DEPARTMENT OF COMPUTER & INFORMATION SYSTEMS ENGINEERING

BACHELORS IN COMPUTER SYSTEMS ENGINEERING

Course Code: CS-324

Course Title: Machine Learning

Open Ended Lab

TE Batch 2021, Spring Semester 2024

Grading Rubric

TERM PROJECT

MUHAMMAD BILAL SHAIKH	CS-21024
IBAD UR REHMAN	CS-21021
HAMNA KHAN	CS-21048
ABDULLAH MAHMOOD	CS-21155

Project Title: DIABETES DECODED

Project Report: Machine Learning Open-Ended Lab

1. Problem Definition

The project aims to predict diabetes using machine learning techniques applied to the "*Pima Indians Diabetes Database*". This report covers data exploration, model training, and a sophisticated graphical user interface (GUI) for intuitive predictions.

2. Data Collection and Exploratory Data Analysis (EDA)

2.1. Dataset: Obtained from Kaggle, the dataset includes crucial health metrics and outcomes.

2.2. Initial Analysis: Conducted thorough checks including handling missing values and exploring feature relationships through visualizations like pair plots and correlation matrices.

3. Feature Engineering and Model Training

3.1. Feature Selection and Split: Utilized all features except 'Outcome', splitting data into training and testing sets (80/20 split).

3.2. Model Preparation: Implemented `StandardScaler` for numerical feature standardization to optimize model performance.

4. Model Building and Evaluation

3.3. Models Implemented: Trained robust models:

3.3.1. Logistic Regression: Leveraged for its interpretability and performance metrics.

Rationale: Chosen for its simplicity, interpretability, and robust performance in binary classification tasks. Logistic regression provides insights into the significance of each feature's impact on diabetes prediction, making it suitable for initial model exploration and understanding.

3.3.2. Decision Tree: Utilized for its ability to handle non-linear relationships in data.

Rationale: Selected for its ability to capture non-linear relationships and interactions among features. Decision trees are advantageous in revealing complex decision-making processes inherent in medical diagnostics, potentially offering insights into intricate health interactions beyond linear assumptions.

3.4. Model Evaluation:

Assessed models comprehensively using accuracy, ROC AUC score, and detailed classification reports to validate efficacy on unseen data. This rigorous evaluation ensures the reliability and generalizability of predictions in clinical settings.

4. Graphical User Interface (GUI) Implementation

4.1. GUI Design Excellence:

4.1.1. Intuitive Navigation: Designed for seamless interaction with radio buttons for model selection, allowing users to switch between logistic regression and decision tree models effortlessly.

4.1.2. Interactive Input: Enhanced with input fields for user-entered health metrics, enabling personalized predictions and fostering user engagement.

4.1.3. Real-Time Feedback: Empowered with real-time prediction feedback, displaying results such as predicted class and probability, promoting transparency and trust in the predictive process.

4.1.4. Educational Value: Positioned as an educational tool, bridging theoretical machine learning concepts with practical application, catering to diverse learning styles and fostering deeper understanding of health analytics.

5. Conclusion

5.1. Findings:

5.1.1. Best-Performing Model: The best performing model is being displayed on the screen based on the current data set being inputted to it in the real time.

5.1.2. Key Insights: Analysis revealed that glucose levels and BMI significantly influence diabetes prediction, highlighting their critical role in health diagnostics.

5.2. Limitations:

5.2.1. Data Limitations: The dataset's size may limit its representativeness, potentially overlooking diverse population variations.

5.2.2. Model Limitations: Some models exhibited signs of overfitting due to the dataset's constraints, suggesting a need for caution in generalizing results.

5.3. Future Work:

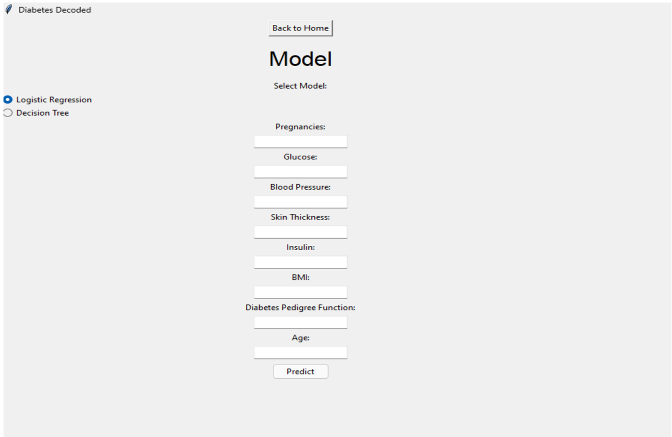
5.3.1. Improvements: Recommend advanced model techniques such as ensemble methods or neural networks for heightened predictive accuracy.

5.3.2. Enhancements: Propose hyperparameter tuning and additional feature engineering to refine model performance, ensuring robustness across diverse datasets.

5.3.3. Extended Analysis: Suggest further exploration into demographic factors and lifestyle variables to uncover deeper insights into diabetes predictors.

7. SAMPLE RUNS:

7.1 USER FRIENDLY DATA INPUT



Model

Select Model:

● Logistic Regression
○ Decision Tree

Pregnancies:

Glucose:

Blood Pressure:

Skin Thickness:

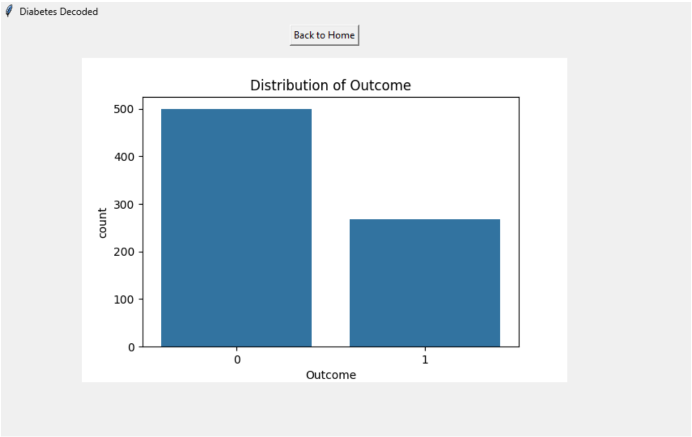
Insulin:

BMI:

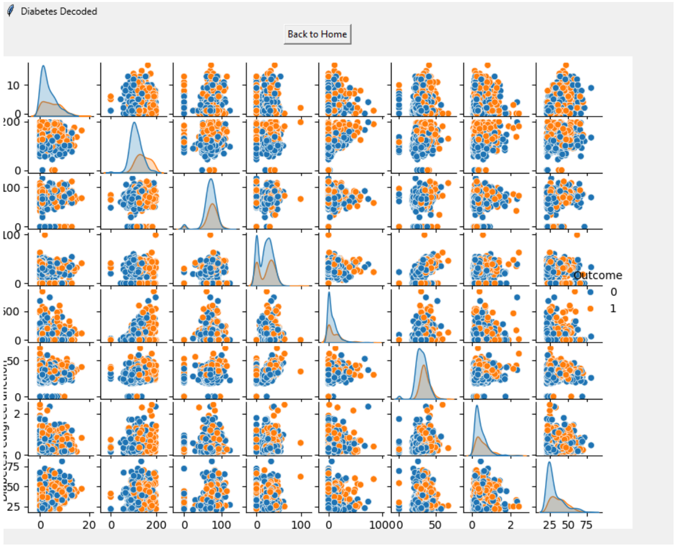
Diabetes Pedigree Function:

Age:

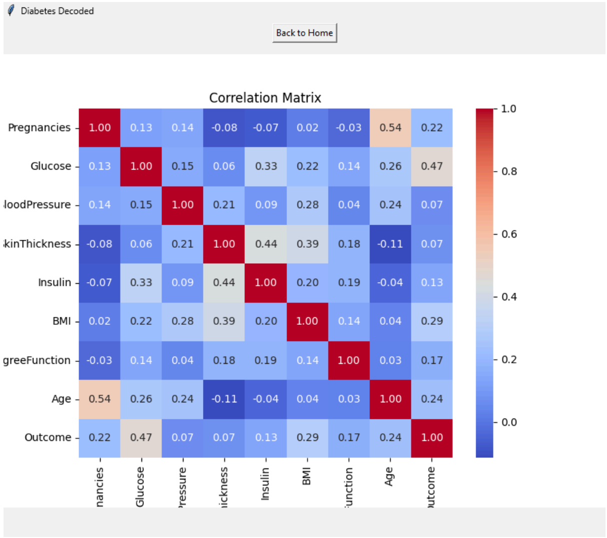
7.2 TARGET VARIABLE



7.3 DATA VISUALIZATION



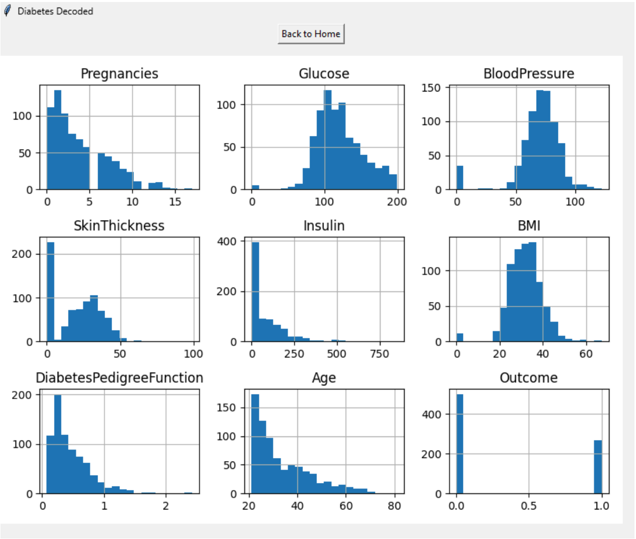
7.4 CORRELATION MATRIX



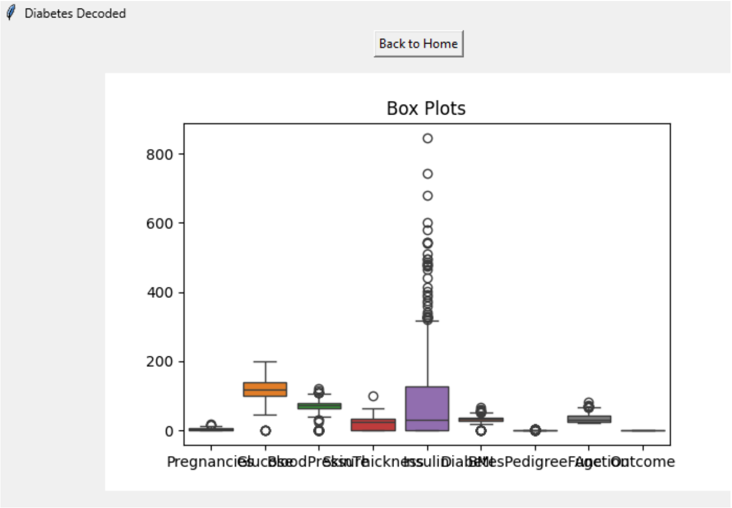
Correlation Matrix

	nancies	Glucose	ressure	ickness	Insulin	BMI	unction	Age	utcome
Pregnancies	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54	0.22
Glucose	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26	0.47
loodPressure	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24	0.07
kinThickness	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11	0.07
Insulin	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04	0.13
BMI	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04	0.29
greeFunction	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03	0.17
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00	0.24
Outcome	0.22	0.47	0.07	0.07	0.13	0.29	0.17	0.24	1.00

7.5 FEATURE HISTOGRAMS



7.6 BOX PLOTS


Diabetes Decoded
Back to Home
Box Plots
0
200
400
600
800
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome