**Report on Diabetes Prediction**

**ABSTRACT**

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like hear disease, kidney disease, stroke, eye problem, nerve damage etc. Some machine learning model helps in predicting the diabetes soon or much better than the doctor and easily. It will help the person to gets earlier. Here we have used some dataset that used earlier and some common factor that helps to predict the disease. The objective of this project is to create a robust system for early diabetes prediction in patients with a heightened degree of accuracy by using different machine learning methodologies. Some major approach involves the utilization of various algorithms, including K-nearest neighbor, Logistic Regression, Random Forest Classifier, Support Vector Machine, and Decision Tree.

Accuracy of model gets calculated and used for the prediction of diabetes.

**Introduction**

Diabetes refers to a chronic (long-lasting) health condition that affects how your body turns food into energy. Early detection allows individuals to take vital precautions and adapt their lifestyles to either prevent the disease or manage it effectively. This proactive approach can significantly enhance one's quality of life, reduce healthcare costs, and mitigate the risk of debilitating complications associated with diabetes.

Types of Diabetes:

1. Type 1 – This form of typically diagnosed in childhood or adolescence. In this type 1 diabetes, immune system is compromised and the cells fail to produce insulin in sufficient amounts.
2. Type 2 – Mostly cases in the world wide are belongs to this type. This type of Diabetes is seen when body cells are not able to use insulin properly.
3. Gestational Diabetes – This type of diabetes mainly occurs in women during pregnancy and affects some pregnant women.

Symptoms-

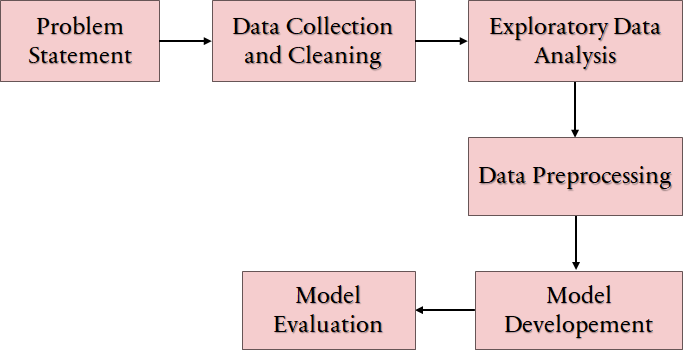
* Increased Thirst
* Tired/Sleepiness
* Weight loss
* Blurred vision
* Frequent Urination

**Existing Method**

In this method, author used dataset with attributes from blood and urine test to classify dataset. Dataset has 200 instances, collected from hospital warehouse. Using WEKA, they used machine learning algorithms, Naïve Bayes, J48, REP Tree, and Random Tree. In this he used 10-fold cross validation approach found J48 best performs and gets highest accuracy 60.2%.

In other paper, author detect the diabetes by analyzing data patterns through classification with Decision Tree and Naïve Bayes algorithms. Author used PIMA dataset and used cross validation approach and study found that J48 achieved 74.8% accuracy and Naïve Bayes achieved 79.5% accuracy.

**Proposed method with Architecture**

****

*Diabetes Prediction Model*

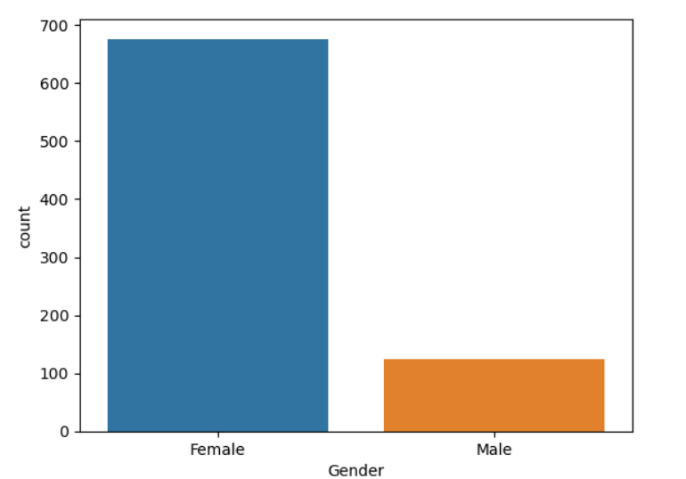
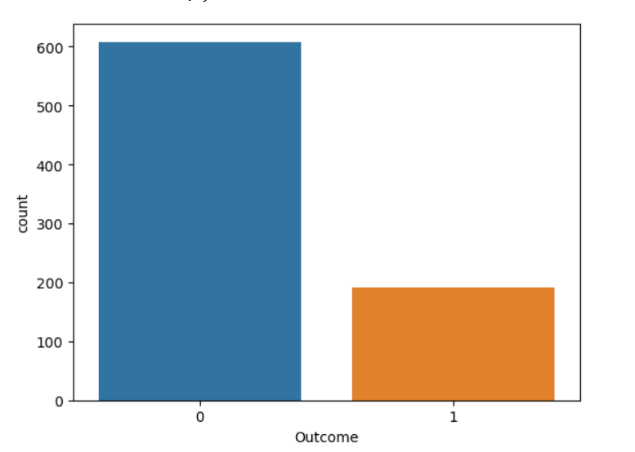
**Data Cleaning**

A Comprehensive way to clean our both dataset, in PIMA diabetes first dataset we saw that a lot of rows having filled with 0, so we remove it using the mean value of the respective column and created a new feature named as Gender as Female as its written dataset contains the female only, and in second dataset diabetes prediction there is no any missing value.

After merging the dataset, we taken random dataset using sample and taken more dataset from first rather as second, as we going to give more importance to first dataset. After merging the dataset we get many null values so I removed by mean and New features Gender having both Male and Female as Second dataset has both Gender.

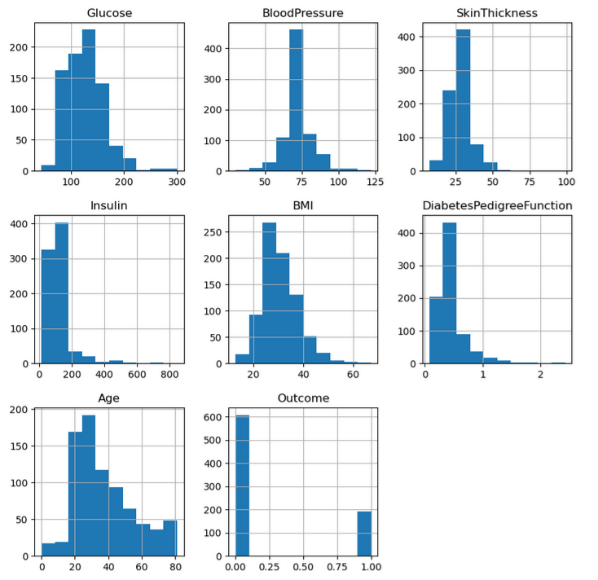
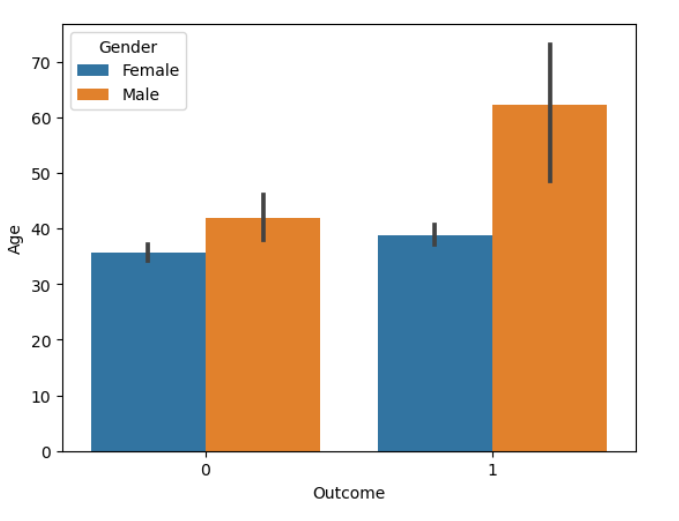
**Exploratory Data Analysis**

EDA is an important step in the data analysis process because it allows us to understand the data before applying any statistical models or making any decisions. There are some EDA diagram below to know the dataset,

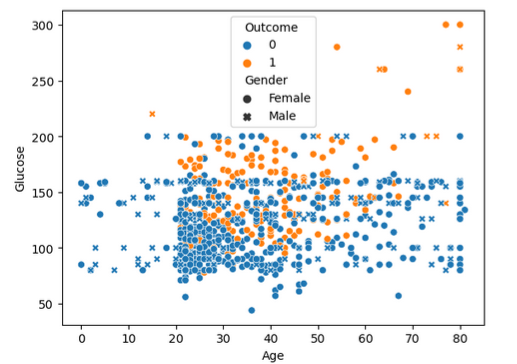
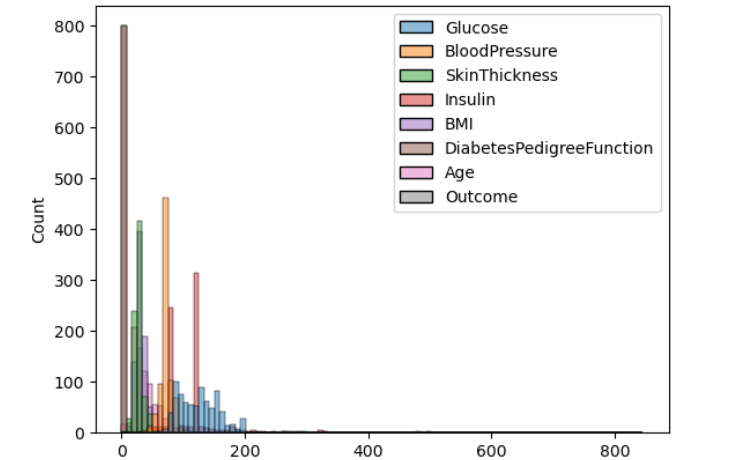


*Fig 2. Patients having diabetic or not*

*Fig 1. Number of Male and Female*

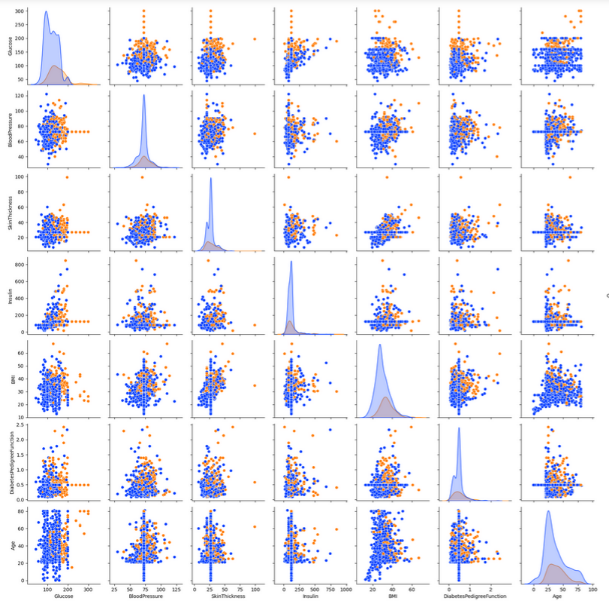


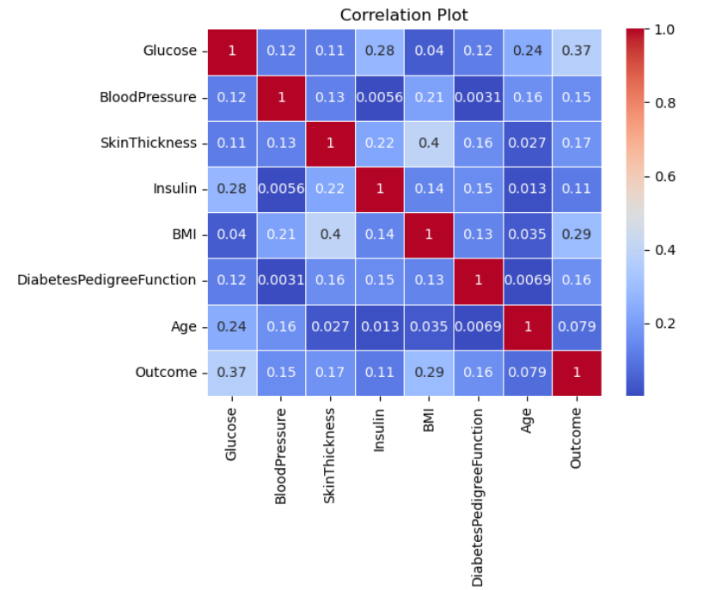
*Fig 3. Distribution Plot of Dataset*



*Fig 6. Hist plot on all dataset*

*Fig 5. Scatter plot showing the relation with outcome and gender with age*





*Fig 8. Pair plot of all feature*

*Fig 7. Correlation between the features*

**Data Preprocessing**

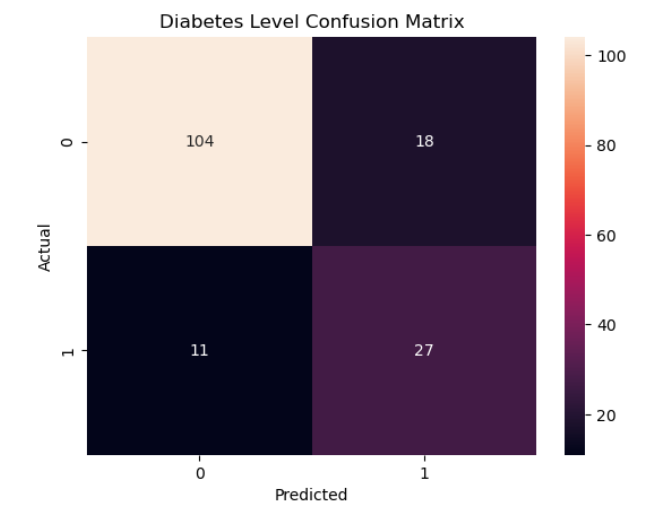
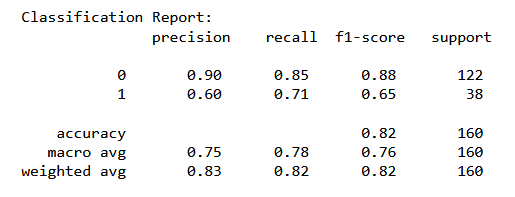
* Data Splitting: The dataset was divided into training and testing sets using the train-test split method. This separation ensures the evaluation of models on unseen data helping to identify potential issues.
* Label Encoding: Label encoding was applied to the categorical “Gender” column, converting them into numerical representations
* Handling Outcome Imbalance: SMOTE (Synthetic Minority Over-sampling Technique) used to rectify Outcome imbalance, enhancing model performance.
* Feature Scaling: MinMaxScaler is used so data to be transformed into the minimum value becomes 0, and the maximum value becomes 1. This scaling method is referred to as 'min-max scaling' or 'feature scaling to a specific range.

**Model Development**

In the critical phase of building a diabetes prediction model, we extensively assessed multiple classification algorithms, including the Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Logistic Regression, K-Nearest Neighbor, and Gaussian Naïve Bayes. Among these contenders, the Random Forest Classifier stood out as the top performer, demonstrating its robust predictive capabilities.

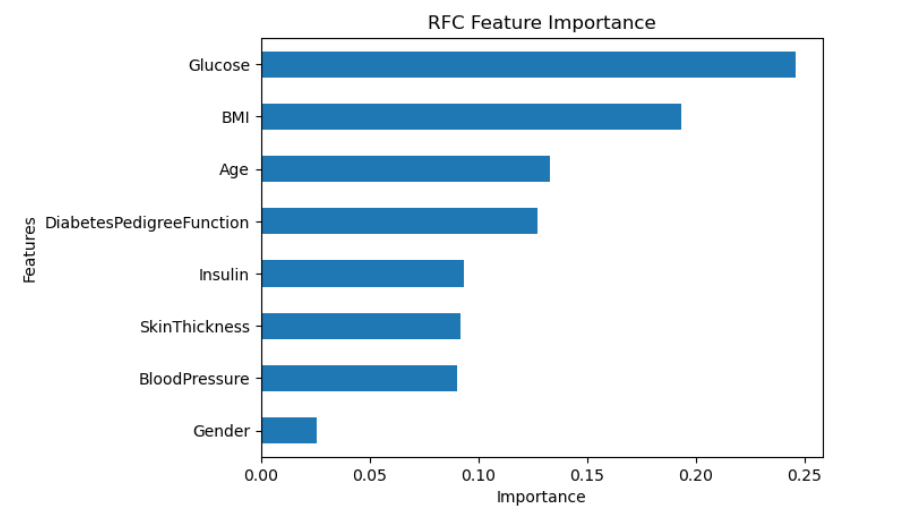
The selected Random Forest Classifier underwent rigorous training on scaled training data, achieving an impressive training accuracy of 99.38%. This high accuracy underscores the model's capacity to capture complex patterns in the data effectively. Moreover, when applied to unseen data in the test phase, it maintained a commendable accuracy rate of 81.87%, confirming its reliability and generalizability in real-world scenarios.

This comprehensive evaluation has firmly established the Random Forest Classifier as the leading choice in the realm of diabetes prediction, offering a robust and dependable solution for early detection and management of this critical health condition.

*Classification Report*

*Confusion Matrix*



*Feature Importance*

**Conclusion**

This research paper focuses on diabetes prediction using various machine learning classification algorithms. After a thorough evaluation, the Random Forest Classifier emerged as the top-performing model. With a remarkable 99% accuracy on the training dataset, the Random Forest Classifier demonstrates its adeptness at capturing intricate data patterns. Furthermore, the model maintains an 82% accuracy rate when tested on new data, showcasing its robust generalization capabilities. In the context of diabetes prediction, where early and accurate detection is paramount for effective intervention, the Random Forest Classifier's outstanding performance positions it as a valuable tool for improving healthcare outcomes and patient well-being.