DEVELOPMENT OF MACHINE LEARNING ALGORITHMS FOR DIABETIC PATIENT PREDICTION

Objective:

Develop a binary classification model to predict whether a patient is diabetic or not based on relevant health features.

Steps:

Data Collection:

- Obtain a dataset containing health-related features such as glucose levels, blood pressure, BMI (Body Mass Index), and age, along with information on whether the patient is diabetic or not.

Data Exploration:

- Explore the dataset to understand the distribution of diabetic and non-diabetic cases and examine the relationships between health features.

Data Preprocessing:

- Handle missing values and outliers.

- Encode categorical variables if applicable.

- Split the dataset into training and testing sets.

Data Visualization:

- Create visualizations to compare the distribution of health features for diabetic and non-diabetic patients.

Model Selection:

- Choose a classification algorithm

- Implement the chosen model using a machine learning library.

Model Training:

- Train the model using the training set.

- Evaluate the model's performance on the testing set.

Model Evaluation:

- Use metrics like accuracy, precision, recall, and F1-score to evaluate the model's performance.

- Consider the confusion matrix for a more detailed analysis and ROC graph.

Hyperparameter Tuning (Optional):

- Experiment with different hyperparameters to optimize the model's performance.

Predictions:

- Make predictions on new health data or the testing set.

Results Interpretation:

- Interpret the results, understand the importance of different health features in predicting diabetes, and discuss potential applications in healthcare.

About Dataset

Diabetes is an opportune disease which has large wealth of data available and has with it huge complications. There is a need for a better and a more accurate approach in the diagnosis of the disease

The data were collected from the Iraqi society, as they data were acquired from the laboratory of Medical City Hospital and (the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital). Patients' files were taken and data extracted from them and entered in to the database to construct the diabetes dataset. The data consist of medical information, laboratory analysis. The data attribute are:

The data consist of medical information, laboratory analysis… etc. The data that have been entered initially into the system are: No. of Patient, Sugar Level Blood, Age, Gender, Creatinine ratio(Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting lipid profile, including total, LDL, VLDL, Triglycerides(TG) and HDL Cholesterol , HBA1C, Class (the patient's diabetes disease class may be Diabetic, Non-Diabetic, or Predict-Diabetic).

• Group 1 - Support Vector Machine and Random Forest

• Group 2 - Decision Tree and KNN

• Group 3 – Catboost and Logistic Regression

• Group 4 - Naïve Bayes and AdaBoosting

• Group 5 – AdaBoosting and Gradient Boostin