CHAPTER 1

INTRODUCTION

1.1 DATAMINING OVERVIEW

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

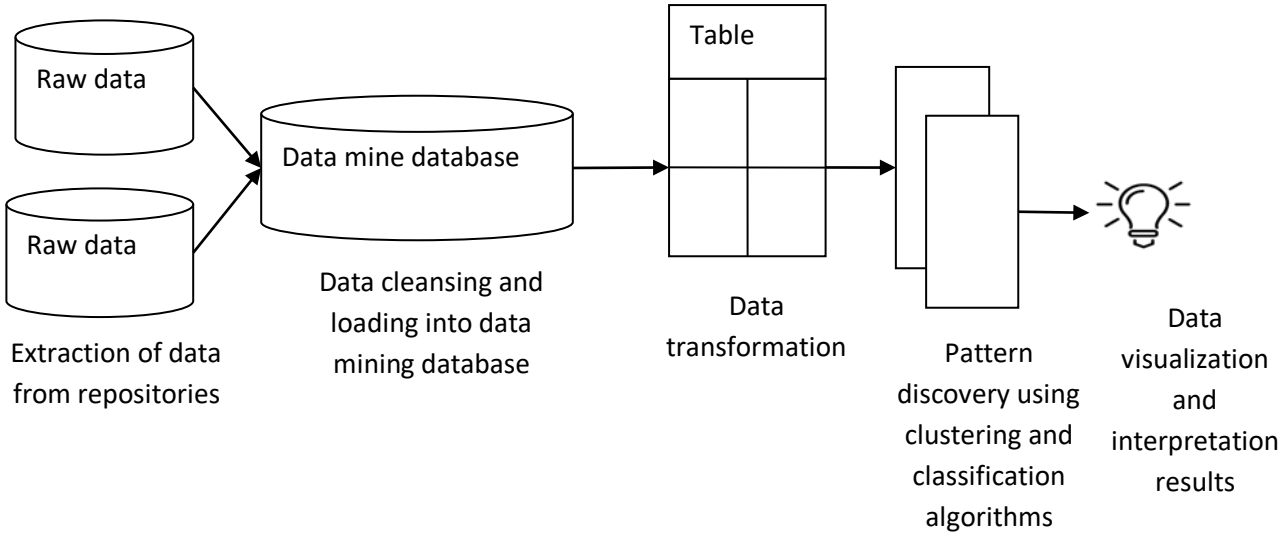

Figure 1 - Background of Data Mining

1.2 ORIGINS OF DATA MINING

Data Mining is the process of posing queries to large amounts of data sources and extracting patterns and trends using statistical and machine learning techniques. It integrates various technologies including database management, statistics and machine learning. Data mining has applications in numerous disciplines including medical, financial, defense and intelligence. Data mining tasks include classification, clustering, making associations and anomaly detection. For example, data mining can extract various associations between people, places or words. During recent years there have been many developments in data mining. The process of digging through data to discover hidden connections and predict future trends has a long history. Sometimes referred to as "knowledge discovery in databases," the term "data mining" wasn't coined until the 1990s. But its foundation comprises three intertwined scientific disciplines: statistics (the numeric study of data relationships), artificial intelligence (human-like intelligence displayed by software and/or machines) and machine learning (algorithms that can learn from data to make predictions). What was old is new again, as data mining technology keeps evolving to keep pace with the limitless potential of big data and affordable computing power. Various data mining techniques have been developed. These include techniques for extracting associations, neural networks, inductive logic programming, decision trees, fuzzy logic and rough sets. Furthermore, data mining has gone beyond mining relational databases to mining text and multimedia data. Also, data mining is being applied to areas such as information security and intrusion detection. While there have been many practical developments, we still have major challenges. One of the most important challenges is scalability. If data mining is to be useful we need to mine very large databases. Therefore, it is critical that we need to understand the limitations of the data mining algorithms. To understand the limitations, we need to study the foundations of data mining. We need to explore the time and space complexity of the algorithms. There are techniques such as inductive logic programming and rough sets that have underpinnings in logic and mathematics. One needs to explore these techniques for data mining and examine the computational complexity aspects. We also need to understand the complexity of the various search algorithms being used for market basket analysis.

1.2.1 PROCESS OF DATA MINING

DATA

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes: Operational or transactional data such as, sales, cost, inventory, payroll, and accounting. Non-operational data, such as industry sales, forecast data, and macro-economic data. Metadata - data about the data itself, such as logical database design or data dictionary definitions.

INFORMATION

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

KNOWLEDGE

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

DATA WAREHOUSES

A data warehouse constructed from integrated data source systems does not require ETL, staging databases, or operational data store databases. The integrated data source systems may be considered to be a part of a distributed operational data store layer. Data federation methods or data virtualization methods may be used to access the distributed integrated source data systems to consolidate and aggregate data directly into the data warehouse database tables. Unlike the ETL-based data warehouse, the integrated source data systems and the data warehouse are all integrated since there is no transformation of dimensional or reference data. This integrated data

warehouse architecture supports the drill down from the aggregate data of the data warehouse to the transactional data of the integrated source data systems.

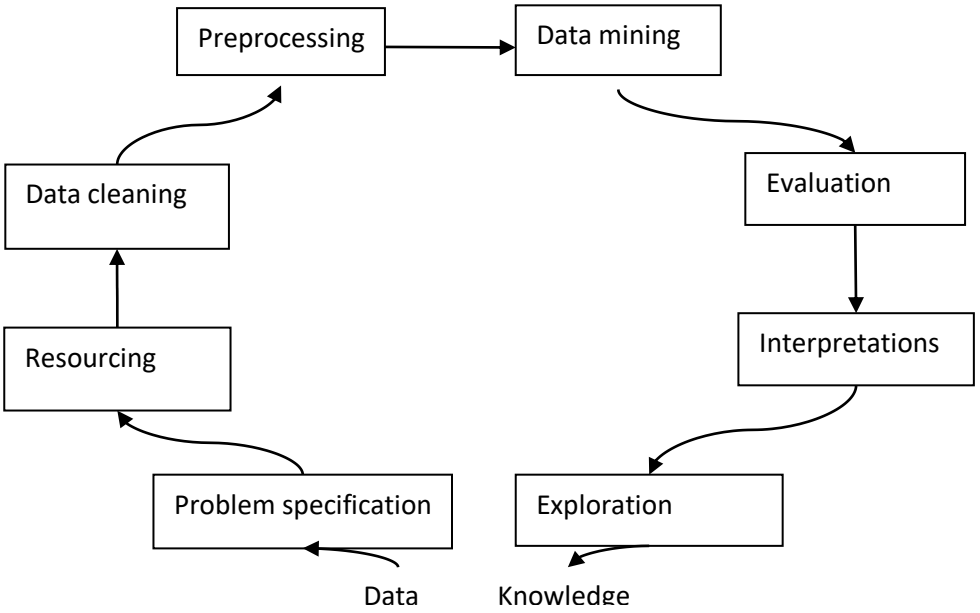

Figure 2 - Process of Data Mining

1.2.2 LEVELS OF DATA MINING

CLUSTERING

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number

of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties. Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, and typological analysis. The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest.

CLASSIFICATION

Classification is a general process related to categorization, the process in which ideas and objects are recognized, differentiated, and understood. A classification system is an approach to accomplishing classification. In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition. In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance. Often, the individual observations are analyzed into a set of quantifiable properties, known variously as explanatory variables or features. These properties may variously be categorical , ordinal, integer-valued or real-valued. Other classifiers work by comparing observations to previous observations by means of a similarity or distance function. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category. Terminology across fields is quite varied.

1.3 MACHINE LEARNING

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

Machine learning tasks are classified into several broad categories. In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. For example, if the task were determining whether an image contained a certain object, the training data for a supervised learning algorithm would include images with and without that object (the input), and each image would have a label (the output) designating whether it contained the object. In special cases, the input may be only partially available, or restricted to special feedback. Semi-supervised learning algorithms develop mathematical models from incomplete training data, where a portion of the sample input doesn't have labels. Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited set of values. For a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email. For an algorithm that identifies spam emails, the output would be the prediction of either "spam" or "not spam", represented by the Boolean values true and false. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range. Examples of a continuous value are the temperature, length, or price of an object. In unsupervised learning, the algorithm builds a

mathematical model from a set of data that contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points. Unsupervised learning can discover patterns in the data, and can group the inputs into categories, as in feature learning. Dimensionality reduction is the process of reducing the number of "features", or inputs, in a set of data.

1.4 DEEP LEARNING

A deep neural network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers. The DNN finds the correct mathematical manipulation to turn the input into the output, whether it will be a linear relationship or a non-linear relationship. The network moves through the layers calculating the probability of each output. For example, a DNN that is trained to recognize dog breeds will go over the given image and calculates the probability that the dog in the image is a certain breed. The user can review the results and select which probabilities the network should display (above a certain threshold, etc.) and return the proposed label. Each mathematical manipulation as such is considered a layer, and complex DNN have many layers, hence the name "deep" networks. DNNs can model complex non-linear relationships. DNN architectures generate compositional models where the object is expressed as a layered composition of primitives. The extra layers enable composition of features from lower layers, potentially modeling complex data with fewer units than a similarly performing shallow network. Deep architectures include many variants of a few basic approaches. Each architecture has found success in specific domains. It is not always possible to compare the performance of multiple architectures, unless they have been evaluated on the same data sets. DNNs are typically feed forward networks in which data flows from the input layer to the output layer without looping back. At first, the DNN creates a map of virtual neurons and assigns random numerical values, or "weights", to connections between them. The weights and inputs are multiplied and return an output between 0 and 1. If the network didn’t accurately recognize a particular pattern, an algorithm would adjust the weights.

1.5 DIABETES INTRODUCTION

Diabetes mellitus is the disease which is persisting for the long time in the human body. It is otherwise called as chronic disease that occurs when the pancreas is no longer. Pancreas is the gland which is present behind the stomach which is responsible for segregating the digestive enzyme into the duodenum. The term diabetes was first coined by Apollonius around the year of 250 BC after that in the year of 1675 the term mellitus was added to the diabetes by Thomas Willis and it was called as diabetes mellitus. The patients will have the symptoms of frequent urination, increased thirst, feeling very tired, always feeling hungry, slow healing of the wounds and patches of dark skin in the body. There are two types of diabetes of type-1 and type-2. The people with type-1 diabetes don't produce insulin and people with type-2 don't response to insulin and this type of diabetes will not produce enough insulin in the human body as mentioned in Figure 1.5. Nearly 95 percent of people were affected by diabetes in the world. In India diabetes affects nearly 62 million people which is more than 7.2 percent of the population. Type-1 diabetes will be common for the people. Metformin is the first medication prescribed for the type-2 diabetes patients. The latest research on diabetes will reduce the segregation of fat build-up in the pancreas and in liver so that the insulin segregation will be more in the pancreas.

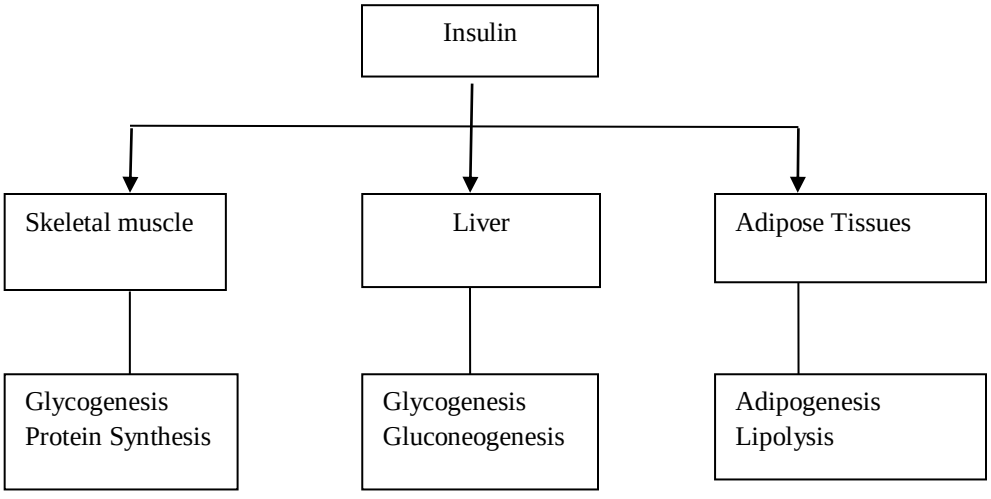

Figure 3 - Insulin Production Throughout The Parts Of The Body

CHAPTER 2

LITERATURE SURVEY

2.1. SURVEY: DIABETES ANALYSES FOR PREGNANT LADIES

AUTHOR: LAHIRU LIYANAPATHIRANA

This Survey says that there are several stages in analyzing diabetes for Pregnant Ladies,

1. Data Preparation
2. Data Exploration
3. Data Cleaning
4. Model Selection

2.1.1. STAGE 1: DATA PREPARATION

In this survey, the own dataset is created instead of using the already existing dataset called the Pima India Diabetes Dataset which was given by the unique client identifier machine learning repository .

2.1.2. STAGE 2: DATA EXPLORATION

The Pregnancy data set was collected to analyze the particular data of the patient to predict the diabetes. The dimensions of the data set were calculated by using the Panda Data Frame. In that the Shape attribute has been used. From this data set the prediction of diabetes for the pregnant ladies has been analyzed. From the result if the column is one the patient is with diabetes or if the patient is with result of column is zero then the patient is not with the diabetes.

2.1.3. STAGE 3: DATA CLEANING

In the process of data cleaning they have used the Better Data Beats Fancier Algorithm which have produced the best result. There were some of the factors to be considered in the process of data cleaning.

1. Duplicate values in the dataset
2. Bad labeling in the dataset
3. Missing value or null data point

In all these factors the unexpected outlier is the most important one because in the dataset the value for the blood pressure of the patients is zero. This data seems to be wrong because a living person cannot have a diastolic blood pressure of zero. In the analyses of the same dataset the plasma glucose level was zero for the patient and the skin thickness for the normal patient will not be less than 10mm but the analyses for the dataset will have the skin thickness as zero. In the rare cases the insulin for the patient will be zero but in the analyses of the dataset it results as zero.

2.1.4. STAGE 4: MODEL SELECTION

The important stage in the data analyses is that algorithm selection. They have used totally of seven classifier of Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Gaussian Naive Bayes, Random Forest And Gradient Boost. Among these seven algorithms they have chosen the best algorithm of Logistic Regression. In the logistic algorithm they have achieved the accuracy of 77.64 percent. This algorithm was considered as the prime candidate for the next phase. The accuracy of the above mentioned algorithm from this survey are,

1. KNN - 0.71 -71%
2. SVM - 0.65 - 65%
3. LR - 0.77 - 77%
4. DT - 0.68 - 68%
5. RF - 0.74 - 74%

Logistic regression is considered as best algorithm in giving accuracy

2.1.5. DATASET

Table 1: Datasets For Pregnant Ladies

Patient	Glucose	Glucose level	Blood pressure level	Skin thickness	Insulin	BMI	Diabetes pedigree	Age of patient	Out Come
0	5	149	71	28	0	32.6	0.638	45	1
1	1	87	64	34	0	24.6	0.347	30	0
2	7	143	64	0	0	22.3	0.681	29	1
3	1	90	76	30	90	21.1	0.214	31	0
4	0	117	45	21	156	40.1	2.355	41	1

2.2. SURVEY: HYPERGLYCAEMIA IN PREGNANCY (HIP)

AUTHOR: LILI YUEN, POUYA SAEEDI, MUSARRAT RIAZ

The Main objective of this analyzes is to reduce the complication of pregnancy. Hyperglycemia is one of the most complications in the pregnancy. This analyzes was done to reduce the complication of pregnancy in the upcoming year of 2030 and 2045.

2.2.1 METHODS

The international diabetes federation (IDF) had used many methods to reduce the complication in the pregnancy which is projected in the year of 2030 and 2045.

1. Carrying the age adjusted prevalence rates in the year by SVM [Figure 4]
2. Applying the Linear Regression to the past four edition of the IDF [Figure 5]
3. Applying the Linear Regression to the previous edition of the IDF with the most consistent trends followed by the extrapolation [Figure 6]

Hyperglycemia is one of the metabolic changes during the pregnancy. Hyperglycemia in pregnancy (HIP) was defined by the world health organization. The WHO had described the HIP, diabetes first detected at any time during the time of pregnancy. It was defined as pre-existing diabetes and it was further classified into two types.

1. Diabetes in pregnancy
2. Gestational diabetes mellitus

Charts were used to describe the analyses of Hyperglycemia in different years. The process of reducing the Hyperglycemia in the upcoming year of 2030 and in 2045 has discussed.

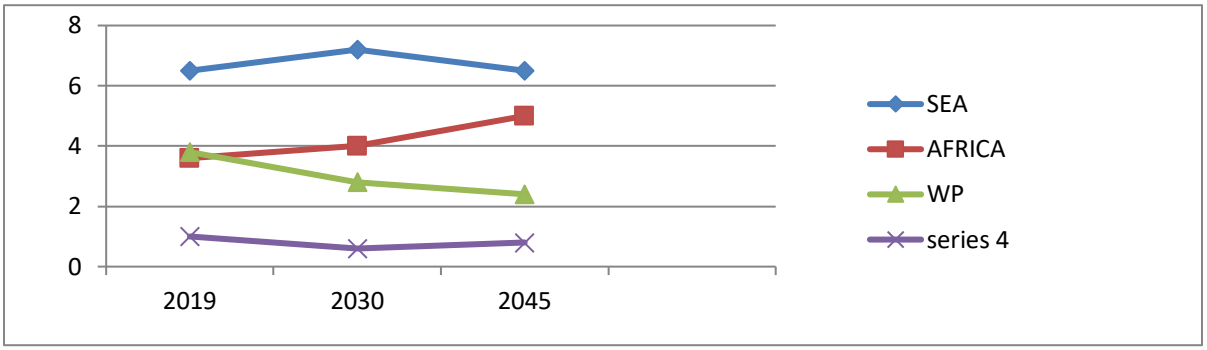

Figure 4 - Future Analyses Of Hyperglycemia In 2030, 2045

Method -1 Analyzes

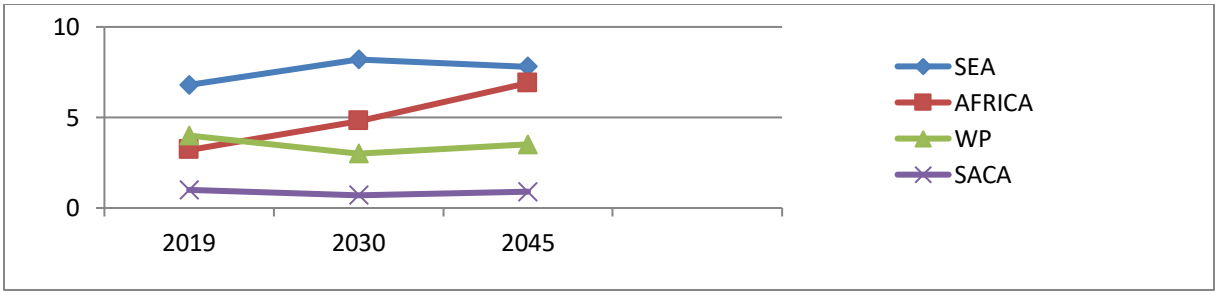

Figure 5 - Future Analyses Of Hyperglycemia In 2030, 2045

Method-2 Analyzes

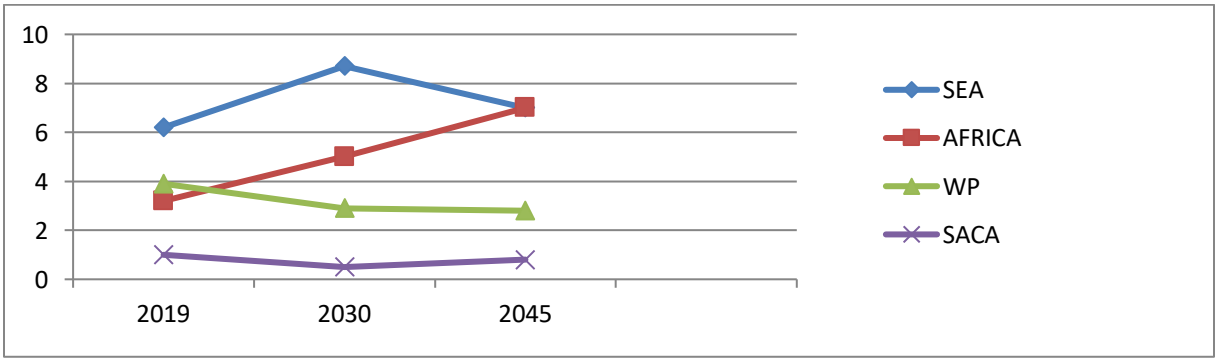

Figure 6 - Future Analyzes Of Hyperglycemia In 2030, 2045

Method- 3 Analyzes

2.2.2. DATASET

Table 2: Datasets For Hyperglycemia

COUNTRY	TOTAL BIRTH 2030	HIP IN 2030	AGE ADJUSTED IN %	TOTAL BIRTH 2045	HIP IN 2045	AGE ADJUSTED IN %
SEA	25,645,284	7,487,451	26.44	23,240,547	6,421,410	26.44
AFRICA	38,341,741	4,048,784	10.24	47,471,241	4,745,466	10.22
WP	25,752,654	6,745,471	10.17	24,741,541	2,540,471	10.18
SACA	64,471,265	7,241,476	10.46	48,596,415	3,462,189	1.44
OVERALL PREVALANCE	1,30,594,412	18,456,546	14.00	135,750,047	17,993,475	13.25

From the above survey, the results for analyzes of Hyperglycemia in the year of 2030 and in 2045 are shown as percentages in the below table.

Table 3: Accuracy For Each Methods Of Hyperglycemia

METHODS	2030 (IN %)	2045 (IN %)
METHOD ONE	12	13.4
METHOD TWO	16.2	18.7
METHOD THREE	16.0	15.4

2.3. SURVEY-3: 52 WEEK OBSERVATIONAL STUDY USING THE FOUR INHIBITORS

AUTHOR: EU JEONG KU, DONG-HWA LEE, HYUN JEONG JEON, TAE KEUN OH

In this observational study the effectiveness of the two distinct inhibitors is compared and results were discussed in the table. The two inhibitors are in the following,

1. Sodium glucose co-transporter 2(SGLT-2)
2. Empagliflozin and Dapagliflozin

The second inhibitor performs as the oral anti diabetic agents. These inhibitors were the controlling remedy for the type-2 diabetic in patient.

2.3.1 METHODS

The observational study was first done the patient with Glycated Hemoglobin (HbA1c). The Glycated Hemoglobin is the content of glucose in the blood. The presence of glucose in the red blood cells is Glycated Hemoglobin. This presence of glucose will be there for about 120 days or for about 3 months. This glucose in red blood cells will be in the range of 7.2 to 12.0 %. It will be present along with the other inhibitors of Metformin, Glimepiride, and Dipeptidyl Peptidase-4. The patients will be classified into two types of categories.

1. Patient with Empaglinflozin (25 mg/day)
2. Patient with Diapaglinflozin (10 mg/day)

The results from these observational studies depend on the changes in the HbA1c, and the fasting plasma glucose (FPG) in the blood.From the above the Research Design proves that the person with adult age of 18-80 will have the common diabetes of type-2. These persons will have the Glycated Hemoglobin in the range of >= 7.5 to < 12.0 % at the baseline along with the three inhibitors.

1. Metaformin - 2000 mg/Day
2. Glimepiride – 8 mg/Day
3. Dipeptidyl Peptidase - (local stay for > 12 weeks)

2.3.2 ENDPOINTS DESIGN

The primary endpoints were designated to two things of:

1. HbA1c Mean changes
2. Fasting plasma glucose

The secondary endpoints were focus on the following parameters.

1. Changes in body weight
2. Systolic (SBP)

The symptoms of the hypoglycemia are sweating the one type of removal of waste water from the body from the skin of the body, tremors the muscle contraction in the body which leads to rhythmic movement in the body or the occurrence of shaking in the hands and in legs, palpitation the excess fasting of heart beat in the patient. Patient with type-2 with oral anti diabetic drug for 12weeks (n=393). The patients classified into two types of Empagliflozin (n=180) and patient with Dapagliflozin (n=213). In both the types the patients with Empagliflozin shows the greater reduction in the HbA1c with the least production of GUT in the number of three (n=3) along with the volume depletion in the number of two (n=2). In this observational study the total analyses was done with 362 patients. Some patients with the Empagliflozin (n=180) and some patients with Dapagliflozin (n=182). The analyses were done for 52 weeks. After the weeks, the final outcome result produces the reduction in the HbA1c and FPG. Both the types of patients will have reducing of HbA1c and in FPG in the final results. But, the reduction of Empagliflozin was greater than the Dapagliflozin in the patients. Along with this reduction the patients had the decrease in their blood pressure, body weight and in lipoprotein cholesterol. From these inhibitors the sodium glucose co-transporter-2 (SGLT-2) acts as the effective remedy for the patients with type-2 diabetic. At the same time the Empagliflozin acts as greater remedy for reducing the HbA1c than the Dapagliflozin.

2.3.3 FLOWCHART

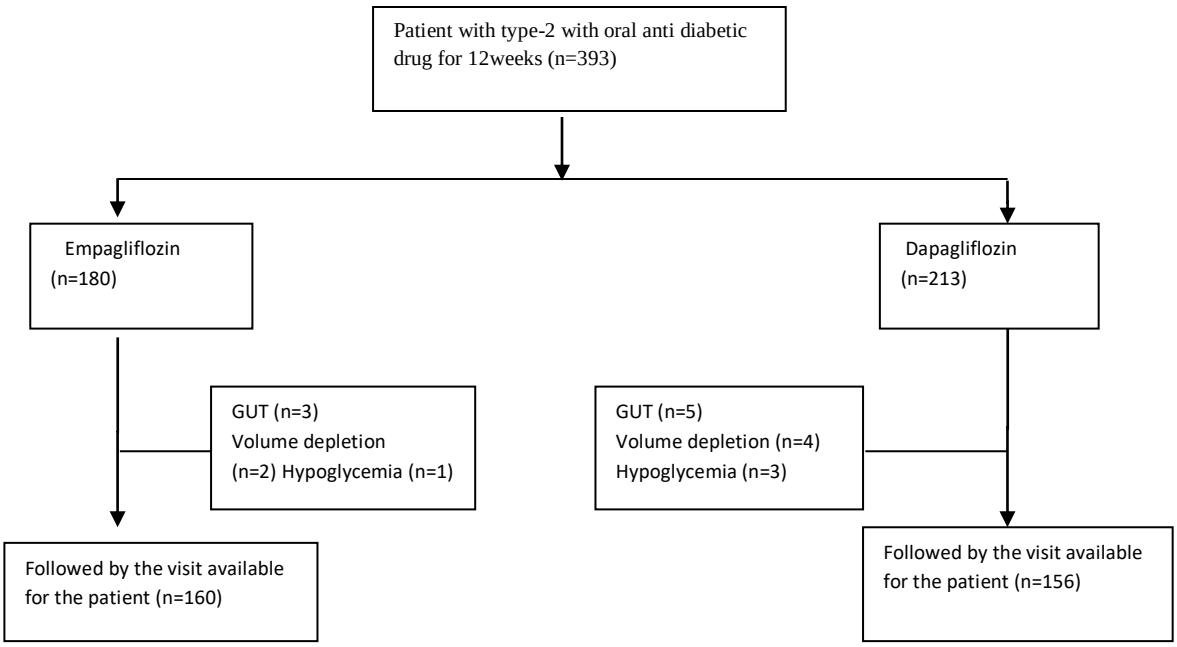

Figure 7 - Flowchart For Inhibitors

2.4 SURVEY-4: A SURVEY ON MEDICAL DIAGNOSIS OF DIABETES USING MACHINE LEARNING TECHNIQUES

AUTHOR: AMBIKA CHOUDHURY

At an early period of diagnosing the diabetic disease, machine learning techniques can help the physicians to diagnose and cure diabetic diseases. This paper presents a comprehensive comparative study on various machine learning algorithms for PIMA Indian Diabetic dataset. The comparative study is based on the parameters such as accuracy, recall, specificity, precision, negative predicted value (NPV), false positive rate (FP rate), rate of misclassification (RMC), F1-measure, and G-mean. Increase in classification accuracy helps to improvise the machine learning models and yields better results. The performance analysis is analyzed in terms of accuracy rate among all the classification methods such as decision tree, logistic regression, k-nearest neighbors, naïve Bayes, and SVM. It is found that logistic regression gives the most accurate results to classify the diabetic and non-diabetic samples. Future work can be done in such a manner that type I and type II diabetes can be made possible to identify in a single classifier. By developing classifier system, machine learning algorithm may immensely help to solve the health-related issues which can assist the physicians to predict and diagnose diseases at an early stage. We can ameliorate the speed, performance, reliability, and accuracy of diagnosing on the current system for a specific disease by using the machine learning classification algorithms. This paper mainly targets the review of diabetes disease detection using the techniques of machine learning. Further, PIMA Indian Diabetic dataset is employed in machine learning techniques like artificial neural networks, decision tree, random forest, naïve Bayes, k-nearest neighbors, support vector machines, and logistic regression and discussed the results with their pros and cons.

2.5 SURVEY-5: PREDICTION AND DIAGNOSIS OF FUTURE DIABETES RISK: A MACHINE LEARNING APPROACH

AUTHOR: ROSHAN BIRJAIS

Machine learning is a discipline where machines are instructed without human intervention by mean of algorithms. We can train them to perform a particular task and based on the training they can be used to handle the similar job without being explicitly programmed. Training the machine with some algorithm and feeding them with the dataset results into the

formation of a classifier and to find the accuracy of the classifier we test it in the testing phase. Accuracy is always major issue in medical science and with different algorithms we can obtain different accuracies on same data set and it is very important to see which algorithm provides best result in order to obtain better classifier to do the better classification. Machine learning can be used in almost every field nowadays. Using it in the field of medical science can prove to be very beneficial in the improvement of healthcare. Health care is always a big issue for any nation and is always challenging thing to provide. Better, the health care of a nation better is the condition of the inhabitants living there. Improvement in the health care can directly result in economic growth because a healthy person can prove to be a big asset to the nation and can conduct activities effectively in the workforce than any unhealthy person. Health care is aggregation and integration of all the measures, which can be taken to improve the health system. Healthcare constitutes prevention, diagnosis and treatment. Improving healthcare should be the first priority and major task to ponder on. Usage of technologies in the improvement of the health care is proved very beneficial

2.6 SURVEY-6: PREDICTION AND PREVENTION OF HYPOGLYCAEMIC EVENTS IN TYPE-1 DIABETIC PATIENTS USING MACHINE LEARNING

AUTHOR: JOSEP VEHÍ

A novel system for the prediction of hypoglycemic events in T-1 patients has been presented. Machine learning methods were applied to different datasets for patient condition assessment, continuous glucose level prediction, and the prediction of postprandial and nocturnal hypoglycemic events. Even though the systems performed effectively, they have been analyzed separately considering only data from CSII therapy. However, most of the methodologies can be adapted for MDI. Models with different prediction goals based on diverse techniques working in parallel provide an increased robustness for the proposed system. Each predictive system performs better in particular scenarios. However, the combination of different models increases the possibility of anticipating events that would probably have been missed if a unique prediction subsystem was considered. The simultaneous utilization of these different personalized prediction models will allow the evaluation of an integrated and robust system for the prevention of hypoglycemic events in both CSII and MDI users. Individuals with T1D face a lifelong challenge to maintain their BG levels within a safe range, by reducing hyperglycemia without increasing the risk of hypoglycaemia.5 However, tightening glycaemic control increases the risk

of hypoglycaemia.6–8 Besides the large intra- and inter-day glycaemic variability, which presents a barrier to achieving optimum insulin therapy, patients' habits play an important role in their glycaemic control. Meals, physical activity, menstruation, illness, and stress are the main challenges for patients and physicians to keep BG levels into normal levels

2.7 SURVEY-7: PREDICTIVE ANALYTICS IN HEALTHCARE FOR DIABETES PREDICTION

AUTHOR: FAIZAN ZAFAR

Diabetes mellitus type 2 is a chronic disease which poses a serious challenge to human health worldwide. Globally, about 8.3% of the population is diagnosed with the disease. The applications of predictive analytics in diagnosis of diabetes are gaining significant momentum in medical research. The aim of this research paper is to aid medical professionals in the early detection and efficient diagnosis of Type 2 diabetes. We utilize bioinformatics theory and supervised machine learning techniques for improving the accuracy in predicting diabetes, based on 8 clinical measurements existing in the widely used PIMA dataset. We outline our methodology and highlight the implementation steps, while reviewing prominent past work in the field. Moreover, this paper fully exploits known machine learning algorithms and provides a detailed comparison of the results obtained from each method. In this paper, we presented our approach in the design and development of a diabetes prediction model, which uses the gradient boosting algorithm to contribute to the actual diagnosis of Type 2 diabetes for local and systemic treatment, along with presenting related work in the field. Experimental results show the effectiveness of the model with an F1 score of 0.853 and out of sample prediction accuracy of 89.94%. The performance and evaluation of the applied machine learning techniques for the problem of diabetes diagnosis were thoroughly investigated. Detection of diabetes in its early stages is the key for treatment. By transforming clinical data into useful results, this analysis utilizes artificial intelligence in bioinformatics for knowledge discovery and predicting future occurrences. As part of the future work, we plan to engage formally in data collection from local sources and derive a more user specific model for diabetes prediction. Furthermore, we aim to research into more features which can be used as potential predictors. In this way, this work can be expanded to aid in fully automating diabetes prediction.

2.8 SURVEY-8: A MACHINE LEARNING APPROACH TO PREDICTING BLOOD GLUCOSE LEVELS FOR DIABETES MANAGEMENT

AUTHOR: KEVIN PLIS

Patients with diabetes must continually monitor their blood glucose levels and adjust insulin doses, striving to keep blood glucose levels as close to normal as possible. Blood glucose levels that deviate from the normal range can lead to serious short-term and long-term complications. An automatic prediction model that warned people of imminent changes in their blood glucose levels would enable them to take preventive action. In this paper, we describe a solution that uses a generic physiological model of blood glucose dynamics to generate informative features for a Support Vector Regression model that is trained on patient specific data. The new model outperforms diabetes experts at predicting blood glucose levels and could be used to anticipate almost a quarter of hypoglycemic events 30 minutes in advance. Patients strive to avoid both hyperglycemia, or high BG levels, and hypoglycemia, or low BG levels. Hyperglycemia can lead to long-term complications including blindness, amputations, kidney failure, strokes, and heart attacks, while hypoglycemia can cause immediate symptoms of weakness, confusion, dizziness, sweating, shaking, and, if not treated in time, seizures, coma or death. Physiological models try to capture the dynamics of glucose relevant variables within different systems in the body. For example, equations have been introduced in the literature for tracking the carbohydrate intake as it is converted to blood glucose which then interacts with the kidneys, liver, muscles, and other body systems. Most physiological models characterize the overall dynamics into three compartments: meal absorption dynamics, insulin dynamics, and glucose dynamics

CHAPTER 3

SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Present days one of the major application areas of machine learning algorithms is medical diagnosis of diseases and treatment. Nowadays Machine learning algorithms widely used in various fields especially of medical field. In medical field Machine Learning had been used for disease diagnosis and its treatment. Two types of machine algorithms had been used are correlation and associations for finding different diseases. In present days some people are dying because of serious disease called diabetics. This diabetes mellitus had been Predicted and diagnosed at the various stages of the patients and this had become one of the challenging factors which is faced by many doctors and hospitals in everywhere. To reduce the seriousness of the diabetes in the patients we have predicted this disease at the initial stage itself of the patients. In this project we have used machine learning algorithms and deep learning algorithms to predict this disease. Many researchers have been developing the software to help doctors to take decision regarding both prediction and diagnosing of heart disease. We have also discussed about the data mining techniques which is used to predict the diabetes disease in advance. The diagnostic system in this process is used to determine the presence of diabetes is not. For this diagnostics system, machine learning algorithms are widely used. We have used machine learning techniques in the medical field because of its characteristics i.e., high performance to deal with missing data, irrelevant data and noisy data, and the ability to explain decisions. As everyone is using more date everywhere and there will be a need for some classifier to classify the newly generated data. The classifiers are used for its accuracy and efficiency. The already existing system had mainly focused on the supervised learning technique called the Random forests by changing the values of different parameters.

3.1.1 EXISTING BLOCK DIAGRAM

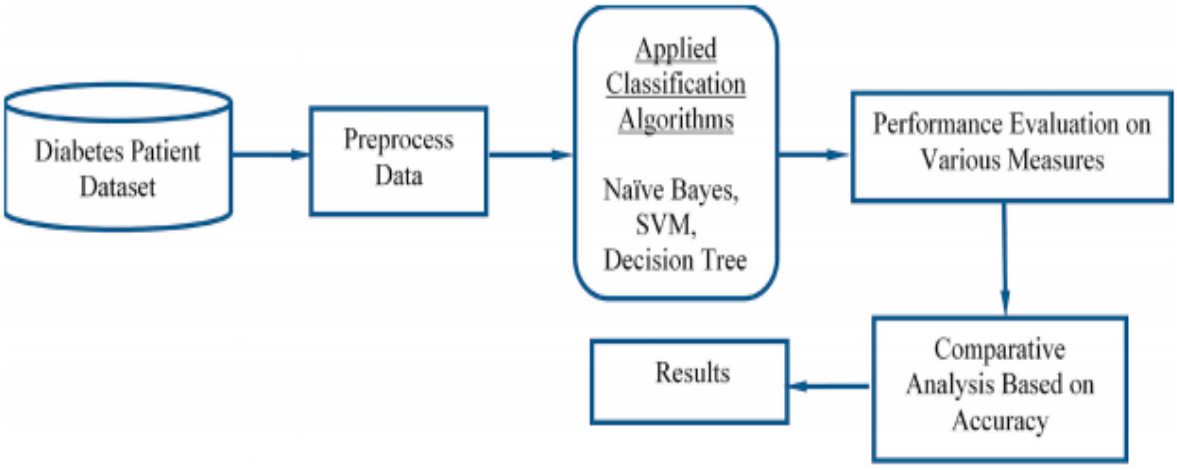

Figure 8 - Machine Learning Based Diabetic Prediction

3.2 DISADVANTAGES

- Labeled data based disease classification
- Provide high number of false positive
- Binary classification can be occurred
- Computational complexity

CHAPTER 4

PROBLEM IDENITIFICATION

Diabetes mellitus is a most common disease faced by most of patients can have uncontrollable glucose level can lead to chronic disease to prevent this risk of higher chances of chronic diseases. Almost 382 million people have diabetes over the world. Diabetes is known as diabetes mellitus a cluster of chronic illness which is faced due to increase in level of blood glucose level and decreased insulin level in body. Symptoms are polyuria - frequent urination, polyphagia - high appetite, polydipsia - increased intake of water. Three types of Diabetes names are Type -1, Type-2 and Gestational Diabetes. Type-1 Diabetes is also called juvenile diabetes mellitus where patient suffers right from childhood as pancreas cannot produce sufficient insulin hence intake of insulin and diabetic medication as per doctor's advice is must. In rare cases, people may suffer from secondary diabetes which is same as type-1 which doesn't affect beta cells but affects immune system by some disease which affects pancreas. The destruction of beta cells prevents entry of glucose into blood without insulin hence it piles up in blood causes rise in blood sugar levels. Patients of Type -1 Diabetes go through Diabetes Ketoacidoisis in which body cannot store glucose hence converts fat cells in the form of ketoses. Existing system developed a system using genetic algorithm oriented semantic features which are used for generalizing text in text classification algorithms and comprises of selecting features from text. First stage of computation is carried out using state of the art algorithm. Second stage is carried out using latent semantic indexing entitled by genetic algorithm but outcome of these where corresponded to larger values. It also developed a model which will be called as predictive risk which integrates various data analysis techniques like data mining, machine learning and statistics as it predicts future with current data in the current case it might predict risk associated with diabetes. The dataset used is Pima-Indian Dataset and algorithm which has been used is machine learning in Hadoop MapReduce in order to find missed values on it. From all these survey, problem occurred both in T-1 and in T-2. In T-1 diabetic, the patients will have beta cells destruction in the pancreas. This will lead to the insulin deficiency in the body. In T-2 diabetic, the patients will have progressive damage, dysfunction and various failures of organs including the kidney, nerves, eyes, blood vessels. This occurrence of risk in patient is due to poor prior notification and proper treatment for their diabetes.

CHAPTER 5

PROPOSED SYSTEM

Several opportunities are used for medical fields because of machine learning models which have high potential advantage for predictive analysis. There are already existing models in machine learning which can predict the chronic illness like heart disorder, infections and intestinal diseases. Several researches had been used in machine learning models to diagnosis and predict the non-communicable diseases, which have more advantage in the medical field. Upcoming researchers have been working on deep learning models to predict specific disease especially of diabetes in the patient more effective in the prevention of the diabetes diseases. So this hospitalization of patients will get reduce. This will help the medical organization by providing more beneficial transformation. Diabetes is a chronic disease which reduces the insulin level and increases the glucose level in the body. In the diabetes patient's body cannot respond to the hormone of insulin production. This results in anomalous metabolism of carbohydrates and increased blood glucose levels. Early detection of diabetes becomes very important. This disease causes increased level of glucose in the body. Glucose will get generated in the body after eating food. The Insulin hormone helps to maintain the balanced state of glucose level and the blood sugar level in the body. T-1 diabetes is a scenario where the body does not produce insulin at all to balance the sugar levels in blood. T-2 is a diabetes type where the body produces insulin but does not utilize this hormone to balance blood sugar levels. T-1 diabetes is most common one. Some certain cases called pre-diabetes. In this situation the person will have high glucose level but they not consider under diabetes patients. But the people who have pre-diabetes are prone to get t-2 diabetes. This disease can cause serious damage to many vital organs in the body like kidneys, heart, nerves and eyes. If a woman suffers from this disease during their pregnancy then it is called gestational diabetes. So implement deep learning based neural network algorithm can be used to predict the diabetic diseases with improved accuracy. Neural Networks has emerged as an important method of classification. Back-propagation has been employed as the training algorithm in this work. This project proposes a diagnostic system for predicting heart disease with improved accuracy. The propagation algorithm has been repeated until minimum error rate was observed. And also provide the diagnosis information to patients through SMS alert and also provide voice information to patients based on trained diabetic datasets.

5.1 PROPOSED BLOCK DIAGRAM

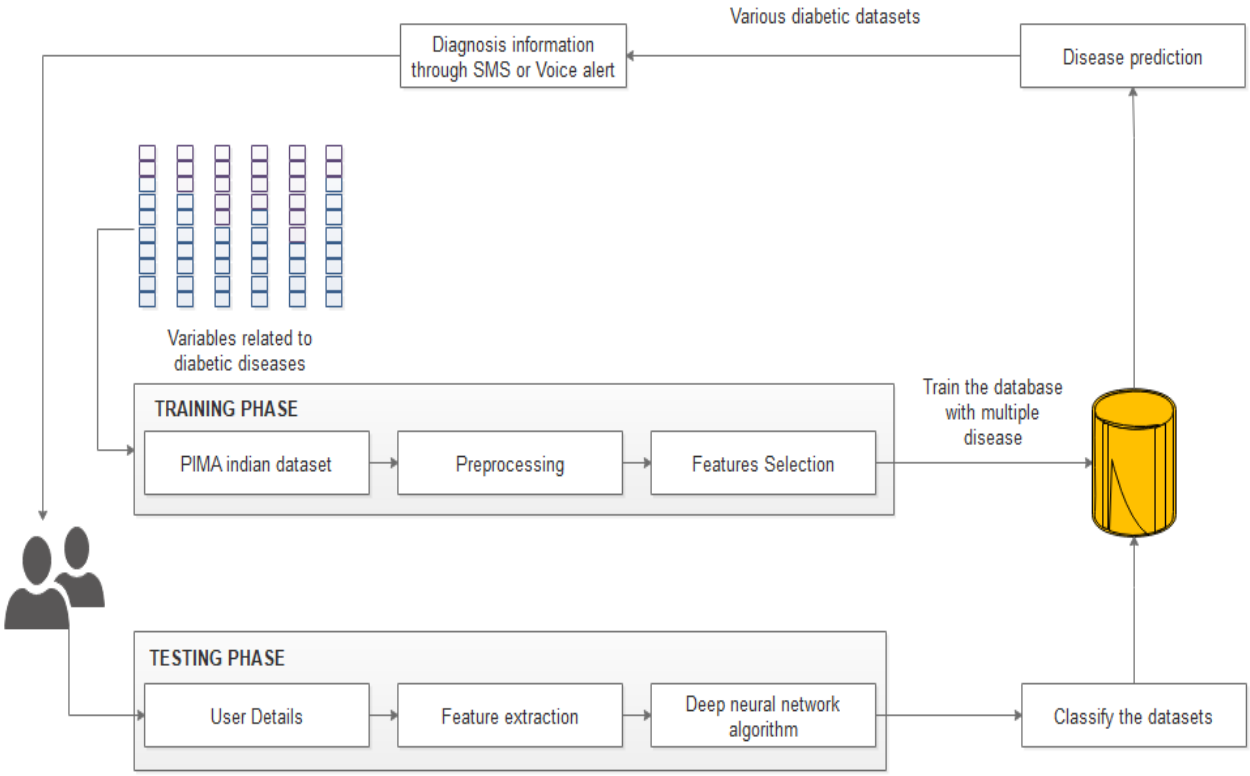

Figure 9 - Proposed Block Diagram

5.2 ADVANTAGES

- Accuracy is high
- Parallel processing
- Multiple diabetic diseases are predicted
- Reduce number of false positive rate

5.3 MODULES DESCRIPTION

- Datasets Acquisition
- Preprocessing
- Features Selection
- Classification
- Disease diagnosis

5.3.1 DATASETS ACQUISITION

A data set is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a variable, and every row represent the particular member of the data set. The variables represented in the dataset produces the information about the height and weight of an object or particular person. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. The data set refers to collection of data which are closely related to the particular events. In this module, we can upload the cardiovascular datasets related to diabetic diseases which includes the attributes such as glucose, insulin level, blood pressure, BMI and so on.

5.3.2 PREPROCESSING

Removing the missing data and irrelevant data is an important step in the process. This is called preprocessing. The two phrases are used in data mining and machine learning techniques are "garbage in and garbage out". Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc. If the data analysis is not done properly then it leads to problems which cause major issues in the medical field. To avoid this major issues the representation and quality of data should be analyzed carefully. If the dataset contains more irrelevant and redundant information, it will causes more difficulties during the training phase. During preprocessing the considerable amount of time will be taken by two process of data preparation and data filtering. In this module, we can eliminate the irrelevant values and also estimate the missing values of data. Finally provide structured datasets.

5.3.3 FEATURES SELECTION

The major process consider during feature selection is to reduce the inputs for processing and analyzing. The useful information or features from already existing data will be extracted by the process of feature engineering. The statistical measures are applied to filter the feature selection according to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often uni-variate and consider the feature independently, or with regard to the dependent variable. It can be used to construct the multiple diabetic diseases. In this module, select the multiple features from uploaded datasets. And train the datasets with various disease labels such as T-1 diabetics with diagnosis information. T-2 diabetics with diagnosis information

5.3.4 CLASSIFICATION

In this module implement classification algorithm to predict the diabetic diseases. To predict the disease we have used one of the deep learning algorithms such as back propagation. A back propagation is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. It consists of multiple layers of nodes in a directed graph, and each layer is fully connected to the next one. Each and every node is considered as a neuron with a function of nonlinear activation except for the input nodes. Supervised learning technique uses the back propagation methods for network called back propagation for training the network. Back propagation distinguishes the data which is not linearly separable and it is entirely modified form of linear perceptron. If back propagation uses some mechanism i.e., if all the neurons contains linear activation function, to determine whether the neurons fires or not, then it can be easily reducible with linear algebra with many number of layers into two standard layers of input-output model. To optimize the methods to adjust the heavy weights by reducing the loss function in the network. This can be done by using gradient techniques. To compute this gradient technique of loss function the algorithm must requires a known and a desired output for all types of inputs. Usually, the generalization of back propagation Feed Forward Networks is done using delta rule which possibly makes a chain of iterative rules to compute gradients for each layer. Back Propagation Algorithm necessitates the activation function to be different between the neurons. Back Propagation algorithms are currently implemented as the major concepts on the researches of parallel and distributed computing. This Back Propagation algorithm plays a vital

role in the pattern recognition domains. They are so convenient in research, because of their ability in solving complex problems, and also for their fitness approximation results even with critical predictions. The architecture of feed forward back propagation for supervised training is same as the architecture of back propagation in the neural network. The back propagation is the most known and most frequently used type of neural network. User can provide the features and automatically predict the diseases.

5.3.5 DISEASE DIAGNOSIS

The disease diagnosis done by the physicians and other health specialists is done by the decision making tasks. Medical decision support system is a set of decision making support program. In this module, provide the diagnosis information based on predicted diabetic diseases. Proposed system provides improved accuracy in diabetic disease prediction. Some of the conditions or habits are considered as the risk factors of the patients which make them to develop the disease. In this module, provide the diagnosis information based on predicted diabetics diseases. Information sends to user in the form of SMS or Voice information. Proposed systems provide improved accuracy in heart disease prediction.

5.4 DATASET USED FOR THIS PROJECT

The dataset, to predict the diabetes in training and testing phase, used in my project is Indian Pima Diabetes Dataset**.**

Table 4: Indian Pima Diabetes Dataset

>>>	Pregnancies	Glucose	...	Age	Outcome
0	6	148	...	50	1
1	1	85	...	31	0
2	8	183	...	32	1
3	1	89	...	21	0
4	0	137	...	33	1
..	...	...	...	...	...
763	10	101	...	63	0
764	2	122	...	27	0
765	5	121	...	30	0
766	1	126	...	47	1
767	1	93	...	23	0

CHAPTER 6

SYSTEM REQUIREMENTS

6.1 HARDWARE REQUIREMENTS

- Processor : Intel processor 2.6.0 GHZ
- RAM : 2 GB
- Hard disk : 160 GB
- Compact Disk : 650 Mb
- Keyboard : Standard keyboard
- Monitor : 15 inch color monitor

6.2 SOFTWARE REQUIREMENTS

- Operating system : Windows OS
- Front End : Python
- Back End : MySQL SERVER
- IDLE : Python 2.7 IDLE

6.3 SOFTWARE DESCRIPTION

FROND END: PYTHON

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. In July 2018, Van Rossum stepped down as the leader in the language community. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of Python's other implementations. Python and CPython are managed by the non-profit Python Software Foundation. Rather than having all of its functionality built into its core, Python was designed to be highly extensible. This compact modularity has made it particularly popular as a means of adding programmable interfaces to existing applications. Van Rossum's vision of a small core language with a large standard library and easily extensible interpreter stemmed from his frustrations with ABC, which espoused the opposite approach. While offering choice in coding methodology, the Python philosophy rejects exuberant syntax (such as that of Perl) in favor of a simpler, less-cluttered grammar. As Alex Martello put it: "To describe something as 'clever' is not considered a compliment in the Python culture". Python's philosophy rejects the Perl "there is more than one way to do it" approach to language design in favor of "there should be one and preferably only one obvious way to do it".

Python's developers strive to avoid premature optimization, and reject patches to non-critical parts of CPython that would offer marginal increases in speed at the cost of clarity.[When speed is important, a Python programmer can move time-critical functions to extension modules written in languages such as C, or use PyPy, a just-in-time compiler. CPython is also available, which translates a Python script into C and makes direct C-level API calls into the Python interpreter. An important goal of Python's developers is keeping it fun to use. This is reflected in the language's name a tribute to the British comedy group Monty Python and in

occasionally playful approaches to tutorials and reference materials, such as examples that refer to spam and eggs (from a famous Monty Python sketch) instead of the standard for and bar.

A common neologism in the Python community is pythonic, which can have a wide range of meanings related to program style. To say that code is pythonic is to say that it uses Python idioms well, that it is natural or shows fluency in the language, that it conforms with Python's minimalist philosophy and emphasis on readability. In contrast, code that is difficult to understand or reads like a rough transcription from another programming language is called unpythonic. Users and admirers of Python, especially those considered knowledgeable or experienced, are often referred to as Pythonists, Pythonistas, and Pythoneers. Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

Python’s initial development was spearheaded by Guido van Rossum in the late 1980s. Today, it is developed by the Python Software Foundation. Because Python is a multi paradigm language, Python programmers can accomplish their tasks using different styles of programming:

object oriented, imperative, functional or reflective. Python can be used in Web development, numeric programming, game development, serial port access and more.

There are two attributes that make development time in Python faster than in other programming languages:

1. Python is an interpreted language, which precludes the need to compile code before executing a program because Python does the compilation in the background. Because Python is a high-level programming language, it abstracts many sophisticated details from the programming code. Python focuses so much on this abstraction that its code can be understood by most novice programmers.
2. Python code tends to be shorter than comparable codes. Although Python offers fast development times, it lags slightly in terms of execution time. Compared to fully compiling languages like C and C++, Python programs execute slower. Of course, with the processing speeds of computers these days, the speed differences are usually only observed in benchmarking tests, not in real-world operations. In most cases, Python is already included in Linux distributions and Mac OS X machines.

BACK END: MY SQL

MySQL is the world's most used open source relational database management system (RDBMS) as of 2008 that run as a server providing multi-user access to a number of databases. The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Oracle Corporation.

MySQL is a popular choice of database for use in web applications, and is a central component of the widely used LAMP open source web application software stack LAMP is an acronym for "Linux, Apache, MySQL, Perl/PHP/Python." Free-software-open source projects that require a full-featured database management system often use MySQL. For commercial use, several paid editions are available, and offer additional functionality. Applications which use MySQL databases include: TYPO3, Joomla, Word Press, phpBB, MyBB, Drupal and other

software built on the LAMP software stack. MySQL is also used in many high-profile, large-scale World Wide Web products, including Wikipedia, Google(though not for searches), ImagebookTwitter, Flickr, Nokia.com, and YouTube.

INTER IMAGES

MySQL is primarily an RDBMS and ships with no GUI tools to administer MySQL databases or manage data contained within the databases. Users may use the included command line tools, or use MySQL "front-ends", desktop software and web applications that create and manage MySQL databases, build database structures, back up data, inspect status, and work with data records. The official set of MySQL front-end tools, MySQL Workbench is actively developed by Oracle, and is freely available for use.

GRAPHICAL

The official MySQL Workbench is a free integrated environment developed by MySQL AB, that enables users to graphically administer MySQL databases and visually design database structures. MySQL Workbench replaces the previous package of software, MySQL GUI Tools. Similar to other third-party packages, but still considered the authoritative MySQL frontend, MySQL Workbench lets users manage database design & modeling, SQL development (replacing MySQL Query Browser) and Database administration (replacing MySQL Administrator).MySQL Workbench is available in two editions, the regular free and open source Community Edition which may be downloaded from the MySQL website, and the proprietary Standard Edition which extends and improves the feature set of the Community Edition.

CHAPTER 7

SYSTEM DESIGN

7.1 UML DIAGRAMS WITH DESCRIPTIONS

7.1.1 USE CASE DIAGRAM

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well.

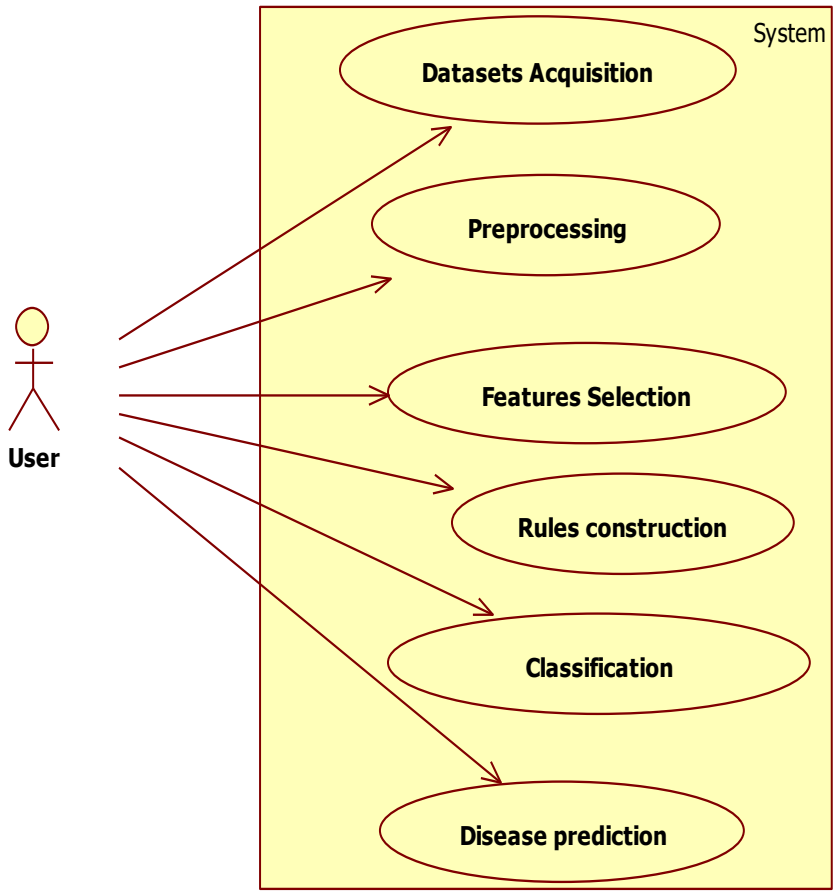

Figure 10 - Use Case Diagram

7.1.2 CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.

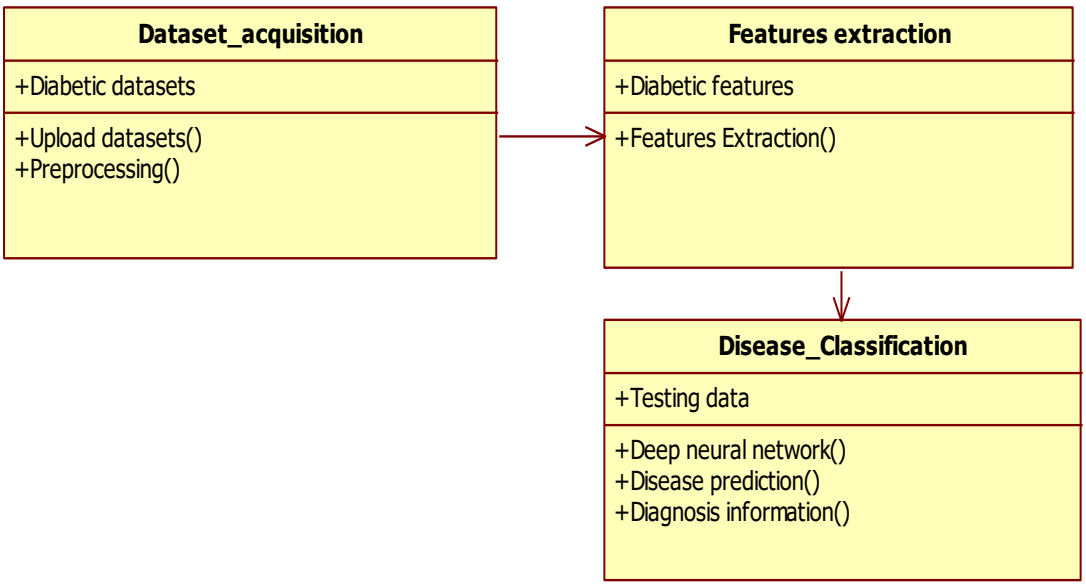

Figure 11 - Class Diagram

7.1.3 Sequence Diagram

A Sequence diagram is an interaction diagram that shows how objects operate with one another and in what order. It is a construct of a message sequence chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams or event scenarios

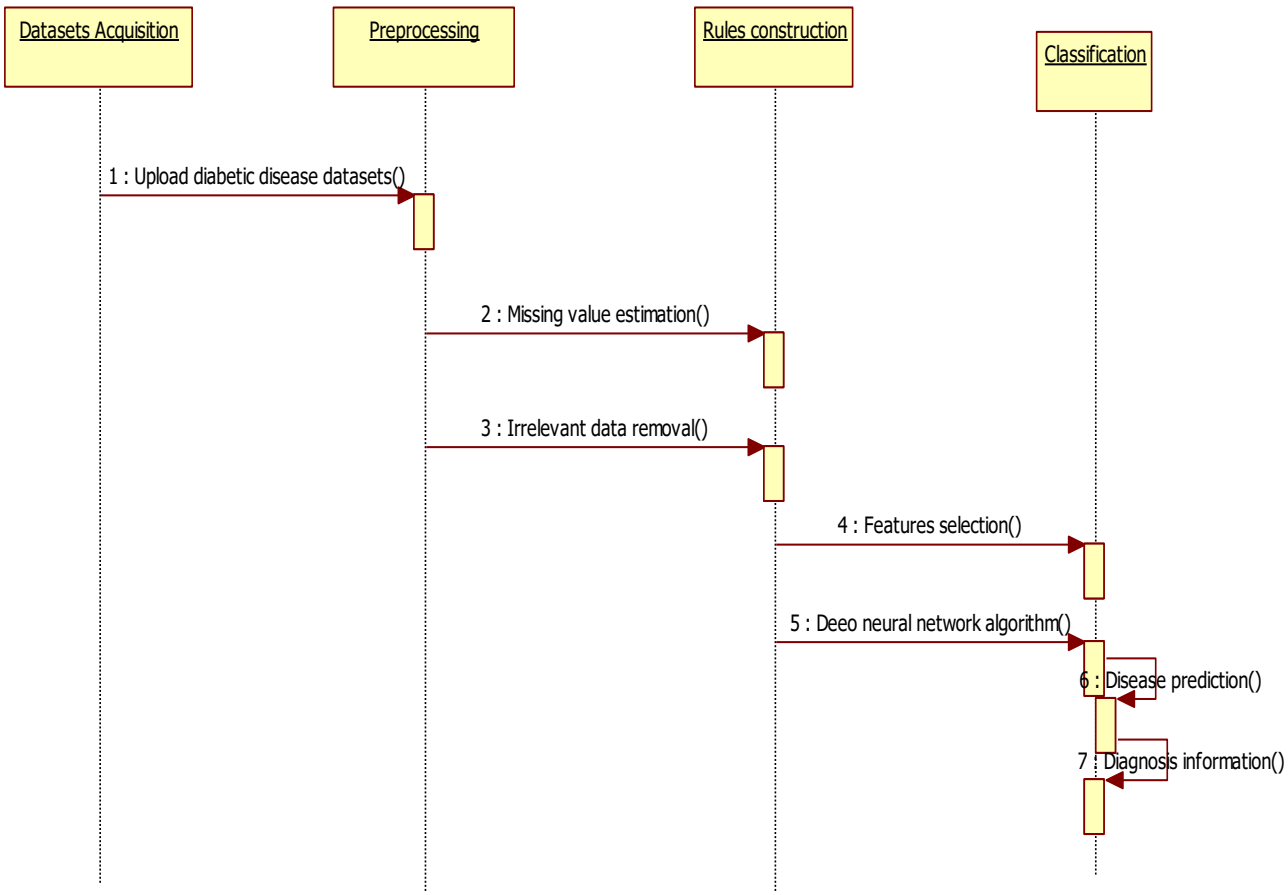

Figure 12 - Sequence Diagram

7.1.4 COLLABORATION DIAGRAM

A collaboration diagram, also called a communication diagram or interaction diagram, is an illustration of the relationships and interactions among software objects in the Unified Modeling Language (UML). The concept is more than a decade old although it has been refined as modeling paradigms have evolved

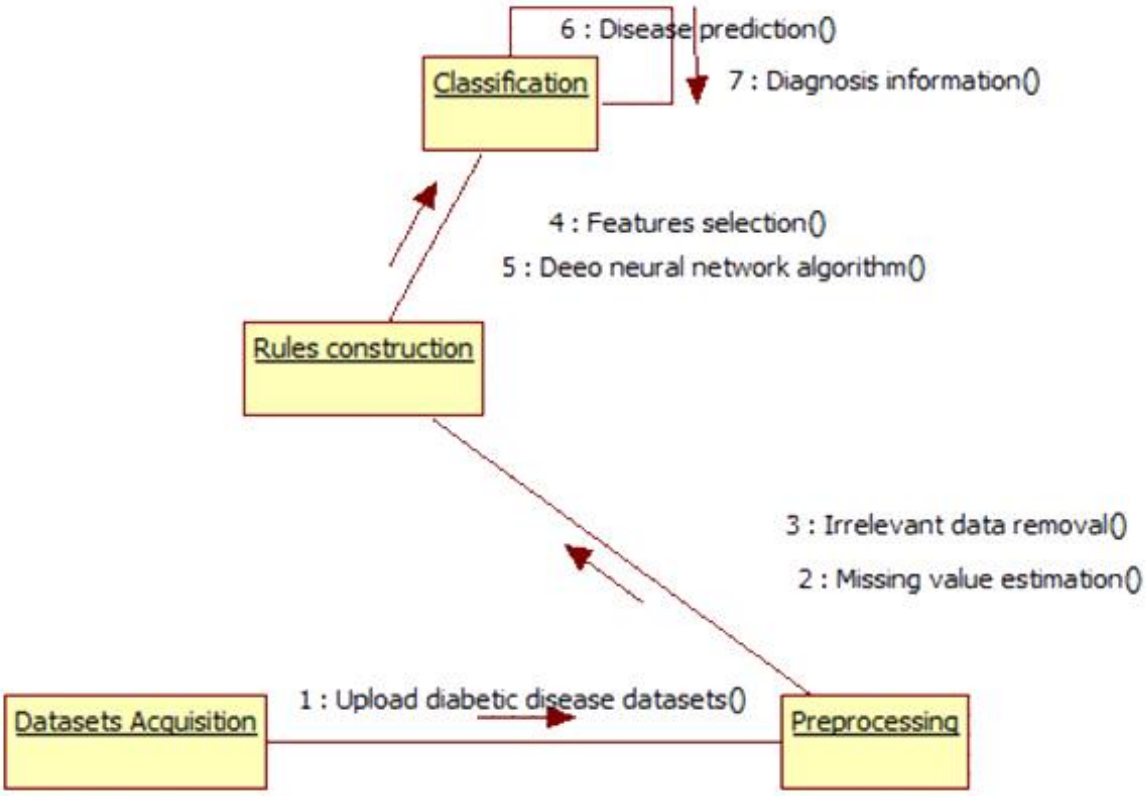

Figure 13 - Collaboration Diagram

7.1.5 ACTIVITY DIAGRAM

Activity diagram is another important diagram in UML to describe dynamic aspects of the system. Activity diagram is basically a flow chart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. So the control flow is drawn from one operation to another. This flow can be sequential, branched or concurrent. Activity diagrams deals with all type of flow control by using different elements like fork, join etc.

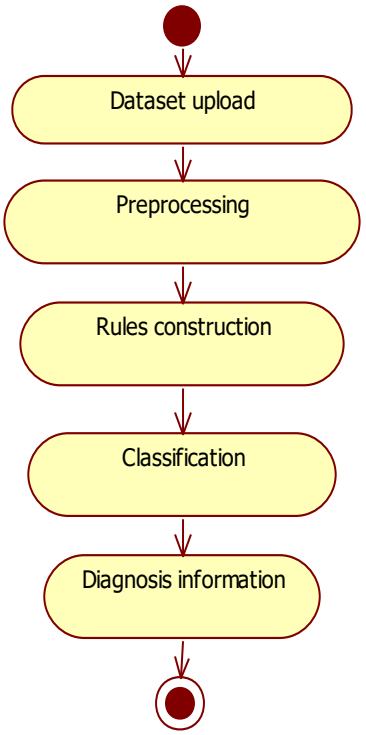

Figure 14 - Activity Diagram

CHAPTER 8

TESTING

8.1 SOFTWARE TESTING

Software testing is a method of assessing the functionality of a software program. There are many different types of software testing but the two main categories are dynamic testing and static testing. Dynamic testing is an assessment that is conducted while the program is executed; static testing, on the other hand, is an examination of the program's code and associated documentation. Dynamic and static methods are often used together.

Testing is a set activity that can be planned and conducted systematically. Testing begins at the module level and work towards the integration of entire computers based system. Nothing is complete without testing, as it is vital success of the system.

Testing Objectives:

There are several rules that can serve as testing objectives, they are

1. Testing is a process of executing a program with the intent of finding an error
2. A good test case is one that has high probability of finding an undiscovered error.
3. A successful test is one that uncovers an undiscovered error.

If testing is conducted successfully according to the objectives as stated above, it would uncover errors in the software. Also testing demonstrates that software functions appear to the working according to the specification, that performance requirements appear to have been met.

There are three ways to test a program

1. For Correctness
2. For Implementation efficiency
3. For Computational Complexity.

Tests for correctness are supposed to verify that a program does exactly what it was designed to do. This is much more difficult than it may at first appear, especially for large programs.

Tests for implementation efficiency attempt to find ways to make a correct program faster or use less storage. It is a code-refining process, which reexamines the implementation

phase of algorithm development. Tests for computational complexity amount to an experimental analysis of the complexity of an algorithm or an experimental comparison of two or more algorithms, which solve the same problem.

The data is entered in all forms separately and whenever an error occurred, it is corrected immediately. A quality team deputed by the management verified all the necessary documents and tested the Software while entering the data at all levels. The development process involves various types of testing. Each test type addresses a specific testing requirement. The most common types of testing involved in the development process are:

• Unit Test.
• System Test
• Integration Test
• Functional Test

8.2 TYPES OF TESTING

8.2.1 UNIT TESTING

The first test in the development process is the unit test. The source code is normally divided into modules, which in turn are divided into smaller units called units. These units have specific behavior. The test done on these units of code is called unit test. Unit test depends upon the language on which the project is developed. Unit tests ensure that each unique path of the project performs accurately to the documented specifications and contains clearly defined inputs and expected results. Functional and reliability testing in an Engineering environment and producing tests for the behavior of components (nodes and vertices) of a product is to ensure their correct behavior prior to system integration.

8.2.2 FUNCTIONAL TESTING

Functional test can be defined as testing two or more modules together with the intent of finding defects, demonstrating that defects are not present, verifying that the module performs its intended functions as stated in the specification and establishing confidence that a program does what it is supposed to do.

8.2.3 INTEGRATION TESTING

Testing the modules are combined and tested as a group. Modules are typically code modules, individual applications, source and destination applications on a network, etc. Integration Testing follows unit testing and precedes system testing. Testing after the product is code complete. Betas are often widely distributed or even distributed to the public at large in hopes that they will buy the final product when it is released.

8.2.4 WHITE BOX TESTING

Testing based on an analysis of internal workings and structure of a piece of software. This testing can be done sing the percentage value of load and energy. The tester should know what exactly is done in the internal program. This includes techniques such as Branch Testing and Path Testing and also known as Structural Testing and Glass Box Testing.

8.2.5 BLACK BOX TESTING

Testing without knowledge of the internal workings of the item is being tested. Tests are usually functional. This testing can be done by the user who has no knowledge of how the shortest path is found.

8.3 TEST RESULT

All the test cases are mentioned above passed successfully. No defects encountered.

CHAPTER 9

RESULTS / SCREENSHOTS FOR OUTPUT

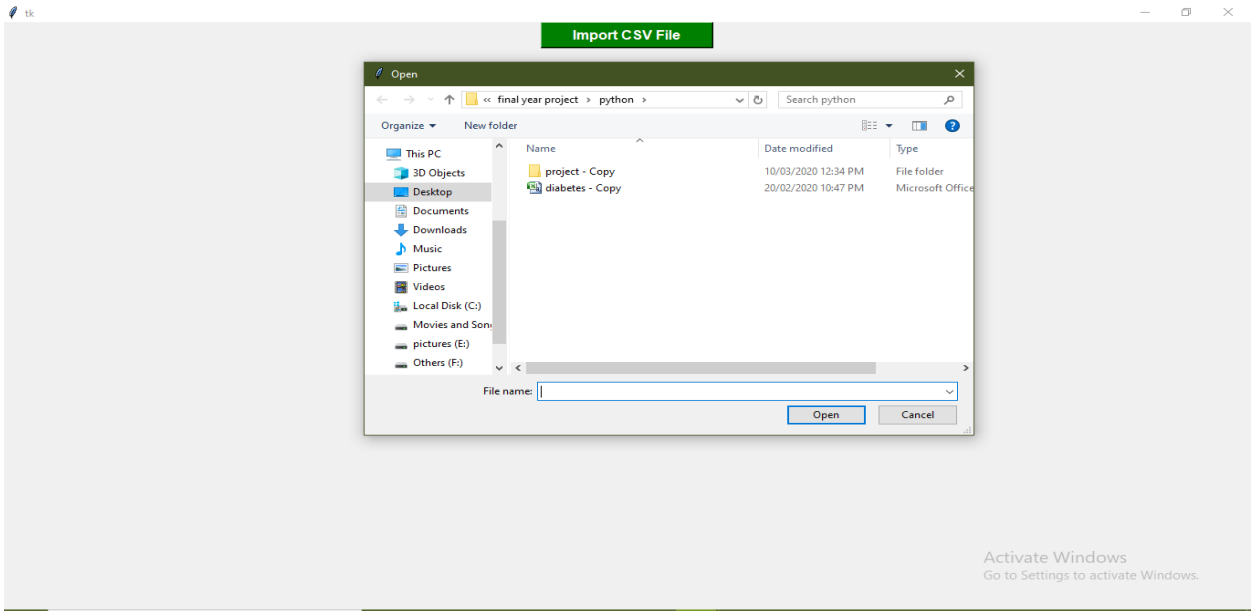

Figure 9.1 – Module 1 Datasets Acquisition

Uploading the dataset in the program for the further process from the system

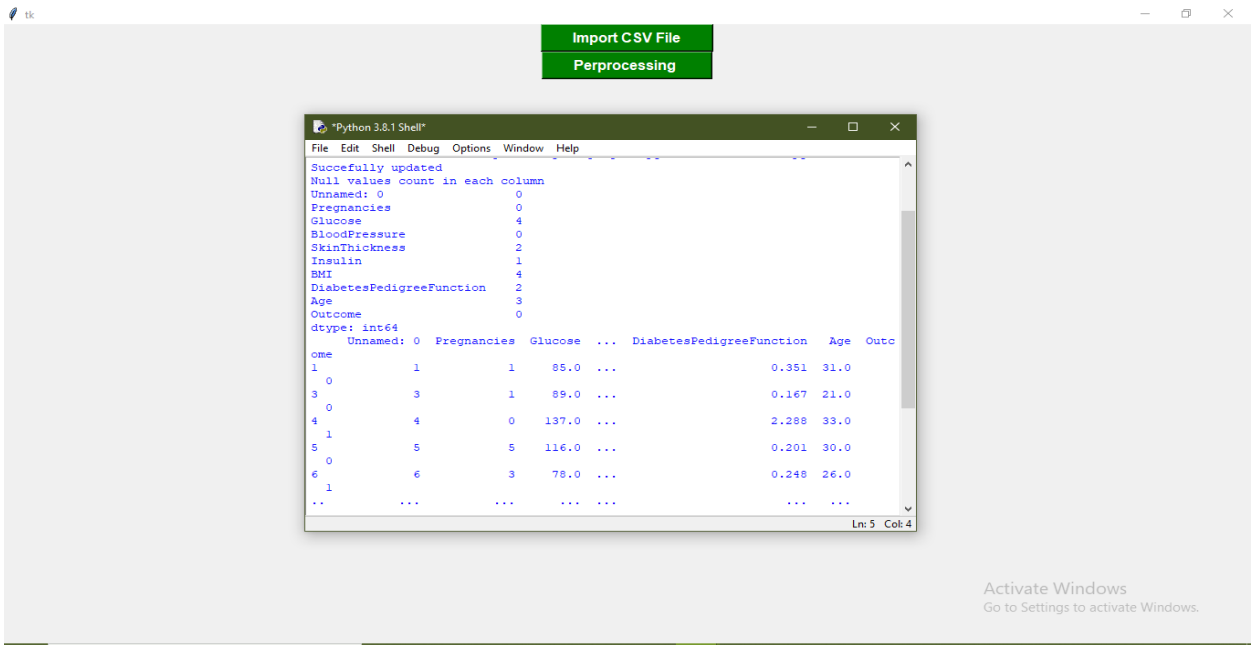

Figure 9.2 – Module 2 Preprocessing

The uploaded dataset is preprocessed for removing the unwanted data or missing values

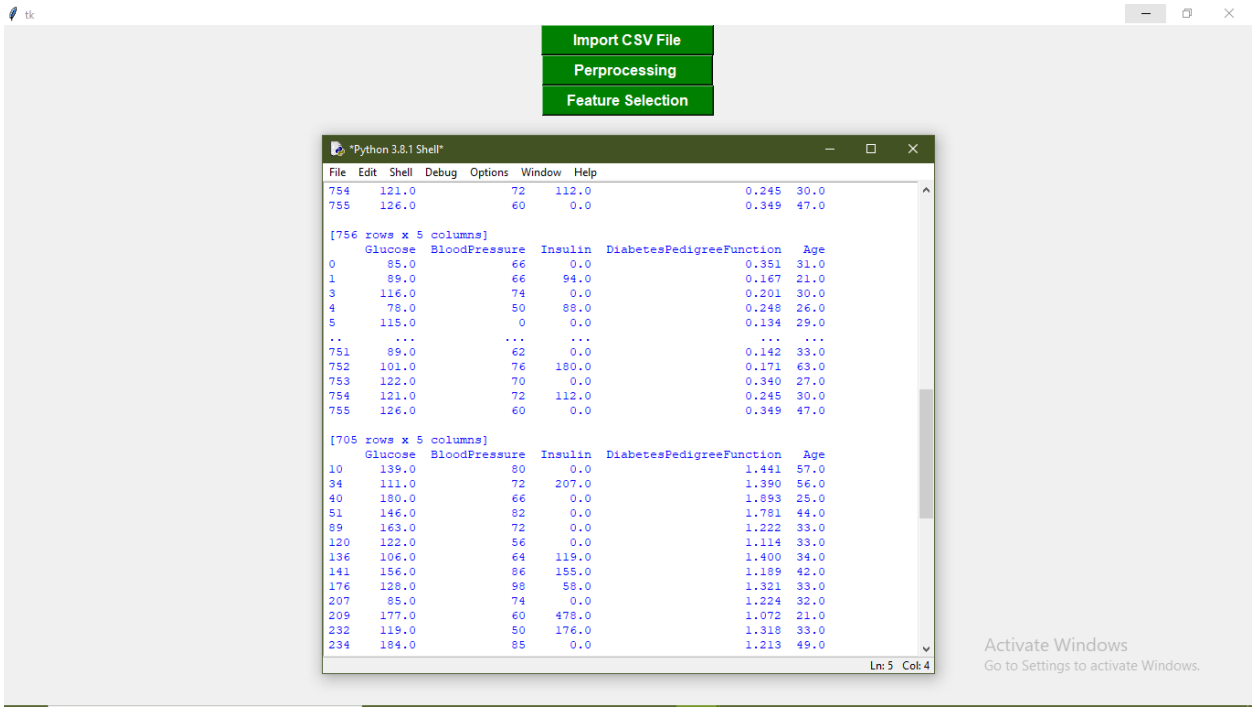

Figure 9.3 – Module 3 Feature Selection

Selecting the data from the preprocessed datasets to give the accurate results

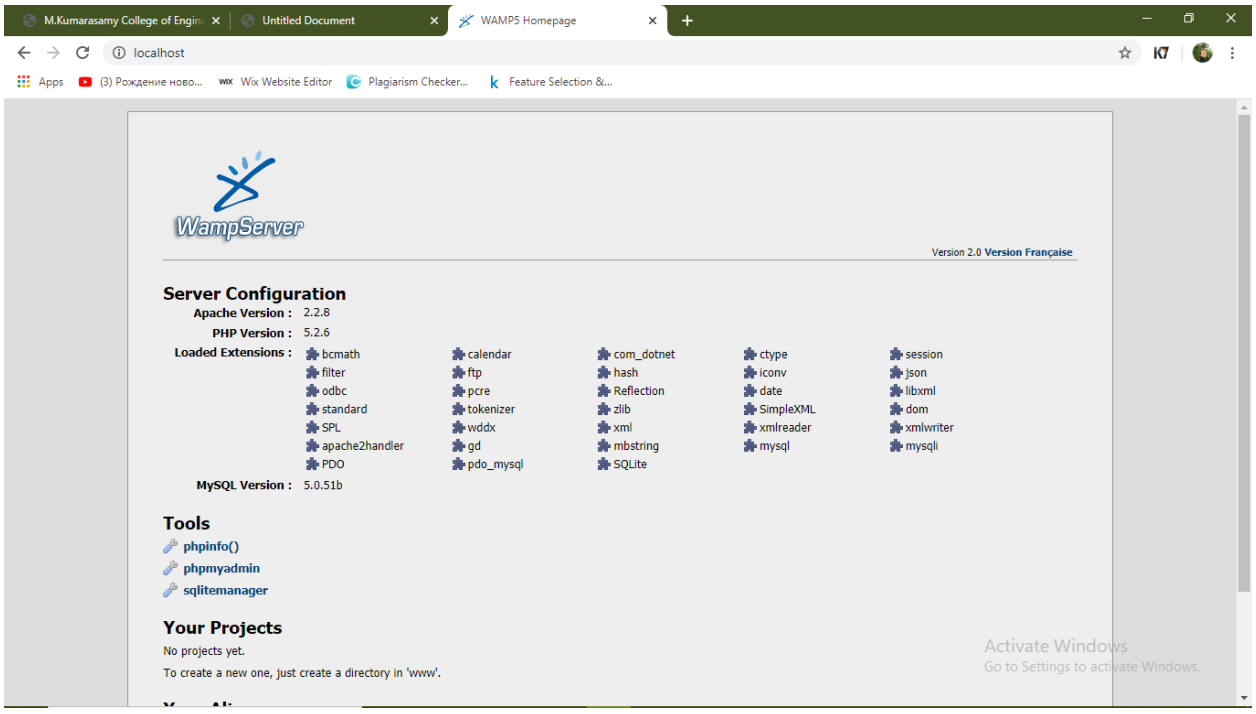

Figure 9.4 – Wamp Server Local Host - Mysql

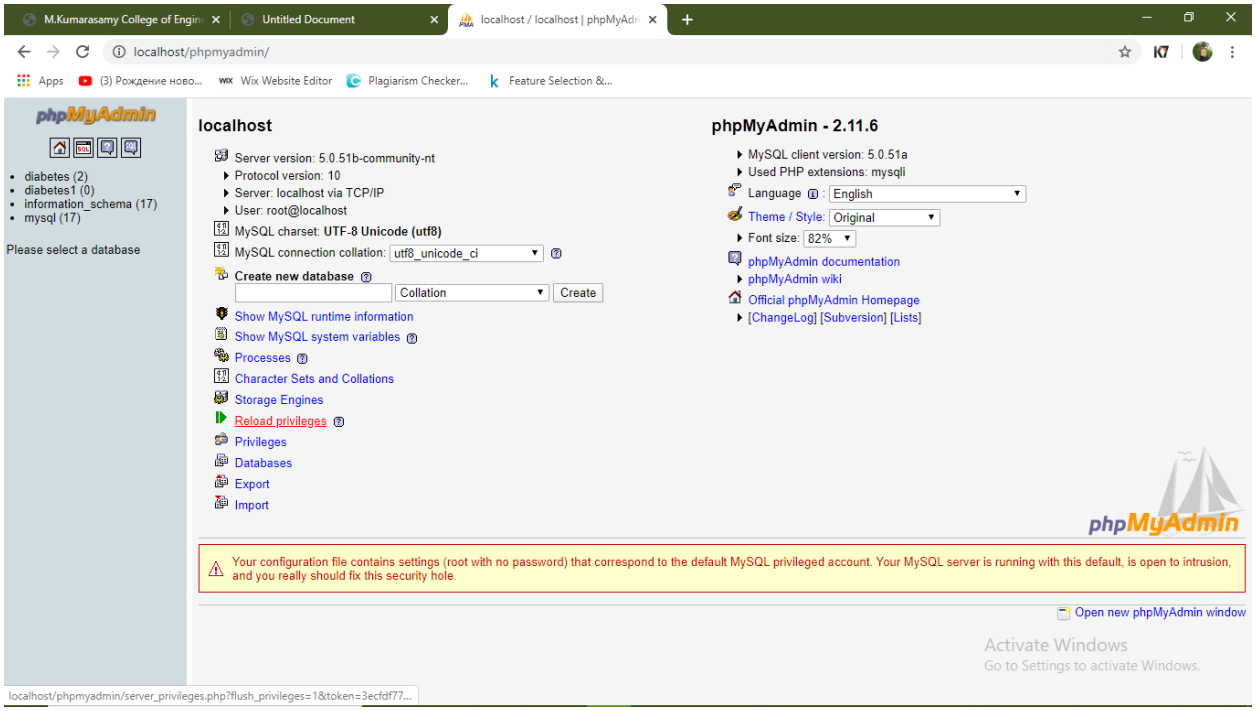

Figure 9.5 – Database Creating In Mysql Server (Wamp Server)

Creating a database in the wamp server to fetch the data from the database

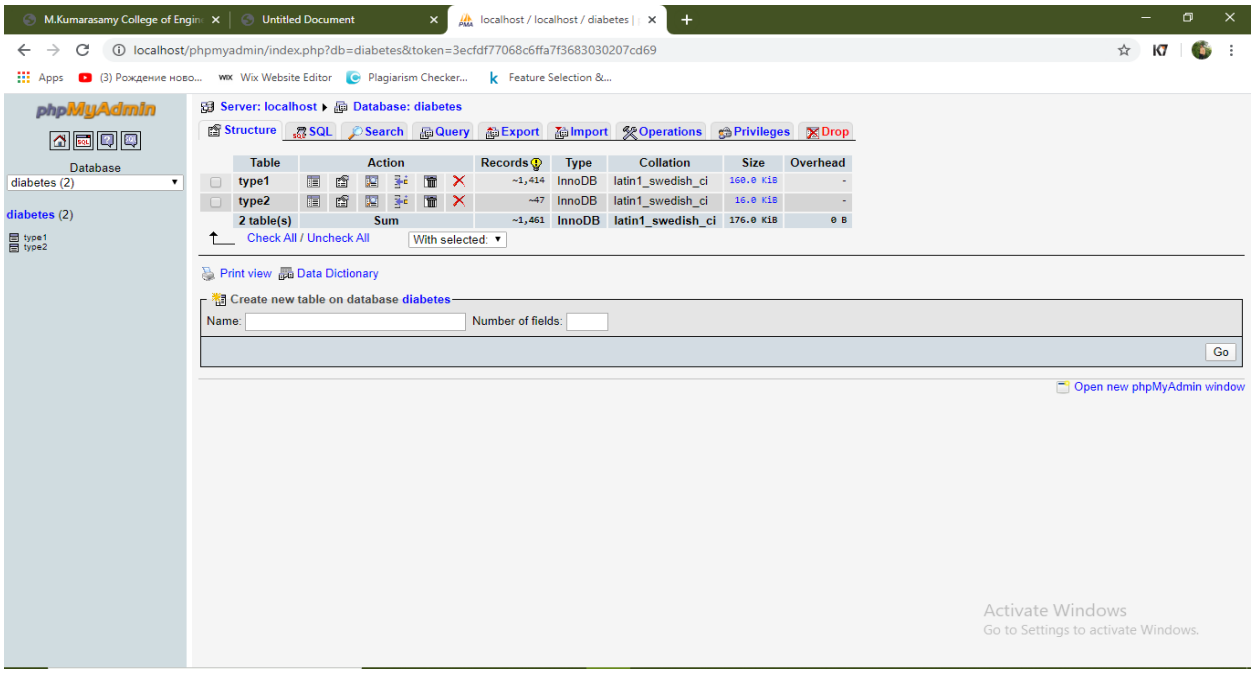

Figure 9.6 – Table Creation In Wamp Server

Table should be created for the extracted attributes from the feature selection module

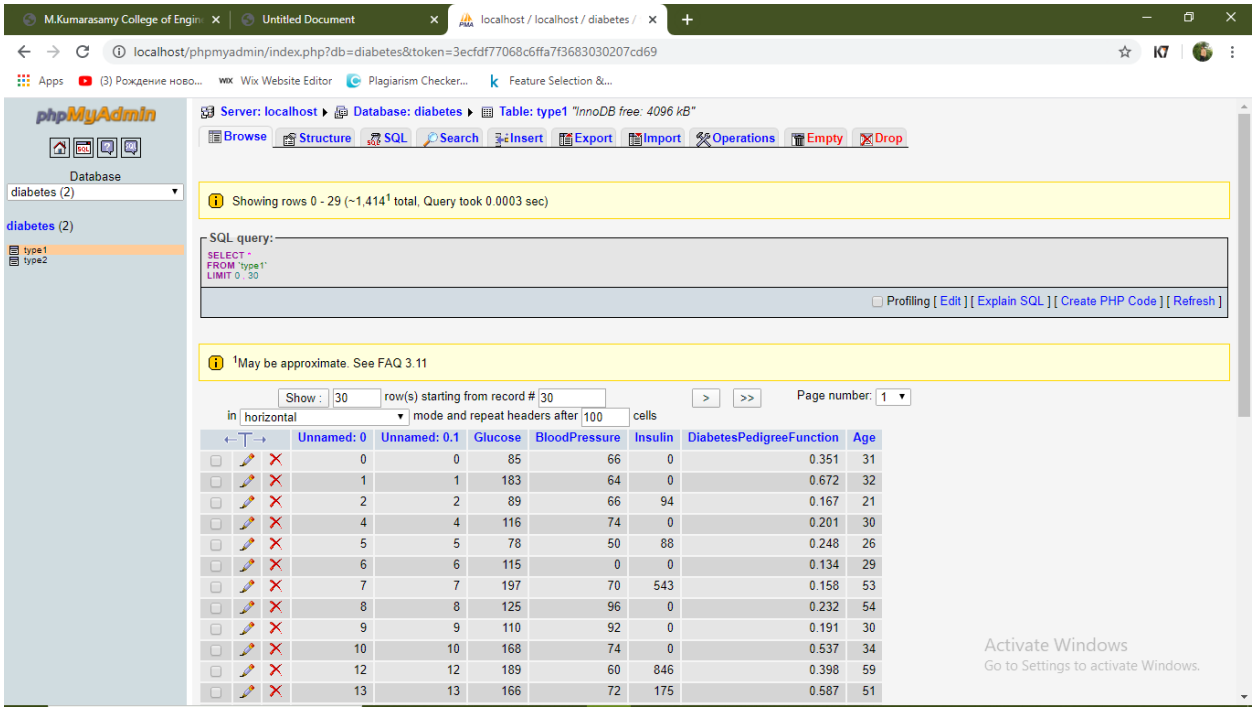

Figure 9.7 – Data Uploaded In The Server From The Module 3

The extracted data from the dataset is uploaded in the database

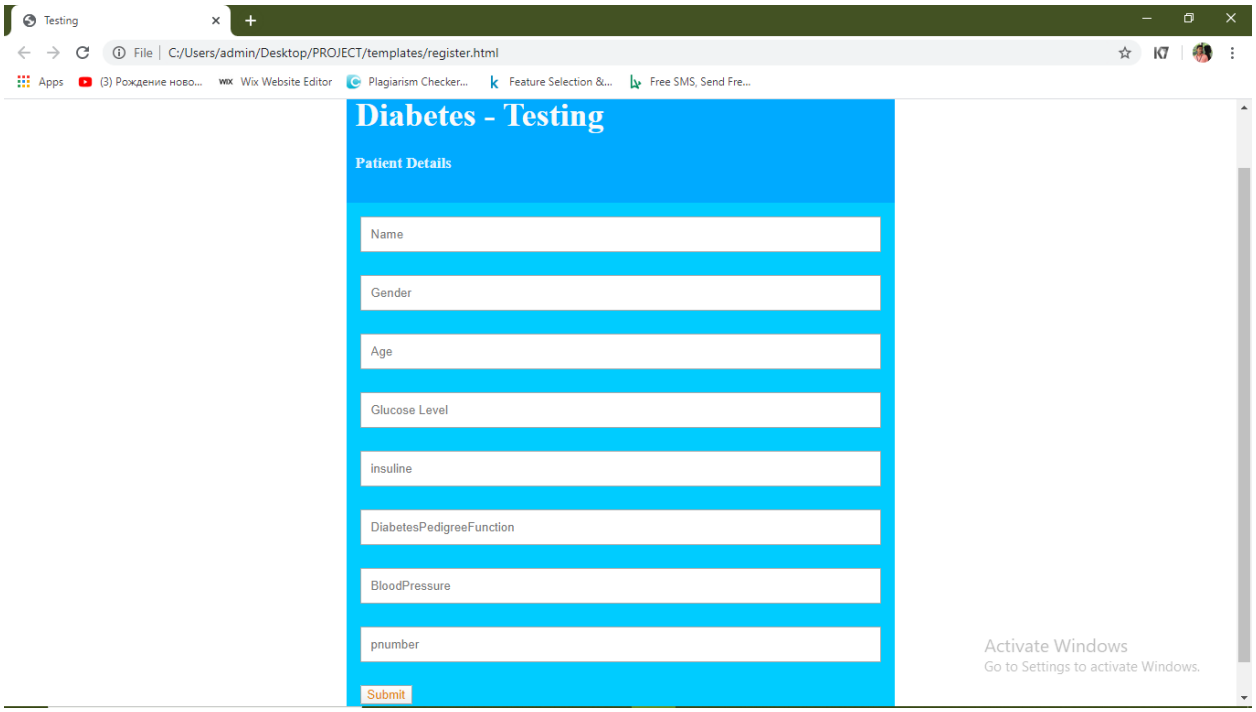

Figure 9.8 – User Input Form

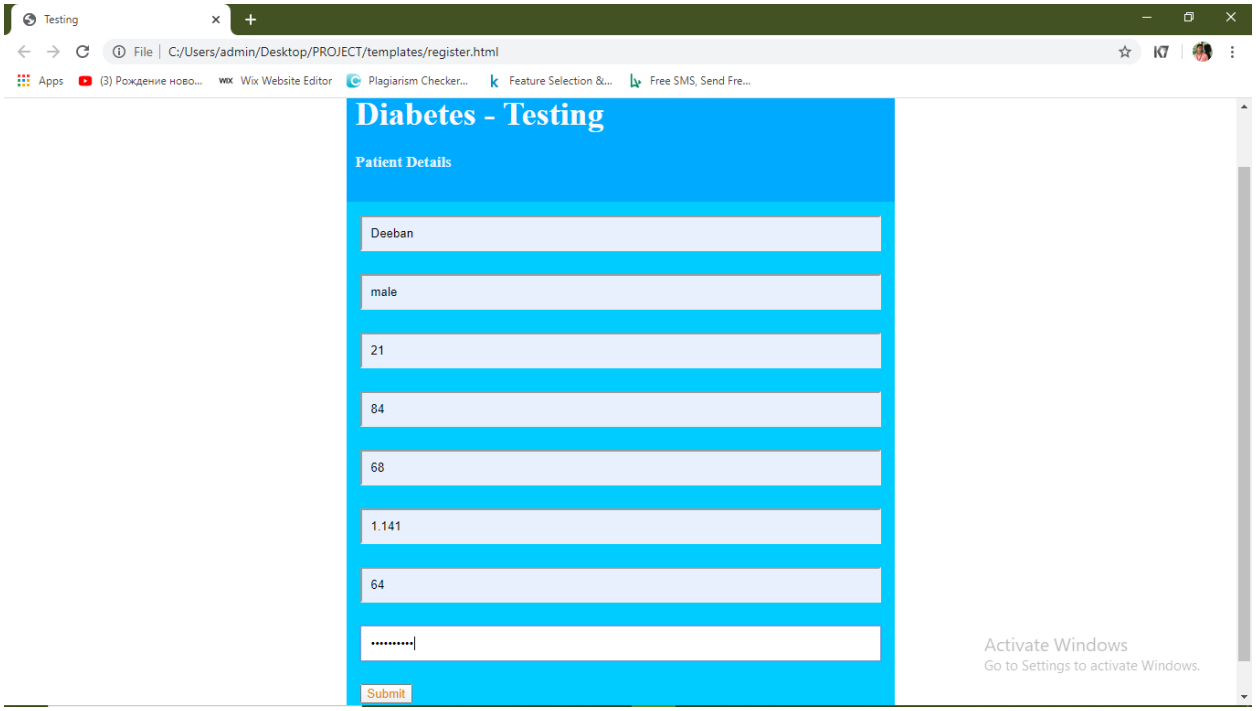

Figure 9.9 – Data Entry By The User To Check Diabetes

It is used for gathering details from the patients to check that they having diabetes or not

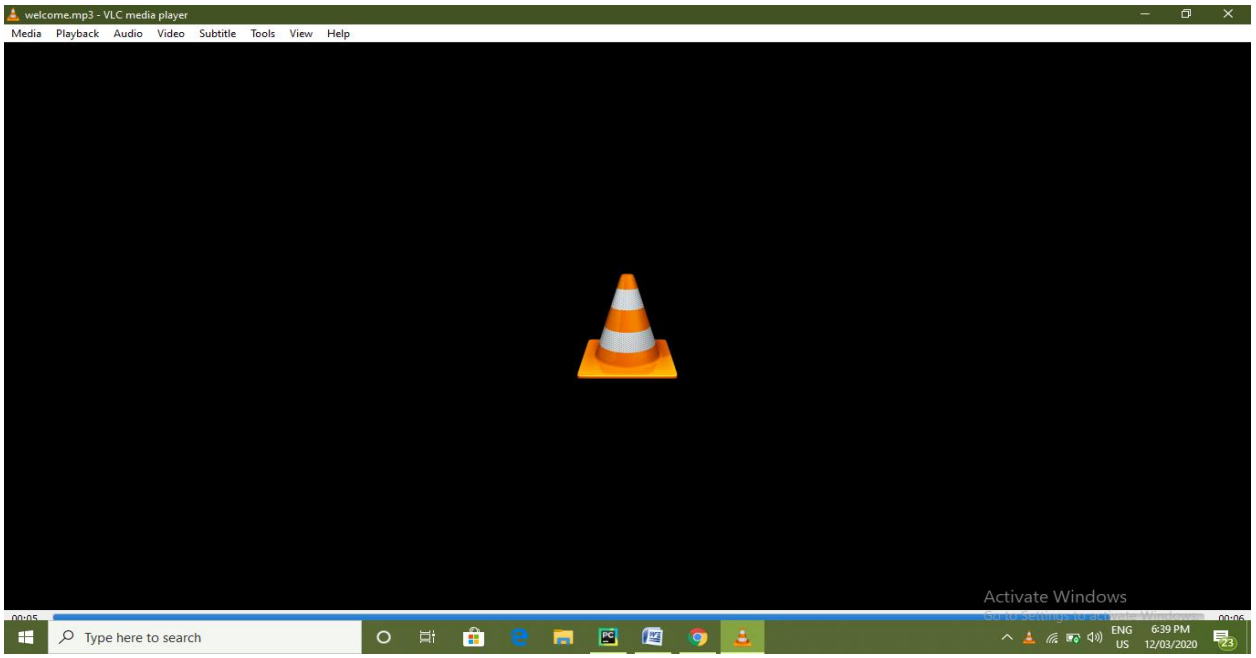

Figure 9.10 – Voice Message In The System

A voice message will be played in the system after the results came

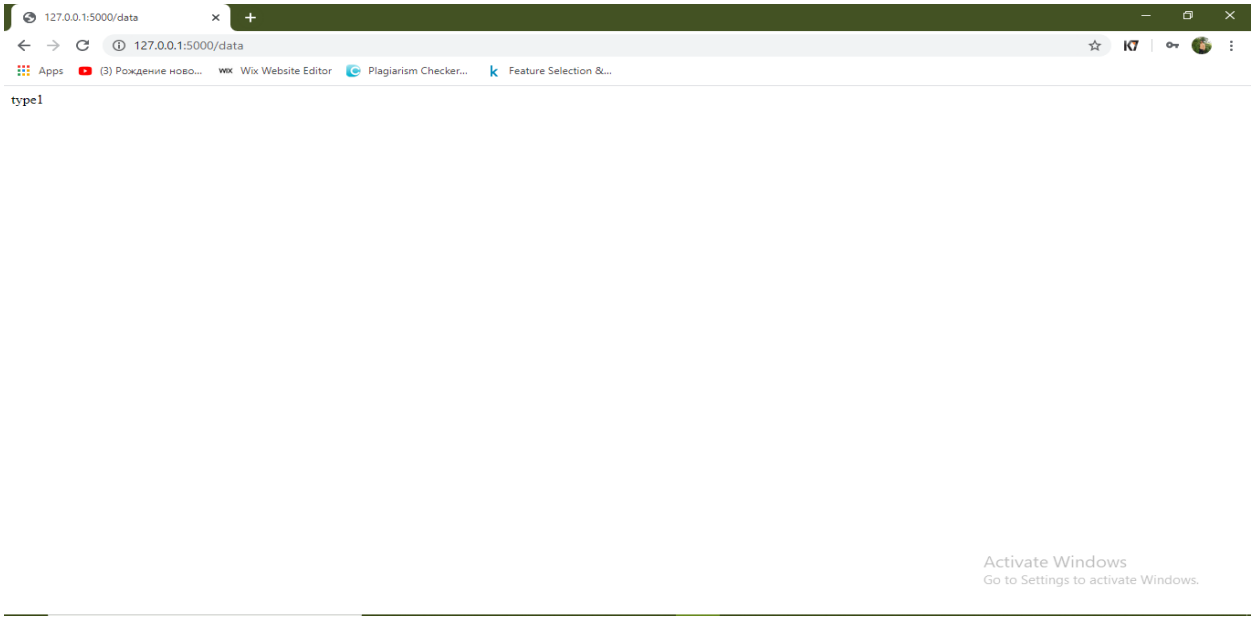

Figure 9.11 - Text Message Viewed In The System

The results will be displayed in the system with respect to the voice message

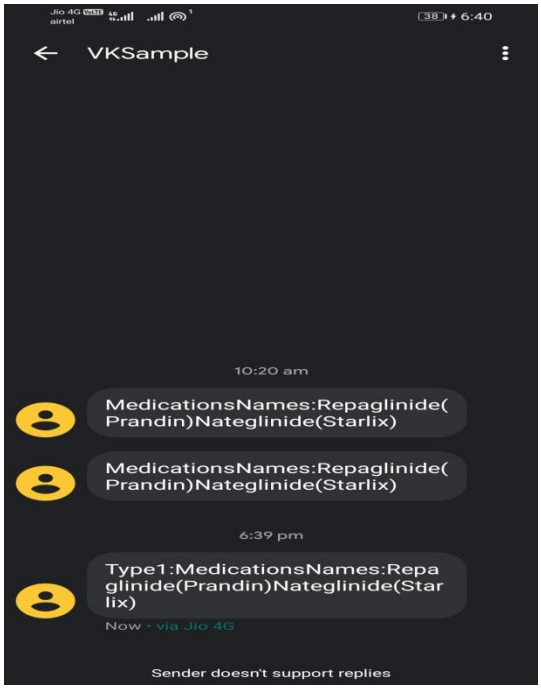

Figure 9.12 - SMS Text To The Patient Mobile

At last the results will be shared to the registered or given mobile number of the patients with the tablets they need to take for their diabetes disease

CHAPTER 10

CONCLUSION AND FUTURE WORK

10.1 CONCLUSION

In this project the problem of constraining and summarizing different algorithms of data mining used in the field of medical prediction are discussed. The focus is on using different algorithms and combinations of several target attributes for intelligent and effective diabetic disease prediction using data mining. Data mining technology provides an important means for extracting valuable medical rules hidden in medical data and acts as an important role in disease prediction and clinical diagnosis. There is an increasing interest in using classification to identify disease which is present or not. In the current study, have demonstrated, using a large sample of patients hospitalized with classification. Classification algorithm is very sensitive to noisy data. If any noisy data is present then it causes very serious problems regarding to the processing power of classification. It not only slows down the task of classification algorithm but also degrades its performance. Hence, before applying classification algorithm it must be necessary to remove all those attributes from datasets who later on acts as noisy attributes. In this project work, we can implement preprocessing steps and implemented the classification rule algorithms namely deep neural network algorithm are used for classifying datasets which are uploaded by user. By analyzing the experimental results it is observed that the deep learning technique has yields better result than other techniques.

10.2 FUTURE ENHANCEMENT

In future we tend to improve efficiency of performance by applying other data mining techniques and algorithms.

APPENDIX 1 – PYTHON PROGRAM

#MODULE 1 TO 3

```
import tkinter as tk
from tkinter import filedialog
import pandas as pd
import csv
def getCSV ():
    global df
    #asking to select the file - filedialog
    import_file_path = filedialog.askopenfilename()
    df = pd.read_csv (import_file_path)
    print(df)
    df.to_csv('Dataset.csv')
    #after read the csv file put it in another file named dataset in csv format
    print(import_file_path)
    print("Succefully updated")
def PreProcess():
    df = pd.read_csv("Dataset.csv")
    #find the missing values count in the dataset -attribute wise
    D=df.isnull().sum()
    print('Null values count in each column')
    print(D)
    #removing the missing data by deleting the row
    df = df.dropna()
    df.to_csv('Preprocessed_dataset.csv')
```

```
    print(df)
def Feature_Selection():
    df = pd.read_csv("Preprocessed_dataset.csv", usecols =
['Age','Glucose','BloodPressure','Insulin','DiabetesPedigreeFunction'])
    print(df)
    df.to_csv('features_selection.csv')
    #extracted the particular details or attributes for prediction and stored in csv file
    df = pd.read_csv("features_selection.csv",usecols =
['Age','Glucose','BloodPressure','Insulin','DiabetesPedigreeFunction'])
    Type1 = df[(df.DiabetesPedigreeFunction >= 0) & (df.DiabetesPedigreeFunction <=1)]
    print(Type1)
    Type1.to_csv('Type1.csv')
    df = pd.read_csv("features_selection.csv",usecols =
['Age','Glucose','BloodPressure','Insulin','DiabetesPedigreeFunction'])
    Type2=df[(df.DiabetesPedigreeFunction >= 1) & (df.DiabetesPedigreeFunction <=2)]
    print(Type2)
    Type2.to_csv('Type2.csv')
    # read CSV file
     column_names = ['slno','Glucose', 'BloodPressure','Insulin','DiabetesPedigreeFunction','Age']
     df = pd.read_csv('Type1.csv', header = None, names = column_names)
     print(df)
     df = pd.read_csv('Type1.csv', header = 0)
     print(df)
     engine = create_engine('mysql://root:@localhost/diabetes')
     with engine.connect() as conn, conn.begin():
         df.to_sql('type1', conn,schema='online', if_exists='append', index=False)
```

```
    # read CSV file
    column_names1 = ['slno','Glucose', 'BloodPressure','Insulin','DiabetesPedigreeFunction','Age']
    df1 = pd.read_csv('Type2.csv', header = None, names = column_names)
    print(df1)
    df1 = pd.read_csv('Type2.csv', header = 0)
    print(df1)
    engine = create_engine('mysql://root:@localhost/Diabetes')
    with engine.connect() as conn, conn.begin():
        df1.to_sql('type2', conn, schema='online',if_exists='append', index=False)
#creating buttons
B0 = tk.Button(text="       Import CSV File       ", command = getCSV, bg='green', fg='white', font=('timesnewroman', 12, 'bold'))
B0.pack()
B1 = tk.Button(text="       Perprocessing        ", command = PreProcess , bg='green', fg='white', font=('timesnewroman', 12, 'bold'))
B1.pack()
B2 = tk.Button(text ="     Feature Selection     ", command = Feature_Selection, bg='green', fg='white', font=('helvetica', 12, 'bold'))
B2.pack()
```

#MODULE 4 AND 5

```
import pyttsx3 as pyttsx3
from flask import Flask, request, json, Response, redirect, url_for, render_template
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import mean_absolute_error
from wtforms import Form, BooleanField, StringField, PasswordField, IntegerField, FloatField, validators
import pandas as pd
```

```
import mysql.connector
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
import urllib
import urllib.request
import urllib.parse
# to speech conversion
from gtts import gTTS
# This module is imported so that we can
# play the converted audio
import os
import requests
import json
app = Flask(__name__)
# rs 2 = 14.6305
print("Calculate accuracy")
def getMae():
    diabetes = pd.read_csv('diabetes.csv')
    print(diabetes)
    cnn = None
    X_train, X_test, y_train, y_test = train_test_split(diabetes.loc[:, diabetes.columns != 
'Outcome'],
                                                        diabetes['Outcome'], stratify=diabetes['Outcome'],
                                                        random_state=70)
```

```
    """neighbors_settings = range(1, 11)
    for n_neighbors in neighbors_settings:
        # build the model
        cnn = KNeighborsClassifier(n_neighbors=n_neighbors)
        cnn.fit(X_train, y_train)"""
    # Make model
    Jarvis = GradientBoostingClassifier(random_state=6, min_samples_split=2,
max_leaf_nodes=15, max_features=7)
    Jarvis.fit(X_train, y_train)
    # Predict
    prediction = Jarvis.predict(X_test)
    answers = y_test
    # Calculate mean absolute error
    mae = mean_absolute_error(answers, prediction)
    return mae
    print(getMae())
def sendPostRequest(reqUrl, apiKey, secretKey, useType, phoneNo, senderId, textMessage):
  req_params = {'apikey':apiKey,'secret':secretKey,'usetype':useType,'phone':
phoneNo,'message':textMessage,'senderid':senderId}
  return requests.post(reqUrl, req_params)
@app.route('/')
def home():
    return render_template("register.html")
@app.route('/data', methods=['GET', 'POST'])
def data():
    global cnn
```

```
    if request.method == 'POST':
        name = request.form['name']
        gender = request.form['gender']
        age = request.form['age']
        Glucose = request.form['Glucose']
        insuline = request.form['insuline']
        DiabetesPedigreeFunction = request.form['DiabetesPedigreeFunction']
        BloodPressure = request.form['BloodPressure']
        pnumber = request.form['pnumber']
        data = getMae()
        print(name)
        print(data)
        if (float(data) < 0.3066):
            conn = mysql.connector.connect(user='root', password='', host='localhost',
database='diabetes')
            cursor = conn.cursor()
            cursor.execute("select * from type1 where
DiabetesPedigreeFunction>="'+DiabetesPedigreeFunction+"'")
            data = cursor.fetchone()
            print(data)
            if data is None:
                conn = mysql.connector.connect(user='root', password='', host='localhost',
database='diabetes')
                cursor = conn.cursor()
                cursor.execute("select * from type2 where Glucose<="' + Glucose + '" and
BloodPressure<="' + BloodPressure + '" and Insulin<="' + insuline + '" and
DiabetesPedigreeFunction<="' + DiabetesPedigreeFunction + "'")
```

```
            data = cursor.fetchone()
            print(data)
            data2 =
"Hi,"+name+",Type2:Metformin,Or,SulfonyLureas,Or,Meglitinides,Or,DppFourInhibitors"
            ph = pnumber
            msg = data2
            URL = 'https://www.sms4india.com/api/v1/sendCampaign'
            response = sendPostRequest(URL, 'R2YVAI5K6HS43M7MRZ35F9ZYB9YIMYW0',
'UFE9MANL6YYBY634', 'stage', ph,'active-sender-id', msg)
            print(response.text)
            mytext = data2
            language = 'en'
            myobj = gTTS(text=mytext, lang=language, slow=True)
            myobj.save("welcome.mp3")
            os.system("welcome.mp3")
            return data2
        else:
            print(data)
            data1 =
"Hi,"+name+",Type1:CereStarCapsule,Or,GoodHealthSugarKnockerVegCapsule,Or,OrganicInd
iaSugarBalanceSixtyCapsule"
            ph = pnumber
            msg = data1
            URL = 'https://www.sms4india.com/api/v1/sendCampaign'
            response = sendPostRequest(URL, 'R2YVAI5K6HS43M7MRZ35F9ZYB9YIMYW0',
'UFE9MANL6YYBY634', 'stage', ph, 'active-sender-id', msg )
            print (response.text)
```

```
            mytext = data1
            language = 'en'
            myobj = gTTS(text=mytext, lang=language, slow=True)
            myobj.save("welcome.mp3")
            os.system("welcome.mp3")
            return data1
if __name__ == '__main__':
    app.run(debug=True, use_reloader=True)
```

APPENDIX 2 – USER INPUT – WEB PAGE

```
<html>
<head>
<title>Testing</title>
</head>
<style>
html {
  height: 100%;
}
body {
  display: flex;
  justify-content: center;
  align-items: center;
  background-image: url('https://www.pharmanewsdaily.com/wp-content/uploads/2019/10/6ccb60c04aed1553999dc44925969df6.gif');
background-repeat: no-repeat;
  background-attachment: fixed;
```

```
  background-size: 100% 100%;
}
/* Login Form */
.Testing {
  align-items: center;
  width: 600px;
  background: #00CCFF;
}
.Testing > header {
  padding: 10px;
  background: #00AAFF;
  font-size: 1.5rem;
  color: #FAFAFA;
}
.Testing > header h2 {
  margin: 50px 0 10px 0;
}
.Testing > header h4 {
  font-size: 0.7em;
  color: rgba(255, 255, 255, 0.9);
}
/* Form */
.Testing-form {
  padding: 15px;
}
```

```
/* Inputs */
.Testing-input {
  width: 100%;
  padding: 10px 10px;
  margin: 0 10px 25px 0;
  }
/* Submit Button */
.Testing-button {
  color: #E37F00;
}
.style3 {color: #009688}
</style>
<body>
<div class="Testing">
    <header>
      <h2>Diabetes - Testing </h2>
      <h4> Patient Details</h4>
    </header>
    <form class="Testing-form" action="/data" method="post" name="form">
      <input type="text" class="Testing-input" placeholder="Name" name="name" required/>
      <input type="text" class="Testing-input" placeholder="Gender" name="gender" required />
      <input type="text" class="Testing-input" placeholder="Age" name="age"  required/>
     <input type="text" class="Testing-input" placeholder="Glucose Level" name="Glucose"
required />
     <input type="text" class="Testing-input" placeholder="insuline" name="insuline"
required/>
```

```
<input type="text" class="Testing-input" placeholder="DiabetesPedigreeFunction" name="DiabetesPedigreeFunction"  required/>

<input type="text" class="Testing-input" placeholder="BloodPressure" name="BloodPressure" required/>

<input type="password" class="Testing-input" placeholder="pnumber" name="pnumber" required/>

<button type="submit" class="Testing-button">Submit</button>

</form>

</div>

</body>

</html>
```

REFERENCES

1. Lahiru Liyanapathirana, Machine **learning workflow in diabetes in the article of towards data science**, A Medium publication sharing concepts, ideas, and codes, Feb-26, 2018.
2. Norbert Freinkel, Diabetes Care,**Classification and diagnosis of diabetes: Standards of medical care in diabetes**, The Journal of Clinical and Applied Research and Education, Volume 41, Supplement 1, January 2018.
3. Lili Yuen, Pouya Saeedi, Musarrat Riaz, **Projection of prevalence of Hyperglycaemia in pregnancy in 2019 and beyond**: results from the International Diabetes Federation Diabetes Atlas ,9th edition: Diabetes research and clinical practice, Volume 157, 107841, November 01, 2019.
4. Eu Jeong Ku, Dong-Hwa Lee, Hyun Jeong Jeon, Tae Keun Oh, **Empagliflozin versus Dapagliflozin in patients with type-2 diabetes inadequately controlled with metformin, glimepiride and dipeptidyl peptide 4 inhibitors,** Volume 151, P65-73, May 01, 2019.
5. Ayesha A. Motala, Jonathan E. Shaw, **Global and regional diabetes prevalence estimates for 2019 and projection for 2030 and 2045**: Results from the International Diabetes Federation Diabetes Atlas, 9th edition, Volume 157, 107843, November 01, 2019.

6. P.Santhi, R.Vikram, **Implementation Of Classification System Using Density Clustering Based Gray Level Co Occurrence Matrix (DGLCM) For Green Bio Technology**, International Journal of Pure and Applied Mathematics,Vol.118, No.8,PP. 191-195,2018.
7. S. Thilagamani, N.Shanthi, **A Novel Recursive Clustering Algorithm for Image Oversegmentation**, European Journal of Scientific Research, Vol.52, No.3, pp.430-436, 2011.
8. Mohammed Akour1 , Osama Al Qasem2 , Hiba Alsghaier3 , Khalid Al-Radaideh4, **The Effectiveness of Using Deep Learning Algorithms in Predicting Daily Activities** , of Advanced Trends in Computer Science and Engineering, pp. 2231- 2235, Volume-8, No.5, September – October 2019.
9. Ahmad al-Qerem1 Arwa Alahmad 2 , **Human Body Poses Recognition Using Neural Networks with Data Augmentation**, Advanced Trends in Computer Science and Engineering, pp. 2117 – 2120,Volume-8, No.5, September – October 2019, ISSN 2278-3091.
10. P.Santhi, G.Mahalakshmi, **Classification of Magnetic Resonance Images Using Eight Directions Gray Level Co-Occurrence Matrix Based Feature Extraction**, International Journal of Engineering and Advanced Technology, ISSN: 2249-8958, Volume-8 Issue-4, April 2019.
11. Dr. Navneet Malik, Nilesh N Wani, Jimmy Singla, **Complications of Sight Threatening Diabetic Retinopathy,** International Journal of Engineering and Advanced Technology, ISSN 2278-3091, Volume 8, No 4, July – August 2019.
12. Choudhury, Ambika, and Deepak Gupta, **A survey on medical diagnosis of diabetes using machine learning techniques,** Recent Developments in Machine Learning and Data Analytics, Springer, Singapore, 2019, page 67-78.
13. Birjais, Roshan, et al. **Prediction and diagnosis of future diabetes risk: a machine learning approach,** SN Applied Sciences 1.9 (2019): 1112.
14. Vehí, Josep, et al. **Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning,** Health informatics journal (2019): 1460458219850682.
15. Zafar, Faizan, et al. **Predictive Analytics in Healthcare for Diabetes Prediction,** Proceedings of the 2019 9th International Conference on Biomedical Engineering and Technology. 2019.

16. Plis, Kevin, et al. **A machine learning approach to predicting blood glucose levels for diabetes management,** Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.

17. Lashari, Saima Anwar, and Rosziati Ibrahim, **Comparative analysis of data mining techniques for medical data classification,** 4th International Conference on Computing and Information. 2013.

18. Wisaeng, Kittipol, **An empirical comparison of data mining techniques in medical databases,** International Journal of Computer Applications 77.7 (2013).

19. Sharma, Anil, and Balrajpreet Kaur, **A Research Review On Comparative Analysis Of Data Mining Tools, Techniques And Parameters,** International Journal of Advanced Research in Computer Science 8.7 (2017).

20. David, Satish Kumar, A. T. Saeb, and Khalid Al Rubeaan, **Comparative analysis of data mining tools and classification techniques using weka in medical bioinformatics,** Computer Engineering and Intelligent Systems 4.13 (2013): 28-38.

21. Pon Periasamy, **A Review on Health Data Using Data Mining Techniques,** International Research Journal of Engineering and Technology (IRJET) 2.07 (2015): 2395-0056.