



Classification – A General theory and Efficiency in Medical Diabetes Dataset

Raja Aswathi R^{*1}, K Pazhanikumar²

^{*1,2}Department of Computer Science and Research Centre, S.T.Hindu College, Manonmaniam Sundaranar University, Tamilnadu, India.

^{*}Corresponding Author.

To Cite this Article: Raja Aswathi R^{*1}, K Pazhanikumar², “Classification – A General theory and Efficiency in Medical Diabetes Dataset”, Indian Journal of Computer Science and Technology, Volume 05, Issue 03 (September-December 2026), PP: 460-468.



Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution License; Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: Mining is the process of extracting efficient and correct information from a large sample of metadata. Data mining indicated gathering data from preexisting data sources. Classification is a branch of data mining method that is based on the supervised learning mechanism of categorizing data. This paper proceeds with the supervised learning algorithms that can exclusively deal with categorical and multi-variated data. The classification algorithms Linear Regression of Statically testing, Random Forest of Decision Tree and Rule Generation Decision Table are tested in 10 cross-fold validation to choose the best performing algorithm. It mainly focuses on the algorithm's tenacity in handling outliers in the medical field using the diabetes dataset. The Rule Generation-based Decision Table algorithm performs well with the multi-variate and categorical diabetes dataset with an accuracy of 81.1% in result calculation.

Key Word: Data Mining, Classification, Prediction, Decision Tree, Perceptron.

LINTRODUCTION

In today's new economy, technology leads nations to have better growth in various fields like information technology, scientific research, medical science, business management, space research, automobile industry, agriculture, etc. Data mining is one of the fast-moving fields in recent times; where it's useful for solving many work-oriented problems and problems based on theories and mathematical concepts. Mining data is the process of filtering out useful information from a large existing potential database. With data mining and its machine learning algorithms, a lot of applications have been developed successfully to enhance the work more effortlessly. The data mining techniques are of various kinds that are best suited for all fields. The data mining techniques are classified into two main branches supervised and unsupervised learning. (1) Supervised learning is the process of understanding the given input and then categorizing them into their specific classes by using the predefined classes that are already classified. (2) Whereas unsupervised learning is the process of grouping the data value into an appropriate class using the existing mechanism [1]). The supervised data mining task includes classification, regression, prediction models and time series analysis as their major subgroup in learning data from already existing predefined categories. Unsupervised learning includes learning the patterns and the existing relationship among the attributes without considering the new outcome in the present data, it groups them based on their similarities. Clustering, association rule mining, sequential mining and summarization are the subgroups of unsupervised learning.

Some of the drawbacks in the traditional problem-solving procedures have been overcome using data mining methods. With an accurate data mining process, the hard and time-driven tasks can be reduced efficiently. The workload that is once handled by humans is now transferred to machines. Even though the purchase of hardware and software that are suitable for a heavy workload and large disk storage space may even end up paying a high price when compared with the traditional model when the problems are provided with a specifically larger dataset to be processed then only the machines with high-end speed and processor are suitable for calculation.

With the advantage of data mining machine learning algorithms, the processing speed and problem-solving techniques are enhanced in computations. A variety of data types like numerical, categorical, sequence, temporal, and sequential data with temporal attributes are available in the real world. The machine learning algorithms in data mining are built to handle these kinds of data specifically. Some algorithms are individually assigned to work with unique data types. Improving innovative proposals in data mining machine learning techniques in recent times has contributed a lot to the reduction of error rate in the outcome. Machines that are employed using logical and computational tasks use machine learning algorithms for the smooth performance of applications. Data mining alone has its part in the current developed technology. This paper discusses the variety of major methodologies of data mining classification algorithms with their unique characteristics in the performance and categorization of data. Every algorithm has its advantages and disadvantages that can be overcome when the algorithm works in conjunction with some feature selection methods, hybrid topology, ensemble methods, etc. The hierarchy of classification algorithms in data mining is represented in figure 1.

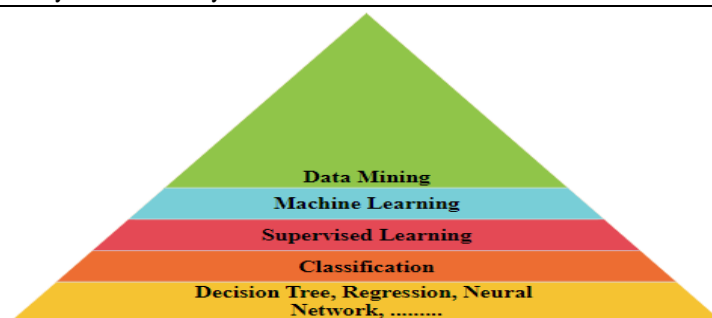


Fig.1.The hierarchy of classification algorithms in data mining.

The data sources for processing the algorithms are taken from the medical dataset repositories to contribute to disease identification at an early stage and take precautions for the further improvement of health. Due to the increase in health-damaging diseases and growth in population, high potent applications are required in hospitals to identify the risk of particular disease far faster using the classification mechanisms. There is a vast range of health problems in humans like heart disease, cancer, thyroid, etc. and the data mining methods also involves predicting pregnancy symptoms and any other clinical factors in any meaningful time without being waiting for hours or days. Major common data sets used in classification problems and most necessary care and speedy technology are required for the problem solving of medical factors are described in figure 2.

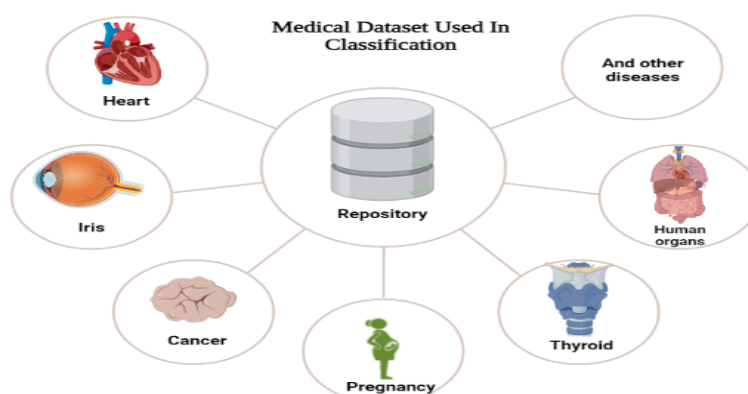


Fig.2.The medical Datasets that are employed in machine learning.

II. CLASSIFICATION

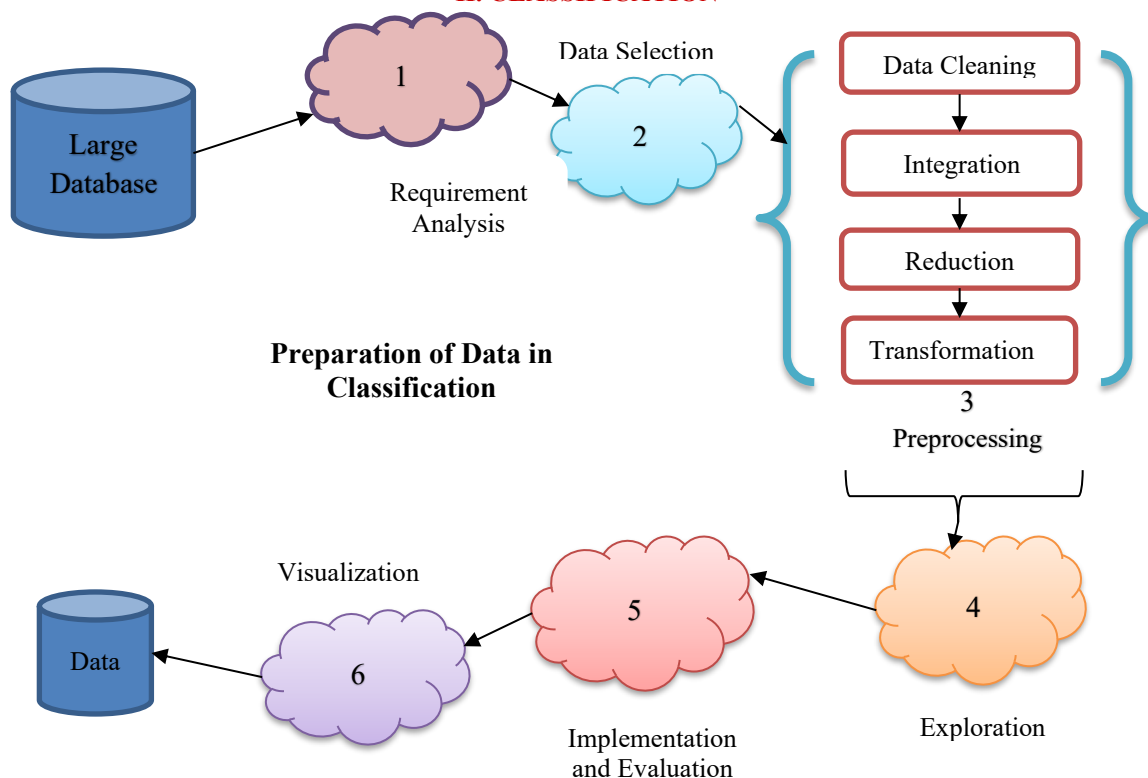


Fig.3. Data Preparation Stages in Data Mining

Definition: It pigeonholes (contains one or more records of data) the provided data according to the features of predefined classes.

Classification is a data mining methodology that is based in the category of supervised machine learning processes to predict the output with better accuracy [2]. In classification, all the records are filtered according to the existing class attributes, so it is hard to find outliers in the classified data. More accurate results are obtained in classification than other cluster-based algorithms. Before applying the data to the classifier algorithms the data must enter the processing stage for better results. The data preparation stages are described in figure 3.

2.1 Requirement Analysis

This phase works in gathering the required data that are relevant for the problem-solving activity. Here the unwanted information and irrelevant data are not taken into consideration. By analyzing the requirement of the working project, the objective is known and it reduces the excess time used in accumulating information.

2.2 Data Selection

Data Selection includes the selection of data with the required data types that match the requirement. Data selection is also regarded as the collection of data from primary sources such as repositories, literature surveys, questionnaires, schedules, etc. For a better understanding and evidence of collected data, a very good data collection instrument with accurate evaluation procedures must be followed. Thorough evidence of documentation of selected data and adding the circumstances in which the data is specified will always help in the future when any survey-related problem arises.

2.3 Preprocessing

Preprocessing is one of the crucial processes in accumulating data that are of high calibre with perfect accuracy, consistency, and completeness in the occurred data. Most of the time that is required in the completion of data preparation is engulfed in this process. The current stage in data machine learning is a technique to gather the data that are fundamentally suited for project building. Data preprocessing steps include,

(i) Data Cleaning

Data cleaning includes the removal of noisy data, identifying the outliers, making incorrect values of data correct, redundant data, fixing the structural errors and replacing the missing value in the presented data by using the data similarity measures. Cleanin large oversized data into a beneficial one can improve the computational rate. Only a robust dataset can avoid future problematic errors.

(ii) Data Integration

A single cohesive type of data obtained was from this process. It uses integration software like on-premise data integration tools, cloud-based and open-source integration tools, and proprietary data integration tools. The main goal of data integration is to avail information to everyone that can be easily accessed, scalable, secure, real-time, and processed by anyone without any other substantial measures.

(iii) Data Reduction

Data Reduction optimizes or compresses the data by reducing the data volume or original content with its attributes into a simple readable format based on the storage capacity allotted for that specific data. The reduction phase in preprocessing is processed by using the freely available compression mechanism and other analytical algorithms.

(iv) Data Transformation

Includes the transformation of data structures, format, mapping, code sequence, code generation and generation, review and values of the source system into another data format. The transformation of data can be complex or simple based on the characteristic of data from the database. Data are transformed from traditional filtered, interactive, and batch data.

2.4 Exploration

After the preprocessing stage where the data is collected and cleaned the exploration stage begins. Once the data is preprocessed then it is subjected to some of the data mining tools that work to fulfil the needs of an enterprise. In this phase, a lot of data mining tools are used to explore the effectiveness of collected data and then validated for their robustness.

2.5 Implementation and Evaluation

Once the information model has been validated, the software model is implemented and evaluated to measure performance. A better evaluation procedure is taken to improve evaluation and choose a good data mining technique. Accuracy and effectiveness are evaluated in the presented model.

2.6 Visualization

Giving out the result obtained in a visualized manner is the major task in data mining that are handled by the decision-makers. Without a quality visualization of how the software model work, all the energy deployed in the project will go unseen. So for a clear-cut understanding of how the model works, how accurate it is in the field survey, and must represent

how it makes the easier must be visualized through screen display or printouts.

The Classification algorithm uses the output from phase three, the preprocessed data and apply them to the developed algorithms. The classification algorithms are categorized as statistical-based algorithms, distance-based algorithms, decision tree-based algorithms, neural network-based algorithms, and rule-based algorithms based on the metrics used to process the information.

III. METHODOLOGIES

3.1 Statistical Classification

Linear Regression

Linear Regression is a classification algorithm that gives an estimated result based on the input values. The regression is used in various other fields for forecasting, building strategies, sales, finding the relationship between variables, etc. The formula for linear regression is

$$y = \lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_n x_n \quad (1)$$

Here $\lambda_0, \lambda_1, \lambda_2 \dots \lambda_n$ are the regression coefficient of the relation. The value of 'y' can be determined by substituting the regression coefficient values to the equation (1). With more accurate data the regression method makes out a fine linear model. If the data presented is contaminated with errors, outliers and noise; the generated linear model may be inadequate. The error rate in the data can be identified by equation (2)

$$y = \lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_n x_n + e \quad (2)$$

Where 'e' is the error rate variable substituted with the linear regression. To minimize the error rate the least square error mechanism is formulated,

$$LSE(e) = \sum_{i=0}^n e_i^2 \quad (3)$$

The value of 'e_i' can be found from equation (2) as,

$$LSE(e) = \sum_{i=0}^n (y_i - \lambda_0 - \lambda_1 x_{1i})^2 \quad (4)$$

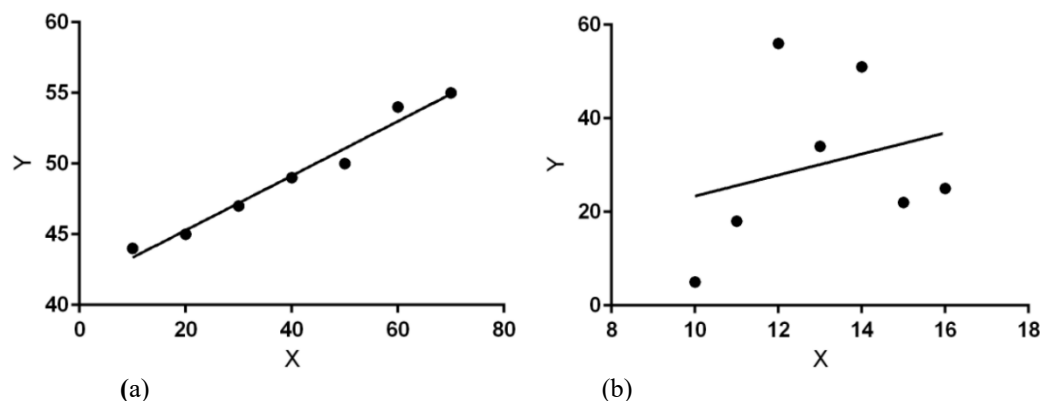


Fig.4. (a) A simple linear regression (b) A poor fit linear regression

Bayesian Classifier

The Naïve Bayes methodology is a classification technique that is based on the rule of conditional probability called the 'Bayes Rule'. Every prediction that is done on classification will always be influenced by the attributes that are available to make the prediction a successful one. The Bayes can solve datasets with high dimensions and predicts based on probability. The formula for naïve Bayes theorem is

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)} \quad (5)$$

Where $P(A|B)$ is the posterior probability.

$P(B|A)$ is the likelihood probability.

$P(A)$ is the prior probability.

$P(B)$ is the marginal probability.

As compared to other algorithms it efficiently does multiclass prediction and is mostly suited for text classification. It can also be used in real-time prediction, medical evaluation of diseases, spam identification, etc.

3.2 Distance-Based Classification

k-NN

The k-NN (k Nearest Neighbor) is a non-parametric classification algorithm based on the distance relationship or

similarity between the attribute values. It can also be worked out with regression methodology for analysis. This methodology categorizes the data into a specific group based on their distance value and gives out a required format that is necessary to understand the data. If the given k data sample is similar to the preexisting data samples, it then categorizes the predicted result set in that group. It prefers the Euclidean Distance to calculate the distance metric.

$$\text{Euclidean Distance, } E(x, y) = \sqrt{\sum (x_i - y_i)^2} \quad (6)$$

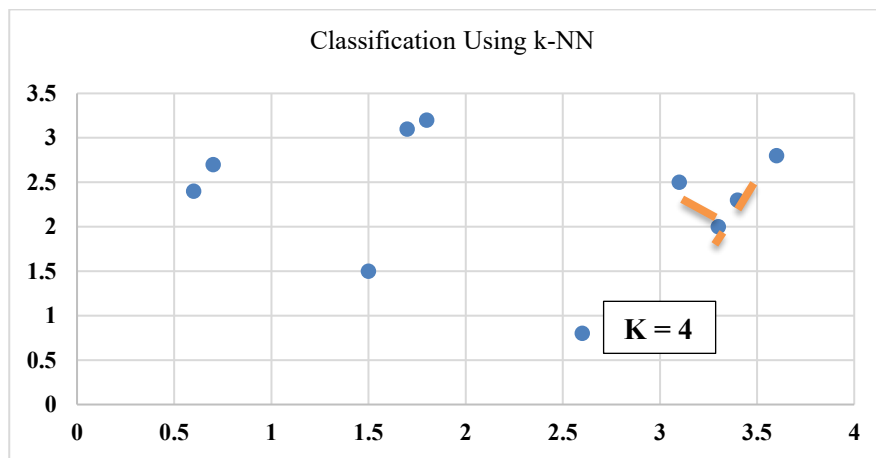


Fig.5. Classification of k-NN with the value k=4.

As shown in figure 5 the closer four values are grouped into one class. When a large group of data are prepared for classification by setting the k value to a smaller one it classifies the data into multiple classes with a very accurate closer relationship between the relations set. By increasing the k value, multiple categorizations can be reduced when proper input classes are specified.

3.3 Decision Tree-Based Algorithms

The decision tree is a classification and regression technique that employs a tree data structure for dataset categorization [3] class. It can deal with categorical and continuous datasets and classifies the data from the root of each tree by providing necessary criteria for classification. The most common decision tree terminology algorithms that are particularly famous in various are ID3, C4.5, CART, PART, Random Forest, etc.

ID3

ID3 is a decision tree based classification algorithm that's based on the information theory of entropy (Ye et al., 2021). ID3 uses the entropy value to predict the Information Gain of the attributes.

$$\text{Entropy (P)} = \sum_{i=1}^s ((p_i) \log \left(\frac{1}{p_i} \right)) \quad (7)$$

$P = (p_1, p_2, p_3, \dots, p_i)$ is the probability distribution value of the attributes. The information gain of attributes is calculated from the following formula,

$$\text{Gain (P, A)} = \text{Entropy (P)} - \sum_{i=1}^s ((p_i) \text{Entropy (p}_i)) \quad (8)$$

The entropy value of attributes is used to identify the uncertainty in the data values which ranges from 0 to 1. If the value of $E(P)$ tends to be 0 then the number of the class is one and there is no probability.

C4.5

The C4.5 is a decision tree-based classification method, which is an extended version of the tree algorithm ID3. Since the ID3 cannot generate a better tree solution for the dataset if the dataset is too large, so here the C4.5 can handle any smaller or larger databases. It can also handle missing values by adjusting matching data values with the missing record. The information Gain value from ID3, Gain (P, A) is followed up in C4.5 then combined with the split information to figure out the Gain Ratio.

The split information of the dataset D is,

$$\text{SplitInfo (D)} = \sum_{i=1}^s (|D_j| / |D|) \log(|D_j| / |D|) \quad (9)$$

From the above equation (9) the split information value is calculated for all attributes in the dataset D. The Gain (P, A) is divided by the SplitInfo (D) to get the GainRatio (D, A) ,

$$\text{GainRatio (D, A)} = \frac{\text{Entropy (P)} - \sum_{i=1}^s ((p_i) \text{Entropy (p}_i))}{\text{SplitInfo (D)}} \quad (10)$$

The C4.5 can deal with continuous data values along with sub-tree replacement where the sub-tree is replaced by a

leaf node and sub-tree raising where the leaf node in the tree is replaced by a sub-tree in the C4.5 tree structure. It then uses the rule generation or decision tree for the grouping of data into their respected classes.

Random Forest

Random Forest is a decision tree algorithm that uses the supervised method of learning and can also be employed with the regression models. Random Forest is mostly preferred to handle large multiclass attributes because of its high accuracy in data classification. The forest tree is a merge of several trees which allows each tree to randomly the information from datasets like bagging. In the decision tree, the data features depend on the possibilities, but in Random Forest the features of sub-trees are chosen in a random format, which reduces the correlation across trees.

3.4 Neural Network-Based Algorithms

The Artificial Neural Network (ANN) is an imitation of the Biological Neural Network to perform the task similarly to how the neurons in the human brain do. The Neural Network (NN) is made up of three layers the input layer, hidden layer, and the output layer [4]. Some of the common supervised learning neural algorithms are Perceptron [5], Back Propagation Neural Network, Radial Basis Function Network, Adaptive Linear Neuron, etc.

Perceptron

The Perceptron is the simplest form of neural network with a single feed-forward system with three layers which constructs the architecture of a perceptron network. The first layer in the system is called the sensory unit or input layer, the second one is the associator unit or hidden layer and the final layer is called the response unit or the output layer [6]. Each layer in the system has its own functionalities to build for the purpose of learning. The input values $x_1, x_2, x_3, \dots, x_n$ in the sensory unit is then transferred to the associator unit with a weight adjustment associated with the input values.

The value of weight added to the sensory layer input values ranges from -1 to 1. After passing the input to the associator unit the perceptron learning rule is employed to calculate the weight adjustment $w_{11}, w_{12}, w_{13}, \dots, w_{1n}$, in the hidden layer. When respective weight adjustment and perceptron learning rule are applied to the data values in the hidden layer and it is transferred to the response unit and then to their according to classes and their error rates. The error rate $E_1, E_2, \dots, E_n$ is the difference between the predicted output values $Y_1, Y_2, \dots, Y_n$ from the response unit to the target value $T_1, T_2, \dots, T_n$ that is already predefined.

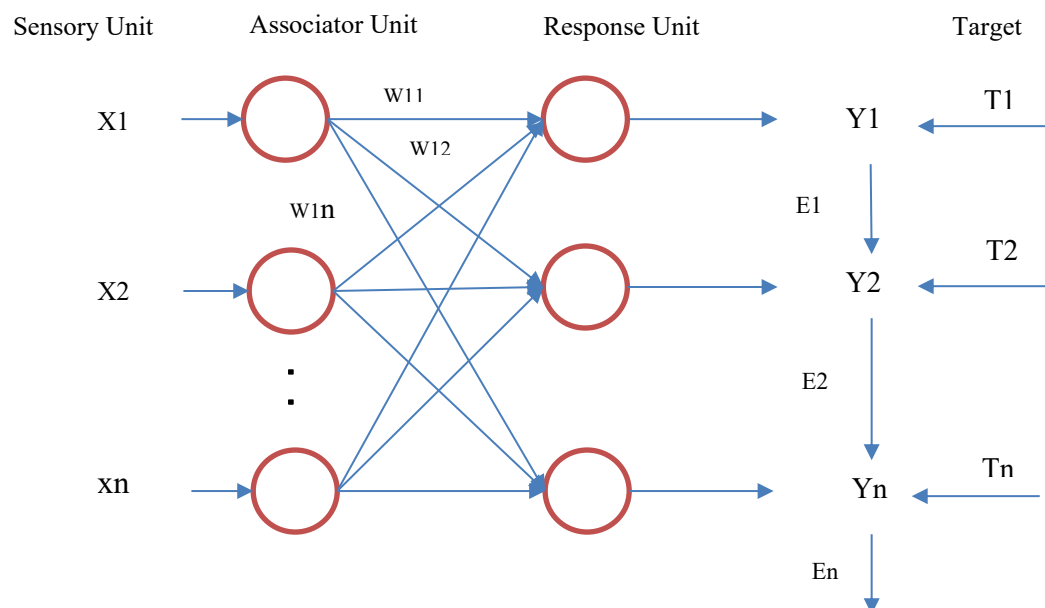


Fig.6. Perceptron Network

A Multi-Layer Perceptron (MLP) is a feed-forward Neural Network that contains single input and output layer but two or more hidden layers. When performing classification with multi-class data, the MLP is the best option to choose to have better accuracy and memory consumption.

3.5 Rule-Based Classification

A rule-based classification can be used in association with other deep learning or machine learning algorithms by setting rule conditions to predict the accurate classes. If a set of rules are generated for a given set of input data, it is quicker to classify the data record in their particular classes in an instant of milliseconds. A rule must be as concise as possible to boost up the accuracy. A larger and messy rule will never give an accurate result rather it will end up in inadequate classification which will mislead the program. The rule-based algorithm can be used in conjunction with a decision tree, neural networks and other methodologies for better output.

Rules are commonly employed with the ‘if ‘condition as ‘if avg_mark $\geq$ 80 then class = A’. A rule has no order and works by choosing a rule that suits best for the current data [7]).

Decision Table

The rule generation method works in conjunction with the decision table to update the generated rules in the form of a table. When a record satisfies a rule in the table, the data values are classified and added to the specified class of the particular rule.

IV. RESULT

The classification algorithms are tested for their efficiency by applying the diabetes dataset to the wide variety of algorithms described in the above section. Not all types of classification terminologies can handle every data type and so the diabetes dataset is best suited with the statistical-based, decision tree-based and rule generation-based algorithms.

Diabetes Dataset	
Dataset Characteristics	Multivariate, Time Series
Attribute Characteristics	Categorical, Integer
Number of Instances	768
Number of Attributes	9

Table.1. Diabetes Dataset

The 9 attributes in the diabetes dataset and their data value representation are specified below in a visualized format.

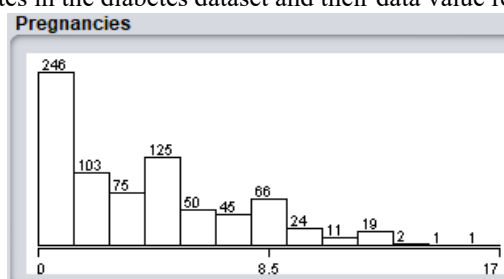


Fig.7. Visualization of Pregnancy Attribute

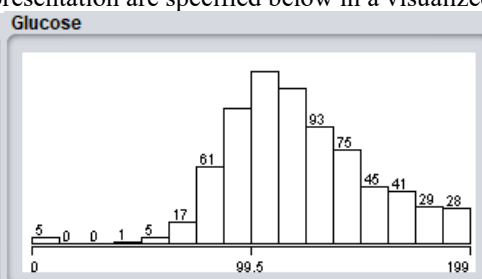


Fig.8. Visualization of Glucose Attribute

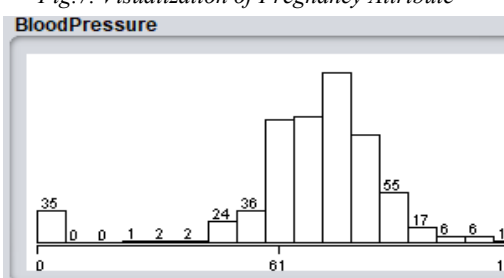


Fig.9. Visualization of BloodPressure Attribute

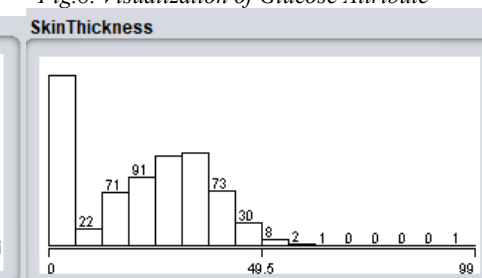


Fig.10. Visualization of SkinThickness Attribute

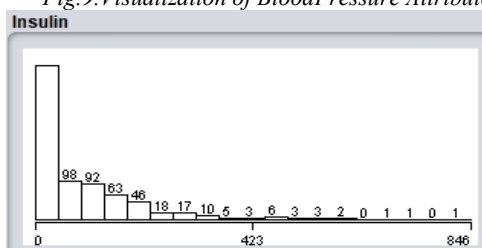


Fig.11. Visualization of Insulin Attribute

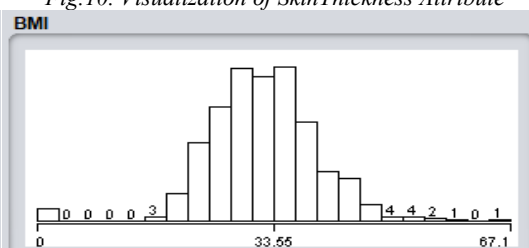


Fig.12. Visualization of BMI Attribute

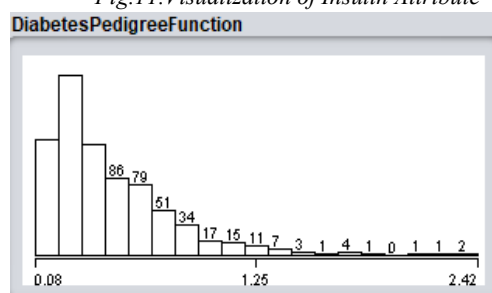


Fig.13. Visualization of PedigreeFunction Attribute

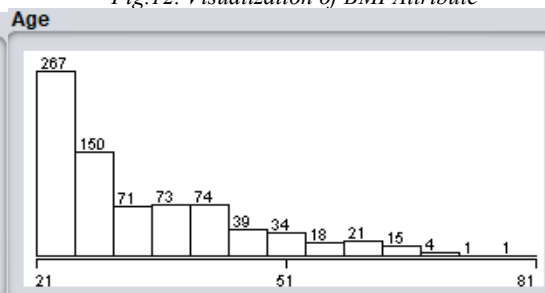


Fig.14. Visualization of Age Attribute

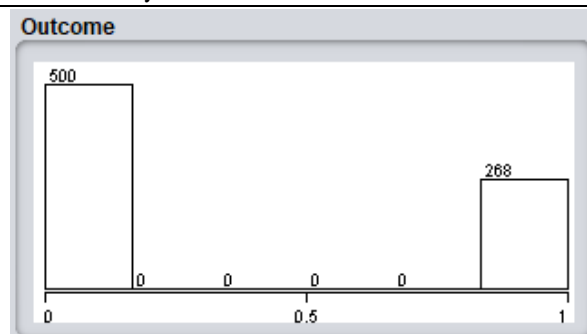


Fig.15. Visualization of Target Value

The execution result of the three best classification algorithms in a 10 fold cross-validation is depicted in Tables 2, 3, and 4 with their visualized format in the below figures.

Statistical-Based Linear Regression	
Correlation Coefficient	0.4998
Mean Absolute Error	0.3263
Root Mean Squared Error	0.4138
Relative Absolute Error	71.7425 %
Root Relative Squared Error	86.7359 %

Table.2.Evaluation of Linear Regression Model

Decision Tree-Based Random Forest	
Correlation Coefficient	0.536
Mean Absolute Error	0.3123
Root Mean Squared Error	0.403
Relative Absolute Error	68.6677 %
Root Relative Squared Error	84.4642 %

Table.3.Evaluation of Random Forest Model

Rule generation-Based Decision Table	
Correlation Coefficient	0.5322
Mean Absolute Error	0.3366
Root Mean Squared Error	0.4036
Relative Absolute Error	74.0119 %
Root Relative Squared Error	84.6013 %

Table.4.Evaluation of Rule Generation Model

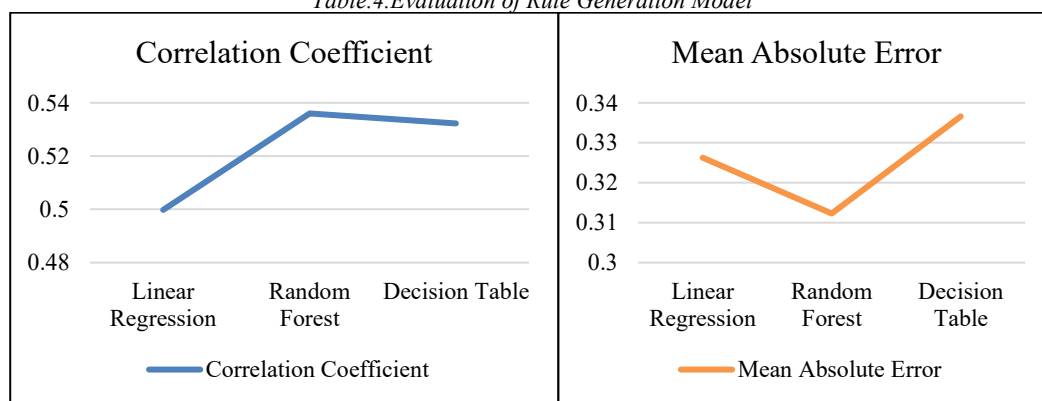


Fig.16. Correlation Coefficient

Fig.17. Mean Absolute Error

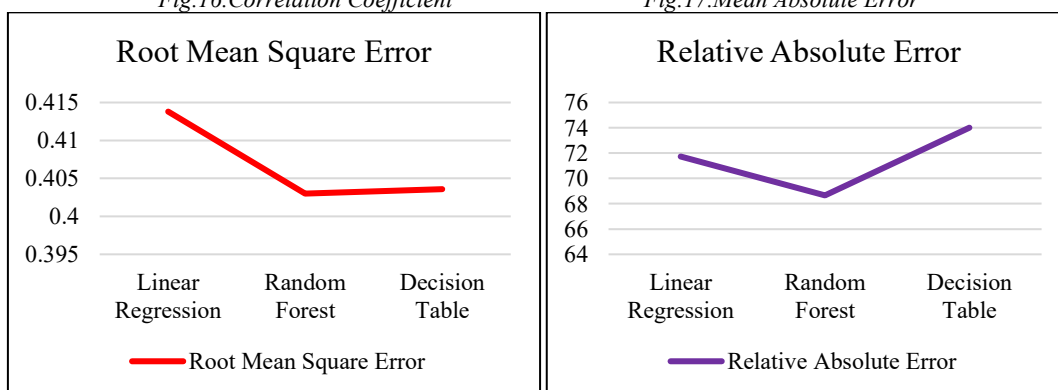


Fig.18. Root Mean Square Error

Fig.19. Relative Absolute Error

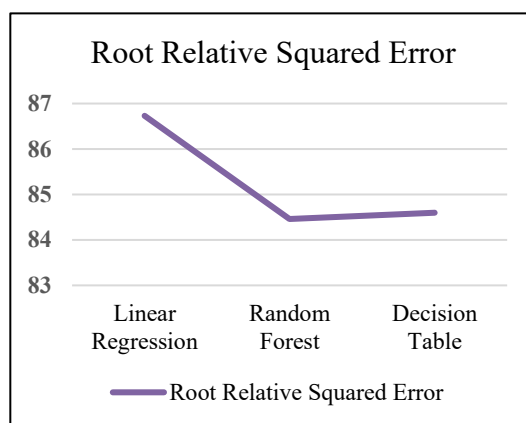


Fig.20.Root Relative Squared Error

Algorithms	Accuracy
Statistical-Based Linear Regression	79.1 %
Decision Tree-Based Random Forest	78.9 %
Rule generation-Based Decision Table	81.1 %

Table.5..Accuracy Level of Algorithms

The algorithms that are chosen to analyze the multivariate and categorical diabetes dataset is the statistical-based Linear Regression algorithm, decision tree-based Random Forest algorithm and rule generation-based Decision Table algorithm, the Rule Generation based Decision Table method performs well the given data set with a higher accuracy range of 81.1% and minimum error rate than the other two chosen algorithms.

V. CONCLUSION

In the study of classification algorithms for the medical dataset, the Statistical, Decision Tree and Rule Generation algorithm based Linear Regression, Random Forest and Decision Table methodologies are tested for the multivariate and categorical diabetes dataset. Among the specified classification algorithms with 10 fold cross-validation, the rule Generation based Decision Table method stands out with an accuracy of 81.1% which is higher than the compared other mechanisms. In future, the chosen rule generation can be further improved by using ensemble methods and by updating the rule table by adding new feature selection procedures. Due to the increase of medical factors in society, a new well-built model is needed for efficient disease classification in machine learning.

REFERENCES

- Shao, G., Tang, L., Zhang, H., 2021. Introducing Image Classification Efficacies. IEEE Access 9, 134809–134816. <https://doi.org/10.1109/ACCESS.2021.3116526>
- Hussein, S., Kandel, P., Bolan, C.W., Wallace, M.B., Bagci, U., 2019. Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches. IEEE Trans. Med. Imaging 38, 1777–1787. <https://doi.org/10.1109/TMI.2019.2894349>
- Hsu, C.-H., 2021. Optimal Decision Tree for Cycle Time Prediction and Allowance Determination. IEEE Access 9, 41334–41343. <https://doi.org/10.1109/ACCESS.2021.3065391>
- Rowley, H.A., n.d. Neural Network-Based Face Detection 149.
- Wang, K., Zhu, Y., Chen, C.-Z., 2020. Perceptron Algorithm and Its Verilog Design, in: 2020 China Semiconductor Technology International Conference (CSTIC). Presented at the 2020 China Semiconductor Technology International Conference (CSTIC), IEEE, Shanghai, China, pp. 1–3. <https://doi.org/10.1109/CSTIC49141.2020.9282536>
- Alla, H., Moumoun, L., Balouki, Y., 2021. A Multilayer Perceptron Neural Network with Selective-Data Training for Flight Arrival Delay Prediction. Scientific Programming 2021, 1–12. <https://doi.org/10.1155/2021/5558918>
- Dunham MH, Data mining: Introductory and advanced topics, Pearson Education, 4-5.2003.