Research Article

Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach

Sourav Kumar Bhoi[a], Sanjaya Kumar Panda[b], Kalyan Kumar Jena[a], P. Anshuman Abhisekh[c], Kshira Sagar Sahoo[d], Najm Us Sama[e,*], Shweta Supriya Pradhan [c], Rashmi Ranjan Sahoo[a]

[a] Department of Computer Science and Engineering, Parala Maharaja Engineering College (Govt.), Berhampur - 761003, India

[b] Department of Computer Science and Engineering, National Institute of Technology, Warangal - 506004, India

[c] Department of Pharmacology Maharaja Krishna Chandra Gajapati Medical College & Hospital (Govt.), Berhampur - 760004, India

[d] Department of Information Technology, VNRVJIET, Hyderabad- 500090, India

[e] Department of Science, Deanship of Common First Year, Jouf University, Sakaka, Saudi Arabia

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract

Nowadays, diabetes is a common disease that affects millions of people over the world, and women are mostly affected by this disease. Recent healthcare studies have applied various innovative and advanced technologies to diagnose people and predict their disease based on clinical data. One of such technologies is machine learning (ML) in which diagnosis and prediction can be made more accurately. In this paper, the designed model predicts the diabetes of females of Pima Indians heritage by taking the clinical dataset. Here, this problem is considered as a binary classification problem. Therefore, supervised learning algorithms have been used, such as classification tree (CT), support vector machine (SVM), k-Nearest Neighbour (k-NN), Naïve Bayes (NB), Random Forest (RF), Neural Network (NN), AdaBoost (AB) and Logistic Regression (LR). We use the female Pima Indians diabetic dataset from Kaggle and UCI data repository and k-fold cross-validation to carry out the process of training and testing. We determine the area under the curve (AUC), classification accuracy (CA), F1, precision and recall results of all the supervised learning algorithms and compare them to determine the best algorithm that is suitable for prediction. For this, we use the Orange 3.24.1 open-source platform to generate the results, which uses Python open-source libraries. From the results, it is concluded that the LR performs better in comparison to other algorithms.

Keywords: Machine Learning; Supervised Learning; Ada-Boost; SVM; k-NN

1. Introduction

Diabetes is a non-communicable disease, which affects healthcare severely by reducing the efficiency of a person [48, 34, 47, 9, 30, 49]. In this disease, the blood glucose level rises more than the normal glucose level in the body [35, 20, 5, 11, 41]. It is noteworthy to mention that glucose is the form of sugar that is needed by the body for better metabolism. All cells need glucose as a source of energy. However, if the blood glucose level increases due to lack of insulin hormone in the body, then it imbalances the blood glucose in the body that results in severe damage to other parts of the body, such as eyes, heart, kidneys and many more [27, 12, 39, 44]. It is controlled by changing the lifestyle, such as food habits, medications, exercises to name a few. If the disease is diagnosed in time by prediction, then a person's health can be improved. Therefore, if the healthcare system uses intelligent 48 prediction mechanisms, then a person's life can be saved [40, 1, 23]. However, in this work, our main focus is to only predict whether a female of Pima Indian heritage has diabetes or not. Diabetes is of three types, namely type-1, type-2 and gestational [24, 8, 42, 46, 17, 45]. In type-1, the immune system destroys the insulin cells. It generally happens to children and adolescents. In type-2, the pancreas makes very little insulin [36, 18, 37, 43]. It generally happens to adults. The former type is also called as insulin resistance, whereas the latter type is called as insulin deficiency. Recent healthcare studies have applied various technologies to diagnose people and predict their disease based on the collected clinical data. Nowadays, the healthcare system can predict diabetes more accurately using ML techniques. ML enables a computer to become intelligent by learning from the experiences or inputs (i.e., clinical data) and predict the output category (i.e., disease) [21, 6, 26, 2, 10]. The ML techniques are categorized into supervised, unsupervised and reinforcement learning. In supervised learning, the features, as well as the target class, are used as input for learning. In unsupervised learning, there is no such target class. Here, the input data is provided without the target class [13, 25, 28]. This data is further used for clustering using a similarity measure. Note that the input data may be unstructured. Reinforcement learning is an approach that uses the hit and trial method, where the computer or machine finds the best outcome. To get the best outcome, award and penalty values are used [16, 7, 15, 19]. In this paper, the diabetes of females of Pima Indians heritage problem has considered as a binary classification problem and mainly focused on the supervised learning algorithms. The supervised algorithms are best suits for classification problem. The main contributions, in this paper, are listed below.

- Various supervised learning algorithms have been utilized, such as classification tree (CT), support vector machine (SVM), k-nearest neighbour (k-NN), naive bayes (NB), random forest (RF), neural network (NN), adaboost (AB) and logistic regression (LR) on female Pima Indians diabetic dataset [14] to predict the diabetes of females.
- In the work, nine different features have considered, namely pregnancies, glucose, blood pressure,

skin thickness, insulin, body mass index (BMI), diabetes pedigree function, age and outcome (i.e., 0 or 1) that are present in the dataset to make the prediction and compare the results in terms of classification accuracy (CA). The supervised learning algorithms use k-fold cross-validation (preferably, k = 10) to split the datasets.

• The comparison results are evaluated in terms of five parameters, namely Area under the curve (AUC), CA, F1, precision and recall using an open-source platform, called Orange 3.24.1 and the results show that logical regression performs better in comparison to other algorithms.

The rest of the paper is presented as follows. The next section presents the related works. Methodology section presents the proposed approaches. The analysis and results section discusses the outcome of the experiments. Finally, in the last section the conclusion and future scope of the work has been highlighted.

2. Related Work

Many diabetes prediction algorithms have been proposed by the researchers to accurately predict the types 84 of diabetes as well as diabetes of Pima Indians [23, 9, 49, 35, 20, 36, 15, 19, 24, 85 8, 10]. These works are briefly discussed as follows. Zolfagri et al. [48] have proposed a method to diagnose diabetes in females' populations of Pima Indians using an ensemble of neural network and SVM. Pham et al. [34] have predicted diabetes by a new data mining approach that balances using fitting and generalization. Wu et al. [47] have proposed a method to diagnose diabetes using a semi-supervised learning method that uses Laplacian SVM. Sanakal et al. [40] have diagnosed diabetes using the prognosis of fuzzy c-means clustering and SVM. Al et al. [1] have used decision tree technique to diagnose type-2 diabetes. Kumari et al. [23] have used SVM for classification of diabetes. Dey et al. [9] and Zou et al. [49] have implemented a web-based approach to predict diabetes using ML approaches. However, none of the paper has addressed all the well-known supervised learning algorithms at a glance. Pradhan et al. have predicted diabetes using an artificial neural network (ANN) [35]. Karthikeyani et al. have proposed a comparison performance of data mining algorithm (CP-DMA) to predict the diabetes disease [97 20] . Karatsiolis et al. have proposed region-based SVM algorithm for medical diagnosis of Pima Indians [43] [18]. Guo et al. have used bayes network to predict type-2 diabetes [11]. Maniruzamman et al. [27] have performed a comparative analysis of diabetes mellitus data using ML techniques. Han et al. [12] have analyzed the data using rapid miner software. Saji et al. [39] have predicted diabetes using a multilayer perceptron. Jahangir et al. have proposed an expert system to predict diabetes using an autotuned multilayer perceptron [13]. Li et al. [25] have proposed a weight-adjusted approach to diagnose diabetes. In [28] have proposed an accurate diabetes risk stratification using ML techniques. Like Pradhan et al. [35], Sivastava et al. [43] have predicted diabetes using ANN approach. Kala et al. have proposed an intelligent hybrid system for a diabetes diagnosis [16]. Chen et al. have proposed a hybrid prediction model to diagnose type-2 diabetes using decision trees and k-means [7]. Kahramanli et al. have proposed a hybrid system for diabetes and heart disease [15]. In [19] authors have proposed a genetic algorithm approach using NN to diagnose Pima Indians diabetes. In [24] authors have proposed a fuzzy technique to diagnose diabetes accurately. Many such algorithms have been presented in [8, 42, 46, 17, 45, 21, 6, 26, 51, 10, 3, 50, 52]. However, they have not combinedly addressed most of the supervised learning algorithms.

3. Methodology

This section describes the methodology used to predict the diabetes of females of Pima Indians heritage. As stated earlier, we visualize this as a binary classification problem. It also describes the dataset, different supervised learning algorithms and the steps to perform the binary classification.

3.1 Dataset

The dataset [14] is taken from the national institute of diabetes and digestive and kidney diseases (NIDDK). The data is stored in Kaggle and UCI data repository. This dataset is mainly used to predict whether a Pima Indians female has diabetes or not. All the patients taken in the dataset are females of Pima Indians heritage of minimum age of 21 years. In order to decide a female in this dataset as diabetic, the following attributes are considered.

- **Age**: It shows the age in years. The range is 21 to 81 and the average age is 33.

- **Pregnancies:** It shows that the number of times a female gets pregnant. The range is 0 to 17 and the average is 4.

- **Glucose:** It shows the plasma glucose concentration level (2 hours). It is from 0 to 199 and the average is 121.

- **Blood pressure:** It shows the diastolic blood pressure in mm Hg. It is from 0 to 122 and the average is 69.

- **Skin thickness**: It shows the triceps skin thickness in mm. The range is 0 to 99 and the average is 21.

- **Insulin:** It ranges from 0 to 846. The average is 80.

- **BMI:** It shows body mass index in Kg/m2. The range is 0 to 67.1 and the average is 32.

- **Diabetes pedigree function:** This function scores the likelihood of diabetes. It is from 0.078 to 2.42 and the average is 0.47.

- **Outcome:** It is either 0 or 1. Here, 0 means that a female has non-diabetic and 1 means that a female is diabetes.

In this dataset, the outcome used as the target class to predict that the Pima Indians female is diabetic or not. It has 768 instances (i.e., number of rows) and 9 columns (i.e., number of attributes).

3.2 Algorithms for Prediction of Diabetes

This section discusses various supervised learning algorithms for classifying the diabetic and non-diabetics Pima Indians females. Note that these algorithms create the training dataset and testing dataset from the original dataset to classify or predict diabetes.

- **Classification tree [1, 7]:** It is a supervised machine learning approach to predict the output parameter (i.e., outcome). It consists of nodes (tests), edges (the outcome of a test) and leaf nodes (outcome). In this, the decision variable is discrete or categorical. It is mainly designed using binary recursive partitioning. This process uses iterations to split the data into partitions. The partitioning of the samples of each node is done utill all samples belong to the same class.
- **Support Vector Machine [47, 40, 23, 5, 38, 18]:** SVM is a supervised learning approach to analyse the data and used it in the classification problem. SVM constructs a hyperplane or set of hyperplanes to classify the data into different classes.
- **K-Nearest Neighbour (k-NN) [8, 29, 33, 32, 4]:** It is a classification algorithm that keeps all available data and classifies new data based on a similarity measure. The new data is classified based on the closest distance among the neighbours. The similarity measure is performed using Euclidean distance, Manhattan distance, Minkowski distance and hamming distance.
- **Naïve Bayes:** This is based on bayes theorem, which is the collection of algorithms. It has independent assumptions for the features. It is a conditional probability model, which considers each feature to contribute separately, regardless of the correlation between the features. The main advantage of this algorithm is that it requires a small dataset for training to classify the categories.
- **Random Forest (RF):** It is a classification approach that uses the average of multiple deep decision trees. The training algorithm used by RF is bootstrapping aggregation or bagging method.
- **Neural Network (NN) [48, 39, 13, 44, 19, 17]:** It is a machine learning approach, which is used for classification purposes by modelling itself as a human brain. It consists of neurons that are arranged layer-wise, which converts the input vector to output. Each unit in NN takes input and applies a non-linear function to generate output, which is further passed to the next layer. Generally, ANN is a feed-forward network. Here, weights are applied to pass from one layer to another. In this way, learning is performed to get the desired output.
- **AdaBoost:** It is an algorithm that is used for binary classification. This is mainly used with short decision trees. It is originally called as adaboost.m1. Each instance of the training set is assigned with a weight value. The initial weight is assigned as 1n, where n is the number of instances in the training set.
- **Logistic Regression:** This is a well-known and popular classification algorithm that estimates the discrete values, such as yes or no, true or false and 0 or 1. It predicts the probability of an event by using the data in a logistic function.

A common algorithm has been developed to perform the classification and prediction as shown in Algorithm 1 and present a pictorial representation in Fig. 1.

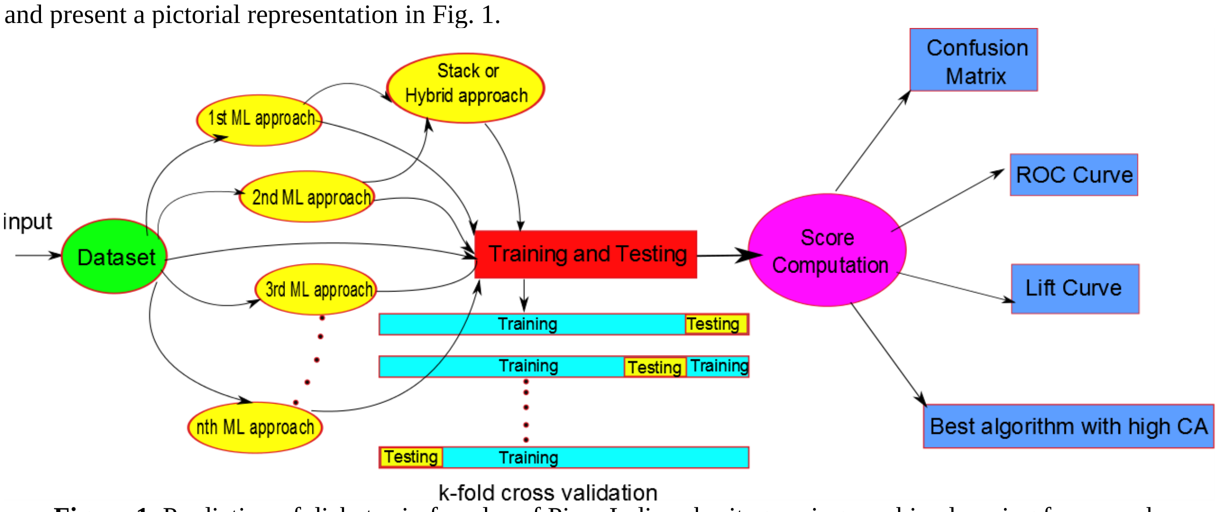


Figure 1. Prediction of diabetes in females of Pima Indians heritage using machine learning framework.

```
Algorithm 1 A common algorithm for classification and prediction

Input: Pima Indians diabetes dataset with 768 instances and 9 columns, and without missing values
Output: AUC, CA, F1, precision, recall, confusion matrix, receiver operating characteristic (ROC) curve
and lift curve Steps:

1: Open Pima Indians diabetes dataset.
2: Create a model.
3: Provide the data to the CT, SVM, k-NN, NB, RF, NN, AB and LR algorithms individually for learning.
        Each algorithm is treated as a learning model.
4: Apply k-fold cross-validation (Note that the dataset is folded k times, where k - 1 fold is used for
testing and the rest for training).
5: Determine the average of the series of tests to find the results or scores, such as AUC, CA, F1,
precision and recall, respectively.
6: Determine the best CA and its corresponding algorithm by comparing the results of all the algorithms.
7: The scores are graphically compared using the confusion matrix, ROC curve and lift curve, respectively.
```

4. Analysis and Result

In this section, results have analyzed, using Algorithm 1. We use Orange 3.24.1 [31] open-source platform to classify and predict diabetes in females of Pima Indians heritage dataset. Note that it uses the open-source libraries of Python. The open-source platform is installed in a machine with the following configurations; i) 64-bit operating system ii) 4 GB RAM iii) Intel(R) 3.40 GHz processor. The dataset is available in [14] and it is divided into the training dataset and the testing dataset as mentioned in Algorithm 1. The data is analyzed by predicting the target class using various supervised learning algorithms, namely CT, SVM, k-NN, NB, RF, NN, AB and LR algorithms. Fig. 2 shows the skeleton representation of the ML framework in Orange 3.24.1 for the prediction of diabetes in females of Pima Indians dataset.

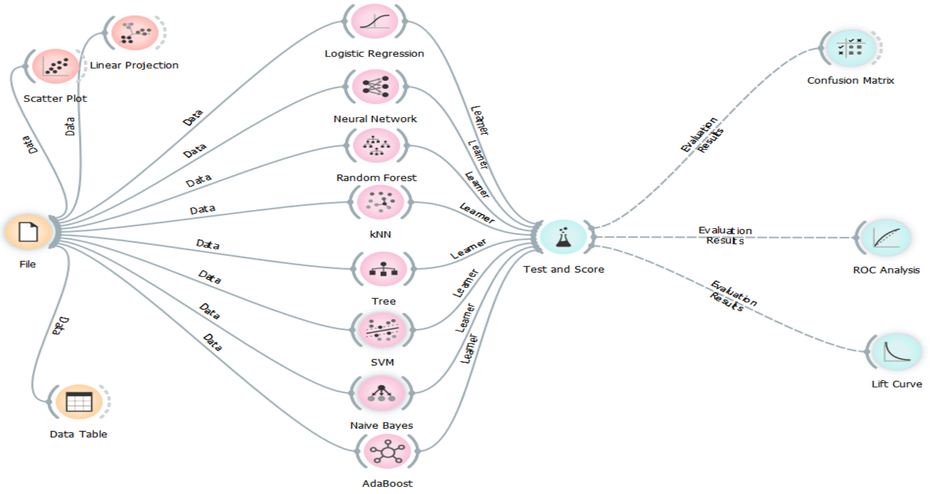
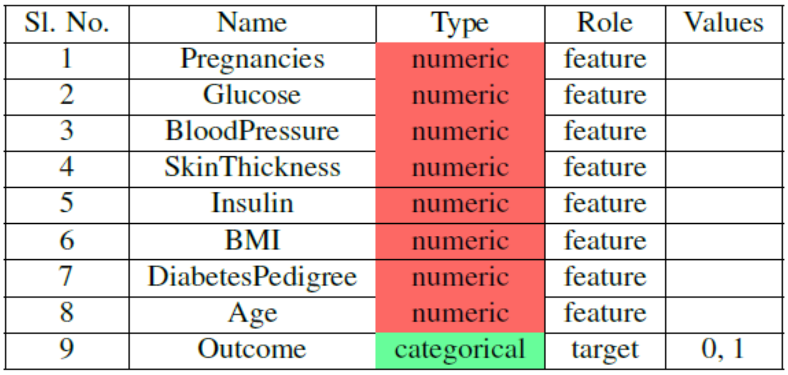


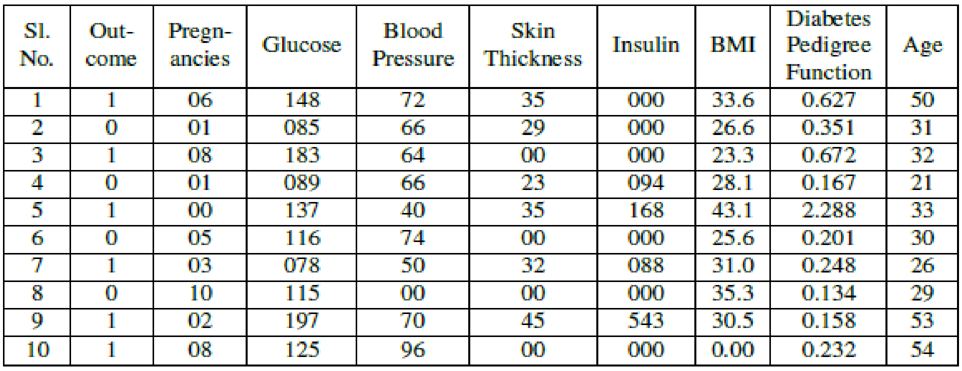
Figure 2. Designed machine learning framework in Orange 3.24.1 for prediction of diabetes in females of Pima Indians.

Firstly, the dataset file (i.e., diabetes-dataset.csv) is taken as input, which is in the form of a comma separated values file. This dataset file is opened using the "File" shown in Fig. 2. As stated earlier, the dataset file contains nine number of attributes in which the outcome attribute is "categorical" type and other attributes are "numeric" type. Here, the outcome attribute is selected as a target value in which category 0 indicates non-diabetic and category 1 is diabetic. The orange platform allows the user to convert the attribute from one type to another type. The following types are available in this platform. 1)
Categorical 2) Numeric 3) Text 4) Datetime. Fig. 3 shows the columns generated from the dataset file and setting the target class. For easy visualization, green color for "categorical" type and red color for "numeric" type has been used. The "File" shown in Fig. 3 is visualized using "Data Table" shown in the left bottom of Fig. 3 for better visualization of dataset file in the row and column format. The snapshot of the "Data Table" is shown in Fig. 4. Note that it shows the first 10 instances out of 768 instances of the dataset file. In "Data Table", we can visualize the numeric values.



Sl. No.	Name	Type	Role	Values
1	Pregnancies	numeric	feature	
2	Glucose	numeric	feature	
3	BloodPressure	numeric	feature	
4	SkinThickness	numeric	feature	
5	Insulin	numeric	feature	
6	BMI	numeric	feature	
7	DiabetesPedigree	numeric	feature	
8	Age	numeric	feature	
9	Outcome	categorical	target	0, 1

Figure 3. Columns generated from the dataset file and setting the target class (represented in green color).



Sl. No.	Out-come	Pregn-ancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
1	1	06	148	72	35	000	33.6	0.627	50
2	0	01	085	66	29	000	26.6	0.351	31
3	1	08	183	64	00	000	23.3	0.672	32
4	0	01	089	66	23	094	28.1	0.167	21
5	1	00	137	40	35	168	43.1	2.288	33
6	0	05	116	74	00	000	25.6	0.201	30
7	1	03	078	50	32	088	31.0	0.248	26
8	0	10	115	00	00	000	35.3	0.134	29
9	1	02	197	70	45	543	30.5	0.158	53
10	1	08	125	96	00	000	0.00	0.232	54

Figure 4. Columns generated to visualize the data file.

The "File" is visualized using "Linear projection" as shown in Fig. 5(a) and Fig. 5(b). Here, 0 (or blue) colour indicates that a female has non-diabetic and 1 (or red) colour indicates that a female has diabetes.

Note that linear projection is represented in the form of principal component analysis in Fig. 5(a) and circular placement in Fig. 5(b). Fig. 5(a) shows the principal components analysis of all the eight "numeric" type attributes. In general, principal component analysis reduces the dimensionality of the dataset. As seen in Fig. 5(a), the attributes of the dataset increase interpretability without information loss. The Orange platform provides a filter to add or remove the attributes. Moreover, the user can select colour, shape, size and label as per their requirement. Fig. 5(b) shows the circular placement of all the eight attributes. As seen in Fig. 5(b), each attribute is represented in the form of an axis. It shows the very informative attributes that are required to perform the classification.

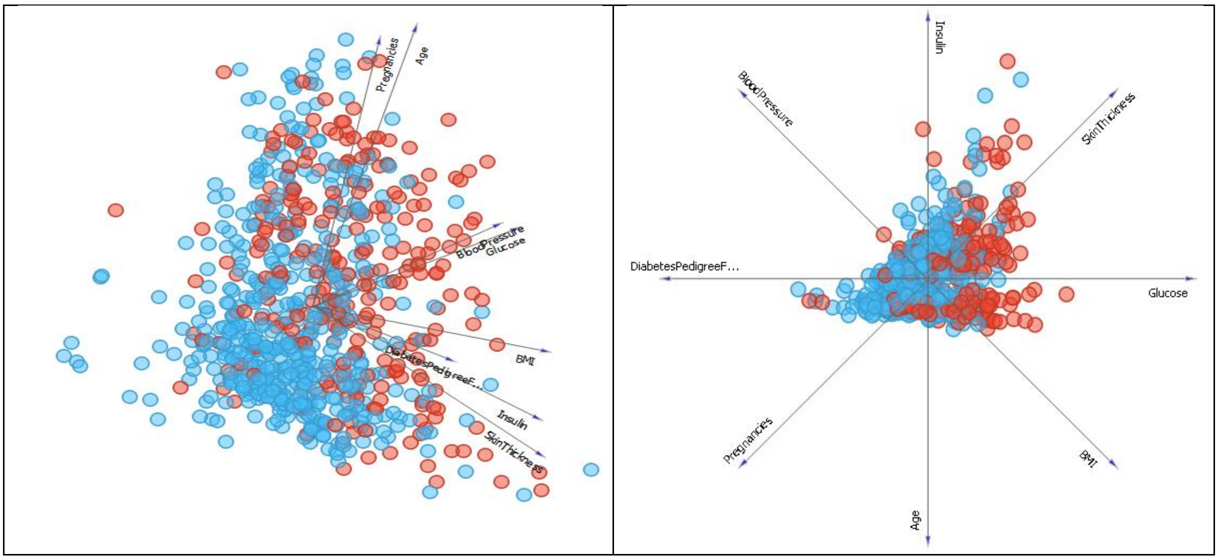
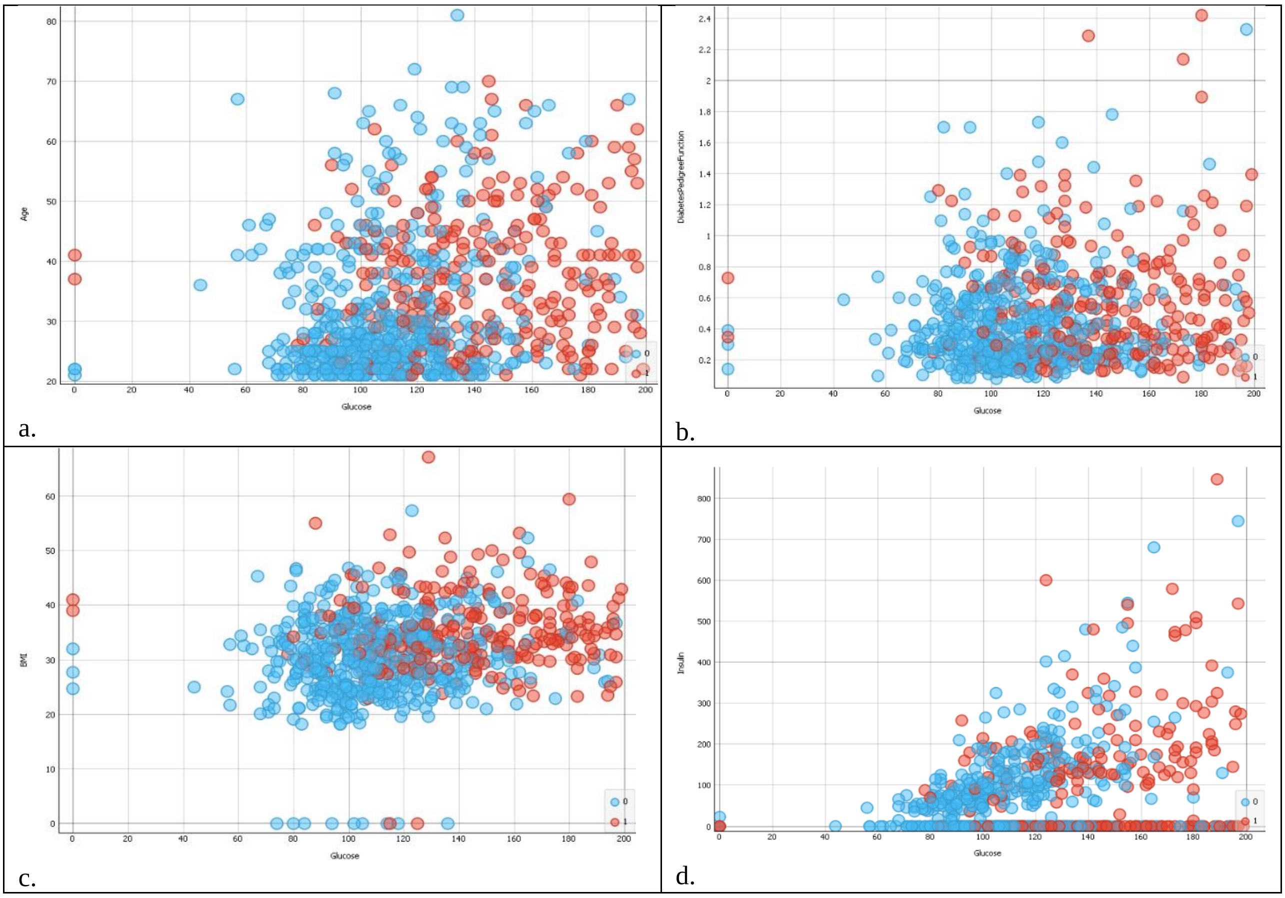


Figure 5. (a) Principal component analysis of all the attributes (blue colour: non-diabetic and red colour: diabetes). **(b)** Linear projection analysis of all the attributes (blue colour: non-diabetic and red colour: diabetes).

Fig. 6 shows the scatter plot of the attributes with respect to glucose level. Here also, blue colour indicates that a female has non-diabetic and red color indicates that a female has diabetes. Note that the visualization of these figures is self-explanatory. The Orange platform provides the option to show the regression line. Table 1 shows the AUC, CA, F1, precision and recall values or scores of different supervised learning algorithms using 10-fold cross-validation. Note that 10-fold cross-validation used, as suggested by various researchers in the field of machine learning. It is clear from Table 1 that LR performs better in terms of CA, i.e., 76.80%. The rationality behind this is that LR divides the linear space into two parts and draws a single line between them. Therefore, LR is a better ML solution for this type of dataset to predict diabetes more effectively.

Table 1. Comparison of supervised learning algorithms using 10-fold cross-validation

Sl no.	Algorithm	AUC	CA	F1	Precision	Recall
1	CT	0.648	0.708	0.703	0.701	0.7082
2	SVM	0.707	0.665	0.671	0.681	0.665
3	k-NN	0.737	0. 711	0. 706	0.703	0.711
4	NB	0.818	0.736	0.739	0.745	0.736
5	RF	0.808	0.754	0.752	0.751	0.754
6	NN	0.824	0.758	0.755	0.754	0.758
7	AB	0.681	0.710	0.710	0.710	0.710
8	LR	0.825	0.768	0.760	0.763	0.768



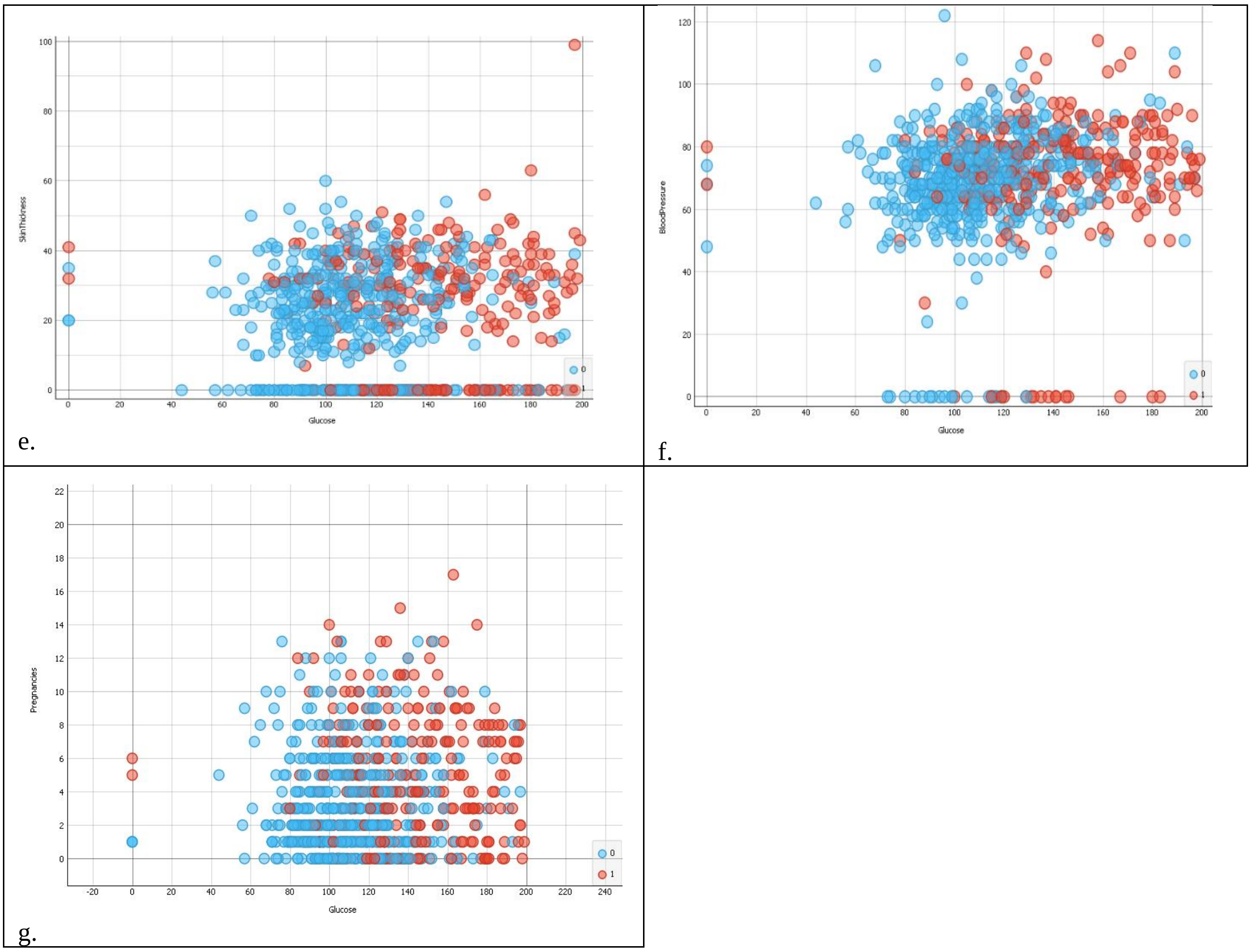
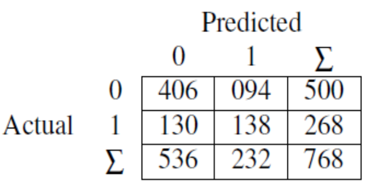


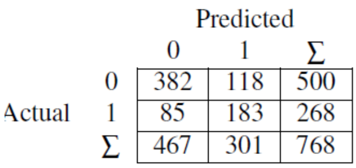
Figure 6. Scatter plot of **(a)** glucose (in x-axis) with respect to age (in y-axis) **(b)** glucose (in x-axis) with respect to diabetes pedigree function (in y-axis) **(c)** glucose (in x-axis) with respect to body mass index (in y-axis) **(d)** glucose (in x-axis) with respect to insulin (in y-axis) **(e)** glucose (in x-axis) with respect to skin thickness (in y-axis) **(f)** glucose (in x-axis) with respect to blood pressure (in y-axis) **(g)** glucose (in x-axis) with respect to pregnancies (in y-axis)

Fig. 7 shows the confusion matrices of different supervised learning algorithms. Here, (0, 0) represents the true positive, (0, 1) represents false positive, (1, 0) represents false negative and (1, 1) represents true negative. The first value represents the actual value and the second value represents the predicted value. A true positive is a kind of outcome where the supervised learning algorithm correctly predicts 222 the positive instances. A false positive is a kind of outcome where the supervised learning algorithm incorrectly predicts the positive instances. A false negative is a kind of outcome where the supervised learning algorithm incorrectly predicts the negative instances. A true negative is a kind of outcome where the supervised learning algorithm correctly predicts the negative instances. Here, the positive instance is a person with diabetes and the negative instance is a person with no diabetes. As seen in Fig. 7(a) to Fig. 7(h), the number of true positive instances is 441 in LR and it is maximum in comparison to other algorithms. Therefore, it is clear from the results that LR is a better ML solution, especially in the used dataset.



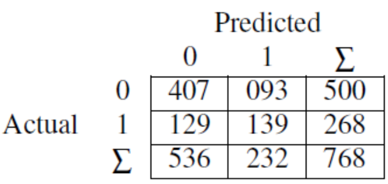
		Predicted		
		0	1	Σ
Actual	0	406	094	500
	1	130	138	268
	Σ	536	232	768

a. CT



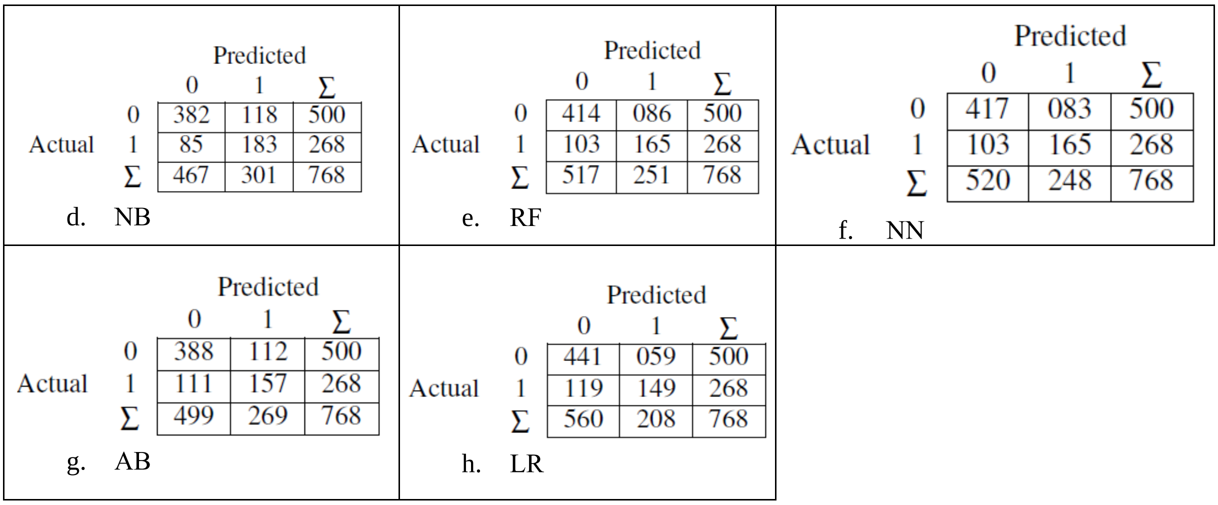
		Predicted		
		0	1	Σ
Actual	0	382	118	500
	1	85	183	268
	Σ	467	301	768

b. SVM



		Predicted		
		0	1	Σ
Actual	0	407	093	500
	1	129	139	268
	Σ	536	232	768

c. k-NN



		Predicted 0	1	Σ
Actual	0	382	118	500
	1	85	183	268
	Σ	467	301	768

d. NB

		Predicted 0	1	Σ
Actual	0	414	086	500
	1	103	165	268
	Σ	517	251	768

e. RF

		Predicted 0	1	Σ
Actual	0	417	083	500
	1	103	165	268
	Σ	520	248	768

f. NN

		Predicted 0	1	Σ
Actual	0	388	112	500
	1	111	157	268
	Σ	499	269	768

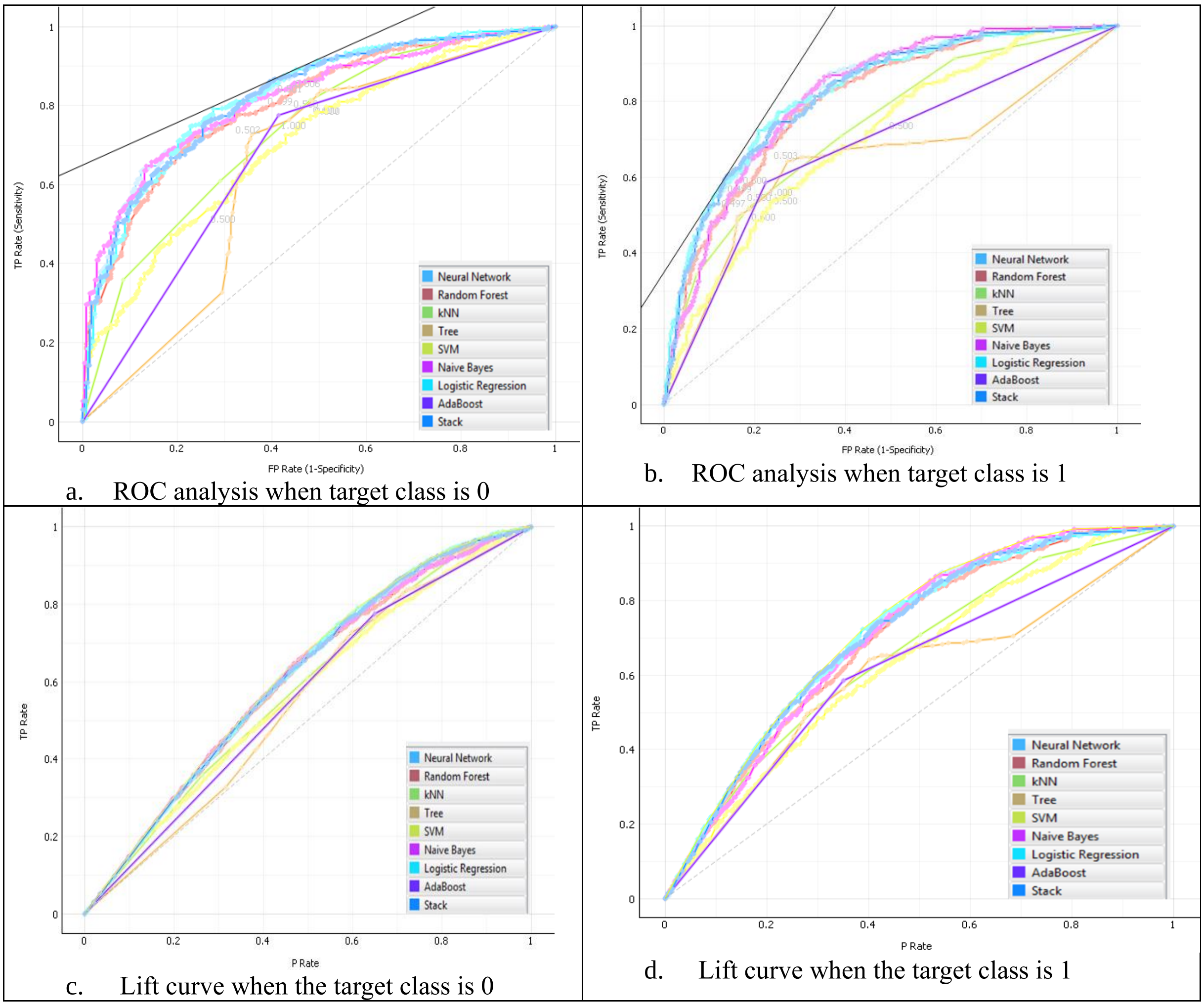
g. AB

		Predicted 0	1	Σ
Actual	0	441	059	500
	1	119	149	268
	Σ	560	208	768

h. LR

Figure 7. Confusion matrix of various supervised learning algorithms.

The ROC analysis of all the algorithms for both target class 0 and target class 1 are separately shown in Fig. 8(a) and Fig. 8(b), respectively. The lift curve analysis of all the algorithms for both target class 0 and target class 1 is separately shown in Fig. 8(c) and Fig. 8(d), respectively. It is clear from Fig. 8(a) and Fig. 8(b) that LR outperforms other algorithms. The rationality behind this is that the number of true positive instances are maximum in the case of LR.



a. ROC analysis when target class is 0

b. ROC analysis when target class is 1

c. Lift curve when the target class is 0

d. Lift curve when the target class is 1

Figure 8. ROC curve analysis and lift curve analysis (x-axis: positive rate and y-axis: true positive rate).

5 CONCLUSIONS

In this paper, we have predicted diabetes in the females by taking the female Pima Indians diabetes dataset. In this dataset, we have taken pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome attributes for performing the prediction. Various supervised learning algorithms have been used such as CT, SVM, k-NN, NB, RF, NN, AB, and LR, and generated the training dataset and testing dataset using k-fold cross-validation with k = 10. The results are compared on the Orange 3.24.1 open-source platform using AUC, CA, F1, precision, and recall parameters. From the comparison, it has been observed that LR performs better in comparison to other algorithms. However, we have not considered any other datasets of diabetes and any other diseases. Moreover, the used dataset has very small and contained limited instances. Therefore, we will consider bigger datasets and apply deep learning algorithms for the same as our future work.

Conflict of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Acknowledgement: Acknowledgement, Funding statement, Conflicts of Interest and References headings should be left justified, bold, not numbered, with the first letter capitalized. Sections following those headings are formatted as the main text. Authors should thank those who contributed to the article but cannot include themselves.

Funding Statement: Authors should describe sources of funding that have supported their work, including specific grant numbers, initials of authors who received the grant, and the URLs to sponsors' websites. If there is no funding support, please write "The author(s) received no specific funding for this study."

References

1. Al Jarullah and A. Asma, "Decision tree discovery for the diagnosis of type II diabetes." 2011 International conference on innovations in information technology. IEEE, pp.303-307, 2011.
2. E. Alpaydin, "Introduction to machine learning," MIT press, 2020.
3. Wantao Wang and Guozhong Sun, "Classification and Research of Skin Lesions Based on Machine Learning," Computers, Materials & Continua, vol. 62, pp. 1187-1200, 2020.
4. W. Xu, Y. Tao, C. Yang and H. Chen, "MSICST: multiple-scenario industrial control system testbed for security research," CMC: Computers, Materials & Continua, vol. 60, no.3, pp. 691-705, 2019.
5. D. Martens, J. Huysmans, R. Setiono, J. Vanthienen and B. Baesens, "Rule extraction from support vector machines: an overview of issues and application in credit scoring." In Rule extraction from support vector machines, Springer, Berlin, Heidelberg, pp. 33-63, 2008.
6. D. Bhulakshmi and G. Gandhi, "The prediction of diabetes in pima indian women mellitus based on xgboost ensemble modeling using data science," Technical report, EasyChair, 2020.
7. W. Chen, S. Chen, H. Zhang and T. Wu, T, "A hybrid prediction model for type 2 diabetes using k-means and decision tree," in 2017 8th IEEE International Conference on Software Engineering
   a. and Service Science (ICSESS), pages 386–390, 2017.
8. Y. A. Christobel and P. Sivaprakasam. "A new class wise k nearest neighbor (cknn) method for the classification of diabetes dataset," International Journal of Engineering and Advanced Technology,
   a. vol. 2, no.3, pp. 396–200, 2013.
9. S.K. Dey, A. Hossain and M.M. Rahman, "Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm," In 21st international conference
   a. of computer and information technology (ICCIT) IEEE, pp–5, 2018.
10. G. George, A.M. Lal, P. Gayathri and N. Mahendran, "Comparative study of machine learning algorithms on prediction of diabetes mellitus disease," Journal of Computational and Theoretical Nanoscience, vol. 17, no. 1, pp. 201–205, 2020.
11. Y. Guo, G. Bai and Y. Hu, "Using bayes network for prediction of type-2 diabetes," in 2012 International Conference for Internet Technology and Secured Transactions, IEEE, pp. 471–472, 2012.
12. J. Han, J.C. Rodriguez and M. Beheshti, "Diabetes data analysis and prediction model discovery using rapidminer," in Second international conference on future generation communication
    a. and networking, IEEE, vol. 3, pp. 96–99, 2008.
13. M. Jahangir, H. Afzal, M. Ahmed, K. Khurshid and R. Nawaz, "An expert system for diabetes prediction using auto tuned multi-layer perceptron," in Intelligent Systems Conference (IntelliSys), IEEE, pp. 722–728.,2017.
14. Kaggle (2020). Pima Indians Diabetes Database. https://www.kaggle.com/uciml/pima-indians-diabetes-database. [Online; accessed 30-January-2020].
15. H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart
    a. diseases. Expert systems with applications," vol. 35, no. 1-2, pp. 82–89, 2008.
16. R. Kala, A. Shukla and R. Tiwari, "Comparative analysis of intelligent hybrid systems for detection of pima indian diabetes," in World Congress on Nature & Biologically Inspired Computing (NaBIC), IEEE, pp. 947–952, 2009.
17. K. Kannadasan, D.R. Edla and V. Kuppili, "Type 2 diabetes data classification using stacked autoencoders in deep neural networks. Clinical Epidemiology and Global Health," vol. 7, no.4, pp. 530–535, 2019.
18. S. Karatsiolis and C.N. Schizas, "Region based support vector machine algorithm for medical diagnosis on pima indian diabetes dataset," in 12th International Conference on Bioinformatics & Bioengineering

(BIBE), IEEE, pp. 139–1442, 2012.

19. A.G. Karegowda, A. Manjunath and M. Jayaram, "Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima indians diabetes," International Journal on Soft Computing, vol.2, no. 2, pp. 15–23.
20. V. Karthikeyani and I.P. Begum, "Comparison a performance of data mining algorithms (cpdma) in prediction of diabetes disease," International journal on computer science and engineering,
    a. Vol. 5, no. 3, pp. 205, 2013.
21. P. Kaur and R. Kaur, "Comparative analysis of classification techniques for diagnosis of Diabetes," in Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals, Springer, pp. 215–221, 2020.
22. A. Kumari, R.K. Behera, K.S Sahoo, A. Nayyar, A. K. Luhach and S.P. Sahoo, "Supervised link prediction using structured-based feature extraction in social network," Concurrency and Computation: Practice and Experience, pp. e5839, 2020.
23. V. A. Kumari and R. Chitra, "Classification of diabetes disease using support vector machine," International Journal of Engineering Research and Applications, vol. 3, no.2, pp. 1797–1801, 2013.
24. S. Lekkas and L. Mikhailov, "Evolving fuzzy medical diagnosis of pima indians diabetes and of dermatological diseases. Artificial Intelligence in Medicine," vol. 50, no. 2, pp.117–126, 2010.
25. L. Li, "Diagnosis of diabetes using a weight-adjusted voting approach," in 2014 IEEE
    a. International Conference on Bioinformatics and Bioengineering, Boca Raton, FL, USA, pp.320–324, 2014.
26. S. Mishra, H.K. Tripathy, P.K. Mallick, A. K. Bhoi and P. Barsocchi, "EAGA-MLP—An Enhanced and Adaptive Hybrid Classification Model for Diabetes Diagnosis," Sensors, vol. 20, no. 14, pp. 4036, 2020.
27. M. Maniruzzaman, M.J. Rahman, B. Ahammed, and M.M Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," Health Information Science and Systems, vol. 8, no. 1, pp. 7, 2020.
28. M. Islam, J. Rahman and D.C. Roy, "Automated detection and classification of diabetes disease based on Bangladesh demography and health survey data, 2011 using machine learning approach," Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 2020.
29. S.K. Nayak and S. K. Panda, "A user-oriented collaborative filtering algorithm for recommender systems," in 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 374–380, 2018.
30. S. Nithya, M. Sangeetha, K.A. Prethi, K. S. Sahoo, S.K. Panda and A.H. Gandomi, "SDCF: A software-defined cyber foraging framework for cloudlet environment," IEEE Transactions on Network and Service Management, 2020.
31. Orange (2020). Orange Data Mining. https://orange.biolab.si/. [Online; accessed 26-January-2020].
32. S. K. Panda, S. K. Bhoi and M. Singh, "A collaborative filtering recommendation algorithm
    a. based on normalization approach," Journal of Ambient Intelligence and Humanized Computing, pp. 1–23, 2020.
33. S. K. Panda, M.R. Senapati and P.K. Sahu, P. K, "An item-oriented collaborative filtering algorithm for recommender systems," 60th Annual Technical Session, pp. 228–23, 2019.
34. H. N. A. Pham and E. Triantaphyllou, "Prediction of diabetes by employing a new data
    a. mining approach which balances fitting and generalization," In Computer and Information Science,
    b. Springer, pp.11–26, 2008.
35. S. Srivastava, L. Sharma, V. Sharma, A. Kumar and H. Darbari, "Prediction of Diabetes Using Artificial Neural Network Approach," in Engineering Vibration, Communication and Information Processing, Springer, Singapore, pp. 679-687, 2019.
36. P. Radha and B. Srinivasan, "Predicting diabetes by cosequencing the various data mining classification techniques," International Journal of Innovative Science, Engineering & Technology, vol. 1, no. 6, pp. 334–339.
37. V. Ravindranath, S. Ramasamy, R. Somula, K.S. Sahoo and A.H. Gandomi, "Swarm intelligence based feature selection for intrusion and detection system in cloud infrastructure," in IEEE Congress on Evolutionary Computation (CEC), pp. 1–6, 2020.
38. K. S. Sahoo, B. K. Tripathy, K. Naik, S. Ramasubbareddy, B. Balusamy, M. Khari and D. Burgos,
    a. "An evolutionary svm model for ddos attack detection in software defined networks," IEEE
    b. Access, vol. 8, pp. 132502–132513, 2020.
39. S.A. Saji and K. Balachandran, "Performance analysis of training algorithms of multilayer perceptrons in diabetes prediction," in International Conference on Advances in Computer Engineering and Applications, IEEE, pp. 201–206, 2015.
40. R. Sanakal and T. Jayakumari, "Prognosis of diabetes using data mining approach-fuzzy c means clustering and support vector machine," International Journal of Computer Trends and Technology, vol. 11, no. 2, pp. 94–8, 2014.
41. L. O. Schulz, P.H. Bennett, E. Ravussin, J.R. Kidd, K.K. Kidd, J. Esparza, and M. E. Valencia, "Effects

of traditional and western environments on prevalence of type 2 diabetes in Pima

a. Indians in Mexico and the US," Diabetes care, vol. 29, no. 8, pp. 1866–1871, 2006.

42. M. Seera and C.P. Lim, C. P, "A hybrid intelligent system for medical data classification. Expert Systems with Applications," pp. 41, no. 5, pp. 2239–2249, 2014.
43. R. Sivanesan and K.D. R. Dhivya, "A review on diabetes mellitus diagnoses using classification on Pima Indian diabetes data set," International Journal of Advance Research in Computer Science and Management Studies, vol. 5, no. 1, 2017.
44. S. Srivastava, L. Sharma, V. Sharma, A. Kumar and H. Darbari, "Prediction of diabetes using artificial neural network approach," in Engineering Vibration, Communication and Information Processing, Springer, pp. 679–687, 2019.
45. J. Thomas, A. Joseph, I. Johnson and J. Thomas, J, "Machine learning approach for diabetes Prediction," International Journal of Information, vol. 8, no. 2, 2019.
46. Q. Wang, W. Cao, J. Guo, J., Ren, Y. Cheng and D. N. Davis, "Dmp mi: an effective
    a. diabetes mellitus classification algorithm on imbalanced data with missing values," IEEE Access,
    b. vol. 7, pp. 102232–10223, 2019.
47. J. Wu, Y.B. Diao, M.L. Li, Y.P Fang and D.C. Ma, "A semi-supervised learning based method: Laplacian support vector machine used in diabetes disease diagnosis," Interdisciplinary Sciences: Computational Life Sciences, vol. 1, no. 2, pp.151–155, 2009.
48. R. Zolfaghari, "Diagnosis of diabetes in female population of pima indian heritage with ensemble of bp neural network and svm," Int. J. Comput. Eng. Manag, vol. 15, pp. 2230–7893, 2012.
49. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, "Predicting diabetes mellitus with machine learning techniques," Frontiers in genetics, vol. 9, pp. 515, 2018.
50. Saeed, Soobia, et al. "Performance Analysis of Machine Learning Algorithm for Healthcare Tools with High Dimension Segmentation." Machine Learning for Healthcare: Handling and Managing Data (2020): 115.
51. Usmani, Raja Sher Afgun, et al. "A spatial feature engineering algorithm for creating air pollution health datasets." International Journal of Cognitive Computing in Engineering 1 (2020): 98-107.
52. Kok, S. H., Abdullah, A., Jhanjhi, N. Z., & Supramaniam, M. (2019). A review of intrusion detection system using machine learning approach. International Journal of Engineering Research and Technology, 12(1), 8-15.