**AI BASED DIABETES PREDICTION**

**SYSTEM**

**TEAM MEMBER**

**712221121001 – Aanesht Raj G**

**Phase 4 Document Submission**

**Project Title:** AI Based Diabetes Prediction System

**Phase 4 : Development Part 2**

**Topic:** Start building , Selecting a machine learning algorithm , Traning the model , Evaluvating its performance .

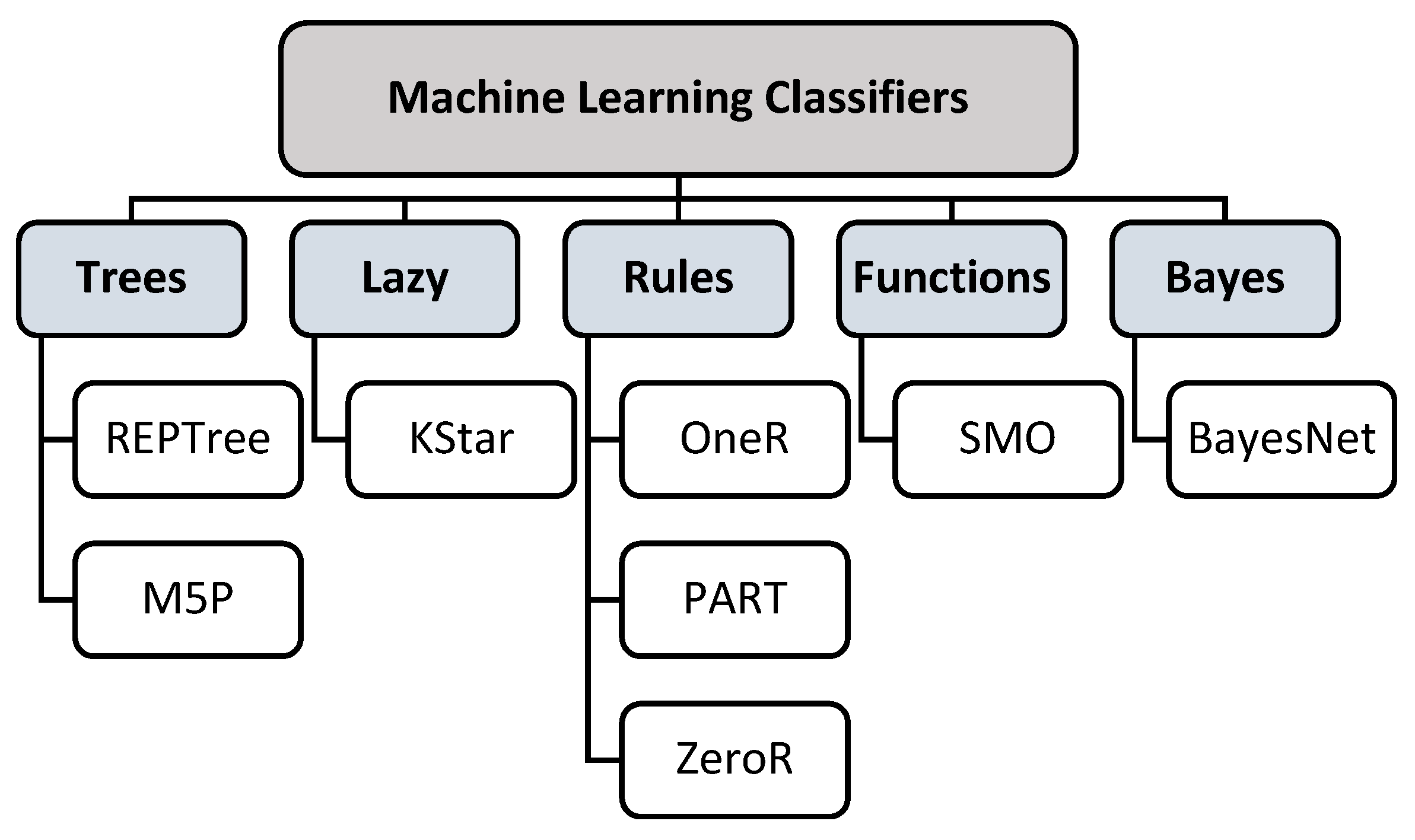
**A machine learning algorithm:**



**Factors to Choose Correct Algorithm**

1. The kind of model in use (problem)
2. Analyzing the available Data (size of training set)
3. The accuracy of the model.
4. Time taken to train the model (training time)
5. Number of parameters.
6. Number of features.
7. Linearity.

BLOCK DIAGRAM:



#### **Feature selection**

Pearson’s correlation method is a popular method to find the most relevant attributes/features. The correlation coefficient is calculated in this method, which correlates with the output and input attributes. The [coefficient value](https://www.sciencedirect.com/topics/computer-science/coefficient-value) remains in the range between −1 and 1. The value above 0.5 and below −0.5 indicates a notable correlation, and the zero value means no correlation. In Weka, the correlation filter is used to find the correlation coefficient, and the results are shown in [Table 3](https://www.sciencedirect.com/science/article/pii/S2405959521000205#tbl3). We used 0.2 as a cut-off for relevant attributes. Hence SkinThickness, BP, DPF features are removed. Glucose, BMI, Insulin, Preg, and Age are our most relevant five input attributes.

#### **Normalization**

We performed [feature scaling](https://www.sciencedirect.com/topics/computer-science/feature-scaling) by normalizing the data from 0 to 1 range, which boosted the algorithm’s calculation speed [[27]](https://www.sciencedirect.com/science/article/pii/S2405959521000205#b27). The mean and standard deviation results for all attributes after normalization are shown in [Table 4](https://www.sciencedirect.com/science/article/pii/S2405959521000205#tbl4).

In [Fig. 2](https://www.sciencedirect.com/science/article/pii/S2405959521000205#fig2), we can see that, after completing preprocessing, we have 699 samples/instances where 466 patients have no diabetes, and 233 patients have diabetes. After preprocessing, the correlation between input and output attributes is shown in [Fig. 3](https://www.sciencedirect.com/science/article/pii/S2405959521000205#fig3). In [Fig. 3](https://www.sciencedirect.com/science/article/pii/S2405959521000205#fig3), we can see that ‘Glucose’ and ‘Outcome’ have a 0.46 correlation coefficient. Hence these are highly correlated.

Table 4. Mean and standard deviation after normalization.

| **Attributes** | **Mean** | **Standard deviation** |
| --- | --- | --- |
| Preg | 0.23 | 0.20 |
| Glucose | 0.48 | 0.19 |
| Insulin | 0.50 | 0.18 |
| BMI | 0.35 | 0.17 |
| Age | 0.20 | 0.19 |

Table 3. The correlation between input and output attributes.

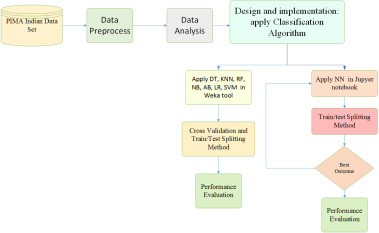
| **Attributes** | **Correlation coefficient** |
| --- | --- |
| Glucose | 0.484 |
| BMI | 0.316 |
| Insulin | 0.261 |
| Preg | 0.226 |
| Age | 0.224 |
| SkinThickness | 0.193 |
| BP | 0.183 |
| DPF | 0.178 |

### Dataset train and test method

After data cleaning and preprocessing, the dataset becomes ready to train and test. We used K-fold cross-validation and 85% train/test splitting method separately to test the different machine learning model’s performance. In the train/split method, we split the dataset randomly into the training and testing set. In the K cross-validation method, the data is divided into K folds. One-fold is used for validation/testing, and the remaining K-1 folds are used for training. The procedure will continue until every single K fold is a test set. The performance is measured by the average of all recorded scores of the Kth test.

### Design and implementation of classification model :

In this research work, comprehensive studies are done on the PIDD applying different ML [classification techniques](https://www.sciencedirect.com/topics/computer-science/classification-technique) like [DT](https://www.sciencedirect.com/topics/computer-science/decision-trees), KNN, RF, NB, AB, [LR](https://www.sciencedirect.com/topics/computer-science/logistic-regression), [SVM](https://www.sciencedirect.com/topics/computer-science/support-vector-machine), and neural network (NN). We used Kth value = 7 for the KNN algorithm.



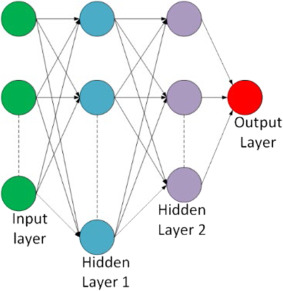
### Neural network model implementation :

We built three different neural network models with varying levels of hidden layers. We implemented the neural network with hidden layers 1, 2, and 3 with different epochs (200, 400, 800), and the results are compared. In ANN, the weighted sum of input is processed by the [activation function](https://www.sciencedirect.com/topics/computer-science/activation-function) in the hidden layer. We used two types of activation functions in our work, sigmoid and RELU. We used Keras and TensorFlow library to create the neural network models. We used a Sequential class from Keras library. The target variable is the ‘Outcome’ attribute. In ANN, the optimizer is required to reduce the output error during the [backpropagation](https://www.sciencedirect.com/topics/computer-science/backpropagation) method. We used SGD (Stochastic Gradient Descent) as an optimizer. The learning rate is a parameter in an [optimization algorithm](https://www.sciencedirect.com/topics/computer-science/mathematical-optimization) that controls the weight adjustment with respect to loss gradient. We used different learning rates to find an effective one. From the scikit-learn library, we used the train\_test\_split function to perform the train/test splitting task. We used the cross\_val\_score function from the scikit\_learn library for the K-fold cross-validation task. As the target variable is binary, the ‘StratifiedKfold’ technique is used in our work.

### Results for neural network :

In the NN with hidden layer 1, with 200 epochs, we changed the learning rate 0.1, 0.01, 0.005, shown in [Table 7](https://www.sciencedirect.com/science/article/pii/S2405959521000205#tbl7). We found that the learning rate at 0.01 provides better accuracy. Hence each case, we used learning rate = 0.01.

[Table 8](https://www.sciencedirect.com/science/article/pii/S2405959521000205#tbl8) shows the impact of epoch in a neural network with hidden layers 1, 2, and 3 at the learning rate of 0.01. We found that the [NN model](https://www.sciencedirect.com/topics/computer-science/neural-network-model) with two hidden layers with 400 epochs at a 0.01 learning rate provides the best accuracy of 88.6%. Moreover, the NN model gives more accuracy, Training accuracy, and Testing accuracy among all neural network models. The neural network of our best model NN, which included two hidden layers, is shown in [Fig. 10](https://www.sciencedirect.com/science/article/pii/S2405959521000205#fig10). The ROC curve (receiver operating characteristic curve) for 2 hidden layers with 400 epochs is shown in [Fig. 11](https://www.sciencedirect.com/science/article/pii/S2405959521000205#fig11). With K = 10-fold cross-validation, we also calculated the accuracy for two hidden layers NN model with 200 epochs. The mean accuracy obtained was 76%.



Traning The Model :

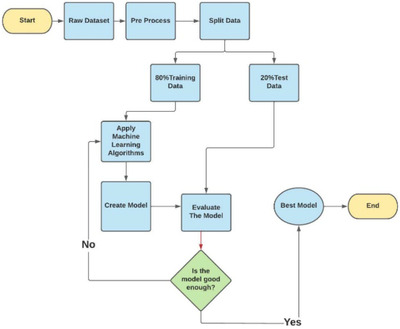
Training a model simply means learning (determining) good values for all the weights and the bias from labeled examples. In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss; this process is called empirical risk minimization.

Steps :

* Step 1: Begin with existing data. Machine learning requires us to have existing data—not the data our application will use when we run it, but data to learn from.
* Step 2: Analyze data to identify patterns.
* Step 3: Make predictions.

## **PROPOSED SYSTEM**

This section describes the working procedures and implementation of various machine learning techniques to design the proposed automatic diabetes prediction system. Figure [1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/figure/htl212039-fig-0001/) shows the different stages of this research work. First, the dataset was collected and preprocessed to remove the necessary discrepancies from the dataset, for example, replacing null instances with mean values, dealing with imbalanced class issues etc. Then the dataset was separated into the training set and test set using the holdout validation technique. Next, different classification algorithms were applied to find the best classification algorithm for this dataset. Finally, the best‐performed prediction model is deployed into the proposed website and smartphone application framework.

[](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/figure/htl212039-fig-0001/)

[Evaluvating Its Performance :](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/figure/htl212039-fig-0001/" \t "figure)

 Performance Evaluation is defined as a formal and productive procedure to measure an employee's work and results based on their job responsibilities.

INTRODUCTION

Globally, many chronic diseases are prevalent in developing and developed countries. Diabetes mellitus (DM) often known as diabetes, is one of the metabolic disorders which may cause blood sugar, by producing more significant or a smaller amount of insulin. Diabetes affects the different parts of the human body parts like eyes, kidneys, heart, and nerves [18]. Hence, the prevention and detection of disease in the early stages could likely save human lives. Most common are Type-I diabetes and Type-II diabetes. In Type I desease, insulin will not be produced by the human body and 10% of diabetes caused by this type. Type II diabetes is non-insulin-dependent diabetes caused because of not enough production of insulin by the pancreas or body cells are resistant to absorb it. Therefore, there is a need to prevent and diagnose the disease to save human life from an early death.Researchers have proved that data mining and machine learning models Decision Tree, Support Vector Machine etc .

**METHODOLOGY :**

The Exploratory Data Analysis (EDA) is conducted to understand the various data sources and collected the contextualized Pima Indians Diabetes dataset from the UCI repository as step-1. Cleaned the dataset, that is, removal of the duplicate and missing values as a step of data preprocessing. Then, identified a set of candidate classification models for the experiment on the training dataset to fit a model. Assessed the models based on various performance metrics such as Accuracy, Recall, f1-Score, Misclassification Rate and ROC- AUC-Score. Compared the models on the various performance metrics to analyze the accuracy of a model for the dataset. Represented the performance evaluation of the candidate classification models based on the metrics graphically. The proposed model diagram is presented in the below figure-1. This framework presents the sequence and flow of the experimental process followed in this work.

**CONCLUSION :**

Diabetes detection and prediction are one of the prevalent medical problems in real-world. The persistence of it in the human body for a long time leads to the microvascular complications of diabetes. In this paper, a systematic experimental study was conducted using various Machine Learning classifiers to predict the likeliness of diabetes in a human being. These models that were used to fit the training set were compared and estimated their performance in terms of precision, recall , accuracy and ROC-AUC-Score. The results show that Logistic Regression (LR) classifier performed better with a with highest accuracy of 77.6 %, f1-score of 75 , ROC-AUC score of 73.6 and with least Misclassification rate as 23.8 in comparison to other algorithms. This work can be extended to improve the prediction accuracy using ensemble machine learning techniques.