© 2019 IJSRST | Volume 6 | Issue 2 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology
DOI : https://doi.org/10.32628/IJSRST196294

Improving Accuracy for Diabetes Mellitus Prediction Using Data Pre-Processing and Various New Learning Models

Garvit Khurana, Prof. Arun Kumar
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India

ABSTRACT

Data mining in medical data has successfully converted raw material into useful information. This information helps the medical experts in improving the diagnosis and treatment of diseases. Type II Diabetes Mellitus is one of silent killer diseases worldwide. According to World Health Organization, 346 million people are suffering from diabetes worldwide. Diagnosis or prediction of Diabetes is done through various data mining technique such as association, classification, clustering and pattern recognition. The study led to related open issues of identifying the need of a relation between the major factors that lead to the development of diabetes. This is possible by mining patterns found between the independent and dependant variable in the dataset. This paper compares classification accuracies of various machine learning models. Objective of paper is to find whether a person has diabetes or not and what features are highly responsible for diabetes. As due to its continuously increasing occurrences more and more families are influenced by diabetes mellitus. Most diabetic people know little about their health. In this study, we have proposed novel model on data mining techniques for predicting type 2 diabetes mellitus. Diabetes often referred to by doctors as metabolic disease in which the person has high blood glucose (blood sugar), because of inadequate insulin production.

Keywords : Machine learning, Diabetes, Sugar, Data Analysis Diabetes, Support vector machines, Prediction algorithms, Classification algorithms

I. INTRODUCTION

In this paper we describe the recent work at diabetes mellitus prediction with the help of machine learning alogrithms and Machine learning is a powerful artificial intelligence tool that make sense of a complicated world. And its transforming a wide variety of industries. It's becoming gradually ubiquitous with more and more and more application that we can't even think of them.

Most people probably already know that email provider uses a machine learning algorithm to identify spam. From past few years Google, Tesla and other are building self-driving system that will soon augment or replace the human drivers. And E-commerce giant like Amazon and technology companies like Braintree are using it in conjunction with other tools to stop credit card fraud.

Mining is one of the most significant applications of Machine learning. More often during analyses or, possibly, when trying to establish relationships between multiple features people become prone to making mistakes.

This makes them tiresome for them to find a solution to certain problems. Here comes Machine learning which can be utilized to apply successfully to these problems, improving the Data efficiency of the

system and designs of machines. The same set of features represents every instance of any dataset that is used by machine learning algorithms. The features can be Continuous, Discreet or Binary. Whenthe instances of the dataset are provided with known labels (corresponding correct output) that learning is known as Supervised Machine learning. While in case of Unsupervised learning instances are without known labels. By implementing these unsupervised (clustering) learning algorithm researchers anticipate to discover unknown, but useful classes of items. In supervised learning, the gathered data after preprocessing feed to the algorithm which analyze the data and build a model which then predicts the result on new data, example problems are Classification and Regression.

In contrast with Unsupervised learning the gathered data is un-label therefore algorithm analyze the data and create the model by deducing the structure present in the input data, example problems are clustering, reduction in dimensionality and learning rule for association. Reinforcement machine learning is another different kind of machine learning technique where learning is taking place by interacting with the environment. Here learner is not provided what actions to take, as in most forms of machine learning, but instead it discovers by it-self which actions produces the most reward by trying them.

A reinforcement learning agent learns from the outcomes of its actions, rather than from being explicitly taught and selection of its actions depends upon its past experiences (exploitation) and also by new choices (exploration), which is mandatorily trial and error learning. In Robotics Reinforcement learning is common, where the collection of sensors readings at one point in time is a data point, and the algorithm decide the robot's next action. In the Internet of Things (IoT) applications also a natural fit into it. Here in this paper, we are predicting if a person will develop diabetes and analyze the model created by using simple but powerful algorithms like Naïve Bayes, Logistic Regression and Decision Tree and Random Forest. The data we have chosen is from a Pima Indian Diabetes study.

1.1 OBJECTIVES OF PAPER

Due to increasing cost of health care, it is useful to assist patients to control diabetes by themselves. In many instances, early information related to diabetes may help in avoidance, curing and appropriate treatment of the disease.

The need for avoidance and better management of type II diabetes has been an important issue since ages. Medical practitioners and researchers have investigated and continue to find solutions to overcome this disease. Various researches and studies are done on predicting the blood glucose levels for Type II diabetes patients for a short term. Most of the predictions helped to decide the diet control and physical activity in order to maintain a good life.

The main focus of our paper is to investigate the possible solutions for the group of people who are at a risk of developing type II diabetes in future.

Improving accuracy will be one main job of our paper

LIMITATIONS OF EXISTING WORK

a) Unfortunately OAD(oral antidiabetic drugs) approach alone fails to achieve adequate glycemic control within 1 year for the majority of patients. It is extremely difficult for patients to modify lifelong habits, and most will ultimately require

pharmacotherapy to restore normoglycemia.

b) Most of the models have been developed to diagnose diabetes and predict the blood sugar level for a short term. However, according to the authors knowledge, there are hardly any system developed to predict the onset of diabetes in the long run.

c) Compulsory to do preprocessing : Many researches are done on the prediction and diagnosis of diabetes.

After the literature survey, we have found that most of the data mining techniques applied on the PIDD were pre-processed.

The PIDD has eight attributes out of that a couple of attributes contain values that square measure out of the traditional vary.
Also there are many missing values i.e. the value is 0 instead of an actual value.
Therefore, the pre-processing of information is critical for economical data processing of patterns within the PIDD.

d) Accuracy of result obtained: The proposed cascaded model with categorical data obtained the classification accuracy of 93.33 % when compared to accuracy of 73.62 % using C4.5 .Hence the accuracy is not greater than 95% which is also one of the drawback of proposed works.

1.2 METHOD AND APPROACHES

Due to rising price of health care, it's helpful to help patients to manage polygenic disorder by themselves.

In several instances, early info associated with polygenic disorder would possibly facilitate in rejection, curing and appropriate treatment of the disease.

Many computer programs or systems were developed and are being developed by emulating human intelligence that could be used to assist the users or patients in managing diabetes.

We assessed completely different systems like AI systems, mobile applications and specially designed devices for the prediction and designation of polygenic disorder.

The focus of this paper is to investigate for a model to predict and diagnose diabetes in the long run.

Most of the models are developed to diagnose polygenic disorder and predict the glucose level for a brief term.

S. No	Attributes	Type
1	Number of Times pregnant	Continuous
2	Plasma glucose concentration 2 hours in an oral glucose tolerance test	Continuous
3	Diastolic blood pressure (mm Hg)	Continuous
4	Triceps skin fold thickness (mm)	Continuous
5	2-Hour serum insulin (mu U/ml)	Continuous
6	Body mass index (weight in kg/(height in m)^2)	Continuous
7	Diabetes pedigree function	Continuous
8	Age (years)	Numeric
9	Class variable (0 or 1)	Discreet

However, as some researchers indicate, there are a number of impossible values, such as 0 body mass index and 0 plasma glucose.

Furthermore, one attribute (2-hour se- rum insulin) contains almost 50% impossible values.
To keep the sample size fairly massive, this attribute is removed from analysis.

There ar 236 observations that have a minimum of one not possible worth of aldohexose, blood pressure, triceps skin thickness, and body mass index.
There are nine variables, including the binary response variable, in this dataset; all other attributes are numeric-valued.

1.3. DATA ANALYSIS METHOD

1.3.1. DATAVISUALIZATION

Data visualization is a very general term that describes any effort to assist individuals perceive the importance ofinformation by putting it in a visual context. Patterns, trends and correlations that may go unobserved in text-based information are often exposed and recognized easier with the help of Data Visualization Techniques. It is also very useful for finding outliers and anamolies in data. The attributes are given below as a correlation matrix:
Correlation Matrix

II. METHODS AND MATERIAL

A. TOOLS USED

2.1 PYTHON

Python is AN taken, object-oriented, high-level programming language with dynamic semantics. Its high-level inbuilt knowledge structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's easy, simple to find out syntax emphasizes readability and so reduces the value of program maintenance.

Python supports modules and packages, which inspires program modularity and code utilise.
The Python interpreter and also the intensive customary library square measure on the market in supplyor binary kind for complimentary for all major platforms, and might be freely distributed.

Often, programmers fall gaga with Python as a result of the augmented productivity it provides. Since there's no compilation step, the edit- test-debug cycle is implausibly quick. Debugging Python programs is easy: a bug or unhealthy input can ne'er cause a segmentation fault.

Instead, once the interpreter discovers a slip-up, it raises AN exception.
When the program does not catch the exception, the interpreter prints a stack trace.
A supply level program permits review of native and world variables, analysis of discretional expressions, setting breakpoints, stepping through the code a line at a time, and so on.
The program is written in Python itself, testifying to Python's introspective power.
On the opposite hand, often the quickest way to debug a program is to add a few print statements

to the source: the fast edit-test-debug cycle makes this simple approach very effective.

2.2 SCIKIT LEARN

What is scikit-learn?
Scikit-learn is maybe the foremost helpful library for machine learning in Python.
It is on NumPy, SciPy and matplotlib, this library contains a lot of effiecient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

Please note that scikit-learn is employed to create models.
It shouldn't be used for reading the information, manipulating and summarizing it.
There are better libraries for that (e.g. NumPy, Pandas etc.)

2.2.1 COMPONENTS OF SCIKIT-LEARN:

Scikit-learn comes loaded with a lot of features. Here are a few of them to help you understand the spread:

- Supervised learning algorithms: Think of any supervised learning algorithm you might have heard about and there is a very high chance that it is part of scikit-learn.
- Starting from Generalized linear models (e.g Linear Regression), Support Vector Machines (SVM), Decision Trees to Bayesian methods – all of them are part of scikit-learn toolbox.
- The unfold of algorithms is one in every of the massive reasons for prime usage of scikit-learn.
- I started exploitation scikit to unravel supervised learning issues and would suggest that to individuals newscikit / machine learning furthermore.

- Cross-validation: There are various methods to check the accuracy of supervised models on unseen data.
- Unsupervised learning algorithms: Again there is a large spread of algorithms in the offering – starting from clustering, factor analysis, principal component analysis to unsupervised neural networks.
- Various toy datasets: This came in handy while learning scikit- learn.
- Feature extraction: Useful for extracting features from images and text

2.3 MACHINE LEARNING

Machine learning (ML) may be a class of algorithmic program that permits package applications to become a lot of correct in predicting outcomes while not being expressly programmed.

The basic premise of machine learning is to create algorithms which will receive {input knowledge|input file|computer file} ANd use applied mathematics analysis to predict an output whereas change outputs as new data becomes available.

2.3.1 TYPES OF MACHINE LEARNING ALGORITHMS

Just as there square measure nearly limitless uses of machine learning, there is no shortage of machine learning algorithms.

They range from the fairly easy to the extremely complicated.
Here are a few of the most commonly used models:

This class of machine learning algorithm involves identifying a correlation -- generally between two variables -- and using that correlation to make predictions about future data points.

Decision trees.
These models use observations concerning bound actions ANd establish an optimum path for incoming at a desired outcome.

K-means clustering.
This model teams a mere range of information points into a selected range of groupings supported like characteristics.

This information set is extracted from a bigger info originally closely-held by the National Institute of polygenic disease and organic process and excretory organ Diseases.

The purpose of the study is to analyze the connection between the polygenic disease diagnostic result and an inventory of variables that represent physiological measurements and medical attributes.

The data set within the UCI repository contains 768 observations and nine variables with no missing values reported.

S. No	Attributes	Type
1	Number of Times pregnant	Continuous
2	Plasma glucose concentration 2 hours in an oral glucose tolerance test	Continuous
3	Diastolic blood pressure (mm Hg)	Continuous
4	Triceps skin fold thickness (mm)	Continuous
5	2-Hour serum insulin (mu U/ml)	Continuous
6	Body mass index (weight in kg/(height in m)^2)	Continuous
7	Diabetes pedigree function	Continuous
8	Age (years)	Numeric
9	Class variable (0 or 1)	Discreet

However, as some researchers indicate, there are a number of impossible values, such as 0 body mass index and 0 plasma glucose.

Furthermore, one attribute (2-hour se- rum insulin) contains almost 50% impossible values.
To keep the sample size fairly massive, this attribute is removed from analysis.

There ar 236 observations that have a minimum of one not possible worth of aldohexose, blood pressure, triceps skin thickness, and body mass index.

There are nine variables, including the binary response variable, in this dataset; all other attributes are numeric-valued.

III. DATA ANALYSIS METHOD

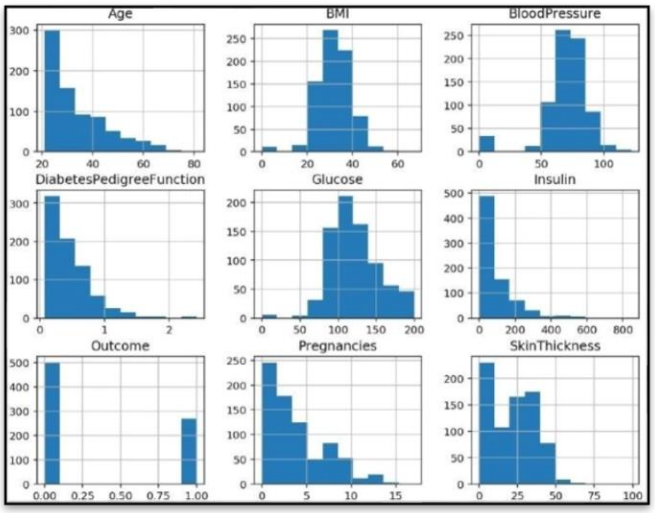
3.1. DATAVISUALIZATION

Data visualization is a very general term that describes any effort to assist individuals perceive the importance ofinformation by putting it in a visual context. Patterns, trends and correlations that may go unobserved in text-based information are often exposed and recognized easier with the help of Data Visualization

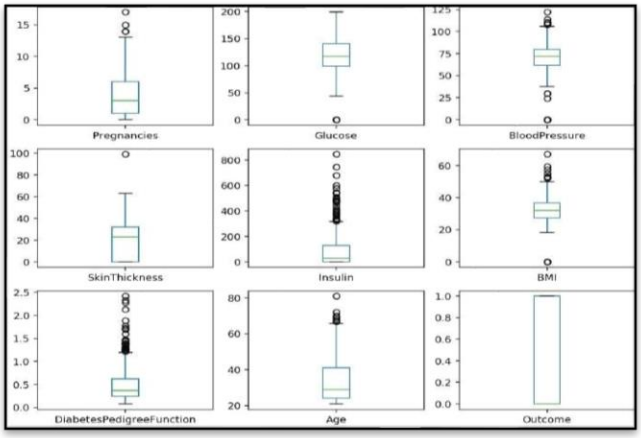
Techniques. It is also very useful for finding outliers and anamolies in data. The attributes are given below as a correlation matrix:

Correlation Matrix

3.1.1. HISTOGRAM



3.1.2.BOX PLOT



3.2. DATA PRE-PROCESSING

To improve the quality of the results obtained after mining and the effectiveness of the complete mining process, data pre-processing is done. Researchers and practitioners realize that in ord er to use data mining tools on the database effectively, data pre-processing is essential for successful data mining. After observing the Pima Indians.

Diabetes dataset, we found the need to pre-process the data in two steps.

Firstly, it's seen that the informationset has the worth zero for missing data.

We removed all the instances that had the worth zero for a selected field wherever having a zero as a worthwas not possible.Therefore, the instances that have missing values were eliminated.

By removing the irrelevant and redundant features we make feature subset and this process termed as feature subset selection. This helps in reducing the dimensionality of the data and help to operate effectively and faster.

In this paper, we have imputed missing data with mean and deleted the externous correlated features which contribute to the better comprehensibility of the produced classifier and the better understanding of the learned concept.

The Prepared data after cleaning, used for training and testing the model.

The data is split seventieth for coaching and half-hour for testing.

Then we train the algorithm on 70% of the data set and keeping the test data aside. This training process will produce the training model based on the logic and the algorithm and the values of the features in the training data. Then test the model on the unseen data to evaluate the model. If we trained the model on the entire set of data, then it produces a good result on the test data as it has seen the biases and when we use this model to the real world data, then it will perform poorly as it is unaware of the biases present in the real world data.

Therefore, we keep the testing data separated from training data so that it produces better results.

3.3. PREDICTIONS AND ANALYSIS BASED ON COLLECTED DATA:

We are using the machine learning workflow to process and transform Pima Indian diabetes data to create prediction model. This model must predict which people are likely to develop diabetes, using Logistic Regression, Decision Tree, AdaBoostClassifier, Support Vector Machine and K Nearest Neighbors.

3.3.1. NAÏVE BAYES

The Naïve Bayes algorithm is a simple probabilistic classifier that calculates a collection of probabilities by investigating frequency and combination of values in a given data set. The algorithm is based on applying Bayes theorem with the "naïve" assumption of independence between every pair of features. Due to simple structure of Naive Bayes, construction of it is very simple and also has several advantages. Moreover, the inference (classification) is achieved in a linear time (while the inference in Bayes networks with a general structure is known to be NP-complete). Also, it does not require much larger data set smaller data set can also be used.

Finally, the construction of naive Bayes is incremental, in the sense that it can be easily updated (namely, it is always easy to consider and take into account new cases in hand). Suppose Ci be diabetes risk group i and N be input variables that are used in a model and under the assumption of all variables are independent.

To predict a class of diabetes risk, a model of Naive Bayes can be defined by

$$P(C_i|\,N) \quad = \quad \frac{P(N\,|C_i) \;\times\; P(C_i)}{P(N)}$$

Where is a posterior probability of a training data set with variable N that will be.

3.3.2. LOGISTIC REGRESSION

Despite its name, Logistic regression is basically a linear model for classification rather than regression. It is also known as the logit regression, maximum-entropy classification (MaxEnt) or the log- linear classifier. In this model, we use logistic regression to model probabilistically described outcomes of a single trial. It is a basic model which describes dichotomous output variables and can be extended for disease classification prediction. Suppose there are N input variables where their values are indicated by m1, m2, m3,…,mN.

Let us assume that the P probability of that an event will occur and 1- P be a probability that event will not occur. Logistic regression model is given by

$$\log\left(\frac{P}{1-p}\right) = \text{log it}\,(P) = \beta_0 + \beta_1 m_1 + \cdots + \beta_N m_N$$

Where β0 is that the intercept and β1, β2, … .
, βN are regression coefficients.

3.3.3 DECISION TREE

It creates a binary tree. The decision tree approach is most useful in the classification problem.

With this method, a tree is constructed to model the classification process.

It consists of three types of nodes root node, child node, and leaf node. The algorithm starts with defining a root node from the most relationship between every input and output variables. Next, the child node is selected by calculating Information Gain (IG). IG(parent, child) = Entropy(parent) – [P(x1) × Entropyx1 + Px1×Entropyx1+…3 Entropy(Ci) = –P(xi) log P(xi) and P(xi) is the probability of child node i. Node having the highest IG will become the parent for next generation. This process is repeated until it gets a leaf node and completed decision tree. The stopping criteria for decision tree is that all the sample for a given node belong to the same class, there aren‟t remaining attributes for any further partitioning and there aren‟t any leftover sample. It requires little data preparation. While different techniques typically require data normalization, creation of dummy variables and blank values to be removed.
Note but that this module doesn't support missing values.

Decision trees tend to over-fit on data having a vast number of features.

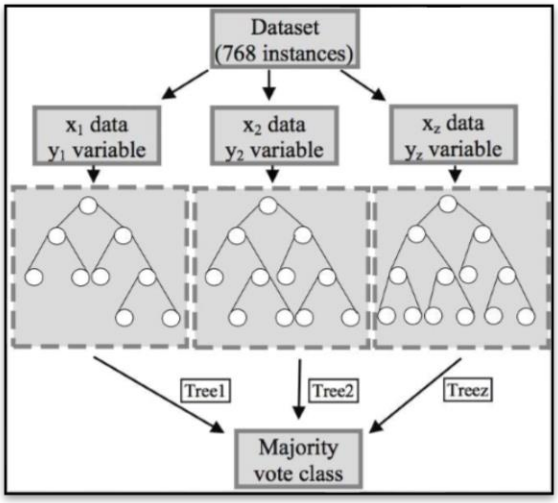
Obtaining the right ratio of samples to a number of features is important since a tree with few samples in high dimensional space is very likely to over-fit.

3.3.4. RANDOM FOREST

Random Forest is an ensemble algorithm which was modeled from trees algorithm and Bagging algorithm. It is developed by Breiman, he found that the algorithm can potentially improve classification accuracy.

It also works well with a data set with a vast number of input variables. The algorithm begins by creating a combination of trees which each will vote for a class. The figure below presents how to model the Random Forest.

Suppose that there are X data and Y input variables in a data set where the real data used in this paper compose of 768 data and 9 input variables. Let z be the number of sampling groups, xi and yi be a number of data and variables in group i where i is equal to 1, 2, ... and z. Each sampling group is as followed:

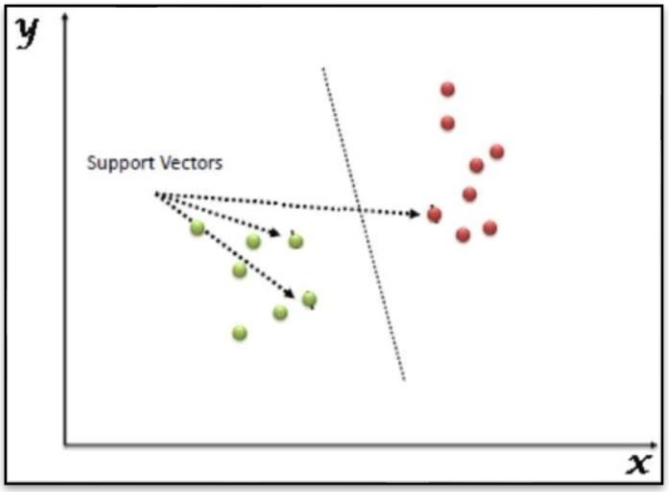


xi variables where xi is not greater than X are selected randomly from X.
yi variables where yi is not greater than Y are selected randomly from Y.

A tree is grown and gives a prediction class. After Step one to three was recurrent for z times, these trees become a forest. Then the classification will be elected by a majority vote of all trees within the forest. Note that all data have to be returned to the data set before selecting a new sampling group.

3.3.5. SUPPORT VECTOR MACHINE

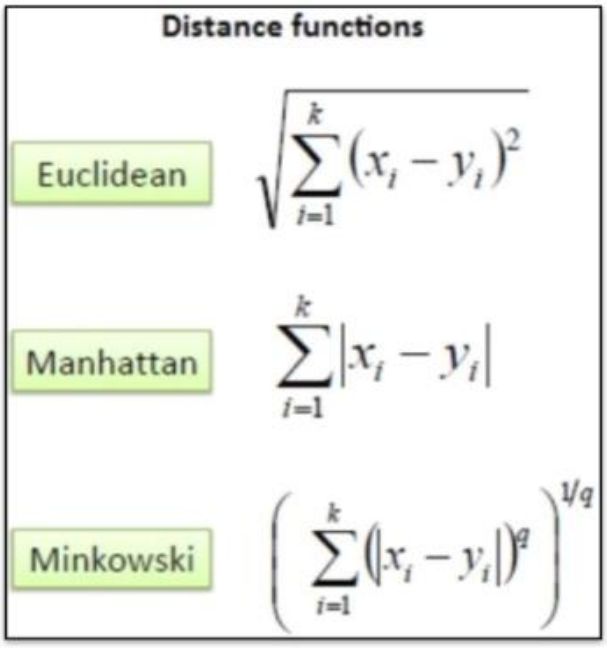
SVM could be a supervised machine learning algorithmic rule which may be used for each classification or regression challenges. However, it is mostly used in classification problems. In this algorithmic rule, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we tend to perform classification by finding the hyper-plane that differentiate the two categories.



3.3.6. K NEAREST NEIGHBORS – CLASSIFICATION

K nearest neighbors could be a straightforward algorithmic rule that stores all offered cases and classifies new cases supported a similarity live (e.g., distance functions).

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function.

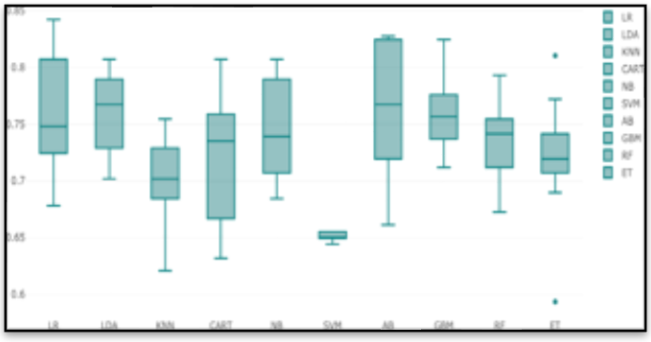


IV. IMPLEMENTATION AND PROPOSED SYSTEM RESULTS

4.1. CLASSIFICATION

ACCURACY ON NON-PROCESSED DATA

1. LR: 0.758904 (0.049893)
2. LDA: 0.760510 (0.034873)
3. KNN: 0.698048 (0.041546)
4. CART: 0.723790 (0.052410)
5. NB: 0.744962 (0.041601)
6. SVM: 0.651037 (0.003678)
7. AB: 0.760687 (0.063215)
8. GBM: 0.757121 (0.031491)
9. RF: 0.737944 (0.035908)
10. ET: 0.720816 (0.053628)

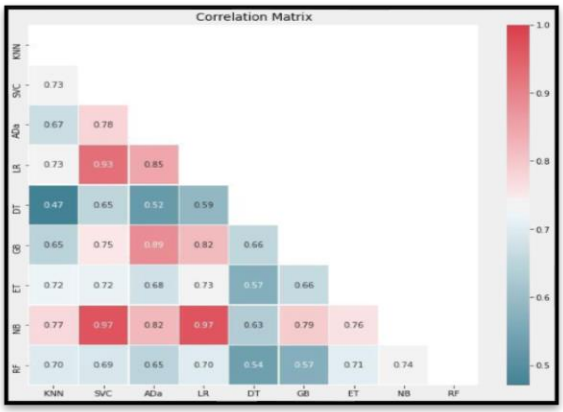


4.2. CLASSIFICATION ACCURACY ON PROCESSED DATA

We have used constant dataset for the comparison. The following table shows the accuracy results on non-processed data for the various techniques used :

TECHNIQUE	ACCURACY
LogisticRegression	81.81%
KNeighborsClassifier	79.22%
SVC	81.81%
DecisionTree	84.41%
AdaBoostClassifier	81.81%
GradientBoosting	76.62%
VotingClassifier	83.11%

4.3. CORRELATIONS MATRIX FOR ERRORS:



V. RESULTS AND DISCUSSION

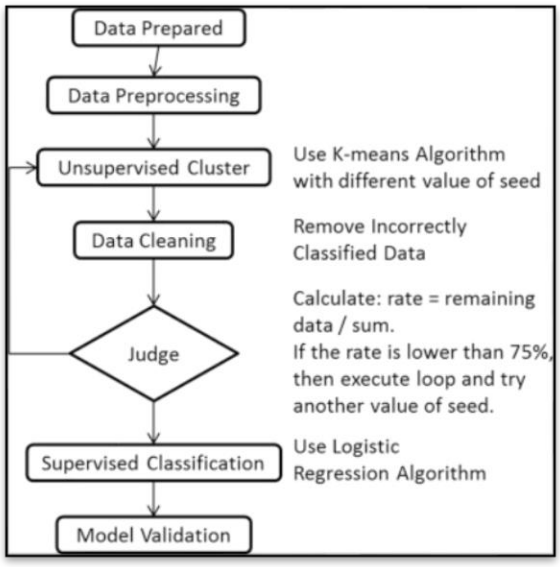
This paper focused on the importance of data pre-processing for data mining. We used the PIMA Indian polygenic disorder Dataset for the study. The data was initial classified while not pre-processing it and therefore the results were noted. Then constant set knowledge of knowledge of information} was pre-processed that's the removal of missing values and data discretization. Classification was done once the 2 step method of information pre- process. After the comparison between the accuracies of classification on non-processed and pre-processed data, it showed that the classification accuracy increases when the data is pre-processed. Hence the info mining accuracy depends tons on the pre-processing of information.

COMPARISON WITH EXISTING STUDY RESULTS

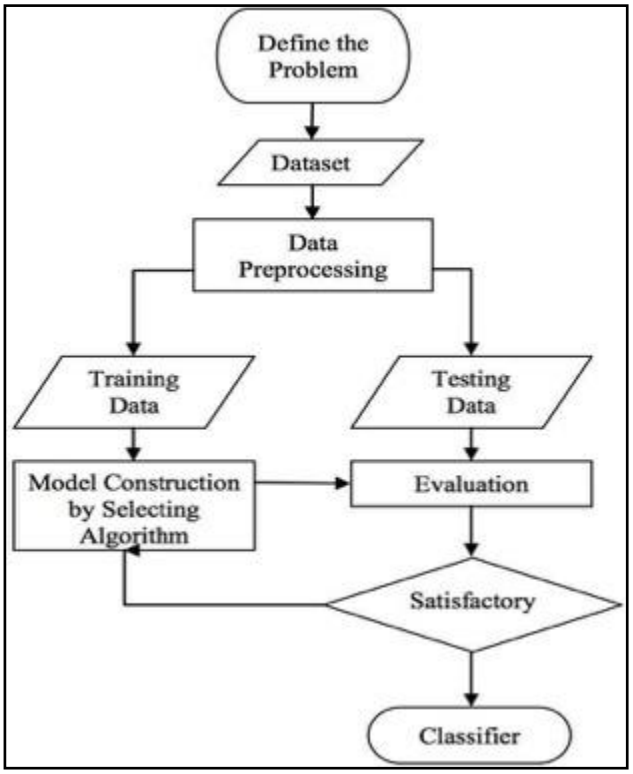
In recent years, using the data mining technique has been used with increasing frequency to predict the possibility of disease. Many algorithms and toolkits have been created and studied by researchers. These have highlighted the tremendous potential of this research field. In this section, a few important works that are closely related to the proposed issue are presented.Based on several studies, we found that a commonly used dataset was the Pima Indians Diabetes Dataset from the University of California, Irvine (UCI) Machine Learning Database. Patil proposed a hybrid prediction model (HPM), which used a K-means clustering algorithm aimed at validating a chosen class label of given data and used the C4.5 algorithm aimed at building the final classifier model, with 92.38% classification accuracy. Ahmad compared the prediction accuracy of multilayer perception (MLP) in neural networks against thde ID3 and J48 algorithms. The results showed that a pruned J48 tree performed with higher accuracy, which was 89.3% compared to 81.9%.

Marcano-Cede˜no proposed artificial metaplasticity on multilayer perceptron (AMMLP) as a prediction model for diabetes, for which the best result obtained was 89.93%. All the studies presented above used the same Pima Indians Diabetes Dataset as the experimental material. The Waikato Environment for Knowledge Analysis (WEKA) toolkit was the primary tool which most researchers chose.In order to obtain more useful and meaningful data, we realized that the preprocessing methods and parameters should be chosen rationally. Vijayan V. reviewed the benefits of different preprocessing techniques for predicting DM. The preprocessing methods were principal component analysis (PCA) and discretization. It concluded that the preprocessing methods improved the accuracy of the naive Bayes classifier and decision tree (DT), while the support vector machine (SVM) accuracy decreased. Wei analyzed risk factors of T2DM based on the FP-growth and Apriori algorithms. Guo proposed the receiver operating characteristic (ROC) area, the sensitivity, and the specificity predictive values to validate and verify the experimental results.On the basis of an effective prediction algorithm, we need an appropriate way to make the model convenient for everyone.

EARLIER METHOD



NEW METHOD



VI. CONCLUSION

In this paper, various investigations on prediction and diagnosis of type II diabetes mellitus using data mining techniques are present. Various classification techniques area unit used once pre-processing of the info. In this paper we have done a comparison of the accuracy of classification done on non- processed and pre-processed data. We have come back to a conclusion that the pre- processed information offers us a higher accuracy results instead of non-processed information.

This shows the importance of pre- processing in the data mining.

TECHNIQUE	ACCURACY
LogisticRegression	81.81%
KNeighborsClassifier	79.22%
SVC	81.81%
DecisionTree	84.41%
AdaBoostClassifier	81.81%
GradientBoosting	76.62%
VotingClassifier	83.11%

USING EARLIER TECHNIQUES

TECHNIQUE	ACCURACY
LogisticRegression	78.2%
Naïvebay	74.9%
KNN	67.6%

FUTURE WORK AND LIMITATIONS

For future work, it is necessary to bring in hospital's real and latest patients' data for continuous training and optimization of our proposed model. The quantity of the dataset should be large enough for training and predicting. Some advanced algorithms and models should be applied in the study of DIABETES MELLITUS. Grading forecasting standards are also necessary for potential diabetes patients. Developing a series of rules and standards is a valid method to prevent people from developing DIABETES MELLITUS. Based on that, a more effective model for predicting DIABETES MELLITUS and grading potential patients is presented. This will help to lower the growth rate of diabetes and eventually decrease the risk of developing DIABETES MELLITUS.

The Pima Indian Diabetes dataset we used in the research have only the numeric parameters. Sometimes medicine dataset may contain different data formats like text, images (x-ray, EGC report) and dates and time. Artificial neural network accepts only numeric data formats. The other machine learning algorithms can be used for handling the text and image data types.

VII. REFERENCES

[1]. Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui ," Application of data mining:Diabetes health care in young and old patients" , 2012

[2]. Asha Gowda Karegowda and M.A. Jayaram, Cascading GA & CFS for Feature Subset Selection in Medical Data Mining , International Conference on IEEE International Advance Computing Conference (IACC‟09), Thapar University, Patiala, Punjab India (Mar 2009).

[3]. Margaret H. Danham,S. Sridhar, “Data mining, Introductory and Advanced Topics”, Person education , 1st ed., pp. 75-84,2006.

[4]. Aman Kumar Sharma, SuruchiSahni, “A Comparative Study of Classification Algorithms for Spam Email Data Analysis”, IJCSE, Vol. 3, No. 5, pp. 1890-1895,2011.

[5]. Barto, A. G. & Sutton, R., “Introduction to Reinforcement Learning”, MIT Press.M. Young, The Technical Writer‟s Handbook Mill Valley, CA: University Science, pp. 45-60,1997.

[6]. S. B. Kotsiantis, I. D. Zaharakis, P. E. Pintelas, “Machine learning: a review of classification and combining techniques”, Springer Science+Business Media B.V., ArtifIntell Rev, Vol. 26, pp. 159–190,2007.

[7]. Leslie Pack Kaelbling, Michael L. Littman,“Reinforcement Learning:A Survey”,

Journal of Atificial Intelligence Research, Vol. 4, pp. 237-285,1996.

[8]. B.M Patil, R.C Joshi, Durga Tosniwal, Hybrid Prediction model for Type-2 Diabetic Patients, Expert System with Applications, 37, 8102-8108 (2010).

[9]. Asha Gowda Karegowda, MA.Jayaram, Integrating Decision Tree and ANN for Categorization of Diabetics Data,International Conference on Computer Aided Engineering, December 13-15, IIT Madras, Chennai, India (2007).

[10]. Humar, K., & Novruz, A. Design of a hybrid system for the diabetes and heart diseases. Expert Systems with Application.

Cite this article as :

Garvit Khurana, Prof. Arun Kumar, "Improving Accuracy for Diabetes Mellitus Prediction Using Data Pre-Processing and Various New Learning Models", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 6 Issue 2, pp. 502-515, March-April 2019. Available at doi : https://doi.org/10.32628/IJSRST196294
Journal URL : http://ijsrst.com/IJSRST196294