**Disease Diagnosis Prediction Using Gradient Boosting**

# Objective

The goal of this task was to build a machine learning model to predict the likelihood of diseases such as diabetes using the PIMA Diabetes dataset. Gradient Boosting was used as the primary classification algorithm. The outcome aimed to provide actionable insights to aid early detection and prevention of the disease.

# Methodology

1. Data was loaded from the PIMA Diabetes Dataset available through Seaborn.  
2. Exploratory Data Analysis (EDA) was conducted to explore relationships between features (e.g., Age, BMI, Glucose).  
3. Data was scaled using StandardScaler for better model performance.  
4. Gradient Boosting Classifier was trained using the scaled features.  
5. Model performance was evaluated using F1 Score and AUC-ROC curves.  
6. Feature importance was visualized to understand the most influential factors in prediction.

# Findings

The Gradient Boosting model showed promising performance for binary classification of diabetes prediction. Important features identified included Glucose, BMI, Age, and Insulin. The F1 Score and AUC-ROC curve indicated that the model can effectively distinguish between positive and negative diabetes cases.

# Challenges and Solutions

A major challenge encountered was a mismatch between the number of model features and the number of feature names when plotting feature importance. This was due to training the model on a reduced feature set but using the full set of column names from the original dataset.  
  
To resolve this, we ensured that the correct set of feature names matched the number of features used to train the model. Alternatively, the model was retrained on the full feature set to align with the original dataset columns.

# Conclusion

This project demonstrates the effective use of Gradient Boosting for early disease prediction. With proper data preprocessing, scaling, and model evaluation, we can build a reliable predictive system to assist healthcare professionals. Future improvements may include testing additional models and integrating real-time patient data for higher accuracy.