PREDICTION AND DIAGNOSIS OF DIABETES MELLITUS USING DATA MINING ALGORITHMS

Divya Agrahari, M.Tech Scholar, Department of Computer Science & Engineering, Buddha Institute of Technology GIDA, Gorakhpur, India.
Dr. Anshu Kumar Dwivedi, Associate Professor, Department of Computer Science & Engineering, Buddha Institute of Technology GIDA, Gorakhpur, India.

Abstract— Diabetes, also referred to as diabetes mellitus (DM), is a potentially fatal disorder that affects people from all different parts of the world. Diabetes can be caused by a number of different risk factors, including but not limited to obesity, high blood glucose levels, a lack of physical activity, and other risk factors. There is a possibility that it can be controlled or mitigated if it is identified at a relatively early stage. An example of machine learning is the creation of a computer system or programme that is capable of modifying itself and learning from prior experiences. This is an example of the field of artificial intelligence. The PIMA dataset is utilized at several points throughout the course of this inquiry. The collection has around 9 distinguishing qualities for each of the 768 cases. Each algorithmic strategy for machine learning can be implemented in a great number of different ways. On the other hand, in order to meet the requirements of these research efforts, we opted to use three unsupervised learning strategies. These algorithms are known by their respective names, such as logistic regression, decision tree, and random forest. Every single one of these algorithms was trained and put through its paces before being used in this model. In the conclusion, we will evaluate the effectiveness of various metric algorithmic methods to machine learning by comparing and contrasting their respective performance levels. There are a number of performance indicators that are analyzed, including accuracy, F-measure, recall, and precision. The Logistic Regression model has the highest overall score, the highest value of 0.68 for their f-measure, and the best accuracy score, which is 74%. Additionally, it has the highest precision value, which is 0.73, and it also has the highest value for their f-measure. Decision Tree came out on top with a recall score of 0.61, which was the highest of any method**.**

Index Terms— Data mining, Diabetes Mellitus Disease, EM algorithm, Random Forest with Feature Selection, Machine Learning Algorithm, etc.

I. INTRODUCTION

Diabetes Mellitus (DM) is a chronic illness that requires continual medical care and education on self-management in order to lower the risk of adverse long-term outcomes and prevent complications from developing. By bringing the patient's blood sugar levels under control and treating diabetes with a combination of diet and medication, one can reduce or eliminate a wide range of diabetes-related symptoms and complications. There are primarily two distinct forms of diabetes that can be recognized from one another: Juvenile diabetes is a kind of diabetes type 1, which is also known as adult-onset diabetes. Insulin dependency is a subtype of diabetes that develops when the body stops producing the hormone known as insulin. Insulin is required for the body to make use of the glucose that is found in food; hence its absence leads to diabetes. This is very common in people of younger ages, especially youngsters. [Cause and effect] Five to ten percent of diabetics can be attributed to this factor. Insulin injections are typically necessary for survival in diabetics who are diagnosed with this form of the disease. Type 2 diabetes, also known as adult-onset diabetes or diabetes that is not dependent on insulin, affects the vast majority of people who have diabetes. Diabetes mellitus type 1 is characterized by the body's inability to produce sufficient amounts of insulin in the appropriate manner. People who are overweight, have a family history of diabetes, and are over the age of 40 are at increased risk for acquiring type 2 diabetes. This is due to the fact that diabetes is becoming more prevalent in adults as a direct result of poor eating habits [1].

Diabetes can develop for a number of different reasons, including but not limited to: high blood pressure, being overweight, kidney failure, high cholesterol levels, blindness, and a lack of physical activity (American Diabetes Association, 2004). It would appear that both heredity and environmental variables, such as being overweight, being of a certain race or gender, reaching a certain age, and not getting enough exercise, all play important roles in the onset of diabetes. Researchers in the fields of artificial intelligence and biomedical engineering who are working in the field of diabetes research have become more interested in the topic as a result of the rise in the number of diabetic patients around the world (Ashwinkumar & Anandakumar 2012).

Diabetes comes in at number seven on the list of disorders that can result in death, as determined by the conclusions of a study that was conducted without bias. In India alone, there are 51 million people who have been identified as having diabetes, and the number of people who have type 2 diabetes much outnumbers the number of people who have type 1 diabetes. In November 2007, around 7% of the population in the United States was affected by diabetes, which affected a total of 20.8 million children and adults. The findings of a global survey that was conducted in 2013 by Boehringer Ingelheim and Eli Lilly and company revealed that 25.8 million people in the United States and 382 million people around the world are afflicted with either Type-1 diabetes or Type-2 diabetes. Both industrialized and developing countries have a significant problem with the prevalence of type-2 diabetes, which is the most common form of the disease and is believed to account for 90–95% of all diabetes cases.

The International Diabetes Federation (IDF) has developed some forecasts that indicate the number of people living with diabetes in the world would reach 592 million by the year 2035. These projections were made in the year 2005. The World Diabetes Atlas estimates that there are currently 285 million people living with diabetes around the globe, and that number has the potential to climb to 438 million by the year 2030. A survey's findings indicated that the number of people suffering from type 2 diabetes would skyrocket by the year 2030, setting off worrying predictions for the future. According to Kenney and Munce (2003). In addition to this, it is a given that by the year 2030, developing countries would be home to 85 percent of the world's diabetic patients. This prediction is based on the fact that the prevalence of diabetes is expected to rise. It is anticipated that the number of individuals residing in India who are affected by diabetes would increase from 31.7 million in the year 2000 to 79.4 million in the year 2030. (Huy Nguyen et al 2004). One of the most critical aspects of successfully treating diabetes is getting a correct diagnosis as quickly as possible (Mythili et al 2003).

Over 62 million individuals in the Republic of India are currently living with diabetes, which means the disease is rapidly approaching the status of a potential epidemic. According to Wild et alresearch, .'s the number of people living with diabetes is expected to more than double from 171 million in the year 2000 to 366 million in the year 2030, with the greatest growth expected in India. It is anticipated that by the year 2020, India would have a diabetic population of up to 79.4 million people, while China will have 42.3 million people and the United States will have 30.3 million people who will also see significant rises in the number of diabetics in their populations. Diabetes has the potential to be a significant burden for India in the future, and the country is already facing an uncertain future because of this possibility. [2].

Diabetes is a group of disorders in which the body either does not generate enough insulin or does not use the insulin that is produced in the correct manner, or a combination of both of these factors. If this were to occur, the body would be unable to get sugar from the blood into the cells, which would lead to elevated levels of blood glucose. The form of sugar that is present in our blood is called glucose, and it is one of the primary sources of energy. The accumulation of sugar in the blood is a symptom of insulin resistance or a lack of insulin production. It will lead to a number of different health issues [5].

The following are the three primary forms of diabetes:

- **Diabetes Type 1**, commonly known as insulin-dependent diabetes, is the most common form of diabetes. It is believed that autoimmune conditions can lead to type 1 diabetes. Diabetes type 1 develops when the immune system in our body mistakenly attacks and kills the beta cells in the pancreas that create insulin, causing the damage to be permanent. This is the most severe form of diabetes. Genetic predisposition is the primary factor in the development of type 1 diabetes [5].
- **Diabetes Type 2,** develops when the body is either unable to manufacture enough insulin or is unable to use the insulin it produces in an effective manner. As a consequence of this, sugar accumulates in the blood rather than being utilized as a source of energy. Diabetes type 2 affects approximately ninety percent of people who have the disease. Despite the fact that adults are more likely to get type 2 diabetes, children are frequently afflicted by the condition.
- **Gestational** diabetes, Diabetes that is just brief and occurs during pregnancy is referred to as gestational diabetes. It is possible to develop diabetes during pregnancy, even in people who have never been diagnosed with diabetes before, and this condition is referred to as gestational diabetes. It affects somewhere between two and four percent of all pregnancies and is associated with an increased risk of diabetes development for both the mother and the child.

Data Mining is the process of gathering useful information from massive datasets, such as associations, patterns, and anomalies, which are stored in databases and other types of data repositories. This can be accomplished through the use of techniques such as pattern recognition and anomaly detection. Data warehouses and other types of data storage facilities typically contain larger databases than other types of facilities. Knowledge discovery is an essential component of data mining, and it is comprised of the procedures that can be found stated below. These procedures include data cleaning, data integration, selection, transformation, mining, pattern evaluation, and the knowledge display of those data. The process of removing noise and missing values from a dataset is referred to as "data cleaning." This procedure also includes gathering information on the model that was used to access the noise and accounting for any adjustments that were applied. The phase that is known as "data integration" is the phase in which the primary focus is on merging data from a number of different sources. In order to retrieve the information that is needed, a subset of the data is chosen to be retrieved. In order to make the data suitable for mining, a process known as data transformation must first involve a variety of data preparation techniques. Once this is complete, the data can be mined. Normalization and aggregation are two examples of the processes that fall within this category.

"Knowledge discovery" refers to the process of automatically creating information in a manner that can be understood by human beings [3]. This process can be carried out by computers. The multiple stages that are involved in the KDD process are depicted in a graphical format in Figure 1.

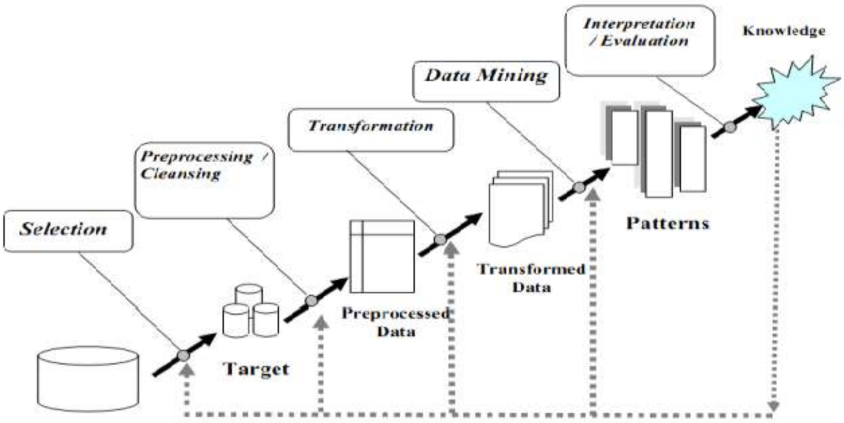


Figure 1: Steps of the KDD Process

The term "data mining" refers to a wide range of tasks, such as classifying, forecasting, analyzing time series, associating, grouping, and summarizing data. These are only some of the activities that fall under this umbrella. Each and every one of these tasks is connected to one facet or another of data mining, either the predictive or descriptive aspects. A data mining system is capable of carrying out each of the actions that were listed above, either on their own or in various combinations, as part of the data mining process.

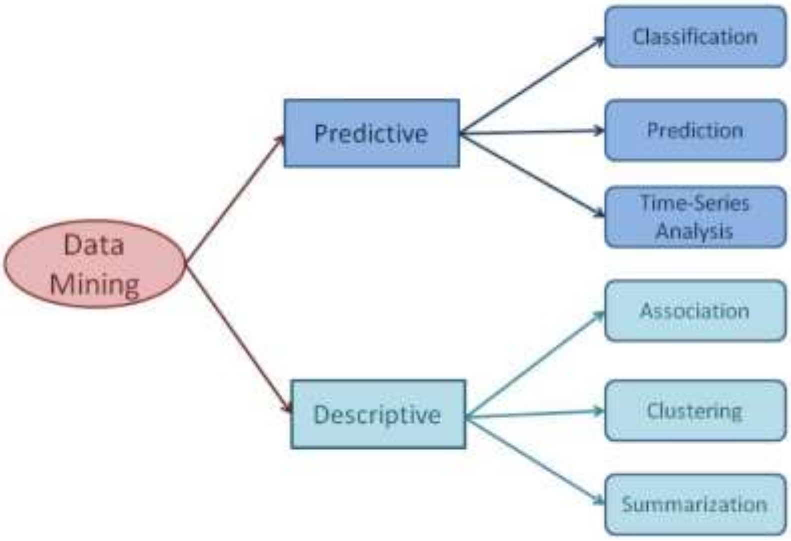


Figure 2: Data Mining Tasks

II. Literature Review

In the field of healthcare, data mining can serve as a useful educational tool for the purpose of identifying instances of fraud and abuse. It is possible to use it to make better decisions on the management of client relationships, which in turn enables hospital staff to provide better and more affordable medical care. In terms of treatment, it enables medical professionals to determine which procedures offer the highest level of care. Data modeling for healthcare applications, executive information systems for healthcare, projecting treatment costs, and demand of resources are all medical applications that frequently use data mining methods. Given the patient's past, Public Health Informatics, e-governance frameworks in healthcare, and Health Insurance data, it is possible to make predictions about the patient's behavior in the future (Dey & Rautaray 2014).

The naive Bayes algorithm is one of the most promising possibilities for extracting relevant information from medical databases. Even though this method has been used in the analysis of medical data, it is not without both benefits and drawbacks. It is a statistically straightforward classifier that operates under the presumption that attributes are independent. The great speed of classification that this algorithm maintains even when applied to big datasets is another notable feature of it. When taken into consideration for other features, its accuracy improves, which in turn makes it more suitable for use in medical data. However, it fails to function well when the degree of independence between two characteristics is difficult to discern. When there is noise present, it suffers significantly as a result. The performance of the decision tree method is comparable to that of this one.

When a physician wants to model his or her decision-making in the form of rules, the decision tree algorithm is the suitable tool to use. The classification of rules is one of the most prominent features of this algorithm (Kuo et al 2001). When attempting to quantify a patient's symptoms, the physician can apply regression on this information to make a prediction about a particular value. Even when the differentiating measure between two classes is quite tiny, it performs admirably. The decision tree method is able to handle fluidly factors such as accuracy, specificity, sensitivity, positive predictive value, and negative predictive value.

The decision tree classifier was utilized in order to achieve the best possible error ratio. Techniques such as feature selection, cross validation, error reduction pruning, and increased model complexity have all been researched and investigated. Dimensionality reduction, also known as shrinking the attribute space of a feature set, can be accomplished by the use of feature selection. This is accomplished through the elimination of data attributes that are irrelevant. Cross-validation provides a more accurate estimate of the predictive value, and it shown an improvement in accuracy of classification despite an increase in model complexity. Cross-validation is a more reliable estimation method. By adopting the approach of reduced error pruning, the over fitting problem that was affecting the decision tree was solved. When compared to the previous system, the improvement consisted of both an increase in accuracy as well as a reduction in the error rate. Construction of the decision tree takes significantly less time [4].

The SVM method is able to deal with medical databases and is an essential component in classification. SVM was developed to prevent over fitting of training samples, and with the right selection of the kernel for example, the Gaussian kernel the algorithms can place a greater emphasis on the degree to which classes are similar to one another.

When SVM is used to classify a new category, the values of its ratios are compared with the support vectors of the training sample that is most comparable to the category that is being classified. After then, this class is categorized according to how closely it resembles the other class. The significance of SVM lies in the fact that it can serve as a universal approximate for a wide variety of kernels, in addition to not having any local minima. Nevertheless, a significant limitation of the SVM is that it does not make it easy to determine which features or combinations of features have the most impact on a forecast.

K Nearest Neighbor (KNN) Algorithms have an intriguing set of qualities that make it appropriate for use on medical datasets and make it ideal for deployment on those databases. Due to the ease with which it can be implemented, the KNN method is the one that is most frequently used for pattern recognition. Despite this, there are some situations in which it is unable to deliver satisfactory results. However, the outcomes could be improved in a variety of contexts by the process of fine-tuning the parameter k in the KNN algorithm, which represents the number of neighbors (Moreno et al 2003). An investigation has been carried out on kNN through the incorporation of voting, and the investigation has been put to the test on the prediction of heart disease. According to the findings, the application of kNN has the potential to attain a greater level of accuracy in the prediction of cardiac disease than neural networks do. The classification accuracy of the dataset relating to heart disease has been improved with the combination of KNN and genetic algorithm.

Evaluation of the patient's skin temperature in all sections of the body as well as their serum levels of asymmetric dimethylarginine (ADMA) was carried out in type-2 diabetes mellitus patients. People were split into two groups: those with no complications and those with them. One group was considered normal. Using a thermography camera that did not require direct contact, thermograms of every portion of the body were taken. Biochemical measurements were taken of various blood parameters as well as thyroid hormones. In addition to that, a score for diabetic risk was computed. The posterior aspect of the sole and the ear were found to have the lowest and highest values of skin temperature, respectively, in normal persons. This was discovered through observation. For diabetes patients, the mean values of skin temperature from head to toe were lower than those of other patients, and the nose and tibia areas had a considerable fall in temperature [3].

According to a number of studies, the diagnosis of a single patient can change dramatically depending on whether or not the patient is checked by a variety of physicians, or even by the same physician at a number of different times. Automated medical diagnosis allows doctors to more accurately forecast patients' diseases in a shorter amount of time. The Naive Bayesian theorem is utilized by this system to assist in the process of obtaining data patterns. The naive Bayesian algorithm not only estimates the likelihood of a variety of dermatological conditions but also calculates the proportion of patients who suffer from each condition.

III. Data Mining Algorithms

The Expectation Maximization (EM) Algorithm

This EM technique can be broken down into two distinct phases. The first stage is to determine what to expect, and the second step is to maximize what you expect through multiple iterations of the process. The expectation begins with the selection of a model, and it continues with the estimation of any missing labels. Choosing labels and then mapping appropriate models to those labels is what the maximizing stage is all about. This is done to ensure that the expected log-likelihood of the data is maximized. The order of operations can be broken down into three stages. [2]

Step 1: The expectation step that determines mean value, denoted by μ and infers the values of x and y such that $x= [(0.5) / (0.5 + \mu) * h]$ and $y= [(\mu/ 0.5 + \mu) * h]$ with conditions of $x / y = (0.5 / \mu)$ and $h = (x + y)$.

Step 2: The maximization step that determines fractions of x and y and then computes the maximum likelihood of μ at first.

Step 3: Steps 1 and 2 are to be repeated for the next cycle. Cross validation of the mean and standard deviation for seven different characteristics was used to define the clusters. After that, everyone in the class was given a test to determine whether they had positive or negative conditions associated with diabetes. In the process of analyzing the results, binary answer variables are represented by 1 and 0 respectively. A 1 indicates that the diabetes test is positive (present), while a 0 indicates that the test is negative (not present) for diabetes. However, because of the numerical imprecision, the EM technique is not very precise when applied to data sets of higher dimensions. [2].

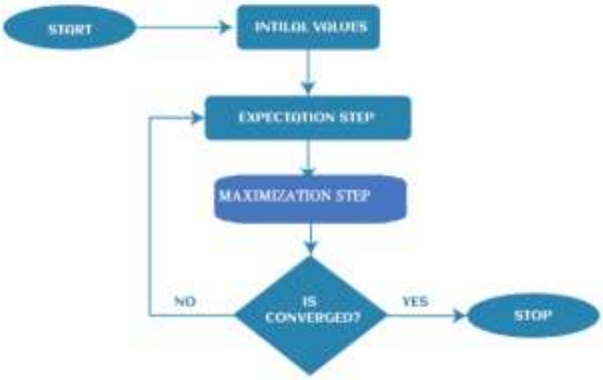


Figure 3: EM Algorithm Steps

K Nearest Neighbour Algorithm

Because of its ease of use and high level of accuracy, the K Nearest Neighbor (KNN) algorithm has been utilised in a variety of applications for the purpose of data analysis. These applications include pattern recognition, data mining, database management, and machine learning. It is one of the best 10 algorithms in the field of data mining, according to recent rankings (Wu et al 2008). KNN is a classification method that belongs to the lazy learning category. In machine learning, this is the simplest possible algorithm. Any kind of label can be predicted with the use of this technology [5].

The KNN classification organizes examples according to their degree of similarity. It is an example of a sort of learning algorithm known as "lazy learning," in which the function is approximated locally and computation is postponed until classification. The primary applications for KNN are in the areas of classification and clustering. In their trials on a wide variety of datasets, numerous researchers have discovered that the KNN algorithm achieves satisfactory results. The Pima Indian diabetes dataset is difficult to understand since it contains many missing variables. The KNN approach fills in missing values in the Euclidean Distance matrix with the appropriate values from the columns that are adjacent to the matrix. In the event that the equivalent value from the closest neighbor is likewise absent, the value from the next immediate column is used instead. When compared to other methods, this approach is not only straightforward but also extremely competitive. The lack of probabilistic semantics that enable the use of posterior predictive probabilities is one of the limitations of KNN, which is a disadvantage.

KNN has been updated by a large number of authors in order to improve its effectiveness. An implementation of the class-wise KNN (C-KNN) algorithm has been developed and validated using the Pima Indian diabetes dataset. In this step, a class label is designated for the testing data by using the class-wise distance that is the shortest. The C-KNN algorithm has reached an accuracy of 78.16%. In order to classify the cases of the Pima Indian diabetes database, the K means and KNN algorithms have been integrated into a single model known as the amalgam KNN model. In this case, the quality of the data is increased by eliminating the noise, which also results in an increase in efficiency. K-means is used to remove the instances that were incorrectly classified, and the KNN algorithm is used to complete the classification.

The choice of K in the KNN algorithm is determined by the data. In classification, having a larger value for k can help reduce noise. The cross-validation method can be used to choose an appropriate value for k. We were able to get a classification accuracy of 97.4% by first calculating the k value, and then doing ten-fold cross validation [6]. Figure 4 presents a visual representation of the KNN algorithm's core idea.

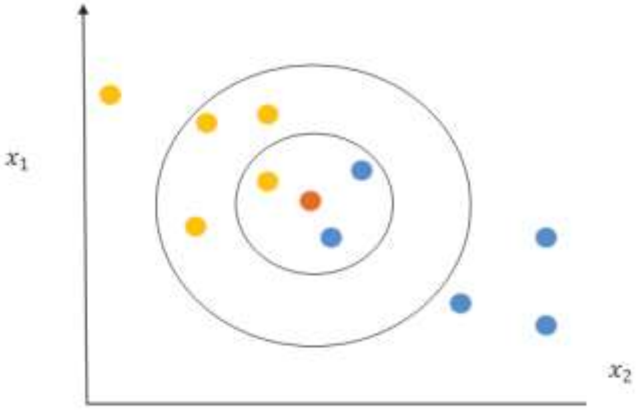


Figure 4: K nearest neighbor algorithm

The KNN algorithm:

Step 1: Each new instance is compared to the ones that are already available cases based on the distance assignment, and it is then classified using the k value.

Step2:. If the instances are more similar to one another, then the distance between them will be less, and vice versa.

Step 3: Take note of the k-value, the distance, and the instance. On the basis of these observations, occurrences are classified into the appropriate category.
Step4: The k-value serves as the foundation for the forecast. So KNN classifier is k-dependent. The number of nearest neighbors is denoted by k in this context, and depending on the value of k, the results may or may not be the same [7].
Step 5: Pima Indian Diabetic Dataset (PIDD) classification accuracy can be improved by determining the value of the parameter k.

K-Means Algorithm

Unsupervised algorithms are algorithms that can function on unlabeled samples without direct supervision. This indicates that the output cannot be predicted even if the input can be identified. The K means algorithm is one of several that fall under the category of unsupervised learning algorithms. They require an input parameter, the number of clusters, as well as n objects in the data collection, which is then partitioned into k clusters. A random selection of k items is made by the algorithm. Each object is given a place in one of the clusters that it belongs to according to how closely it is located to its related cluster. The following step is to locate the locations that are the most adjacent to one another. It is recommended to use the Euclidean distance while trying to locate the object's most central location. After the items have been divided up into k clusters, the new centers of the clusters are determined by taking the average of the objects within each of the k clusters in turn. This procedure is carried out until there is no longer any variation in the k cluster centers. The sum of squared error (SSE) is the objective function that the K-means algorithm seeks to minimize in order to achieve its goal [8]. The acronym SSE stands for

$$\mathrm{argmin}_C \quad \sum_{i=1}^{k} \sum_{p \in Ci} |p - m_i|^2 \qquad (1)$$

Here, E stands for the total squared error of the objects that have been assigned cluster means for the kth cluster, p is the item that has been assigned to the *C*ith cluster, and mi is the mean of the *C*ith cluster. The total number of records in the dataset is denoted by the letter n, while the value k indicates the number of clusters.

Input: D is input -data set.
Output: Output is k clusters.
Step 1: Set the initial values for the cluster centers to D.
Step 2: Pick k items at random from the collection D.
Step 3: Repeat the steps below until there is no change in the cluster means and the minimum error E has been obtained.
Step 4: Take into consideration each of the k clusters. When it comes to the initialization process, compare the objects' mean values across the clusters.
Step 5: Create the initial state of the object by assigning the value that is most similar to D to one of the k clusters.
Step 6: Find the average value of the objects in each of the k different clusters.
Step 7: Make the necessary adjustments to the cluster means based on the object value.

Amalgam KNN

When utilized prior to the mining process, data pre-processing techniques have the potential to either reduce the amount of time required for mining or dramatically improve the overall quality of the patterns that are mined. The pre-processing of data is a key phase in the process of knowledge discovery. This is due to the fact that quality decisions need to be predicated on quality data.

This strategy involves cleaning up noisy data, employing k-means, and substituting means and medians for values that are absent from the dataset. After the data has been preprocessed, the KNN classification is applied to the data so that it can produce better results [9].

The PIDD database contains a total of 768 examples to choose from. 192 patients had measurements taken of their skin fold thickness, 5 patients had measurements taken of their glucose levels, 11 patients had measurements taken of their body mass index, 28 others had data for their diastolic blood pressure, and 140 patients had measurements taken of their serum insulin levels. The aforementioned values are checked during the pre-processing stage, and if they are found to be inconsistent, the pre-processing step will remove them (values with a value of '0' are considered to be empty values).

- As a preliminary stage in the processing, the inconsistent values are eliminated.
- In order to lessen the amount of computing effort required by k-NN, the K-means clustering technique is applied to locate and get rid of instances that were improperly classified.
- The means and medians are substituted for the values that are missing.
- Using KNN, the final step of the procedure, which is the fine-tuned classification, is carried out by using the successfully clustered instance along with the preprocessed subset as inputs for the KNN.

- Following that, the model is tested using a variety of variables for k.

Random Forest Algorithm

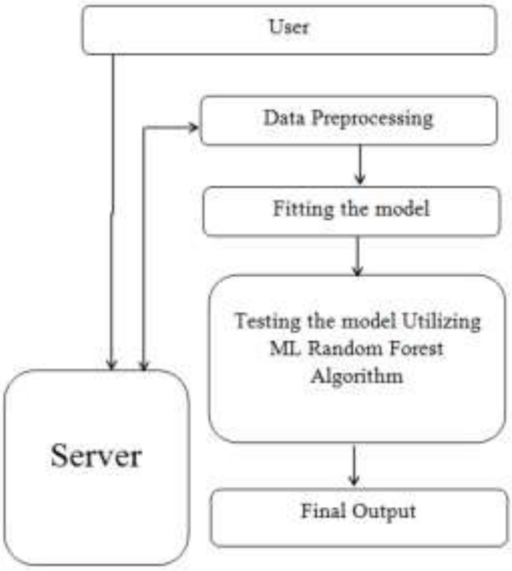


Figure 5: Flow graph of Random Forest Algorithm

To begin, the Random Forest algorithm is a form of supervised classification. The goal, which is implicit in the game's moniker, is to generate a random forest using whatever means possible. There is a correlation between the quantity of trees in a forest and the findings that it is able to produce: the more trees there are, the more precise the results will be. However, one thing to keep in mind is that the process of constructing the decision with information gain or the gain index approach is not the same as the process of constructing the forest.

The author provides readers with four links that can assist those who are dealing with decision trees for the first time in learning about them and gaining a solid comprehension of them. The decision tree is a tool that helps with decision making. To illustrate the many outcomes, it employs a graph in the form of a tree. If you provide the decision tree with a training dataset that contains targets and features, it will generate some kind of rule set for you. These rules can be used to perform predictions. The author gives one example to illustrate this point: imagine you want to anticipate whether or not your daughter will appreciate an animated movie. If so, you should compile a list of previous animated movies that she has like and utilize certain characteristics as the input for your prediction. After that, you may go ahead and generate the rules by using the decision tree technique. You are then able to input the characteristics of this movie in order to determine whether or not your daughter will enjoy it. Calculations involving information gain and the Gini index are utilized throughout the process of determining these nodes and developing the regulations.

Leo Bremen was the one who initially designed Random Forest. The Random Forest rule could be an example of a supervised classification rule [11], the Random Forest rule consists of two stages, the first of which is the creation of the random forest, and the second of which is the decision to make a prediction based on the random forest classifier that was developed in the first stage [9]. The pseudo code for Random Forest is rf, and the Random Forest rule's supervised classification counterpart is [11].

- The first thing you need to do is pick the "R" features out of the total "m" features, where R<<m.
- The node that makes use of the most optimal split point among the "R" features.
- Step Three: Using the most effective split; divide the node into daughter nodes.
- Continue to repeat steps a to c until the desired number of nodes has been achieved.
- Construct the forest by performing steps a to d a "a" number of times in order to produce a "n" number of trees.

IV. DATA SET DESCRIPTION

Since 1965, the Pima Indians of the Gila River Indian Community in Central Arizona have taken part in the study of diabetes mellitus, which has been examined every two years. The majority of the information regarding the prevalence, incidence, risk factors, and pathogenesis of diabetes in the Pima Indian population is provided by these examinations, which also include an oral glucose tolerance test and various assessments of complications of diabetes and other medical conditions (Leslie et al

2004). Numerous study findings that are pertinent to the Pima people appear to be common. Obesity, insulin resistance, insulin secretion, and an increased rate of endogenous glucose synthesis, which are the traits that identify diabetes, are metabolic features of Pima Indians with type 2 diabetes [10].

The Pima Indian diabetes dataset includes data on 768 individuals' various measures as well as a prediction of whether they would eventually develop diabetes. All of the patients in this facility were Pima Indians and at least 21 years old. This consists of eight qualities, which determine whether the tested data falls into the category of people with diabetes (tested positive) or those without diabetes (tested negative). 500 patients without diabetes (class = 0) and 268 patients with diabetes (class = 1) make up the dataset.

Table 1: Characteristics of PIMA Indian Dataset

Data Set	No. of Example	Input Attributes	Output Classes	Number of Attributes
Pima Indian Diabetes	768	8	2	9

This data set's goal was to identify Pima Indians who had diabetes. Try to determine whether a Pima Indian person had diabetes positive or not based on personal information such as age, the number of pregnancies, and the results of medical examinations such as blood pressure, body mass index, glucose tolerance test results, etc. The qualities are listed below:

1. The number of pregnancies.
2. In an oral glucose tolerance test, plasma glucose levels at two hours.
3. Diastolic pressure (mm Hg)
4. Thickness of the triceps skin fold (mm)
5. Insulin 2-hour serum (mu U/ml)
6. Body mass index (BMI) (weight in kg/ (height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

V. Result and Discussions

The effectiveness of the suggested method is assessed in this section. The proposed Protocol is subjected to simulations of the proposed algorithms. Our work is Python-based, and Tensorflow and other python libraries can be utilized for this.

Synthetic Minority Over-Sampling Technique (SMOTE)
In order to balance the number of samples in each class SMOTE analysis is been carried out. Below figures shows the item count before and after the SMOTE analysis.

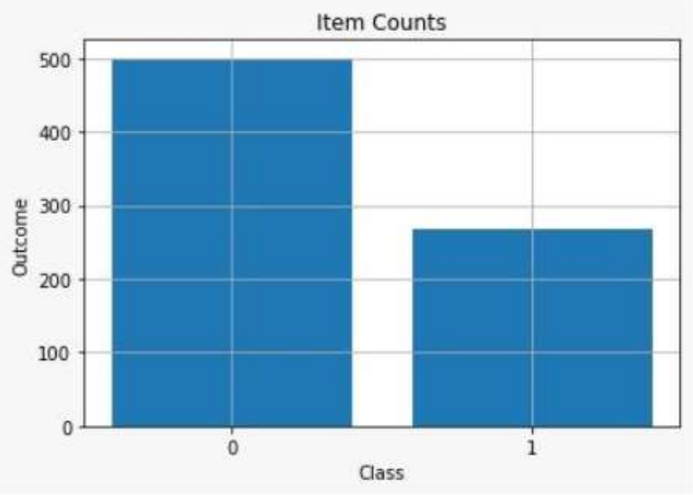


Figure 6: Item Counts

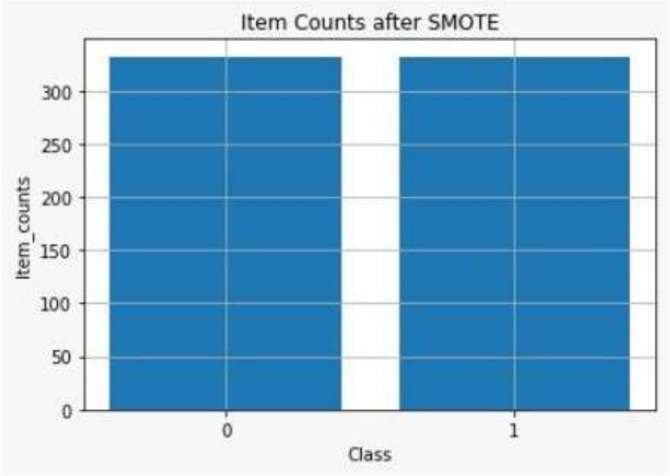
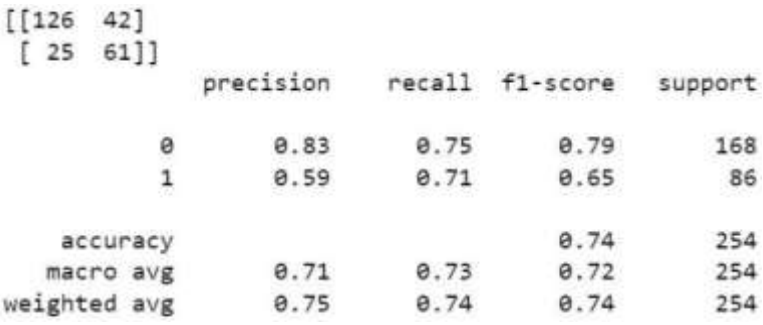


Figure 7: Item Counts

Ensemble Learning

The below performance chart shows that the accuracy of ensemble learning model to identify normal and abnormal diabetic cases is 0.74.



```
[[126  42]
 [ 25  61]]
              precision    recall  f1-score   support

           0       0.83      0.75      0.79       168
           1       0.59      0.71      0.65        86

    accuracy                           0.74       254
   macro avg       0.71      0.73      0.72       254
weighted avg       0.75      0.74      0.74       254
```

The figure 8 chart shows the Confusion matrix of ensemble learning model. The diagonal elements show the correctly classified item count and off diagonal elements show the count of misclassified elements.

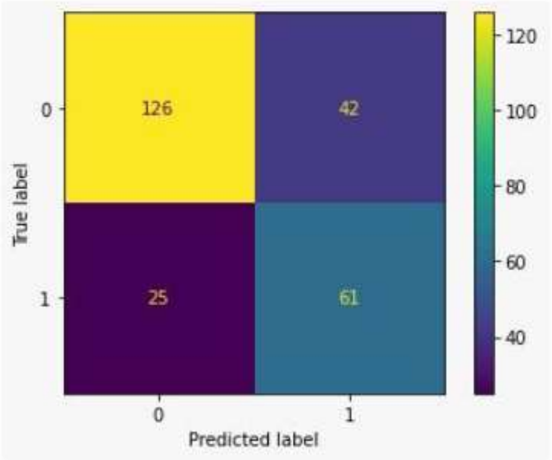
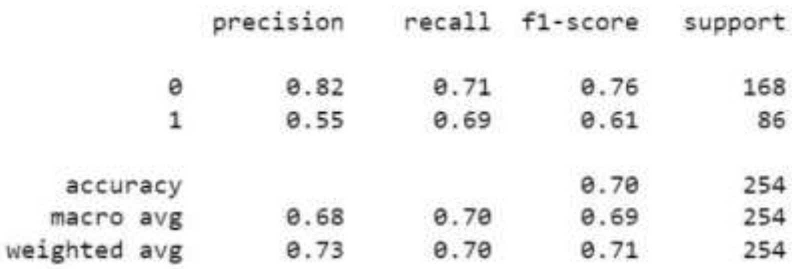


Figure 8: Confusion matrix

Logistic Regression

The below performance chart shows that the accuracy of Logistic Regression model to identify normal and abnormal diabetic cases is 0.70.



```
              precision    recall  f1-score   support

           0       0.82      0.71      0.76       168
           1       0.55      0.69      0.61        86

    accuracy                           0.70       254
   macro avg       0.68      0.70      0.69       254
weighted avg       0.73      0.70      0.71       254
```

The figure 9 chart shows the Confusion matrix of Logistic Regression model. The diagonal elements show the correctly classified item count and off diagonal elements show the count of misclassified elements.

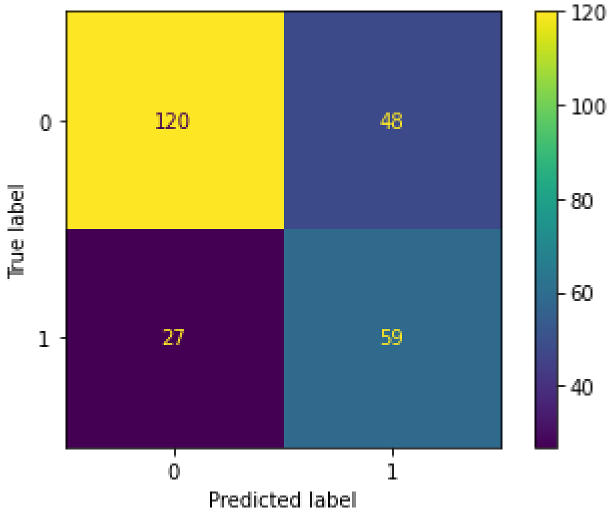
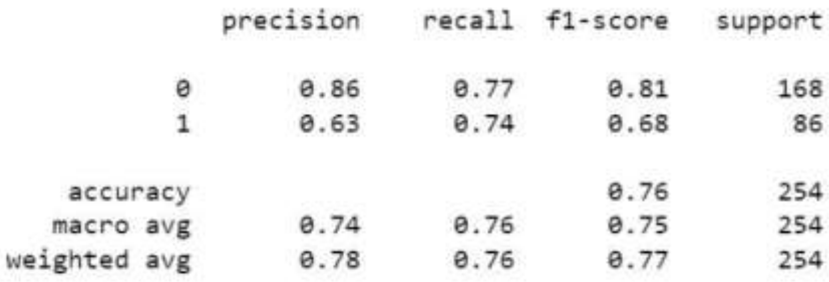


Figure 9: Confusion matrix

Random Forest

The below performance chart shows that the accuracy of Random Forest model to identify normal and abnormal diabetic cases is 0.76.



```
              precision    recall  f1-score   support

           0       0.86      0.77      0.81       168
           1       0.63      0.74      0.68        86

    accuracy                           0.76       254
   macro avg       0.74      0.76      0.75       254
weighted avg       0.78      0.76      0.77       254
```

The figure 10 chart shows the Confusion matrix of Random Forest model. The diagonal elements show the correctly classified item count and off diagonal elements show the count of misclassified elements.

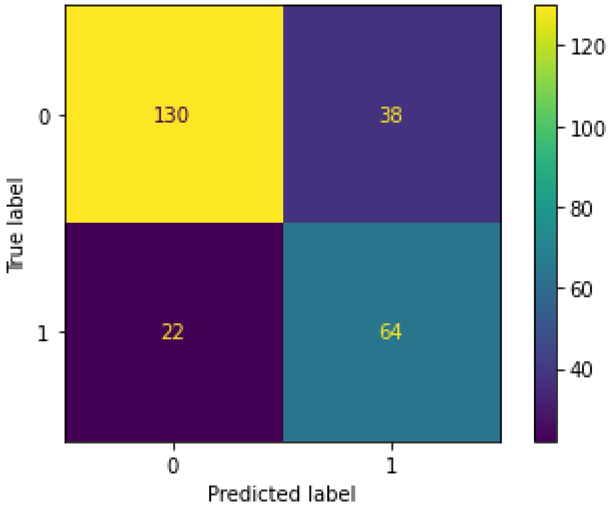
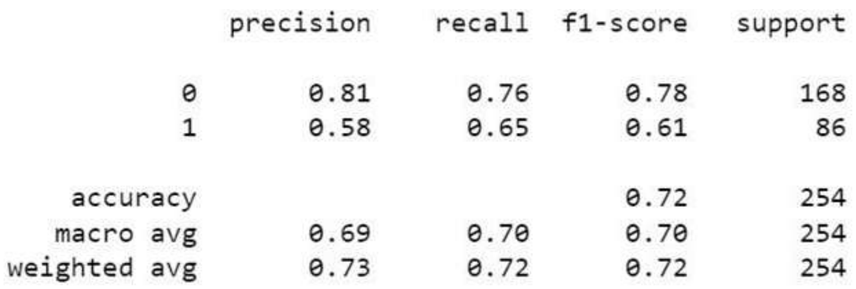


Figure 10: Confusion matrix

Gaussian NB

The below performance chart shows that the accuracy of Gaussian NB model to identify normal and abnormal diabetic cases is 0.72.



```
              precision    recall  f1-score   support

           0       0.81      0.76      0.78       168
           1       0.58      0.65      0.61        86

    accuracy                           0.72       254
   macro avg       0.69      0.70      0.70       254
weighted avg       0.73      0.72      0.72       254
```

The figure 11 chart shows the Confusion matrix of Gaussian NB model. The diagonal element show the correctly classified item count and off diagonal elements shows the count of misclassified elements.

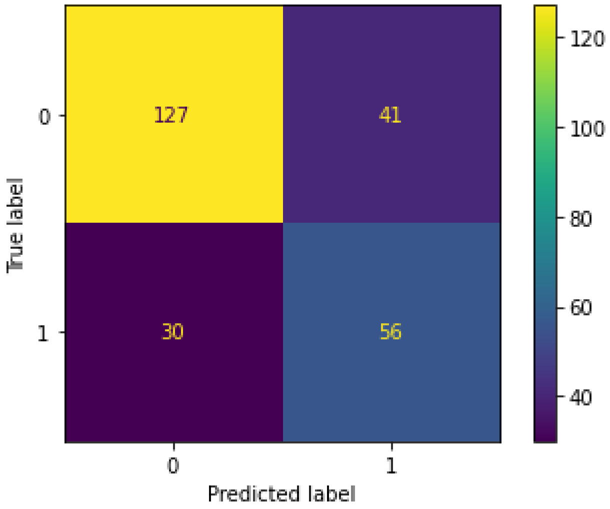
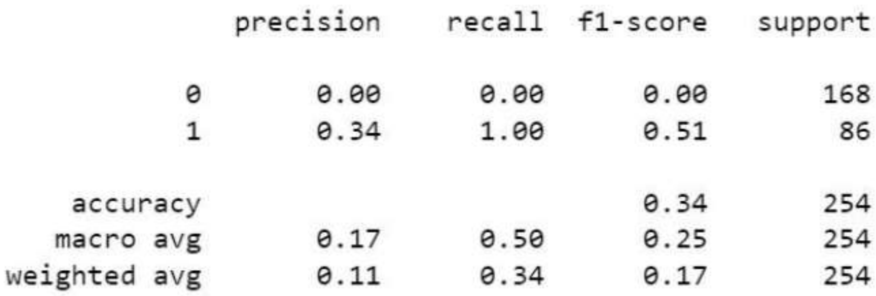


Figure 11: Confusion matrix

Artificial Neural Network (ANN)

The below performance chart shows that the accuracy of ANN model to identify normal and abnormal diabetic cases is 0.34.



```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       168
           1       0.34      1.00      0.51        86

    accuracy                           0.34       254
   macro avg       0.17      0.50      0.25       254
weighted avg       0.11      0.34      0.17       254
```

Table 2: Comparison Algorithms

S. No.	Method Name	Accuracy (%)	Sensitivity (%)	Specificity (%)
1	Ensemble learning	0.74	0.75	0.71
2	Logistic Regression	0.70	0.71	0.69
3	Random Forest	0.76	0.77	0.74
4	Gaussian NB	0.72	0.76	0.65
5	ANN	0.34	0.65	0.67

VI. Conclusion

The number of data mining tools is growing, and with it, machine intelligence algorithms. Data mining can be tested on healthcare data. In the field of healthcare, a lot of data has been acquired and arranged. The diabetes dataset is the one that has been examined the least. In this thesis, the problem of diabetes prediction has been successfully solved by using data mining methods. Three frameworks for predicting diabetes have been shown to be useful; all three are based on the well-known classification method known as the Random Forest algorithm. We can clearly see an increase in the performance of the suggested classification algorithms from the trials done on the Pima Indian diabetes dataset using Python software.

References

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

[2] Velu C.M, K.R.Kashwan,2013.”Visual Data Mining Techniques forClassification of Diabetic Patients”, IEEE International Advance Computing Conference (IACC),pp-1070-1075.

[3] Sapna. S,Tamilarasi. A and Pravin Kumar.M, 2012 “Implementation of genetic algorithm in predicting diabetes”, IJCSI, International Journal of Computer Science Issues, Vol. 9, Issue 2, No 4, pp. 393-398

[4] Nirmala Devi M.,Appavu alias Balamurugan S.,Swathi U.V., 2013.”,An amalgam KNN to predict Diabetes Mellitus”, IEEE International Conference on Emerging Trends in Computing ,Communication and Nanotechnology(ICECCN),pp 691-695.

[5] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996) ‘From Data Mining to Knowledge Discovery in Databases’ AAAI Vol.17, No.3 pp 37-54.

[6] Zoran Bosnic, Petar Vracar, Milos D. Radovic, Goran Devedzic, Nenad D. Filipovic and Igor Kononenko(2012) 'Mining Data From Hemodynamic Simulations for generating Prediction and Explanation Models' IEEE Vol. 16, No. 2,pp 248-254.

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

[8] Polat, K., Gunes, S., & Aslan, A., (2008) A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert Systems with Applications, 34(1), 214–221.

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

[10] UCI machine learning repository and archive.ics.uci.edu/ml/datasets.html.

[11] Cwiklinska-Jurkowska, M 2009, 'Performance of the support vector machines for medical classification problems', Biocybernetics and Biomedical Engineering, vol. 29, no. 4, pp. 63-81.

[12] Al-Sakran, HO 2015, 'Framework architecture for improving healthcare information systems using agent technology', International Journal of Managing Information Technology, vol. 7, no.1, pp. 17-31.

[13] Galathiya, AS, Ganatra, AP & Bhensdadia CK 2012, 'Improved decision tree induction algorithm with feature selection, cross validation, model complexity and reduced error pruning', International Journal of Computer Science and Information Technologies, vol. 3, no. 2, pp. 3427-3431.

[14] Kumar, DS, Sathyadevi, G & Sivanesh, S 2011, ' Decision support system for medical diagnosis using data mining',. International Journal of Computer Science Issues, vol. 8, no.3, pp. 147-153.

[15] D. Menon, K. Schwab, D.W. Wright, A.I. Maas, and the Demographics and Clinical Assessment Working Group of the International and Interagency Initiative toward Common Data Elements for Research on Traumatic Brain Injury and Psychological Health, Position statement: definition of traumatic brain injury, Arch. Phys. Med. Rehabil., vol. 91, pp. 1637– 40, Nov 2010.