**Identification of Diabetes Based on Healthcare Statistics**

Chiluka Akshitha   
16-12-2024

**Abstract**

This project aims to predict diabetes status (Healthy, Pre-diabetic, Diabetic) using healthcare and lifestyle data. By leveraging machine learning algorithms, we identified critical features impacting diabetes risk and developed a highly accurate model (99%) for early detection. This documentation captures the journey from data collection to final recommendations.

**Introduction**

**Problem**: Diabetes is a growing global health concern. Early prediction can significantly reduce risks through timely intervention.  
**Goal**: Build a machine learning model to predict diabetes status based on healthcare and lifestyle statistics.

**Data Collection**

**Source**:

Age, BMI, Glucose Level, Physical Activity, Family History.

**Data Preprocessing**

1. **Missing Values**: Imputed with mean/mode.
2. **Outliers**: Handled using IQR method.
3. **Encoding**: Label encoding for categorical data.
4. **Scaling**: StandardScaler for numerical features.

**Results**: Dataset prepared with balanced classes and clean features for training.

**Feature Selection**

**Techniques**:

* Correlation Analysis: Removed features with correlation > 0.8.

**Outcome**: Final features included Age, BMI, Glucose Level, etc.

**Model Building**

**Algorithms Tested**: Logistic Regression, Random Forest, Gradient Boosting, SVM, XGBoost.

**Best Model**: Extra Trees

* Accuracy: 100%.
* F1-Score: 100%.
* AUC: 1.0.

**Hyperparameter Tuning**:

* Optimized estimators, depth, and splits for best performance.

**Evaluation Metrics and Results**

* **Metrics Used**: Accuracy, Precision, Recall, F1-Score, AUC.

