ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/353694753

Comparative Analysis of Different ML Classification Algorithms with Diabetes Prediction through Pima Indian Diabetics Dataset

Conference Paper · June 2021

DOI: 10.1109/CONIT51480.2021.9498361

CITATIONS
22

READS
172

2 authors:

Vijaykumar Patil
Bharati Vidyapeeth College of Engineering Navi Mumbai
22 PUBLICATIONS **124** CITATIONS
SEE PROFILE

Dayanand Ragho Ingle
Bharati Vidyapeeth College of Engineering Navi Mumbai
35 PUBLICATIONS **99** CITATIONS
SEE PROFILE

All content following this page was uploaded by Dayanand Ragho Ingle on 13 October 2023.

The user has requested enhancement of the downloaded file.

2021 International Conference on Intelligent Technologies (CONIT)
Karnataka, India. June 25-27, 2021

Comparative Analysis of Different ML Classification Algorithms with Diabetes Prediction through Pima Indian Diabetics Dataset

Vijaykumar Patil
Department of Computer Engineering,
Bharati Vidyapeeth College of Engineering,
Navi Mumbai, India
vijay.patil.karad@gmail.com

Dr. D. R. Ingle
Department of Computer Engineering,
Bharati Vidyapeeth College of Engineering,
Navi Mumbai, India
vijay.patil.karad@gmail.com

***Abstract*— Diabetes is a disease which is increasing in occurrence day by day worldwide. It triggered due the expanded level of sugar within the blood. This type of disease could be a chronic that happens when the patient's body isn't ready to prepare proper insulin level, or also when the body hasn't utilized it. The various computer or electronic based systems were delineated employing miscellaneous classification techniques for anticipating and diagnosing diabetes. Selecting appropriate classifiers from set of various classification algorithms from machine learning family is an undoubtedly expands the accuracy and expertise of the system. This paper proposed an implementation and comparison statistics of well-known supervised ML classification methods such as K-NN, Logistic Regression which based on Regression problem, Naïve Bayes probabilistic classifier, SVM with both linear and non-linear kernel, Decision Tree with Random Forest classifier for statistical modelling and accuracy verification. The dataset downloaded from kaggle.com, it is a Pima Indians Diabetes Database which includes 9 different attributes and 768 records. The maximum accuracy obtained for Machine Learning classification algorithms is 80.20% which fall under very good category model. The CAP Curve Analysis is another performance measure which shows 92.26 % accuracy by Logistic Regression. The ANN framework is designed to improve reliability on system which designed to computer-based diagnosis. The proposed research compares accuracy of ML classification algorithms and ANN, in which ANN having noteworthy improvement in accuracy which is around 97.66 %.**

Keywords — Diabetes, Machine Learning (ML), K-NN, SVM, ANN

I. INTRODUCTION

Diabetes could be an unrelenting metabolic syndrome, in such diseases the inappropriate or irregularity in organization of the blood glucose level, so this will the prime chance of different infections like renal failure, kidney illness and heart attack. The well-known types of the diabetes can be primarily categorized into three different levels, having forms like Type 1(T-I), Type 2(T-II) and Gestational type of diabetes i.e. GDM [4]. The T-I too called as Adolescent Onset type of diabetes is arises when the individuals body declined to produce insulin [2]. The T-II diabetes disorder is depicted by body when it resists to generate insulin and also individuals' body not utilize it [4]. From this T1 diabetes can arrives at anything stage in lifecycle however happens most each presently and over in young people and developmental a long time. T-II diabetes is more regular in grown-ups and means around 90% of all diabetes cases. The pregnant women generally found the GDM type of diabetics. This type may be a type of diabetes that integrates of elevated blood glucose only during pregnancy time and is related with their issues only with mother but some time cases with kid as well. GDM as a run the show vanishes after pregnancy in any case ladies preferential, and their youths are at expanded peril of making type 2 diabetes at some point down the street. Diabetes impacts the human body; it can be a badly unending disorder that bearings the individual body by weakening the attack into the platelets if body permits on glucose or sugar level. Diabetic patients show up mishap of weight, obscured vision, contaminations, visit urination etc. This paper is going to attempts to upgrade the symptomatic precision of the proceeding glucose levels of individuals blood within the diabetic disorder patients, at first by development of Exploratory Data Investigation and Machine Learning classification algorithms. The proposed system makes blood glucose prediction using dataset downloaded from kaggle.com it is a Pima Indians Diabetes Database which includes various nine different attributes related to life style of an individual.

The different frameworks designed for health, which process huge amount of different patient's data related to various types of diseases where this process of extricating is done using data mining and it can be covered up different diseases related patterns. Such frameworks designed to distinguish the relationship between clinical data and also the variations within it. [3]. In diabetes patient, any instance the sugar level within an individuals' body can't be controlled, it could be a genuine health issue in which the degree of sugar in blood varies according daily routing and eating habit of an individual. Regardless the leading reason of the sugar unbalancing in reality that the different variables or factors like genetic factor, height and weight of an individual and affront moreover clasps an obvious portion for influencing diabetes. The early prediction of such diseases helps to an individual with age, that they may help it to live better life and do preclusions accordingly [5][6][7]. There are different DM or ML techniques developed various frameworks to support such preclusive decision to avoid such illness. The selection of such decision-making system that ampleness by its accuracy. To predict and analyze such type of aging disabilities, the decision-making framework are designed which proposes most prominent degree of precision.

Machine Learning based framework used to perform analysis which extract appreciable evidences and different hidden patterns from the given dataset for decision supportive prediction. The proposed work performs the univariant and bivariant analysis initially with the attributes

like age and BMI. The diabetes philosophy is named as the decisions and associations between the individual thoughts got from medical services territories. This diabetes learning commitments for associating with other conventional therapeutic administrations settings. There can be a few issues for distinguishing the kind of diabetes due some unacceptable passages or absence of understanding information about patients [8]. Dithering or vulnerability can likewise be set off because of the more prominent sizes of information. Regularly, surmising or decisions about a particular pollutant are made by the expert.

In this paper the implementation and comparison statistics of Machine Learning techniques or algo such as K-NN, SVM with linear and non-linear-Kernel SVM, Naïve Bayes, Logistic Regression and also the Decision Tree and Random Forest Classification for accuracy validation with statistical modelling. The CAP curves are drawn for each model. The Artificial Neural Network also build as prediction model and compare its accuracy with above different Machine Learning Models. The designing of data analysis framework which gives higher precision is the primary objective of this proposed research paper is to set up a for the classification or predicting has individual is having diabetes or not. The same dataset utilizes to does comparisons of the ML and neural-network model classification with counterfeit are designed. The accuracy or precision comparison comes about appeared generally made strides exactness, which demonstrates the adequacy of this proposed framework. The remaining research paper presented as Section II contains brief picture about the Related Work from similar domain. The Material and Implementation Methods includes in next Section III. Section IV includes Results and Discussion it includes brief about the statistical variances of various ML classifiers which are used for later in prediction. Finally, conclusion and future scope that lay down the outcomes of proposed work.

II. Related Work

The proposed framework designed two distinct approach RBFNN and MLP approach it might be capable show up the blood sugar, and it outflanked [1]. It designed with combination of different wavelet layers with the WNN and it can be controlled with wavelet limits an also can catch the lead of the turbulent time-sharing. The MLP and RBFNN designed with sigmoid and Gaussian functions respectively. The PCA used as promising exploratory analysis, likely it contributes by the info affirmation. It firmly used to view the different factors from dataset that blood glucose and helps it in design of prediction framework [1]. But it has problem with the performance issues. PCA consider the historic records from the dataset, and having range up to 36 hrs to draw prediction or result.

The system was proposed by [2] that businesses AdaBoost algo with decision support as a base classifier for classification [2]. Also, SVM, Naïve Bayes and Decision Tree are moreover reified as base classifiers for AdaBoost order for precision affirmation. The proposed methodology by [2] used the base classifier to design new optimized one called as AdaBoost having commercial or businesses approach as decision support for classification. The design of AdaBoost is a refinement of SVM, Naïve Bayes and Decision Tree as the base classifiers for it. The best performance is given by AdaBoost with Decision Tree used as the base classifier and having 80.72% accuracy and which is more than other as Naïve Bayes and SVM.

The diabetes prediction can be proposed in [4] with different clustering and classification methods here the SVM was applied to analyze diabetes on the Pima Indian diabetes dataset. [4] It proposes Versatile Neuro Fluffy Deduction Classification, that offer a high precision for the conclusion of diabetes and expectation of cancer. [4] too gives the prove approximately the precision of Naïve Bayes K –means classifier. The accuracy or exactness picked up by these approaches is around 80 %.

The proposed framework in [9] design the hereditary structure known as Belief–Intensification, Loyd's H implies+ algo and gained. They also used the Decision tree classifier for accuracy verification of sugar level in diabetes patients, it achieved accuracy round 78.17% [9]. For similar kind of classification [9] also implement ANN with back propagation with verify accuracy and performance of classification model.

Amalgam KNN-a designed in [10] as despising support system for diabetes with multistep pre-preparing, it has used the KNN and K-means with cross variety [10].The promising results are shown by K means with the surrounding of KNN classifier algorithm [10]. Framework design in [8] with half variety Hereditary Algorithm, Fluffy Logic and Information calculus principally focus on Neural Network. The NN designed here having feathery rationale to use for end with the innovative diabetic dataset, it has major contribution and assumption supported in the design of neural network.; and to surmise alliance between unmistakable indications of diabetics and finally precision is surveyed by crossbreed inherited calculation. [8] besides shows practically the limit and constringent of neural network and imprecision in outcomes when is data contains upheaval segments.

Life-threatening learning computer framework was utilized for determination and investigation of diabetes in [11]. Feed Forward Fake Neural Network or ELM contains a greater number of covered up layers which offer way better execution for the expectation of diabetes. Adjusted extraordinary learning machine and Back engendering neural network is indicated for expectation of Diabetes Mellitus having distinctive algorithms like Apriori, Frequent Design and Affiliation Run the show Mining are indicated in [12]. Association rules are too created for Apriori and Frequent Pattern Development calculations [12]. Here distinctive clustering and classification methods have been considered. as K means, Boosting, Bagging.

The SVM, Naive Bayes, Random Forest, AdaBoost, Random woodland, etc [13]. Three noteworthy strategies are refining the anticipation the precision of diabetes in [13] are Case Based Thinking, Fluffy framework, and Neural Network. It too indicates almost the strategies like Complex Valued Neural Network, Spiral Premise work, Genuine Valued Neural Network, Decision tree and their disadvantages [13].

III. Material and Implementation Methods

A. *Dataset description*

The dataset utilized amid proposed test investigate comprises of 768 occurrences which download from Kaggle. This dataset is at first from the National Founded of Diabetes and Stomach and Kidney related Illnesses. The goal is to

anticipate based on demonstrative estimations whether an understanding has diabetes. A few imperatives were put on the determination of these occurrences from a bigger database. Dataset contains data of different patients which all are females having age between 21 to 81 years, which is collected at Pima Indian heritage. It incorporates the following properties –

- Pregnancies: No. of times pregnant
- Glucose_: Plasma glucose fixation 2 hours in an oral glucose resistance test
- Blood_Pressure: Diastolic circulatory strain (mm Hg)
- Skin_Thickness: Triceps skin overlap thickness (mm)
- Insulin_Level: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight vs height ratio)
- Diabetes_Pedigree_Function: Predefined diabetes related household work
- Age_: Age of an individual (a long time)
- Outcome(final): Diabetic or not (0 or 1)

Data or dataset pre-preparing might be an essential term in ML that helps progress the idea of information to help the extraction of basic bits of information from the dataset. Pre-handling is the term which suggests the strategy of preparation (data cleaning and putting together) the unrefined data to make it fitting for a turn of events and planning ML models. Information pre-processing is the primary step plan of handle ML show. Actually, real-world information is blemished, questionable, wrong (it contains blunders or some time included exceptions), and frequently needs particular include values/trends. Amid anticipated consider dataset having a few values are non within the run where they are theoretical to be, ought to be treated as the lost esteem which are filled by cruel esteem from space values whereas a few records are erased which contains a more prominent number of lost values.

The percentage option used to divide dataset for processing through training and testing step. Out of the total occurrences (768) the 75 percentage are utilized for preparing training model and 25 percentage are utilized for testing.

B. *Exploratory Data Analysis (EDA)*

- *Univariate Data Analysis:*

This kind of investigation includes of as it were single variable. The best form of investigation may be a univariate examination, here the data bargains with as it were one amount that changes. Within proposed analysis Age, Body mass index (BMI) and Class variables are selected for the univariate analysis shown in Fig. 1and Fig. 2 –

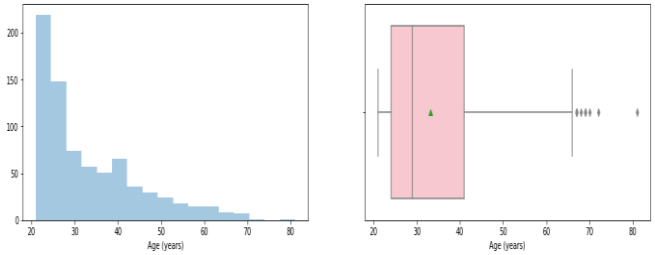

Fig. 1. Univariate Data Analysis of Age

- *Bivariate Data Analysis:*

The relationships, causes and the analysis and also to find out the association, Bivariate analysis is used in which data deals with the two variables. Thus, type of data analysis contains comparisons, causes, relationships and explanations. Proposed analysis of dataset tried to find the relationship between Age vs BMI and Age vs Diastolic Blood Pressure which is shown in Fig. 3 –

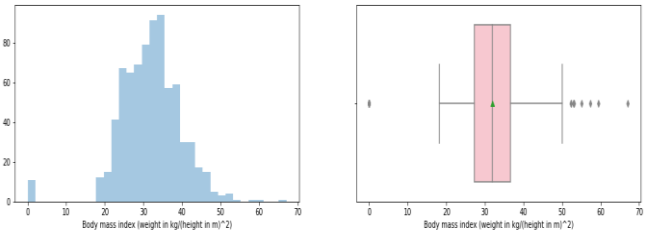

Fig. 2. Univariate Data Analysis of BMI

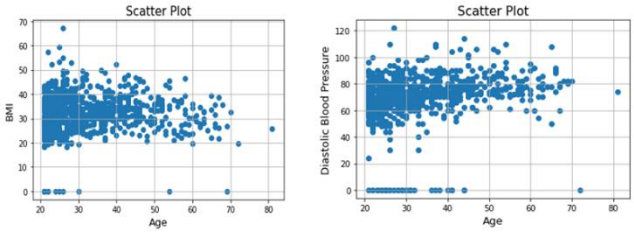

Fig. 3. Univariate Analysis of BMI vs Age

- *Multivariate Data Analysis:*

Multivariate analysis is created on the ideologies of multivariate statistics, which includes analysis of more than one statistical outcome and observations of variable at a time. Numerous conclusions are constructed using univariate analysis, but only the multivariate analysis reveals associations or relationship that help you notice problems that are not noticeable by looking at the variables separately. The Diastolic_Blood Pressure, BMI, Age, Pedigree Function and class or outcome variable are considered for Multivariate analysis which is shown in Figure 4

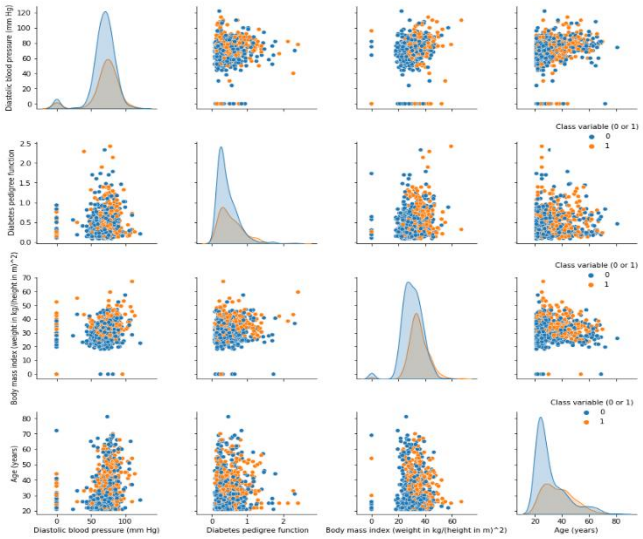

Fig. 4. Multivariate Data Analysis

C. *Logistic regression*

The regression or logistic regression is one of the foremost as a model utilized as ML algorithms for binary classification issues, which are issues with only two values, counting expectations such as either "yes or no" [14], so also inside proposed strategy the logistic regression utilized to

anticipate, a person is diabetic or not. Logistic regression may be a classification or expectation calculation based on the thought of likelihood of categorical subordinate variable. Consider given a data (X, Y), X being a matrix of values with m examples which are input parameters from dataset required to make prediction and n features (Age, BMI, …etc) and Y being a vector with m examples which are the sample outcomes based on X. The objective is to train the model LRML to predict which class the imminent values belong to it.

$$Y = b_0 + b_1 * X \quad (1)$$

Eq. 1 is for Linear Regression in which sigmoid function is added then obtained eq. 2 –

$$p = \frac{1}{1 + e^{-y}} \quad (2)$$

Then we add the above probability sigmoid function to equation 1 instead of classification dependent variable Y and derived final logistic regression function –

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * X \quad (3)$$

D. K-Nearest Neighbor (KNN)

It is a simple but very powerful method for construction of prediction model. KNN algorithm is referred for both for regression as well as classification problems [15]. But the KNN could be a non-parametric and little bit apathetic learning calculation. Non-parametric implies there's no speculation for fundamental information dispersion. In other words, the commonplace structure of KNN decided from the dataset itself. This will be actual supportive in arrangement where most of the genuine datasets don't take after numerical hypothetical suspicions.

K-Nearest Neighbors Classifier Learning

Basic Assumption:

1. All the occurrences compare to points within the n-(n-D) dimensional space from dataset where the n means the no. of highlights in any case.
2. The closest neighbors of an event are characterized in terms of the Euclidean distance [15].

An instance can be represented by < X_1, X_2, ..., X_n >, which is input matrix from dataset. Euclidean distance between two instances like age and BMI of an individual i. e. X_a and X_b is given by d (X_a, X_b) :

$$\sqrt{\sum_{j=1}^{n}\left(X_j^a - X_j^b\right)^2} \quad (4)$$

Pseudocode for KNN Classifier:

- First store all training samples < X1, X2, ..., Xn >.
- Recurrence in steps 3, 4, and 5 for each of test sample.
- Find the K number of training examples nearest to the current test sample.
- y_pred for current test sample = most common class among K-Nearest training instances.
- Go to step 2.

E. Support Vector Machine (SVM)

The most of times it identified or called as SVM, which is one of the foremost common and exceptionally well-known Supervised ML algorithms [15], which is utilized for Regression issues as well as for Classification. But, for the most part, it is utilized for the Classification issues in ML. The main goal of this Supervised algorithms is to create the unsurpassed decision boundary or line that can isolate n-dimensional space into individual classes so that we can effortlessly put the new data instance in the correct group in the future. This unsurpassed decision boundary is called a hyperplane. It can be possible to construct multiple decision boundaries (shown in Fig. 5) to separate out the classes in n-dimensional space, but to find out the finest decision boundary that benefits to classify the data points.

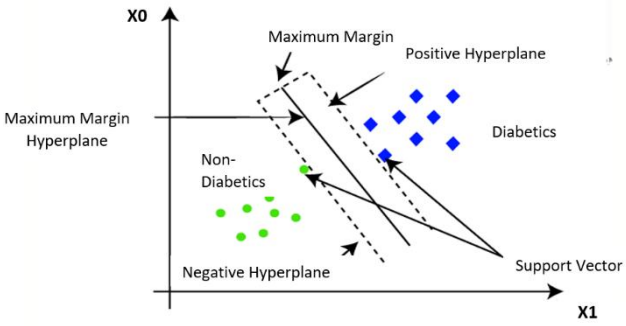

Fig. 5. Hyperplanes Illustration [Source: Internet]

Well-known Forms of SVMs : There are two distinctive kinds of SVMs, each utilized for assorted things:

Simple SVM: Regularly, it is utilized for classification & direct regression problems.

Kernel SVM: It has more flexibility for non-linear information since able to include more highlights or parameters to fit a hyperplane input of a two-dimensional space.

The primary goal is to the test instances from dataset, according to properties supported by instance decide it is either positive or negative or it belongs from positive zone which is diabetics class it is characterized by formulation W * X + B = 1 and the non- diabetics or negative zone is formulated by W * X + B = - 1 which shown in figure 5.

F. Naïve Bayes classifier

This prediction or classification model (Naïve Bayes) built on the basics of probability that is Bayes theorem [16]. The development of Naïve Bayes is built on the statistical demonstrating with linear work. Hypothetically, it ordinarily implies the apparatus of unreasonable suspicion that the qualities are essentially critical and self-governing. The genuine dataset incorporates of attributes that are surely not correspondingly imperative or autonomous. Credulous Bayes techniques are a bunch of administered learning calculations dependent on put on Bayes' speculation with the "naive" hypothesis of contingent independence between each feature given the estimation of the class variable [16]. Bayes' hypothesis expresses the subsequent affiliation, given class variable Y and ward include vector X_1 through X_n

$$\boldsymbol{P}(\boldsymbol{Y}|(X_1, X_2, \ldots\ldots X_n) = \frac{\boldsymbol{P}(\boldsymbol{Y})\,\boldsymbol{P}((X_1, X_2, \ldots\ldots X_n \mid \boldsymbol{Y})}{\boldsymbol{P}(X_1, X_2, \ldots\ldots X_n)} \quad (5)$$

The final classifier model makes as got to discover the likelihood of given set of inputs for all likely values of the class variable y and preference up the yield with greatest likelihood. This could be expressed scientifically as in equation 6. This kind of order having unique classes like Balanced Naive Bayes, Multi-nomial, Categorical aswell as popular one Gaussian, and Bernoulli Gullible Bayes. In proposed paper the Gaussian NB classifier is used. Gaussian Gullible Bayes expect that relentless qualities are analyzed from a gaussian conveyance and expects to be the taking after:

$$Y' = argmax_2 p(Y) \prod_{i=0}^{n} P(X_i|Y) \quad (6)$$

$$P(X_i Y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(X_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (7)$$

The parameters σ_y and μ_y are estimated using maximum likelihood.

G. Decision Tree Classifier

It is a supervised ML algorithm, it's also castoff for both for regression and classification task. It constructs a model that usages set of rules to does the classify something. It the samples are more randomness in nature from data it will create more entropy, that must choose an algorithm that must maximize the information gain and minimizes the entropy [17]. In a few of the setting, DT classifier should be executing for classification. DT might be type of classification algorithm with the single root node and more than one inside or leaf nodes. The class name which contains genuine data will be consigned to the leaf. The root and centre contain particular - distinctive test settings, to parcelled unmistakable properties the decision of in the first-place node(root) is developed on the information get. Thinking about the essential node, the condition is significant to every one of the records and every division is constructed dependent on the outcome. A property determination in DT degree could be a heuristic for picking the agonizing model that's most competent of making choice that how to perform segment on information in such way that would influence in person class. Attribute finalisation measures are also called as splitting rules because they describe how the data points from dataset are splits on a certain level of DT [17].

Here are some major attribute selection measures:

Information Gain: This attribute offers amount of information required that uses splitting attribute in relationships of the to further tree. It minimizes the info which is needed to categorize the data points into the respective dividers and reflects the least arbitrariness or "impurity" in these partitions.

The information gain can be calculated as follows:

$$Gain\ (A) = Info(D) - Info_A(D) \quad (8)$$

Here Info(D) is the mean amount of information required to classify the patient is diabetics or not of a data fact in set of data point D, it is calculated from independent variables of dataset like Age, BMI, Body glucose level, …, etc.

Gain Ratio: The info gain measure is unfair toward tests with many results. Thus, it prefers pick out attributes that have an appropriate number of values. The Gain ratio is purely used to improve this problem.

$$SplitInfo_A(D) = -\sum_{j=i}^{v} \frac{|D_j|}{|D|} * log_2\left(\frac{|D_j|}{|D|}\right) \quad (9)$$

$$GainRatio(A) = \frac{Gain\ (A)}{SplitInfo(A)} \quad (10)$$

Gini Index: The Gini index is calculated in the following manner:

$$Gini\ (D) = 1 - \sum_{i=0}^{m} C_i^2 \quad (11)$$

where c_i is the possibility or probability that a given tuple from data values which belongs from D to class which either from diabetics or non-diabetics and is projected by $|C_i, D| / |D|$. The sum is calculated over m classes.

H. Random forests

Random forests (RF) develop numerous individual decision trees at training step. Forecasts from all trees are pooled to form the final forecast; the mode of the classes for classification or the cruel forecast for regression. As they utilize a get together of comes about to create a final decision, they are raised to as Ensemble procedures. Ensemble learning may be a ML strategy that combines various base models in arrange to abdicate one ideal prediction or decision-making model which could be a powerful model. The different types of Ensemble Methods: AGGregating (Random forest), Bootstrap Boosting or BAGGing, Cascading and Stacking which are used in Random forests.

RF Model = DT + bagging model + feature_bagging model + aggregation model

Here our base learner having very low bias inclination and high fluctuation so, to prepare DT with full perceptiveness length. There's no ought to stressed almost profundity, let them develop as at conclusion change diminish in demonstrate conglomeration.

Here base learner taking high discrepancy and little bias to train out DT with full of depth, so let them raise at termination discrepancy reduction in model aggregation that why there is no any need to concerned about depth of DT or RF Model.

I. Artificial Neural Network (ANN)

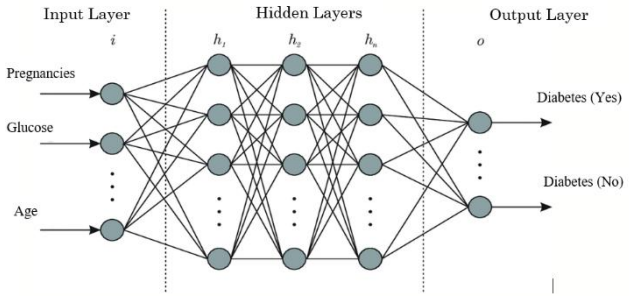

Fig. 6. Artificial Neural Network Architecture

ANNs are effective data-driven modelling framework which is popularly used for nonlinear systems for identification and dynamic modelling, due to their common approximation abilities and flexible structure in nature which let to capture complex nonlinear behaviours [18]. Moreover, it called as response surface approximation model or proxy

model for the reason that of its heftiness to solve nonlinear and multivariate modelling problems, similar to the function classification and approximations. The ANN framework or classifier is built to expand the accuracy or precision value of given classification problem of diabetes patients [20] proposed in this paper shown in Fig 6. It contains three different layers, first is Input Layer which accept all the input values like Pregnancies, Glucose, BMI, etc. which are from input dataset. Next layer which is hidden layer could be a covered-up layer found between the input and output of the ANN classifier, in which the faction applies weights to the inputs and coordinates them through an activation function as the output.

The ANN preparing process was made utilizing the Levenberg-Marquardt (LM) backpropagation algorithm [18], bearing in intellect the mean squared error (MSE) as joining indicator and a most extreme of 150 epochs. Let us comment that the LM strategy has second-order convergence rate and it was proposed by Hagan and Menhaj [18] since of its capability for ANN with no more than a number of hundred weights, because it is right now the case.

ANN Design Paradigm

- For mathematical modelling, let us first vectorize all input samples including the input, the output and their intermediate weights.
- Keeping L2 as the reference layer, the input vector arises from output of layer L1.
- Whereas the output vector of L2 is fed as the input of L3.
- The X used to specify the input vector, where X= [x1, x2, x3].
- The output vector ŷ (prediction vector) create the corresponding outputs of the perceptron model in L2 layer, where ŷ = [y1, y2, y3, …….. , yn].
- The components of weights vector i. e. W, are concerning the input element to the matching perceptron's in a given layer.
- The average of weighted input vector with that of the individual weights can be statistically modelled as the dot product (multiplication) of the equivalent vectors.
- Multiply X by the transpose of the weight matrix W.
- Transposing of W and X is performed in such way where dimensions both matrixes must match. After this procedure add the bias vector by matrix addition.
- These steps are together termed and called as forward propagation.
- The mathematical equation turns out to be:

$$Z = WT.\ X + b$$

- Then present a new module in the network which imposes non-linearity to the data. This new accumulation to the architecture is called the activation function.
- The proposed ANN usages ReLU, and Sigmoid two distinct activation functions at input layer at output layer respectively.
- Each neural layer gives out an output vector ŷ, after the forward propagation then to the next layer as the input vector and it continues till the last layer.
- At a time of the initial forward propagation process, it randomly initialises the weights and biases.
- Now proposed ANN tune these constraints in accordance with the dataset then it can call "the problem statement". Thus, the tuning of the biases and weights is finalized with the help of backward propagation which is another set of algorithms.
- The prediction vector ŷ at the end the network. The error of this specific o/p from the network is intended with admiration to the actual predictable o/p.

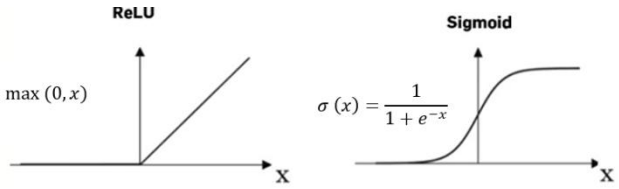

Fig. 7. ANN Activation Functions (Source: Internet)

- Each neural layer gives out an output vector ŷ, after the forward propagation then to the next layer as the input vector and it continues till the last layer.
- At a time of the initial forward propagation process, it randomly initialises the weights and biases.
- Now proposed ANN tune these constraints in accordance with the dataset then it can call "the problem statement". Thus, the tuning of the biases and weights is finalized with the help of backward propagation which is another set of algorithms.
- The prediction vector ŷ at the end the network. The error of this specific o/p from the network is intended with admiration to the actual predictable o/p.

$$Loss(y, \hat{y}) = \sum_{i=1}^{n} (y - \hat{y})^2 \quad (12)$$

Where,
y : actual or original o/p value
ŷ : projected o/p
n : all data-points

- where ŷ is given by:

$$f\left(b + \sum_{i=0}^{n} x_i\, w_i\right) \quad (13)$$

- The above equation f(z) is the activation function used on the weighted sum of i/p vector. then derive the idea that loss function obviously tells the total loss suffered by a model.

IV. Results and Discussion

As discussed in Section III, for prediction of diabetic's patients there are six different Machine Learning

classification methods are applied as well as for improving accuracy the seventh method has been constructed as an Artificial Neural Network (ANN) which construct 97.20 % accuracy. Prior to building a model for breaking down information, and assessing the classification execution, getting ready and testing tests measure is led to examination in consent to the technique depicted in [19]. With the given number of instances from dataset of diabetes patients are from Pima Indian legacy or heritage are analyze through the various ML algos, and the obtained average accuracy level of 80%, the most of misclassification occurred between 1 to 25 %, that occurs because of the limited number of tests to examinees, it is required to be in the reach from of dataset in the form of number of truly validated instances. The Table I shows the number of accurately and false classified records, which is of estimated by the seven classifiers of ML along the no of instances used for testing and training set during algo design. Whereas observing the meticulousness, the most noteworthy accuracy is given by Logistic Regression, K-NN and SVM (kernel=linear) which is 80.20 and the most reduced value is appeared by Random Forest is 72.40%. The Random Forest regard of error rate is high and low for Logistic Regression, K-NN and SVM linear kernel.

TABLE I. CLASSIFICATION AND THE ACCURACY WITH CONFUSION MATRIX

Classifiers	**No. of Records**	**Training / Test set**	**Correctly classified no. of instances**		**Incorrectly classified no. of instances**		**Accuracy**
			TP	***TN***	***FP***	***FN***	
Logistic Regression	768	576/ 192	118	36	12	26	80.20 %
K-NN	768	576/ 192	114	40	16	22	80.20 %
SVM (kernel=linear)	768	576 /192	117	37	13	25	80.20 %
SVM (kernel=rbf)	768	576/ 192	117	32	13	30	77.60%
Naive Bayes	768	576/ 192	114	33	16	29	76.56%
Decision Tree	768	576/ 192	105	44	25	18	77.60%
Random Forest	768	576/ 192	110	29	20	33	72.40%

The notations- TP, TN, FP and FN suggest the quantity of data that represent the true positive, it represents patients count those have actual, second true negative indicate patients which are not diabetics in reality. The other true measures as false positive and false negative represent incorrect prediction or called as false prediction which shows diabetic patients as non-diabetic or vice-versa.

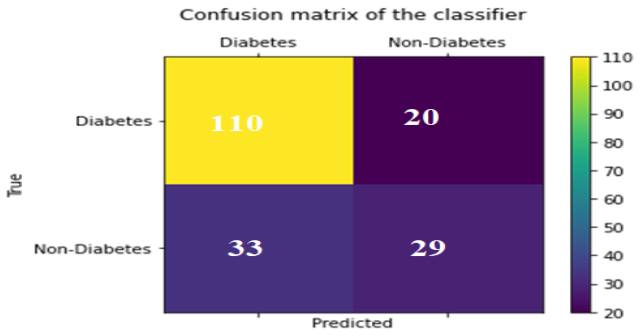

Fig. 8. Confusion Matrix with Accuracy 80.50 %

The others measures are also used for performance measure of the different classifiers which are implemented as the Precision, second Recall and f1-score with support. However, the precision of classifier is the amount of the facts from our model it says that model is where relevant or not. Recall can be supposed as of a model's aptitude to find all the data points of concentration in a dataset, its overnight delivery the ability to find all appropriate instances in a dataset. To discover an ideal balance of accuracy and recall, they joined together as F1 score to the performance utilizing. The F1 is the consonant mean of recall and precision taking both measurements under consideration. These measures are calculated as given below:

$$precision = \frac{true\ positives\ (TP)}{true\ positives\ (TP) + false\ positive\ (FP)}$$

$$recall = \frac{true\ positives\ (TP)}{true\ positives\ (TP) + false\ negative} \quad (15)$$

$$F1 = 2 * \frac{precision * recall}{precision + recall} \quad (16)$$

The outcomes relating to the precision and recall shown in Table II. having classes such as individual is diabetics (1) or nondiabetics (0).

TABLE II. PRECISION, RECALL AND F1 SCORE

Classifiers	**Precision (Non-diabetic)**	**Recall (Non-diabetic)**	**F1-Score (Non-diabetic/ Class No)**	**Precision (Diabetic/ Class Yes)**	**Recall (Diabetic/ Class Yes)**	**F1-Score (Diabetic/ Class Yes)**
Logistic Regression	0.82	0.91	0.86	0.75	0.58	0.65
K-NN	0.84	0.88	0.86	0.71	0.65	0.68
SVM (kernel= linear)	0.82	0.90	0.86	0.74	0.60	0.66
SVM (kernel=rbf)	0.80	0.90	0.84	0.71	0.52	0.60
Naive Bayes	0.80	0.88	0.84	0.67	0.53	0.59
Decision Tree	0.85	0.81	0.83	0.64	0.71	0.67
Random Forest	0.77	0.85	0.81	0.59	0.47	0.52

There's another way to evaluation of classification model could be a ROC Curve which is a Receiver Operating **Characteristic Curve, it is an incredible strategy for estimating the execution of an order model. The Cumulative Accuracy Profile (CAP) is also utilized as a method in ML through which the severe force of a grouping show is imagined. The CAP of a model** infers the cumulative number of optimistic happens along the y-hub versus the x-hub which having the comparing absolute number. The CAP is unique in relation to the ROC curves, it plots the true positive rate in separate to the false positive pace of arrangement.

The CAP Curve Analysis model appeared in Fig. 9 which shows the particular modes, in the first place, plot an irregular model which depends on the truth that the correct disclosure of class 1.0 will develop straight. Extraordinary mode which has way preferable estimate result over the any irregular model. Another, plot the ideal model. An ideal model is one which is capable distinguish all class 1.0 information focuses inside similar number of attempts as

there are class 1.0 information point. Finally, plot the happens from the Logistic Regression Classifier showed up in Fig 10 by blue shading which having higher accuracy through CAP Curve which is 92.26 %. Accuracy examination outline of all classifiers showed up in Fig 11. The essential procedure to dissect the CAP Curve is using Area Under Curve. We should consider region under curve show by then ascertain the Accuracy rate using the following steps:

- Estimate the (aP) which is area under the perfect model shown in Fig. 9
- Estimate the (aR) that is area below or under the actual model shown
- Compute (AR), i. e. Accuracy Rate

$$AR = \frac{a_P}{a_R} \qquad (17)$$

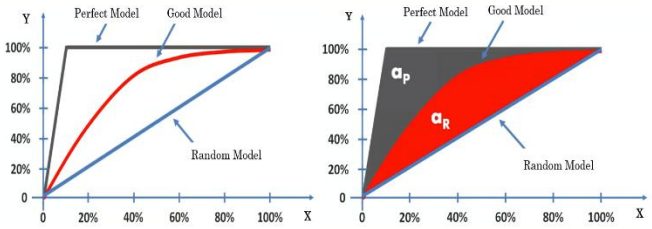

Fig. 9. CAP Curve Analysis [Source : Internet]

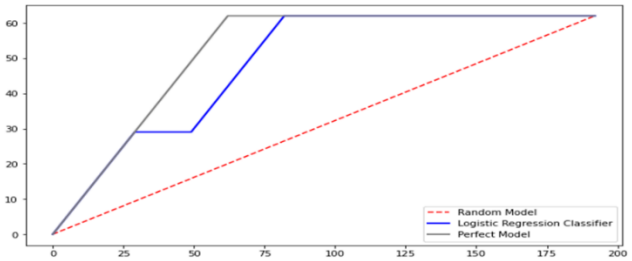

Fig. 10. CAP Curve Analysis of Logistic Regression

As discussed in Section III, the ANN classification model builds to improve the prediction accuracy, which include input layer having all input attributes from input dataset. As discussed above from Machine Learning classification models Logistic Regression, K-NN, and SVC provide higher accuracy (80.20) through confusion matrix performance measure. CAP Curve Analysis is another performance measure which shows of Logistic Regression having higher accuracy (92.26) than all other models. The ANN model provide accuracy from 75.58 % to 97.66 % and loss from 18.54 % to 0.03with 150 epochs and batch size defined as 10. The model and accuracy graph of ANN classification model shown below in Fig. 11.

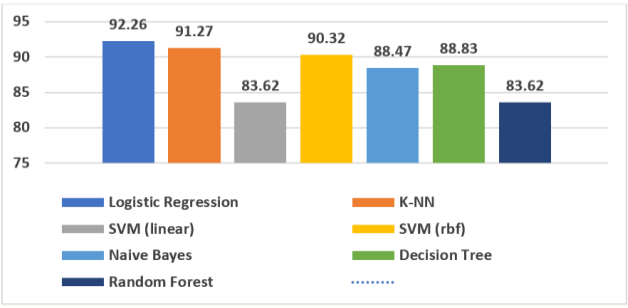

Fig. 11. Comparison of Accuracy through CAP Curve Analysis

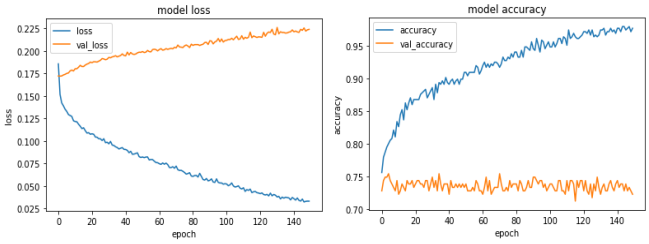

Fig. 12. Comparison of Accuracy through CAP Curve Analysis

V. Conclusion

The blood glucose level out of the desired range will make the diabetic patients to go into a risk of various different ailments, sooner or later patients having vision issues or it goes into trance state or coma too. The unbalanced management of the blood glucose levels of an individuals' body leads the serious difficulties or issues related with his/her health, for avoiding such issued. It is required to design such prediction or classification framework that helps to an individual. The Machine Learning can be utilized for evaluating distinctive diseases pattern, helpful data taking out to identify risk also it to design patient support system and it required to find such impactful clinical parameters and association between them. Consequently, Machine Learning techniques are proposed that anticipate diabetes which employments different decision parameters included in dataset plan. The proposed framework used Pima Indians Diabetes dataset taken from Kaggle store consist of nine different attributes, after applying pre-processing it remains 768 records or instances. The various ML classification algorithms are proposed to verify the accuracy or precision of prediction of the blood glucose. CAP curve method of performance analysis of ML algorithms evaluates the 92.26 % accuracy of Logistic Regression which is higher in all other designed algorithms. Moreover, the precision of decision support framework designed, it can be improved with the usage of other effective classifiers like ANN, which shows noticeable improvement in accuracy which is 97.66%

Future work concerns more deeper investigation and classification of way of life illnesses like hypertension, joint pain and diabetes. Ordinarily, these common clinical illnesses emerge with age but, presently within the current time, these are no more as it were pertinent to the age[21]. Due to the active plan or way of life of a person, they emerge at any organize of life. Future work will centre on examining unique finger impression images and blood groups of a person and the dataset which incorporates the outside qualities work nature, eating propensities (non-vegetarian or vegetarian), locale (urban or rural), enslavement (in the event that they may like drink, smoke), etc.

References

[1] Zarita Zainuddin, Ong Pauline and Cemal Ardil, "A Neural Network Approach in Predicting the Blood Glucose Level for Diabetic Patients", World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:3, No:2, 2009

[2] Veena Vijayan V., Anjali C. "Prediction and Diagnosis of Diabetes Mellitus -A Machine Learning Approach", 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 10-12 December 2015

[3] Zhilbert Tafa, Nerxhivane Pervetica, Bertran Karahoda, "An Intelligent System for Diabetes Prediction", 4th Mediterranean Conference on Embedded Computing Budva, Montenegro, 2015

[4] C.kalaiselvi, G.M.Nasira, “A New Approach of Diagnosis of Diabetes and Prediction of Cancer using ANFIS”, IEEE Computing and Communicating Technologies, p p 188-190, 2014.

[5] Bellazzi, R. and B. Zupan. “Predictive data mining in clinical medicine: Current issues and guidelines.” International journal of medical informatics 77 2 (2008): 81-97 ..

[6] Ning Wang and Guixia Kang, "A monitoring system for type 2 diabetes mellitus," 2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom), Beijing, China, 2012, pp. 62-67, doi: 10.1109/HealthCom.2012.6380067.

[7] Sarwar, Abid and V. Sharma. “Intelligent Naïve Bayes Approach to Diagnose Diabetes Type-2.” (2012).

[8] G. Thangarasu and P. D. D. Dominic, "Prediction of hidden knowledge from Clinical Database using data mining techniques," 2014 International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, Malaysia, 2014, pp. 1-5, doi: 10.1109/ICCOINS.2014.6868414.

[9] Asma A. Al Jarullah , “Decision discovery for the diagnosis of Type II Diabetes”, IEEE conference on innovations in information technology,pp-303-307,2011.

[10] M. NirmalaDevi, S. A. alias Balamurugan and U. V. Swathi, "An amalgam KNN to predict diabetes mellitus," 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), Tirunelveli, India, 2013, pp. 691-695, doi: 10.1109/ICE-CCN.2013.6528591.

[11] Jefri Junifer Pangaribuan , Suharjito “Diagnosis of Diabetes Mellitus Using Extreme Learning Machine”.pp-33-38,2014

[12] Sankaranarayanan.S, Pramananda Perumal T. “Diabetic prognosis through Data Mining Methods and Techniques” International Conference on Intelligent Computing Applications, pp.162-166,2014

[13] P. Undre, H. Kaur and P. Patil, "Improvement in prediction rate and accuracy of diabetic diagnosis system using fuzzy logic hybrid combination," 2015 International Conference on Pervasive Computing (ICPC), Pune, India, 2015, pp. 1-4, doi: 10.1109/PERVASIVE.2015.7087029.

[14] J. Ran, G. Zhang, T. Zheng and W. Wang, "Logistic Regression Analysis on Learning Behavior and Learning Effect Based on SPOC Data," 2018 13th International Conference on Computer Science & Education (ICCSE), Colombo, Sri Lanka, 2018, pp. 1-5, doi: 10.1109/ICCSE.2018.8468834.

[15] C. A. Ul Hassan, M. S. Khan and M. A. Shah, "Comparison of Machine Learning Algorithms in Data classification," 2018 24th International Conference on Automation and Computing (ICAC), Newcastle Upon Tyne, UK, 2018, pp. 1-6, doi: 10.23919/IConAC.2018.8748995.

[16] R. Muhamedyev, K. Yakunin, S. Iskakov, S. Sainova, A. Abdilmanova and Y. Kuchin, "Comparative analysis of classification algorithms," 2015 9th International Conference on Application of Information and Communication Technologies (AICT), Rostov on Don, Russia, 2015, pp. 96-101, doi: 10.1109/ICAICT.2015.7338525.

[17] A. Navada, A. N. Ansari, S. Patil and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," 2011 IEEE Control and System Graduate Research Colloquium, Shah Alam, Malaysia, 2011, pp. 37-42, doi: 10.1109/ICSGRC.2011.5991826.

[18] Krogh, A. “What are artificial neural networks”. Nat Biotechnol 26, 195–197 (2008). https://doi.org/10.1038/nbt1386

[19] J. E. Bartlett, J. W. Kotrlik, and C. C. Higgins, “Organizational research: Determining appropriate sample size in survey research,” Information Technology, Learning and Performance. Journal., vol. 19, no. 1, pp. 43– 50,2001.

[20] Patil, V., Ingle, D.R. An association between fingerprint patterns with blood group and lifestyle based diseases: a review. Artif Intell Rev 54, 1803–1839 (2021). https://doi.org/10.1007/s10462-020-09891-w

[21] P. N. Vijaykumar and D. R. Ingle, "A Novel Approach to Predict Blood Group using Fingerprint Map Reading," 2021 6th International Conference for Convergence in Technology (I2CT), 2021, pp. 1-7, doi: 10.1109/I2CT51068.2021.9418114.

View publication stats