.894737	0.971429	0.931507	0.971429
3	Random Forest	0.921569	0.918919	0.971429	0.944444	0.975893

- **EVALUATION METRICS AFTER HYPERPARAMETER TUNING:**

	Model	Accuracy	Precision	Recall	F1-Score
0	Decision Tree	0.946154	0.947939	0.946154	0.946515
1	Logistic Regression	0.930769	0.930538	0.930769	0.930594
2	Random Forest	0.969231	0.970005	0.969231	0.969373
3	SVM	0.930769	0.930538	0.930769	0.930594

MODEL COMPARISON:



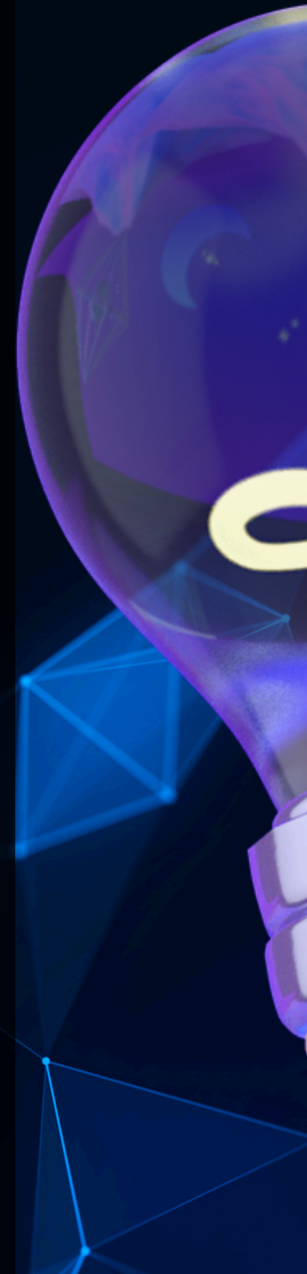
MODEL SELECTION:

Random Forest emerged as the top model, achieving the highest accuracy (96.92%) and F1-Score (96.94%). It demonstrated superior performance with high precision (97%), indicating fewer false positives

Decision Tree performed well (accuracy: 94.62%, F1-Score: 94.65%), but Random Forest's ensemble approach provides an edge.

Logistic Regression and SVM showed similar results (accuracy: 93.08%, F1-Score: 93.06%). While effective, they may not capture complex patterns as well as Random Forest.

Random Forest is preferred due to its superior performance. However, Decision Tree offers a simpler and faster alternative.



CONCLUSION

This project successfully led to the creation of a predictive model for assessing diabetes risk, leveraging the capabilities of Artificial Intelligence and machine learning techniques. The model demonstrated exceptional accuracy in predicting an individual's likelihood of developing diabetes. It also provided actionable insights into key lifestyle and healthcare factors that play a critical role in influencing diabetes risk.

FUTURE RECOMMENDATIONS AND EXTENSIONS:

Future recommendations include:

1. Expanding the dataset to incorporate a wider range of lifestyle and environmental factors for a more comprehensive analysis.
2. Investigating advanced machine learning models to improve predictive accuracy and model performance.
3. Integrating real-time health data to enable continuous and dynamic risk assessment.
4. Conducting further research to evaluate the model's generalizability across diverse populations and healthcare environments, ensuring its effectiveness and applicability in varied contexts.

THANK YOU!