Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

Optimal Machine Learning Model for Diabetes Disease Prediction

Dr. G. Naveen Sundar[1], Dr. D. Narmadha[2*], Dona Davis[3], Helan Selvin[4], Dr. V. Ebenezer[5]
[1,2,3,4,5]Karunya Institute of Technology and Sciences, Coimbatore
naveensundar@karunya.edu
*narmadha@karunya.edu
dona7295@gmail.com
helenselvin12343@gmail.com
ebenezerv@karunya.edu

ABSTRACT

Diabetes is a metabolic disease that is caused as a result of increased blood sugar levels. Elevated levels of blood glucose, influence all vital organs and other organs of the body. To own healthy life, early diabetes detection is very essential. As the occurrence of diabetes is rapidly growing, this condition is a matter of global concern. The exponential development in Machine Learning has been applied to many fields of medical health. We have gathered a dataset called Pima Indian diabetes dataset, followed by feature extraction and evaluation of performance, precision, accuracy, recall, and f1score of various Machine Learning models.

Keywords
Artificial Neural Network, Type II diabetes

1. Introduction

Type 2 diabetes is a customary disease that can be mainly observed in adults and the elderly and merely in children. It is a constant disease that makes the body maintain the insulin level. Type 2 diabetes can be caused by genetic disorders or by an unhealthy lifestyle and food habits. The impediments of diabetes can lead to heart and brain diseases and even can lead to cancers. Hence it's vital to spot the disease at an early stage most accurately. [1-3]

The proposed method is based on machine learning and deep learning which are computational learning techniques. These techniques are derived from Artificial Intelligence (AI) making the system impulsively learn and upgrade based on the experience without human interference. These algorithms work by building a model that eventually takes the input from instances and makes predictions. [4]

The primary step in the machine learning process involves collecting the data from various sources. The dataset being collected is preprocessed by cleaning which involves filling out the missing values with the median value of the corresponding column. The next step is normalization which converts the higher value input to a lower value within the range of 0 to 1. Then the unnecessary data fields are eliminated using feature extraction and selection algorithms for making the model more accurate. Since the prepared dataset is too large, it is split into two as training and testing datasets one for training and another for evaluating the efficiency of the model. Several machine learning and deep learning models are constructed and the training dataset is passed to analyze the data based on previous knowledge. The final step is passing the test dataset to estimate the accuracy and performance of the model. Machine learning has various applications in prediction, classification, and image recognition like disease predictions, email filtering, character recognition, malware detection, proctoring, traffic management, etc. [5]

Our preferred method relies on deep learning cognition in python. The dataset being used is Pima Indians Diabetes Dataset(PIDD) which includes medical predictive attributes like Pregnancies,

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

Glucose, BloodPressure, SkinThickness, insulin, BMI, DiabetesPedigreeFunction, and Age, and one target attribute called Outcome. The overview of the dataset is portrayed in Fig.1. We worked with the machine learning models like Support Vector Classifier(SVC), Logistic Regression, Decision Tree, Gradient Boost, Random Forest, Naïve Bayes, and K-Nearest Neighbour(KNN) and deep learning model Artificial Neural Network(ANN). The execution of these models is judged for predicting diabetes in Pima Indian women.

2. Literature Review

The Pima Indian Diabetes Dataset(PIDD) being acquired from the UCI machine learning repository, emerged from the National Institute of Diabetes and Digestive and Kidney Diseases, which is collected to accurately speculate either the person is diabetic or not based on the features available in the dataset. The dataset consists of 9 columns and 768 instances and its details can be viewed in Figure.1. This dataset constitutes only female patients of age greater than 21 years old. It consists of the forecaster variables like Pregnancies, which depicts how many pregnancies the patient had, their Glucose level, BloodPressure, SkinThickness, BMI, Insulin level, DiabetesPedigreeFunction, and Age, and one target variable known as Outcome having values 0 or 1 in which 0 concludes the person is non-diabetic and 1 concludes diabetic. The 500 instances are 0 and 268 instances depict 1, evident in Figure.2.

0	Pregnancies	768 non-null	int64
1	Glucose	768 non-null	int64
2	BloodPressure	768 non-null	int64
3	SkinThickness	768 non-null	int64
4	Insulin	768 non-null	int64
5	BMI	768 non-null	float64
6	DiabetesPedigreeFunction	768 non-null	float64
7	Age	768 non-null	int64
8	Outcome	768 non-null	int64

Figure 1. Information about Pima Indian Diabetes Dataset

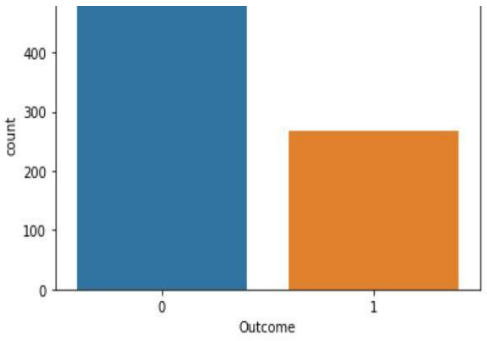

Figure 2. Outcome variable description

We evaluated this particular dataset with the help of Python libraries. Feature extraction, the most predominant section of our proposed model is applied for picking out the best attributes that contribute to predicting the target variable most accurately and efficiently. We accomplished this procedure by envisioning the correlation matrix of the attributes using a function called heatmap that will statistically report the association among variables. The heatmap is portrayed as a graph of data in which the corresponding values are displayed in colors for better readability of values,

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

as in Fig.9. This surrendered 5 dominant features contributing to the better enhancing of the model and those are used for training the model.

Numerous machine learning procedures exist for implementing the process of training and testing the model. The considerably used machine learning mechanisms are supervised and unsupervised. The supervised learning technique can be implemented in particular where we can acquire antecedent data based on the experience. Supervised techniques include algorithms like Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, Artificial Neural Network, etc. And the unsupervised machine learning technique can be applied in case of training date with unlabeled attributes. Principle Component Analysis, k-means clustering, hierarchical clustering, and hidden Markov model are some of the widely used unsupervised learning modes.

The supervised machine learning algorithms like Support Vector Classifier, Decision Tree, Random Forest, k-Nearest neighbor, Logistic Regression, Naïve Bayes, Gradient Boost, and Artificial Neural Network are chosen to predict whether the Pima Indians in the dataset is affected by diabetes or not.

2.1 Support Vector Classifier

A support vector classifier(SVC) belongs to a supervised algorithm that can be implemented in multi-class classification. It finds the best fit hyperplane in N-dimensional(N-number of features) space which classifies data in distinct points. SVC focuses on maximizing the gap between the hyperplane and the two classes which is suitable for the dataset which can be linearly classified as in Figure.3. SVC has several regulatable parameters like C known as the retribution parameter for error that is 1 by default. A low value of C gives the more accurate hyperplane and with an increasing C value, the variance is also increased in the classifier. The hypothesis function h can be expressed as follows:

$$h(x1) = \begin{cases} +1 \; if \; a.x + b \geq 0 \\ -1 \; if \; a.x + b < 0 \end{cases}$$

The point on the hyperplane and above is considered as class +1 and below it is considered as -1. [6,7]

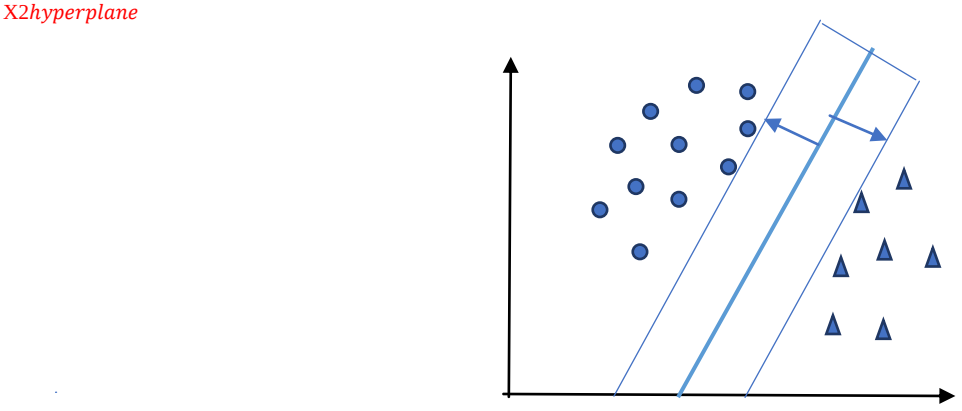

Figure 3. Maximum margin hyperplane

2.2 Decision Tree

A decision tree classifier is a predictive model that performs the classification of instances depending on their attribute values starting from the root node splitting into subsets to generate

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

subtrees as in Figure.4. This type of classifier learns in a tree system whose values can be noted as a pile of rules making it easier to understand. A decision tree consists of a root node, many internal and leaf nodes, and edges. The internal nodes of a typical tree represent features, leaf nodes indicate the classes which are to be allotted to a sample, and the edges represent the concurrence of features that are leading to classification. [8].

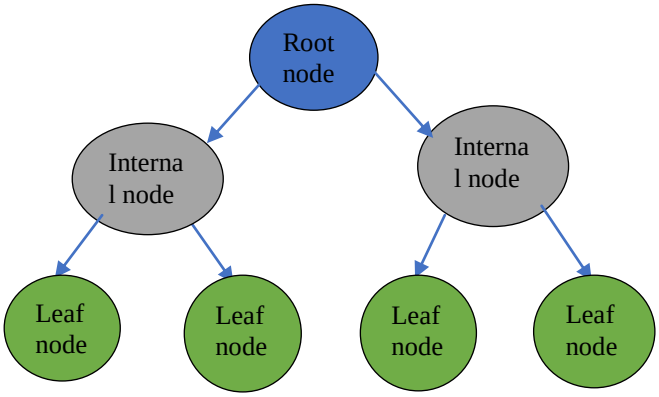

Figure 4. Decision tree structure

The parents in each hierarchy are determined by two criteria Gini index and Entropy.[9]

Gini index: $G(Attribute) = 1 - \sum_{j=1}^{C} P_j^2$

Entropy: $H(Attribute) = -\sum_{j=1}^{C} P_j \log_{10} P_j$

2.3 Random Forest

A random forest(RF) algorithm is a supervised machine learning approach that can be applied both in classification and regression processes. A random forest classifier is the collection of numerous decision trees within it. [9]. They imprecisely produce a large number of decision trees making it to control the election of the best-suited attribute in every node to a fairly tiny subset of randomly selected attributes as shown in Figure.5. Every decision tree in the random forest is created from the training set and is used for classifying new instances through a consistent selection process. [10].
The equation of the RF tree is:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x')$$

$\hat{f}$ represents terminal tree prediction; B represents the aggregate number of trees; b represents the current tree; x' represents the training set.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

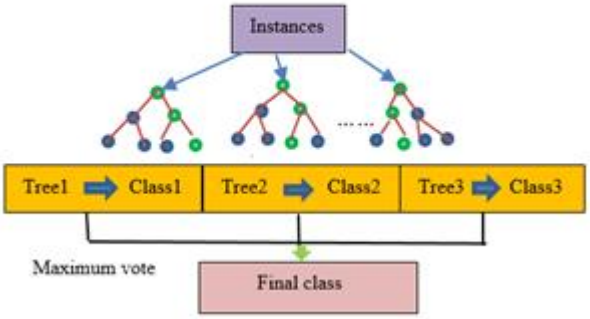

Figure 5. Random Forest structure

2.4 k- Nearest neighbor (k-NN)

A k-Nearest neighbor algorithm is a machine learning technique that can be used for the classification of different classes. k-NN classifiers are heavily used because they make it easier to predict the output, takes less calculation time, and have a stronger predictive power. k-NN classifier prefers to take the vote from its nearest neighbor by a distance function as in Figure.6. This algorithm has a parameter called K which has a crucial role in segregating the classes, different values of K are used to find the best segregation. The K value that best separates classes is chosen and is used for all predictions. [11].

k-NN classifies data by finding the distance between the nearest neighbor that can be calculated using Euclidean distance, Manhattan distance, and Hamming distance. The equation of Euclidean function to compute distance is as follows:

$$d(a,b) = d(b,a)$$

$$= \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + \cdots + (b_n - a_n)^2}$$

$$= \sqrt{\sum_{i=0}^{n} (b_i - a_i)^2}$$

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

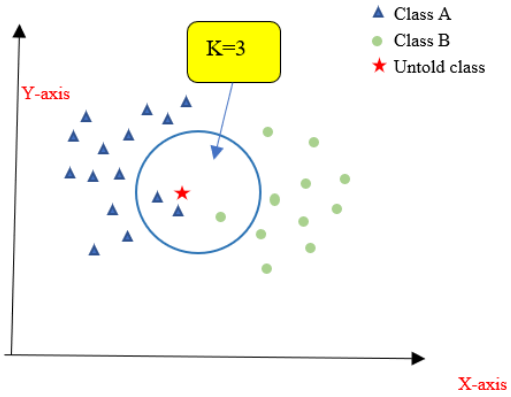

Figure 6. k-NN Classifier

2.5. Gradient Boost Classifier

Gradient Boost is an ensemble learning boosting algorithm that can be used for classification. This algorithm treats bias-variance compromise unlike bagging algorithms and hence considered to be most efficient. The ensemble is the key technique for the working of boosting algorithms and it is achieved by merging the weak learners and producing improvised prediction accuracy by constructing a strong model. The GradientBoost parameters are 3 subcategories: tree-specific parameters focusing on trees constituted, boosting parameters focus on the boosting function, andmiscellaneous parameters for inclusive operation. [12].

The probability of finding the existence of weak learners is as follows:

$$P(surviving) = \frac{e^{\log(odds)}}{1 + e^{\log(odds)}}$$

log(odds) is similar to average in classification problems.

2.6. Artificial Neural Network

Artificial neural networks are prophesying methods that are based on mathematical models of the brain. It works like the way the human brain processes information. ANN detects relationships and patterns from the data and trains through experience to gather information. ANN includes a large number of connected processing components called artificial neurons and works together to develop information. The artificial neurons are connected with the coefficients, which create neural structure and are layered.Each processing element has weighted inputs, one output, and transfer functions. ANN structure is divided into layers; here data moves from the first layer and reaches the middle layers then finally the output, each layer converts the data to specific information and at last gives the valid output. [13].

The activation function or transfer function is a function that translates the input signals to output signals. The weighted inputs are summed up by this function as shown in Figure.7. [14].

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

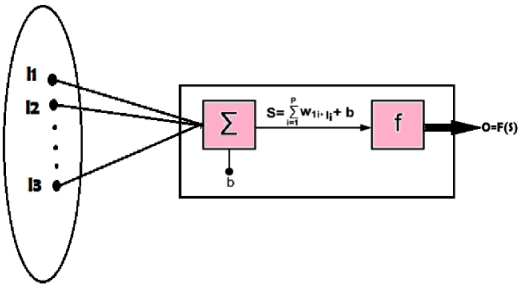

Figure 7.ANN Model

2.7 Logistic Regression

Logistic regression is a supervised learning classification Machine Learning algorithm that is categorical. It gives outputs that can be binary that the data converted to either 1 (denotes success) or 0 (denotes failure) or multi-class. Mathematically, this model predicts P(Y=1) as a function of X. For prediction purposes, first, we train the dataset in the Logistic Regression model on (X_train, y_train), and then we evaluate the model generated using (X_test, y_test). Finally, build the Logistic Regression model and predict for X_test and compare the prediction to the y_test. [15].

2.8 Naïve Bayes Classifier

Naïve Bayes Classifier included in the family, probabilistic classifiers is a classification algorithm that considers features as unrelated and independent and it predicts according to the probability of an object**.** It works based on Bayes Theorem to calculate the target class's posterior probability P(B/C). [16]

P(B|C) = (P(C|B) P(B))/P(C)
where ,P(C|B) = probability of predictor class's.
P(B) = class B's probability is true.
P(C) = prior probability of predictor

Firstly, the dataset is converted into frequency tables. Then,a table is generated by finding the probabilities of each feature present. At last, the Bayes theorem is used to find the posterior probability. There are three types of Naïve Bayes classifier named Gaussian, Multinomial, and Bernoulli.

3. Predictive Model

In the proposed model Figure.8. the collected data is first visualized. visualization is a technique used to grasp the structure of data and its relations. This data is cleaned to fill the null values followed by normalization of data to convert values to a small range of 0-1. Then the proportionalities of the attributes of the data are identified by the correlation coefficient. They help to know which attributes are highly dependent on the prediction variable. Later we used the heatmap feature extraction technique for getting better results. The datais branched to training and test data to reduce complexity. Finally, data is passed through the Machine Learning prediction model and the accuracy is evaluated.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

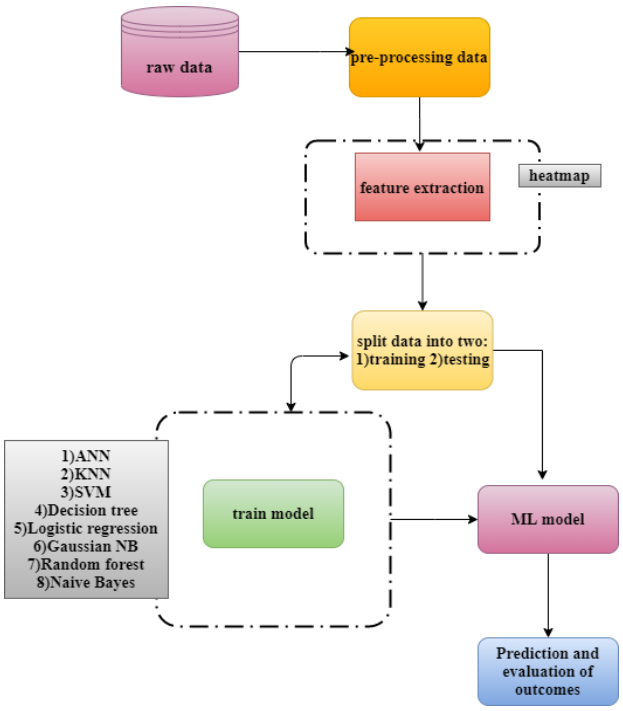

Figure 8. The framework of the proposed model

4. Results and Discussion

Using different supervised ML algorithms we developed various diabetes, prediction models. These models developed is applied to the Pima Indian dataset, which contains data of whether a patient will have an outbreak of diabetes within positive years includes the times of pregnancy, plasma glucose count, blood pressure(mmHg), BMI (weight in kg/(height in m)^2), insulin (mu U/ml), skin thickness (mm), diabetes pedigree function, age (years) and the Outcome (1 for tested positive for diabetes and 0 for tested negative for diabetes).[17] The feature extraction technique called heatmap has been applied to identify which attributes are vastly related to the target variable, we have plotted the heatmap of related features as shown in Figure.9.

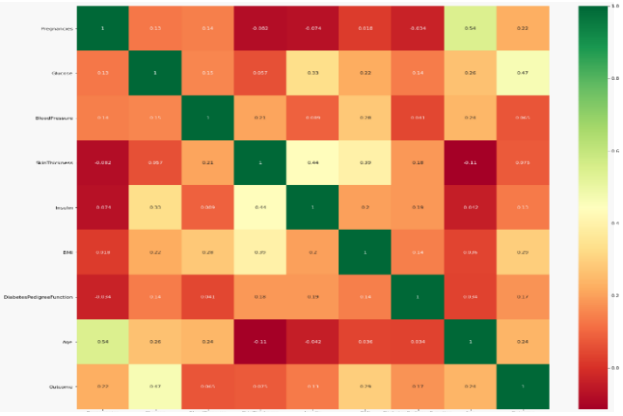

Figure 9. Heatmap Feature Extraction

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

The feature selection reduced the complexity of the dataset by reducing the column numbers from 8 to 5. The five columns are glucose concentration, insulin, BMI, diabetes pedigree function, and diastolic blood pressure. Eight predictive models are developed to detect whether a person is diabetic or not, using ML supervised algorithms with optimized hyperparameters to tune as in Table.1.

Table.1. Tuning Parameters

S.NO.	Prediction model	Tune parameters
1	ANN	max_iter=1000, alpha=1
2	KNN	n_neighbors=14
3	SVM	C=0.2
4	Logistic regression	C=100
5	Decision tree	max_depth=6
6	Random forest	random_state=2, n_estimators=100
7	Gradient boosting	learning_rate=0.5, max_depth=6, max_features=0.4

To detect diabetes in Pima Indians at an early stage we have estimated the model using distinct metrics such as accuracy, precision, recall, and F1 score as in Table.2. Accuracy can be used in classification models to find better predictions. It can be computed byusing the formula given below:

$$Accuracy = \frac{number\ of\ right\ pedictions}{total\ number\ of\ predictions}$$

That is:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TotalSample}}$$

TP is True Positive depicting that what we have predicted to be diabetic and the actual prediction is also diabetic, and TN is true Negative depicting what we have predicted to be non-diabetic and

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

the actual prediction is also non-diabetic. Accuracy has two classifications, training and testing accuracy.

From Table.2, it is evident that the accuracy of ANN is 0.88 and of Logistic Regression is 0.87. For SVC and Random Forest, the accuracy is found to be 0.84 while for k-NN and Naïve Bayes it is 0.83. The GradientBoost is having an accuracy of 0.80 and that of the Decision Tree is 0.77.

Precision can be concluded by fractionating positive results and the total number of positive results made by the classifier. It is calculated by using the following formula:

$$Precision = \frac{TP}{TP + FP}$$

FP is a False Positive depicting that we have predicted to be diabetic but the actual prediction came to be non-diabetic.

As in Table.2, the precision of the ANN classifier is observed as 0.86, that of Logistic Regression is 0.85. SVC is having a precision of 0.81. It is found to be 0.78 in k-NN and 0.76 in Naïve Bayes. Random forest is having a precision of 0.75. 0.73 and 0.74 are the precision for GradientBoost and Decision Tree classier respectively.

Recall can also be termed as sensitivity, measuring true positive to total right predictions made. The method to calculate recall is as follows:

$$Recall = \frac{TP}{TP + FN}$$

FN is False Negative that we have predicted the output to be non-diabetic and the actual output was diabetic.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

Table 2. Tabulation of Results

S.NO.	ML models	Evaluated without feature extraction				Evaluated with feature extraction			
		accuracy	precision	recall	F1 score	accuracy	precision	recall	F1 score
1	AN	0.85	0.82	0.73	0.77	0.88	0.86	0.76	0.81
2	KNN	0.84	0.81	0.73	0.77	0.83	0.78	0.69	0.73
3	SVM	0.85	0.82	0.73	0.77	0.84	0.81	0.69	0.75
4	Logistic regression	0.87	0.86	0.73	0.79	0.87	0.85	0.76	0.81
5	Decision tree	0.81	0.73	0.73	0.73	0.77	0.64	0.76	0.7
6	Naïve Bayes classifier	0.83	0.76	0.86	0.75	0.83	0.76	0.73	0.74
7.	Random Forest	0.83	0.74	0.76	0.75	0.84	0.75	0.8	0.77
8	Gradient boosting	0.83	0.7	0.84	0.77	0.8	0.73	0.65	0.69

From Table.2, we can observe that ANN and Logistic Regression are having a recall value of 0.76. That of SVC and k-NN is 0.69. Naïve Bayes is found to have a recall value of 0.73. The Random Forest, GradientBoost, and Decision Tree have 0.80, 0.65, and 0.76 respectively.

F1 score, calculated by finding the Harmonic mean of precision and recall, quantifies the test's accuracy. It describes how precise the system is. [18]. The formula is as follows:

$$F1\ score = 2 * \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$$

In Table2, the F1 score of ANN and Logistic Regression is observed as 0.81. Tree classifiers like Decision Tree and Random Forest have an F1 measure of 0.70 and 0.77 respectively. The classifiers like SVC, k-NN, and GradientBoost are spotted having an F1 score of 0.75, 0.73, and 0.69 respectively. Naïve Bayes has an F1 measure, 0.74.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

Thereby it is apparent that the ANN system proved as the finest model envisioning the person is diabetic or not. However, the Logistic Regression is discovered to have an accuracy of 0.87 which is lesser than that of the ANN model. Since the outcome class outlines 500 instances to be non-diabetic and 268 to diabetic there is an imbalance in the data. Hence it is advised to focus on other evaluation parameters also rather than pointing to a single metric accuracy. Here, since all the metrics checked for ANN and Logistic Regression are almost producing the same results but the accuracy and precision being lesser by 1% for Logistic Regression than ANN, it can be concluded that ANN can be used as an optimal system to predict diabetes. The accuracies evaluated before and after feature extraction in different machine learning and deep learning algorithms are portrayed in Figure.11 and Figure.12 respectively.

Figure.10 depicts the confusion matrix of the ANN MLP classifier. The value in the first row first column is TN(True Negative) that is the model says 48 numbers of data from the dataset are true and Negative (0 as 0). The 3 numbers in the first-row second column specify the FP(False Positive) value that is 3 numbers of data is positive and is false (0 as 1). In the next row first column we have 6 numbers of data as FN(False Negative) that is it is negative and false (1 as 0). The second column of the second row has 20 numbers of data as TP(True Positive) that is positive and true (1 as 1).

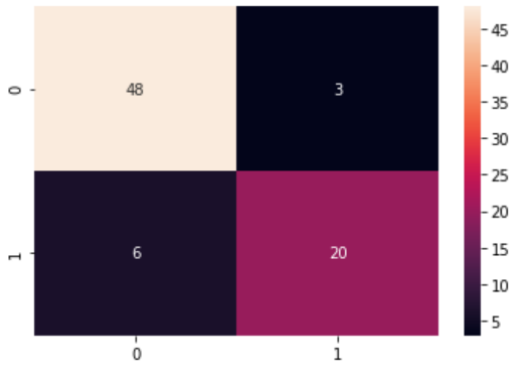

Figure 10. The confusion matrix of ANN MLP classifier using the test dataset

Figure 11. Comparing accuracy of different machine learning and deep learning algorithms before feature extraction. (ANN(85%), KNN(84%), SVM(85%), Logistic Regression(87%), Decision Tree(81%), Naïve Bayes(83%), Random Forest(83%), Gradient Boosting(83%))

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

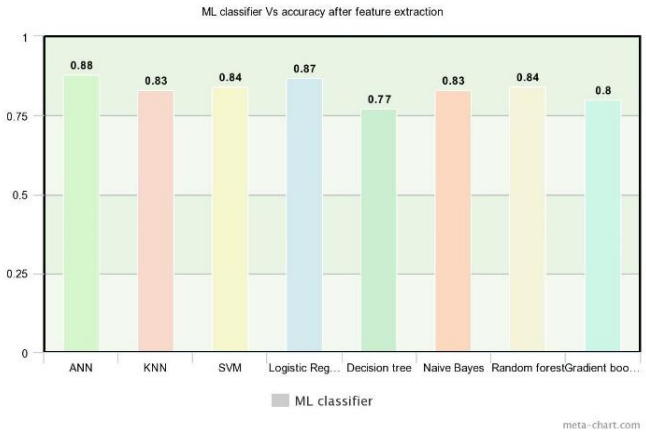

Figure 12. Comparing accuracy of different machine learning and deep learning algorithms after feature extraction. (ANN(88%), KNN(83%), SVM(84%), Logistic Regression(87%), Decision Tree(77%), Naïve Bayes(83%), Random Forest(84%), Gradient Boosting(80%))

5. Conclusion

Diabetes is a chronic disease that causes many health issues to become a major concern to the medical field. In this research work, we used the Machine Learning approach for the prediction of chronic disease. We have developed eight prediction models using SVM, Random Forest, k-NN, Gradient boosting, ANN, Decision tree, Naïve Bayes classifier, and Logistic regression Machine Learning algorithms. The selection of the best among these eight models was done by evaluation and comparison of parameters such as accuracy, precision, recall, and f1 score. We concluded that ANN which has an MLP classifier to feed forward is the best model for diabetes prediction.

The feature extraction using the heatmap feature extraction technique is applied to the dataset for reducing the complexity of the data and for better results. Thus we used a minimal amount of data and produced more accurate and precise results.

References

[**Error! Reference source not found.**] WebMD Medical Reference, Type 2 diabetes (2019)

[2] Leslie J Baier, Robert L Hanson, Genetic studies of the etiology of type 2 diabetes in Pima Indians (2004)

[3] Yanlng Wu, Yanping Ding, Yoshimasa Tanaka, Wen Zhang, Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention (2014)

[4] Annina Simon, Mahima Singh Deo, Venkatesan Selvam, Ramesh Babu, An overview of machine learning and it's applications (2016)

[5] Harleen Kaur, Vinita Kumari, Predictive modelling and analytics for diabetes using a machine learning approach, (2018)

[6] Rohith Gandhi, Support Vector Machine-Introduction to machine learning algorithms, (2018)

[7] Ben Alex Keen, Support vector classifiers in python using scikit-learn, (2017)

[8] Cheng-Jin Du, Da-Wen Sun, Computer vision technology for food quality evaluation, (2008)

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 14925 - 14938
Received 05 March 2021; Accepted 01 April 2021.

[9] Christoph Reinders, Bodo Rosenhahn, Learning convolutional neural networks for object detection with very little training data, (2019)

[10] Igor Kononenko, MatjazKukar, machine learning basics, (2007)

[11] Tavish Srivastava, Introduction to k-Nearest neighbors: a powerful machine learning algorithm(with implementation in python and R), (2018)

[12] Aarshay Jain, Complete machine learning guide to parameter tuning in Gradient Boosting(GBM) in python, (2016)

[13] SnezanaKustrin, Rosemary Beresford, Journal of Pharmaceutical and Biomedical Analysis, (2000)

[14] Jalpa Shah, Biswajit Mishra, Analytical Equations based Prediction approach for PM2.5 using Artificial Neural Network, (2020)

[15] Changsheng Zhu, Christian Idemudia, Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques, (2019)

[16]Deepti Sisodia, Dilip Singh Sisodia, Prediction of Diabetes using classification Algorithms, (2018)

[17] pima-indian-diabetes dataset, Kaggle

[18] Aditya Mishra, Metrics to evaluate your machine learning algorithm, (2018)