GlucoSence- AI Powered Diabetes Detection for Early Intervention

Presented by:-

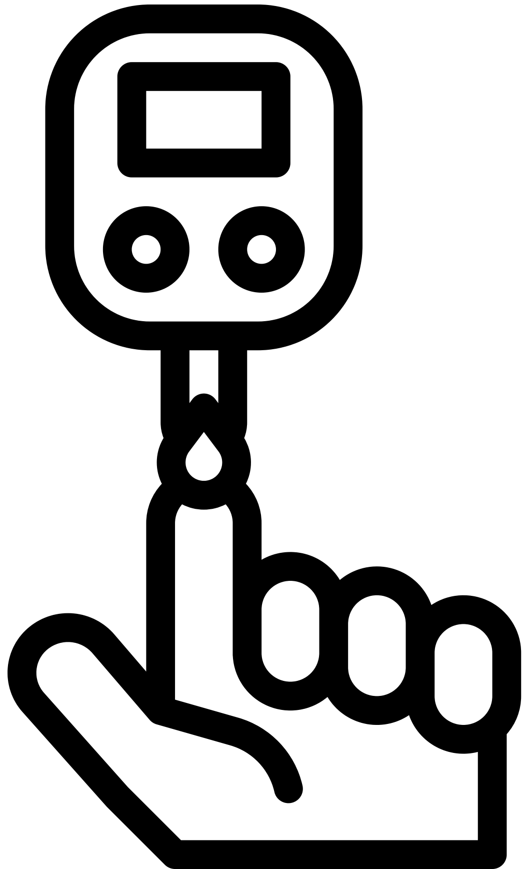
Aditya Kishor

Infosys Springboard Internship 5.0 Batch 2

Github Link

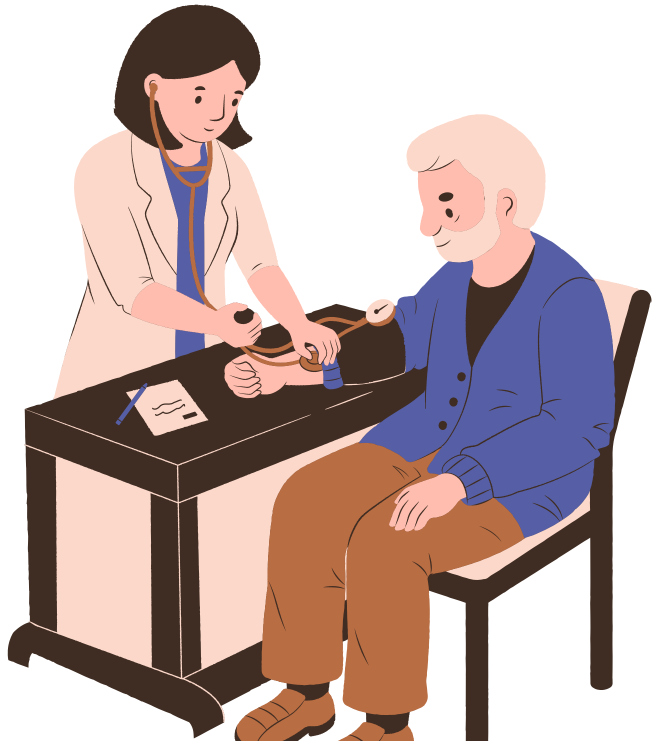
Problem Statement:

- Diabetes has become a global health concern, with cases rising due to lifestyle changes.
- Early diagnosis is crucial to reduce risks of severe complications.
- Many people delay testing, leading to late detection and poor outcomes.
- The goal is to create an AI-based model to detect diabetes early and promote timely intervention.



Objectives:

- **Goal:** Create a model to predict if a person is healthy, pre-diabetic, or diabetic.
- **Why It's Important:** Early detection helps in taking steps to avoid serious health problems.
- **Focus:** Understand how lifestyle habits affect diabetes.
- **Result:** Support doctors and healthcare teams in preventing diabetes.



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."

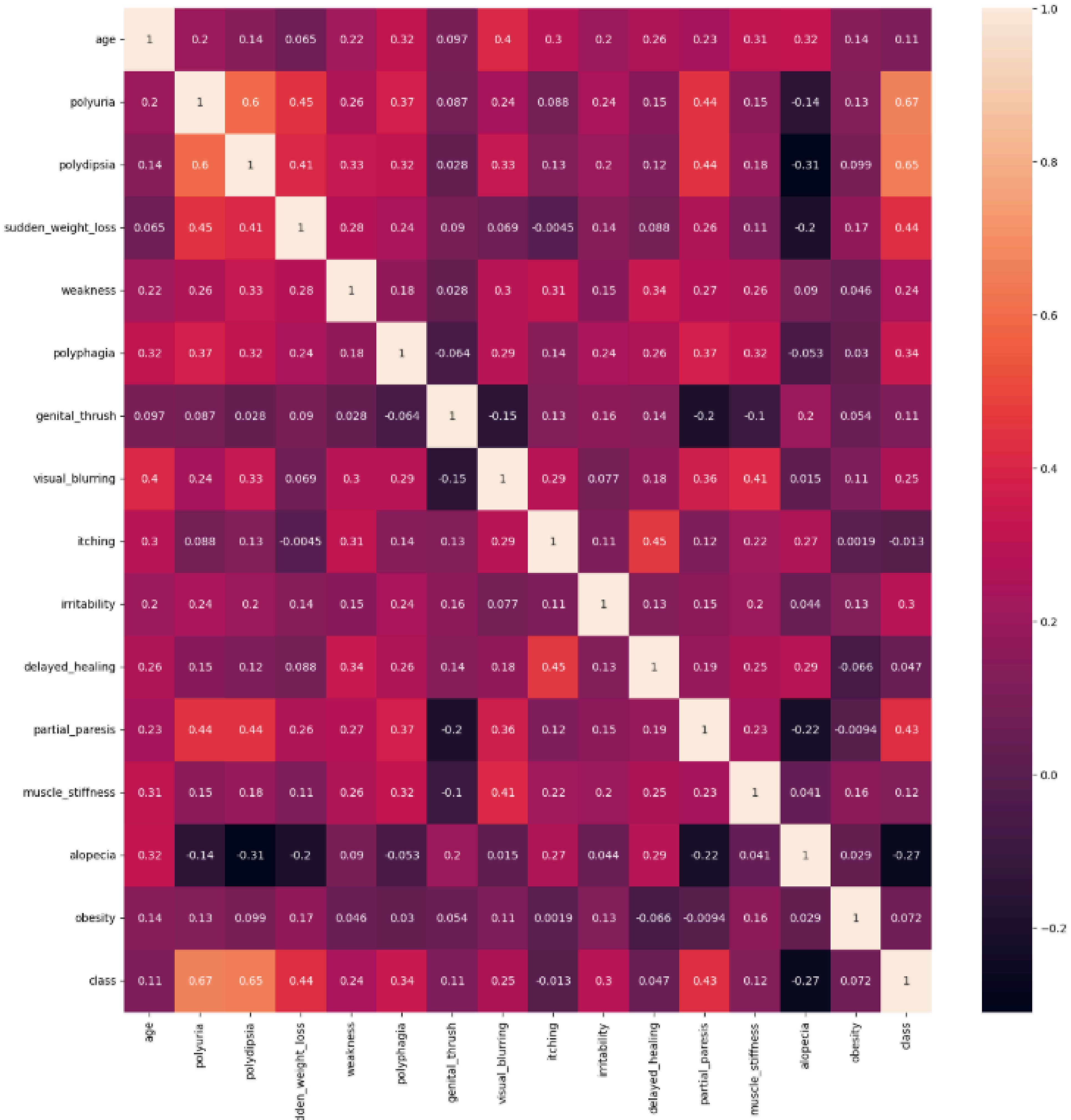
Features in the Dataset:

- 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).
- Individuals with diabetes are generally older, with a higher median age compared to the non-diabetes group.

Corelation Matrix



	age	polyuria	polydipsia	dden_weight_loss	weakness	polyphagia	genital_thrush	visual_blurring	itching	irritability	delayed_healing	partial_paresis	muscle_stiffness	alopecia	obesity	class
age	1	0.2	0.14	0.065	0.22	0.32	0.097	0.4	0.3	0.2	0.26	0.23	0.31	0.32	0.14	0.11
polyuria	0.2	1	0.6	0.45	0.26	0.37	0.087	0.24	0.088	0.24	0.15	0.44	0.15	-0.14	0.13	0.67
polydipsia	0.14	0.6	1	0.41	0.33	0.32	0.028	0.33	0.13	0.2	0.12	0.44	0.18	-0.31	0.099	0.65
sudden_weight_loss	0.065	0.45	0.41	1	0.28	0.24	0.09	0.069	-0.0045	0.14	0.088	0.26	0.11	-0.2	0.17	0.44
weakness	0.22	0.26	0.33	0.28	1	0.18	0.028	0.3	0.31	0.15	0.34	0.27	0.26	0.09	0.046	0.24
polyphagia	0.32	0.37	0.32	0.24	0.18	1	-0.064	0.29	0.14	0.24	0.26	0.37	0.32	-0.053	0.03	0.34
genital_thrush	0.097	0.087	0.028	0.09	0.028	-0.064	1	-0.15	0.13	0.16	0.14	-0.2	-0.1	0.2	0.054	0.11
visual_blurring	0.4	0.24	0.33	0.069	0.3	0.29	-0.15	1	0.29	0.077	0.18	0.36	0.41	0.015	0.11	0.25
itching	0.3	0.088	0.13	-0.0045	0.31	0.14	0.13	0.29	1	0.11	0.45	0.12	0.22	0.27	0.0019	-0.013
irritability	0.2	0.24	0.2	0.14	0.15	0.24	0.16	0.077	0.11	1	0.13	0.15	0.2	0.044	0.13	0.3
delayed_healing	0.26	0.15	0.12	0.088	0.34	0.26	0.14	0.18	0.45	0.13	1	0.19	0.25	0.29	-0.066	0.047
partial_paresis	0.23	0.44	0.44	0.26	0.27	0.37	-0.2	0.36	0.12	0.15	0.19	1	0.23	-0.22	-0.0094	0.43
muscle_stiffness	0.31	0.15	0.18	0.11	0.26	0.32	-0.1	0.41	0.22	0.2	0.25	0.23	1	0.041	0.16	0.12
alopecia	0.32	-0.14	-0.31	-0.2	0.09	-0.053	0.2	0.015	0.27	0.044	0.29	-0.22	0.041	1	0.029	-0.27
obesity	0.14	0.13	0.099	0.17	0.046	0.03	0.054	0.11	0.0019	0.13	-0.066	-0.0094	0.16	0.029	1	0.072
class	0.11	0.67	0.65	0.44	0.24	0.34	0.11	0.25	-0.013	0.3	0.047	0.43	0.12	-0.27	0.072	1

Feature Selection:

- 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.
- Step ensured optimized data without extra transformations.

Classification model:

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

- **Logistic Regression:** A model estimating diabetes likelihood using a sigmoid function.
- **Decision Tree:** A tree-based model for decision-making, partitioning data based on feature values.
- **Random Forest:** An ensemble method using multiple decision trees for improved accuracy and overfitting resistance.
- **Support Vector Machines (SVM):** An algorithm finding optimal hyperplanes to separate data points in high-dimensional spaces.

Performance metrics

Accuracy: This metric quantifies the overall correctness of the model's predictions by measuring the proportion of instances where the model correctly predicted the true class label.

Recall: Recall measures the model's ability to identify all actual positive instances. It quantifies the proportion of actual positive instances that the model correctly predicted.

AUC-ROC (Area Under the Receiver Operating Characteristic curve): This metric measures the model's ability to distinguish between positive and negative classes across all possible classification thresholds.

Precision: Precision focuses on the correctness of the model's positive predictions. It measures the proportion of instances predicted as positive that are actually positive.

F1-score: The F1-score is the harmonic mean of precision and recall, 1 providing a balanced measure of the model's 2 performance that considers both aspects. It is particularly useful when there is an imbalance between positive and negative classes.

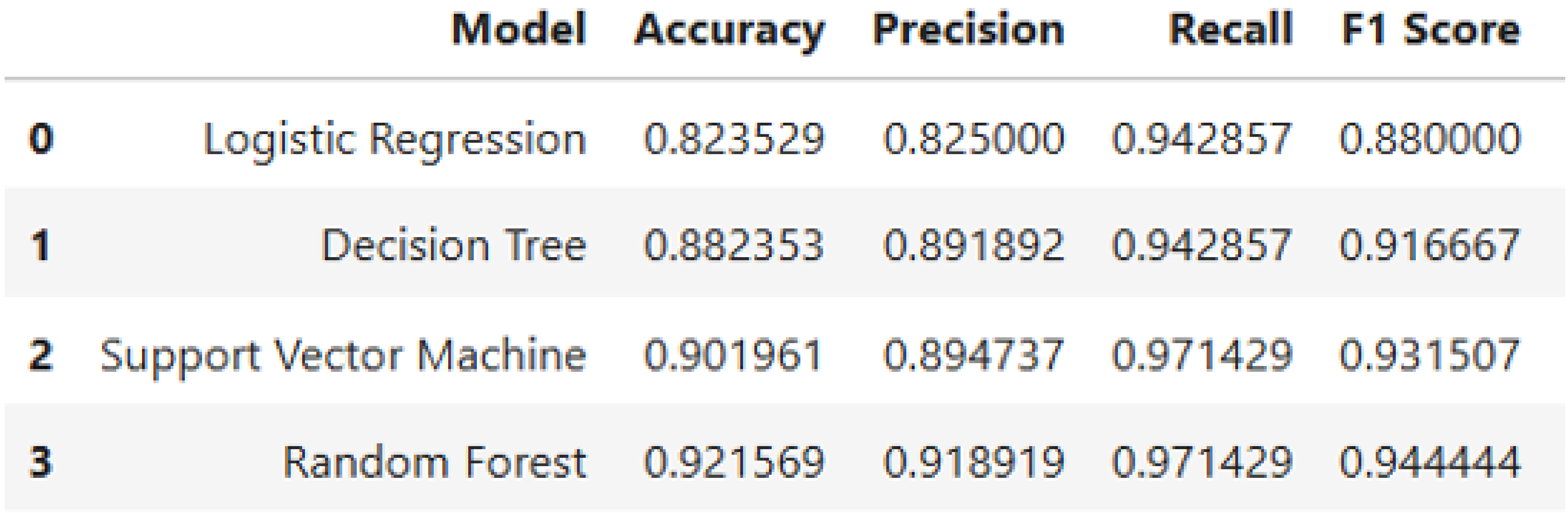
Hyperparameter tuning:

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:

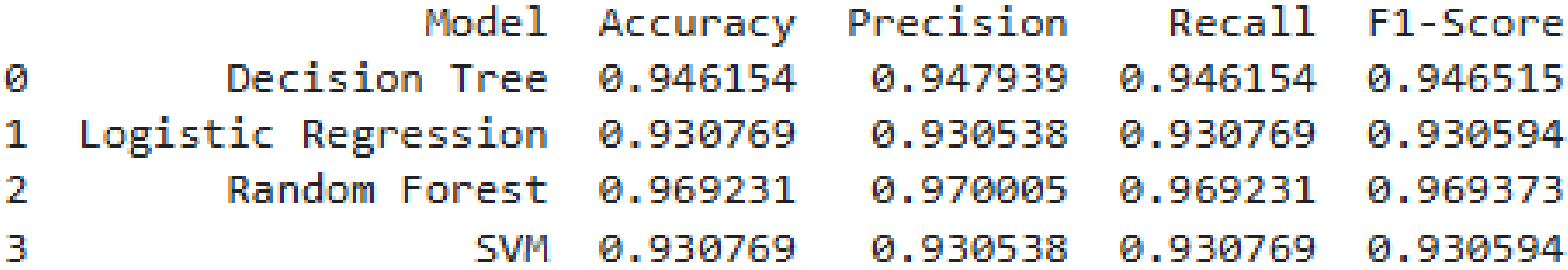
- **Grid Search:** Employed a methodical approach to evaluate diverse combinations of hyperparameters, systematically examining the performance of each configuration to identify the optimal values.
- **Random Search:** Explored a randomly selected subset of the hyperparameter space, providing a more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

Evaluation metrics before hyperparameter tuning:

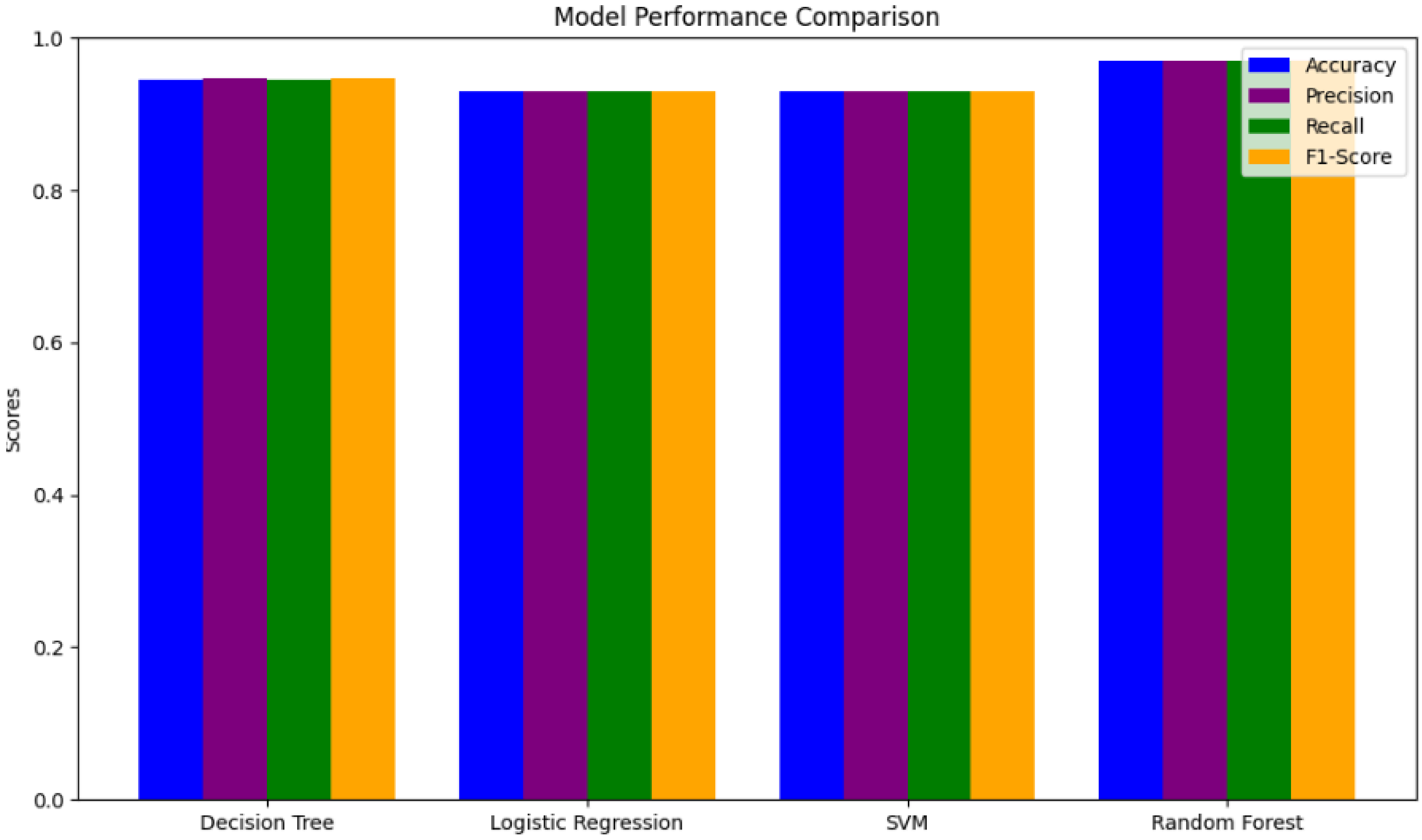


	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.823529	0.825000	0.942857	0.880000
1	Decision Tree	0.882353	0.891892	0.942857	0.916667
2	Support Vector Machine	0.901961	0.894737	0.971429	0.931507
3	Random Forest	0.921569	0.918919	0.971429	0.944444

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 Performance Comparison
Scores
1.0
0.8
0.6
0.4
0.2
0.0
Decision Tree
Logistic Regression
SVM
Random Forest
Accuracy
Precision
Recall
F1-Score

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 culminated in the development of a predictive model for diabetes risk assessment, harnessing the power of Artificial Intelligence and machine learning techniques. The model exhibited a high degree of accuracy in forecasting the likelihood of an individual developing diabetes, providing valuable and actionable insights into the lifestyle and healthcare factors that significantly influence diabetes risk.

Future recommendations and extensions:

Future recommendations encompass expanding the dataset to encompass a broader spectrum of lifestyle and environmental factors, exploring cutting-edge machine learning models with the potential to enhance predictive accuracy, and integrating real-time health data for continuous and dynamic risk assessment. Further research could also concentrate on evaluating the model's generalizability across diverse populations and healthcare settings to ensure its applicability and effectiveness in various contexts.

Thank You

Aditya Kishor
adikishor67@gmail.com

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