Kurdistan Journal of Applied Research (KJAR)
Print-ISSN: 2411-7684 | Electronic-ISSN: 2411-7706

Website: Kjar.spu.edu.iq | Email: kjar@spu.edu.iq

Predict Diabetes Using Voting Classifier and Hyper Tuning Technique

Chra Ali Kamal
Database Department
Computer Science Institute
Sulaimani Polytechnic University
Sulaimani,Iraq
chra.kamal@spu.edu.iq

Manal Ali Atiyah
Database Department
Computer Science Institute
Sulaimani Polytechnic University
Sulaimani,Iraq
manal.ali@spu.edu.iq

Article Info

Volume 7 - Issue 2 - December 2022

DOI: 10.24017/Science.2022.2.10

Article history:

Received: 21/11/2022
Accepted: 04/01/2023

Keywords:

Diabetic Prediction, Graphical User Interface, Hyper Tuning,Machine Learning Algorithm,Voting Classifier.

ABSTRACT

Diabetes is one of the most common chronic diseases in the world due to the sedentary lifestyle which led to many health issues as heart attack, kidney frailer, and blindness. Additionally, the majority of people are not aware of the early-stage diabetes symptoms. The above reasons encourage developing a diabetes prediction system using machine learning techniques. The Pima Indian Diabetes Dataset (PIDD) was utilized for this framework as it is common and appropriate dataset in CSV format. In addition, the proposed framework is divided into two phases of model selection. In the first phase, six different algorithms namely, Logistic Regression, Decision Tree, Random Forest, K-nearest neighbor, Support Vector Machine and Naïve Bayes are applied, then two different hyper parameter techniques namely, Randomized Search and TPOT (autoML) are used to increase the accuracy level for each algorithm. In the second phase, the four best performed algorithms with best estimated parameters are chosen and used as input for the voting classifier. The results show that the Random Forest is the best performed performance algorithm to predict diabetes via a simple graphic user interface with 98.69% accuracy level.

1. INTRODUCTION

Blood is one of the most vital place which includes many complex processes relevant to human's food, e.g., transferring carbohydrates to energy. When a piece of carbohydrate is eaten, the blood converts it to glucose (i.e., sugar) and carries the glucoses as a source of energy for all body cells. One of the major types of sugar is blood glucose; it is the main source of energy at the same time [1]. Moreover, insulin is a hormone that is produced by pancreas allowing sugar to enter into the cells and make organs work properly (see Figure 1). In addition, there is a health condition called a diabetes mellitus which occurs when pancreas does not make enough insulin and/or cells do not adjust properly to insulin and this results in having too much glucose in blood [2].

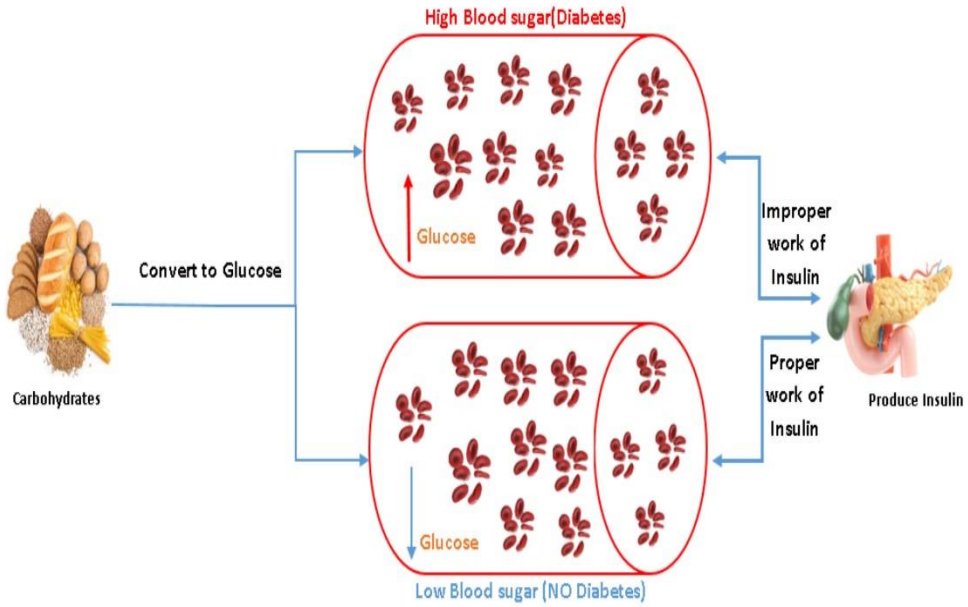

Figure 1: Work of Pancreas.

Artificial Intelligence (henceforth AI) is a common field in computer science; it is a rule-based and operated by complicated statistical algorithms [3]. One of the vital part of AI is machine learning (henceforth ML) and any computer systems and/or applications are developed using ML that have capability to automatically learn from available data or experiences in the absence of writing code for each case [4]. There are many ML types but two major kinds are supervised and unsupervised learning and both categories utilize available datasets in order to improve data understanding and realize valuable knowledge [5]. ML algorithms are considerably applying in many scientific fields and have role in uprising businesses in the world. For example, for diabetes early prediction, risk management and diabetic diagnosis, ML is used. In other words, in the past ten years, ML based system is developed to works on managing diabetes as a chronic diseases and providing clinical decision support [6].

There is no doubt that people's lifestyle has been changed due to industrialization and globalization and as a result they prefer to consume sugary food, choose a sedentary work and ends with increasing diabetes [7]. Diabetes is multifunctional disease which causes many other health issues in human's body. According to International diabetes federation (IDF), the number of adults aged between 20-79 years having life with diabetes and it is predicted to be 783 million in 2045 [8]. Hence, a ML based system is essential to identify risk factors, early diabetes detection and reduce number of cases or death. Recently, diabetes is classified as the largest public health challenges all over the world because it is a long-lasting disease with no

indefinite recovery. Furthermore, diabetes is one of the top ten sequences of deadly diseases [9]. According to the World Health Organization (WHO), in 2019, diabetes is a dead cause for nearly 1.5 million people and 48% of them are not 70 years yet [10]. Thus, early diabetes diagnosis is important since increasing the glucose level leads to many fatal diseases such as kidney failure, heart attack, and nerve damage [11]. Regular diabetes detection is also required for those who have diabetes with no noticeable symptoms and this is usually harmless [12].

The aim of the proposed framework is to develop an accurate system to predict diabetes using machine learning algorithms. Further, to achieve this aim, the following steps have been taken into consideration.

Firstly, the system is fed with Pima Indian Diabetes Dataset (PIDD) which is the common diabetes dataset with 768 records. Secondly, some data are performed cleaning and preprocessing processes like data standardization and removing (i.e., null, duplicate, zero) values. Thirdly, seven classification algorithms (Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Support Vector Machine, Naïve Bayes and Voting Classifier) with two hyper parameter techniques (Randomized Search and Tree based Pipeline Optimization Tool (TPOT)) in two different stages are applied. Fourthly, using the best four performed algorithms as an input for averaging classifier. Finally, the best performed algorithm (Random Forest) is chosen to fit the selection model and results are predicted through a simple graphical interface.

2. LITERATURE REVIEW

In order to shed light on diabetes relevant frameworks, reviewing the most up to date diabetes based prediction frameworks which use ML techniques is required. Park et al. [13] proposed a system to predict the best model of insulin resistance for people over 40 years in South Korea. The dataset was taken from two different cities, namely Ansan and Ansung. Seven ML algorithms (e.g., Linear Regression, Support Vector Machine, XGBoost, Decision Tree, Random Forest, K-Nearest Neighbors and Artificial Neural Network) were applied to 8842 records; each of them had 1411 various features. Additionally, AUC and ROC, Accuracy and k-fold were used to measure the skills of ML models. After selecting optimal features from 99 to 15 and from 15 to 9 features, they realized XGBoost with AUC = 0.86, Random Forest with AUC = 0.84 and Artificial Neural Network with AUC = 0.86 were the best performed algorithms to find the risk factor and predict the disease.

A multi-classification problem is generated with five various ML algorithms (e.g., Multinomial logistic Regression, Naive Bayes, Decision Tree, Random Forest and Stochastic Gradient Boosting) by Rajput et al. [14]. in which its dataset is extracted from Mendeley Diabetes types in Iraq. Due to the fact that selecting strong neighborhood features in their local datasets produce an acceptable ML model performance [15] the framework depends on the human's medical features especially excessive Body Mass Index (BMI) and age. Moreover, it discovered reasons that make patients suffering from no-diabetes to pre-diabetes and from pre-diabetes to diabetes. Many evaluation matrixes were used in the system, however the F1-score was the best matrix for analyzing multi class classifier and both Decision Tree (0.8967) and Stochastic Gradient Boosting (0.89) gained preferable accuracy level. Mushtaq et al. [16] proposed a prediction framework with PIMA diabetes dataset. In creating optimal model, they used six ML algorithms, namely Random Forest, Logistic Regression, Support Vector Machine, k-Nearest Neighbors, Naïve Bayes Theorem and Gradient Boosting with an ensemble learning technique called voting classifier. The dataset was slightly imbalanced, therefore both undersampling (TomekLinks) and oversampling (SMOTE) were used. The three best performed algorithms (e.g., Naïve Bayes Theorem, Gradient Boosting and Random Forest) were utilized in voting classifier and the accuracy results of both balanced and imbalanced data were taken.

Examining the data analytics and ML based framework were the main goal of Krishnamoorthi et al.'s study [17]. Four ML algorithms (e.g., Logistic Regression, Random Forest, Support Vector Machine and k-Nearest Neighbors) were selected to develop an intelligent diabetes

mellitus prediction framework (IDMPF). After preprocessing of PIMA dataset, Grid Search and Random Search as two hyper parameter techniques with cross validation were applied to find out the ideal parameters. As a result, the ROC value for Logistic Regression was 86%. Several researchers including R et al. [18] have generated a heterogeneous model to predict a diabetes called stacked ensemble model. Furthermore, several ML algorithms were used including stacking methods and achieved the highest accuracy score, i.e., 93%. The framework focused on the advantages of using multiple ML techniques rather than the single one in gaining the best accuracy levels.

According to Ahamed et al. [19], Light Gradient Boost Machine (LGBM) was the best performed algorithm with accuracy of 95.2 % comparing to other the ML algorithms while developing a diabetes prediction framework. Additionally, researchers concentrated on using one evaluation matrix only and choosing diabetes mellitus on type 2 for prediction. Designing two applications was the main concern of Bano et al's research [20]. All required steps for diabetes prediction framework were implemented through cloud and user application. In user application interface, dataset was uploaded and send to cloud based application. The cloud application interface contained several buttons for applying five ML algorithms (e.g., Support Vector Machine, Artificial Neural Network, Decision Tree, Logistic Regression and Father First) and demonstrated accuracy results by graph.

Jian et al. [5] generated a framework to classify eight different problems caused by diabetes. Different record numbers were used for each complication after intensive preprocessing and feature selection process. Six various ML algorithms used in evaluating traditional and ensemble learning ML model. Lliha and Rista [21], on the other hand, developed a decision-making framework to detect diabetes depending on analyzing data. A combination of data mining methods and ML algorithms were utilized in the system. Decision Tree was the best preferable algorithm with the accuracy ratio of 79%. Moreover, Jaggi et al. [22] proposed a framework for diabetes prediction through applying one algorithm only, i.e., Artificial Neural Network with six layers. The whole experiment consisted of feed-forward layers. The framework implementation was done by a web portal; a patient was able to enter all the medical information. The accuracy level of the ANN algorithm was 77%.

In terms of comparing this framework to the past related works and particularly aforementioned researches. It can be seen that any prediction model-based frameworks follow the similar steps as thinking of correct dataset for the prediction and utilizing ML algorithms to train the data and then creating a proper model [23]. However, the differences appear with types chosen, number, and accuracy level of each algorithms. Due to use the same dataset (PIMA) and choose ensemble learning algorithm (i.e., voting classifier), Mushtaq et al. [16] was the closest study to the present framework. Meanwhile, many important details confirmed the distinction between their study and the current one. Firstly, in the number and type of algorithms, they used five classification algorithms with voting classifier, while the present proposed framework used seven algorithms including voting classifier. Secondly, in the cleaning and preprocessing data, this framework used both replace method of Python Panda (i.e., remove zeros) and the scatter plot diagram (i.e., eliminate outliers); Mushtaq et al. [16] utilized both undersampling (i.e, TomekLinks) and oversampling (i.e, SMOTE) techniques to remove outliers. Thirdly, they used Naïve Bayes Theorem, Gradient Boosting and Random Forest algorithms as an input for voting classifier and the final accuracy level was 81.5% for voting classifier, while, the present framework used Logistic Regression, Decision Tree, Random Forest, Support Vector Machine as input for voting classifier after performing two hyper parameter tuning techniques, e.g., Randomized Search and TPOT(autoML), and the terminal accuracy level was 98.69% for Random Forest. Finally, in contrast to this framework, which demonstrated the prediction results via an interface using PyCharm and Djungo, Mushtaq et al. [16] did not take this dimension into consideration.

Table 1 demonstrates the summary of all papers that reviewed above based on the dataset, names of the algorithms were used and the best performed ones among them.

Table 1: Summary of reviewed papers based on algorithms and accuracy

Researcher Names	Published Year	Algorithm(s)	Dataset	Best Accuracy
Park et al.	2022	RF, KNN, ANN, LR, SVM,XGBOSST,DT	Ansan and Ansung cities in Korea	XGBOSST & ANN 86%
Rajput and Khedgikar	2022	Multinomial LR, NB, DT, RF, Stochastic GB.	Mendeley Diabetes types in Iraq	DT & GB 89%
Mushtaq et al.	2022	LR,RF,SVM,KNN, NB,GB, Voting classifier.	Pima Indian Diabetes Dataset (PIDD)	Voting Classifier 81.5%
Krishnamoothi et al.	2022	LR,RF,SVM,KNN	PIDD	LR 86%
R et al.	2022	RF,KNN,LR,GB, AdaBoost ,SVM, Stacking	PIDD	Stacking 93%
Ahamed et al.	2022	LR,XGB, GB,DT, Extra DT, RF, LGBM	PIDD	LGBM 95.20%
Bano et al.	2021	SVM,ANN,DT,LR, Father First	PIDD	Father First 99.9%
Jian et al.	2021	LR, SVM, DT Cart, RF, AdaBoost, XGB	Rashid Center for Diabetes and Research(RCDR) in Ajman (UAE).	XGB 97.8%
Llaha and Rista	2021	NB, DT, SVM,LR	Public Health Institute.	DT 79%
Jaggi et al.	2021	ANN	PIDD	ANN 77%

3. METHODOLOGY

The current section is basically divided into two parts. In the first part (see Figure 2.a), six classification algorithms (e.g., Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Support Vector Machine and Naïve Bayes) were selected, however, to boost the accuracy level for algorithms, two hyper parameter tuning techniques, e.g., Randomized Search and TPOT(autoML), were used. Each algorithm has various type of parameters with different values; to discover the optimal parameter for each, both techniques were repeated several times and then aforementioned algorithms were applied.

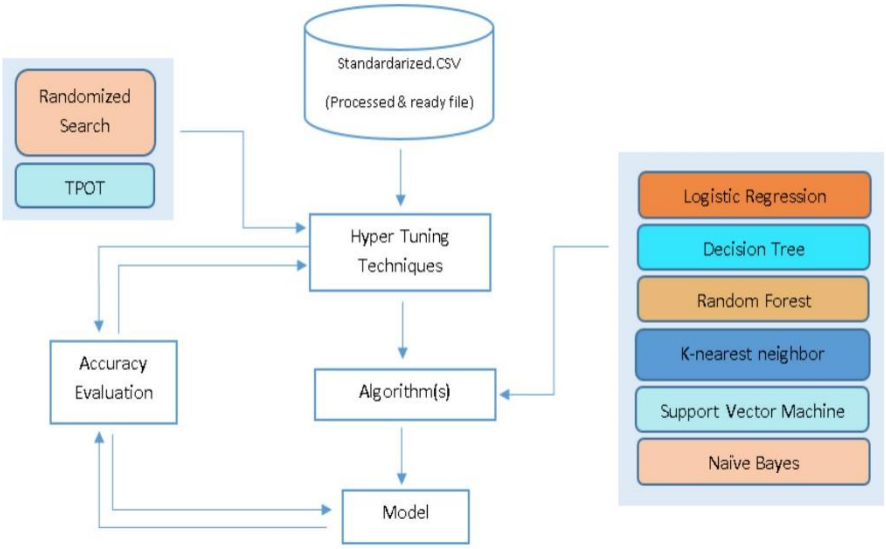

Figure 2.a: The first phase of the system.

In the second part (see Figure 2.b), the best four performed algorithms with best estimated parameters were used as input for the voting classifier, since it applies to discover the preferable algorithm between a group of multiple choices.

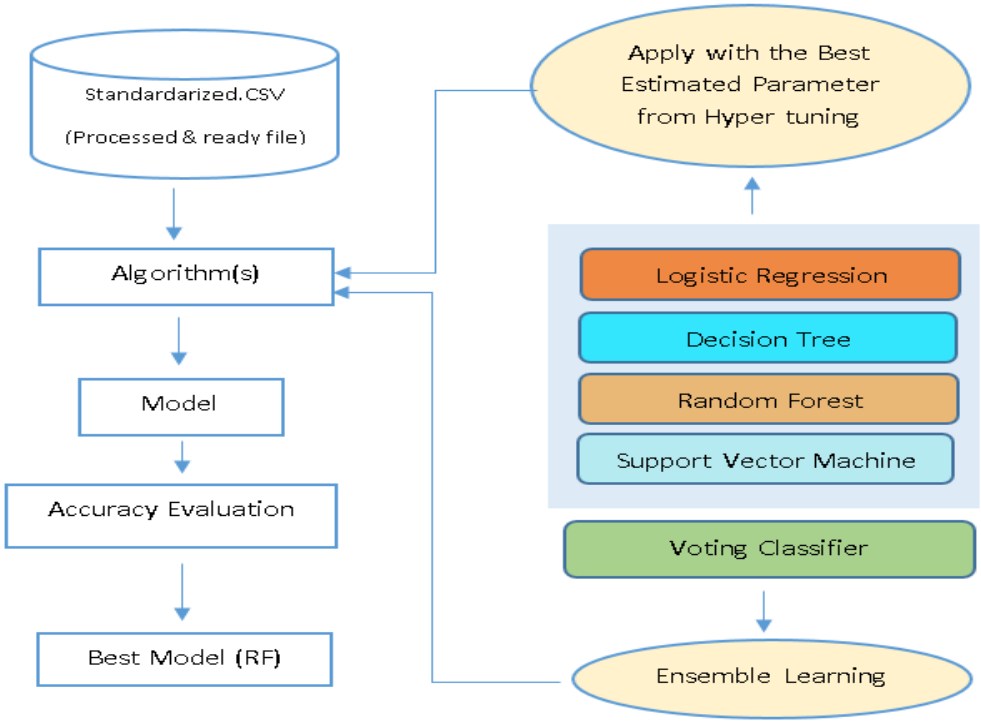

Figure 2.b: The second phase of the system.

3.1 Dataset

Data from Pima Indian Diabetes Dataset (PIDD) was utilized in CSV format and it is from national institute of diabetes and digestive and kidney diseases [24]. The dataset consisted of 768 medical records with one dependent (i.e., target) and eight independent variables. The target variable predicts when a patient has Positive or Negative diabetes. The positive target numbers were 268 and negative numbers were 500, as it is presented in Table 2.

Table 2: The PIDD data description.

Dataset name	Number of records	Number of independent variables	Number of dependent(target) variables	Positive target numbers	Negative target numbers
PIDD	768	8	1	268	500

In addition, Figure 3 demonstrated the histogram of the dataset every attributes, namely, pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index (BMI), diabetes pedigree function, age and outcome. The histogram, further, presented the distribution rate for each medical variable; for example, in age attribute, it showed the highest contribution number for those who are between 20 and 40 years. However, people between 60 and 80 have the lowest participating rate in the PIMA dataset.

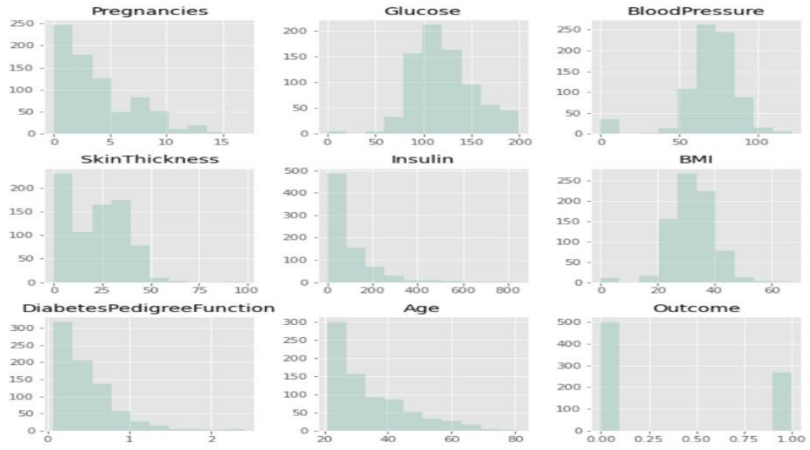

Figure 3: The histogram of attributes.

3.2 Data Cleaning and Preprocessing

After finishing the data exploration of the PIDD dataset, data cleaning and preprocessing were essential. Data cleaning consisted of removing duplicate, null, zero and outlier values. When there were not any duplicate and/or null values in the dataset, some zero values and/or outlier records were removed. The replace method of the Python Pandas data frame was used to relocate all the zero values with the average of their specific column. Table 3 illustrates zero numbers in each column.

Table 3: Zero values in each column.

Column Name	Zero Values
Glucose	5
Blood Pressure	35
Skin Thickness	227
Insulin	374
BMI	11

Furthermore, the scatter plot diagram was used to identify outliers (see Figure 4) and four outlier records were detected. One record in skin thickness (< 80) and three records in insulin (<= 600) were eliminated.

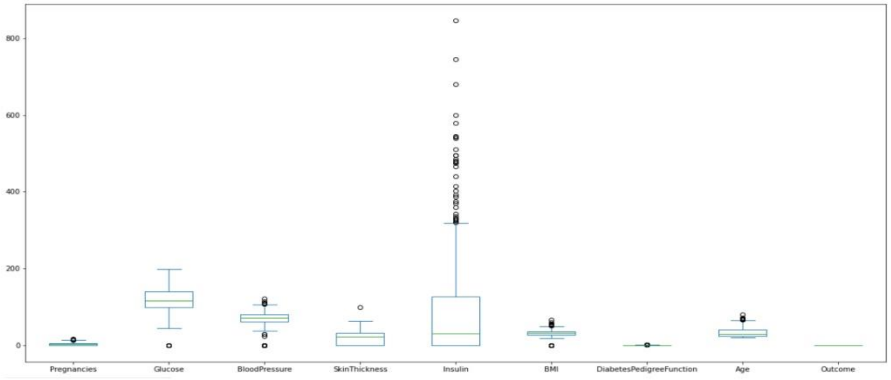

Figure 4: Identification of outliers.

As for preprocessing, data standardization strategy was performed which made all the record values in the same specific range between 0 and 1. Hence, the dataset was understandable, easy to use, and all in the same format. Table 4 and 5 demonstrates the PIDD data set before and after standardization technique.

Table 4: Datasets before standardization process.

S.	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	**6**	148.0	72.0	35.000000	77.246073	33.6	0.627	50	1
1	**1**	85.0	66.0	29.000000	77.246073	26.6	0.351	31	0
2	**8**	183.0	64.0	20.390052	77.246073	23.3	0.672	31	1
3	**1**	89.0	66.0	23.000000	94.000000	28.1	0.167	21	0
4	**0**	137.0	40.0	35.000000	168.000000	43.1	2.288	33	1

Table 5: Datasets after standardization process.

S.	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	0.352941	0.743719	0.590164	0.555556	0.128743	0.500745	0.259091	0.617284	1.0
1	0.058824	0.427136	0.540984	0.460317	0.128743	0.396423	0.142041	0.382716	0.0
2	0.470588	0.919598	0.524590	0.323652	0.128743	0.347243	0.277686	0.395062	1.0
3	0.058824	0.447236	0.540984	0.365079	0.156667	0.418778	0.069008	0.259259	0.0
4	0.000000	0.688442	0.327869	0.555556	0.280000	0.642325	0.945455	0.407407	1.0

3.3 Feature Engineering

In order to discover useful features in available data which help in learning target variable and creating successful ML models, it is necessary to present data in an effective manner. To achieve this, Correlation Matrix as a numerical measure was used to find the relation between all features. The corr method of the Python Seaborn was utilized to visualize the PIDD variables via a heatmap. Figure 5 shows a heatmap for the dataset; all darker color squares had low or negative correlation and by contrast all lighter color ones indicated high or positive correlation coefficient. For example, the relation between (pregnancy, age) and (skin thickness, BIM) were high, whereas correlation between insulin and blood pressure was low.

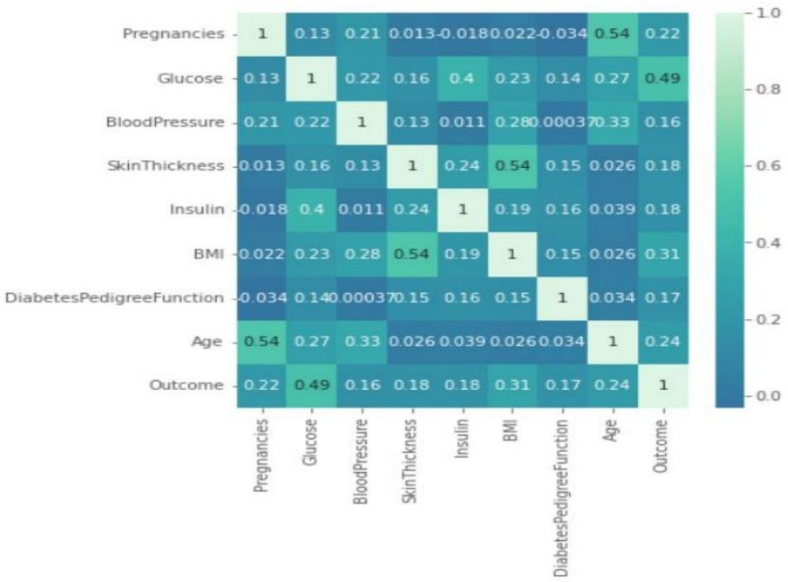

Figure 5: The PIDD correlation matrix.

3.4 Hyper Parameter Tuning

Generally speaking, one or more ML algorithms can be used in creating any prediction framework. Every algorithm has one or more parameters with its own value. In diabetes prediction, for choosing an algorithm, it is necessary to think about list of parameters since they are directly related to boost prediction accuracy level. Hyper parameter tuning is the process of discovering the optimal parameter type with it is value through applying some methods like grid search, random search, genetic algorithm, tree based pipeline optimization tool (TPOT) and many more. In this framework, Randomized Search CV and TPOT strategies were applied.

3.4.1 Randomized Search CV

The best hyper parameter optimization technique that utilizes random combination of optimal features to discover the best way for creating a model is referred to Randomized Search cv. It was applied since it was time and cost effective narrowed down the results with less time and computing power consuming [25]. In order to find the region that the diabetes prediction model performed well, the predefined set of parameter combination for each algorithm were tested depending on the number of iterations and cross validation folds. As a result, the best estimated features were captured.

3.4.2 Tree based Pipeline Optimization Tool(TPOT)Classifier

Automated ML pipeline is a common process that consists of data preprocessing, feature exploration and selection, ML algorithm optimization and hyper parameter tuning. Additionally, AutoML concentrates on automating all the previously mentioned ML processes [26]. Tree based Pipeline Optimization Tool (TPOT) is a type of the AutoML methods and

Python library simultaneously and intellectually discover the most suitable ML pipeline for the existing dataset by using genetic programming. In this framework, the predefined best parameters from Randomized Search CV were applied as input for TPOT technique to perform the hyper parameter tuning and improve the accuracy level. Table 6 demonstrates best hyper parameters for each algorithm.

Table 6: Best parameters for each algorithm.

Algorithms	Best Hyper Parameters
LR	'random_state':800, 'multi_class': 'multinomial', 'max_iter': 120, 'intercept_scaling': 9, 'fit_intercept': True
DT	'random_state':894, 'min_samples_split': 2, 'max_leaf_nodes': 8, 'max_features':'log2', 'max_depth':70, 'criterion': 'entropy'
RF	'n_estimators':1800, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features':'log2', 'max_depth':340, 'criterion': 'gini'
KNN	'weights': 'distance', 'p':2, 'n_nieghbours': 10, 'n_jobs':6, 'algorithm': 'auto'
SVM	'kernel': 'poly', 'gamma': 'auto', 'degree': 3, 'c': 52
NB	'var_smoothing': 1e-09, 'priors': None

3.5 Classification Algorithms

In the framework, six different classification algorithms (e.g., Logistic regression, decision tree, random forest, k-nearest neighbors, support vector machine, naïve bayes) were applied in the first phase. Whereas, in the second phase, the best four performed algorithms were used as an input for voting classifier.

3.5.1 Logistic Regression (LR)

Logistic Regression is a statistical and probability based supervised learning algorithm. LR can apply for those kind of dataset which its target variable is categorical and it is best align with the PIDD dataset chosen for this study. The PIDD outcome variable was binary and represent by 1 for diabetes and 0 for non-diabetes. Due to its effectiveness, this algorithm was utilized in many prediction diseases such as diabetes, cancer, as so on. [16-20].

3.5.2 Decision Tree (DT)

Decision Tree gives researchers a graphical representation for solving classification and regression problems; nodes and leaf nodes refer to dataset variables and outcomes respectively. The DT results can be easily comprehended by humans since they were built based on if-else pattern sets. Therefore, it is appropriate to apply in prediction disease with a high accuracy result [18-21-25].

3.5.3 Random Forest(RF)

Random Forest is also a supervised learning algorithm which generates from several decision trees coping with data overfitting. RF takes mean and/or average from all DTs to produce outcomes. Increasing DT numbers means achieving high accuracy in prediction. Furthermore, the n_estimators is one of the random forest hyper parameter which indicates the number of DTs. Thus, the number of n_estimators was increased to 1800 for this framework so as to gain a better diabetes prediction [14-17].

3.5.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a distance-based supervised learning algorithm; the working strategy of KNN is computing nearly the closest distance between unknown and available data points. The letter 'K' represents the number of the nearest existing values around new ones in a given dataset. This classifier is simple, fast and can be utilized in solving classification and regression problems [17].

3.5.5 Support Vector Machine (SVM)

Support Vector Machine is simple and general classification algorithm which uses one or more lines to categorized data points. In addition, the hyperplane is for lines that their

dimension number are being determined according to the features of the existing dataset. This algorithm classifies the data points by measuring the distance between hyperplanes [27].

3.5.6 Naive Bayes (NB)

Naive Bayesis a numerical classifier used to solve classification problem either binary or multi class classification. This classifier works based on Bayer's Theorem theory and it is a probabilistic formula. The NB's main role is to recognize dependent variable from various independent features which matches the PIDD data sets and predicts diabetes from different medical records [4].

3.5.7 Voting Classifier

Ensemble learning is one of the ML techniques which works based on algorithm diversification concept. This means that it uses a group of various ML algorithms (i.e., base models) as input to produce a high prediction accuracy output Although there are different types of ensemble learning techniques, the simplest and basic method is voting classifier. In the present framework, averaging as a powerful kind of voting classifier was chosen. Four of the well performed algorithms act as a parameter list for the averaging classifier. Finally, the algorithm took the mean of predictions from every model and utilized it to produce the ultimate prediction. For example, in this framework, the approximate accuracy for Logistic Regression, Decision Tree, Random Forest and Support Vector Machine were 79, 81,98,83 respectively and the final prediction accuracy model for averaging classifier was 85.

3.6 Performance Evaluation

Measuring model performance is vital to realize whether the prediction framework is working well; repeating some important steps as preprocessing or feature selection in case the framework has no preferable performance level. There are several mechanisms to evaluate model performance namely precision, accuracy, recall, ROC, AUC, sensitivity and specificity. Accuracy as a confusion matrix based method was applied in this framework. Separating the accurate and inaccurate prediction rate was the main role for the accuracy matrix. In other words, depending on the following formula [28], the matrix demonstrates how the diabetes prediction framework detect that a person has diabetes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Where, [29]:

- True Positive (TP): The number of predicted samples that have diabetes correctly.
- False Positive (FP): The number of predicted samples that have diabetes incorrectly.
- True Negative(TN): The number of predicted samples that have no diabetes correctly.
- False Negative (FN): The number of predicted samples that have no diabetes incorrectly.

3.7 Framework Tools

The programming language, tools, software packages used in implementing this framework along with the reasons of choosing each of them are described below.

3.7.1 Python

Python is the programming language with many useful libraries in comparison to other programming languages and available at [30]. The features make python able to decrease the code to one-third for developers and programmers especially in ML area [31].

3.7.2 Anaconda

Choosing Python as a language to create a ML model and predict diabetes needs an interpreter. Therefore, downloading Anaconda available at [32] as a free and an open-source software was necessary since is allows programmer to use Python and all libraries belongs to it.

3.7.3 Jupyter Notebook

Jupyter Notebook is a browser-based software under the Anaconda distribution available at [33]. It is main goal is to give programmers a permission to execute code lines partially in both Python and R programming languages. In addition, it helps visualizing dataset features easily [34]. In this framework, several important Python libraries such as Pandas, Numpy, Skitlearn, Seaborn were applied via Jupyter.

3.7.4 PyCharm and Djungo

After completing all ML prediction codes in the back end using Python, Anaconda and Jupyter, it was being useful to present results through a graphical user interface. PyCharm as a Python IDE available at [35] for designing interface and Djungo available at [36] as Python web framework were chosen.

4. RESULTS AND DISCUSSION

All the framework outcomes of in terms of experimental execution and overall benchmarking, presenting performance results before and after hyper parameter optimization were presented. Additionally, the performance result for voting classifier and all input algorithms were demonstrated. Lastly, implementing the best performed algorithm through an interface was also illustrated.

4.1 Experimental Execution and Overall Benchmarking

Regarding to experimental execution, the dataset separated into testing and training parts then all six classification algorithms (see Figure 2.a) applied to training data and each algorithm with default list of parameters had a basic performance (see Table 7). Then, two different hyper parameter techniques (see section 3.4) were implemented in all six algorithms except averaging classifier to boost the default accuracy level for each algorithm. Next, the best parameters (see Table 6) obtained for each algorithm and the algorithm performances changed (see Table 8). In addition, the best four performed algorithms (see Figure 2.b) were entered as input for voting classifier and again the performance modified (see Table 9). Finally, the best performed algorithm was Random Forest which achieved 98.69% of accuracy.

Concerning overall benchmarking, only one dataset group (PIMA Diabetes Dataset) and one performance measurement (Accuracy) used with seven different ML algorithms. Due to the reason that the dataset consisted of a small amount of records, execution time and memory consumption cannot be measured and all the implementation processes have done swiftly. In other words, all algorithms together applied on the training data and it did not take a noticeable execution time or spend huge memory space.

4.2 Default Performance Analysis

Each algorithm has its own basic accuracy level. This accuracy was directly achieved after data cleaning and preprocessing to find out their default performance and compare it with the accuracy gained after applying hyper tuning process. Table 7 illustrated the default performance for applying algorithms; the lowest accuracy level was 75.81% for DT and KNN algorithms. By contrast, each of SVM and LR algorithms reached the height performance which was 81.69%.

Table 7: Default algorithm performances.

Algorithm Names	Basic Accuracy %
Logistic Regression	81.69%
Decision Tree	75.81%
Random Forest	79.08%
K-Nearest Neighbors	75.81%
Support Vector Machine	81.69%
Naïve Bayes	79.08%

4.3 Performance with Hyper Tuning Analysis

Randomized Search CV and TPOT (AutoML) techniques were applied to enhance the accuracy level. There was an increase in accuracy for all algorithms except NB which was in TOPT method (see Table 8). The algorithms reached more than 80% in accuracy, according to TOPT technique, were selected and utilized as an attribute for averaging voting classifier.

Table 8: Accuracy percent for algorithms using Hyper Tuning.

Algorithm Names	Randomized Search Accuracy %	TPOT Accuracy %
Logistic Regression	81.69	82.35
Decision Tree	79.73	81.04
Random Forest	80.39	81.04
K-nearest neighbor	77.12	79.73
Support Vector Machine	82.35	82.35
Naïve Bayes	79.08	77.12

4.4 Performance with Averaging Classifier Analysis

Algorithms with red accuracy in the above table, were used as input for voting classifier. As suggested from it is name, it takes the prediction average from all algorithms and produces the final prediction for the model; not all clustered algorithms that have performed well, it is vital to demonstrate how an algorithm responds fast to solve various real-world problem [37]. According to Table 9, RF was the best performed among others; it achieved 98.69% and was applied in this framework to predict diabetes.

Table 9: Accuracy present for algorithms using Averaging Classifier.

Algorithm Names	Voting classifier Accuracy %
Logistic Regression	79.08
Decision Tree	81.69
Random Forest	98.69
Support Vector Machine	83.66
Voting classifier	85.62

4.5 Implementation Results

To implement the prediction results while applying RF algorithm, a simple web-based graphical interface was created using PyCharm as an IDE and Djungo as a Python web framework. The interface consisted of one html page designed via PyCharm tool called predict.html and opened through a local host (127.0.0.1). In the predict page, there were multiple columns which correspond to the independent variables and data entry. In addition, there was a submit button that contained the prediction steps using RF algorithm. When submit button was clicked, the prediction result of the data was presented. In case of having diabetes, it showed a positive. Figure 6.a demonstrates the prediction diabetes as positive according to the PIMA Diabetes Dataset and the outcome is 1. On the other hand, Figure 6.b shows negative prediction result depending on the target variable ‘outcome’ which is 0.

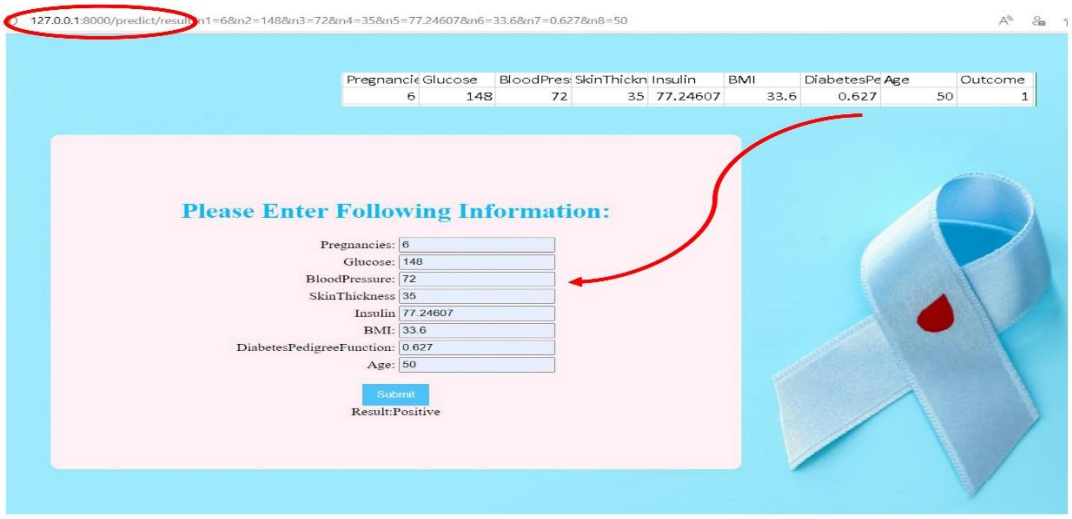

Figure 6.a: The positive result.

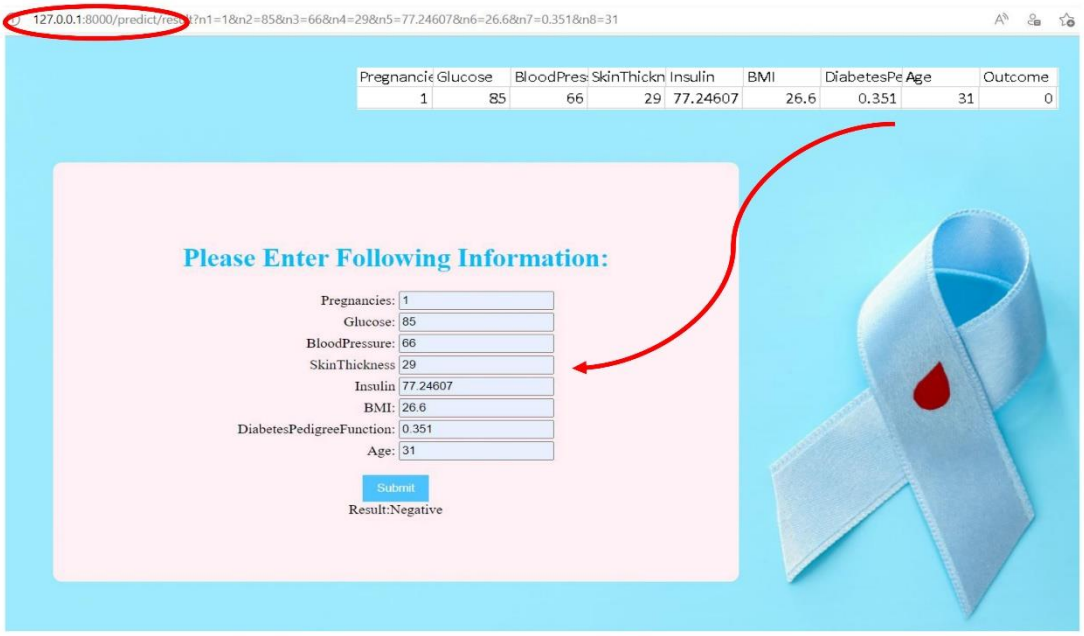

Figure 6.b: The negative result.

5. CONCLUSION

As a matter of fact, early prediction and estimation is necessary in diabetes treatment since it is a chronic disease and causes many fatal heath issues. Hence, the current study has proposed a diabetes prediction framework through working on Pima Indian Diabetes Dataset (PIDD) consisting of 768 records with eight independent and one dependent variable. The pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index (BMI), diabetes pedigree function and age were the independent medical features that the framework has worked on. The outcome value was either 0 or 1; the framework has focused on solving binary classification problem and selecting seven different classification algorithms. In order to make the PIDD ready to use, data cleaning, preprocessing and standardization are performed. In

addition, all outlier records from skin thickness and insulin attributes are eliminated. Then, two different hyper parameter techniques namely, Randomized Search and TPOT (autoML) are used and comparative analysis is done to show the rising accuracy between default and optimized parameters for each algorithm. The averaging classifier is applied and four best algorithms are taken as attribute for it. As a result, RF is the best algorithm to fit the model with accuracy 98.69% and it is used to predict diabetes via a simple graphic user interface. The dataset utilized in this framework was taken from one source and restricted to a small amount of samples. For further studies, a dataset can be collected from multiple sources and/or hospitals to obtain more generalized dataset records. Furthermore, more medical features like physical activity, smoking status, taking alcohol, life style and emotional status could be collected and analyzed in another study.

REFERENCE

[1] S. Kousar, "Type 1 Diabetes: Causes, Symptoms and Treatments, Review with Personal Experience," Current Research in Diabetes & Obesity Journal, vol. 11, issue 4, pp. 555817, 2019.

[2] U. Galicia-Garcia, A. Benito-Vicente, Sh. Jebari, A. Larrea-Sebal, H. Siddiqi, K.B. Uribe, H. Ostolaza and C. Martin, "Pathophysiology of type 2 Diabetes Mellitus," International Journal of Molecular Science, vol. 21, issue 17, pp.6275, 2020.

[3] R. Singla,A. Singla, Y. Gupta and S. Kalra,"Artificial Intelligence/Machine Learning in Diabetes Care," Indian Journal of Endocrinology and Metabolism, vol. 23, issue 4, pp. 495-497, 2019.

[4] M. Makroum, M. Adda, A. Bouzouane and H. Ibrahim,"Machine Learning and Smart Devices for Diabetes Management: Systematic Review,"Sensors, vol. 22, issue 5, pp.1843, 2022.

[5] Y. Jian, M. Pasquier, A. Sagahyroon and F. Aloul," A Machine Learning Approach to Predicting Diabetes Complications," Healthcare, vol. 9, issue 12, pp. 1712, 2021.

[6] A. Tuppad and Sh.D. Patil,"Machine learning for diabetes clinical decision support: a review," Advances in Computational Intelligence, vol. 2, issue 22, pp.2022, 2022.

[7] L.N. Liyanage,"DIABETES MELLITUS AND ITS RISK FACTORS, Epitome," International Journal of Multidisciplinary Research, vol. 4, issue 9, pp.114 – 119, 2018.

[8] International Diabetes Federation, "Diabetes facts & figures," idf.org, Dec. 9, 2021. [Online]. Available: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html[Accessed: Sep. 10, 2022].

[9] X. Lin, Y. Xu, X. Pan, J.Xu, Y. Ding, X. Sun, X. Song, Y. Ren and P. Shan,"Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025," Scientific Report, vol. 10, pp.14790, 2020.

[10] World Health Organization, "Diabetes," who.int, Dec. 10, 2022. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1 [Accessed: Sep. 22, 2022].

[11] J.J. Khanam and S.Y. Foo,"A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, issue 4, pp. 432-439, 2021.

[12] E. Begic, A. Arnautovic and I. Masic, "ASSESSMENT OF RISK FACTORS FOR DIABETES MELLITUS TYPE 2," Mater Sociomed, vol. 28, issue 3, pp.187-90, 2016.

[13] S. Park, Ch. Kim and X. Wu, "Development and Validation of an Insulin Resistance Predicting Model Using a Machine-Learning Approach in a Population-Based Cohort in Korea," Diagnostics, vol. 12, issue 1, pp.212, 2022.

[14] M.R. Rajput and S.S Khedgikar, "Diabetes prediction and analysis using medical attributes: A Machine learning approach," Journal of Xi'an University of Architecture & Technology, vol. XIV, issue 1, pp. 98-103, 2022.

[15] Sh. Pourbahrami, M. Balafar, L. Khanil and Z.Kakarash, "A survey of neighborhood construction algorithms for clustering and classifying data points," Computer Science Review, vol. 38, pp. 100315, 2020.

[16] Z. Mushtaq, M.F. Ramzan, S. Ali, S. Baseer, A. Samad and M. Husnain,"Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques," Mobile Information Systems, vol. 2022, pp.6521532, 2022.

[17] R. Krishnamoorthi, Sh. Joshi, H.Z. Almarzouki, P.K. Shukla, A. Rizwan, C. Kalpana and B. Tiwari,"A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques," Journal of Healthcare Engineering, vol. 2022, pp.1684017, 2022.

[18] S. R, S. M, M.K. Hasan, R.A. Saeed, S.A. Alsuhibany and S. Abdel-Khalek, "An Empirical Model to Predict the Diabetic Positive Using Stacked Ensemble Approach," Front. Public Health, vol. 9, pp.792124, 2022.

[19] B.S. Ahamed, M.S. Arya and A.O. Nancy V, "Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques," Front. Comput. Sci, vol. 4, pp. 835242,2022.

[20] F. Bano, M. K and R. MadanaMohana, "Predict Diabetes Mellitus Using Machine Learning Algorithms,"Journal of Physics: Conference Series, vol. 2089, pp.012002, 2021.

[21] O. Llaha and A. Rista,"Prediction and Detection of Diabetes using Machine Learning," CEUR Workshop Proceedings, vol. 2872, pp. 94-102, 2021.

[22] A.K. Jaggi, A. Sharma, N. Sharma, R. Singh and P.S. Chakraborty,"Diabetes Prediction Using Machine Learning," Intelligent System, vol. 185, pp. 383-392, 2021.

[23] N. Ahmed, K. Hamakarim and Z.Kakarash,"A Temporal and Social Network-based Recommender using

Graph Clustering," Passer Journal, vol. 4, issue 2, pp. 180-18, 2022.
[24] R. Patra and B. Khuntia,"Analysis and Prediction Of Pima Indian Diabetes Dataset Using SDKNN Classifier Technique," IOP Conference Series: Materials Science and Engineering, vol. 1070, pp. 012059, 2021.
[25] E. Elgeldawi, A. Sayed, A. Galal and A. Zaki,"Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis," Informatics, vol. 8, issue 4, pp.79, 2021.
[26] Y. Zhao, R. Zhang and X. Li,"AutoDESS: AutoML Pipeline Generation of Classification with Dynamic Ensemble Strategy Selection,"arXiv, vol. 2, pp. 2201.00207, 2022.
[27] M. Soni and S.Varma, "Diabetes Prediction using Machine Learning Techniques," International Journal of Engineering Research & Technology (IJERT), vol. 9, issue 9, pp. 921-925, 2022.
[28] B. Hassan and T. Rashid, "A multi-disciplinaryEnsemble Algorithm for Clustering Heterogonous Datasets," Neural Computing and Applications, vol. 33, pp. 10987-11010, 2021.
[29] A. Baratloo, M. Hosseini, A. Negida and G. El Ashal, "Part 1: Simple Definition and Calculation of Accuracy," Sensitivity and Specificity, Emergency (Tehran),vol.3, issue 2, pp. 48-9, 2015.
[30] Python, "Download the latest version for Windows," python.org, Dec. 11, 2022. [Online] Available: https://www.python.org/downloads/ [Accessed: March 11, 2022].
[31] A. Dhruv, R. Patel and N. Doshi, "Python:The Most Advanced Programming Languages for Computer Science Application," Science and Technology Publications, Lda, pp.292-299, 2021.
[32] ANACONDA, "Data Science Technology for a better world," anaconda.org, Dec. 11, 2022. [Online]. Available:https://www.anaconda.com/ [Accessed: March 11, 2022].
[33] Jupyter, "Installing Jupyter," jupyter.org, Dec 11,2022. [Online]. Available:https://jupyter.org/install[Accessed: March 11.2022].
[34] B. Randles, I. Pasquetto, M. Golshan and Ch. Borgma, "Using the Jupyter Notebook as a tool for Open Science: An Empirical Study," ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 1-2, 2017.
[35] JetBrains, "Download PyCharm," jetbrains.com, Dec. 11,2022. [Online]. Available: https://www.jetbrains.com/pycharm/download/#section=windows[Accessed: March 11, 2022].
[36] ANACONDA, "Installers," anaconda.org, Dec. 11,2022. [Online]. Available: https://anaconda.org/anaconda/django [Accessed: March 11,2022].
[37] B. Hassan, T. Rashid and H. Hamarashid, "A novel cluster detection of COVID-19 patients and medical disease conditions using improved evolutionary clustering algorithm star," Computer in Biology and Medicine, vol. 138, pp. 104866, 2021.