DIABETES MELLITUS DETECTION AND DIAGNOSTICS USING DATA MINING TECHNIQUES

Arpana Shukla, M.Tech Scholar, Department of Computer Science & Engineering, Kanpur Institute of Technology, Kanpur, India.
Kanchan Gautam, Assistant Professor, Department of Computer Science & Engineering, Kanpur Institute of Technology, Kanpur, India.

Abstract— Diabetes, also known as diabetes mellitus (DM), is a condition that has the potential to be fatal and affects people from all various corners of the world. Diabetes is characterized by high blood sugar levels. Diabetes is a condition that can be brought on by a variety of risk factors, some of which include but are not limited to being overweight, having high blood glucose levels, not getting enough exercise, and other risk factors. If it is detected at a stage when it is still relatively easy to do so, there is a chance that it can be controlled or that its effects can be lessened. The development of a computer system or programme that is capable of self-improvement and the ability to gain knowledge from previous experiences is one illustration of machine learning in action. An illustration of what the field of artificial intelligence entails is shown here. Throughout the course of this investigation, the PIMA dataset is exploited in a number of different contexts. There are approximately nine characteristics that set each of the 768 instances in this collection apart from the others. There are an almost infinite number of possible methods in which each algorithmic strategy for machine learning can be put into practice. On the other hand, in order to fulfil the prerequisites of these research endeavours, we decided to implement three different unsupervised learning methods. These algorithms are well-known by their individual names, such as the logistic regression algorithm, the decision tree method, and the random forest algorithm. Before being incorporated into this model, each and every one of these algorithms underwent extensive training and testing to ensure that it was ready for usage. In the end, we will analyze the usefulness of various metric algorithmic strategies to machine learning by comparing and contrasting their respective performance levels. This will allow us to determine which of these methods are the most effective. Accuracy, F-measure, recall, and precision are some of the performance measures that are examined. There are also other performance indicators. 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%. All of these accolades come from having the finest score possible. Additionally, it has the highest value for their f-measure as well as the highest value for their precision, which is 0.73. In addition, it has the highest value. The Decision Tree technique was victorious, achieving the maximum possible recall score of 0.61 out of all the approaches.

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

I. INTRODUCTION

Diabetes Mellitus (DM) is a chronic illness that calls for ongoing medical care and education on self-management in order to reduce the risk of unfavourable long-term outcomes and the development of complications. One is able to lessen or get rid of a wide variety of diabetes-related symptoms and consequences by bringing the patient's blood sugar levels under control and treating diabetes with a mix of food and medicine. The following are the two primary types of diabetes that can be distinguished from one another: Type 1 diabetes, commonly referred to as adult-onset diabetes, is another name for the form of diabetes that affects children and adolescents. Insulin dependence is a form of diabetes that occurs when the body ceases generating the hormone known as insulin. This causes the body to become dependent on an outside source for its insulin needs. Insulin is necessary for the body to be able to use the glucose that is obtained from meals; as a result, diabetes can develop when insulin is lacking. This is very common in individuals of younger ages, particularly children and teenagers. [The chain of causation] This factor is responsible for between five and ten percent of all cases of diabetes. Injections of insulin are normally required for those diabetics who have been diagnosed with this kind of the disease in order for them to be able to survive. The overwhelming majority of people who are diabetic are diagnosed with type 2 diabetes, which is also referred to as adult-onset diabetes or diabetes that does not require the use of insulin. Diabetes mellitus type 1, often known as juvenile diabetes, is defined by an inability of the body to create sufficient quantities of insulin in the correct manner. Having a history of diabetes in one's family, being overweight, and being over the age of 40 are all factors that put a person at an elevated risk for developing type 2 diabetes. This is because diabetes is growing increasingly common in adults as a direct result of bad dietary habits [1], which explains why this is the case.

Diabetes is a condition that can be brought on by a variety of factors, some of which include, but are not limited to, the following: 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 factors, such as being overweight, being of a given race or gender, reaching a certain age, and not getting enough exercise, all play key roles in the beginning of diabetes. Some of these factors include: Researchers in 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. This is due to the fact that the number of diabetic patients around the world has increased in recent years (Ashwinkumar & Anandakumar 2012).

According to the findings of an investigation that was objectively carried out, diabetes comes in at number seven on the list of conditions that can end in mortality. These findings were used to come to this conclusion. Only in India have 51 million individuals 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 by a significant margin. Diabetes impacted approximately 7.0% of the population in the United States as of November 2007, with a total of 20.8 million people, including children and adults, being diagnosed with the condition. According to the findings of a global survey that was carried out in 2013 by Boehringer Ingelheim and Eli Lilly and company, there are 25.8 million people in the United States who are afflicted with Type-1 diabetes and 382 million people throughout the world who are plagued with Type-2 diabetes. The prevalence of type 2 diabetes, which is the most frequent form of the disease and is considered to account for 90–95% of all instances of diabetes, is a significant problem in both industrialised and developing countries. This is because type 2 diabetes is the most common form of the disease.

According to some projections that were created by the International Diabetes Federation (IDF), the number of individuals in the globe who are currently living with diabetes would increase to 592 million by the year 2035. These forecasts were originally developed in the year 2005. According to the World Diabetes Atlas, there are around 285 million individuals living with diabetes across the globe at the present time, and this number has the potential to increase all the way up to 438 million by the year 2030. The results of a poll suggested that the number of persons suffering from type 2 diabetes will rise by the year 2030, which sparked alarming forecasts for the future. In accordance with the findings of 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. This forecast is based on the fact that there is an anticipated increase in the number of people who have diabetes. It is projected that the number of people living in India who are afflicted with diabetes would rise from 31.7 million in the year 2000 to 79.4 million in the year 2030. This projection is based on current numbers. (Huy Nguyen et al 2004). Obtaining an accurate diagnosis as fast as possible is one of the most essential components of diabetes treatment that will lead to success (Mythili et al 2003).

There are already more than 62 million people in the Republic of India who are afflicted with diabetes, which indicates that the condition is in the process of fast approaching the status of a potential epidemic. According to study conducted by Wild et al., the number of people living with diabetes is projected to more than double from 171 million in the year 2000 to 366 million in the year 2030. India is anticipated to experience the biggest growth in this epidemic. By the year 2020, it is anticipated that India will 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. The number of diabetics in India's population is expected to increase significantly by the year 2020. Diabetes has the potential to become a big burden for India in the future, and because of this likelihood, the country is now facing an uncertain future. [2].

Diabetes is a collection of disorders in which the body either does not create enough insulin or does not use the insulin that is generated in the correct manner, or a combination of both of these factors. Diabetes can also occur when the body does not use insulin in the correct manner. If this were to take place, the body would be unable to transport sugar from the blood into the cells, which would result in an increase in the amount of glucose found in the blood. Glucose is the name given to the type of sugar that is found in our blood, and it is one of the key sources of energy that our bodies use. Insulin resistance or a lack of insulin production can both lead to a buildup of sugar in the blood, which is a sign of diabetes. It will result in a variety of negative effects on one's health. [5].

The following are the three primary forms of diabetes:

- **Diabetes Type 1**, Diabetes mellitus, more frequently referred to as insulin-dependent diabetes, is the most prevalent form of the disease. The development of type 1 diabetes is thought to be linked to autoimmune disorders. Diabetes type 1 arises when the immune system in our body erroneously assaults and kills the beta cells in the pancreas that make insulin, causing the damage to be permanent. This results in the development of diabetes type 1. Diabetes mellitus type 1 is the most severe form of the disease. The presence of a genetic predisposition is the most important contributor to the onset of type 1 diabetes [5].
- **Diabetes Type 2,** diabetes mellitus is a condition that either results from the body's inability to create enough insulin or from its inability to make good use of the insulin it does produce. Because of this, sugar builds up in the blood rather than being used as a source of energy, and this can lead to serious health problems. About ninetieth of persons who have diabetes are diagnosed with diabetes type 2, which is the most common form of the condition. Children are frequently affected by diabetes type 2, despite the fact that adults are more prone to get the ailment.
- **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.

The process of extracting usable information from enormous datasets, such as associations, trends, and anomalies that are held in databases and other types of data repositories is referred to as "data mining." Pattern recognition and the identification of anomalies are two methods that can be utilized in order to attain this goal. It's common for data warehouses and other sorts of data storage facilities to have databases that are far larger than those found in other types of facilities. Knowledge discovery is a crucial part of data mining, and it is made up of the processes that can be found outlined below. This section of the article can be found here. These processes involve cleaning the data, integrating the data, selecting the data, transforming the data, mining the data, evaluating the patterns found in the data, and displaying the information obtained from the data. "Data cleaning" refers to the process of removing unwanted elements, such as noise and values that are missing, from a dataset. Gathering information on the model that was used to access the noise and accounting for any adjustments that were performed are also part of this method. The phase that is known as "data integration" is the phase in which the primary focus is placed on merging data from a number of various sources. This phase is also known as "data integration phase." It is necessary to select a subset of the data to obtain in order to retrieve the specific information that is required. In order to make the data acceptable for mining, a process known as data transformation must first incorporate a number of approaches for data preparation. Once this is complete, the data will be mining-ready. The data can be mined once this step has been finished. This category includes a wide variety of different procedures, some examples of which are normalization and aggregation.

The process of automatically creating information in a format that can be comprehended by human beings is referred to as "knowledge discovery" [3]. Computers are able to complete this operation successfully. Figure 1 is a diagram that presents the many steps that are involved in the KDD process. These steps are shown in a sequential order.

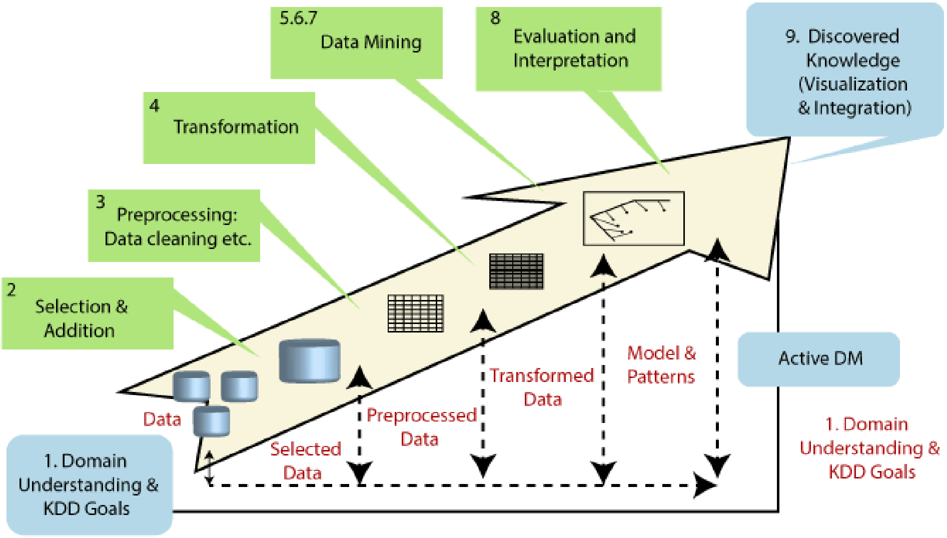

5.6.7
Data Mining
8
Evaluation and Interpretation
9. Discovered Knowledge (Visualization & Integration)
4
Transformation
3
Preprocessing: Data cleaning etc.
2
Selection & Addition
Active DM
Model & Patterns
Transformed Data
Preprocessed Data
Data
Selected Data
1. Domain Understanding & KDD Goals
1. Domain Understanding & KDD Goals

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.

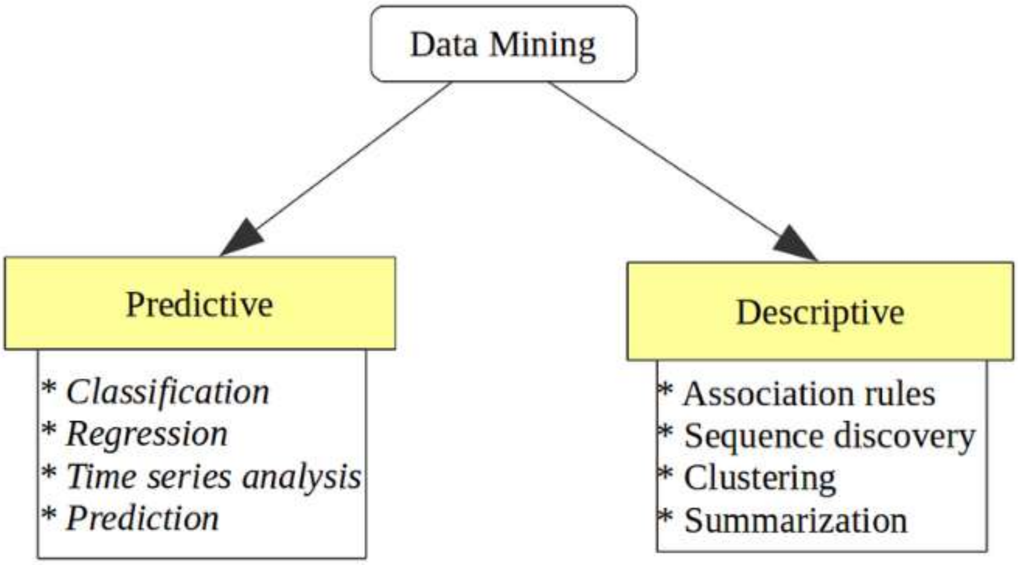


Figure 2: Data Mining Tasks

II. Literature Review

Data mining can be a very helpful instructional tool in the field of healthcare, particularly with regard to the objective of uncovering cases of fraud and abuse. Because of this, it is feasible to use it to make better judgments on the management of client relationships, which in turn enables hospital staff to provide better and more affordable medical care. It enables medical practitioners to identify which methods offer the best degree of care, which is beneficial in terms of treatment. Medical applications commonly make use of data mining techniques, including data modelling for healthcare applications, executive information systems for healthcare, forecasting treatment costs, and demand of resources. It is possible to develop forecasts about a patient's behaviour in the future by looking at data from the patient's past, as well as information from Public Health Informatics, e-governance frameworks in healthcare, and health insurance (Dey & Rautaray 2014).

In the process of gleaning useful information from medical databases, the naive Bayes algorithm stands out as one of the most interesting and potentially fruitful options. In spite of the fact that this methodology has been utilized in the process of analyzing medical data, it is not devoid of either benefits or drawbacks. It is a statistically uncomplicated classifier that functions on the premise that qualities are free to change without influence from other factors. Another important quality of this approach is that it is able to retain a high rate of classification accuracy even when applied to very large datasets. Its accuracy improves when other features are taken into consideration, which, in turn, makes it more appropriate for use in medical data. On the other hand, it does not perform very well in situations in which determining the degree of independence between two qualities is challenging. It endures a large amount of detriment as a direct consequence of the presence of noise. The performance of this method and the decision tree method are roughly equivalent to one another.

The decision tree algorithm is the appropriate instrument to employ in situations in which a medical practitioner desires to represent his or her decision-making in the form of rules. One of the most notable characteristics of this algorithm is the organization of the rules into categories (Kuo et al 2001). When a doctor is attempting to quantify a patient's symptoms, they can use regression on the collected data to produce a prediction about a specific value. It operates admirably even in situations in which the differentiating measure between two groups is quite minute. Accuracy, specificity, sensitivity, positive predictive value, and negative predictive value are some of the criteria that can be smoothly handled by the decision tree technique.

The decision tree classifier was used in order to get the best error ratio that was possible. Methods that include researching and investigating techniques such as feature selection, cross validation, error reduction pruning, and increasing model complexity. Utilizing feature selection is one method for accomplishing dimensionality reduction, which is also referred to as the process of condensing the attribute space of a feature collection. This is performed by removing data attributes that are of no use and are deemed to be irrelevant. Cross-validation provides a more accurate assessment of the predictive value, and it has shown an improvement in accuracy of classification despite an increase in model complexity. This is the case even though cross-validation was performed on a more complicated model. The estimation method known as cross-validation is one that is more trustworthy. The overfitting problem that had been harming the decision tree was successfully resolved by adopting the strategy of reduced error pruning as a solution. When compared to the previous system, the improvement included both an increase in accuracy as well as a reduction in the mistake rate. In other words, the improvement was a win-win situation. The amount of time required to construct the decision tree is drastically reduced [4].

The Support Vector Machine (SVM) technique may work with medical databases and is an important part of the categorization process. SVM was created to prevent overfitting of training samples, and with the appropriate selection of the kernel, for instance the Gaussian kernel, the algorithms can place a larger focus on the degree to which classes are similar to one another in comparison to other degrees of similarity.

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. This ensures that the new category may be accurately categorized. Following that, this class will be classified further according to the degree to which it is comparable to the other class. In addition to the fact that it does not contain any local minima, the relevance of SVM resides in the fact that it may operate as a universal approximate for a wide variety of kernels. This is just one of the reasons why it is so important. However, a fundamental disadvantage of the SVM is that it does not make it easy to discover which features or combinations of data have the most impact on a forecast. This is one of the most serious limitations of the SVM.

The K Nearest Neighbor, or KNN, Algorithm, possesses an exciting mix of properties that make it suitable for usage on medical datasets and make it excellent for deployment on those databases. These qualities also make it appropriate for use on other types of databases. The KNN method is the one that is utilized for pattern recognition the majority of the time because of how simple it is to put into action. This is the case because it is the most reliable. In spite of this, there are certain circumstances in which it is unable to give outcomes that are satisfactory. However, the results could be improved in a variety of settings by the process of fine-tuning the parameter k in the KNN algorithm. This parameter represents the number of neighbours, and it is responsible for determining how similar a given value is to its neighbours (Moreno et al 2003). An study into kNN has been carried out through the implementation of voting, and the investigation has been tested on the forecasting of cardiovascular illness. According to the findings, the implementation of kNN has the potential to achieve a higher degree of accuracy in the prediction of cardiac disease than neural networks do. This is the case despite the fact that neural networks are now the industry standard. The use of KNN in conjunction with a genetic algorithm has resulted in an increase in the dataset's ability to be classified more accurately in terms of heart disease.

Patients with type 2 diabetes mellitus had their skin temperature measured across the entirety of their bodies, and their serum levels of asymmetric dimethylarginine (ADMA) were analyzed as well. Both of these factors were considered in the diagnostic process. People were divided into two groups: those who did not have any complications and those who did have complications. One category of people were thought to be typical. Thermograms were taken of every part of the body by employing a thermography camera that did not require any direct touch with the subject. Several blood parameters including thyroid hormones were measured biochemically, as well as other blood components. In addition to that, a score indicating the likelihood of developing diabetes was determined. In normal people, the values of skin temperature that were discovered to be the lowest were located on the posterior portion of the sole, and the values that were found to be the highest were found on the ear. This was learned through the process of observation. Patients who had diabetes had lower mean values of skin temperature from head to toe than other patients did, and the nose and tibia areas had a significant drop in temperature [3]. This was the same for all areas of the body.

According to the findings of a number of studies, the diagnosis of a single patient can shift significantly depending on whether or not the patient is examined by a variety of physicians, or even by the same physician at a number of different times. This is true even if the patient is checked out by the same physician on multiple occasions. The use of automated medical diagnostics enables physicians to predict the diseases of their patients with more accuracy and in a shorter amount of time. This method employs the Naive Bayesian theorem in order to facilitate the process of identifying patterns in the data that it collects. Not only does the naïve Bayesian algorithm calculate the percentage of patients who suffer from each dermatological problem, but it also assesses the chance of a wide range of conditions that can affect the skin.

III. DATA MINING STRATEGIES

The Expectation Maximization (EM) Algorithm

This electromagnetic method can be divided down into two separate parts. The first step is to figure out what to anticipate, and the second step is to optimize what you anticipate by going through the procedure several times. The process of estimating any missing labels follows the selection of a model as the first step in the expectation, which also includes the selection of the model. During the maximizing stage, you will select labels and then map relevant models to those labels. This will be done in order to maximize your results. This is done in order to maximize the expected log-likelihood of the data, which is the goal of the procedure. There are three distinct stages that can be distinguished within the operational order [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. The clusters were defined through the application of cross validation of the mean and standard deviation for a total of seven different features. After that, a test was administered to each student in the group to assess whether or not they had any positive or negative conditions connected with diabetes. In the course of performing an analysis of the data, binary response variables are alternately represented by the numbers 1 and 0. If the test for diabetes returns a 1, it means that the test is positive (present), and if it returns a 0, it means that the test for diabetes is negative (not present). When used to data sets of larger dimensions, the EM approach, however, is not very exact as a result of the numerical imprecision [2].

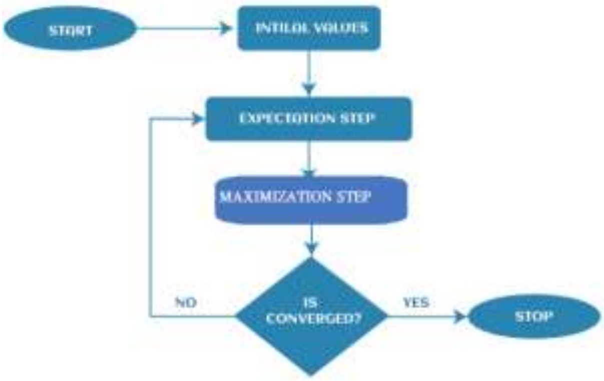


Figure 3: EM Algorithm Steps

K Nearest Neighbour Algorithm

The K Nearest Neighbor (KNN) method has found use in a wide variety of applications for the purpose of data analysis due to the fact that it can be implemented with relative ease and provides a high level of accuracy. Pattern recognition, data mining, database administration, and machine learning are some of the applications that fall into this category. According to the most recent rankings, it is among the top 10 algorithms that can be used in the field of data mining (Wu et al 2008). The KNN algorithm is a classification approach that falls under the umbrella of "lazy learning." This is the simplest form of the algorithm that can be used in machine learning. Using this technology, it is possible to forecast the appearance of any sort of label [5].

The KNN classification arranges samples in accordance with the degree to which they are similar to one another. It is an illustration of a type of learning algorithm known as "lazy learning," in which the function is approximated locally and computation is postponed until classification. In this kind of learning method, the function is approximated. K-Nearest Neighbors is most useful for applications in the fields of classification and clustering. Numerous researchers have found, after testing the KNN algorithm on a wide variety of datasets, that it produces results that meet or exceed their expectations. Because there are so many factors that are missing from the Pima Indian diabetes dataset, it is quite difficult to comprehend. The KNN method utilizes the columns of data that are immediately surrounding the matrix to determine which values should be substituted for those that are absent in the Euclidean Distance matrix. In the event if the value that is equal from the column that is the closest neighbour is also absent, the value from the column that is the next immediate neighbour is utilized instead. In contrast to other approaches, not only is this strategy uncomplicated, but it also offers a significant advantage in terms of competition. One of the drawbacks of KNN, which can be seen as a negative, is that it does not make use of probabilistic semantics, which would allow for the application of posterior prediction probabilities.

In an effort to make KNN more useful, a huge number of its writers have contributed to its most recent upgrade. The class-wise KNN (C-KNN) algorithm has been implemented, and its performance on the Pima Indian diabetes dataset has been validated. A class label is assigned to the testing data at this stage by making use of the class-wise distance that is the shortest. A level of accuracy of 78.16% has been achieved by the C-KNN algorithm. The K means and KNN classification algorithms have been combined into a single model known as the amalgam KNN model in order to facilitate the categorization of the diabetes cases contained inside the Pima Indian database. In this instance, the quality of the data is improved by getting rid of the noise, which also leads to an increase in the amount of work that can be accomplished in the same amount of time. The K-means algorithm is used to exclude the instances that were improperly classified, and the KNN classification algorithm is used to finish the classification.

The data will dictate the value of K should be used by the KNN algorithm. When it comes to categorization, having a higher value for k can assist cut down on the amount of noise. A suitable value for k can be decided upon by the application of the cross-validation approach. By first determining the k value and then carrying out ten-fold cross validation [6,] we were able to achieve a classification accuracy of 97.4%. Figure 4 provides a graphical illustration of the fundamental concept underlying the KNN algorithm.

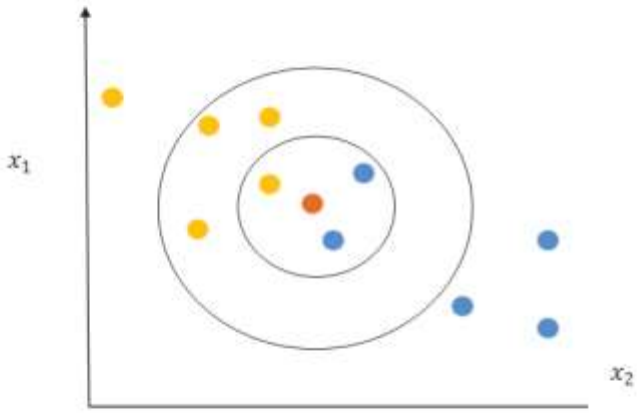


Figure 4: K nearest neighbor algorithm

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 those that are able to function well on unlabeled samples even in the absence of direct supervision. This suggests that it is impossible to forecast the output even if it is possible to determine the input. Unsupervised learning algorithms include a number of different methods, including the K means algorithm, which is one of these methods. They require an input parameter, which is the number of clusters, as well as n objects in the data collection, which is then partitioned into k clusters in order to work properly. The algorithm makes a choice out of the available k items based on a random selection. According to how close an item is situated to the linked cluster to which it belongs, it is assigned a specific location within one of the clusters that it belongs to. The subsequent stage is to determine which areas are the closest in proximity to one another. When trying to find the location of the object that is the most central to it, it is recommended that you utilize the Euclidean distance. After the items have been organized into k clusters, the new centres of the clusters are found by averaging the items contained inside each of the k clusters in turn. This process is repeated until all of the clusters have been exhausted. Following this technique up until the point where there is no longer any fluctuation in the k cluster centres is done. In order for the K-means algorithm to be successful in accomplishing its mission, the objective function that it aims to decrease is the sum of squared error (SSE) [8]. The abbreviation SSE refers to the following:

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

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

Data pre-processing techniques, when utilized prior to the mining process, have the potential to either reduce the amount of time required for mining or significantly improve the overall quality of the patterns that are mined. If both of these outcomes are achieved, then the utilization of data pre-processing techniques is highly recommended. The part of the process known as knowledge discovery known as pre-processing of the data is an important step. This is owing to the fact that quality judgments have to be based on quality data in order to be valid.

To implement this method, you will need to clean up noisy data, use k-means, and substitute means and medians for values that are missing from the dataset. The KNN classification is applied to the data after the data has been preprocessed in order for the classification to be able to generate 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

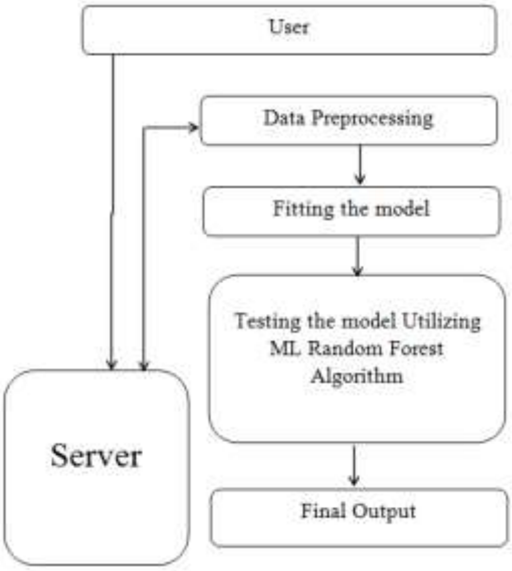


Figure 5: Flow graph of Random Forest Algorithm

To get things started, the algorithm known as Random Forest is a type of supervised categorization. The title of the game gives away the objective, which is to produce a random forest by whatever methods available. This objective can be accomplished in a number of ways. There is a relationship between the number of trees in a forest and the discoveries that it is able to make; the more trees there are, the more accurate the findings will be. However, one thing to bear in mind is that the process of creating the forest is not the same as the process of constructing the choice with information gain or the gain index approach. Keep this in mind.

The author gives readers access to four websites that might be of assistance to individuals who are working with decision trees for the first time in terms of learning about them and obtaining a thorough knowledge of what they are all about. A useful instrument for assisting in decision-making is called a decision tree. It uses a graph in the shape of a tree to represent the many different outcomes that can occur. If you give the decision tree a training dataset that includes targets and features, it will come up with some form of rule set for you to follow on its own. Utilizing these guidelines will allow one to make accurate forecasts. Consider the following scenario, which the author uses to illustrate his thesis with an example: you are trying to determine whether or not your daughter will like watching an animated film. If this is the case, you should make a list of past animated movies that she has like and use particular aspects of those movies as inputs for your forecast. After that, you are free to proceed with the generation of the rules by utilizing the technique of decision trees. You will then be able to determine whether or not your daughter will love this movie by inputting the qualities of the film and seeing what results you get. Throughout the entirety of the process of finding these nodes and developing the laws, calculations involving information gain and the Gini index are applied.

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 [16]:

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

In this part of the article, the usefulness of the approach that was suggested is evaluated. The simulated implementations of the proposed algorithms are used to test the validity of the proposed Protocol. Tensorflow and various other Python libraries can be utilized for this purpose; our work is based on the Python programming language.

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.

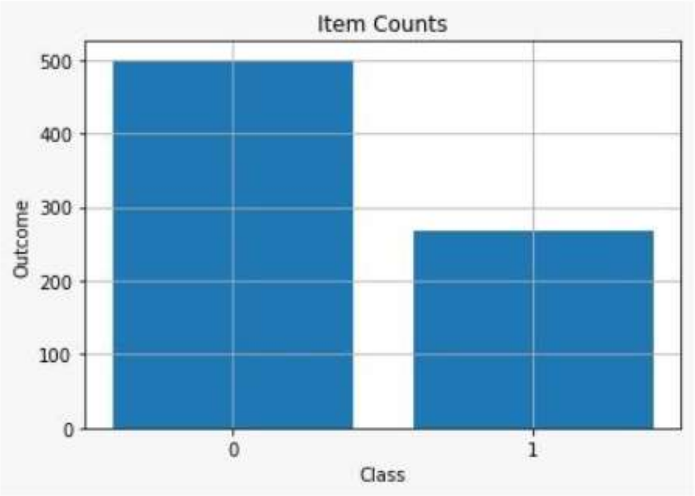


Figure 6: Item Counts

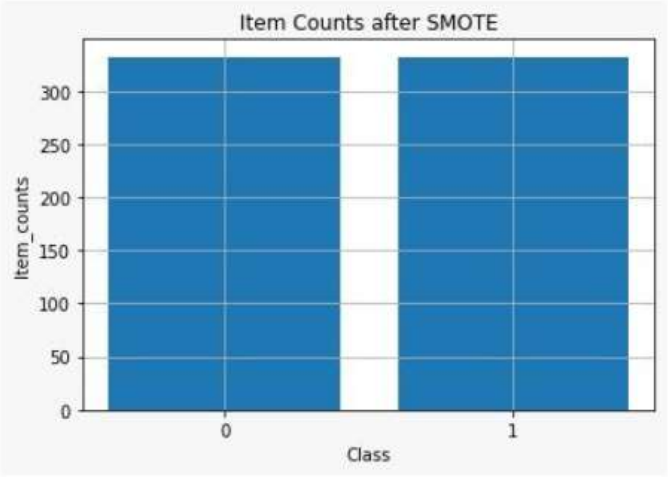
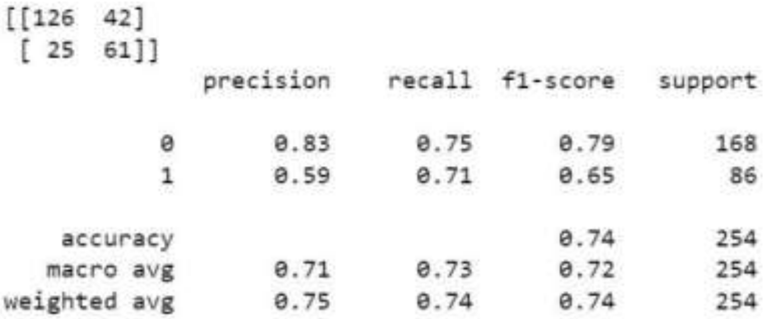


Figure 7: Item Counts

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.

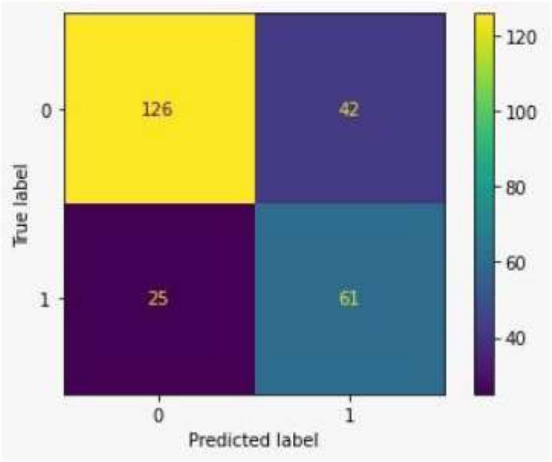
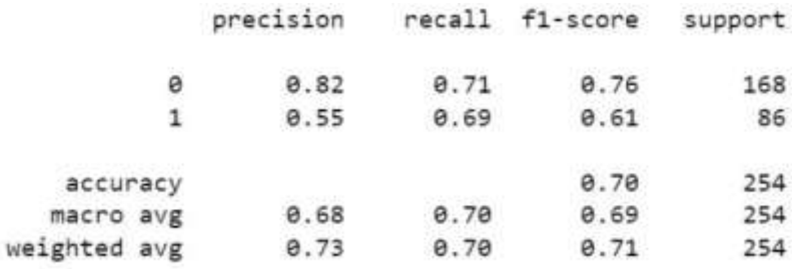


Figure 8: Confusion matrix

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.

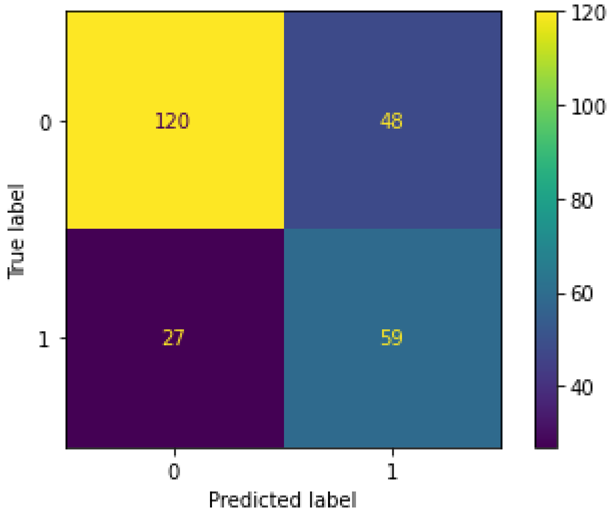
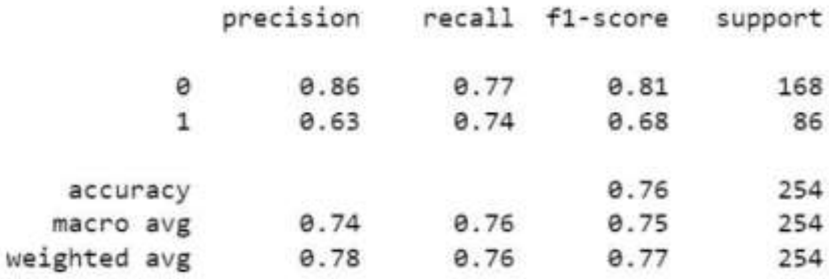


Figure 9: Confusion matrix

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.

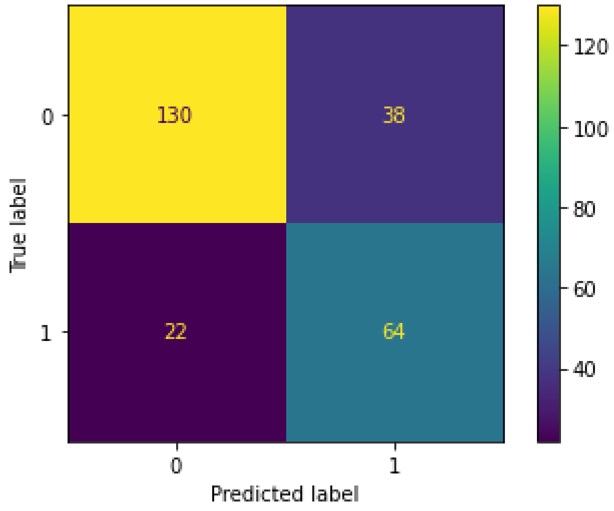
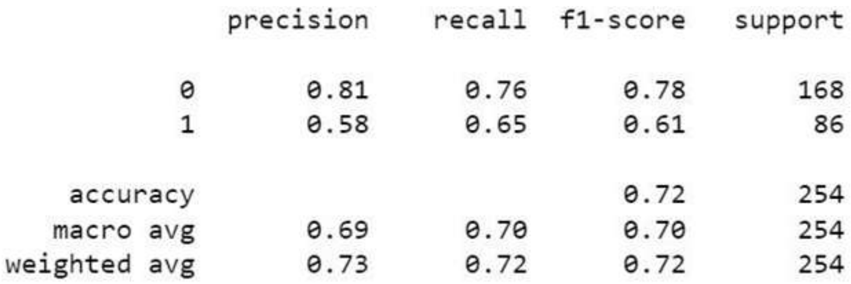


Figure 10: Confusion matrix

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.

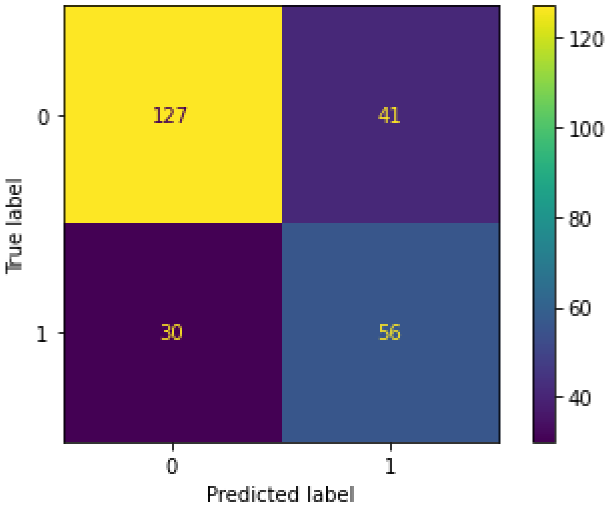
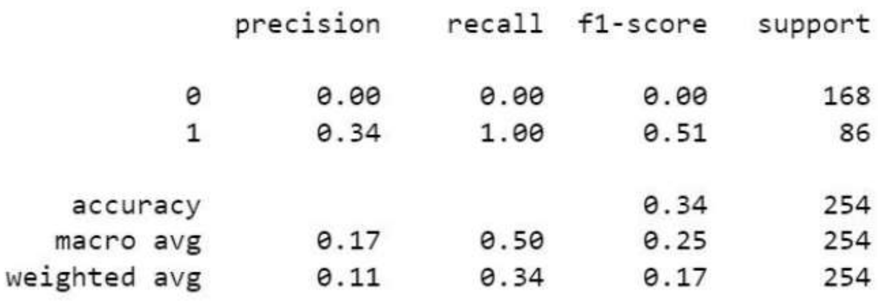


Figure 11: Confusion matrix

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 quantity of data mining tools, and along with them, the number of algorithms for machine intelligence, is expanding. The mining of data can be practiced on patient medical records. A significant amount of information has been gathered and organized in the field of healthcare. The dataset pertaining to diabetes is the one that has received the fewest number of analyses. The subject of diabetes prediction is successfully tackled and resolved throughout the entirety of this thesis by utilizing data mining techniques. It has been demonstrated that three different predictive models for diabetes are useful, and each of these models is founded on the same well-known algorithm for classification, which is known as the Random Forest algorithm. From the tests that were run on the data set containing Pima Indians with diabetes using the Python programme, it is abundantly obvious that the performance of the suggested classification algorithms improved significantly.

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] Kenney, WL & Munce, TA 2003, 'Invited review: aging and human temperature regulation', Journal of Applied Physiology, vol. 95, no. 6, pp. 2598-2603.

[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] Khaing, HW 2011, 'Data mining based fragmentation and prediction of medical data', Proceedings of the third international conference on computer research and development, vol. 2, pp. 480-485.

[5] Kuo, WJ, Chang, RF, Chen, DR & Lee, CC 2001, 'Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images', Breast Cancer Research and Treatment, vol. 66, no. 1, pp.51-57.

[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] Lakshmi, KR & Kumar, SP 2013, 'Utilization of data mining techniques for prediction of diabetes disease survivability', International Journal of Scientific and Engineering Research, vol. 4, no. 6, pp. 933-942.

[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] Manjusha, KK, Sankaranarayanan, K. & Seena, P 2014, 'Prediction of different dermatological conditions using naïve Bayesian classification', International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 1, pp. 864-868.

[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] Mythili, T, Mukherji, D, Padalia, N, & Naidu, A 2013, 'A heart disease prediction model using SVM-decision trees-logistic regression (SDL)', International Journal of Computer Applications, vol. 68, no.16, pp. 11-15.

[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] Palaniappan, S & Awang, R 2008, 'Intelligent heart disease prediction system using data mining techniques', Proceedings of the IEEE in computer systems and applications, pp. 108-115.

[16] Suguna, N & Thanushkodi, K 2010, 'An improved K-nearest neighbor classification using genetic algorithm' ' International Journal of Computer Science, vol. 7 no. 2, pp. 18-21.