International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252

Evaluation and Implementation of a Diabetic Prediction Method based on Machine Learning

Raman Kumar Mondal, Rajneesh Kumar, Nirbhay Mishra, Sarbajit Roy , Shailesh Pandey, Vishal Kumar, Purushotram Kumar, Sudhir Kumar

Ramgovind institute of technology, koderma, jharkhand

Submitted: 25-06-2021 Revised: 06-07-2021 Accepted: 09-07-2021

ABSTRACT

In a recent survey it was discovered that the healthcare industry needs accurate data of increasing cases of diabetes mellitus. Today it is a big question mark for the whole world in our health care industry to predict and prevent diabetic mellitus, which is a chronic and fatal disease for a human being. Primarily it occurs in our body when the pancreas does not produce enough insulin or when our body cannot utilize the insulin after formation. According to the surveying data of WHO, 78 million people in India would be diabetic patients, which is ranked second among Asian countries next to China. In recent years Computer Science has provided different machine learning techniques for examining specific data sets. We observed that the data mining technique will be a very efficient pattern which specifies the dataset and gives valuable patterns to the applied classifiers.

The purpose of this research is to identify or predict diabetes with the help of several machine learning classifiers. In this study we used WEKA software as mining tools for better diagnosis of diabetes. Here the PIMA INDIAN dataset is acquired from UCI repository. Our dataset is analyzed for the purpose of creating an effective model that may give the greatest precision and accuracy. In this research we also applied bootstrapping resample methodology to enhance the accuracy level. Furthermore, we used classifiers like naïve bayes, random forest, ANN, KNN, SVM and LR for accuracy. Also we analyze the comparative result along with these classifiers.

Keywords: Diabetic Mellitus; Machine Learning; Data Mining; ANN; KNN; Naïve Baye; S.V.M; LR; Bootstrap sampling.

I. INTRODUCTION

Diabetes is a common chronic disease in a recent times and poses a great threat to human beings. These types of diseases affected our quality of life, which is a major adverse effect. Diabetes is one of the most acute diseases and is present world-wide. A major reason for deaths in adults across the globe includes this chronic condition. The characteristic of diabetes is that the blood glucose is higher than the normal level, which is caused by defective insulin secretion or its impaired biological effects or both (Lonappan et al 2007) [1]. Diabetes can lead to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves (Krasteva et al 2011) [2]. Generally chronic conditions are also cost associated. A major portion of the budget is spent on chronic diseases by governments and individuals [3,4].The world-wide statistics for diabetes in the year 2013 revealed around 382 million individuals had this ailment around the world [5]. It was the fifth leading cause of death in women and eight leading cause of death for both sexes in 2012 [4].

As of 2019, an estimated 463 million people had diabetes world-wide (8.8% of the adult population) with type-2 diabetes, making up about 90% of the cases [8]. Rates are similar in women and men [9]. Trends suggest that rates will continue to rise [8]. Diabetes at least doubles a person's risk of early death [10].

In 2019, diabetes resulted in approx. 4.2 million deaths [8]. It is the 7th leading causes of death both globally and in the U.S [11].The global economic cost of diabetes related health expenditure in 2017 was estimated at US$ 727 billion [7]. In the U.S diabetes cost nearly U.S$327 billion in 2017 [12]. Average medical expenditure among people with diabetes is about 2.3 times higher[13]. Research on biological data is limited but with the passage of time enables computational and statistical models to be used for analysis. A sufficient amount of data is also being gathered by health care organization. New knowledge is

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252

gathered when models are developed to learn from the observe data using data mining techniques.

Data mining is the process of extracting from data and can be utilized to create a decision making process with efficiency in the medical domain. Several data mining techniques have been utilized for disease prediction as well as for knowledge discovery from bio-medical data [14, 15]. Diagnosis of diabetes is considered a challenging problem for quantitative research.

Many complications occur if diabetes remains untreated. Some of the severe complications include diabetic ketoacidosis and nonketotic hyperosmolar coma [16].Diabetes is examined as a vital serious health matter during which the measure of sugar substance can't be controlled.

Diabetes is not only affected by various factors like height, weight, hereditary factor and insulin but the major reason considered is sugar concentration among all factors. The early identification is the only remedy to stay from the complication [17]. With the development of living standards; diabetes is increasingly common in people's daily life.

Therefore, how to quickly and accurately diagnose and analyze diabetes is a topic worth studying.

In medicine the diagnosis of diabetes is according to fasting blood glucose, glucose tolerance, and random blood glucose levels (Iancu et al 2008, COX and Edelman 2009 [18,19], American diabetes Association 2012) [20].The earlier diagnosis is obtained the much easier we can control it. Machine learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data and it can serve as a reference for doctors (Lee and Kim 2016) [21], Alghamdi et al 2017 [22], Kavakiotis et al 2017 [23] for machine learning method, how to select the valid features and the correct classifier are the most important problem. Data mining [24, 25] and machine learning algorithms gain its strength due to the capability of managing a large amount of data to combine data from several different sources and integrating the background information in the study [26]. Mainly machine learning methods are widely used in predicting diabetes and they get preferable results.

This research work focuses on pregnant women suffering from diabetes. In this study ANN, RF, KNN, NAÏVE BAYES, J48-DECISION TREE and S.V.M machine learning classification algorithms are used and evaluation on the PIDD data set to find the prediction of diabetes in a patient. Here we presented several parameters like accuracy, misclassification rate, precision, recall, f-score as well as resampling which could be given the better efficiency and diagnosis of this disease.

II. MATERIALS AND METHODS

Diabetes or diabetes mellitus is a metabolic disorder in the body. In this disease is destroy the ability to produce insulin in the patient's body or the body becomes resistant to insulin and therefore the produced insulin can't perform its normal function. The primary role of insulin is to blood sugar by different mechanisms. Symbols often include frequent urination, increased thirst, and increased appetite. If left untreated, diabetes can cause much complication [2]. Acute complication can include diabetic ketoacidosis, hyperosmolar hyperglycemic state or death [3]. Serious long term complications include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, damage to the nerves, damage to the eyes and cognitive impairment [2, 5]. Diabetes is due to either the pancreas not producing enough insulin or the cells of the body not responding properly to the insulin produced [12]. There are mainly two types of diabetic patients which are type 1 and type 2.

In type 1, diabetes results from the pancreas failure to produce enough insulin due to loss of beta cells [2]. This form was previously referred to as insulin dependent diabetes mellitus [IDDM] or juvenile diabetes [2]. The loss of beta cells occurs by an autoimmune response [13]. The cause of this autoimmune response is unknown [12].

In type 2, diabetes begins with insulin resistance, a condition in which cells fail to respond to insulin properly [2]. As the disease progresses a lack of insulin may also develop.

This form was previously referred to as non-insulin dependent diabetes mellitus (NIDDM) or adult onset diabetes [2]. Type 2 diabetes also known as genetic factors, obesity and lack of physical activity have an important role in a person.

Data Mining

There are two types of data mining task, the predictive model and descriptive model which are explained below:

1. Predictive model: The predictive data mining model predicts the future out comes based on past records, present in the database or with known answers. Data mining will help figure-out the future credit risk of the applicant and predict future credit history of the applicant by using past data. Classification is known as the

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252

procedure used to locate a model that best suits identified data sets or ideas. The model helps predict the class of objects when class labels are not available. The resultant model is focused on analyzing a set of identified classes. Regression is a mathematical and statistical tool used widely in using numeric values for forecasting time series analysis. Prediction as the term implies means correctly envisioning the future using logical computation of available data.

2. Descriptive Model: This model is to discover patterns in data and understand the relationship between the data attributes. Descriptive model represents the main feature of the data, and summarizes .The collected knowledge can be used to develop marketing programs for targeting audiences. Clustering examines data objects without referring to an identified class label. Summarization is to categorize the distinctive properties of data and point out if the data values are to be categories of noise or outliers. This research classifiers data mining as shown in Figure 4.
3.

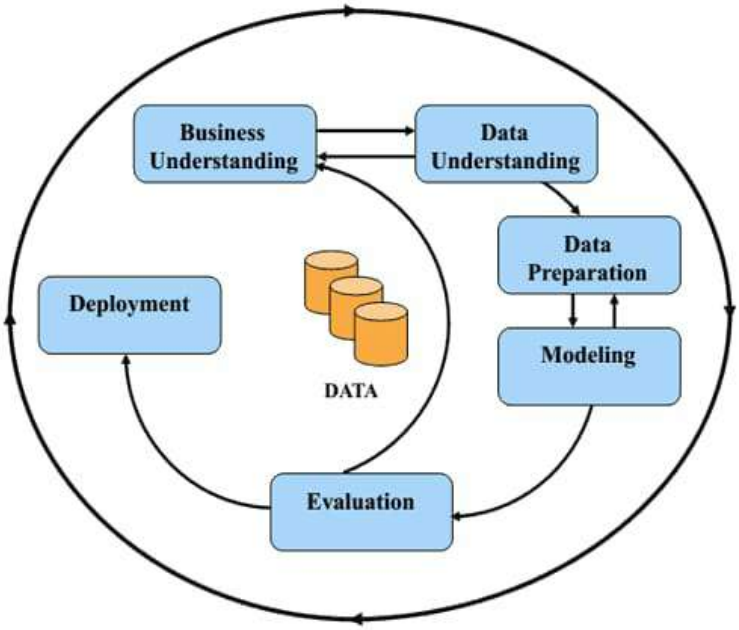


Fig. 4: Data mining task (research gate.com)

Bootstrapping Technique

The bootstrap is a recently developed technique for making certain kinds of statistical inferences .It is only recently developed because it requires modern computer power to simplify the often intricate calculation of traditional statistical theory. The bootstrap is a data based simulation method for statistical inference which can be used to produce inference [27]. The use of the term bootstrap derives from the phrase to pull oneself up by one's bootstraps, widely thought to be based on one of the eighteen century adventures of Baron Munchausen by Rudolph Erich Raspe. In this terminology of statistical summaries and inferences like, regression, correlation analysis of variance, discriminant analysis, standard error, significance level and confidence interval has become the lingua franca of all disciplines that deal with noisy data.

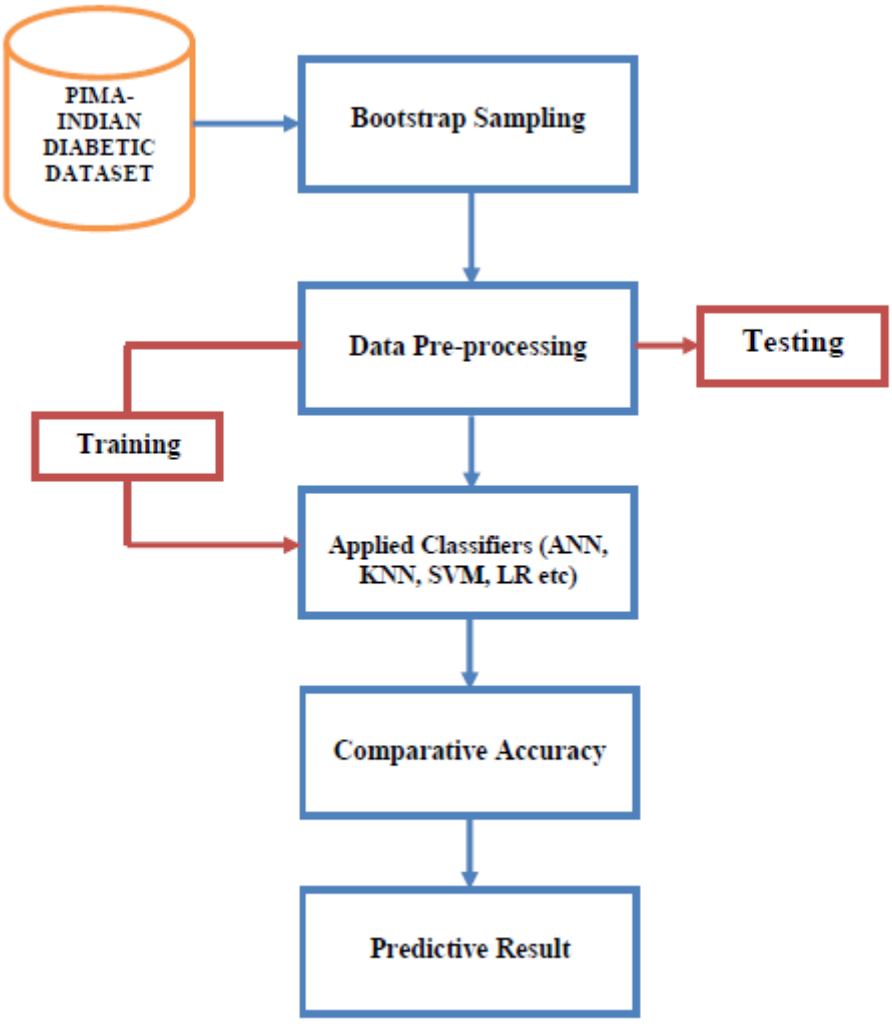
Bootstrapping techniques are rapidly entering mainstream data analysis; some statisticians believe that bootstrapping procedures will soon support common nonparametric procedures and may displace most parametric procedures as well. This paper introduces the vocabulary logic and demonstrates basic application of permutation and bootstrap resampling methods. Resampling methods have become practical with the general availability of cheap rapid computing and new software. Compared to standard methods of statistical Inference these modern methods often are simpler and more accurate, require fewer assumptions and have greater generalizability. Bootstrapping provides especially clear advantages when assumptions of traditional parametric tests are not met, as with small samples from non- normal

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252

distributions. Here we used a supervised sampling technique to the preprocessed dataset. As the class attribute is of Nominal data type therefore we are using supervised resample filter in WEKA, which produces a random subsample of a dataset using either by sampling with replacement or sampling without replacement. Resampling is a series of methods used to reconstruct your sample datasets, including training sets and validation sets [28].

The original dataset must fit completely in memory. The amount of instances in the generated dataset May be identified. This filter helps to preserve the class distribution in the subsample or to bias the class Distribution to a near balanced distribution. It can provide more useful different sample sets for learning Process. In this study we adopt a bootstrapping method of resample on the dataset which obtains a random Sample with replacement from a sample. In order to achieve balance classes WEKA can use a resample with Replacement which replicates some instance.

Model Diagram:



International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252

Data Set

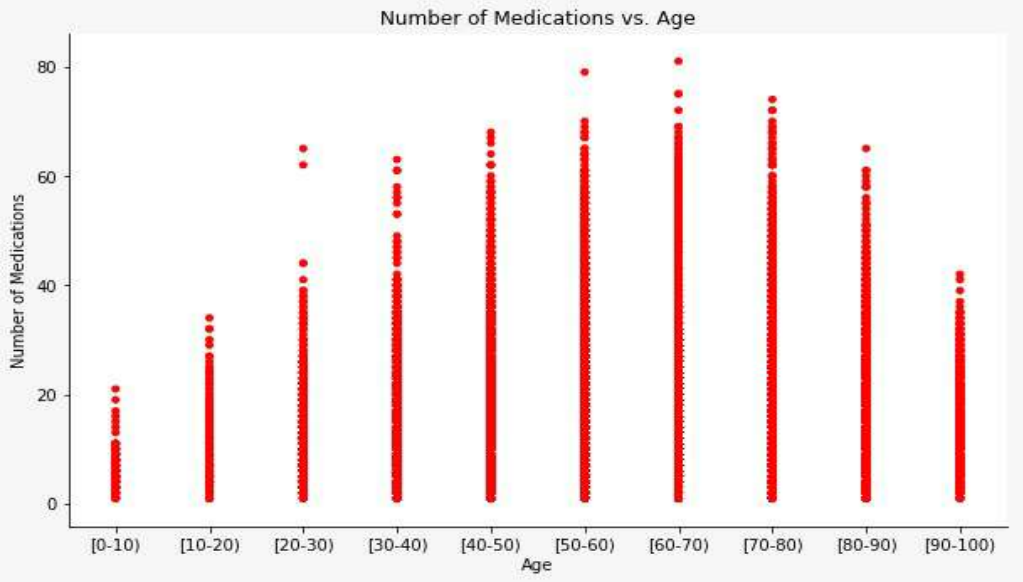
Table 1: Pima Indian Diabetic Data Set

Attributes	Min Value	Max Value	Mean Value	S.D (Standard Deviation)
No of Time Pregnant	0	17.00	3.84	3.37
Plasma Glucose Concentration 2Hrs. In Oral Glucose Tolerance Test	0	199.00	120.85	31.97
Diastolic Blood Pressure	0	122.00	69.50	19.35
Triceps Skin Fold Thickness(Mm)	0	99.00	20.53	15.95
2-Hrs Serum Insulin(Mu/Ml)	0	846.00	79.79	115.24
Body Mass Index(Bmi) Weight In Kg/Height In M) ^2	0	67.10	31.99	7.88
Diabetes Pedigree Function	0.07	2.42	0.42	0.33
Age	21.00	81.00	33.24	11.76
Class Variable				

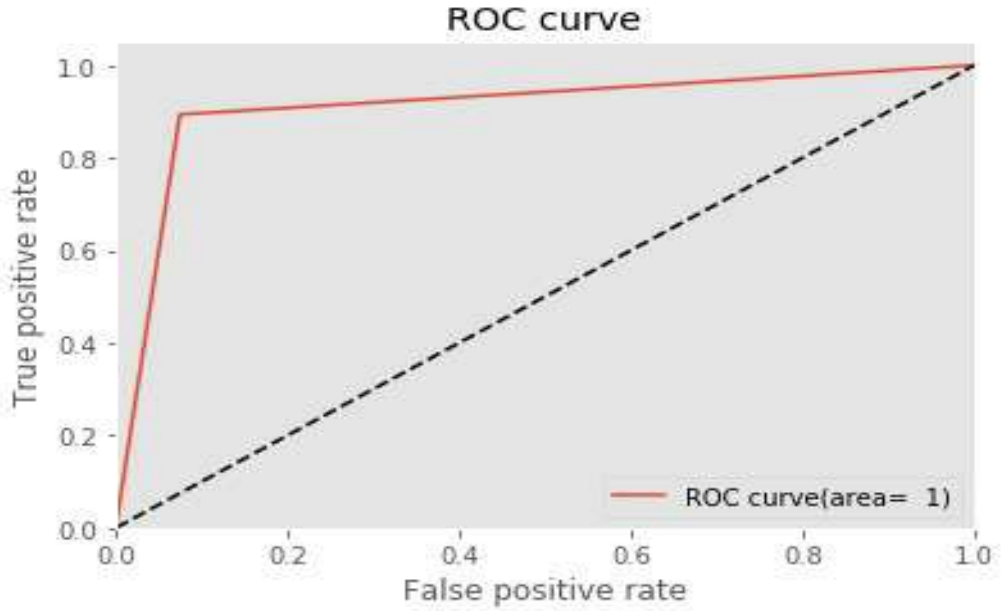
The data set used in this work WEKA Tool [3, 9] is used for performing the experiment. We studied that WEKA is software which is designed in the country of New Zealand by the University of Waikato, which includes a collection of various machine learning methods for data classification, clustering, regression, visualization etc. One of the biggest advantages of using WEKA is that it can be personalized according to the requirements.

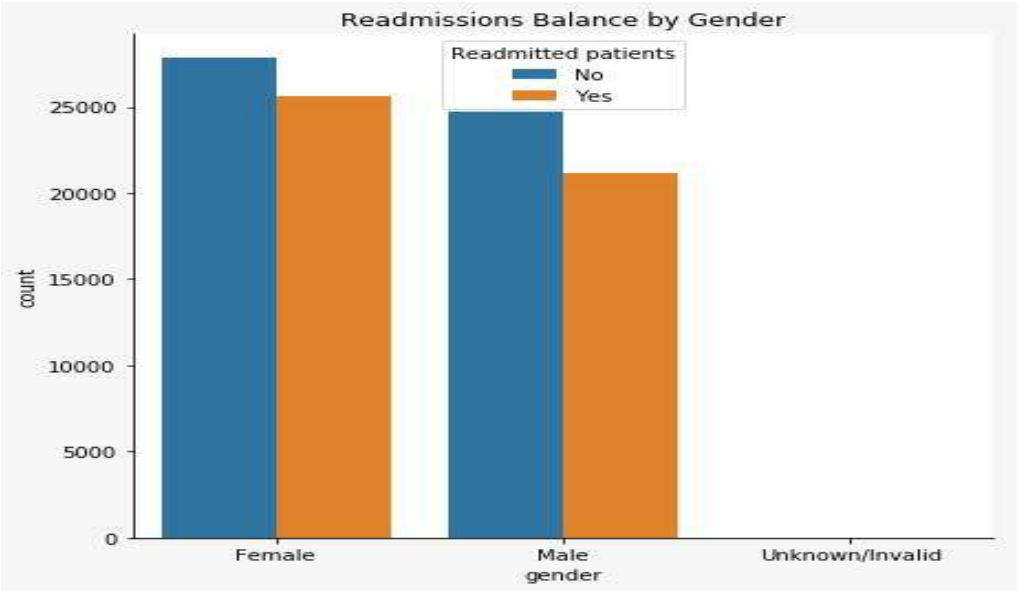
The main objective of using this dataset was to predict through diagnosis whether a patient has diabetes based on certain diagnostic measurements included in the data set. Many limitations were faced during the selection of the occurrences from the bigger dataset. The type of dataset and problem is a classic supervised binary classification.

The PIMA Indian Diabetes (PID) Data set having 9 (8+1 class attribute), 768 records describing female patients of which were 500 negative instance (65%) and 268 positive (35%). The detailed description of all attributes is given in table.

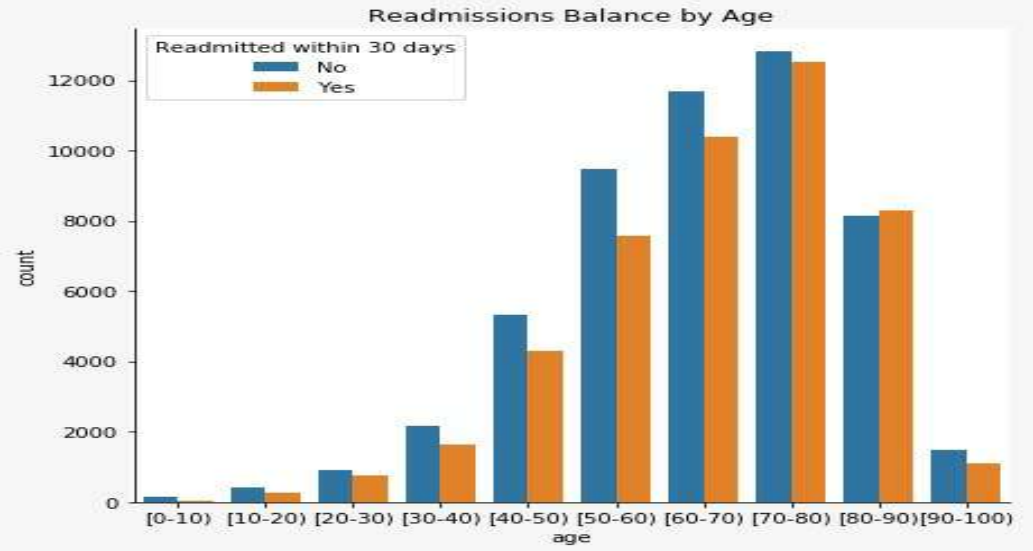


International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252





International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252



Data Preprocessing

Data preprocessing is an important step in the data mining process. The phrase garbage in and garbage out are particularly applicable to data mining and machine learning projects.

So, data is truly considered as a resource in today's world. As per the world economic forum by 2025 we will be generating about 463 exa-bytes of data globally per day, but is all this data fit enough to be used machine learning algorithms. In real world data are generally:

- Incomplete: Lacking attribute values, lacking certain attributes of interest or containing only aggregate data.
- Noisy: containing errors or outliers.
- Inconsistent: Containing discrepancies in codes or names.
- Mainly data pre- processing in machine learning is a crucial step that helps enhance the quality of data to promote the extraction of meaningful insights from the data .Data preprocessing in machine learning refers to some specific task in the real worlds which is :-

Data cleaning: it fills in the missing values; smooth noisy data identify or remove outliers and resolve inconsistencies.

- Data integration: It uses multiple databases, data cubes or files.
- Data transformation: Normalization and aggregation.
- Data reduction: reducing the volume but producing the same or similar analytical results.
- Data discretization: part of data reduction, replacing numerical attributes with nominal ones. So in simple words, data preprocessing in machine learning is a data mining technique that transforms raw data into an understandable and readable format. It is important to make the data more appropriate for data mining and analysis with respect to time, cost and quality [29].

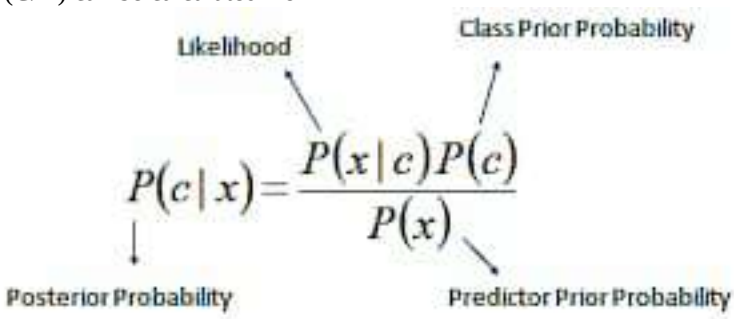
Applied Algorithms

1. Naïve Bayes classifier: Naïve Bayes is one of the powerful machine learning algorithms that is used for classification.it is an extension of the Bayes Theorem wherein each feature assumes independence. It is used for a variety of tasks such as spam filtering and other areas of text classification. It is based on conditional probability. It is considered as a powerful algorithm employed for classification purposes. It works well for the data with imbalance problems and missing values. Naïve Bayes [30] is a machine learning classifier which employs the Bayes Theorem. The Naïve Bayes model is a heavily simplified Bayesian probability model [31]. Mainly Naïve Bayes classification operates on a strong independence assumption [31].This means that the probability of one attribute does not affect the probability of the other. Nevertheless the results of the naïve bayes classifier are often correct .The work reported in [32] examines the circumstances under which the Naïve bayes classifier perform well and why? It states that the error is a result of three factors: 1.Training data noise 2.Bias and 3.Variance.

Training data noise can only be minimized, by choosing good training data. The training data must be divided into various groups by the machine learning algorithms. Bias is the error due to grouping in the training data being very large. Variance is the error due to those groupings being

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252

too small. We can use these bayes theorem posterior probability P(C/X) can be calculated from P(C),P(X),and P(X/C) [33].SO We can write;



$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

2. ANN: The Artificial Neural Network (ANN) is a specific area of artificial Intelligence and an important technique which is used in data mining. We can say that the Artificial Neural Network or ANN is an information processing paradigm that is inspired by the way the biological nervous system such as brain process Information. It is composed of large number of highly interconnected processing elements (neurons) working in unison to solve a specific problem. This Model depends on the complexity of the systems to achieve the purpose of processing Information by adjusting the relationship (between the internal Node) [Mukaietas 2012] [34].

Artificial Neural network (ANNs) as "Biologically Inspired Computing code with the number of simple highly interconnected processing elements for simulating (only an attempt) human brain working & to process Information Model". A Neural Network is an oriented graph. It consists of nodes which in the biological analogy represent neurons, connected by arcs. It corresponds to dendrites and synapses. Each arc is associated with a weight while at each node. Apply the values received as Input by the node and define. Activation functions along the incoming arcs, adjusted by the weights of the arcs. A neural network is a machine learning algorithm based on the model of a human neuron. The human brain consists of millions of neurons.

It sends and processes signals in the form of electrical and chemical signals. These neurons are connected with a special structure known as synapses. Synapses allow neurons to pass signals.

An Artificial Neural Network is an information processing technique. It works like the way the human brain processes Information. ANN includes a large number of connected processing units that work together to process Information. They also generate meaningful results from it.

We can apply Neural Network not only for classification. It can also apply for regression of continuous target attributes.

A neural network may contain the following 3 layers:

a) Input Layer: The activity of the Input units represents the raw information that can feed into the network. Usually the number of Input nodes in an Input Layer is equal to the number of explanatory variables.

b) Hidden Layer: To determine the activity of each hidden unit. The activities of the input units and the weights on the connections between the input and hidden units. There may be one or more hidden layers.

c) Output Layer: The behavior of the output units depends on the activity of the hidden units and the weights between the hidden and output units. It returns an output value that corresponds to the prediction of the response variable. The objective of ANN is to covert input into significant output. Input is the combination of a set of Input values that are associated with the weight vector, where the weight can be negative or positive. This is a function that sums the weight and maps the result to the output as for, Y=W1X1+W2X2.

The influence of a unit depends on the weighting: Where the input signal of neurons meets is called the synapse. ANN works for both technologies, supervised and unsupervised learning. Here we used supervised learning in our present study because the output is given to the model. In supervised learning is the machine learning task of learning a function that maps an Input to an output based on example Input-output pairs [35]. In this learning technique both Input and output are

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252

known. After processing the actual output compared with required outputs. Errors are then back propagated to the system for adjustment. During tracing the data is processed many times. So that, the network can adjusts the weights and refine them [36].

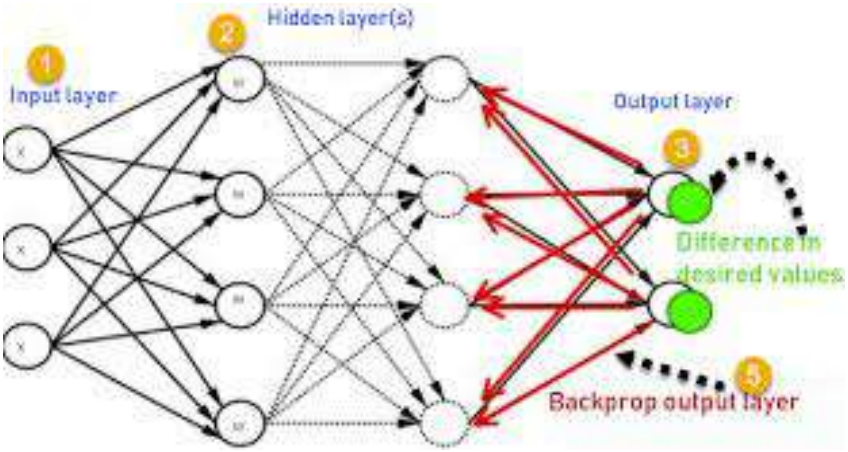


Fig :(Resource fig data flair trining)

3. Random Forest: Random Forest Method is a flexible, easy to use machine learning algorithm that produces, even without hyper parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get. The larger the number of trees, the more accurate the result. Random forests are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multiple of decision trees at training time and out-putting the class that is the mode of the classes. Classification or mean prediction (regression) of the Individual tree [37]. Random decision forest corrects for decision trees' habit of over fitting to their training set [38].

4. Logistic regression (L.R): Logistic regression is a mathematical modeling approach that can be used to describe the relationship of several dichotomous dependable variables. In other words we can say that it is a regression which is used to assign observation to a discrete set of classes, mainly logistic regression is a type of machine learning which is based on concept of probability.

Other modeling approaches are possible also but logistic regression is for the most modeling procedure used to analyze epidemiologic data. The fact that the logistic function (f) ranges between '0' and '1' is the primary reason the logistic model is so popular. The model is designed to describe a probability which is always some number between '0' and '1'. In epidemiologic terms such a probability gives the risk of an individual getting a disease.

The logistic model therefore is set-up to ensure that whatever estimate of risk we get it will always be some number between '0'and '1'.An important feature of the logistic model is that it is defined with a follow up study orientation. This model describes the probability of developing a disease of interest expressed as a function of independent variables presumed to have been measured at the start of fixed follow up period [39], for this reason it is natural to wonder whether the model can be applied to case control or cross sectional studies.

5. S.V.M: Support vector machine algorithm can classify both linear and non-linear data. It first maps each data item into n- dimensional features. It then identifies the hyper-plan that separates the data items into two classes while maximizing the marginal distance for both classes and minimizing the classification errors [40].The marginal distance for a class is the distance between the decision hyper-plane and its nearest instance which is a member of that class. More formally, each data point is plotted first as a point in an n-dimensional space (where 'n 'is the number of features) with the value of each feature being the value of a specific coordinate. To perform the classification, we then need to find the hyper-plane that differentiates the two classes by the maximum margin.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252

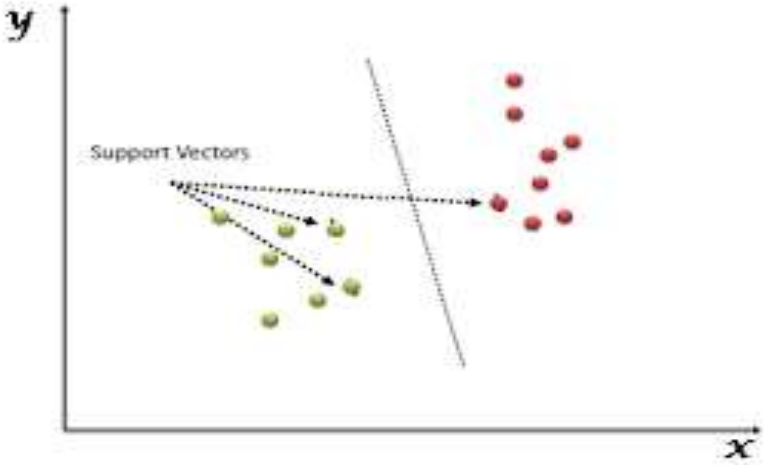


Fig 3. (Resource analyticsvidhya.com)

7. KNN: KNN is a very simple data mining technique and used for classification and parametric methods. Mainly KNN is a sort of instance based learning, also referred as lazy learning, which basically aims with estimating the function locally and all computation is postponed classification [40]. It can be beneficial to allocate weight to the contributions of the neighbors, so as to the closer neighbors continue more to the average than those who reside more far away . The distance is mostly measured by using the Euclidean distance formula. Here k is static value and mostly it takes an odd value like 1, 3, and 5. K –folds cross validation technique is used for training data. This technique is mostly used in circumstances wherever the aim is prediction and we wish to evaluate how a predictive model in practice will perform especially in terms of accuracy. In the prediction problem, a model is generally fed with a dataset that contains known data instances on which training is done , as well as a data set of anonymous data against the model being tested, the so -called testing dataset. This technique is used to assess predictive models by dividing the original sample data set into a training set that is used to train the model and a test set on which testing to evaluate it.

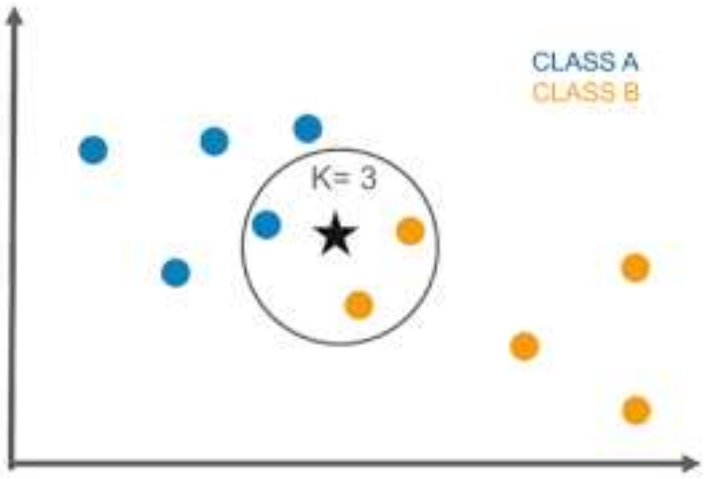


Figure: (Resource edureka.com)

Accuracy Measurement tool with Formula

In this study we used various measurement identity such as sensitivity (SN), SPECIFICITY (SP), accuracy (ACC), PRECISION (P), RECALL (R), F-MEASURE, Receiver operating curve (ROC) to measure the classified effectiveness and the formulas as given below:

a) SN = TP/TP+FN
b) SP = TN/TN+FP
c) ACCURACY (ACC) = TN+TP/TN+TP+FP+FN

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252

d) TP (TOTAL SAMPLE) COUNT = TN+TP+FP+FN
e) PRECISION (P) = TP/TP+FP
f) RECALL (R) = TP/TP+FN
g) F- MEASURE = 2*(P*R)/P+R

Experimental result

Table 2: Random forest

	Predicted			
	POSITIVE		**NEGATIVE**	
TRUE	TRUE POSITIVE	175	FALSE NEGATIVE	93
FALSE	FALSE POSITIVE	94	TRUE NEGATIVE	406

Table 3: NAÏVE BAYES

Actual	**Predicted**			
	POSITIVE		**NEGATIVE**	
TRUE	TRUE POSITIVE	133	FALSE NEGATIVE	135
FALSE	FALSE POSITIVE	96	FALSE NEGATIVE	404

Table 4: ANN

	Predicted			
	POSITIVE		**NEGATIVE**	
TRUE	TRUE POSITIVE	178	FALSE NEGATIVE	90
FALSE	FALSE POSITIVE	100	TRUE NEGATIVE	400

Table 5: S.V.M.

	Predicted			
	POSITIVE		**NEGATIVE**	
TRUE	TRUE POSITIVE	180	FALSE NEGATIVE	86
FALSE	FALSE POSITIVE	115	TRUE NEGATIVE	385

Table 6: KNN (K=3)

	Predicted			
	POSITIVE		**NEGATIVE**	
TRUE	TRUE POSITIVE	183	FALSE NEGATIVE	85
FALSE	FALSE POSITIVE	116	TRUE NEGATIVE	384

Table 7: logistic regression

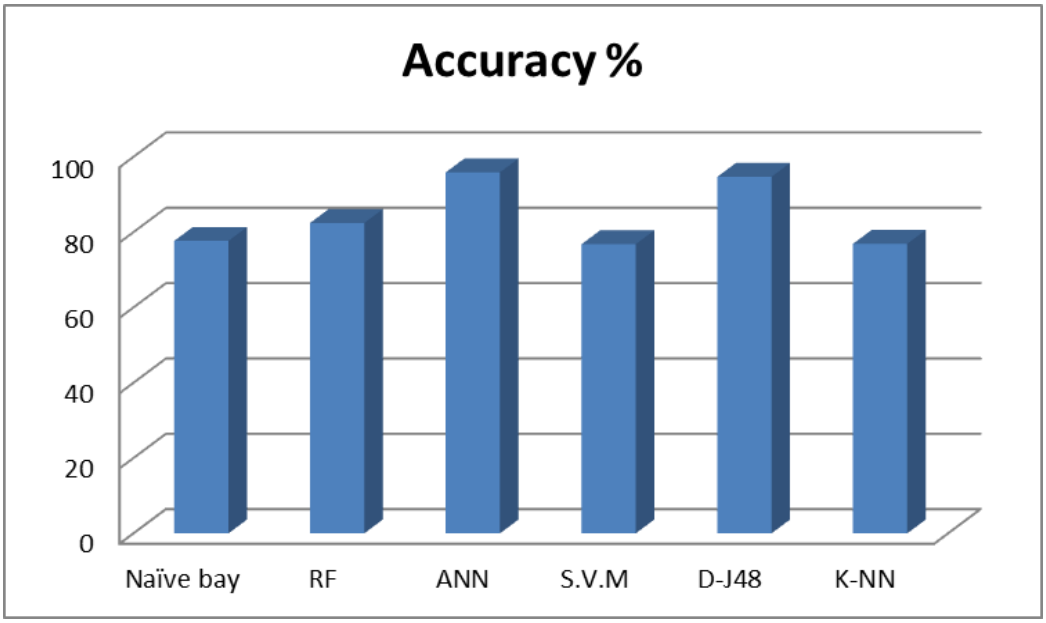
	Predicted			
	POSITIVE		**NEGATIVE**	
TRUE	True positive	186	False negative	82

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252

FALSE	False positive	98	True negative	402

Table 8: Comparison of all classifiers performance

Classifiers	Precision	Recall	F-measure	Accuracy	ROC–AUC	Misclassification
Naïve bay	0.850	0.857	0.848	85.71	0.836	14.29
RF	0.830	0.928	0.89	92.85	0.94	7.15
ANN	0.938	0.941	0.949	94.14	0.93	5.86
S.V.M	0.790	0.896	0.84	89.61	0.90	10.9
L.R	0.88	0.88	0.88	83.11	0.80	16.89
K-NN	0.78	0.91	0.84	90.1	0.89	9.9



International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1547-1562 www.ijaem.net ISSN: 2395-5252

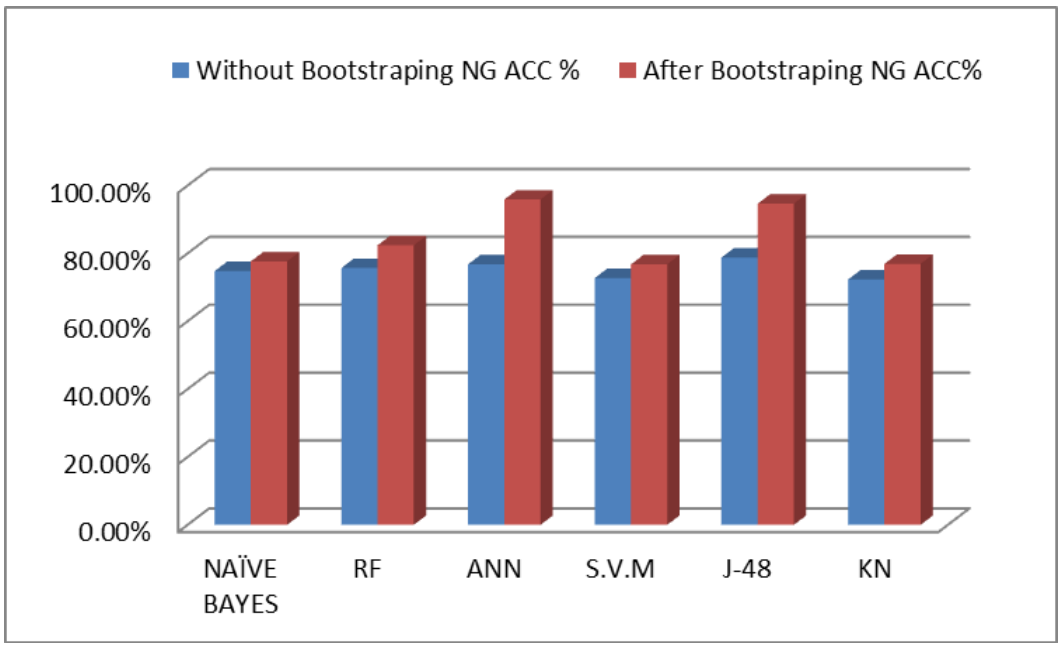


Table 9: Comparison before and after applying resample methodology

Classifiers	Without Bootstrapping ACC %	After Bootstrapping ACC%
NAÏVE BAYES	74.8	85.71
RF	75.7	92.85
ANN	76.8	94.14
S.V.M	72.75	89.61
L.R	78.77	83.11
KN	72.32	90.1

We compare our experimental results achieved in the study with the results reported by other researchers in the existing literature

Methodologies	Accuracy With Finding	Reference Originator
D.T, KNN, RF and S.V.M	Before preprocessing D.T Showed accuracy of 73.82%	Pradeep et al. [41]
K-NN, LR and J-48	j-48 gave accuracy of 78.27%	Xue-Hui-Min et al. [42]
S.V.M, J48, Cart and KNN	j-48 prediction showed better output with accuracy 67.15%	Saravananathan and velmurugan [43]
D.T	D.T has accuracy 74.80%	Iyer et al. (44)
D.T, N.B and S.V.M	N.B has accuracy 73%	Sisodia et al. [45]
N.B, RF and ANN	ANN has highest accuracy 75.60%	Alam TM et al. [27]

N.B, RF, ANN, KNN, S.VM and D.T	ANN has ACC 95. 87%	Current Observation Result

III. CONCLUSION AND FUTURE WORK

Our main purpose of this research paper is how to detect diabetes and their prediction with the help of various machine learning classifiers. We have also used data mining techniques for the enhancement of the accuracy of the predictive model. We enhance the accuracy by improving the data in the preprocessing phase that really works well. For the purpose of better accuracy we used bootstrapping resample technique on the PIMA INDIAN data set, which will increase the accuracy of almost all classifiers but the decision trees lead over others. Accuracy of models is highly dependent on the dataset, so this technique works very well on PIMA INDIANS Diabetic data set but may not guarantee the same results on a different dataset. In this study we used various evaluation matrix parameters in terms of precision, recall, accuracy, f-score, and misclassification. The results show that ANN classifier performed better with the highest accuracy 94.14%. This technique can be extended to enhance the accuracy of higher classifiers like, G.A, E.A, and ensemble machine learning technique.

REFERENCES

[1] Lonappan A, Bindu G, Thomas V, Jacob J, Rajasekaran C, Mathew KT. (2007) Diagnosis of diabetes mellitus using microwaves. J. Electromagnetic wave 21, 1393-1401, doi :10.1163/156939307783239429.

[2] Krasteva A, Panov V, Kisselova A, Krastev Z. (2011) Oral cavity and systemic diseases Diabetic Mellitus Biotechnol EQUIP 25, 2183-2186 doi: 10.5504/BBEQ 2011.0022.

[3] Tao Z, Shi A, Zhao J. (2015) Epidemiological perspectives of diabetes cell Biochem, BIOPHYS 73:1815

[4] Organization WHO, world health statistics (2016) monitoring health for the SDGS SUSTAINABLE DEVELOPMENT GOALS ,WHO.

[5] Falvo D and Holland BE (2017) Medical and psychosocial aspects of chronic illness and disability jones $ Bartlett learning.

[6] Skyler JS, Bakris GL, Bonifacio E, Darsow T, Eckel RH, Groop L. (2017) differentiation of diabetes by pathophysiology, natural history and prognosis, Diabetes, 66:241-55.

[7] Diwani S, Mishol S, Kayange DS, Machuve D, Sam A. (2013) Overview applications of data mining in health care, the case study of Arusha region, International journal comp Engg research 373-7.

[8] IDF DIABETES ATLAS Ninth edition 2019 (www.diabetic atlas.org)

[9] Vost, Flaxman AD, Naghavi M, Lozano R, Michand C, Ezzaim. (2012) for 1160 sequence of 289 diseases and in juries (1990-2010) a systematic analysis for the global burden of disease study 2010 Lancet 380 (9859)2163-96 doi: 10/1016/s0140-6736 (12)/617292

[10] Shobac KDG, Gardner D, eds (2011) chapter 17 green spans basic and clinical endocrinology, new york mcgraw hill medical .ISBN- 978-0-07-1622431.

[11] The top ten causes of death (www.WHO.int) 2020.

[12] American diabetic association (2018) economic cost of diabetes in the USA 2017 (doi: 10/2337/dci-18-0007 ISSN 0149-5992.

[13] Death and cost data & statics diabetic cds.gov (2019).

[14] Alam TM and Awan MJ. (2018) Domain analysis of information extraction technique, Int. journal multi disease engg. 91-9

[15] Alam TM, Khan MMA, Iqbal MA, Wahab A, Mushtaq M,Cervical. (2019) cancer prediction through different screening methods using data mining INTJ ADV COMPUTER SCI APPLI -10388-96.

[16] Kumar DA, Govindasamy R. (2015) performance and evaluation of classification data mining techniques in diabetes, INTERNATIONAL JOURNAL OF COMP SC AND INFORMATION TECHNOLOGIES 1312-1319.

[17] Vijayan V and Anjali C. (2015) Prediction and diagnosis of diabetes mellitus, A machine learning approach IEEE recent Advances in intelligent computational system (RAICS) 122-127 DOI: 10.1109/ RACIS2015.7488400

[18] Iancu I, Mota M, Jancu E (2008). Method for the analysis of blood glucose dynamics in diabetes mellitus patients, in proceedings of the IEEE INTERNATIONAL conferences on automation, quality and testing Robitics cluj Napoca DOI: 10.1109/AQTR 4588883.

[19] Cox M.E and Edelman D. (2009) Tests for screening and diagnosis of types -2 diabetes clin diabetes 27:132-138 doi:10.2337/diaclin 27.4.132.
[20] American diabetes association (2012): Diagnosis and classification of diabetes mellitus. Diabetes care 35 (supply 1), S64-S71 DOI: -10.2337/DCI 2-S064.
[21] Lee B.J and Kim J.y (2016) Identification of type-2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning IEEE J-Biomed health Inform 20, 39-46 doi:10.1109/JBHI 2015.
[22] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N V, lahavas I, Chouvarda I. (2017), machine learning and data mining methods in diabetes research computer structure biotech oj. 15,104-116 doi: 10.1016/j.csbj.2016-12.005.
[23] Alghamdi M, AL Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. (2017) predicting diabetes mellitus using SMOTE and ensemble machine learning approach the henry ford exercise testing (FIT). Project PL one 12.0179805 doi:10.1371/ journal.pone 0179805.
[24] Kumar P.S and Umatejaswi V. (2017) Diagnosing Diabetes using data mining technique, International journal of science and research publication 7, 705-709.
[25] Aljumah A A, Ahmad M.G, Siddique M.K. (2013) Application of data mining diabetes health care in young and old patients journal of kings and university – computer and information science 25, 127-136 doi: 10.1016/j.jksuci 2012. 10.003.
[26] Fatima M and Pasha M. (2017) Survey of machine learning algorithms for disease diagnostic journal of intelligent learning systems and application 09-1-16 doi:-10.4236/JILSA 2017.91001.
[27] ALAM TM, Iqbal M A, ALI A W, IJAZ S (2019) A model for early prediction of diabetes .www.elsever.com 16 2019 - 100204.
[28] A Gentle Introduction to Resembling technique Dale Berger Claremont graduate university.
[29] Xue-Hui, Meng, Yi-Xiang, Huang, Dong ping-Rao, Qiug Liu (2013) Comparison of three Data mining model for predicting diabetes of prediabetes by risk factors, Kaohsiung journal of medical science 2013,29,93-99.
[30] Rish I (2001) An empirical study of the Naïve Bayes classifier in IJCAI 2001 work shop on empirical methods in artificial intelligence IBM, pp-41-46.
[31] Wafa S, AI-sharafat, Riyadh Naoum (2009) development of Genetic based machine learning for Network intrusion detection world academy of science, Engg and technology 55, 2009.
[32] Naidu N and RV Dharaskar. (2010) An effective approach to network intrusion detection system using genetic algorithm , International journal of computer application 0975-8887 volume 1 ,N0-2 2010.
[33] Ray S (2017) 6 Easy steps to learn NAÏVE Bayes Algorithm.
[34] Mukal Y, Tanaka H, Yoshizawa M, Ikeda M. (2012) A computational Identification method for GPI anchored proteins by artificial neural network, Bio inform 7,125-131 doi: 10.2174/157489312/800604390.
[35] Stuart j. Russell, Peter Norvig. (2010) Artificial intelligence a model approach. Third edition prentice hall ISBN 9780136042594.
[36] Schalkoff RJ, artificial neural network vol1 new York Megraw hill 1997.
[37] Hotk (1998) The Random subspace method for constructing decision forests ieee transaction on pattern analysis and machine intelligence. 832-834 doi: 10.1109/34.709601.
[38] Hotinkam (1995) Random decision forest proceeding of the 3rd international conference on document analysis and recognition Montreal QC- 14-16 aug 1995 pp-278-282.
[39] Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome (2008) The elements of statistical learning. ISBN 0-387-952845.
[40] JOHN George H, patlangley, Estimating continuous distribution in Bayesian classifiers proceeding of the Eleventh conference on uncertainly in Artificial Intelligence morgan Kaufmann publication.
[41] Kandhasamy J, pradeep, Balamurali S. (2015) performance analysis of classifier models to predict diabetes mellitus, procedia computer science 47, 45-51.
[42] Olaniyi E.O, Adnan K. onset diabetes diagnosis using artificial neural network international j.sc eng 2014 ,5,754-759.
[43] Farran.B, Channanath A.M, Behbehani K, Thanaraj T.A. (2013) Predictive Models to assess risk of type-2 diabetes ,hypertension and comorbidity, machine learning

algorithm and validation using national health data from kuwat A Cohort study BMJ OPEN 2013 3, 24-57.

[44] A. Iyer J.S and R. sumbaly (2015) Diagnosis of diabetes using classification mining technique, IJDKP vol-5 no -1 pp-01-14-2015.

[45] Siodia D and Sisodia DS (2018) prediction of diabetes using, classification algorithms procedia computer sci- 2018 132/1578-85.