

DIABETES PREDICTION USING MACHINE LEARNING

A COURSE PROJECT REPORT

Submitted by

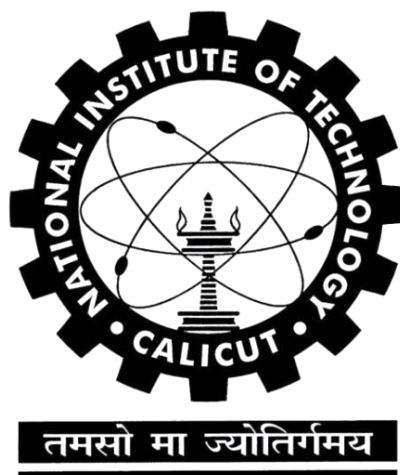
Balaji R
(B230865EE)

Electrical & Electronics engineering

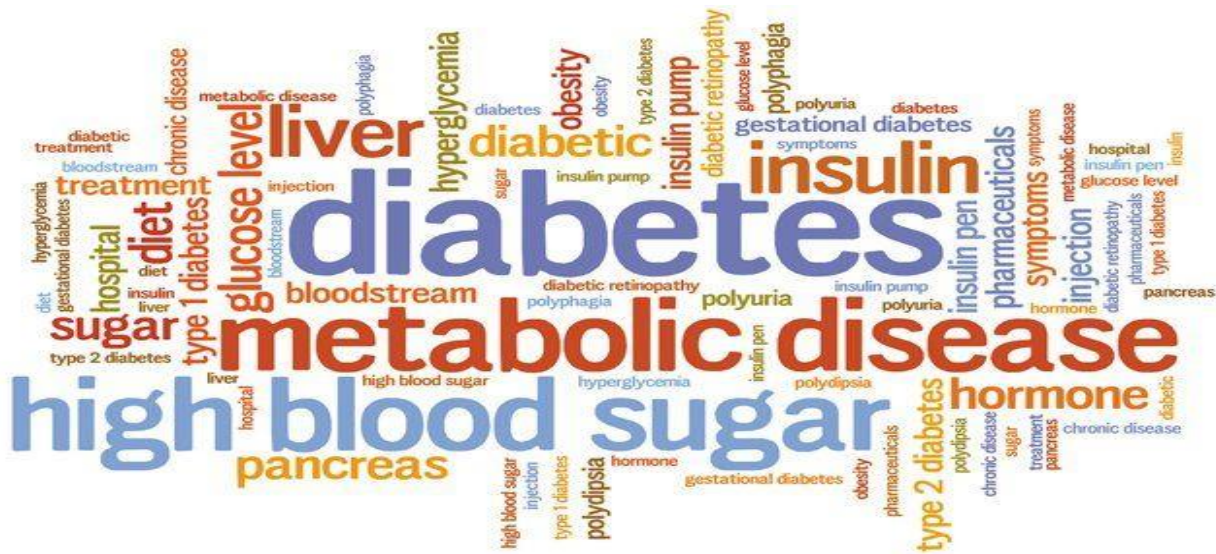
IN COMPLETION OF THE COURSE

EE3058E: Essentials of AI and Machine Learning

Course Faculty: Dr. Shihabudheen K V (EED)



NATIONAL INSTITUTE OF TECHNOLOGY CALICUT
Date – 11th April 2025



ABSTRACT

Diabetes is a chronic condition that impairs the body's ability to process blood glucose, leading to severe complications if not diagnosed and managed early. This project leverages machine learning techniques to develop a predictive model for diabetes diagnosis using the Pima Indian Diabetes dataset. The project involves data preprocessing, exploratory data analysis, model training, evaluation, and comparison across different algorithms. Among the models explored—

- Logistic Regression,
- K-Nearest Neighbors (KNN),
- Support Vector Machine (SVM),
- Random Forest Classifier
- Neural Network
- MLP regressor
- Decision tree classifier

the Random Forest model demonstrated the best performance with an accuracy of 97.9% while validating. This study highlights the potential of machine learning in clinical decision-making and healthcare automation.

INTRODUCTION

Diabetes mellitus, a metabolic disorder characterized by chronic hyperglycaemia, is a leading cause of death and disability worldwide. Early diagnosis is essential to prevent long-term complications such as cardiovascular disease, kidney failure, and neuropathy. Traditional diagnostic methods, though reliable, are often time-consuming and dependent on continuous clinical oversight.

In this context, machine learning (ML) offers an efficient alternative by identifying hidden patterns in patient data, thereby automating early detection. This project implements multiple supervised learning algorithms to classify individuals as diabetic or non-diabetic, based on physiological and medical parameters. The study not only focuses on building accurate models but also aims to understand the influence of each health parameter on the model's decision-making process.

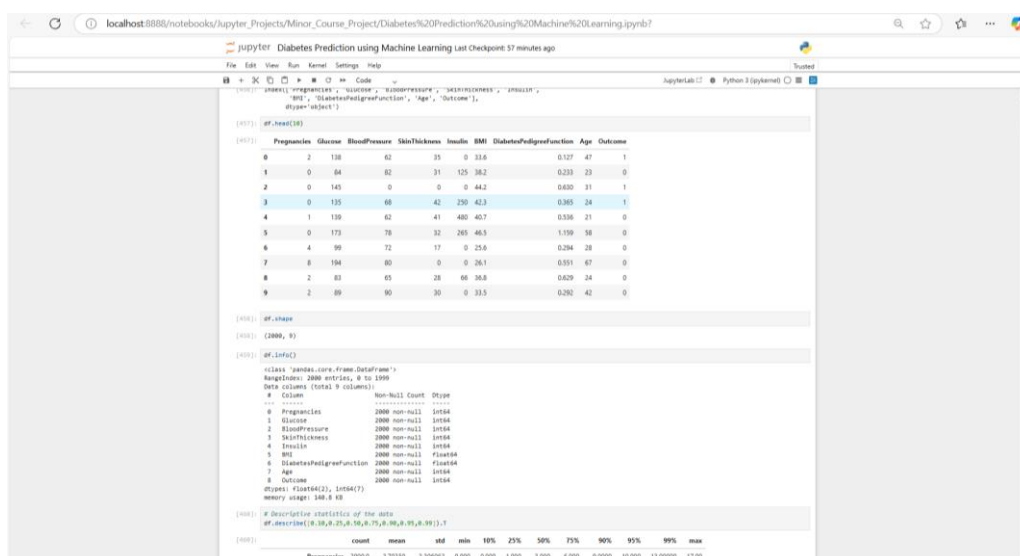
PROBLEM STATEMENT

The objective of this project is to build and evaluate various machine learning classification models for predicting the presence of diabetes in patients using a set of medical features. The challenge lies in selecting the most effective model that delivers high accuracy, precision, and recall, while also ensuring interpretability for potential integration into real-world healthcare systems.

DETAILS ABOUT THE DATASET

The dataset used for this study is the Pima Indian Diabetes dataset, sourced from the UCI Machine Learning Repository. It contains data from 768 female patients of Pima Indian heritage, aged 21 and above. The dataset has 8 independent variables and 1 dependent variable:

- **Pregnancies:** Number of times pregnant
- **Glucose:** Plasma glucose concentration 2 hours in an oral glucose tolerance test
- **BloodPressure:** Diastolic blood pressure (mm Hg)
- **SkinThickness:** Triceps skin fold thickness (mm)
- **Insulin:** 2-Hour serum insulin (mu U/ml)
- **BMI:** Body mass index (weight in kg/ (height in m) ^2)
- **DiabetesPedigreeFunction:** Diabetes pedigree function
- **Age:** Age (years)
- **Outcome:** Class variable (0 or 1), 1 indicates diabetic, 0 indicates non-diabetic



```
[007]: df.head(10)
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	82	62	35	0.336	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.403	31	1
3	0	135	68	40	290	42.3	0.365	24	1
4	1	130	82	41	480	40.7	0.336	21	0
5	0	173	78	32	265	46.3	1.179	58	0
6	4	99	72	17	0	25.6	0.204	28	0
7	8	194	80	0	0	26.1	0.591	67	0
8	2	83	65	28	66	38.8	0.429	24	0
9	2	89	90	30	0	33.5	0.282	42	0

```
[008]: df.shape
```

```
[009]: (2808, 9)
```

```
[010]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 2808 entries, 0 to 2807
```

```
Data columns (total 9 columns):
```

#	Column	non-null count	Dtype
0	Pregnancies	2808 non-null	int64
1	Glucose	2808 non-null	int64
2	BloodPressure	2808 non-null	int64
3	SkinThickness	2808 non-null	int64
4	Insulin	2808 non-null	float64
5	BMI	2808 non-null	float64
6	DiabetesPedigreeFunction	2808 non-null	float64
7	Age	2808 non-null	int64
8	Outcome	2808 non-null	int64

```
memory usage: 148.8 KB
```

```
[011]: # Descriptive statistics of the data
```

```
df.describe([0.19,0.25,0.76,0.75,0.76,0.75,0.75,0.75])
```

```
[012]:
```

	count	mean	std	min	10%	25%	50%	75%	90%	95%	max
Pregnancies	2808.0	3.70710	3.30600	0.000	0.000	1.000	3.000	6.000	9.000	10.000	17.00

DATA PREPROCESSING

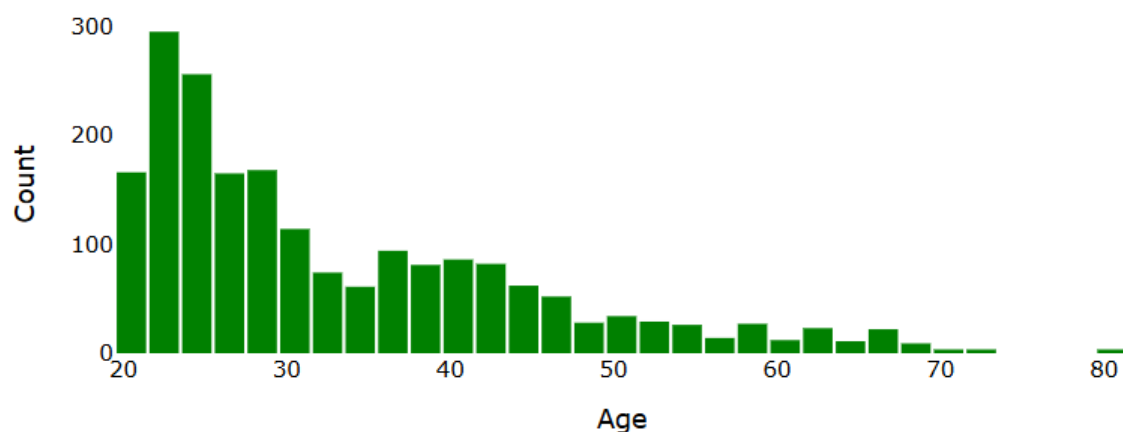
Data preprocessing was critical to ensure model reliability. The following steps were implemented:

- **Missing Value Handling:** Zero values in Glucose, BloodPressure, SkinThickness, Insulin, and BMI were replaced with median values, treating them as missing.
- **Normalization:** Min-Max scaling was applied to all features to bring them to a common scale of [0, 1].
- **Train-Test Split:** The dataset was divided into training and testing sets using an 80-20 split to evaluate model generalizability.

EXPLORATORY DATA ANALYSIS

DISTRIBUTION OF AGE GROUP

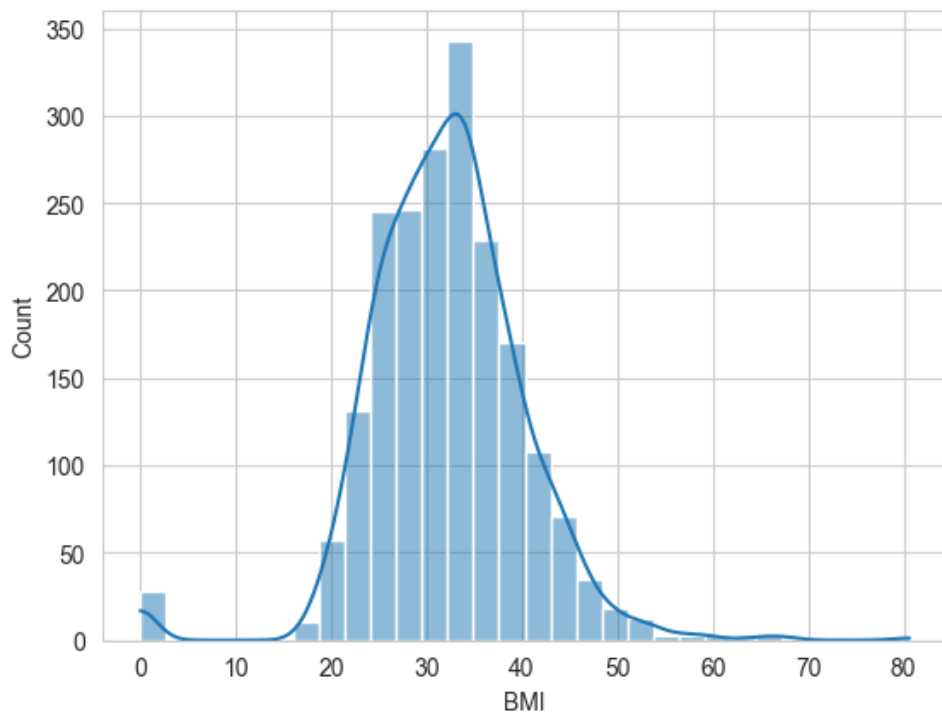
Histogram of Age



Insights:

1. **Age Range:** The ages of individuals in the dataset span from 20 to 80 years, with a noticeable concentration in younger age groups.
2. **Most Frequent Age Group:** The highest frequency is observed in the 20–30 age range, with over 300 individuals aged between 20 and 25.

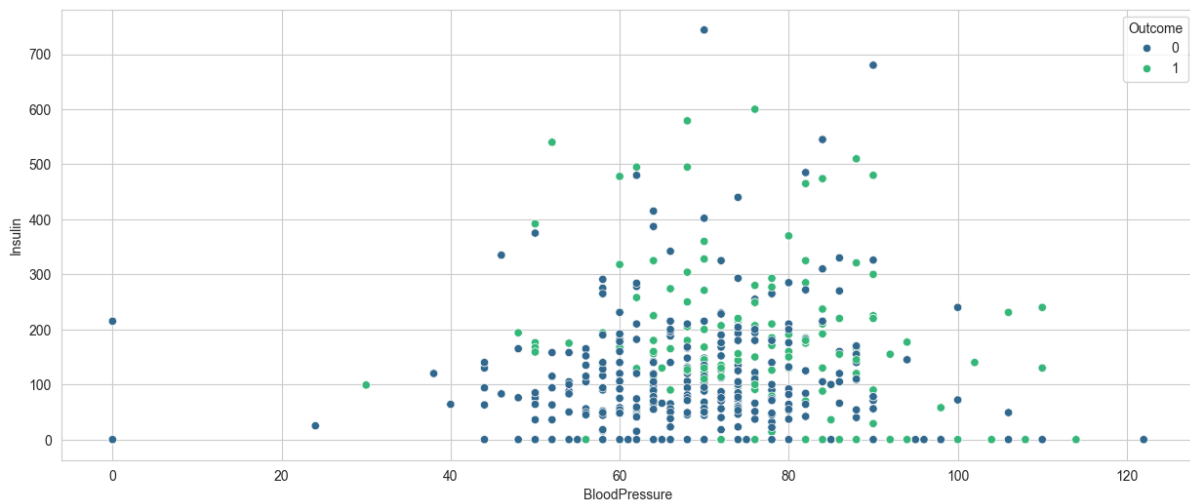
DISTRIBUTION OF BMI



Insights:

1. **Range of BMI Values:** BMI values range from 0 to approximately 80, but most values fall between 20 and 50.
2. **Peak Frequency:** The highest frequency occurs around a BMI of 30, indicating that many participants fall into the overweight category.
3. **Shape of Distribution:** The distribution is approximately normal, with a slight right skew due to higher BMI values beyond 40.

SCATTER PLOT : INSULIN LEVEL (VS) BLOOD PRESSURE

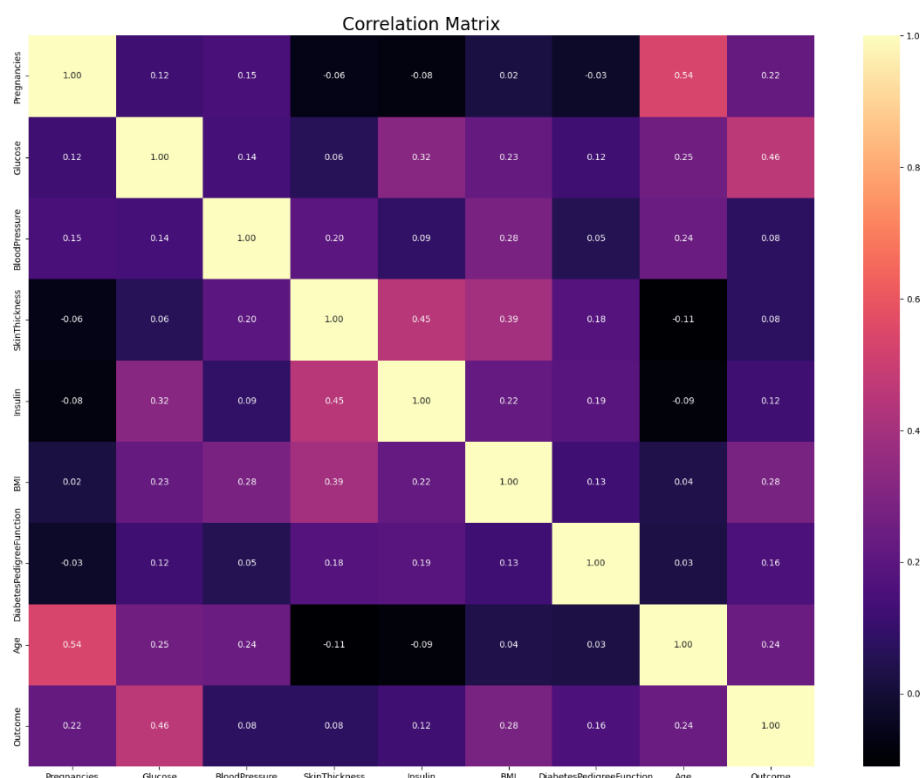


1. **Blood Pressure Range:** Most values are between 60–90 mmHg.
2. **Insulin Levels:** Concentration is below 200, with some outliers above 500.
3. **Outcome Distribution:** Diabetic (green) points are more frequent at higher insulin levels (>200).
4. **Clusters:** A dense cluster exists between 60–80 mmHg blood pressure and <200 insulin levels.

CORRELATION MATRIX

Insights from Correlation Matrix

1. **Strong Correlations:**
 - **Glucose** shows the highest positive correlation with diabetes outcome (Outcome), with a correlation coefficient of **0.46**.
 - **BMI** also has a moderate positive correlation with diabetes outcome at **0.28**.
2. **Weak Correlations:**
 - Features like **Skin Thickness**, **Insulin**, and **Blood Pressure** show weak correlations with diabetes outcome, indicating they may have limited predictive power individually.
3. **Feature Relationships:**
 - **Skin Thickness** and **Insulin** have a moderate correlation (**0.45**) with each other, suggesting some interdependence.
 - **Age** and **Pregnancies** are positively correlated (**0.54**) but show weaker connections to diabetes outcome



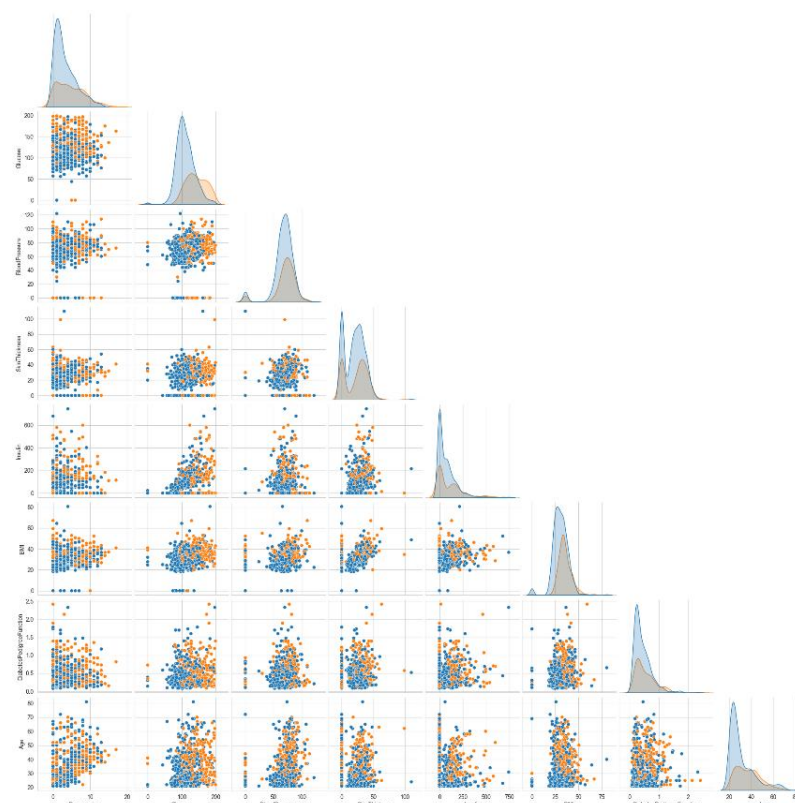
PAIR PLOT

1. Individual Feature Distributions

- **Glucose:**
 - Diabetic individuals (orange points) tend to have higher glucose levels (>120 mg/dL).
 - Non-diabetic individuals (blue points) are concentrated in lower glucose ranges (<120 mg/dL).
- **BMI:**
 - Diabetic individuals generally have higher BMI values (>30), indicating a strong association with diabetes risk.
- **Insulin:**
 - Insulin levels show significant variation, with diabetic individuals often having elevated levels (>200 units).
- **Age:**
 - Diabetic individuals are more frequent in older age groups (>40 years), while younger individuals are predominantly non-diabetic.

2. Feature Relationships

- **Glucose vs. Insulin:**
 - A positive trend is visible, with higher glucose levels often accompanied by elevated insulin levels, particularly for diabetic individuals.
- **BMI vs. Glucose:**
 - Higher BMI values correlate with higher glucose levels, especially for diabetic individuals.
- **Age vs. Pregnancies:**
 - A positive correlation is observed, as older women tend to have more pregnancies.



OUTLIER DETECTION

METHODS USED FOR OUTLIER DETECTION

STATISTICAL METHODS:

- **Interquartile Range (IQR):** Identifies outliers as points outside $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$.
- **Z-Score:** Flags data points with Z-scores greater than a threshold (e.g., $|Z| > 3$).

VISUALIZATION TECHNIQUES:

- Box plots were used to visually identify outliers in features like insulin and BMI.

MODEL-BASED METHODS:

- **Isolation Forest:** Utilized to detect anomalies by isolating observations that deviate significantly from the norm.
- **Local Outlier Factor (LOF):** Applied to measure the local density deviation of a given data point compared to its neighbors.

HANDLING OUTLIERS

- Outliers were handled using:
 - **Capping:** Replacing extreme values with upper and lower thresholds derived from IQR.
 - **Removal:** Removing data points flagged by Isolation Forest or LOF as anomalies.
- Example: Insulin values above a calculated upper threshold were capped to reduce skewness.

IMPACT ON DATASET

- After handling outliers:
 - The dataset became more balanced and representative of real-world conditions.
 - Reduced the risk of overfitting by eliminating extreme values that could disproportionately influence model training.

FEATURE ENGINEERING

CATEGORICAL BMI RANGES:

- BMI values were categorized into ranges to represent weight classifications:
 - *Underweight*: $\text{BMI} < 18.5$
 - *Normal*: $18.5 \leq \text{BMI} \leq 24.9$
 - *Overweight*: $25 \leq \text{BMI} \leq 29.9$
 - *Obesity Classes*: $\text{BMI} > 30$, divided into Obesity Class I, II, and III.
- This transformation captures non-linear relationships between BMI and diabetes risk.

CATEGORICAL GLUCOSE LEVELS:

- Glucose levels were divided into categories based on medical thresholds:
 - *Low*: $\text{Glucose} \leq 70$
 - *Normal*: $70 < \text{Glucose} \leq 99$
 - *Prediabetic*: $99 < \text{Glucose} \leq 126$
 - *High*: $\text{Glucose} > 126$
- This feature highlights the correlation between glucose levels and diabetes outcomes.

INSULIN SCORE:

- Insulin values were categorized into ranges to differentiate between normal and abnormal insulin levels:
 - *Low*: $\text{Insulin} < 50$
 - *Normal*: $50 \leq \text{Insulin} \leq 200$
 - *High*: $\text{Insulin} > 200$
- This feature helps identify individuals with potential insulin resistance.

STEPS TAKEN

DATA TRANSFORMATION:

- Continuous variables (e.g., BMI, glucose, insulin) were converted into categorical variables using predefined thresholds.
- Label encoding and one-hot encoding were applied to convert categorical features into numerical formats suitable for machine learning models.

FEATURE SELECTION:

- Recursive Feature Elimination with Cross-Validation (RFECV) was used to identify the most informative features.
- Correlation analysis ensured that derived features contributed meaningful information without redundancy.

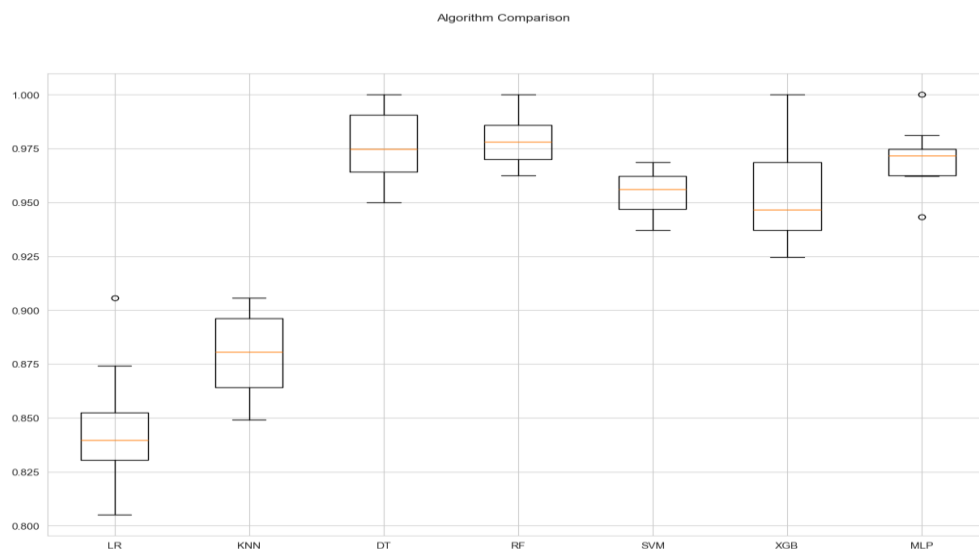
MODEL SELECTION

Base models are evaluated with Training data using cross validation

Model	Accuracy (%)	Std. Dev
Logistic Regression	84.48	± 2.68
K-Nearest Neighbors	87.81	± 1.96
Decision Tree	97.61	± 1.55
Random Forest	97.93	± 1.12
SVM	95.54	± 1.02
XGBoost (Gradient Boosting)	95.29	± 2.32
MLP (ANN)	97.04	± 1.40

Observation:

Ensemble models (Random Forest, XGBoost) and MLP significantly outperformed the simpler ones like LR and KNN.



FINAL MODELS

To optimize model performance, **GridSearchCV** was used to search for the best hyperparameters by evaluating multiple combinations through cross-validation. This helped in selecting the optimal values for parameters like `n_estimators`, `max_depth`, and `learning_rate` across different models, improving accuracy and reducing overfitting.

SUPPORT VECTOR MACHINE

Performance:

- **Cross-validation Accuracy:** 94.22%
- **Test Set Accuracy:** 97.24%
- **Precision:** 97.76%
- **Recall:** 94.24%
- **F1 Score:** 95.97%

RANDOM FOREST

Performance:

- **Cross-validation Accuracy:** 97.17%
- **Test Set Accuracy:** 99.49%
- **Precision:** 99%
- **Recall:** 99%
- **F1 Score:** 99%

GRADIENT BOOSTING (XGBOOST):

Performance:

- **Cross-validation Accuracy:** 98.17%
- **Test Set Accuracy:** 100%
- **Precision:** 100.00%
- **Recall:** 100.00%
- **F1 Score:** 100.00%
- **ROC AUC Score:** 1.00

DECISION TREE

Performance:

- **Cross-validation Accuracy:** 97.30%
- **Test Set Accuracy:** 99.40%
- **Precision:** 99.00%
- **Recall:** 99.00%
- **F1 Score:** 99.00%
- **ROC AUC Score:** 1.00

CONCLUSION

In this project, several machine learning models were evaluated for their performance. After evaluating different machine learning models, it's clear that **Random Forest** outperforms the others, achieving the highest accuracy of **97.93%**. It's closely followed by **Decision Tree** (97.61%) and **MLP (ANN)** (97.04%), all of which performed well with relatively low standard deviations, showing consistent results across different test splits.

On the other hand, **K-Nearest Neighbors** (87.81%) and **Logistic Regression** (84.48%) didn't perform as well, but they could still be useful depending on the specific requirements and computational resources available.

While **Random Forest** clearly leads in accuracy, models like **SVM** (95.54%) and **XGBoost** (95.29%) also showed strong performance with **SVM** having the smallest variability in results, which could be important in certain applications.