Sule Lamido University Journal of Science & Technology Vol. 6 No. 1&2 [March, 2023], pp. 205-213
https://doi.org/10.56471/slujst.v6i.319

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Prediction of Diabetes Mellitus using Machine Learning Algorithms: Comparative Analysis of K-Nearest Neighbor, Random Forest and Logistic Regression

A.M. Adeshina[1], G.C. Ogwume[2]
[1,2]Department of Information Technology, University of Ilorin, Ilorin, Nigeria.
adeshina.am@unilorin.edu.ng[1], ogwumechinonso@gmail.com[2]

Abstract

Diabetes Mellitus is a chronic and one of the deadliest diseases. Diabetes disease increases the risk of long-term complications, including heart diseases and kidney failures, among others. Undoubtedly, Diabetes Mellitus patients may live longer and lead healthier lives if the disease is detected early. Over the years, several efforts have been on more accurate and early detection procedures to safe patients of Diabetes Mellitus. Interestingly, with the applications of Information Technology to the disease diagnoses and therapy managements, more attention has been on using machine learning in the predictions and early detection of Diabetes Mellitus. Unfortunately, determining the most appropriate machine learning algorithm with the best performance in terms of optimum accuracy still remains a challenge. The study proposes a framework for Diabetes Mellitus detection using Machine Learning Algorithms. The proposed framework was evaluated using K-nearest neighbor (KNN), Random Forest (RF), and Logistic Regression (LR). Extensive experiments were conducted to analyze the performance of the framework focusing on four distinct different clinical datasets. To ensure robust, web compatible framework, Python and its popular data science related packages, Pandas, Numpy, Seaborn, Matplotlib and Pickle were used for the implementation. Significantly, using the standard datasets obtained from the National Institute of Diabetes and Kidney Disease, Random Forest was able to predict Diabetes Mellitus in the datasets with the best accuracy of 93.4 %.

Keywords: *Diabetes Mellitus, Machine Learning, K-Nearest Neighbor, Random Forest, Logistic Regression*

1. Introduction

Diabetes mellitus is a chronic disease having the potential of causing a worldwide health care crisis. In so many instances, diabetes has been recorded the cause of death of a number of patients, following other diseases of heart and cancer. Unfortunately, previously proposed computer-assisted diagnosis frameworks, such as SurLens (Adeshina et al., 2012), CAHECA (Adeshina et al., 2014), ConnectViz (Adeshina & Hashim, 2016), though resourceful in detecting features of datasets, they are inefficient in predicting diabetes mellitus. However, the rise in automation procedures brings about a possible dimension to the diabetes mellitus predictive procedures. The aim of machine learning and data mining is to depict relevant and required information stored in datasets and generate clear and understandable description of patterns for further usage. Multiple opportunities are created for healthcare through machine learning models being a viable potential for achieving advanced predictive analytics

With this study, comparison of the accuracy of three machine learning algorithms were studied towards detecting Diabetes Mellitus, which could possibly be at advantageous at the early stage of the chronic disease. Unfortunately, without computer-aided applications, doctors' procedures would have been completely based on common knowledge for the treatment. Indeed, the common knowledge processes take time. With machine learning, patterns could be identified earlier (Warke et al., 2019), and therefore utilized accordingly. Those experiments for this study, the three algorithms of machine learning, the Logistic Regression, the Random Forest Classifier, and the K Nearest Neighbors were used to predict early diabetes detection. The study was to find the best optimal algorithm to predicting diabetes faster and quicker, and possibly with the appropriate supporting web application.

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 6 No. 1&2 [March, 2023], pp. 205-213
https://doi.org/10.56471/slujst.v6i.319
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

2. Related Works

A number of studies attempted developing an application to detect diseases. Dey, Hossain & Rahman (2018) used KNN, ANN, SVM, and Naive Bayes algorithms for disease detection. The datasets were divided into two for usages in the training and testing. Min-Max Scaler (MMS) normalization method was introduced so as to increase the accuracy. The study made used of Tensor Flow as a machine learning model. Similarly, Sisodia & Sisodia (2018) worked on diabetes detection using classification algorithms. The study evaluated how efficient the different algorithms were in achieving classifications taking into account those measurements such as accuracy, precision, recall, as well as F-measure. The calculations of the accuracy used were with instances appropriately and inappropriately classified. Naïve Bayes was observed exceptional with the accuracy of 76.30%. Further verifications of the findings were established with Receiver Operating Characteristic (ROC) curves. In the same vein, Wei, Zhao & Miao (2018) presented a study of Deep Neural Network (DNN), Logistic Regression (LR), Support Vector Classifier (SVC), Naive Bayes and Decision Tree techniques, towards identifying diabetes mellitus, in which the outcome eventually aligned with the submissions of previous studies.

According to Ballewar et al. (2021), diabetes mellitus was described as a series of conditions that include hyperglycemia and insulin resistance, grouped into two distinct categories of type 1 and type 2. Perhaps with this in mind, Kowsher et. al. (2019) gave a detection model to predict type-2 diabetes. The researchers emphasized on machine learning algorithms, the KNN, Logistic Regression, Decision Tree, Random Forest, Naive Bayes, Artificial Neural Network, and Linear Discriminant Analysis for the diabetes predictions. The study worked precisely on data collection, preprocessing, training of the datasets, and the detection and predictions. While 80% of the dataset were considered for the training and the remaining were used as the testing dataset.

Ullah et al. (2022) proposed machine learning methods to detect the risk factors of diabetes and also provided decision support for medical practitioners in diagnosing diabetes. The study used synthetic minority over-sampling technique integrated with the edited nearest neighbor for balancing datasets. The evaluation of the study shows the reliability of the proposed model and was found outperformed other methods.

The study of Kulkarni et al. (2023) was a combination of non-invasive nature of Electrocardiogram (ECG) with the power of machine learning to detect diabetes and pre-diabetes. The study was evaluated using datasets that include time-aligned heartbeats recoded digitally. The algorithm accurately predicted diabetes-related classes and was found to be resourceful in the early detection of diabetes and pre-diabetes. Bhattacharya & Datta (2023) proposed a focused algorithm targeting classification for the prediction of diabetes. The researchers utilized ensemble classifiers which are based on machine learning. Random Forest, Bagging, AdaBoosting and Gradient Boosting were the models used. The results reported showed Gradient Boosting is the best among all other ensemble classifiers attempted to predict diabetes. The team of Fabrian et al. (2023) developed artificial intelligent model that can predict diabetes following comparison between the K-Nearest Neighbor (KNN) and Naïve Bayes algorithms to see which of the algorithms are best for diabetes prediction. The study was evaluated using datasets containing several health attributes. Evaluation of the study using confusion matrix confirmed Naïve Bayes algorithm out-performed KNN. Similarly, lately, Himi et al. (2023) proposed smartwatch-based prediction system for multiple diseases including diabetes using machine algorithms integrated with sensors to collect bodily statistics. The evaluation of the study proved it resourceful in predicting diseases that are considered early and ability to predict multiple disease vulnerabilities before reaching an irrecoverable stage.

Undoubtedly, AI in medicine of today has a significant role to play towards a greater improvement in diagnosis for the greater benefits of the medics and the patients. However, another possible clinical challenge could be on achieving personalized treatments. Undoubtedly, realization of personalized treatments falls into those general paradigms of precision medicine.

2.1 Materials and Methods

The framework proposed for this study, the methods and procedures are discussed in this section. The datasets used were collected from Kaggle and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Data preparations includes converting the data to appropriate format, the data pre-processing and the use of the Machine

Sule Lamido University Journal of Science & Technology Vol. 6 No. 1&2 [March, 2023], pp. 205-213
https://doi.org/10.56471/slujst.v6i.319
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Learning to model and manipulate the data has previously discussed. Several tests were conducted using different datasets and the results were discussed.

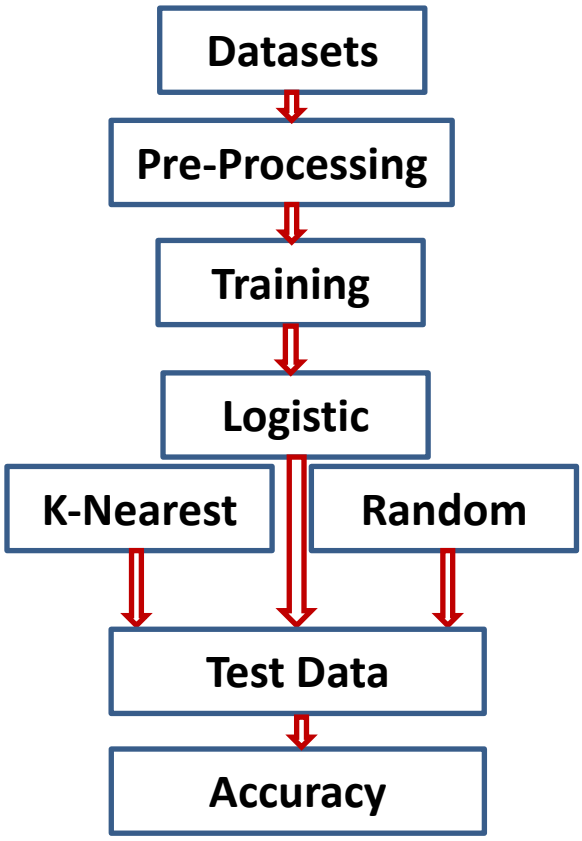

Figure 1: The Proposed Framework

3.2 The Proposed Framework

The proposed framework is presented in Fig. 1. The significant stages include acquisition of datasets, datasets pre-processing and training the datasets. The application of Logistic Regression, K-Nearest Neighbor and Random Forest are the core of the framework's procedures. Testing of the datasets and documentation of their respective performance evaluations conclude the framework stages.

3.2.1 Dataset

All the disease dataset used in the evaluation of this study were obtained from National Institute of Diabetes and Kidney Diseases (NIDDK). The datasets contain 9 attributes in total, as presented in Table 1. Though an ordered collection of data, while handling the data, the dataset could be a bunch of tables, schema and other objects. The analyses were focused on detecting whether or not a patient has diabetes, based on diagnostic measurements included in the dataset.

3.2.2 Pre-Processing Data

The raw datasets have to be carefully transformed into a preferred format for further processing within the framework. In the actual sense, the datasets in their raw forms are usually incomplete, sometimes not consistent or contains other categorized forms of error. Preprocessing stages are alternative procedures to improving the datasets for use. Manipulation of data before it is being used is also significant in order to ensure or enhance its performances. Without any doubt, preprocessing the data is considered important in the data mining processes.

Sule Lamido University Journal of Science & Technology Vol. 6 No. 1&2 [March, 2023], pp. 205-213
https://doi.org/10.56471/slujst.v6i.319
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Table 1: Diabetes dataset

S/N	Attribute	Description
1.	Pregnancies	Pregnancy times
2.	Plasma Glucose	Plasma glucose concentration of 2hours in an oral glucose tolerance test
3.	Diastolic Blood Pressure	Diastolic Blood Pressure measured in mmHg
4.	Triceps Thickness	Triceps skin fold thickness measured in mm.
5.	Serum Insulin	2-Hour serum insulin measured in µU/ml
6.	BMI	Calculation of the body mass index.
7.	Diabetes Pedigree	Diabetic history of the family and other factors.
8.	Age	How old is the Patient, in years.
9.	Outcome	Presence of diabetes. 1 – Yes, 0 –No

3.2.3 Training Data

Data that are labeled are called training data. Such labeled data are used in the training of machine learning algorithms towards ensuring proper decisions. Large dataset is used to teach machine learning model. For supervised machine learning models, the training data is labeled.

4. Implementation

Python was used in the development of the proposed framework. Using CSV format, *Pandas* was imported to read into the proposed framework. *Numpy* was deployed in converting the data into a suitable format for classification model. *Seaborn* and *Matplotlib* were used for the visualizations. Furthermore, Logistic Regression, K Nearest Neighbor and Random Forest algorithms were imported from *sklearn*. Finally, *Pickle*, which is available in *sklearn* was used to save the model for use. In order to achieve good results, large quantities of data must be ingested into the machine learning data exploration models. Without any doubt, if the data is not thoroughly explored, it would affect the accuracy of the model. Furthermore, with Pandas, data exploration is made easier. Figures 2, 3, 4 and 5 show the correlation populated by the algorithm of the study focused datasets 1, 2, 3 and 4 respectively.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	796.000000	796.000000	796.000000	796.000000	796.000000	796.000000	796.000000	796.000000	796.000000
mean	3.871859	121.177136	69.250000	20.633166	79.163317	32.054146	0.472907	33.395729	0.353015
std	3.373021	31.972158	19.105969	16.012687	115.140585	7.833318	0.331674	11.792778	0.478208
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	63.500000	0.000000	0.000000	27.375000	0.244000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	22.500000	32.250000	0.374500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	126.000000	36.600000	0.627500	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 2: Populated correlation values from the framework for Dataset_1

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
ount	1009.000000	1009.000000	1009.000000	1009.000000	1009.000000	1009.000000	1009.000000	1009.000000	1009.000000
ean	3.845391	121.039643	69.355798	20.392468	78.966303	32.087116	0.462710	33.094153	0.346878
std	3.369724	32.650756	19.277713	15.906740	112.640216	7.967603	0.321895	11.726634	0.476213
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.500000	0.238000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	29.000000	32.300000	0.364000	29.000000	0.000000
75%	6.000000	140.000000	80.000000	32.000000	130.000000	36.600000	0.612000	41.000000	1.000000
max	17.000000	383.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 3: Populated correlation values from the framework for Dataset_2

Sule Lamido University Journal of Science & Technology Vol. 6 No. 1&2 [March, 2023], pp. 205-213
https://doi.org/10.56471/slujst.v6i.319
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	3.703500	121.182500	69.145500	20.935000	80.254000	32.193000	0.470930	33.090500	0.342000
std	3.306063	32.068636	19.188315	16.103243	111.180534	8.149901	0.323553	11.786423	0.474498
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	63.500000	0.000000	0.000000	27.375000	0.244000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	40.000000	32.300000	0.376000	29.000000	0.000000
75%	6.000000	141.000000	80.000000	32.000000	130.000000	36.800000	0.624000	40.000000	1.000000
max	17.000000	199.000000	122.000000	110.000000	744.000000	80.600000	2.420000	81.000000	1.000000

Figure 4: Populated correlation values from the framework for Dataset_3

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	1699.000000	1699.000000	1699.000000	1699.000000	1699.000000	1699.000000	1699.000000	1699.000000	1699.000000
mean	3.663920	121.444379	68.910536	21.022955	81.250147	32.186051	0.474017	33.140671	0.343143
std	3.290875	32.374515	19.323427	16.137040	112.248548	8.104375	0.326025	11.886125	0.474899
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	64.000000	0.000000	0.000000	27.300000	0.244500	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	43.000000	32.000000	0.380000	29.000000	0.000000
75%	6.000000	142.000000	80.000000	32.000000	130.000000	36.750000	0.627500	40.000000	1.000000
max	17.000000	199.000000	122.000000	110.000000	744.000000	80.600000	2.420000	81.000000	1.000000

Figure 5: Populated correlation values from the framework for Dataset_4

4.2 Data Cleaning

The significant of data cleaning in the workflow of machine leaning analysis cannot be over-emphasized due its impact in the overall output of the study. Some of the factors to be considered in data cleaning processes include:

i. duplicate observations (sometimes referred to as irrelevant observations),
ii. data labeling,
iii. missing data points (sometimes referred to as null data points)
iv. unexpected outliers

Usually, if standard datasets were used for the study, possible issues of duplicate observations of irrelevant information, data labeling and probably the unexpected outliers would have been taken care of.

4.2.1 Missing (Null) Data Points

Missing data points occur during implementation, when no information is provided for one or more items or for a whole unit. Such challenges in datasets were managed using *Pandas* functions, *Dataframe.isnull( )* and *dfCopy.isnull( ).sum( ).* Through eliminating irrelevant features or instances, we make feature subset, referred to as the features subset selection process. The process greatly reduces dimensionality of data and greatly improves.

```
Pregnancies                 0
Glucose                     5
BloodPressure              35
SkinThickness             236
Insulin                   391
BMI                        11
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Figure 6: Missing populated data values from the framework for Dataset_1

Sule Lamido University Journal of Science & Technology Vol. 6 No. 1&2 [March, 2023], pp. 205-213
https://doi.org/10.56471/slujst.v6i.319
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

```
Pregnancies                 0
Glucose                     5
BloodPressure              43
SkinThickness             302
Insulin                   492
BMI                        15
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Figure 7: Missing data values from the framework for Dataset_2

```
Pregnancies                 0
Glucose                    13
BloodPressure              90
SkinThickness             573
Insulin                   956
BMI                        28
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Figure 8: Missing data values from the framework for Dataset_3

```
Pregnancies                 0
Glucose                    13
BloodPressure              90
SkinThickness             573
Insulin                   956
BMI                        28
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Figure 9: Missing data values from the framework for Dataset_4

Ahead the applications of machine learning to the datasets, categorical variable were determined. The outcome labels have two classes, *0* for no disease and *1* for disease for all cases in the particular dataset. The categorical variables of the datasets were presented in Fig. 10 as categorical variable outcome plot for Dataset 1 and for Fig. 11, Fig. 12 and Fig. 13 for the datasets 2, 3 and 4 respectively.

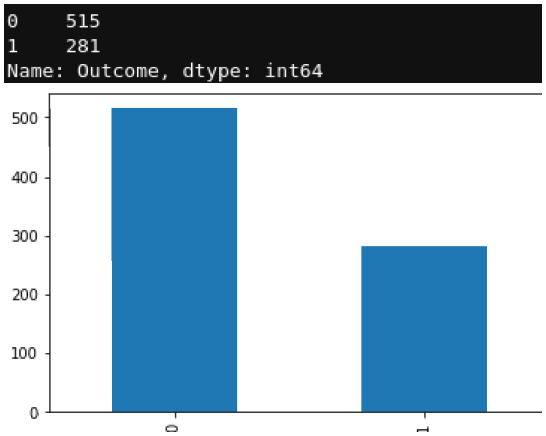

Fig. 10: Categorical variable outcome plot for Dataset 1

Sule Lamido University Journal of Science & Technology Vol. 6 No. 1&2 [March, 2023], pp. 205-213
https://doi.org/10.56471/slujst.v6i.319
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

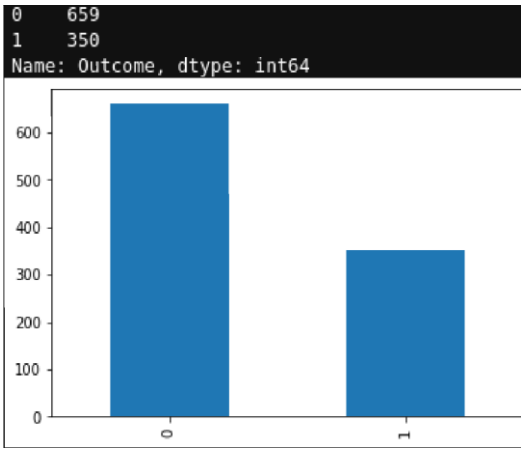

Fig. 11: Categorical variable outcome plot for Dataset 2

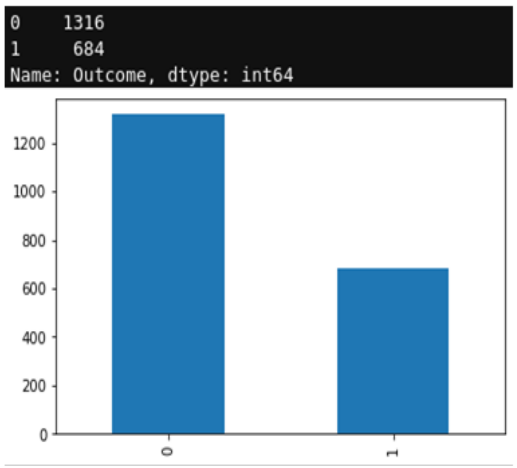

Fig. 12: Categorical variable outcome plot for Dataset 3

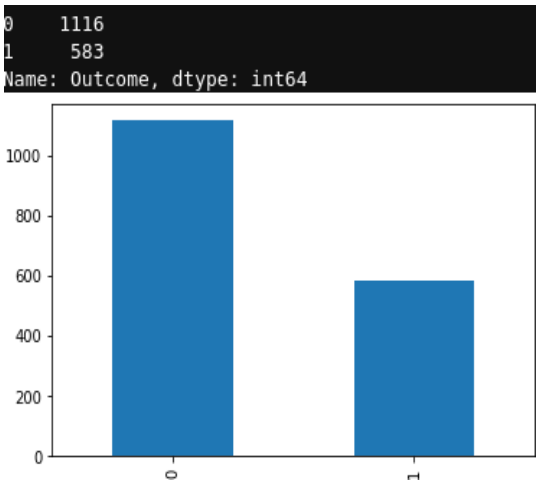

Fig. 13: Categorical variable outcome plot for Dataset 4

Accuracy evaluations of the K-Neighbors Classifier, Logistic Regression and Random Forest in the prediction of Diabetes Mellitus were evaluated with specific focus on datasets 1, 2, 3 and 4. The accuracy performance scores of the classifiers with respect to the datasets are presented in Table 2, Table 3, Table 4 and Table 5.

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 6 No. 1&2 [March, 2023], pp. 205-213
https://doi.org/10.56471/slujst.v6i.319
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Table 2: Accuracy performance score of classifiers with Dataset_1

S/N	Classifier	Accuracy
1	K-Neighbors Classifier(KNN)	72%
2	Logistic Regression (LR)	74%
3	Random Forest (RF)	75%

Table 3: Accuracy performance score of classifiers with Dataset_2

S/N	Classifier	Accuracy
1	K-Neighbors Classifier(KNN)	73%
2	Logistic Regression (LR)	77%
3	Random Forest (RF)	78%

Table 4: Accuracy performance score of classifiers with Dataset_3

S/N	Classifier	Accuracy
1.	K-Neighbors Classifier (KNN)	77%
2.	Logistic Regression (LR)	79%
3.	Random Forest (RF)	93%

Table 5: Accuracy performance score of classifiers with Dataset_4

S /N	Classifier	Accuracy
1	K-Neighbors Classifier(KNN)	77%
2	Logistic Regression (LR)	79%
3	Random Forest (RF)	91%

5. Conclusion and Future Work

Efforts have been put together in this study towards predicting diabetes using machine learning techniques. The LR, KNN and RF algorithms were used in the study and their performances were carefully studied on adult population datasets. Categories of risk factors related to the disease were cautiously taken into consideration. The results of the experiments proved RF best among the three algorithms considered with the best prediction accuracy of 93.4% relative to the datasets used.

Future work intends to focus on solving the imbalances in datasets, with under sampling and/or oversampling algorithms which would most likely increases the chances of achieving better accuracy in the classification of the respective classes even with fewer samples. Furthermore, it would be more resourceful to use real and latest datasets of patients from hospital or clinics for continuous training, and possibly optimization of the proposed framework.

References

Adeshina, A.M., Hashim, R., Khalid, N.E.A., Abidin ,S.Z.Z. (2012). Locating Abnormalities in Brain Blood Vessels using Parallel Computing Architecture. Interdiscip Sci Comput Life Sci 4,161–172.

Adeshina, A.M., Hashim, R., Khalid, N.E.A. (2014). CAHECA: Computer Aided Hepatocellular Carcinoma Therapy Planning. Interdiscip Sci Comput Life Sci 6, 222–234.

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.edu.ng

Adeshina, A.M., Hashim, R. (2015). ConnectViz: Accelerated Approach for Brain Structural Connectivity using Delaunay Triangulation. In: Interdisciplinary sciences: computational life sciences, pp 1–13.

Ballewar, A., Shukla, Y., Yumnam, R., Shihan, A., Mandawkar, U. (2021). Diabetes Prediction Using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), Volume 7(3), pp. 487-493.

Bhattacharya, M., Datta, D. (2023). Development of Predictive Models of Diabetes using Ensemble Machine Learning Classifier. Advancements in Smart Computing and Information Security: First International Conference, ASCIS 2022. Part 1, 377-388.

Dey, S.K., Hossain, A., Rahman, M.M. (2018). Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm. 21st International Conference of Computer and Information Technology (ICCIT), 21- 23 December.

Fabian, M.E., Ferdinan, F.X., Sendani, G.P., Suryanigrum, K.M., Yunanda, R. (2023). Procedia Computer Science 216, 21-30.

Himi, S.T., Monalisa, N.T., Whaiduzzaman, M.,Barros, A., Uddin, M.S. (2023). MedAi: A Smartwatch-based Application Framework for Common Diseases Prediction using Machine Learning. IEEE Access, 2023.

Kowsher,M., Turaba, M.Y., Sajed, T., Rahman, M.M. (2019). Prognosis and Treatment Prediction of Type-2 Diabetes Using Deep Neural Network and Machine Learning Classifiers. 22nd International Conference on Computer and Information Technology (ICCIT).

Kulkarni, A.R., Patel, A.A., Pipal, K.V., Jaiswal, S.G., Jaisinghani, M.T., Thulkar, V., Gajbhiye, L., Gondane, P., Patel, A.B., Mamtani, M., Kulkarni, H. (2023). Machine-Learning Algorithm to Non-Invasively Detect Diabetes and Pre-Diabetes from Electrocardiogram. BMJ Innovations 9(1).

Sisodia, D., Sisodia, D.S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science 132: 1578-1585.

Ullah, Z., Saleem, F., Jamjoom, M., Fakieh, B., Kateb, F., Ali, M.A., Shah., B. (2022). Detecting High-Risk Factors and Early Diagnosis of Diabetes using Machine Learning Methods. Computational Intelligence and Neuroscience 2022.

Warke, M., Kumar, V.,Tarale, S., Galgat, P., Chaudhari, D.J. (2019). Diabetes Diagnosis using Machine Learning Algorithms. International Research Journal of Engineering and Technology (IRJET). 6(3).

Wei, S., Zhao, X., Miao, C. (2018). A Comprehensive Exploration to the Machine Learning Techniques for Diabetes Identification. IEEE 4th World Forum on Internet of Things (WF-IoT).