Revue d'Intelligence Artificielle
Vol. 36, No. 3, June, 2022, pp. 347-352
Journal homepage: http://iieta.org/journals/ria

Diagnostic Analysis of Diabetes Mellitus Using Machine Learning Approach

Navya Pratyusha Miriyala[1], Rajya Lakshmi Kottapalli[1], Geetha Pratyusha Miriyala[2], Giulio Lorenzini[3*], Charankumar Ganteda[1], Venkata Apparao Bhogapurapu[1]

[1] Department of Mathematics, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur 522302, India
[2] School of Electronics Engineering, Vellore Institute of Technology, Amaravati, Guntur 522237, India
[3] Department of Engineering and Architecture, University of Parma, Parma 43124, Italy

Corresponding Author Email: giulio.lorenzini@unipr.it

https://doi.org/10.18280/ria.360301

Received: 28 March 2022
Accepted: 26 May 2022

Keywords:
diabetes mellitus, machine learning, decision making, performance metrics, accuracy

ABSTRACT

Diabetes Mellitus (DM) is caused due to the elevated levels of blood sugar i.e., said to be hyperglycemia. The DM is a metabolic chronic disease; therefore, early diagnosis and treatment is necessary to avoid life-threatening risks. According to the World Health Organization (WHO), the diabetes cause high mortality rate with 1.5 million deaths in a year. With the remarkable improvisations in the technology, the disease can be diagnosed earlier. In this paper, we have developed a decision-making support with the machine learning algorithms for DM diagnosis. The Pima Indians Diabetes dataset was chosen to train with Machine Learning algorithms. Our approach begins with Exploratory data analysis, and later the data is sent for data pre-processing and perform the feature Selection techniques. The important features are selected and finally, the data is trained with six various Machine learning (ML) algorithms such as Naïve Bayes, KNN, Random Forest, Logistic Regression, Decision Tree, and eXtreme gradient boosting. The Experimental results of the ML algorithms are calculated by the performance metrics in which that the eXtreme Gradient Boosting has scored highest with 88.2% accuracy than other machine learning algorithms.

1. INTRODUCTION

With the increase in population, it is important to have an initiative for developing computational systems to improve health outcomes and reduce global challenges. The current studies of health-care sector continue to advance, and the development of the decision-making systems were becoming much more efficient and reliable. In the recent years, the healthcare systems were designed to provide better decision making and diagnose the disease accurately with higher efficacy.

In most developing countries, DM has become a very severe disease and it is classified as a non-communicable disease. According to 2017 statistics, 425 million people were diagnosed with diabetes, and every year approximately two to five million patient's death occurs due to diabetes [1]. One in ten individuals in the United States was affected by DM and the new occurrences of type-1 and type-2 diabetes have increased dramatically among young people is stated by National Diabetes Statistics Report 2020 [2, 3]. As health care system is an important pillar to the society, it is necessary to utilize the capabilities of methods and technologies such as artificial intelligence, machine learning etc. for developing new methods and applications in medical sector.

The rise of technology had a significant impact on the medical field. For people who are unable to approach a clinic or receive emergency treatment, health consequences may be determined in a matter of seconds. For all people who benefit from technology, it bridges the gap in distance and resources. Related to the resources, the data collection in the healthcare sectors have large volumes of database, which have structured and unstructured data. Our proposed methodology deals with the unstructured data where our model performs exploratory data analysis and data pre-processing techniques to convert the data as structured. The feature selection model determines and selects the relevant important features. These features are trained with Machine learning algorithms for predicting diagnostic capable metrics such as accuracy. Machine learning (ML) acts as a discipline method operated without a user interface through the algorithms. ML can perform a particular task without a formulation namely Naïve Bayes, KNN, Random Forest, Logistic Regression, the algorithms used in our methodology are Decision Tree, and eXtreme gradient boosting. The data is differentiated into training, and the algorithms were applied for determining the accuracy. The best-performed classifier obtained is considered as a Qualified algorithm. The paper is organized into four sections, where the section 2 explains about literature survey, section 3 details about methodology, section 4 focus on the experimental results and section 5 is about the conclusion.

2. LITERATURE SURVEY

Every patient has various risk factors and complications related to diabetes disease. The Machine Learning technology has been increasingly popular in recent years for predicting several diseases. Researchers have developed many algorithms and software tools. These aspects have shown huge potential in the medical care sector. In this section, past literature works directly related to the proposed methodology is presented in the below.

Several researchers uses the popular dataset i.e., PIMA Indians Diabetes Dataset (PIDD) from the ‘Kaggle’ or ‘data.world’ repository. Wu et al. [4] had focused on primary difficulties of the classification such as accuracy of the prediction model and generating the adaptability for two or more datasets at equal time, using PIDD Dataset on two ML algorithms such as modified K-means method and the logistic regression algorithm and programmed the research on the Waikato (WEKA) tool with 95.42% accuracy. Nia-arun and Moungmai [5] used four different ML models such as logistic regression, artificial neural network, random forest and naïve bayes with the combination of bagging and ensemble. For this methodology, the author used 30,122 instances collected from 26 primary care units in regional hospital of Sawanpracharak. In the experimental results, the random forest has scored higher than other algorithms with 85.55% accuracy. Meng et al. [6] proposed to predict diabetes mellitus using ML algorithms. For modelling, the author collected the primary data from a private medical source located in Guangzhou, China. The author implemented the diagnostic system with three ML algorithms used for training namely logistic regression, artificial neural network, and decision trees. The author programmed the methodology using the R environment and the experimental analysis shows that the decision tree (C5.0) has scored highest accuracy with 80.68% out of all ML models. Tigga and Garg [7] used the logistic regression algorithm on PIDD dataset where the data was split into training and testing and been implemented using R language. The metric results of the model have scored 75.32% accuracy and the rate error is of 24.68%. Mani and Shraddha used for Random Forest, and Neural Networks to visualize and predict diabetes using PIDD dataset [8]. The experimental result in the papers shows that the random forest has scored with highest accuracy and the observed limitation of the highest accuracy was obtained due to the minimum number of features.

Temurtas et al. [9] used PIMA dataset and proposed models used for training with multilayer neural networks with Levenberg–Marquardt (LM) algorithm and probabilistic neural networks. The author trained the model 10-fold cross validation applied on the algorithms showed a better performance. The model accuracy scores were 82.37% and 78.13% for two models. The study [10] used comparative popular algorithms such as deep neural network, support vector machine etc. for predicting the diabetes along with various data pre-processing techniques such as principal component analysis (PCA) and linear discriminant analysis (LDA). The author used PIDD data and was keen to observe the metrics based on the 10-fold cross validation. During the experiment, the author observed that the deep neural network (DNN) has outperformed other algorithms with highest accuracy of 77.8 6%. Predicted diabetes disease using the soft computing technique aims to provide insights into the vast data applied with ML algorithms. The PIDD dataset was used by Bhat et al. [11], since the data imputation is necessary to remove, therefore they applied data pre-processing technique with the hybrid combination of the Decision tree (CART) technique and Genetic Algorithm. Later the author used classic neural network and scored the accuracy of 82.33%. From these literature survey, it can be understood that the PIDD dataset scores vary with and without feature selection. In the following section, the Methodology is explained as brief.

3. METHODOLOGY

The research methodology is proposed based on four main phases i.e., data preparation, data transformation, Feature selection, and machine learning modelling. The data preparation imports the data, and the exploratory data analysis is carried out to understand the features of PIDD. Later, the data pre-processing and feature selection is used to convert the data from unstructured to structured. Using ML algorithms, we classify and find the best accuracy-based algorithm for diagnosing diabetes. In this paper, we use six various ML algorithms namely Naïve Bayes, KNN, Random Forest, Logistic Regression, Decision Tree, and eXtreme gradient boosting were used to predict the diabetes. Further, the machine learning algorithms efficiency is calculated with performance indices, such as Accuracy, Specificity, and Sensitivity etc. From the metric score, the highest accuracy model is considered as the best classifier for diagnosing diabetes. The workflow of the proposed methodology is depicted in the Figure 1.

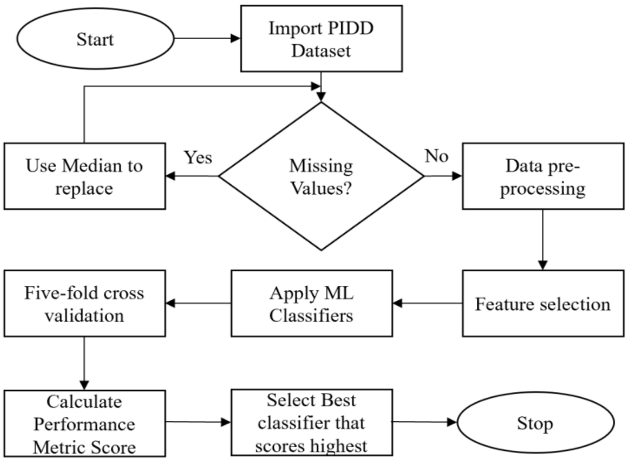


Figure 1. Workflow methodology for diagnostic analysis of diabetes mellitus using machine learning approach

3.1 Data preparation

The PIDD data is collected from the Kaggle website to diagnose diabetes Mellitus. The data consists of 768 patient occurrences with eight numbers of features. With the eight features, this classifier identifies whether the patient is suffering from diabetic or non-diabetic. In the Figure 2. the data of diabetic and healthy patient instances are shown, where the patients with diabetic are 268 and the healthy patients count is 500.

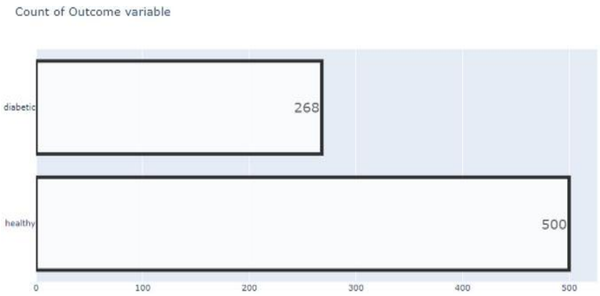


Figure 2. Diabetes and non-diabetes dataset count

3.2 Data pre-processing

Data pre-processing is an important step that enhances the data to promote the extraction of meaningful data. The Medical data-collection has commonly null and not available (NA) data. This data is said to be missing data. In general, there are ‘n’ number of ways to handle missing data values. Some of the statistical methods are median, mean and mode etc. In our work, the missing values are replaced with ‘median’ such that the glucose, blood pressure, skin, thickness, insulin, and BMI could be balanced. The outliers also have noisy values, and form inconsistency. This is solved at the data transformation [12].

3.3 Data transformation

Generally, normalization and standardization fall under data transformation. In this paper, we have applied normalization technique. Normalization aims to change the dataset values into a standard scale without disturbing and distorting the scale range differences. In our work, the features have different ranges. In Data Normalization, Min-Max normalization is applied where the data scales all the features in the same range of values between 0 and 1, ignoring the outliers in the data. The Min- Max scalar equation is given by Eq. (1):

$$z = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where, *Z*=Normalized Feature, x has the input feature values, min(*x*) is feature minimum value and max(*x*) is maximum feature value. With the normalization, the outliers are removed.

3.4 Feature selection

The feature selection measures on the performance observed by the correlation [13]. There are two types of approaches measure the correlation between the same set of features. The two approaches are classical linear correlation and information theory. The linear correlation approach is considered mostly in our feature selection, as it defines and measures the random parts of features. Correlation coefficient sim(*x*, *y*) defined by the Eq. (2):

$$\mathrm{si}\,m(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}} \tag{2}$$

where, var() is the feature variance where *x* is input and *y* is output and cov(*x*, *y*) is the covariance between *x* and *y*. The sim(*x*, *y*) values range between -1 and 1. The sim(*x*, *y*) value takes the values from 1 or -1, when they are correlated completely, and the sim(*x*, *y*) takes the value 0, if *x* and *y* are independent. According to the correlation mapped with the dataset target; Glucose, BMI, and Age, Pregnancies are the selected features.

3.5 Machine learning algorithms

The binary classification problem is identified based on the given PIDD where the features are used for diagnosing diabetes. For the diagnostic analyses, the machine learning technique were applied to the dataset. The dataset is trained with the following six machine learning algorithms as Naïve Bayes, KNN, Random Forest, Logistic Regression, Decision Tree, and eXtreme gradient boosting is applied to the data for predicting diabetes based on the parameter indices, i.e., Accuracy, Specificity, Sensitivity, Precision, F1 Score, and ROC-AUC. The highest algorithm accuracy is selected for the prediction of diabetes. The algorithmic data flow of the Machine learning algorithms is structured in Algorithm 1, and the Machine Learning model flow for training the data is shown in Figure 3.

```
Algorithm 1: Prediction of Diabetes with Machine Learning models
1. Generate the testing and training dataset from the original dataset
2. Specify the Machine Learning Algorithms
Model=[GuassianNB(), KneighborsClassifier(),
RandomForestClassifer(), LogisticRegression(),
DecisionTreeClassifer(),XGBClassifier()]
3. Perform Five folded cross-validation (i)
4. For (i=0; i≤5; i++) do
        Model = Model[i];
        Model.fit();
        Model.predict();
5. Print parameter indices (Accuracy, Recall, Precision, F1-Score, ROC AUC)
6. END
```

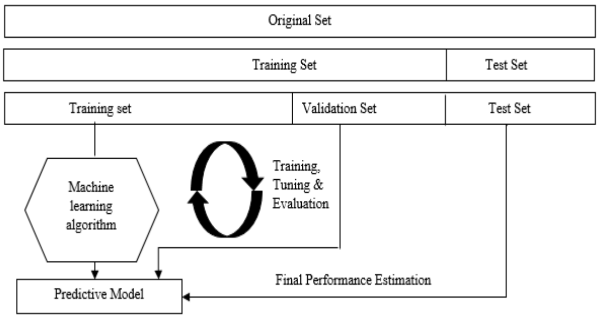

Original Set
Training Set
Test Set
Training set
Validation Set
Test Set
Machine learning algorithm
Training, Tuning & Evaluation
Predictive Model
Final Performance Estimation

Figure 3. Machine learning data training and testing flow

3.5.1 Extreme gradient boosting

Yu and Liu [14] have proposed advanced supervised learning known as Extreme Gradient Boosting (XGBoost). The XGBoost algorithm was recognized, after the superior performance observed at the Kaggle dataset community competition. The XGBoost has the advantages of high efficiency and tractability for prediction. The regularization has improved the prediction quality and reduced the log loss function, which soothes the final learned weights to avoid over-fitting. To prevent over-fitting of the classifier, XGBoost supports the sampling of rows and columns. If the data is fitted with the target, a tree smaller value with the outcome is preferred. The possible tree structure with infinity outcomes could be eliminated, with a “greedy algorithm” to find and optimize the tree structure in practical applications.

3.5.2 Naive bayes

Naive Bayes (NB) is a classification approach obtained from Bayes theorem [15]. The Naive Bayes is termed as Naïve, as it is considered between the two anomalies. The first is the assumption of predictive features that are independent, and the second is they are no concealed features. The naive Bayes classification technique predicts the probability of different classes. The naive Bayes technique is best for text analytics, to indicate substance, categorize standards, etc. The naïve bayes is defined by Eq. (3):

$$p(c/x) = \frac{p(x/c) * p(c)}{p(x)} \tag{3}$$

3.5.3 K-Nearest neighbor

K-Nearest neighbor technique (KNN) is a machine learning algorithm with a simple method to implement the dataset [16]. KNN technique maps the prediction of diabetes into a similarity-based classification. The vectors are mapped with the medical data, and these vectors depict the features in N-dimensional space. The decision is finally computed with the Euclidean distance for 'N' matrix. The closest k-record with the highest similarity to the test is considered, and KNN is an easy learning method. The selected k-records are observed on the performance of diagnosing the target by the majority rule. The KNN is given by Eq. (4):

$$\begin{aligned} &d(p,q) = d(q,p) \\ &= \sqrt{(q_1 - p_1)^2 + (q_1 - p_1)^2 + .. + (q_1 - p_1)^2} \\ &= \sqrt{\sum_{i=0}^{n} (q_i - p_i)^2} \end{aligned} \tag{4}$$

The prediction of diabetes is computed using KNN are considered depending upon these steps:

a) By determining the number of K-Nearest neighbors.

b) By calculating the distance between the trained samples.

c) Sorting all the prepared records according to distance values.

d) Regarding the majority class labels of k-nearest neighbors, by assigning its predictive value for diabetes.

3.5.4 Logistic regression

Logistic regression also called logit regression, or the logit model, is the other supervised learning technique [17] from the statistic area carried off by predictive machine learning analysis. The logistic technique algorithm provides a binary output in 0/1, yes/no, and true/false. Logistic regression depicts the possible outcome for the two possible (*y*) dependent attributes i.e., output and two or more nominal essential (*x*) independent variables i.e., input. In Mathematical terms, the Logistic curve is used to make a prediction. For the prediction, LR needs an independent activity feature, and the activity score is the numerical value of the separate attribute.

Logit is also used for corresponding multiplicate weights and the activity scores. For the achievement of the probability target class, the score should be passed to the Logistic function. A protected class was provided for each of its given features to predict whether the person has diabetes or not by the calculated logit—the simple Logistic regression form is given by:

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \alpha + \beta(x)$$

The logit is said to be a natural logarithm (log) with the probabilities of input ratio (*x*) and (*y*) happening (i.e., diabetic patient) to the probabilities of (1-*p*(*x*)) of (*y*) (i.e., non-diabetic patient).

3.5.5 Decision tree

In medical diagnosis, the Decision tree is mostly used as a good algorithm [18]. The algorithm evaluation starts with the root node, and the root node follows the branch node. The decision nodes are connected to the branches with termination in the leaf nodes. The preliminary data with the target class is predicted using the decision rule. For classification and regression, the decision tree uses the root node and internodes for branch and leaf for connecting. A root node has two or more branches satisfying the instances with different features, while leaf nodes represent classification. In every stage, the Decision tree chooses each node to evaluate the highest information gain through the 'Gini index' or 'entropy' based on classification and regression among all the attributes. In this model, we are using Gini index and is given by Eq. (5):

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{5}$$

where, *i* is the number of classes; P_i is the Proportional of the samples that belong to class *n* for a particular node.

3.5.6 Random forest

The random forest comes under the ensemble technique called "Bagging" which is used in both the classification and regression problems. Lakshmi et al. [19] has first proposed the random decision forest algorithm, and the algorithm is further improved by Ho [20]. In the various developing sectors of prognosis and diagnosis, the random forest has proved satisfactory performance [21]. With the random change in the decision tree combinations, the RF model increases DT possibilities by having a put-back sample.

3.6 Performance evaluation

By Categorizing the algorithms based on the performance, these parameters are to be observed for evaluation. The performance metrics are Accuracy, Sensitivity, Specificity, Precision, and F1 Score which are in the Eq. (6) to Eq. (10).

(a) Accuracy (Acc.,): Accuracy Provides overall performance and observes correctly predicted data of the classifier and formulates as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

(b) Sensitivity (Sn.,): The metrics describe the positive result identified from the classifier and is given by:

$$Sn = \frac{TP}{TP + FN} \tag{7}$$

(c) Specificity (Sp.,): The specificity describes the negative result identified by the classifier and is given by:

$$Sp = \frac{TN}{TN + FP} \tag{8}$$

(d) Precision (Pr.,): Precision is the number of the total target positive results by predicted positive results is expressed by:

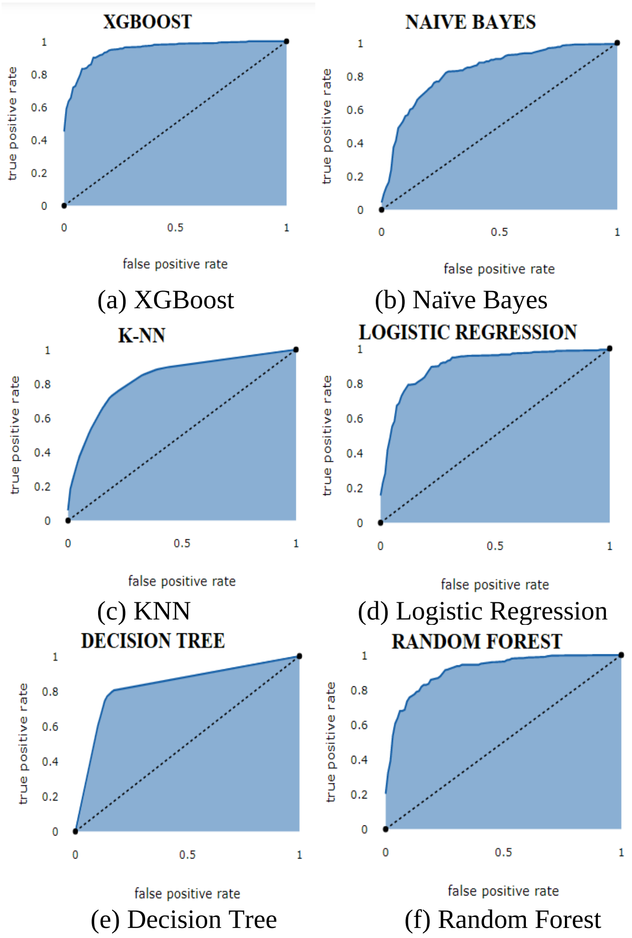
$$\Pr = \frac{TP}{TP + FP} \tag{9}$$

(e) F1-Score (F1.,): F1-Score is the Precision and Recall harmonic mean, and [0,1] is the range. The F1-Score suggests the classifier robustness and the mathematical expression is:

$$\mathrm{F1} = 2 * \frac{1}{\left(\frac{1}{\Pr ecision} + \frac{1}{\mathrm{Re}\,call}\right)} \tag{10}$$

4. EXPERIMENTAL RESULTS

The PIDD dataset is applied to different ML classification algorithms, and the results of the techniques are observed. After testing, the Five folded cross-validation is evaluated to provide mean accuracy for every algorithm. Table 1 shows the results of cross-validation various ML algorithms. From the Experimental Results, Extreme Gradient Boosting gained the highest accuracy with 88.2%, and second highest accuracy is gained by Decision tree algorithms with 85.3%.



(a) XGBoost (b) Naïve Bayes

(c) KNN (d) Logistic Regression

(e) Decision Tree (f) Random Forest

Figure 4. Mean ROC-AUC for the Six Machine Learning Algorithms with k=5

By maintaining a better prognosis of diabetes mellitus, the descriptive capability assessment is observed from the receiver operating curve (ROC). The ROC curve is a graphical plot relation with the true-positive rate (TPR) and false-positive rate (FPR), which improves the classification of algorithmic quality. The ROC Curve of Machine Learning Algorithms is independently shown in the Figure 4(a) to 4(f). The higher ROC-AUC provides better performance. In these figures, the highest ROC are from the XGBoost, Random Forest, and Logistic Regression are scored with 0.95, 0.91, and 0.9.

Table 1. Performance indices observed with the machine learning algorithms

Model	Acc %	Sn %	Sp %	Precision %	F1-Score %	ROC
XGBoost	88.2	80.9	92	84.7	82.7	0.95
Naive Bayes	75.3	60.9	87.6	61.2	89.6	0.82
K-NN	79	70.4	83.4	70.9	69.6	0.83
Logistic Regression	83.7	73.8	89	78.6	76.1	0.90
Decision Tree	85.3	79.3	90	76.5	78.3	0.82
Random Forest	83.6	86.2	82.69	86.6	72.7	0.91

5. CONCLUSIONS

In the traditional Healthcare system, diabetes was evaluated by the physicians through diagnostic tests. The procedure of the diagnostic test in the health care system consumes high economic costs and loss of time for the patient to detect diabetes. Therefore, our paper proposes the diagnostic capability to predict the diabetes. In this work, the PIDD dataset is trained and tested with various ML techniques. These models are cross-validated with five k-folds in which the mean accuracy of five folds were calculated. The observed experimental results shows that the eXtreme gradient boosting (XGBoost) provides the highest accuracy of 88.2%, followed by the Decision tree provide 85.3% accuracy. Since the data is slight imbalanced, hence the future scope can cope with the sampling technique to balance the data.

ACKNOWLEDGMENT

The author thanks and acknowledge the staff working in the Department of Mathematics of Koneru Lakshmaiah Education Foundation (KLEF).

REFERENCES

[1] World Health Organization. (2022). Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes, Accessed on May 2022.

[2] National Diabetes Statistics Report. (2022). https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf, Accessed on May 2022.

[3] Paul, S., Riffat, M., Yasir, A., Mahim, M.N., Sharnali, B. Y., Naheen, I.T., Rahman, A., Kulkarni, A. (2021). Industry 4.0 applications for medical/healthcare services. Journal of Sensor and Actuator Networks, 10(3): 43. https://doi.org/10.3390/jsan10030043

[4] Wu, H., Yang, S., Huang, Z., He, J., Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10(1): 100-107. https://doi.org/10.1016/j.imu.2017.12.006

[5] Nai-arun, N., Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science, 69(1): 132-142.

https://doi.org/10.1016/j.procs.2015.10.014

[6] Meng, X.H., Huang, Y.X., Rao, D.P., Zhang, Q., Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung Journal of Medical Sciences, 29(2): 93-99. https://doi.org/10.1016/j.kjms.2012.08.016

[7] Tigga, N.P., Garg, S. (2021). Predicting type 2 diabetes using logistic regression. In proceedings of the Fourth International Conference on Microelectronics, Computing and Communication Systems. Lecture Notes in Electrical Engineering, Singapore, 673: 491-500. https://doi.org/10.1007/978-981-15-5546-6_42

[8] Mani, B., Shraddha, K. (2021). A data mining approach for the diagnosis of diabetes mellitus using random forest classifier. International Journal of Computer Applications, 120(8): 36-39. https://doi.org/10.5120/21249-4065

[9] Temurtas, H., Yumusak, N., Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with Applications, 36(4): 8610-8615. https://doi.org/10.1016/j.eswa.2008.10.032

[10] Wei, S., Zhao, X., Miao, C. (2018). A comprehensive exploration to the machine learning techniques for diabetes identification. In 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, pp. 291-295. https://doi.org/10.1109/WFIoT.2018.8355130

[11] Bhat, V.H., Rao, P.G., Shenoy, P.D., Venugopal, K.R., Patnaik, L.M. (2009). An efficient prediction model for diabetic database using soft computing techniques. In International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing, pp. 328-335. https://doi.org/ 10.1007/978-3-642-10646-0_40

[12] Abidin, N.Z., Ismail, A.R., Emran, N.A., (2018) Performance analysis of machine learning algorithms for missing value imputation. International Journal of Advanced Computer Science and Applications, 9(1): 442-447. https://doi.org/0.14569/IJACSA.2018.090660

[13] Yu, L., Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. Proceedings of the Twentieth International Conference on Machine Learning, Washington DC, pp. 856-863.

[14] Yu, L., Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 5(1): 205-1224.

[15] Chen, T., Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 785-794. https://doi.org/10.1145/2939672.2939785

[16] McCallum, A., Nigam, K., (1998). A comparison of event models for Naive Bayes text classification. In AAAI-98 workshop on learning for text categorization, 752: 41-48.

[17] Aha, D.W., Kibler, D., Albert, M.K. (1991). Instance-based learning algorithms. Machine Learning, 6(1): 37-66. https://doi.org/10.1007/BF00153759

[18] Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82: 115-121. https://doi.org/10.1016/j.procs.2016.04.016

[19] Lakshmi, B.N., Indumathi, T.S., Ravi, N. (2016). A Study on C. 5 decision tree classification algorithm for risk predictions during pregnancy. Procedia Technology, 24: 1542-1549. https://doi.org/10.1016/j.protcy.2016.05.128

[20] Ho, T.K. (1995). Random decision forests. In Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, Quebec, Canada, Montreal, QC, Canada, pp. 278-282. https://doi.org/10.1109/ICDAR.1995.598994

[21] Breiman, L. (2001). Random forests. Machine Learning, 45(1): 5-32. https://doi.org/10.1023/a:1010933404324