DIABETES MELLITUS DETECTION AND DIAGNOSTICS USING DATA MINING TECHNIQUES: A Review

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

Keywords— Data mining, Diabetes Mellitus, Random Forest with Feature Selection, Machine Learning Algorithm, etc.

I. INTRODUCTION

Diabetes is a condition that can ultimately result in death. Diabetes can be caused by a combination of causes, including but not limited to being overweight, having high blood glucose levels, leading a sedentary lifestyle, and not getting enough exercise. Because of this, the hormone insulin is affected, and as a result, the crabs' metabolism becomes erratic, and there is an increase in the amount of sugar in their blood. Diabetes is a condition that develops when the body is unable to produce an adequate amount of the hormone insulin. Diabetes affects around 422 million people around the world, as reported by the World Health Organization (WHO). This percentage is abnormally high in countries where the standard of living is poor or nonexistent. It's also likely that by the year 2030, this figure will have increased to 490 billion people all over the world. On the other hand, the prevalence of diabetes is observable in a variety of distinct nations, including Canada, China, and India, amongst others. Diabetes is a major factor that contributes to the mortality rates of people all over the world. Patients who are diagnosed with diseases such as diabetes at an earlier stage are more likely to undergo treatment, which could ultimately save their lives. In order to accomplish this objective, the purpose of this study is to evaluate the feasibility of diabetes prediction. To do so, a number of variables that are connected with the diabetes condition will be taken into consideration. In order to accomplish this goal, we make use of the Diabetes Dataset and implement a wide range of machine learning classification and ensemble methods. The conclusion is that an accurate diagnosis of diabetes can be made.

Diabetes can be managed using a variety of approaches, including insulin injections and changes in diet. As quickly as is humanly possible, this condition should be identified for what it is, and then the appropriate treatment should be administered. Testing that is both chemical and physical forms the basis for the majority of the techniques that are used for categorization, identification, and diagnosis. Using the conclusion that may be derived from these results, a particular disease can be expected. It's possible that our predictions will be off. This is because the numerous testing parameters all have varied degrees of uncertainty, which leads to this result [2]. These unanswered questions make it more difficult to find a treatment for the condition and contribute to erroneous forecasts. The computing facility has seen a significant amount of progress during the past few years. The advancements that have been made in information technology have made it possible to more accurately classify data, forecast outcomes, and diagnose diseases in many different situations. The most important advantage that comes from the use of information technology is that it enables medical facilities to continually maintain and monitor massive data storages containing records of former patients for several references [3]. With the use of these medical data, the doctors are able to explore the various patterns contained within the data set. The patterns that are found in data sets may be of assistance in the process of disease classification, as well as prognosis and diagnosis [4].

The training method that may be applied to either computers or machines is referred to as "machine learning," and the phrase is used interchangeably. Constructing a wide variety of classification and ensemble models with the data that has been acquired is one of the many ways in which a wide variety of machine learning approaches create effective outcomes for the accumulation of knowledge. With the use of such gathered data, the risk of developing diabetes can be approximated. The field of machine learning encompasses a wide variety of techniques, each of which is able to make predictions; despite this, it can be challenging to decide which technique is the most effective. Because of this, in order for us to accomplish this aim, we apply well-known classification and ensemble algorithms to the dataset in order to create a prediction. Consequently, in order to accomplish this goal:

The process by which computers figure out how they can execute tasks without being formally educated to do so is referred to as "machine learning." [5] The term "machine learning" refers to this process. It refers to the process whereby computers learn specific tasks based on the data that is supplied to them in order for them to do those tasks. It is possible to write algorithms that instruct a computer how to carry out each step that is necessary to solve the issue at hand; in this case, the computer does not need to engage in any form of learning; however, this option is only available for tasks that are relatively straightforward and are delegated to computers. When it comes to actions that are more sophisticated, it can be difficult for a human to create the necessary algorithm manually. It is possible that assisting the machine in the development of its own algorithms will prove to be more productive in practice than having human programmers manually specify each step that is required. This hypothesis is based on the hypothesis that assisting the machine in the development of its own algorithms will increase productivity.

The pursuit of artificial intelligence served as the driving force behind the establishment of machine learning as a distinct academic discipline in the first place. During the early phases of artificial intelligence's growth as an academic discipline, a number of academics demonstrated an interest in instructing computers to learn from data. This interest continued throughout the field's later stages of development. They attempted to solve the problem by employing a number of symbolic approaches in addition to what was at the time known as a "neural network." The "neural network" was essentially made up of the perceptron in addition to a few other models that were shown to be, with further investigation, simply reimagining’s of the generalized linear model of statistics.

In addition to that, we made use of probabilistic reasoning, in particular for computerized medical diagnosis.

Diabetes is a disease that is responsible for the death of a considerable number of people all over the world. The proliferation of different technologies is directly proportional to the enhancement of the quality of life for people. As a result, there is no valid excuse for not taking use of the technology that exist to make leading a healthy lifestyle more

convenient. Deep learning and other types of machine learning algorithms are used in a wide variety of prediction skills. These capabilities can range from stock market forecasting to medical diagnosis. The question of how mankind might gain from the application of these technologies has been addressed to us here [6]. These technologies are frequently used by the most successful organizations in the world to boost their profits and sales. In order to create a forecast about something that is so specialized that only experts are capable of doing it, we are going to put the various algorithms that we have developed and learned throughout the course of our history to the test. In order for the machine to learn the complexities of the many different aspects of the biomechanics of human beings and to accurately foresee the difficult challenges faced by live beings, it needs to be trained with the minds of medical professionals. This will allow the machine to learn how to accurately predict the difficult challenges faced by live beings. Implementing these algorithms is necessary in order to provide accurate predictions regarding complex diseases by making use of a wide variety of internal and external characteristics that are derived from a reliable dataset [7]. This is accomplished through the utilization of a large number of variables that are measured internally as well as externally.

Diabetes is a disease that reduces the amount of insulin in the body, which prevents glucose from entering the cells of the body's bloodstream. Diabetes is a disorder that affects the human body and can be lifelong and chronic. This causes the sugar level in the body to grow, which can lead to a range of health problems and even death. Some of these health problems and diseases include stroke, heart disease, blindness, and kidney failure. Patients who are afflicted with diabetes frequently present with the following symptoms.

- An increase of thirst experienced
- Nausea and vomiting
- Infections with a sluggish recovery time
- A greater degree of hunger
- Haze in the eyes
- Reduced body weight
- Frequent urinating

The following set of medical tests and procedures are typically performed in order to establish a diabetes mellitus diagnosis.

- Urine test
- Fasting blood glucose level
- Random blood glucose level
- Oral glucose tolerance test
- Glycosylated hemoglobin(HbAlc)

II. Literature Review

In the following part of this article, we are going to have a look at some of the earlier research that pertains to this subject in greater detail.

The research work of Jitranjan Sahool et al. [3] predicting diabetes using Machine Learning Classification Algorithms and this research work shows that, Logistic regression was found to outperform all of the machine learning algorithm showing the maximum accuracy of 72.17% in comparison to other algorithm. [Citation needed] [Citation needed] [Citation needed] Logistic regression was found to outperform all of the machine learning algorithm showing the maximum accuracy of 72.17% in comparison to

Nonso et al. [4] presented a novel approach to the problem of diabetes prognosis, which consists of the following steps: in the supervised learning approach, five widely employed classifiers are used for the ensembles, and a meta-classifier is used to aggregate the outputs of the individual classifiers. The results that are offered are compared with the findings of other research that has been published that was conducted using the same dataset. This was done so in order to validate the findings that are presented. It has been established that the proposed technique can result in a higher level of accuracy

when it is used to the prediction of the start of diabetes.

Tejas et al. reported Diabetes Prediction in an investigation that had been carried out [2]. It is possible to make an accurate diagnosis of diabetes by employing a number of different supervised machine learning approaches such as support vector machine (SVM), logistic regression, and artificial neural network. This is the goal of using machine learning techniques (ANN). Their work project presents a method that is not only effective but also efficient for identifying diabetes disease at an earlier stage.

In a different study that Deeraj and his colleagues conducted [1], data mining was mentioned as a potential new approach that could be used to predict diabetic disease. An Intelligent Diabetes Disease Prediction System is currently in the process of being created. This system will provide an analysis of diabetes disease by utilizing a database that is comprised of people who have diabetes. They suggest using algorithms in this system such as Bayesian and KNN (K-Nearest Neighbor) to apply on a database of diabetes patients and evaluate them by taking into consideration a variety of diabetes-related characteristics in order to make a prediction regarding the diabetes disease. This would be done for the purpose of making a diagnosis of the diabetes condition.

Comparisons were done between the several machine learning approaches (support vector machine, logistic regression, decision tree, K-nearest neighbour, and random forest) that were utilized in the diabetes prediction study carried out by Mitushi Soni et al. [5].

Our team made use of a wide variety of machine learning algorithms. Which of These Three Models Is the Most Accurate—Decision Tree, Logistic Regression, or Random Forest?

Further investigation has been put into developing techniques for estimating the age at which a person would be diagnosed with type 1 diabetes for the first time. Exposure to respiratory infections in early childhood has been shown to be associated with an increased risk of autoantibody seroconversion in children who come from families with a history of type 1 diabetes [8]. [Note: This was found out by the method of continuously monitoring the subject's islet auto antibodies over the first three years of their existence [8]. Longitudinal autoantibody measurements have also been applied as a risk predictor in families that have a first-degree relative with type 1 diabetes [13], in general populations [38], and in individuals who have been identified as being at risk [11]. In addition, genetic factors as well as genetic risk scores were applied in order to evaluate whether or not islet autoantibodies were present in children who possessed high-risk HLA genotypes [9]. Six months before a diagnosis, the levels of post-challenge C-peptide start to drop rapidly after a challenge [6]. An investigation of the variations in metabolic processes supports this conclusion. For children who are at high risk, a composite risk score model was devised. This model took into account clinical, genetic, and immunological factors (who were tracked from birth until 9 years of age). When compared to autoantibodies by themselves, this model displayed a prediction of T1D that was significantly more accurate [10]. However, beyond the scope of the study that was mentioned above, there is a lack of application of machine learning methodologies to the process of building models of the age at which type 1 diabetes first appears in a person's life. This is despite the fact that a great number of studies have utilized a wide array of machine learning strategies for type 2 diabetes [10]. As a result, the work that is being suggested is distinct from the research that has been done in the past since it replicates the age at which children experience their first symptoms of type 1 diabetes (T1D). In order to accomplish this, it employs statistical and machine learning techniques to ascertain the risk factors and construct a prediction model.

III. DATASET CLASSIFICATION

The Pima Indian Diabetic Set, which can be located in the Repository of Machine Learning datasets at the University of California, Irvine (UCI), served as the foundation for the data set that was selected for categorization and experimental simulation. The Pima Indian Diabetic Set can be found in the Repository of Machine Learning datasets at UCI. Patients that are being considered for this study are natives of the Pima Indian tribe who currently make their homes in the state of Arizona, in the United States. More than half of the Pima Indian community suffers from diabetes, and the ailment is nearly exclusively brought on by the people's excessive levels of body fat. Obesity has been decisively proven as the key contributor to the development of diabetes in a number of research that have been carried out on these populations. The data collection in question is mostly made up of 9 characteristics, and there are a total of 768 instances [4]. A listing of these eight traits, together with the symbols that correlate to them, can be found in Table 1.

A medical dataset was obtained from the machine learning data repository at the University of California, Irvine for the purpose of evaluating the performance of three distinct approaches to diagnosing diabetes mellitus using PIDD. These approaches were evaluated with regard to how well they were able to produce accurate results. The Pima Indians Diabetes dataset can be found at the following URL: archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes. The data collection contains a total of nine different properties.

Table 1: Dataset description

Type	Classification	Origin	Laboratory
Features	8	(Real / Integer / Nominal)	(8 / 0 / 0)

Instances	768	Classes	2

Every algorithm necessitates that the information be provided in a particular arrangement. The unprocessed data must first be transformed into a format that can be read by a computer; this stage of the process is referred to as "pre-processing." During preprocessing, the tasks that need to be carried out include converting the attributes in the database to a single scale and replacing any and all missing values in the data. The raw data can be saved in a variety of formats, including text files, Excel spreadsheets, and other database file types. The vast majority of the time, raw data does not conform to any particular format. It is possible that the processing of the data will take less time if the data are formatted into a manner that the algorithms can understand. In most circumstances, each row in the table represents a single case, and the attributes of that case are indicated in the columns. Some databases store their information in a format known as Comma Separated Values (CSV for short). That is, each attribute is denoted by a comma, and the presence of two commas in a row indicates that there is an absence of data attribute. When there is a missing attribute, it is possible to find a question mark instead of a blank space in some cases.

The distribution of attribute values in relation to the class attribute labelled "0 or 1" is shown in Figure 1. The incidence of diabetes is shown by the number of times the colour blue appears. It is clear from looking at the figure that the vast majority of diabetic patients who are pregnant have values between 0 and 1.5, have plasma in the range of 99.5 to 103.5, have pressure in the range of 65 to 71, have skin fold thickness between 0 and 7, have insulin levels between 0 and 50, have a BMI between 27 and 30, have pedigree function between 0.25 and 0.50, and are between the ages of 21 and 25.

Table 2: Dataset Description

S. No.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness
5	Insulin
6	BMI(Body Mass Index)
7	Diabetes Pedigree Function
8	Age

IV. DATA PREPROCESSING

The phase that is believed to be one of the most important is the one in which the data is prepared. The vast majority of the data that pertains to healthcare lacks values and also contains other impurities, both of which might diminish the usefulness of the data. The purpose of performing data preprocessing is to improve the overall quality and usefulness of the findings obtained after the mining procedure has been completed.
In order to make the most of the opportunities presented by the dataset by applying Machine Learning Techniques, it is essential to carry out this approach in order to provide findings that can be trusted and to generate predictions that are spot on. As a direct consequence of this, various processes are carried out in order to convert the data into a data set that is condensed and error-free. You will first finish this strategy before starting the process of iterative analysis, which is a sequential process. The collection of procedures that are carried out is what is meant by the phrase "Data Preprocessing" [11].
Included in it are

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

It is required to undertake data preprocessing since there is access to data from the real world that has not been prepared. Erroneous information (missing data), which can arise for a variety of different reasons, makes up the vast majority of the data that comes from the real world. There are a wide variety of reasons for this, some of which include the fact that data is not continuously collected, an error in the process of entering data, technological challenges with biometrics, and many

more.
The existence of noisy data, which can also include erroneous data and outliers - The reasons for the existence of noisy data could be due to a technological problem with the device that collects the data, a human error made when inputting the data, or a variety of other variables.
Data Inconsistent: The occurrence of inconsistencies can be linked to a number of issues, some of which include the existence of duplication within the data, the entry of the data by humans, the inclusion of mistakes in codes or names, a violation of the data limitations, and a great deal of other things.
We will need to perform preprocessing in two stages if we are going to adequately prepare our dataset on diabetes.
Elimination of Missing Values: Eliminate all of the instances in which there is a value of zero (0). Having no value at all is not a realistic option for anyone. As a direct consequence of this, the episode in question is no longer pertinent. A feature subset is created by a process known as feature subset selection, which involves the elimination of features or instances that are deemed to be unimportant. This helps to reduce the dimensionality of the data, which, in turn, makes it easier and more efficient to get things done in a timely manner.
Data Splitting: In both the training and the testing phases of the model, the data are partitioned and normalized after being cleaned. This ensures that the model is accurate. After the data have been partitioned, the algorithm is educated with the help of the training data set, and the test data set is kept in a separate location. This process will build the training model based on the logic and approaches that are applied, as well as the values of the features that are included in the training data. The major goal of normalization is to realize the objective of bringing all of the traits to the same level, and this purpose drives the process.

V. FORMATION OF TRAINING AND TESTING DATA

Generally speaking, the data that we deal with is split up into two distinct groups: the training data, and the test data. In order for the model to be able to generalize its findings to more data in the future, a known output is included in the training set, and the model is trained using this data. Because we have the test dataset, which is also referred to as the test subset, we are able to determine how accurate our model's prediction is by using this particular subset. The size of the training set is going to be raised, while the size of the test set is going to be decreased.
On the training set, both training and testing will be carried out with the assistance of the test set. The objective of the dataset is to produce a diagnostic prediction as to whether or not a patient suffers from diabetes based on certain diagnostic measurements that are contained within the dataset. This prediction can be made on the basis of the contents of the dataset.

VI. ALGORITHMS

MACHINE LEARNING

Machine learning is a method of statistical learning in which each instance in a dataset is described by a collection of features or qualities. This method was developed by Jerome Bruner and has become increasingly popular in recent years. The IBM Watson Research Center is responsible for the development of this method. In contrast, the phrase "Deep Learning" refers to a statistical learning process that extracts traits or attributes from raw data. This technique is sometimes referred to as "machine learning." This strategy is sometimes referred to as "guided learning."
Deep learning, which makes use of neural networks that incorporate many hidden layers, massive amounts of data, and significant processing resources, is what makes this possible and is how it is performed. However, when applying the Deep Learning approach, the programme will automatically build representations of the data. Although the words may appear to be interchangeable to some extent, this is not the case. Data representations in machine learning methods, on the other hand, are hard-coded to take the form of a set of features. This requires additional processes, such as the selection and extraction of features, which are not included in traditional learning algorithms (such as PCA).
Both of these words stand in stark contrast to another group of conventional AI algorithms that is referred to as rule-based systems. In these types of systems, each decision is manually programmed in such a way that it resembles a statistical model. This is done in order to ensure that the system functions correctly.
When it comes to machine learning and deep learning, there is a wide variety of models that can be utilized, and these models may be separated into two main categories: supervised and unsupervised. Unsupervised learning employs techniques like as k-means, hierarchical clustering, and Gaussian mixture models in an effort to find meaningful structures hidden within the data. This type of learning is not guided by an instructor. Learning through supervision necessitates the application of an output label to every single instance that is present within the dataset.
This output may have values that are discrete or categorical, or it may have values that are real-valued. On the other hand, the outputs of classification models are estimated to have discrete values, whereas the outputs of regression models are predicted to have real values [16].
The simplest type of classification model is called a binary classification model, and it only has two possible output labels: 1 (positive) and 0. (negative). The terms "linear regression," "logistic regression," "decision trees," "support vector machines," and "neural networks" are all examples of popular supervised learning algorithms that fall under the umbrella term "machine learning." Other examples include "support vector machines," "decision trees," and "support vector regression." This category includes contains non-parametric modelling techniques, such as "k-nearest Neighbors," amongst others. We shall be applying a supervised learning strategy for the purpose of this inquiry, which is the scope of it.

LOGISTIC REGRESSION

Logistic regression is a method of statistical analysis that is utilized for the purpose of determining the relationship between a given dataset and one or more independent variables that serve as predictors of an outcome. A wide variety of determining factors could be present in this kind of dataset. Because there are only two plausible interpretations of the data, the result is evaluated using a dichotomous variable, which indicates that there are only two possible outcomes. It is used to produce a prediction about whether the outcome will be binary (1/0, Yes/No, True/False), given a number of independent elements that are being considered. Dummy variables are put to use whenever we need to express results in a manner that is either binary or categorical. Logical regression is another name for a particular type of linear regression that is used when the outcome variable that is being studied is categorical and the log of probabilities is being used as the dependent variable in the analysis. Logical regression can also be seen as another name for logistic regression, which is a form of logistic regression. Putting it another way, it applies the data that was collected to a statistical function that is considered to be reliable in order to derive an estimate of the probability that a particular event will take place.

In the year 1958, a statistician by the name of David Cox was the first person to come up with the idea of logistic regression. With the assistance of this binary logistic model, which takes into consideration one or more predictor elements, also known as independent variables, it is possible to arrive at an estimate of the likelihood of obtaining a binary response (features). It makes it feasible to assert that the existence of a risk factor elevates the possibility of a given consequence by a predetermined percentage.

Sigmoid function P = 1/1+e - (a+bx)

Here P = probability, a and b = parameter of Model.

RANDOM FOREST

In addition to being a part of the ensemble learning family of techniques, this method also performs the functions of classification and regression analysis. This method is very effective when it comes to dealing with really large datasets. Leo Bremen is credited with developing the algorithm known as Random Forest. It is a well-known approach to education that involves learning in groups. Random Forest is able to increase the performance of Decision Tree by bringing the overall variance down to a lower value. It constructs a large number of decision trees while it is in the training phase, and at the end of that phase, it determines the class that corresponds to the mode of the classes, which is also referred to as classification, or it determines the mean prediction, which is also referred to as regression, of the individual trees [17].

1) The first step is to select the “R” features from the total features “m” where R<<M.
2) Among the “R” features, the node using the best split point.
3) Split the node into sub nodes using the best split.
4) Repeat a to c steps until ”l” number of nodes has been reached.
5) Built forest by repeating steps a to d for “a” number of times to create “n” number of trees.

Examining the different possibilities that are open to choose from is the first step that needs to be taken. Next, an educated guess needs to be made regarding the result, using the foundation provided by each arbitrarily generated decision tree. Finally, this educated guess needs to be saved in a number of different places dispersed across the area of interest. In the second stage, the votes that were given for each of the anticipated goals are added up. In the third and final step, the forecast that received the most votes is used as the foundation for the ultimate prediction that is produced by the random forest approach. The user has access to a few of Random Forest's settings, any one of which has the potential to generate accurate forecasts for any one of a number of diverse applications.

DECISION TREE

One specific type of approach that can be used for classifying data is called a decision tree. It is the process of gaining knowledge through observation and guidance. The decision tree is used in situations where the variable being responded to is categorical. A decision tree, which is based on a model that is structured like a tree and describes input features, is used to represent the process of categorization. This tree is based on a model. Input variables can take any form, including graphs, texts, discrete numbers, continuous values, and any other form imaginable.

1) Build the tree with the nodes serving as the input feature.
2) Select the feature to forecast the output from the input feature that has the largest information gain.
3) The information gain that is considered to be the greatest is determined for each characteristic and each node of the tree.
4) Repeat step 2 in order to build a sub tree utilizing the feature that was not used in the node that was previously examined.

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) [14]. The acronym SSE stands for

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

Here, E is the total of the square errors of all of the objects that have cluster means for each of the k clusters, and p is the object that belongs to one of the clusters. The combination of *Ci* and *mi* represents the cluster mean. The total number of records in the dataset is denoted by "n," and "k" indicates the number of clusters.

Input: D is input -data set.
Output: Output is k clusters.
Step 1: Initialize cluster centers as D.
Step 2: Randomly choose k objects from D.
Step 3: Repeat the following steps until no change in cluster means/ min error E is reached.
Step 4: Consider each of the k clusters. Compare the mean value of the objects in the clusters for initialization.
Step 5: Initialize the object with most similar value from D to one of k clusters.
Step 6: Take the mean value of the objects for each of k cluster.
Step 7: Update the cluster means with respect to object value.

VII. PERFORMANCE MEASUREMENT

On the basis of the dataset that we have created, we will evaluate the performance of the algorithms that we have developed [15]. In addition, we will test our model on the dataset that we have prepared. In order to evaluate the efficacy of our newly created classification system and make it comparable to other methods that are currently in use, we use Accuracy as a measure of the effectiveness of classifiers.

True Positives (TP) - These are the positively predicted values that turned out to be correct, which indicates that both the value of the actual class and the value of the predicted class are true. For example, if the actual class value indicates that the passenger survived and the predicted class tells you the same thing, you can assume that the passenger did indeed survive.

True Negatives (TN) - These are the negative values that have been accurately predicted, which indicates that the value of the real class is no and that the value of the predicted class is also no. For example, if the actual class reports that the passenger did not make it out alive while the projected class reports the same thing. These results, known as false positives and false negatives, are produced when your actual class is in conflict with the class that was predicted.

False Positives (FP) – When the class that was actually taken was not the class that was projected to be taken. For example, if the actual class indicates that the passenger did not survive, but the forecast class indicates that they will, this passenger will survive.

False Negatives (FN) – The situation in which the actual class is yes while the projected class is no. For example, if the actual class value indicates that the passenger survived while the predicted class implies that the passenger would die, the actual class value should be used. When you have a firm grasp on these four characteristics, we will be able to proceed with the calculation of accuracy, precision, recall, and F-measure.

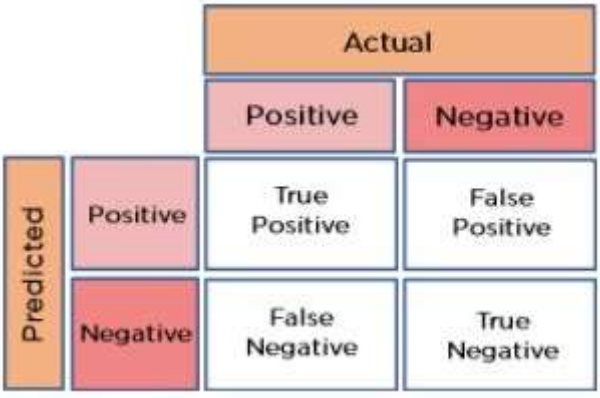

Figure 2: Confusion Matrix

VIII. CONCLUSION

Concerns have been voiced by many who work in the medical field over the best way to diagnose diabetes in its earliest stages. During the course of this research project, an attempt was made to construct a system that would predict diabetes by making use of a number of different algorithms and analyzing how well they performed. This was done in an effort to create a system that could potentially predict diabetes. The research involved the application of three distinct machine learning

algorithms, and the performance of each of these methods was assessed based on a number of different criteria. During the trial, the PIMA Indian Diabetes dataset was utilized, and the findings indicated that logistic regression had the best performance overall. This method of machine learning is also flexible, and it can be applied to the prediction of diseases other than the one that it was initially developed to analyze. The findings have the potential to be enhanced further by the incorporation of other machine learning algorithms, which would result in an increased ability to forecast diabetes.

References

[1] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

[2] 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.

[3] American Diabetes Association 2013, Diagnosis and Classification of Diabetes Mellitus: Diabetes Care, Available from: http://care.diabetesjournals.org/content/36/Supplement_1/S67.full. [January 2013].

[4] Anburajan, M, Sivanandam, S, Bidyarasmi, S, Venkatraman, B, Menaka, M & Raj, B 2011, Changes of skin temperature of parts of the body and serum asymmetric dimethylarginine (ADMA) in type-2 diabetes mellitus Indian patients, Proceedings of the annual international conference of the engineering in medicine and biology society, pp. 6254-6259.

[5] Arif, M & Akram, MU 2010, 'Pruned fuzzy K-nearest neighbor classifier for beat classification', Journal of Biomedical Science and Engineering, vol. 3, no. 4, pp. 380-389.

[6] Ashari, A, Paryudi, I & Tjoa, AM 2013, 'Performance comparison between naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool' vol. 4, pp. 33-39.

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

[8] Asha Gowda Karegowda and Jayaram. A. M., 2009"Cascading GA & CFS for feature subset selection in medical data mining", IEEE International Advance Computing Conference, Patiyala, India

[9] Krzysztof J.Cios, G.William Moore (2002) 'Uniqueness of Medical Data Mining', Artificial Intelligence in Medicine Journal pp 1-19.

[10] Cios, KJ & Moore GW 2002, 'Uniqueness of medical data mining', Artificial Intelligence in Medicine, vol. 26 no. 1, pp. 1-24.

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

[12] 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.

[13] Freudenberg, J & Propping, P 2002, 'A similarity-based method for genome-wide prediction of disease-relevant human genes', Bioinformatics, vol. 18, no. 2, pp. 110-115.

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

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

[16] Guo, Y, Bai, G & Hu, Y 2012, 'Using Bayes network for prediction of Type-2 diabetes, 'Proceedings of the IEEE international conference on internet technology and secured transactions, pp. 471-472.

[17] Hossin, M & Sulaiman, MN 2015, 'A Review on evaluation metrics for data classification evaluations', International Journal of Data Mining and Knowledge Management Process, vol. 5, no.2, pp. 1-11.

