IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION AND DIAGNOSIS OF DIABETES DISEASES UTILIZING RANDOM FOREST CLASSIFIER

Mohammad Saif Raza, M.Tech Scholar, Department of Computer Science & Engineering, Rameshwaram Institute of Technology & Management, Lucknow, India.

Shubham Mishra, Assistant Professor, Department of Computer Science & Engineering, Rameshwaram Institute of Technology & Management, Lucknow, India.

Abstract— Diabetes, usually referred to as diabetes mellitus (DM), is a potentially dangerous disorder that affects people all over the world. High blood sugar levels are a sign of diabetes. Numerous risk factors, such as being overweight, having high blood glucose levels, not exercising enough, and other risk factors, can lead to the development of diabetes. There is a potential that it can be managed or that its effects may be diminished if it is discovered when it is still fairly simple to do so. One example of machine learning in action is the creation of a computer programme or system that has the ability to learn from its past experiences and better itself. Here is an example of the scope of the artificial intelligence field. The PIMA dataset is used in a variety of scenarios throughout the course of this inquiry. Each of the 768 cases in this collection differs from the others in about nine different ways. Each algorithmic machine learning strategy has an almost endless number of implementations that can be used. On the other hand, we choose to use three different unsupervised learning techniques to satisfy the requirements of these research initiatives. The logistic regression algorithm, the decision tree method, and the random forest algorithm are all well-known algorithms by their respective names. Each and every one of these algorithms underwent rigorous training and testing before being put into this model to make sure it was fit for use. By contrasting and analysing the various metric algorithmic techniques' respective performance levels, we will ultimately analyse the applicability of each one to machine learning. This will enable us to decide which of these approaches is the most successful. Some of the performance measures that are looked at include accuracy, F-measure, recall, and precision. Other performance measures are also available. The best accuracy score is 74% for the Logistic Regression model, which also has the highest overall score of 0.68 and the highest f-measure value. Getting the highest score possible is how all of these honours came to be. Additionally, with a precision value of 0.73, it has the greatest f-measure and f-measure precision values. It also has the greatest worth. Out of all the methods, the Decision Tree methodology won, obtaining the highest recall score of 0.61.

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

I. INTRODUCTION

Diabetes Mellitus (DM) is a chronic condition that necessitates constant medical attention as well as instruction in self-management to lower the risk of negative long-term results and the emergence of complications. By keeping the patient's blood sugar levels under control and treating diabetes with a combination of food and medication, one is able to minimise or eliminate a wide variety of diabetes-related symptoms and effects. The two main forms of diabetes that can be recognised from one another are as follows: The type of diabetes that affects children and teenagers is known as type 1 diabetes, sometimes known as adult-onset diabetes. A kind of diabetes known as insulin dependency develops when the body stops producing the hormone known as insulin. As a result, the body starts to rely on external sources of insulin. Diabetes can occur when there is insufficient insulin because the body needs insulin to be able to utilise the glucose from food. This happens frequently to those who are younger, especially kids and teenagers. [The causal relationship] Five to ten percent of all instances of diabetes are brought on by this cause. For diabetics who have been diagnosed with this type of condition, insulin injections are typically necessary for them to be able to survive. Type 2 diabetes, also known as adult-onset diabetes or diabetes that does not require the use of insulin, is the most common type of diabetes and affects the vast majority of diabetics. Juvenile diabetes, also known as diabetes mellitus type 1, is characterised by the body's inability to produce enough insulin in the right amounts. A person's risk of getting type 2 diabetes is increased by variables like being overweight, having a family history of the disease, and being over 40. This is due to the fact that diabetes is becoming more and more prevalent in adults due to poor eating habits [1], which explains why this is the case.

A number of variables, including but not limited to high blood pressure, being overweight, kidney failure, high cholesterol, blindness, and a lack of physical activity, can lead to diabetes (American Diabetes Association, 2004). It would seem that the onset of diabetes is influenced by both hereditary and environmental variables, like being overweight, belonging to a particular race or gender, reaching a specific age, and not exercising enough. Among these elements are: The increase in the number of diabetic patients worldwide has piqued the interest of researchers in artificial intelligence and biomedical engineering who are working in

the field of diabetes research. This is because there are more diabetic people worldwide now than ever before (Ashwinkumar & Anandakumar 2012).
According to the results of an objectively conducted assessment, diabetes is ranked seventh on the list of illnesses that can cause death. This conclusion was reached using these findings. 51 million people worldwide have been diagnosed with diabetes, and type 2 diabetes is significantly more common than type 1 diabetes, with a difference of more than two to one. As of November 2007, 20.8 million persons in the United States, including both adults and children, had been diagnosed with diabetes, which affected around 7.0% of the country's population. According to the results of a global survey conducted in 2013 by Boehringer Ingelheim and Eli Lilly and Company, there are 382 million people suffering from Type-2 diabetes worldwide and 25.8 million people with Type-1 diabetes in the United States. Type 2 diabetes is the most common form of the disease and is thought to account for 90–95% of all cases of diabetes, making it a serious issue in both developed and developing nations. This is due to the fact that type 2 diabetes is the most prevalent variety of the illness.
By 2035, the number of people worldwide who currently have diabetes is expected to rise to 592 million, according to some forecasts made by the International Diabetes Federation (IDF). These predictions were first created in the year 2005. The World Diabetes Atlas estimates that there are currently 285 million people living with diabetes worldwide, with the possibility of this number rising to 438 million by the year 2030. According to poll findings, the number of people with type 2 diabetes will increase by the year 2030, prompting ominous predictions for the future. based on the conclusions of Kenney and Munce (2003). Additionally, it is a guarantee that 85% of the world's diabetic patients will reside in poor countries by the year 2030. This forecast is supported by the expectation that diabetes prevalence will increase. The assumption behind this estimate is that the number of persons with diabetes would likely rise. According to projections, there would be 79.4 million diabetics in India by the year 2030, up from the 31.7 million who had the disease in 2000. This forecast is based on recent data. (2004) Huy Nguyen et al. One of the most crucial elements of diabetes treatment that will ensure success is obtaining an accurate diagnosis as soon as possible (Mythili et al 2003).
Diabetes already affects more than 62 million individuals in the Republic of India, indicating that the disease is rapidly moving towards the position of a potential epidemic. The number of persons with diabetes is expected to more than double from 171 million in the year 2000 to 366 million in the year 2030, according to research done by Wild et al. The disease is expected to spread most quickly in India. India is expected to have up to 79.4 million diabetics by the year 2020, compared to China's 42.3 million and the United States' 30.3 million, both of which would witness considerable increases in the number of diabetes in their populations. By the year 2020, it is anticipated that there will be a considerable increase in the number of diabetes in India. India is currently facing an uncertain future due to the possibility that diabetes could become a significant burden in the future. [2].
Diabetes is a group of illnesses in which the body either produces insufficient amounts of insulin, fails to use the insulin that is produced properly, or exhibits a combination of both of these symptoms. Diabetes can also develop when the body uses insulin incorrectly. If this happened, the amount of glucose in the blood would rise because the body wouldn't be able to move sugar from the blood into the cells. One of the main sources of energy that our bodies need is a form of sugar called glucose that is found in our blood. A accumulation of sugar in the blood, a symptom of diabetes, can be caused by either insulin resistance or a deficiency in the synthesis of insulin. It will have a number of detrimental repercussions on one's health. [5].

The three main types of diabetes are as follows:

- **Diabetes Type 1**, The most common type of the condition is diabetes mellitus, often known as insulin-dependent diabetes. It is believed that autoimmune diseases contribute to the onset of type 1 diabetes. Diabetes type 1 develops when the beta cells that produce insulin in the pancreas are mistakenly attacked and killed by the immune system in our body, resulting in irreversible damage. Diabetes type 1 develops as a result of this. The most serious type of diabetes is type 1 diabetes mellitus. The most significant factor in the development of type 1 diabetes is the existence of a genetic predisposition [5].
- **Diabetes Type 2,** Diabetes mellitus is a disorder that develops either when the body is unable to produce enough insulin or when it is unable to utilise the insulin that is produced properly. Because of this, sugar does not function as an energy source and accumulates in the blood, which might have a negative impact on one's health. Diabetes type 2, the most prevalent form of the disease, is diagnosed in about 90% of people with the disease. Despite the fact that adults are more likely to develop type 2 diabetes than children, the condition regularly affects children.
- **Gestational** diabetes, Gestational diabetes is a short form of diabetes that affects pregnant women. Gestational diabetes is the medical term for the condition that can occur during pregnancy in people who have never been diagnosed with diabetes. It affects between two and four percent of all pregnancies and is linked to a higher risk of both the mother and the child developing diabetes.

The activity of identifying correlations, trends, and anomalies from massive datasets stored in databases and other types of data repositories is known as "data mining." Two techniques that can be used to accomplish this aim are pattern recognition and anomaly identification. Larger databases than those found in other types of facilities are frequently found in data warehouses and other forms of data storage facilities. Data mining's essential component, knowledge discovery, is made up of the procedures listed below. You may find this part of the article here. These procedures include data cleansing, integration, selection, transformation, mining, evaluation of patterns discovered in the data, and presentation of information derived from the data. "Data cleaning" is the process of removing undesirable components from a dataset, such as noise and missing information. This strategy also includes gathering data on the model that was applied to access the noise and accounting for any adjustments made. The "data integration" phase is the stage where the primary emphasis is on combining data from a variety of different sources. Another name for this stage is "data integration phase." In order to access the precise information that is needed, a subset of the data must be chosen. A procedure called as data transformation must first integrate a number of techniques for data preparation in order to make the data suitable for mining. The data will be prepared for mining after this is finished. Once this stage is complete, the data can be mined. The terms normalization and aggregate are just two examples of the many distinct methods that fall under this category.

"Knowledge discovery" is the process of autonomously generating information in a manner that humans can understand [3]. Computers are able to successfully execute this task. The numerous processes that make up the KDD process are depicted in a schematic in Figure 1. These actions are displayed one after the other.

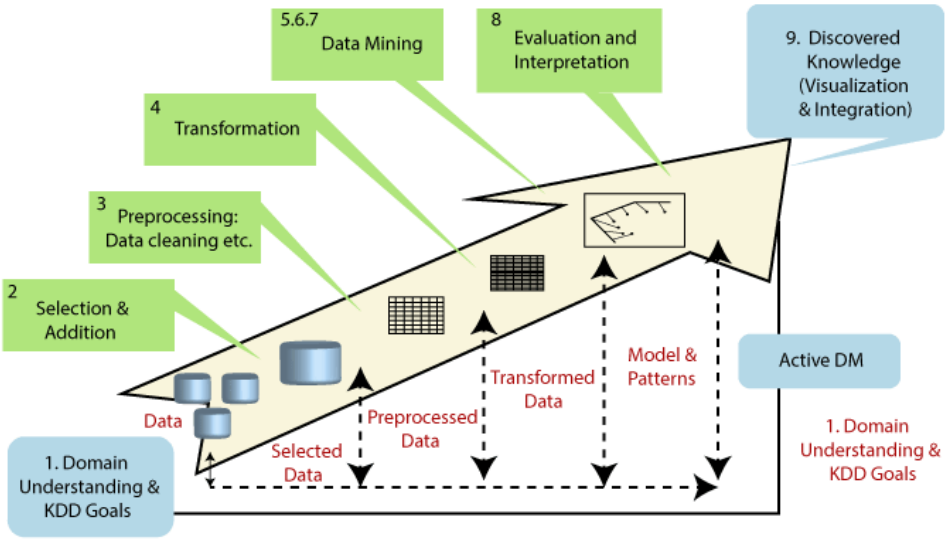

Figure 1: Steps of the KDD Process

The phrase "data mining" covers a broad variety of operations, including categorization, forecasting, time series analysis, association, grouping, and summarization of data. These are but a few of the activities included in this category. Each and every one of these jobs has some connection to the descriptive or predictive components of data mining. Each of the aforementioned actions can be performed by a data mining system as part of the data mining process, either separately or in various combinations.

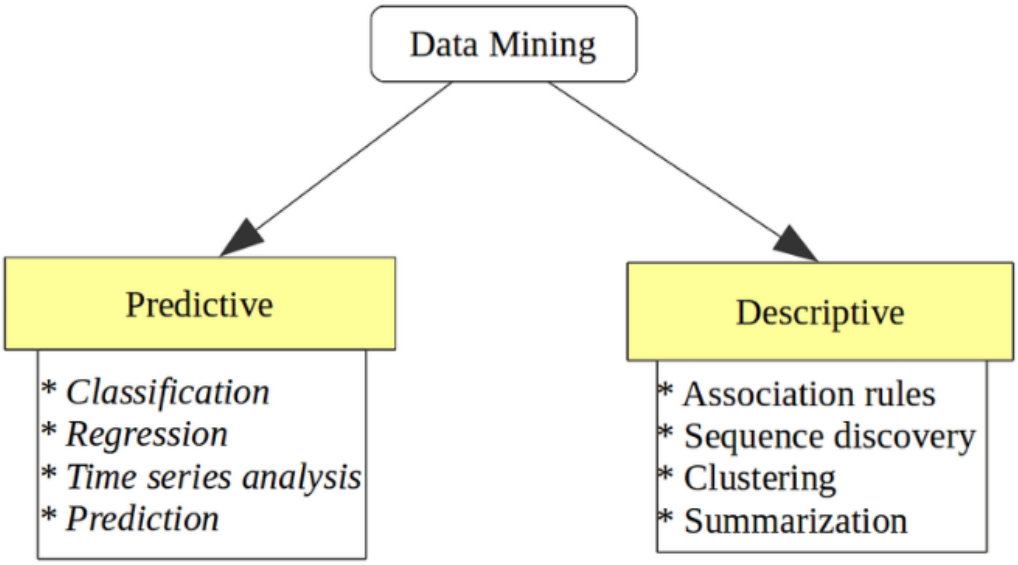

Figure 2: Data Mining Tasks

II. LITERATURE REVIEW

In the sphere of healthcare, data mining can be a very useful teaching tool, especially when it comes to the goal of finding instances of fraud and abuse. As a result, it is practical to utilise it to improve client relationship management decisions, allowing hospital staff to deliver better and more reasonably priced medical care. It enables medical professionals to determine which techniques provide the best quality of care, which is advantageous for therapy. Data mining techniques are frequently used in medical applications, such as data modelling for healthcare applications, executive information systems for healthcare, projecting treatment costs, and resource demand. By examining data from the patient's past as well as information from Public Health Informatics, e-governance frameworks in healthcare, and health insurance, forecasts regarding the behaviour of a patient in the future can be made (Dey & Rautaray 2014).

The naive Bayes algorithm stands out as one of the most intriguing and potentially profitable solutions for mining meaningful information from medical datasets. Despite the fact that this methodology has been used to analyse medical data, it is not without advantages or disadvantages. It is a statistically straightforward classifier that operates under the assumption that attributes are subject to change independently of other factors. This method's ability to maintain a high rate of classification accuracy even when used with very big datasets is another crucial feature. When other features are taken into account, its accuracy increases, which makes it more suitable for use with medical data. On the other hand, it struggles to determine the level of independence between two attributes when it is difficult to do so. It suffers significantly from harm as a direct result of noise. This method's performance and the decision tree method's performance are comparable.

The decision tree algorithm is the best tool to use in circumstances where a medical professional wants to express their decision-making as rules. The categorization of the rules in this algorithm is one of its most significant features (Kuo et al 2001). When a physician is seeking to quantify a patient's symptoms, regression analysis of the gathered information can be used to generate a prediction about a particular value. Even when there is a very small difference between two groups, it still performs wonderfully. Among the criteria that the decision tree technique can easily manage are accuracy, specificity, sensitivity, positive predictive value, and negative predictive value.

To get the lowest possible error ratio, the decision tree classifier was employed. methods such as feature selection, cross validation, error reduction pruning, and increasing model complexity are being researched and investigated. Dimensionality reduction, also known as the process of compressing the attribute space of a feature collection, can be achieved in part through the use of feature selection. This is accomplished by eliminating data attributes that are judged irrelevant and useless. The predictive value can be evaluated more accurately thanks to cross-validation, which has also shown an improvement in classification accuracy despite an increase in model complexity. This is true even when a more complex model was used for cross-validation. Cross-validation is an estimating technique that is more reliable. Reduced error pruning was used as an approach to successfully address the overfitting issue that had been damaging the decision tree. The enhancement over the prior system includes both an increase in accuracy and a decrease in the error rate. In other words, both parties benefited from the improvement. The decision tree is built in a significantly shorter amount of time [4].

The Support Vector Machine (SVM) method is a crucial step in the categorising process and can be used with medical datasets. SVM was developed to avoid overfitting training data, and with the right kernel choice, such the Gaussian kernel, the algorithms can concentrate more on how similar classes are to one another than on other levels of similarity.

The support vectors of the training sample that is most similar to the category being categorised are compared to the values of the SVM's ratios when it is used to classify a new category. This guarantees that the new category can be classified correctly. The degree to which this class is comparable to the other class will then determine how further this class is categorised. The relevance of SVM is due to the fact that it can function as a universal approximation for a large range of kernels in addition to the absence of any local minima. It is crucial for a number of reasons, including this. The SVM's primary drawback, however, is that it makes it difficult to identify the characteristics or data sets that have the most influence on a forecast. One of the SVM's most serious shortcomings is this.

The K Nearest Neighbour, or KNN, Algorithm is well suited for use on medical datasets and makes effective use of those databases thanks to a fascinating combination of features. It is suitable for usage on other kinds of databases due to these characteristics as well. Because it is so straightforward to use, the KNN approach is the one that is most frequently used for pattern recognition. Because it is the most dependable, this is the situation. Despite this, there are some situations when it is unable to produce satisfactory results. However, by fine-tuning the parameter k in the KNN algorithm, the outcomes might be enhanced in a number of contexts. According to Moreno et al. (2003), this parameter, which denotes the number of neighbours, determines how similar a particular value is to its neighbours. Voting was used to conduct a study into kNN, and the results of the investigation were evaluated on the prediction of cardiovascular disease. The results show that kNN implementation can potentially achieve a higher level of accuracy in heart disease prediction than neural networks. Despite the fact that neural networks are currently the industry standard, this is the case. The dataset may now be identified more precisely in terms of heart disease thanks to the usage of KNN in combination with a genetic approach.

All-over body skin temperature measurements and asymmetric dimethylarginine (ADMA) blood levels analysis were performed on patients with type 2 diabetes mellitus. These two elements were taken into account during the diagnosis procedure. The population was split into two groups: those with no issues and those with complications. One group of individuals was regarded as typical. A thermography camera was used to take thermograms of every part of the subject's body without having to touch them directly. Thyroid hormones and other blood constituents were measured biochemically together with a number of other blood parameters. Additionally, a score reflecting the propensity to acquire diabetes was established. The areas of the sole with the lowest skin temperature readings in healthy individuals were found to be the back of the foot, while the areas with the highest readings were found to be the ear. Through the process of observation, this was discovered. Diabetes patients showed lower mean skin temperatures overall than non-diabetic patients, and the nose and tibia areas saw significantly lower skin temperatures [3]. The entire body experienced this in the same way.

The results of numerous studies indicate that whether or not a patient is seen by a variety of doctors, or even by the same doctor at various times, the diagnosis of that patient can change dramatically. This holds true even if the patient receives several examinations from the same doctor. The use of computerized medical diagnostics enables doctors to diagnose their patients' illnesses more quickly and accurately. In order to make it easier to spot patterns in the data it gathers, this strategy makes use of the Naive Bayesian theorem. The naive Bayesian algorithm estimates the likelihood of a wide range of disorders that can affect the skin in addition to calculating the proportion of patients that have each dermatological issue.

III. DATA MINING STRATEGIES

The Expectation Maximization (EM) Algorithm

This electromagnetic technique can be broken down into two distinct components. The first stage is to decide what to expect, and the second is to maximize that expectation by repeatedly going through the process. After choosing a model as the first stage in the expectation, which also includes choosing a model, comes the process of estimating any missing labels. You will choose labels and then map pertinent models to those labels throughout the maximization phase. Your outcomes will be maximized by doing this. The purpose of the process is to maximize the expected log-likelihood of the data, therefore this is done. Within the operational order, three separate phases can be identified [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 + μ) * h] and y= [(μ/ 0.5 + μ) * h] with conditions of x / y = (0.5 / μ) 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: For the following cycle, repeat steps 1 and 2. Cross-validation of the mean and standard deviation for a total of seven different features was used to establish the clusters. Each student in the group was then given a test to see whether they had any positive or negative conditions related to diabetes. Binary answer variables are alternatively represented by the numbers 1 and 0 during the data analysis process. If the diabetes test yields a 1, it indicates that diabetes is present (positive), and if it yields a 0, it indicates that diabetes is absent (negative). The EM technique, however, is not very accurate when applied to data sets with bigger dimensions because of the numerical imprecision [2].

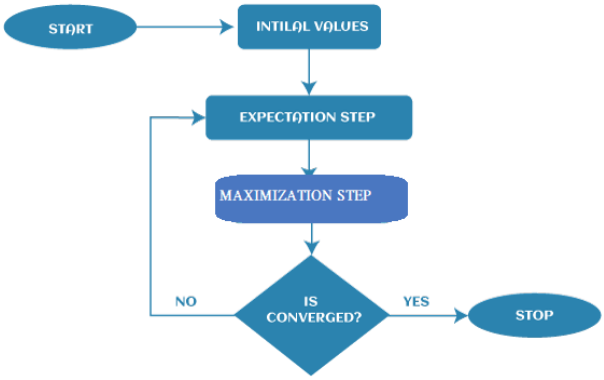

Figure 3: EM Algorithm Steps

K Nearest Neighbour Algorithm

Due to its relative simplicity and high degree of accuracy, the K Nearest Neighbour (KNN) method has been used in a wide range of applications for the goal of data analysis. The applications that fall under this heading include machine learning, pattern recognition, data mining, and database management. It is one of the top 10 algorithms that can be utilised in the field of data mining, according to the most recent rankings (Wu et al 2008). The categorization method known as the KNN algorithm is referred to as "lazy learning." The machine learning algorithm can be employed in this most basic version. This system makes it feasible to predict when any kind of label will appear [5].

Using the KNN classification, samples are arranged based on how similar they are to one another. It serves as an example of a learning method known as "lazy learning," which approximates the function locally and defers execution until classification. This type of learning methodology approximates the function. Classification and clustering applications make best use of K-Nearest Neighbours. Numerous researchers have discovered that the KNN algorithm generates outcomes that meet or surpass their expectations after testing it on a wide range of datasets. It is really challenging to understand the Pima Indian diabetes dataset since there are so many elements that are missing. The Euclidean Distance matrix's missing values are determined via the KNN method by looking at the columns of data that are immediately surrounding the matrix. If the equal value from the closest neighbouring column is likewise missing, the value from the subsequent immediate neighbouring column is used in its place. In contrast to other ways, this method is not only straightforward but also provides a sizable competitive edge. The fact that KNN does not use probabilistic semantics, which would enable the use of posterior prediction probabilities, is one of the disadvantages of KNN, which might be considered as a negative.

A large number of KNN's writers have contributed to its most recent upgrade in an effort to make KNN more useful. The class-wise KNN (C-KNN) technique has been used, and the Pima Indian diabetes dataset has been used to validate its performance. At this point, a class label is applied to the testing data using the shortest class-wise distance. The C-KNN algorithm has reached an accuracy level of 78.16%. To make it easier to classify the diabetes cases found in the Pima Indian database, the K means and KNN classification methods have been integrated into a single model called as the amalgam KNN model. By eliminating the noise in this case, the quality of the data is enhanced while simultaneously increasing the quantity of work that can be completed in the same period of time. The cases that were incorrectly classified are excluded using the K-means algorithm, and the classification is completed using the KNN algorithm.

The KNN algorithm's K value will be determined by the data. A greater value for k can help reduce the noise in categorization by assisting in the categorization process. The cross-validation approach can be used to choose an appropriate value for k. We were able to attain a classification accuracy of 97.4% by first figuring out the k value and then performing ten-fold cross validation [6,] all of which were necessary steps. A graphic representation of the basic idea underpinning the KNN algorithm is shown in Figure 4.

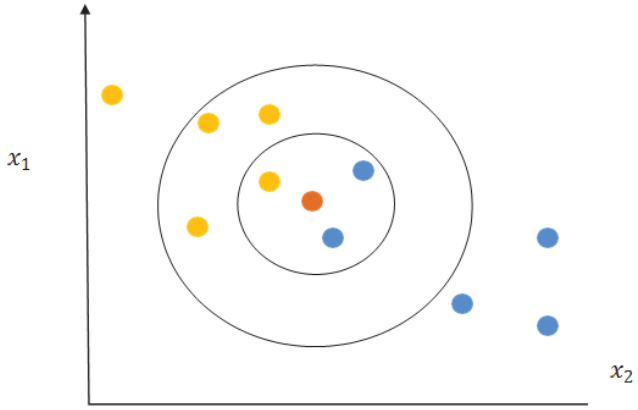

Figure 4: K nearest neighbor algorithm

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

Algorithms that can perform effectively on unlabeled samples even in the absence of direct supervision are known as unsupervised algorithms. This implies that even if the input can be determined, the output cannot be predicted. The K means algorithm is one of a variety of unsupervised learning algorithms that include many techniques. In order to function properly, they need n objects in the data collection, which are then divided into k clusters, as well as an input parameter, which is the number of clusters. Based on a random selection, the algorithm selects one of the k options. An item is given a specific placement within one of the clusters to which it belongs based on how close it is to the linked cluster to which it belongs. The next step is to identify the regions that are in close proximity to one another. It is advised to use the Euclidean distance while attempting to determine the position of the object that is most central to it. The new cluster centres are identified by averaging the items contained within each of the k clusters after the items have been grouped into clusters. Up until all of the clusters have been used, this process is repeated. It is done by using this method up until there is no longer any fluctuation in the k cluster centres. The sum of squared error (SSE) is the objective function that the K-means algorithm seeks to reduce in order to successfully carry out its goal [8]. The acronym SSE stands for the following:

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

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 *Ci*th cluster, and mi is the mean of the *Ci*th 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.

Random Forest Algorithm

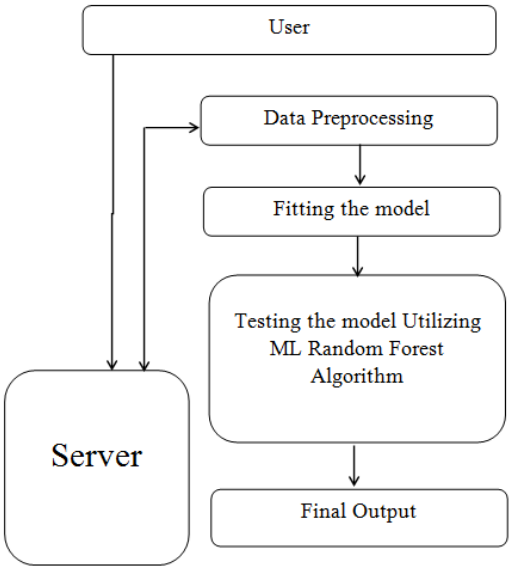

Figure 5: Flow graph of Random Forest Algorithm

To get things going, let's look at the supervised classification method known as Random Forest. The goal of the game is to create a random forest using every means possible, as indicated by the game's title. There are several approaches to accomplish this goal. A forest's ability to make discoveries is correlated with the amount of trees present; the more trees present, the more precise the results will be. One thing to keep in mind, though, is that building the forest is not the same as building the choice using information gain or using the gain index approach. So have that in mind.
In order to learn more about decision trees and get a full understanding of what they are all about, the author provides readers with access to four websites that may be helpful to anyone dealing with them for the first time. A decision tree is a helpful tool for aiding in decision-making. To illustrate the numerous potential possibilities, a graph in the form of a tree is used. The decision tree will automatically generate some sort of rule set for you to follow if you provide it with a training dataset that includes targets and features. Making precise forecasts is possible by following these rules. As an example to support his argument, the author gives the following scenario: You're attempting to ascertain if your daughter will enjoy watching an animated movie. If this is the case, you should compile a list of previous animated films she like and incorporate specific elements from those films as inputs for your forecast. After that, you are free to carry on with the decision tree technique for generating the rules. Once you've entered the movie's qualities and seen the results, you'll be able to tell whether or not your daughter will enjoy it. Information gain and the Gini index computations are used throughout the entire process of identifying these nodes and creating the regulations.

Random Forest was initially created by Leo Bremen. The Random Forest rule, which 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—could be an example of a supervised classification rule [11]. The supervised classification counterpart of the Random Forest rule is [11], while the pseudo code for Random Forest is rf.

- 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

The study of diabetes mellitus has been conducted on the Pima Indians of the Gila River Indian Community in Central Arizona since 1965. The study is repeated every two years. These tests, which also include an oral glucose tolerance test and various assessments of complications of diabetes and other medical conditions, provide the majority of information regarding the prevalence, incidence, risk factors, and pathogenesis of diabetes in the Pima Indian population (Leslie et al 2004). There are many research discoveries that appear to be relevant to the Pima people. Metabolic characteristics of Pima Indians with type 2 diabetes include obesity, insulin resistance, insulin secretion, and a higher rate of endogenous glucose production, which are the characteristics that distinguish diabetes [10].

The Pima Indian diabetes dataset contains information on 768 people's various measurements and a forecast of whether diabetes will eventually strike them. The patients at this hospital were all Pima Indians and had reached the age of 21. These eight characteristics determine whether the tested data belongs to the group of those with diabetes (tested positively) or those without diabetes (tested negatively). The dataset consists of 268 patients with diabetes (class = 1) and 500 patients without diabetes (class = 0).

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

The aim of this data set was to identify diabetic Pima Indians. Based on personal details like age, the number of pregnancies, and the results of medical tests like blood pressure, body mass index, glucose tolerance test results, etc., try to ascertain if a Pima Indian person had diabetes or not. The attributes are detailed 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

The usefulness of the suggested technique is assessed in this section of the article. To evaluate the viability of the suggested Protocol, simulated implementations of the proposed algorithms are employed. For this, Tensorflow and a number of other Python libraries can be used; the Python programming language is the foundation of our work.

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.

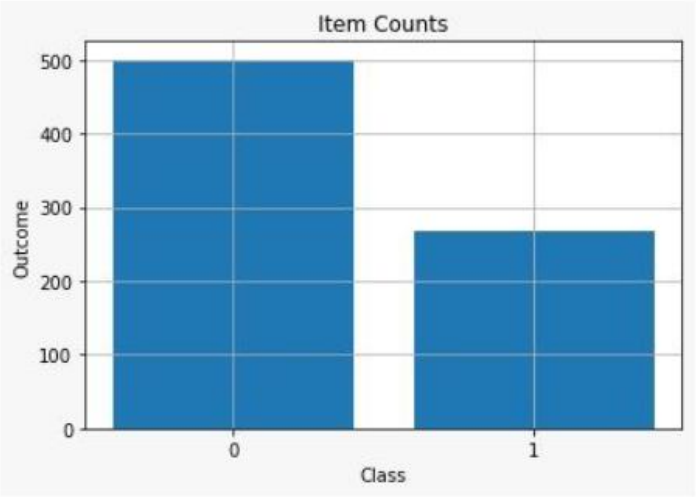

Figure 6: Item Counts

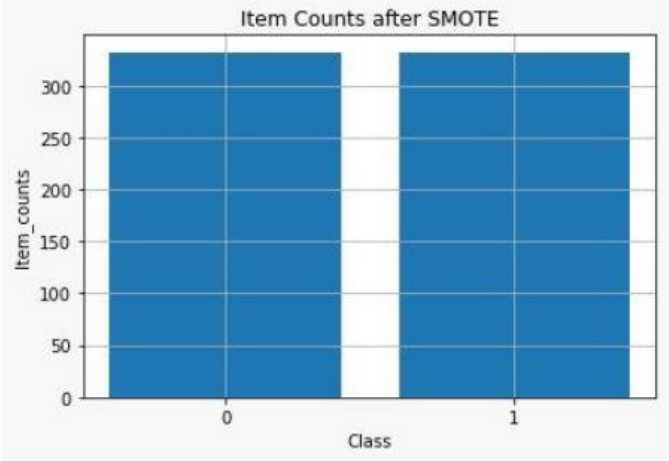

Figure 7: Item Counts

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.

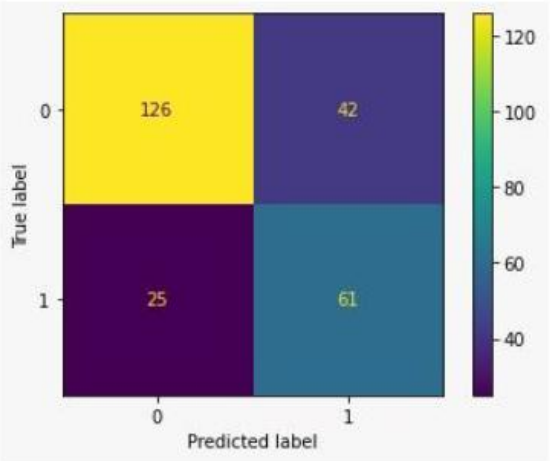

Figure 8: Confusion matrix

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.

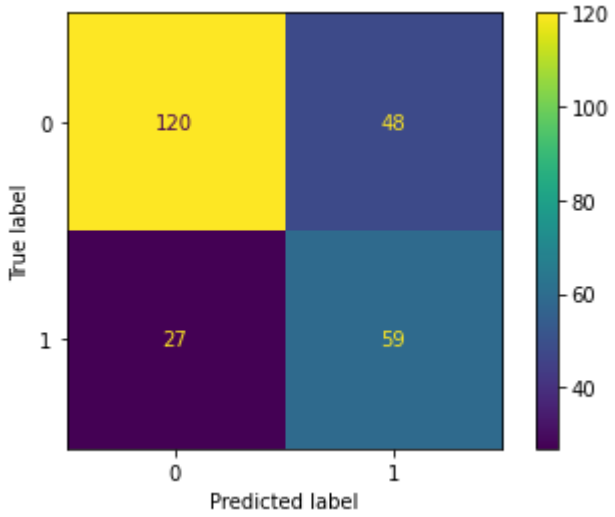

Figure 9: Confusion matrix

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.

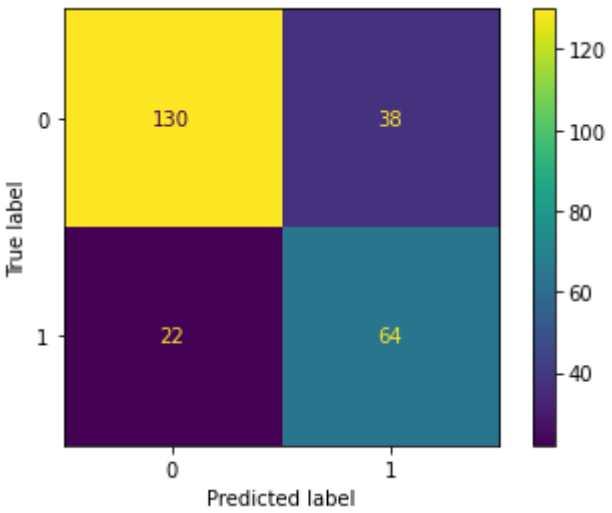

Figure 10: Confusion matrix

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.

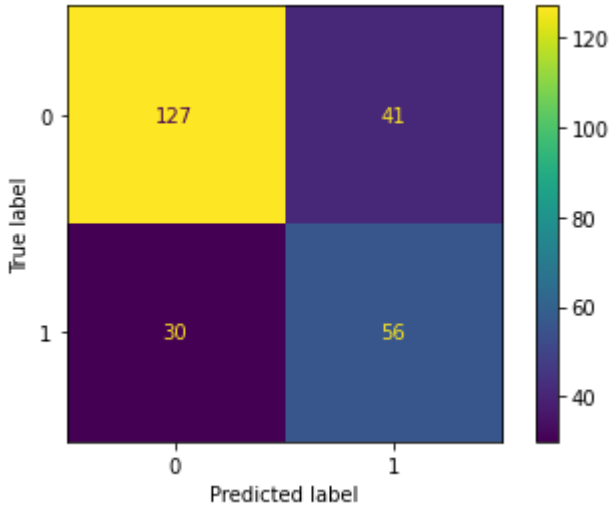

Figure 11: Confusion matrix

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 them, the number of machine intelligence algorithms. On patient medical records, data mining can be used. In the area of healthcare, a substantial amount of data has been acquired and organized. The diabetic dataset is the one that has undergone the least amount of analysis. Data mining approaches are used to successfully address and resolve the topic of diabetes prediction throughout the entire thesis. Three distinct predictive models for diabetes have been shown to be beneficial, and each of these models is based on the same well-known classification technique, known as the Random Forest algorithm. It is abundantly clear from the tests carried out on the data set containing Pima Indians with diabetes using the Python programme that the performance of the suggested classification methods greatly increased.

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] R. Kaundal, A. S. Kapoor, and G. P. Raghava, “Machine learning techniques in disease forecasting: a case study on rice blast prediction,” BMC Bioinformatics, vol. 7, no. 1, p. 485, Nov 2006.

[4] S. W. Franklin and S. E. Rajan, “Diagnosis of diabetic retinopathy by employing image processing technique to detect exudates in retinal images,” IET Image Processing, vol. 8, pp. 601–609, October 2014.

[5] L. Zhou, Y. Zhao, J. Yang, Q. Yu, and X. Xu, “Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images,” IET Image Processing, vol. 12, pp. 563–571, April 2018.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, p. 10971105, 2012.

[7] W. Sandham, E. Lehmann, D. Hamilton, and M. Sandilands, “Simulating and predicting blood glucose levels for improved diabetes healthcare,” IET Conference Proceedings, pp. 121–121(1).

[8] L. C. Shofwatul Uyun, “Feature selection mammogram based on breast cancer mining,” IJECE, vol. 8, pp. 60 – 69, Feb 2018.

[9] N. K. A. Hussein Attya Lafta, Zainab Falah Hasan, “The classification of medical datasets using back propagation neural network powered by genetic-based features elector,” IJECE, vol. 9, Apr 2019.

[10] V. D. Komal Kumar N, R. Lakshmi Tulasi, “An ensemble multi-model technique for predicting chronic kidney disease,” IJECE, vol. 9, Apr 2019.

[11] F. J. Rini Widyaningrum, Sri Lestari, “Image analysisof periapical radiographfor bone mineral density prediction,” IJECE, vol. 8, pp. 2083–2090, Aug 2018..

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.