|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| **[Pima Indians Diabetes Database]**  [Machine learning course] |
| |  |  |  | | --- | --- | --- | | Aya Elneanaei fouda Abdallah Ashraf Elsaadany |  | Ahmed Ashraf Mohamed Youssif Amr Shaaban | |

**Table of Contents**

1 Abstract ……………………………………………….…………………………………….

2 Introduction…………………………………………………………………………………….

3 Data Explanation………………………………………………………………………………

4 Literature Review……………………………………………………………………………...

4.1 Intelligent Methods of Diabetes Prediction………………………………………….……….

4.2 The Selected Machine Learning Algorithms………………………………….

4.2.1 AdaBoost ………………………………………………………………

4.2.1.1 the steps for performing the AdaBoost algorithm……………………

4.2.2 Random Forest………………………………………………………….

4.2.2.1 the steps for performing the Random Forest algorithm………….

4.2.3 K Nearest Neighbors……………………………………………………

4.2.4 Logistic Regression…………………………………………………….

4.2.4.1 the steps for performing the Logistic Regression algorithm………….

4.2.5 Ensemble Classifier…………………………………………………...

5 Data and Methodology………………………………………………………….

5.1 Dataset Exploration and Pre-Processing……………………………….

5.1.1 Replace Missing Value……………………………………………

5.1.2The dataset is unbalanced……………………………………….

5.1.3 Exploratory Data Analysis………………………………….

5.1.3.1 Data Before cleaning………………………………………

5.1.3.1.1 Preprocessing……………………………….

5.1.3.2 Data After cleaning……………………………………

5.2 Experiment and Results……………………………………………….

5.2.1Results of Machine Learning Algorithms……………………..

5.2.1.1 AdaBoost ……………………………………………......

5.2.1.1.1 AUC-ROC Curves\_AdaBoost ……………………………….

5.2.1.2 Random Forest………………………………………………………….

5.2.1.2.1 AUC-ROC Curves\_ Random Forest ……………………………….

5.2.1.3. K Nearest Neighbors……………………………………………………

5.2.1.3.1 AUC-ROC Curves\_ K Nearest Neighbors……………………….

5.2.1.4 Logistic Regression…………………………………………………….

5.2.1..4.1 AUC-ROC Curves\_ Logistic Regression ………………………

5.2.1.5 Ensemble Classifier…………………………………………………...

5.2.5.1 AUC-ROC Curves\_ Ensemble Classifier …………………………

5.3 Final Results…………………………………………………………………

6 Discussion and Conclusion………………………………………………………..

7 References………………………………………………………………………..

Abstract

Diabetes is a condition that can lead to numerous illnesses, including heart attack, kidney disease, blindness, renal failure, etc. The endocrine (hormone) system issue, which is the most prevalent, develops when the body's blood sugar levels persistently remain over simple. There are two different varieties of diabetes: one occurs when the body is unable to produce insulin, and the other happens when the body does not react to the effects of insulin. In other words, either insufficient insulin is produced, or the body's cells do not interact with insulin as intended. This research delineates the use of four interpretable supervised ML models: AdaBoost, Random Forest Classifier, K Nearest Neighbors, and Logistic Regression models to be trained and tested using the Pima Indians Diabetes dataset in python programming language. The performance of each algorithm is analyzed to determine the one with the best accuracy, precision, sensitivity, and specificity.

Introduction

Diabetes is not only an illness, but it also causes various other disorders. Diabetes can lead to several complex conditions, such as kidney problems, heart issues, and blindness. Additionally, it raises the affected individual's blood sugar levels. When the body's glucose level rises by more than 200 mg/dL, micro, and macro coronary heart problems manifest. This circumstance arises when diabetes is out of control. Type-2 diabetes is becoming more common at an alarming rate in both industrialized and developing nations worldwide. It is one of the primary causes of death and a host of medical issues.

Patients with diabetes are consequently a financial burden on both their families and the healthcare system. Numerous studies have shown that modifying one's diet and lifestyle can easily prevent this condition. In wealthy nations, obesity has affected many people. Western nations have rather strange dining customs. Most individuals consume junk food, which is high in carbohydrates and saturated fats. These foods don't have any fiber. They are therefore more susceptible to many different ailments, particularly diabetes.

Diabetes type 1 is a metabolic condition marked by elevated blood sugar levels. Type 2 diabetes frequently leads to illnesses like heart attacks, strokes, blindness, and renal failure. Therefore, there is an unbreakable link between the worldwide obesity epidemic and the enormous rise in type 2 diabetes. Diagnosing diabetes is a difficult but necessary task.

This paper seeks to develop an e-diagnosis system for detecting and classifying diabetes. Through the use of machine learning algorithms (AdaBoost, Random Forest Classifier, K Nearest Neighbors, and Logistic Regression ), based on a number of risk factors, the system will be able to determine if a person is at risk for developing diabetes, offer a preliminary diagnosis to doctors, and relay to patients the doctor's recommendations for diet, exercise, and blood glucose monitoring. To determine which classifier performed the best, these classification models were tested using a variety of techniques, including Accuracy, Precision, Sensitivity, F-measure, and Area Under Receiver Operating Characteristics (AUROC) Curve.

The top classification model yielded several relevant characteristics that may be utilized to predict the severity of diabetes. For this experiment, the Pima Indian Diabetes dataset is used. Native Americans known as Pima Indians reside in Arizona and Mexico. State of America. It was determined that the incidence of diabetes mellitus was high in this group. They were therefore considered important to and emblematic of world health in the studies conducted around them. A well-known benchmark dataset is the Pima Indian Diabetes dataset, which includes Pima Indian females aged 21 and older. Members of indigenous or underrepresented minority groups value this organization as well. The dataset's characteristics include measurements that don't need a lot of testing. This capability is crucial in emergency circumstances and patient self-care, both of which are growing in popularity.

Data Explanation

* **Pregnancies**: Number of times pregnant
* **Glucose**: Plasma glucose concentration 2 hours in an oral glucose tolerance test
* **BloodPressure**: Diastolic blood pressure (mm Hg)
* **Skin Thickness**: 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**: indicates the function which scores likelihood of diabetes based on family history.
* **Age**: Age (years)
* **Outcome**: Class variable (0 or 1)

Literature Review

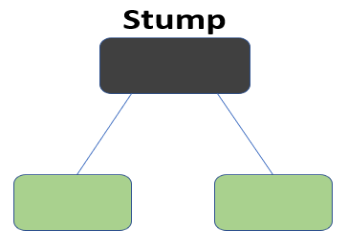
4.1 Intelligent Methods of Diabetes Prediction

By illustrating common problems, emerging technologies in data science can bring benefits to other fields of science, including medicine. Much research has used various machine learning or AI methods for diabetes prediction, such as random forest, logistic regression, and KNN, to predict diabetes at an early stage or onset. Early diagnosis of diabetes and appropriate treatment influence costs and mortality in the later stage. Early diagnostic and testing expenses are critical.

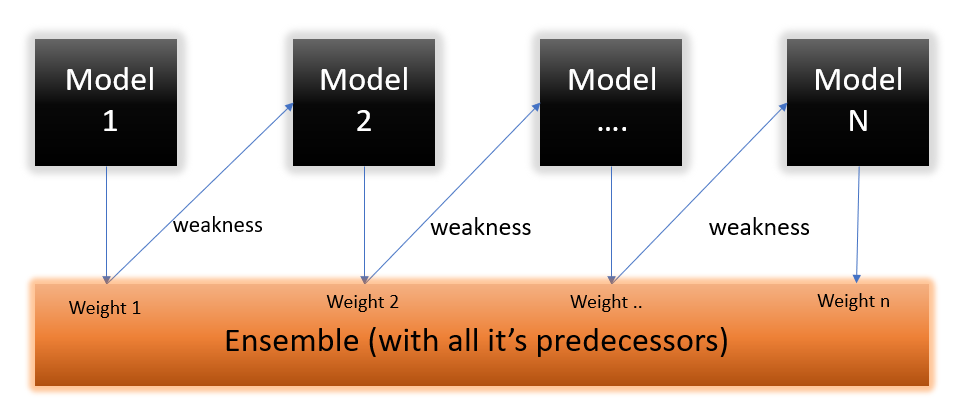
4.2 The Selected Machine Learning Algorithms

4.2.1 AdaBoost

Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called Decision Stumps.



What this algorithm does is that it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. Now all the points which have higher weights are given more importance in the next model. It will keep training models until and unless a lower error is received.



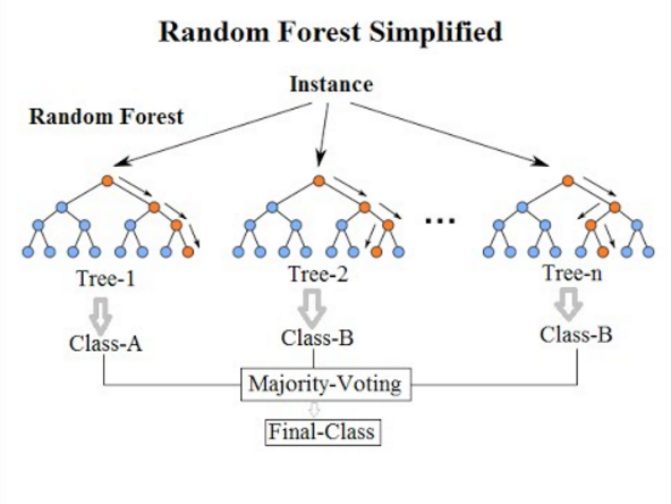
4.2.1.1 the steps for performing the AdaBoost algorithm:

* 1. Initially, all observations are given equal weights.
  2. A model is built on a subset of data.
  3. Using this model, predictions are made on the whole dataset.
  4. Errors are calculated by comparing the predictions and actual values.
  5. While creating the next model, higher weights are given to the data points which were predicted incorrectly.
  6. Weights can be determined using the error value. For instance, the higher the error the more is the weight assigned to the observation.
  7. This process is repeated until the error function does not change, or the maximum limit of the number of estimators is reached.

4.2.2 Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

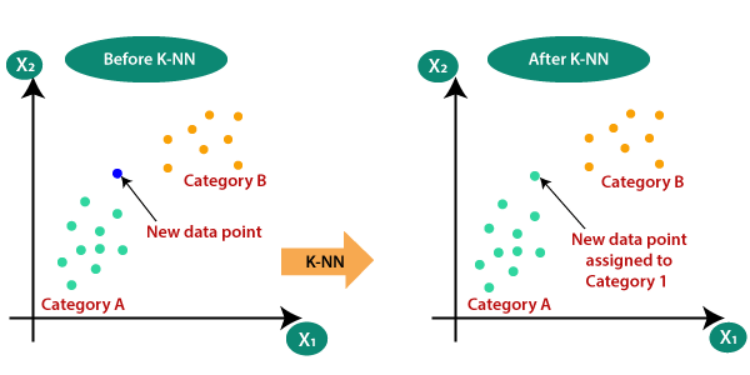


4.2.2.1 the steps for performing the Random Forest algorithm.

1. In Random Forest n number of random records are taken from the data set having k number of records.
2. Individual decision trees are constructed for each sample.
3. Each decision tree will generate an output.
4. Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

the random forest can generalize over the data in a better way. This randomized feature selection makes random forest much more accurate than a decision tree.

* + 1. K Nearest Neighbors

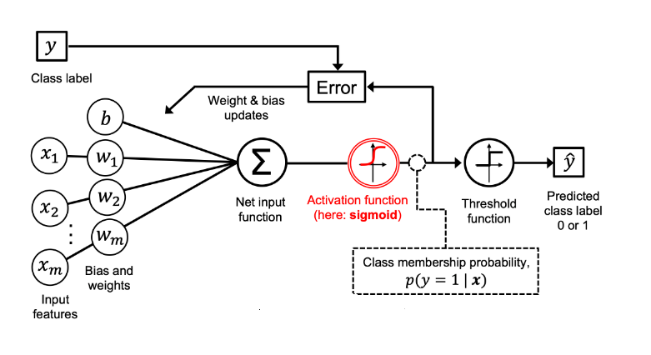
Yes, this is the principle behind K Nearest Neighbors. Here, nearest neighbors are those data points that have minimum distance in feature space from our new data point. And K is the number of such data points we consider in our implementation of the algorithm. Therefore, distance metric and K value are two important considerations while using the KNN algorithm. Euclidean distance is the most popular distance metric. You can also use Hamming distance, Manhattan distance, Minkowski distance as per your need. For predicting class/ continuous value for a new data point, it considers all the data points in the training dataset. Finds new data point’s ‘K’ Nearest Neighbors (Data points) from feature space and their class labels or continuous values.

Then: For classification: A class label assigned to the majority of K Nearest Neighbors from the training dataset is considered as a predicted class for the new data point.

For regression: Mean or median of continuous values assigned to K Nearest Neighbors from training dataset is a predicted continuous value for our new data point.

* + 1. Logistic Regression

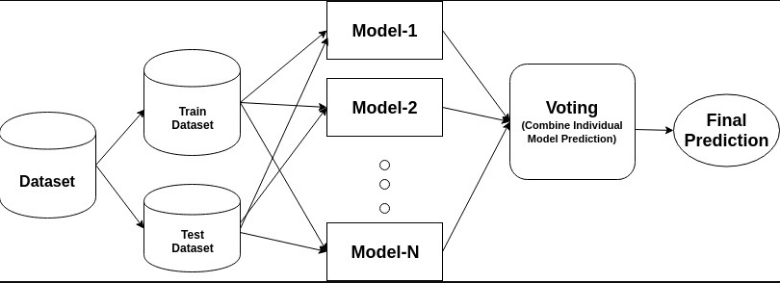
Logistic Regression is a supervised machine learning Algorithm that is used widely in Classification and Regression problems. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. Logistic regression could be used to predict whether a person is diabetic or not. These binary outcomes allow straightforward decisions between two alternatives. Logistic regression has become an important tool in the discipline of machine learning. It allows algorithms used in machine learning applications to classify incoming data based on historical data. As additional relevant data comes in, the algorithms get better at predicting classifications within data sets.



* + - 1. the steps for performing the Logistic Regression algorithm.

1. Data Pre-processing step.
2. Fitting Logistic Regression to the Training set.
3. Predicting the test result.
4. Test accuracy of the result (Creation of Confusion matrix)
5. Visualizing the test set result.
   * 1. Ensemble Classifier

Ensemble learning is a way of generating various base classifiers from which a new classifier is derived which performs better than any constituent classifier. These base classifiers may differ in the algorithm used, hyperparameters, representation or the training set.



For every iteration,

1. A base model is created on each of these samplings.
2. The models run in parallel and are independent of each other.
3. The final predictions are determined by combining the predictions from all the models.

These models collectively form a higher graded model to produce more accuracy. The final model is averaged by:

e= (**Σ** eᵢ)/n

where eel₂…. eₙ = base classifier

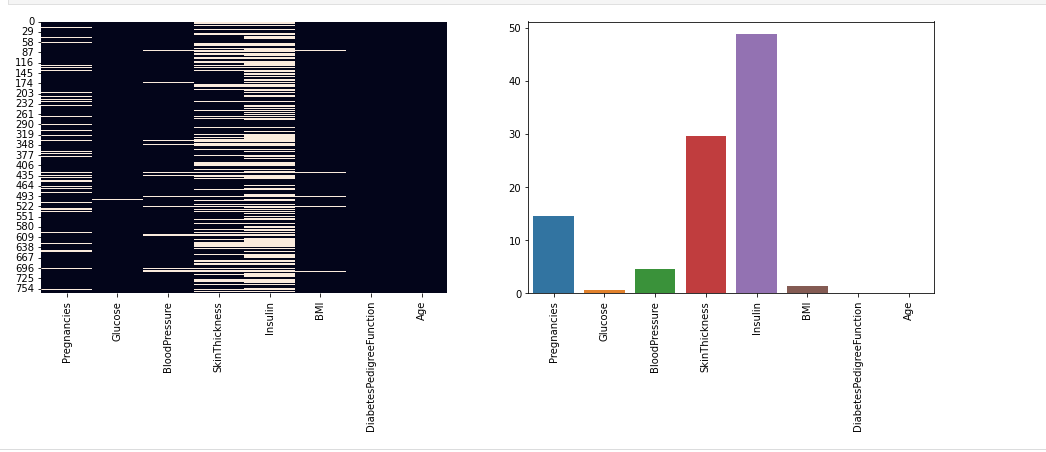
e = final classifier

Data and Methodology

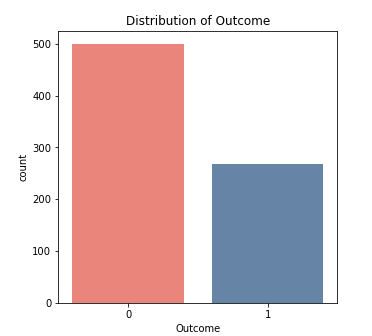
5.1 Dataset Exploration and Pre-Processing

Although there are now larger, more complex diabetes datasets, the Pima Indian Diabetes dataset has remained a benchmark for diabetes classification research. Given the presence of a binary outcome variable, the dataset naturally lends itself to supervised learning and logistic regression. However, various ML algorithms have been employed to produce classification models based on this dataset for not being limited to a singular type of model. In this research, our focus is to analyze the Pima Indian Dataset. The dataset was downloaded from Kaggle (<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>). The dataset has 9 columns and 768 rows (500 non-diabetics and 268 diabetics). The goal of this pre-processing stage is to transform already-obtained raw data into qualified data. As a result, the accuracy and error rate ratio in the data mining process will both be improved. Methods like Replace Missing Value, Normalization, and Feature Extraction are used during the preprocessing step. However, according to domain knowledge, there are inconsistent values for the attributes: Glucose Concentration (Glucose), Blood Pressure (Blood pressure), Skin Fold Thickness (Skinthickness), Insulin and BMI, whereby zero values are not within the normal range and are therefore inaccurate.

5.1.1Replace Missing Value



* Insulin ------ 50% nan values. so, we may remove insulin column.
* Age and DiabetesPedigreeFunction ------ NO nan values.



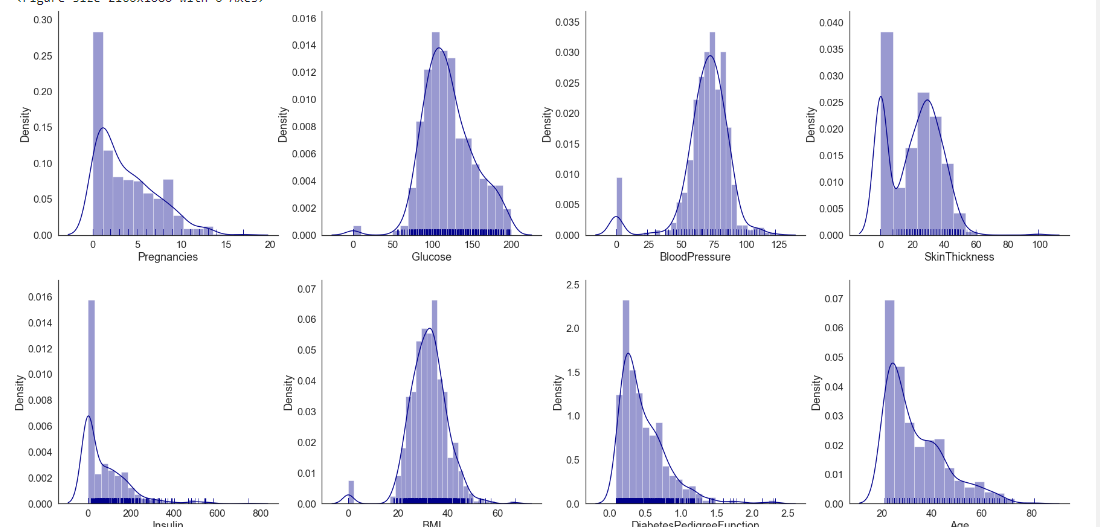
5.1.2The dataset is unbalanced:

Due to this, predictions will be biased towards non-Diabetes cases. So, we must balance this class. We will reduce the text to zero. because they prefer to say that a person is diabetic while he is not ill and not the other way around.

5.1.3 Exploratory Data Analysis:

EDA refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

5.1.3.1Data Before cleaning:



Pregnancies, Insulin, DiabetesPedigreeFunction, Skin Thickness and Age have positively or rightly skewed data distribution.

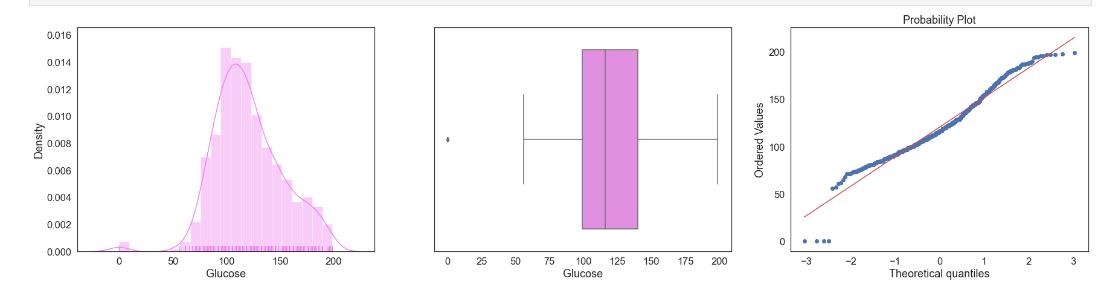
Data distributions of Glucose, BloodPressure & BMI are near a normal distribution.

5.1.3.1.1 Preprocessing

1.Pregnancies

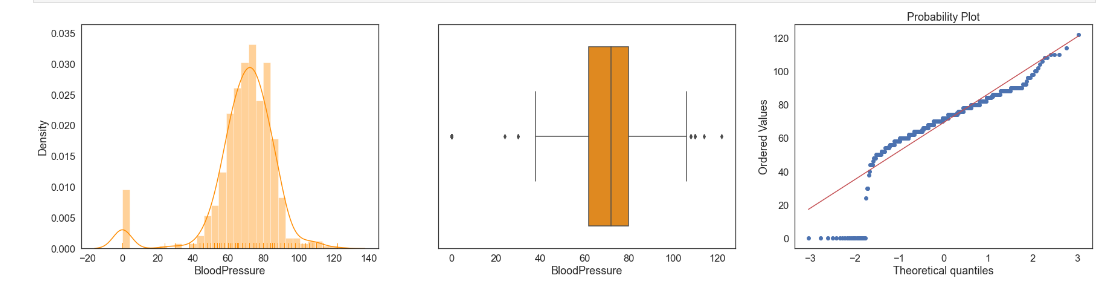
It is normal to have **zero values** and there are **outliers**.

2. Glucose



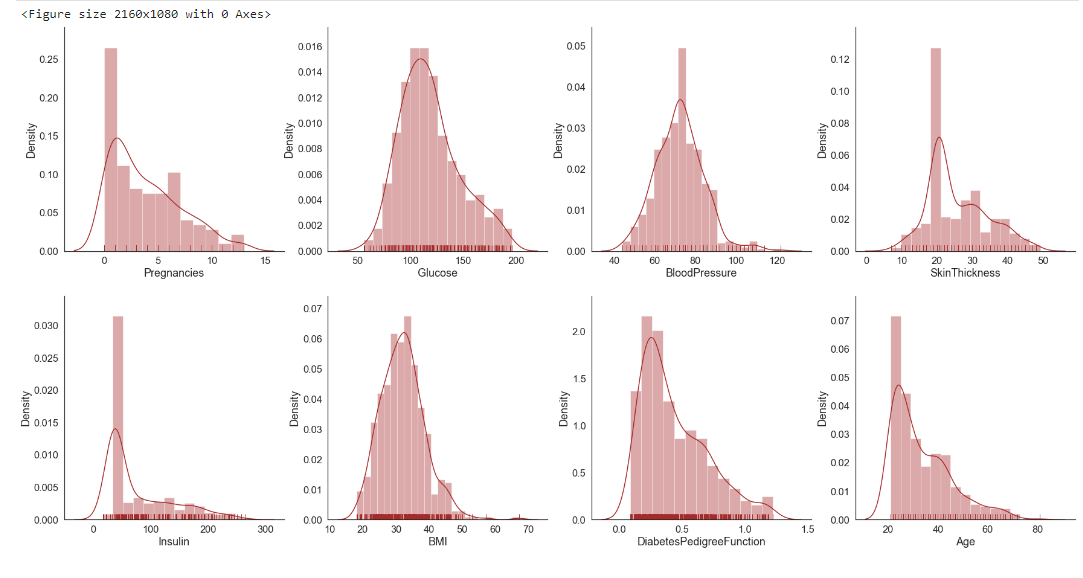
There is few outliers and also distribution is normal, So we decided to fill zeroes with mean value.

3. BloodPressure



It looks like there are few Outliers at both higher end and lower end. But at higher end maximum Blood Pressure is 122, So it is considerable. Now at lower end BP near 25 is not making sense. so we decided to replace zeroes with median and remove the outliers.

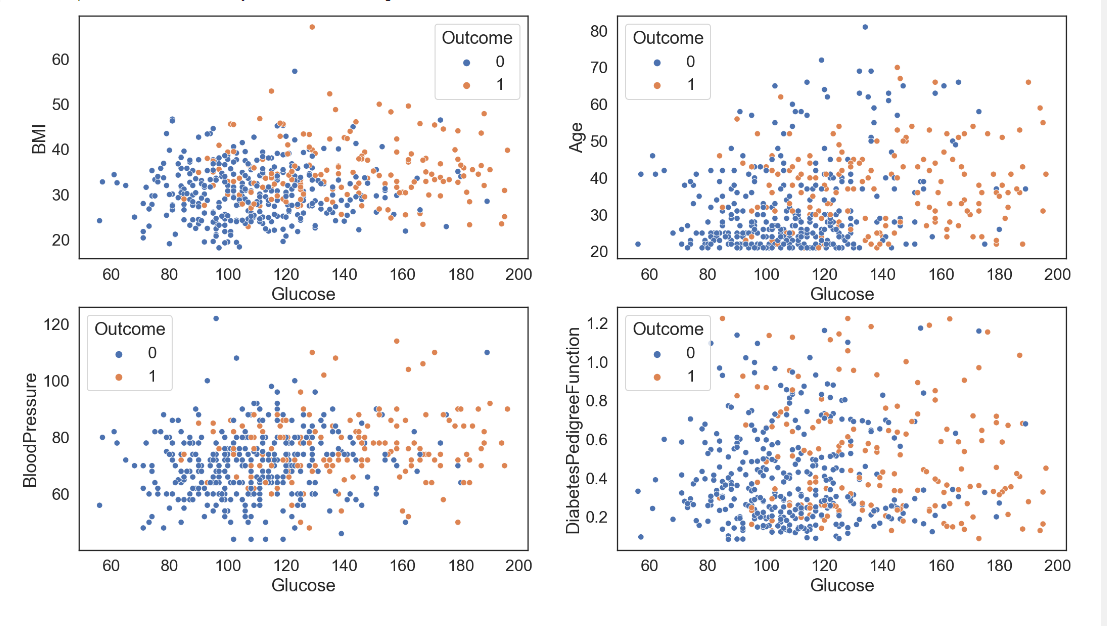
5.1.3.2Data after Cleaning:



* 1. Experiment and Results

This section demonstrates to the end-users how we choose which elements are critical in order to make our model understandable and transparent. This study runs k-means clustering, principal component analysis (PCA), and relevance ranking on the dataset for feature selection purposes inside the trials with the machine learning models. Using PCA as a starting point, the feature groups were shown in the plot of Figure 6, where arrows that are near to one another indicate closely linked features. The following can be considered as being connected:

* The correlation between Outcome and Glucose is high.
* High correlation coefficient with Pregnancies and Age.
* High correlation coefficient with Skin Thickness and BMI.



5.2.1Results of Machine Learning Algorithms

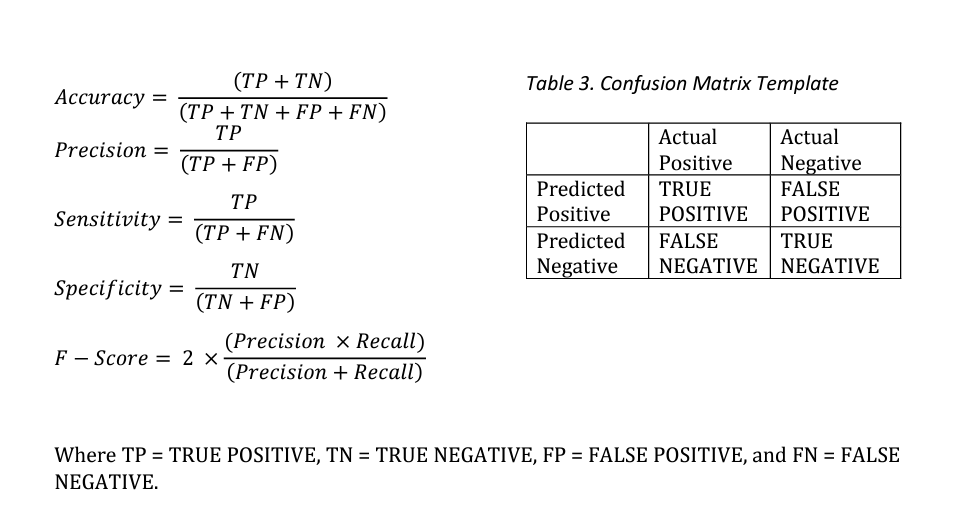
The accuracy, precision, sensitivity, specificity, F-score, and Area Under the Curve were six measures used to assess the outcomes (AUC). The confusion matrix, a matrix that displays the values of the actual result classes and the anticipated outcome classes on the testing set, is used to compute these variables (See Table 3 and formulas below).

1.The percentage of all samples for which predictions were accurate is known as the accuracy. It is the proportion of total forecasts to the sum of true positives and true negatives.

2. The percentage of samples that have been accurately predicted as true out of all samples that have been forecasted as true even when they were false is known as precision.

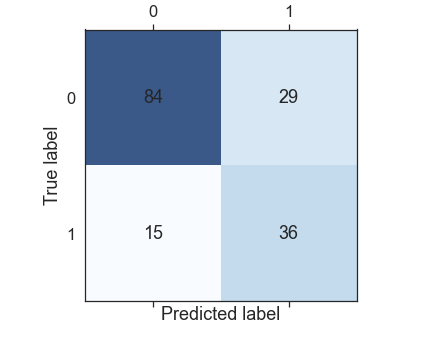
3. The weighted average of the precision and sensitivity is used to determine the standard F-Score (F1-score), which is a measure of a binary classification model's accuracy. It is determined specifically by multiplying the outcome by two and dividing the product of the precision and sensitivity by the sum of the precision and sensitivity.

4. The Receiver Operating Characteristics (ROC) Curve and the resulting Area Under the Curve (AUC) provide a vital performance measurement for classification models and represent the degree of separability of classes.

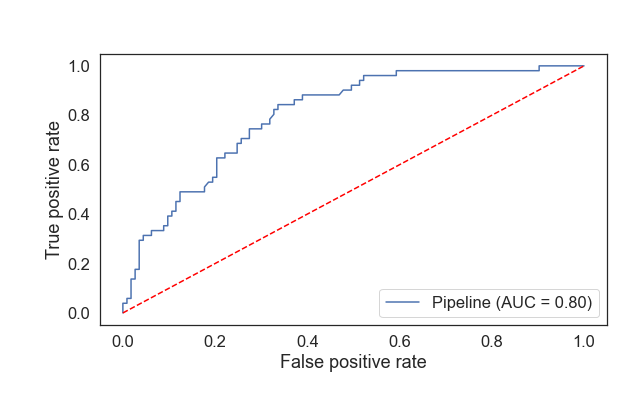


5.2.1.1 AdaBoost

AdaBoost, also called Adaptive Boosting, is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called Decision Stumps.

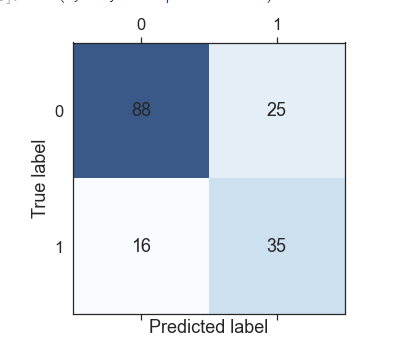


5.2.1.1 AUC-ROC Curves\_AdaBoost

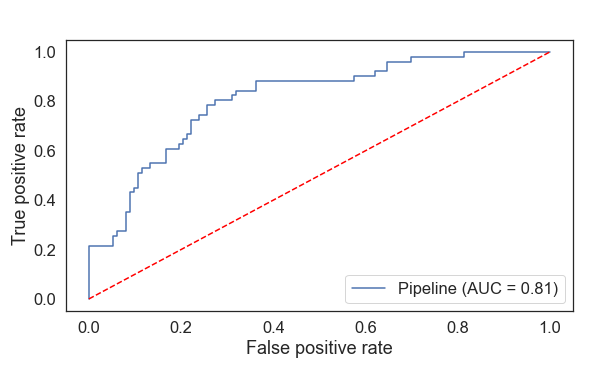


5.2.2 Random Forest

A Random Forest model refers to a tree ensemble that works similarly to a decision tree but, instead of splitting at a single attribute, forms random groups of attributes to make classifications. As a result, more processing is done, improving the accuracy of this model over a single tree model. the table show the confusion matrixes of Random Forest

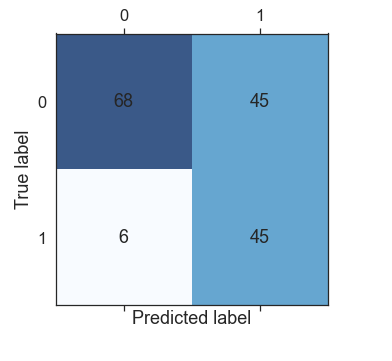


5.2.1.2.1 AUC-ROC Curves\_ Random Forest

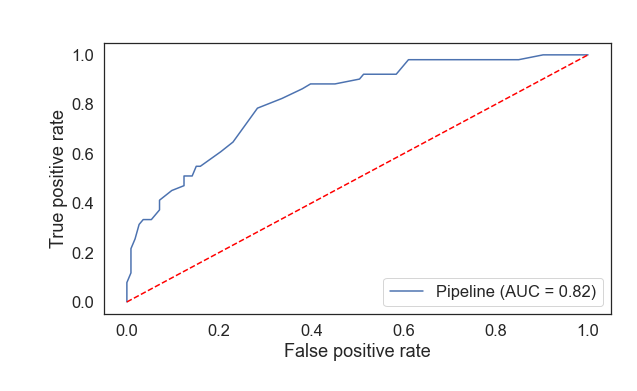


5.2.1.3. K Nearest Neighbors

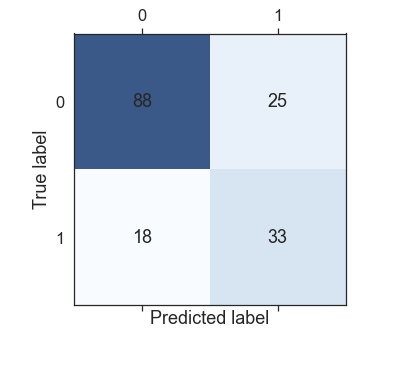
KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.



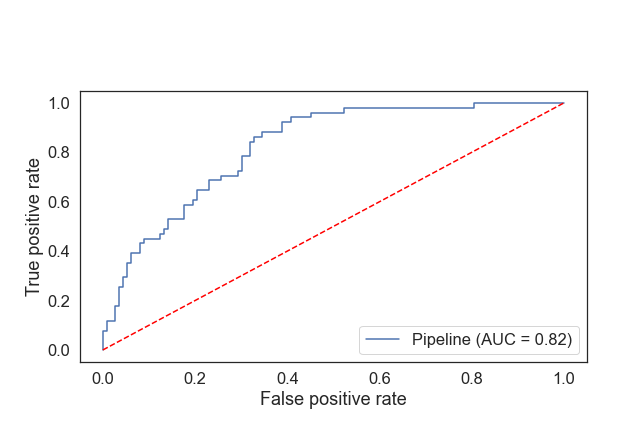
5.2.1.3.1 AUC-ROC Curves\_ K Nearest Neighbors



5.2.1.4 Logistic Regression

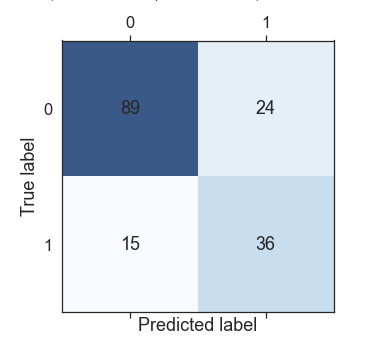
Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.

5.2.1.4.1 AUC-ROC Curves\_ Logistic Regression



5.2.1.5 Ensemble Classifier

a machine learning approach to combine multiple other models in the prediction process. These models are referred to as base estimators. Ensemble models offer a solution to overcome the technical challenges of building a single estimator.



5.2.5.1 AUC-ROC Curves\_ Ensemble Classifier



* 1. Final Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL | accuracy | precision | f1-score | AUC |
| AdaBoost Model | 0.72 | 0.54 | 0.59 | 0.79 |
| Random Forest Model | 0.76 | 0.58 | 0.66 | 0.81 |
| KNN Model | 0.70 | 0.51 | 0.64 | 0.81 |
| Logistic Regression | 0.74 | 0.56 | 0.62 | 0.82 |
| Ensemble Classifier | 0.73 | 0.55 | 0.62 | 0.82 |

Chart

Description automatically generated

Discussion and Conclusion

This paper presented classification models suitable. The models were trained using four AdaBoost, Random Forest Classifier, K Nearest Neighbors, and Logistic Regression models learning algorithms and evaluated to predict whether is positive according to eight given attributes. The experimental results on the full Pima Indian Diabetes dataset, the KNN Classifier outperformed both the AdaBoost and , Random Forest Classifier, with accuracy metric (0.70), precision (0.51), f-score (0.64) and AUC (0.81) . The use of exploratory data analysis to categorize, model, and interpret clinical data of disease patients has many facets, including the prediction of disease using classification methods among large patient populations, modeling disease trajectories through interpretation of indicators, identification of metabolic and genetic biomarkers in disease patients and associated cardio-vascular factors using chemometric approaches, the study of diabetic complications, etc. The goal of the current study is to identify key indicators among the typical clinical and anthropometric data for disease patients. The main objectives of the study can be accomplished thanks to the variable reduction offered: to accurately separate healthy volunteers from patients in the control group into various classes. quick tests to ascertain the Statistically significant patterns (clusters) of patients can be treated and cared for differently based on their individual health status. providing discriminant criteria for each recognizable, unique pattern of disease patients. to establish a statistical foundation for the customized approach to the care of individuals who have an illness and associated diseases. Intelligent data analysis has been employed in the current work to explain the varying characteristics of diabetes patients relative to the control group, which may reflect variations in the pathophysiological mechanisms underlying the disease. Additionally, other patient phenotypes have been discovered in other investigations, which may call for a different therapy strategy and set of objectives.

# **References**

<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

<https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7270283/>

<https://link.springer.com/article/10.1007/s00521-022-07049-z>

<https://ieeexplore.ieee.org/document/9917992>

<https://www.researchgate.net/publication/326303390_Influence_of_Logistic_Regression_Models_For_Prediction_and_Analysis_of_Diabetes_Risk_Factors>

<https://www.news-medical.net/health/What-is-Diabetes.aspx>

<https://pubmed.ncbi.nlm.nih.gov/8422798/>

<https://pubmed.ncbi.nlm.nih.gov/7988310/>

<https://vitalflux.com/python-train-model-logistic-regression/>

<https://www.researchgate.net/publication/359447724_Pima_Indians_diabetes_mellitus_classification_based_on_machine_learning_ML_algorithms?enrichId=rgreq-34d4d3d4c5c29c5f4f7d85976b35427d-XXX&enrichSource=Y292ZXJQYWdlOzM1OTQ0NzcyNDtBUzoxMTU2MzU3NDc0NTIxMDg4QDE2NTI3MDg1MTQyODc%3D&el=1_x_3&_esc=publicationCoverPdf>

<https://www.researchgate.net/publication/359447724_Pima_Indians_diabetes_mellitus_classification_based_on_machine_learning_ML_algorithms?enrichId=rgreq-34d4d3d4c5c29c5f4f7d85976b35427d-XXX&enrichSource=Y292ZXJQYWdlOzM1OTQ0NzcyNDtBUzoxMTU2MzU3NDc0NTIxMDg4QDE2NTI3MDg1MTQyODc%3D&el=1_x_3&_esc=publicationCoverPdf>