DIABITIES PREDICTION USING MACHINE LEARNING

SEMINAR REPORT

Submitted by

V VISWA RA1911026020076

NAVEEN CHAITANYA RA1911026020090

M .NAVEEN RA1911026020103

Under the guidance of

Mr. M GOWTHAM SETHUPATHI

Mrs. ANTONY VIGIL

In partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

RAMAPURAM CAMPUS, CHENNAI-600089

MAY 2022

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Deemed to be University Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that the Seminar-II report titled "**DIABITIES PREDICTION USING MACHINE LEARNING**" is the bonafide work of **"V VISWA [RA1911026020076]**" submitted for the course 18CSP106L Seminar – II. This report is a record of successful completion of the specified course evaluated based on literature reviews and the supervisor. No part of the Seminar Report has been submitted for any degree, diploma, title, or recognition before

SIGNATURE
Mr. M. Gowtham Sethupathi
Assistant Professor,
Department of Computer Science and Engineering,
SRM IST Ramapuram Campus,
Chennai.

SIGNATURE
Dr. K. Raja
Head of the Department,
Department of Computer Science and Engineering,
SRM IST Ramapuram Campus,
Chennai.

Submitted for the Viva Voce Examination held on.....at SRM Institute of Science and Technology, Ramapuram Campus, Chennai-600089.

EXAMINER 1 **EXAMINER 2**

CHAPTER – 1

INTRODUCTION

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. Diabetes is the fast growing disease among the people even among the youngsters. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience.

The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. The glucose moves around the body in the bloodstream. Some of the glucose is taken to our brain to help us think clearly and function. The remainder of the glucose is taken to the cells of our body for energy and also to our liver, where it is stored as energy that is used later by the body. In order for the body to use glucose for energy, insulin is required. Insulin is a hormone that is produced by the beta cells in the pancreas. Insulin works like a key to a door. Insulin attaches itself to doors on the cell, opening the door to allow glucose to move from the blood stream, through the door, and into the cell. If the pancreas is not able to produce enough insulin (insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycaemia) and diabetes develops. Diabetes Mellitus means high levels of sugar (glucose) in the blood stream and in the urine.

Types of Diabetes

Type 1: diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

Type 2 : diabetes means that the cells produce a low quantity of insulin or the body can’t use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living. Gestational diabetes appears in pregnant women who suddenly develop high blood sugar. In two thirds of the cases, it will reappear during subsequent pregnancies. There is a great chance that type 1 or type 2 diabetes will occur after a pregnancy affected by gestational diabetes.

Symptoms of Diabetes

- Frequent Urination
- Increased thirst
- Tired/Sleepiness
- Weight loss
- Blurred vision
- Mood swings
- Confusion and difficulty concentrating

frequent infections Causes of Diabetes Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with viruses such as rubella, Coxsackievirus, mumps, hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

Currently, the medical condition has no permanent cure. However, the effectiveness of managing it is dependent on its early diagnosis. According to the U.S Department of Health and Human Services, early diagnosis of diabetes is essential in keeping patients with the disease healthy . Despite the fundament of early diagnosis and the breakthrough in its management, the currently available tools are ineffective for timely diagnosis. The rapid development of computational intelligence has made it possible to increase the diagnosis accuracy and to expedite the process. Computational intelligence techniques are used to study patterns and behavior of the medical condition and to build an accurate logic that is applicable in diagnosing the disease . The most common machine learning algorithms used for diabetes diagnosis include Support Vector Machine and Random Forest. The accuracy and efficiency of diabetes diagnosis using computational intelligence techniques vary with the type of classification algorithm used, and the quantity, quality, and accuracy of the training dataset

1.1 OVERVIEW

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, five machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on john Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 99% using Decision Tree algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

1.2 PROBLEM STATEMENT

In this present era, where huge quantity of information is generating on the internet day by day. So it is necessary to provide the better mechanism to extract the useful information fast and most effectively. Text summarization is one of the methods of identifying the important meaningful information in a document or setrelated document and compressing them into a shorter version preserving its overall meanings. It reduces the time required for reading whole document and also it space problem that is needed for storing large amount of data. Automatic text summarization problem has two sub-problems that is single document and multiple documents. In single document the single document is taken as the input and summarized information is extracted from that particular single document. In Multiple document the multiple documents of single topic is taken as an input andthe output which is generated should be related to that topic.

1.3 OBJECTIVE

- With such a big amount of data circulating in digital space there is need to develop machine learning algorithms that can automatically predicts the diabetes for the persons having symptoms
- The data should be clear so that we can predict the certain person having the diabetes weather he is positive or negative for the disease by these methods of the machine learning
- The predominant aim of this project is to propose novel predictive model to predict diabetes mellitus using the clinical and e-diabetic Big Data. The objectives of the oposed work are formulated
- Machine learning classification approaches are well accepted by researchers for developing disease risk prediction models
- The objective is to use those approaches and develop a prediction model for Diabetes disease detection

2. LITERATURE SURVEY

Existing system

In existing system Mostly they are using the Naive Bayes algorithm.uses the classification on diverse types of datasets that can be accomplished to decide if a person is diabetic or not. The diabetic patient's data set is established by gathering data from hospital warehouse which contains two hundred instances with nine attributes. These instances of this dataset are referring to two groups i.e. blood tests and urine tests. In this study the implementation can be done by using WEKA to classify the data and the data is assessed by means of 10-fold cross validation approach, as it performs very well on small datasets, and the outcomes are compared. The naïve Bayes, J48, REP Tree and Random Tree are used. It was concluded that J48 works best showing an accuracy of 60.2% among others. Volume 6, Issue 4, May-June-2020 | http://ijsrcseit.com KM Jyoti Rani Int J Sci Res CSE & IT, July-August-2020; 6 (4) : 294-305 296 Aiswaryaet al. [2] aims to discover solutions to detect the diabetes by investigating and examining the patterns originate in the data via classification analysis by using Decision Tree and Naïve Bayes algorithms. The research hopes to propose a faster and more efficient method of identifying the disease that will help in well-timed cure of the patients. Using PIMA dataset and cross validation approach the study concluded that J48 algorithm gives an accuracy rate of 74.8% while the naïve Bayes gives an accuracy of 79.5% by using 70:30 split

2.1 ISSUES IN EXISTING SYSTEM

- Accuracy of the algorithm is very low as we can see in above existing system and the data sets used in the existing system are very less
- By using the Naive bayes algothis the accuracy that we are getting is approximately above 50% which is not sufficient to predict the diabetes accurately
- Extractive summarizers only extracts the important sentences from text where abstractive summarizer forms its own sentences in order

2.2 Predicting the diabtes using naive bayes algorithm

AUTHOR: J.N.Madhuri, Ganesh Kumar.R

JOURNAL PUBLISHED: IEEE

TECHNOLOGIES USED: Naive bayes algorithm

OBJECTIVE:

In this present era, where huge quantity of information is generating on the internet day by day. So it is necessary to provide the better mechanism to extract the useful information fast and most effectively. Text summarization is one of the methods of identifying the important meaningful information in a document or set related document and compressing them into a shorter version preserving its overall meanings. It reduces the time required for reading whole document and also it space problem that is needed for storing large amount of data. Automatic text summarization problem has two sub-problems that is single document and multiple documents. In single document the single document is taken as the input and summarized information is extracted from that particular single document. In Multiple document the multiple documents of single topic is taken as an input and the output which is generated should be related to that topic.

ADVANTAGES:

explained in detail some of the remarkable works in the arena of text summarization

2.3 An Overview of Diabets predicion techiniques using machine learning techniques

AUTHOR:K.VijiyaKumar

JOURNAL PUBLISHED: IEEE

TECHNOLOGIES USED: Random forest algorithm

OBJECTIVE: Proposed random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly. Nonso Nnamoko et al. [13] presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy

ADVANTAGES: This system contains containing good amount of accuracy.

DISADVANTAGES: The data sets used in the system are very less.

2.4 Diabetes prediction model using machine learning - A Survey

AUTHOR: . **Muhammad Azeem Sarwar**

JOURNAL PUBLISHED: IEEE

TECHNOLOGIES USED: Bayesian and KNN (K-Nearest Neighbor)

OBJECTIVE: Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of the diabetes disease. Deeraj Shetty . proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algorithms Performance and accuracy of the applied algorithms is discussed and compared

ADVANTAGES: Performance and accuracy of the applied algorithms is discussed and compared

DISADVANTAGES: Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset

2.5 Diabetes prediction model Analysis

AUTHOR: Yasodhae

JOURNAL PUBLISHED: IEEE

TECHNOLOGIES USED: The naïve Bayes, J48, REP Tree and Random Tree are used **OBJECTIVE:** Yasodhaet a uses the classification on diverse types of datasets that can be accomplished to decide if a person is diabetic or not. The diabetic patient's data set is established by gathering data from hospital warehouse which contains two hundred instances with nine attributes. These instances of this dataset are referring to two groups i.e. blood tests and urine tests. In this study the implementation can be done by using WEKA to classify the data and the data is assessed by means of 10-fold cross validation approach, as it performs very well on small datasets, and the outcomes are compared. The naïve Bayes, J48, REP Tree and Random Tree are used. It was concluded that J48 works best showing an accuracy of 60.2% among others.

ADVANTAGES: . The class imbalance is a mostly occur in a dataset having dichotomous values, which means that the class variable have two possible outcomes and can be handled easily if observed earlier in data preprocessing stage and will help in boosting the accuracy of the predictive model

DISADVANTAGES: In order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset

2.6 A Diabetes Predicton Model Using Machine Leraning Reviews

AUTHOR: A.A. Aljumah, N.M. Saravana Kumar, Saumya

JOURNAL PUBLISHED: IEEE

TECHNOLOGIES USED: Logistic Regression, SVM, KNN and Decision Tree methods.

OBJECTIVE: A.A. Aljumah suggested a predictive analysis of diabetic treatment using a regression based data mining technique. Oracle Data Miner (ODM) tool was deployed for predicting diabetics and support vector machine algorithm was employed for experimental analysis on Datasets of Non Communicable Diseases (NCD) risk factors in Saudi Arabia.
Mohammed presented a review of existing applications of the Map Reduce programming framework and its implementation platform Hadoop in clinical big data and related medical health informatics. N.M. Saravana Kumar presented Predictive Analysis System Architecture with various stages of data mining. Prediction was carried out in Hadoop / Map Reduce environment. Predictive Pattern matching system was deployed to compare the analyzed threshold value with the obtained value. Saumya applied analytical techniques to reduce the hospital readmission of diabetic patients. In the proposed methodology, Hive was used as the preprocessing tool and R Hadoop as the analytics and predictive modeling tool. Classification was done using Logistic Regression, SVM, KNN and Decision Tree methods. Miss-classification error rates were also calculated.

ADVANTAGES: Hive was used as the preprocessing tool and R Hadoop as the analytics and predictive modeling tool

2.7 Diabetes prediction model using machine learning

AUTHOR: Peter Augustine,Sadhana,K. Sharmila

JOURNAL PUBLISHED: IEEE

TECHNOLOGIES USED: Deployment of big data algorithms

OBJECTIVE: Peter Augustine presented a concept paper on analyzing the data flowing from health monitoring devices. The present status of healthcare in India was presented. The application of Hadoop's map reduce in healthcare data was expounded. An interface HIPI (Hadoop Image Processing Interface) in Hadoop environment was also explicated. Sadhana analyzed the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases data set using a proposed architecture which comprised of Hive and R. The raw data (CSV file) was given as input to .K. Sharmila presented a survey paper on the advancement in the field of data mining, the latest adoption of Hadoop platform, deployment of big data algorithms and consequently the open challenges in the Indian medicinal data set.

ADVANTAGES: This system contains containing good amount of accuracy.

DISADVANTAGES: it may get many errors using this model

Summary:

The synthesized literature postulates the following major observation:

• The Main task before extractive summarization is to find important information to be comprised in the summary.

• The predominant aim of this project is to propose novel predictive model to predict diabetes mellitus using the clinical and e-diabetic Big Data. The objectives of the proposed work are formulated as below:

•To create an e-diabetic portal

•To build data warehouse using cloud computing technology

•To apply Random forest to derive patterns

•To predict diabetes using the generated patterns

CHAPTER-3

SYSTEM DESIGN

3.1 INTRODUCTION

Diabetes is a very common disease all over the world.Many of the patients do not even know that they have this disease, so it is important to have a means through which we can know that we have diabetes.Most of the time people make an assessment based on their experience and symptoms and consult a doctor.So to make it easy and accurate, we need a prediction model to predict diabetes at an early stage.In this modern era, human beings encounter different health issues. Most of the health issues are due to the food habits of the individuals. In this project work, a predictive approach is proposed to pre-treat Diabetic Mellitus. The proposed approach has three phases namely data collection, data storage and analytics. This approach plays an important role in predicting diabetes and pre-treating diabetic patients. The phases in the proposed approach for diabetic prediction are presented.

3.2 SYSTEM ARCHITECTURE

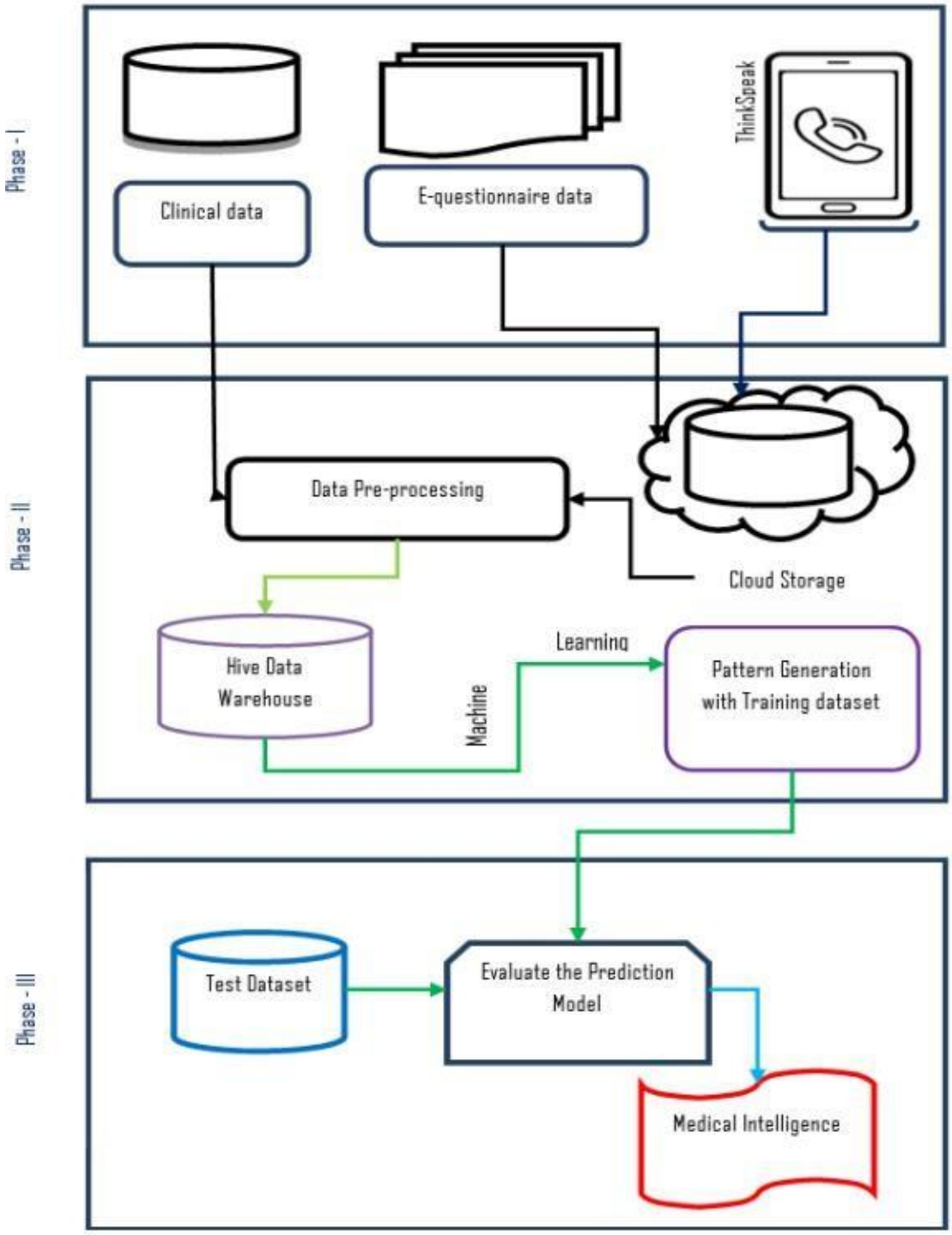

FIGURE 1

CHAPTER-4

MODULE DISCRIPTION

4.1 INTRODUCTION

Diabetes mellitus is one of the non-communicable diseases that pose a threat to human health. It has become a major global health problem. It is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin which it produces. It is found that diabetes causes blindness, amputation and kidney failure. Lack of awareness about diabetes, insufficient access to health services and essential medicines can lead to the above mentioned complications. According to a study by the World Health Organization (WHO), number of diabetic patients will raise to 552 million by 2030, which means that one in 10 adults will have diabetes by 2030. In 2014, the global prevalence of diabetes was estimated to be 9 % among adults aged 18+ years [1]. WHO insisted with an alarm that Diabetes is the 7th leading cause of death in the world. In 2012, an estimated 1.5 million deaths were directly caused by diabetes. Total deaths due to diabetes are projected to rise by more than 50 % in the next 10 years

4.2 DATA COLLECTION

The proposed approach has three phases namely data collection, data storage and analytics. This approach plays an important role in predicting diabetes and pre-treating diabetic patients.

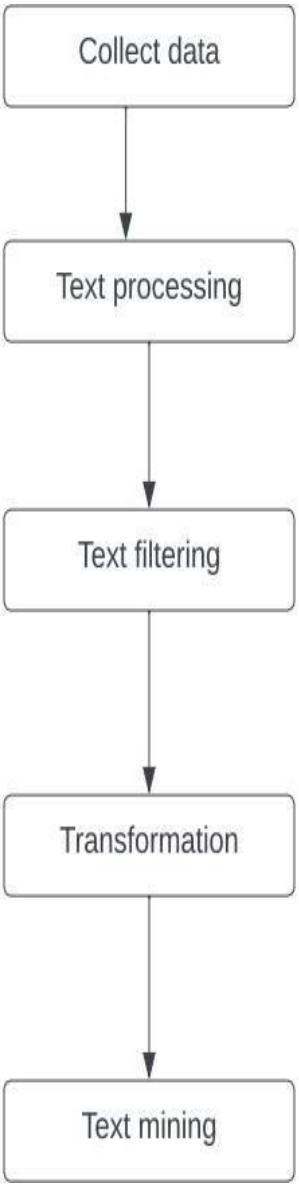

FIGURE 2

- Dataset used in this paper is based on the various symptoms faced by diabetes affected people
- Include glucose levels, hunger feelings, vision loss, healing speed etc
- Using a mixed-methods approach can provide the most comprehensive information to answer specific evaluation questions

4.3 Data pre processing

Remove Irrelevant data :-

In the data collected from the dataset, the less important features will be eliminated from the data. For example, in the dataset collected, the least important feature is given as skin thickness.

Dropping missing values:-

Through the analysis of our data, we have found few null values under many features. We remove them using "dropna" function.

Removing Outliers:-

From the dataset we removed all the data which are not in actual range.

Dimensionality Reduction:-

We checked the correlation between data and eliminated them if there is high correlation coefficient.

Data pre-processing is one vital step in data discovery methodology. Most health care information contain missing value, wheezy and inconsistency information.

4.4 Data cleaning

It is that the tactic of detection and correcting (or removing) corrupt or inaccurate records from a record set, table, or data and refers to distinguishing incomplete, incorrect, inaccurate or tangential parts of the knowledge some substitution, modifying, or deleting the dirty or coarse data. [8] Data cleansing is additionally performed interactively with data twenty five haggle tools, or as execution through scripting. Information cleansing is in addition said as information clean-up or information cleansing

4.5 Data integration

It could be a method within which heterogeneous knowledge is retrieved Associate in Nursing combined as an incorporated kind and structure. Knowledge integration permits fully completely different information kinds (such as information sets, documents and tables) to be integrated by users, organizations and applications, to be used as personal or business processes and or functions.

4.6 Data reduction

It is that the transformation of numerical or alphabetical digital data derived through empirical observation or by experimentation into a corrected, ordered, and simplified type. The fundamental construct is that the reduction of undeterminable amounts of data all the means right down to the purposeful components

4.7 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy

their probabilistic expectation, rather than giving every classifier a chance to vote in favor of a single class It is capable of handling large datasets with high dimensionality.

.It enhances the accuracy of the model and prevents the overfitting issue.

It takes less training time as compared to other algorithms.It predicts output with high accuracy, even for the large dataset it runs efficiently It can also maintain accuracy when a large proportion of data is missing.

In random forests, each tree in the group is worked from an example drawn with substitution (i.e., a bootstrap sample) from the training set. Likewise, while splitting a node amid the construction of the tree, the split that is picked is not any more the best split among all highlights. Rather, the split that is picked is the best split among an random subset of the features. Because of this randomness, the bias of the forest marginally increases (as for the bias of a single non-random tree) in any case, because of averaging, its fluctuation likewise diminishes, generally more than making up for the expansion in predisposition, thus yielding a general better model.

The scikit-learn usage joins classifiers by averaging their probabilistic expectation, rather than giving every classifier a chance to vote in favor of a single class. Below is an image of how the splitting looks like in a Random Forest:

The scikit-learn usage joins classifiers by averaging. Below is an image of how the splitting looks like in a Random Forest

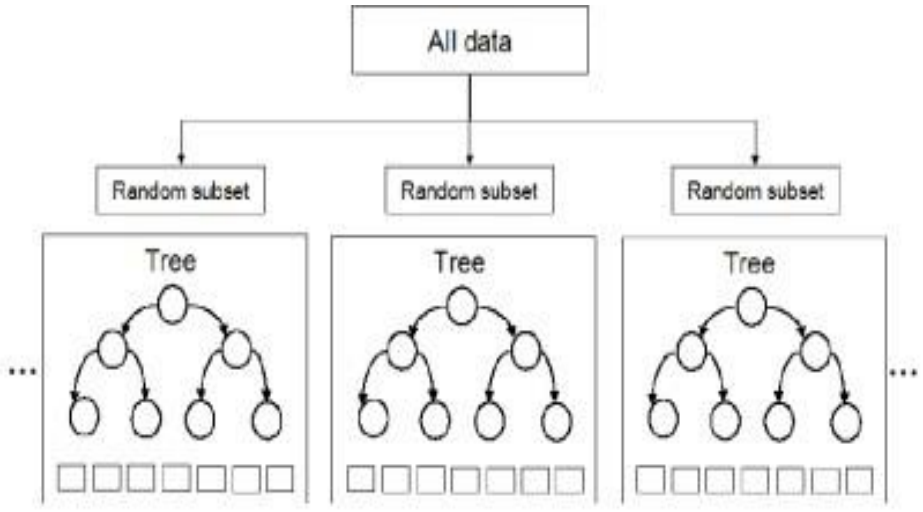

FIGURE 3

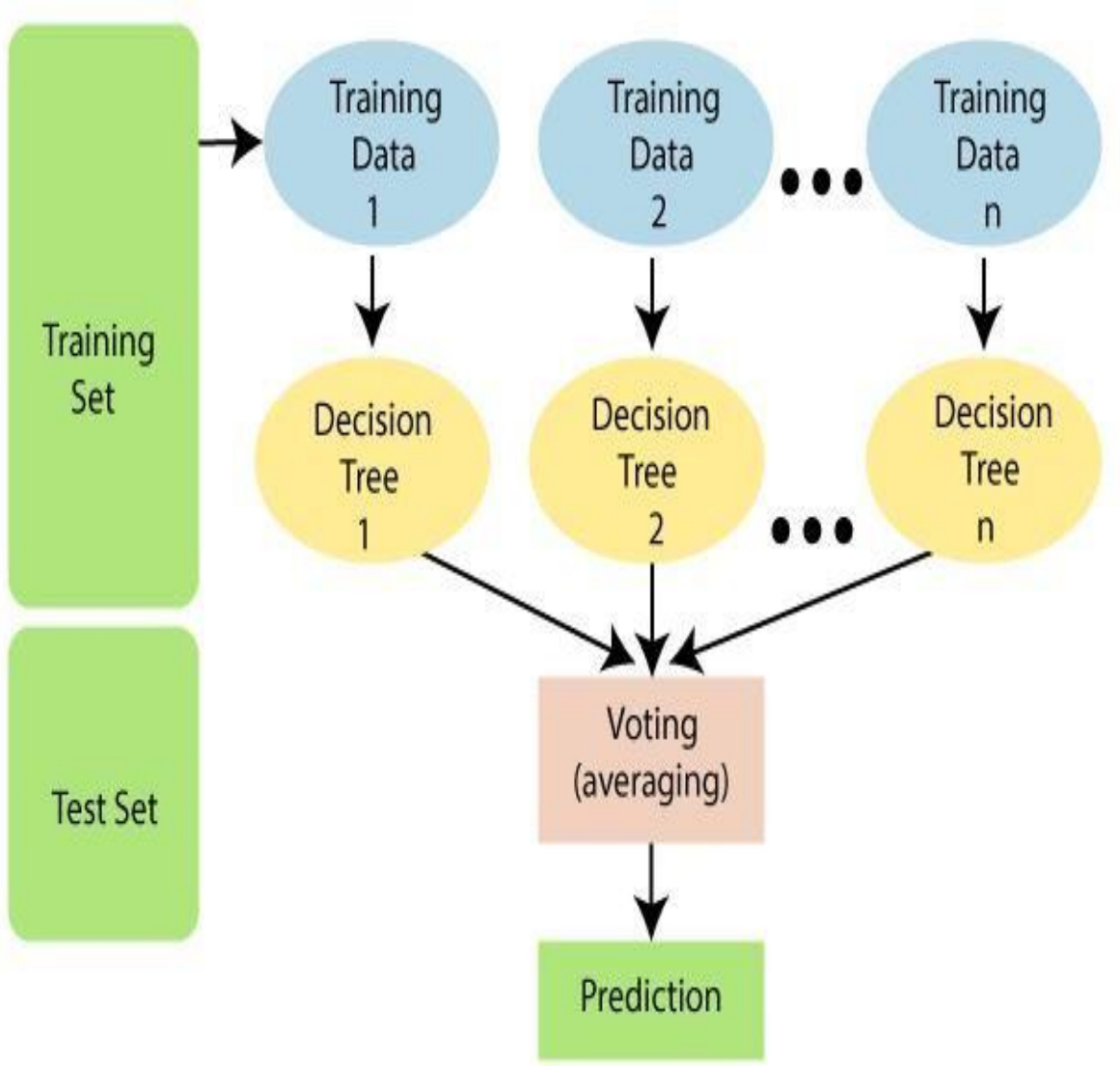

FIGURE 4

We made a model to predict whether the blood test detection for diabetes is going to turn out positive or negative. However ,the dataset was slightly imbalanced having around 262 class 0, i.e.class negative and 130 for class 1, i.e. positive. Below is an image of the class distribution

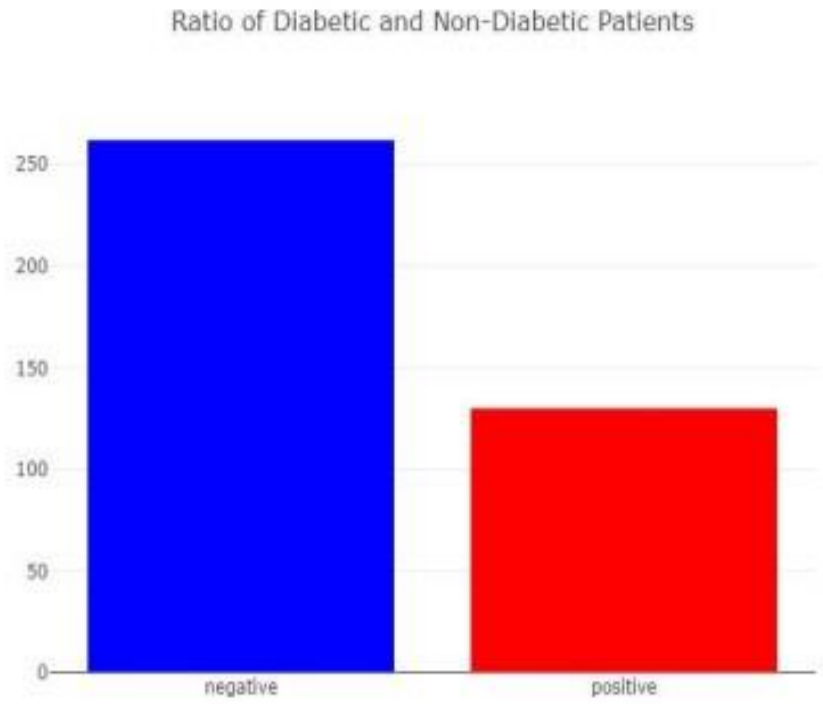

FIGURE 5

To fix the imbalancing we first used some upsampling techniques on the train dataset, to balance the two classes . Also precision is not a good evaluation metric for imbalanced dataset, and we used f1-score and recall score [10] for evaluation which is discussed in the next section. Following are the list of features that we had in our dataset:

i) **Pregnant**: It represents the number of times the woman got pregnant during her life

ii) **Glucose:** It represents the plasma glucose concentration at 2 hours in an oral glucose tolerance test

iii) **Diastolic**: The blood pressure is a very well-known way to measure the health of the heart of a person, there are too measure in fact, the diastolic and the systolic. In this data

set, we have the diastolic which is in the fact the pressure in (mm/Hg) when the heart relaxed after the contraction.

iv) **Triceps:** It is a value used to estimate body fat (mm) which is measured on the right arm halfway between the olecranon process of the elbow and the acromial process of the scapula.

v) **Insulin:** It represents the rate of insulin 2 hours serum insulin (mu U/ml).

vi) **BMI**: It represents the Body Mass Index (weight in kg / (height in meters squared), and is an indicator of the health of a person.

4.8 MODEL BUILDING

The preprocessed data is now deployed in the analysis to build prediction model. In this project work, three type –II Diabetes predicative models are developed using BigML tool. BigML provides machine learning algorithms as Software as a Services (SaaS). It can be accessed through three models namely web interface, command line and Restful API. The Model is developed using the available web interface. The model creation process using BigMLIn standard statistical analysis, log_lenier is complex due to exponential number of variables.

Association discovery approaches are very much suitable for high dimensional data. The relationship between the values is discovered using association discovery rather than the relation between the variables. FilteredTop_K Technique is used for association discovery. The created model is named as Diabetes Diagnosis

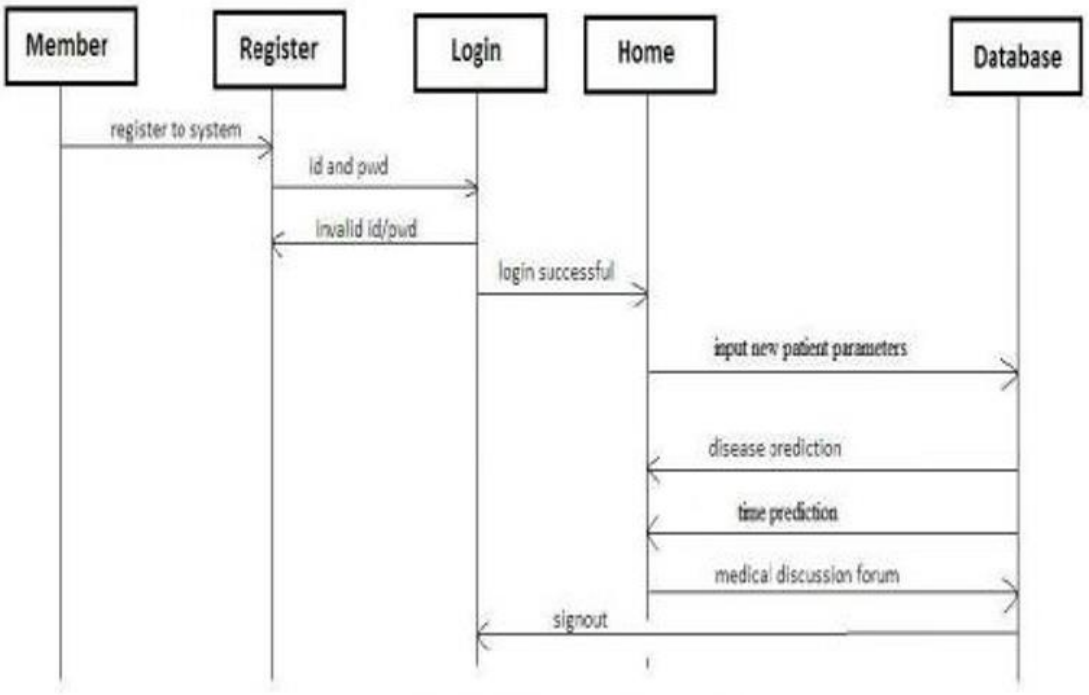

FIGURE 6

Implementation can be described as the realization of an application, or execution of the plans, ideas, models, design and system development, specification of the model, standard, algorithms used in the system, or authority. In computer science, an implement is explained as the realization of technically specified or algorithms' as a programed, a software component, or anyothers computer systems through computer programming and deployment

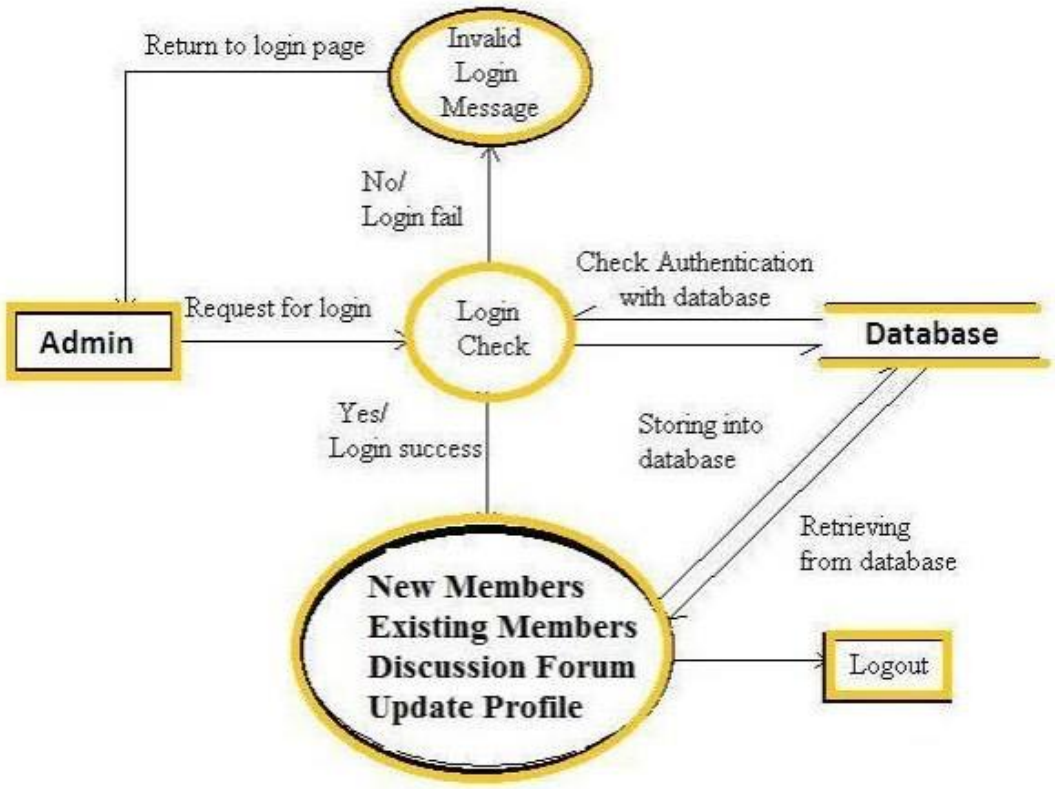

Figure 4: Data Flow for Admin.

FIGURE 7

CHAPTER 5
Calculation and Graphs

First of all we found the correlation between each feature columns, to check whether there is any highly correlated features, and as per our threshold of 0.7 there was none.. Below is the correlation

FIGURE 8

The surprising fact is that, even on a small dataset, it was the bagging algorithm, Random Forest that surpassed all the other algorithms in the predictions. Following is a classification report, which signifies each of the algorithms' predictive accuracy, False +ve is also known as Type 1 error and False - ve is also known as Type 2

Name of Algorithm	True +ve	True -ve	False +ve	False -ve
Logistic Regression	73	64	18	18
Support Vector Machines	68	70	23	12
Random Forest	75	70	12	16

Following table is the f1-score and recall score for each class in case of Random Forest**.**

Class	F1-Score	Recall
0 (Negative)	0.84	0.86
1 (Positive)	0.83	0.81
Avg./Total	0.84	0.84

The most important part for an algorithm to make predictions are the features which it is using for making the predictions, and some features play exceptionally important part for predicting. Below is the table, to how much importance has Random Forest put to each feature, following a graphical representation of the same.

Feature Name	Importance
Pregnant	0.07
Glucose	0.21

Symptoms values based on the algorithm

Diastolic	0.08
Triceps	0.09
Insulin	0.19
BMI	0.12
Diabetes	0.11
Age	0.13

The sum of the importance of each feature will result to one. Above, only the features playing major role for diabetes have been plotted, where X-Axis represents the importance of each feature and Y-Axis the names of the features.

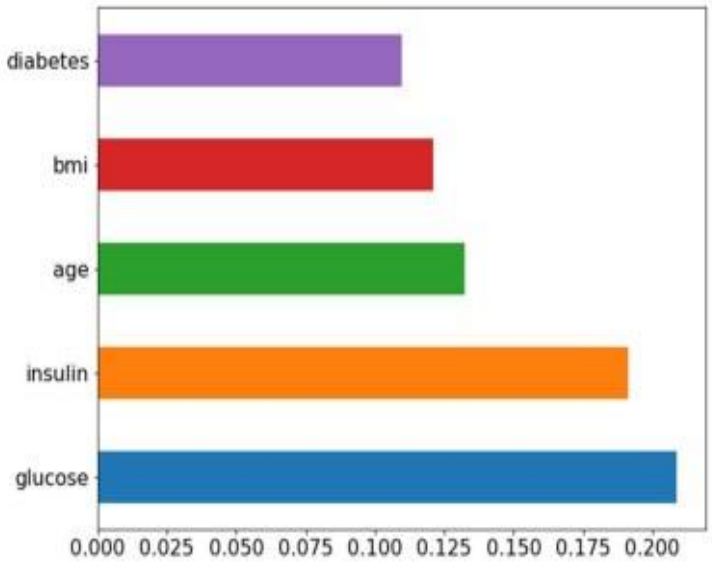

FIGURE 9

CHAPTER 6

CONCLUSION

In our project the result is classified into Yes or No. If the result is classified into No then we use time prediction module. Time Prediction - here we predict the "time" of getting the diabetes disease. We analyze the result of the diabetes prediction and check the accuracy of the diabetes prediction, time taken to compute the accuracy of the diabetes prediction, correctly classification and incorrectly classification of result of the diabetes prediction. We have used KNN Algorithm to predict the diabetes where result is classified into Yes or No and also for time prediction module same KNN Algorithm is used. We compared the testing data and actual data to get the accuracy of our project.The prediction of diabetes is vital in today's scenario, and this is related to its causes and its threatening complications. The world's biggest cause of death is diabetes.

The System model is designed to detect diabetes using a small number of indicators. System allows physicians to predict the likelihood of a person developing diabetes. So that medicines will be prescribed to the patients. System uses some of the machine learning strategies for the prediction, so as to obtain more accurate results. There has been a lot of investigation on diabetes profiles. Building a diabetes disease prediction system is really helpful for hospital administrations and physicians. When an illness is identified at an early stage, proper treatment becomes available. It is the multi hospital real time disease prediction system. Machine Learning algorithms will improve on disease prediction techniques

SCOPE :

"Random forest algorithm" has been used to detect diabetes diseases, many different classification algorithms have been used and will continue to be in the future. We can add module for visitors to ask questions for administrators and administrators can reply to those questions. We can include treatment module where doctors upload information about treatments for patients. Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning

REFERENCES:

[1] “Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus” Md Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, IEEE 2019.“A Comprehensive Exploration to the Machine Learning Techniques for Diabetes Identification” Sidong Wei1, Xuejiao Zhao, Chunyan Miao Shanghai Jiao Tong University, China.

[2] “Association Rule Extraction from Medical Transcripts of Diabetic Patients” Lakshmi K S, G Santhosh Kumar, 2014.

[3] “Diabetes Care Decision Support System” 2nd International Conference on Industrial and Information Systems IEEE 2010.

[4] “An Intelligent Mobile Diabetes Management and Educational System for Saudi Arabia: System Architecture” M.M. Alotaibi, R.S.H. Istepanian, A.Sungoor and N. Philip, IEEE 2014.

[5] “Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases” by BerinaAlic, Lejla Gurbeta, IEEE 2017.

[6] Performance Analysis of Classification Approaches for the Prediction of Type II Diabetes” by M. Durgadevi, M. Durgadevi, IEEE 2017.

[7] “Cloud-Based Diabetes Coaching Platform for Diabetes Management” Elliot B. Sloane Senior Member IEEE, Nilmini Wickramasinghe, Steve Goldberg 2016.

[8] Minyechil Alehegn and Rahul Joshi, “Analysis andprediction of diabetes diseases using machine learning algorithm”:International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct -2017

[9] P. Suresh Kumar and V. Umatejaswi, “Diagnosing Diabetes using Data Mining Techniques”,International Journal of Scientific and ResearchPublications, Volume 7, Issue 6, June 2017 705 ISSN 2250-3153.

[10] [11] “Clustering Medical Data to Predict the Likelihood of Diseases” by Razan Paul, Abu Sayed Md. Latiful Hoque, IEEE 2010. “Robust Parameter Estimation in a Model for Glucose Kinetics in Type 1 Diabetes Subjects” Proceedings of the 28th IEEE EMBS A nnual International Conference New York City, USA, Aug 30-Sept 3, 2006.

[11] Anjali C And Veena Vijayan V, Prediction and Diagnosis of Diabetes Mellitus, "A Machine Learning Approach" ,2015 IEEE in Intelligent Computational Systems (RAICS) | Trivandrum.

[12] Ridam Pal ,Dr. Jayanta Poray, and Mainak Sen, ,"Application of Machine Learning Algorithms on Diabetic Retinopathy", 2017 2nd IEEE International Conference On Recent Trends In Electronics Information & Communication Technology, May 19- 20, 2017, India.

[13] Dr. M. Renuka Devi and J. Maria Shyla, "Analysis of Various Data Mining Techniques toPredict Diabetes Mellitus", International Journal ISSN 0973-4562 Volume 11, Number 1 (2016) pp 727-730.