**DATA MINING PROJECT- DIABETES PREDICTION**

Professor: Srinivasan Radhakrishnan

Course Name: Data Mining in Engineering

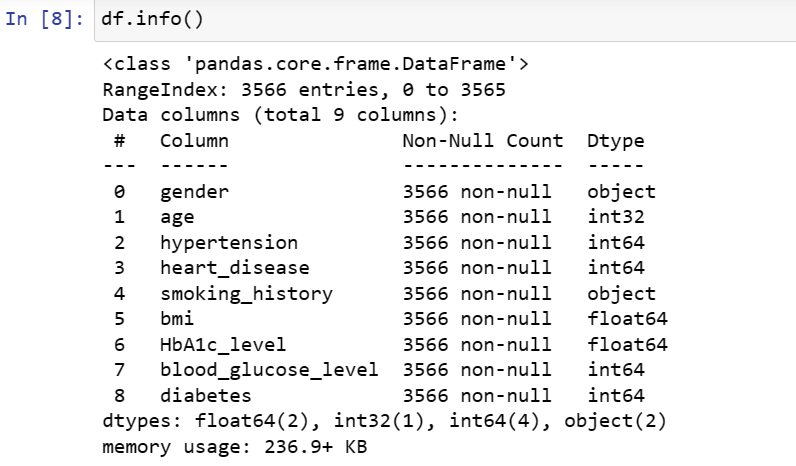
Due Date of Proposal: 03-12-2024

**Project Background and Objective**

This project aims to develop a predictive model for diabetes using machine learning algorithms on a dataset comprising health-related attributes like gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, blood glucose level, and diabetes status. The primary objectives include evaluating the performance of algorithms such as Logistic Regression, Random Forest, and Decision Tree in diabetes prediction, comparing their accuracy metrics, and identifying the most effective algorithm for this task. The insights gained from this analysis can contribute to better healthcare decision-making and management of diabetes-related risks.

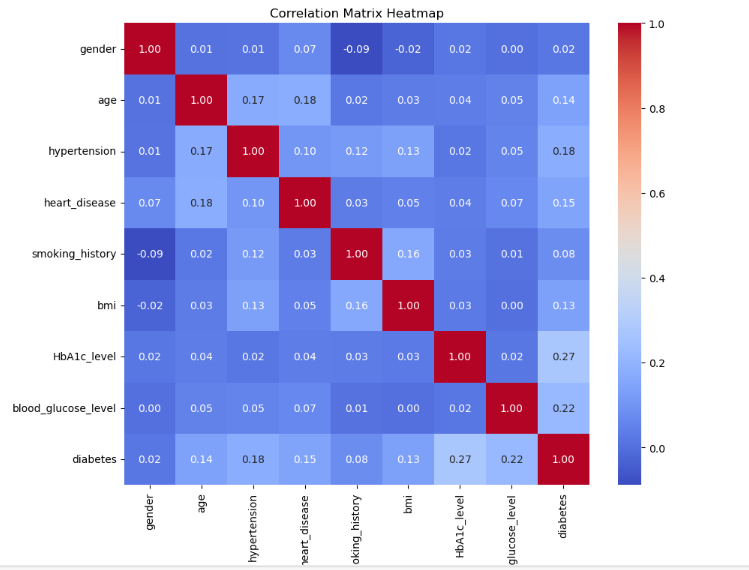
**Data Selection and Preprocessing**

The dataset used is taken from Kaggle website. The required dataset is cleaned by handling missing values and removing unnecessary columns.



**Exploratory Data Analysis (EDA) and Feature Selection**

In Phase 2 of the project, Exploratory Data Analysis (EDA) and Feature Selection is conducted. Used box plots and bar graphs to visualize numerical variables and encoded categorical variables to understand their distribution and relationships with the target variable. This step is crucial for gaining insights into the dataset's structure and identifying important features for further analysis. And the confusion matrix is displayed.



In this specific heatmap, it shows the correlation between a person's age and whether they have heart disease. A darker color indicates a stronger correlation. In this case, there is a weak positive correlation between age and heart disease, which means that as age increases, the likelihood of heart disease also increases. There is a weak positive correlation between HbA1c level and heart disease. HbA1c is a measure of blood sugar control over the past few months. There is a moderate positive correlation between diabetes and heart disease.

**Model Implementation and Baseline Evaluation**

In Phase 3, three supervised learning algorithms are implemented Logistic Regresson, Decision Trees, Random Forest. These algorithms were evaluated using the initial features without hyperparameter tuning to establish baseline performance metrics.

The results acquired are :

Logistic Regression Accuracy: 0.9612403100775194

Logistic Regression Classification Report:

precision recall f1-score support

0 0.97 0.99 0.98 614

1 0.69 0.35 0.47 31

accuracy 0.96 645

macro avg 0.83 0.67 0.72 645

weighted avg 0.95 0.96 0.96 645

Logistic Regression ROC-AUC Score: 0.6733476936009247

Random Forest Accuracy: 0.9705426356589147

Random Forest Classification Report:

precision recall f1-score support

0 0.97 1.00 0.98 614

1 1.00 0.39 0.56 31

accuracy 0.97 645

macro avg 0.98 0.69 0.77 645

weighted avg 0.97 0.97 0.96 645

Random Forest ROC-AUC Score: 0.6935483870967742

Decision Tree Accuracy: 0.9503875968992248

Decision Tree Classification Report:

precision recall f1-score support

0 0.98 0.97 0.97 614

1 0.48 0.52 0.50 31

accuracy 0.95 645

macro avg 0.73 0.74 0.74 645

weighted avg 0.95 0.95 0.95 645

Decision Tree ROC-AUC Score: 0.7442208679205632

**Phase 4: Hyperparameter Tuning**

During this phase, random search is used to discover the optimal hyperparameters for each model.

Results acquired are as follows:

Logistic Regression Classification Report:

precision recall f1-score support

0 0.97 1.00 0.98 614

1 0.83 0.32 0.47 31

accuracy 0.96 645

macro avg 0.90 0.66 0.72 645

weighted avg 0.96 0.96 0.96 645

Decision Tree Classification Report:

precision recall f1-score support

0 0.97 1.00 0.98 614

1 1.00 0.39 0.56 31

accuracy 0.97 645

macro avg 0.98 0.69 0.77 645

weighted avg 0.97 0.97 0.96 645

Random Forest Classification Report:

precision recall f1-score support

0 0.97 1.00 0.98 614

1 1.00 0.39 0.56 31

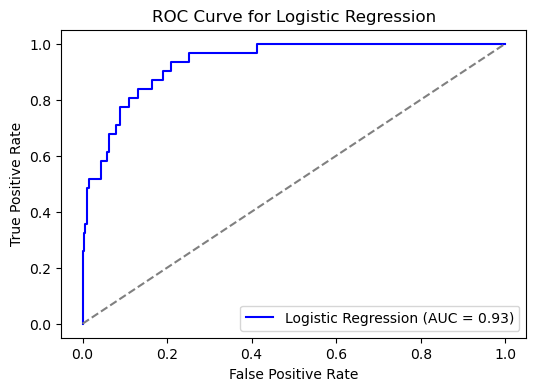
accuracy 0.97 645

macro avg 0.98 0.69 0.77 645

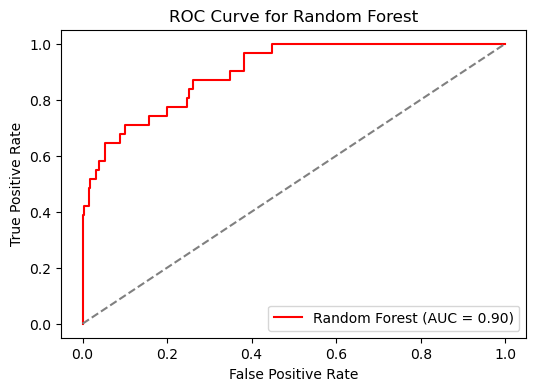
weighted avg 0.97 0.97 0.96 645

The ROC-AUC score is calculated and the curve is displayed:

Logistic Regression ROC-AUC Score: 0.9332247557003258



Random Forest ROC-AUC Score: 0.900178627718819



Decision Tree ROC-AUC Score: 0.900178627718819

**Conclusion and Recommendations**

In conclusion, the comparative analysis of three supervised learning algorithms - Logistic Regression, Random Forest, and Decision Tree - for predicting diabetes revealed valuable insights into their performance, computational efficiency, and applicability to the dataset.

Before hyperparameter tuning, Logistic Regression showed promising accuracy but lacked in precision and recall for positive diabetes cases. Random Forest exhibited high accuracy and stability, ideal for capturing complex patterns in healthcare data, while Decision Tree provided a simpler model with good interpretability.

After hyperparameter tuning, all models showed improvements, with Random Forest maintaining its high accuracy and becoming the preferred choice due to its robust performance and versatility.

Furthermore, fine-tuning and exploring other algorithms may further enhance model performance. Logistic Regression remains suitable for straightforward interpretability, while Random Forest proves effective for capturing intricate relationships in the data. Decision Tree, although simpler, can be valuable for understanding feature importance.

Ultimately, the choice of algorithm should align with specific analysis goals and the desired balance between accuracy, interpretability, and computational efficiency.