PREDICTION OF DIABETES USING DIFFERENT MACHINE LEARNINGS

**A MINI PROJECT**

**REPORT SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF**

**MASTER OF TECHNOLOGY IN**

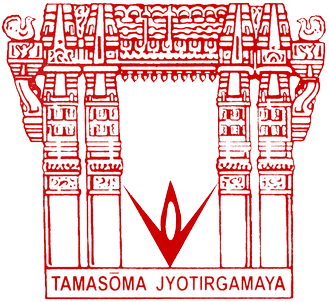
**COMPUTER SCIENCE &ENGINEERING**

**BY**

**V. RAJENDRA**

**H.T.NO. (20071D5811)**

**Under the esteemed guidance of Mr. P. Bharath Kumar Chowdary Assistant Professor**



**COMPUTER SCIENCE AND ENGINEERING DEPARTMENT**

**VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY**

### An Autonomous Institute NAAC Accredited with 'A++' Grade NBA Accredited for CE, EEE, ME, ECE, CSE, EIE, IT B. Tech Courses

**Approved by AITCE, New Delhi, Affiliated to JNTUH Recognized as "College with potential for Excellence" by UGC Bachupally (v), Hyderabad, Telangana, India.**

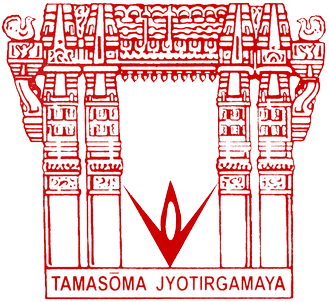
**November 2021.**

**VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY**

**An Autonomous Institute NAAC Accredited with 'A++' Grade NBA Accredited for CE, EEE, ME, ECE, CSE, EIE, IT B. Tech Courses**

**Approved by AITCE, New Delhi, Affiliated to JNTUH Recognized as "College with potential for Excellence" by UGC Bachupally (v), Hyderabad, Telangana, India.**

#### DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



**CERTIFICATE**

This is to certify that **V. Rajendra (20071D5811)** have successfully completed the mini project at Computer Science and Engineering Department of VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad entitled **“PREDICTION OF DIABETES USING DIFFERENT MACHINE LEARNINGS” in** partial fulfillment of the requirements for the award of M. Tech degree during the academic year 2021-2022. This work is carried out under my supervision and has not been submitted to any other University/Institute for award of any degree/diploma.

#### Project Guide

Mr. P. Bharath Kumar Chowdary Assistant Professor

CSE Department VNR VJIET

#### Head of the Department

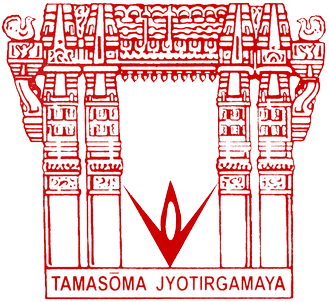
Dr. S. Nagini Professor & HOD CSE Department VNR VJIET

**NR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY**

**An Autonomous Institute NAAC Accredited with 'A++' Grade NBA Accredited for CE, EEE, ME, ECE, CSE, EIE, IT B. Tech Courses**

**Approved by AITCE, New Delhi, Affiliated to JNTUH Recognized as "College with potential for Excellence" by UGC Bachupally (v), Hyderabad, Telangana, India.**

#### DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



**DECLARATION**

I hereby declare that the mini project work entitled **“PREDICTION OF DIABETES USING DIFFERENT MACHINE LEARNINGS** is carried out by me during the year 2021-2022 in partial fulfillment for the award of the Degree of **MASTER OF TECHNOLOGY** in **COMPUTER SCIENCE & ENGINEERING** from

VNRVJIET, Hyderabad. I have not submitted the same to any other university or organization for the award of any other Degree.

**V.Rajendra (20071D5811)**

### ACKNOWLEDGMENT

I sincerely thank **Dr. C. D. Naidu**, Principal of our college and **VNR Vignana Jyothi Institute of Engineering and Technology College Management** for the facilities provided in the college premises.

I earnestly thank **Dr. S. Nagini**, Professor and HOD, CSE Dept., VNR VJIET, for the extreme cooperation and encouragement to do the project work.

I am immensely thankful to my project guide Mr. **P. Bharath Kumar Chowdary** Assistant Professor of CSE Dept., VNRVJIET, for the exhaustive technical help and supervision, many hours of devoted guidance during this course of work.

I extend my heartfelt thanks to my mini project coordinators

**Dr. T. Sunil Kumar, Professor & Dr. Deepak Sukheja,PArossfoescsiaotre** of CSE

Dept., VNRVJIET, for excellent supervision and guidance. Without his supervision this project would never come out in this form.

I express my gratitude to my **Parents, CSE faculty** and **Classmates** for their suggestions and moral support.

**V. Rajendra (20071D5811)**

**ABSTRACT**

The diabetes is one of lethal diseases in the world. It is additional a inventor of various varieties of disorders foe example: coronary failure, blindness, urinary organ diseases etc. In such case the patient is required to visit a diagnostic center, to get their reports after consultation. Due to every time they have to invest their time and currency. But with the growth of Machine Learning methods, we have got the flexibility to search out an answer to the current issue, we have got advanced system mistreatment information processing that has the ability to forecast whether the patient has polygenic illness or not. Furthermore, forecasting the sickness initially ends up in providing the patients before it begins vital. Information withdrawal has the flexibility to remove unseen data from a large quantity of diabetes associated information. The aim of this analysis is to develop a system which might predict the diabetic risk level of a patient with a better accuracy. Model development is based on categorization methods as Decision Tree, ANN, Naive Bayes and SVM algorithms. For Decision Tree, the models give precisions of 85%, for Naive Bayes 77% and 77.3% for Support Vector Machine. Outcomes show a significant accuracy of the methods.

### TABLE OF CONTENTS

1. [INTRODUCTION 1](#_TOC_250010)
   1. [Introduction 2](#_TOC_250009)
2. [EXISTING SYSTEM AND PROPOSED SYSTEM 4](#_TOC_250008)
   1. [Existing System 5](#_TOC_250007)
   2. [Proposed System 6](#_TOC_250006)
3. [LITERATURE Survey 7](#_TOC_250005)
   1. [Literature Survey 8](#_TOC_250004)
4. [SYSTEM ANALYSIS AND ARCHITECTURE 25](#_TOC_250003)
   1. System Architecture 26
   2. Requirement Specifications 28
   3. [Implementation 3](#_TOC_250002)0
   4. [Software Overview 36](#_TOC_250001)
5. [SOFTWARE DESIGN 46](#_TOC_250000)
   1. Introduction to Software Design 47
   2. Introduction to UML 48
   3. UML& Class Diagram 49
      1. Use Cases 49
      2. Sequence Diagrams 51
      3. Class Diagram 51

### 6.SOFTWARE TESTING 61

* 1. Software Testing 62
  2. Levels of Testing 63

### CONCLUSIONS 70

* 1. Conclusion 71

### REFERENCES 72

* 1. References 73

### LIST OF FIGURES

|  |  |
| --- | --- |
| Fig: 2.1 Block diagram of prediction system | 6 |
| Fig: 4.1 Diagram of waterfall model | 26 |
| Fig 4.2 Working method of neural network | 30 |
| Fig: 4.3 diagram of S.V.M | 35 |
| Fig:4.4 testdata.csv | 44 |
| Fig:4.5 PID.xlsx | 44 |
| Fig:4.6 PIDtest.xlsx | 45 |
| Fig: 5.1 Use Case Diagram | 55 |
| Fig: 5.2 State chart | 56 |
| Fig: 5.3 Sequence Diagram | 57 |
| Fig:5.4 Activity Diagram | 58 |
| Fig:5.5 Class Diagram | 59 |
| Fig:5.6 Deployment Diagram | 60 |
| Fig: 6.1 Level of testing | 63 |
| Fig:6.2 Black Box Diagram | 65 |
| Fig:6.3 Diagram of result display | 69 |

# 1. INTRODUCTION

### INTRODUCTION

Diabetes is a situation which causes deficiency due to less amount of insulin in the blood. Warning sign of high blood sugar results in frequent urination, feeling thirsty, increased hunger. If it is not medicated, it will lead to many difficulties. This difficulty leads to death. Severe difficulties lead to cardiovascular disease foot sores, and eye blurriness. When there is a rise within the sugar level within the blood, it is referred to as prior diabetes. The prior diabetes isn't therefore great than the traditional worth. Diabetes is appreciations to the exocrine gland not manufacturing plentiful hypoglycemic agent not responding properly to the hypoglycemic agent created. Various information mining algorithms presents different decision support systems for assisting health specialists. The effectiveness of the decision support system is recognized by its accuracy. Therefore, the objective is to build a decision support system to predict and diagnose a certain disease with extreme amount of precision. The AI consist of ML which is its subfield that resolves the real-world difficulties by "providing learning capability to workstation without supplementary program writing. 1.1 Types of Diabetes 1) Type one diabetes outcomes due to the failure of pancreas to supply enough hypoglycemic agent. This type was spoken as "insulin-dependent polygenic disease mellitus" (IDDM) or "juvenile diabetes". The reason is unidentified. The type one polygenic disease found in children beneath twenty years old. People suffer throughout their life because of the type one diabetic and rest on insulin vaccinations. The diabetic patients must often follow workouts and fit regime which are recommended by doctors.

1. The type two diabetes starts with hypoglycemic agent resistance, a situation inside which cells fail to response the hypoglycemic agents efficiently. The sickness develops due to the absence of hypoglycemic agent that additionally built. This type was spoken as "non-insulin-dependent polygenic disease mellitus". The usual cause is extreme weight. The quantity of people affected by type two will be enlarged by 2025. The existences of diabetes mellitus are condensed by 3% in rural zone as compared to urban zone. The prehypertension is joined with bulkiness, fatness and diabetes mellitus. The study found that an individual United Nations agency has traditional vital sign. 3) Type 3 Gestational diabetes occurs when a woman is pregnant and develops the high blood sugar levels without a previous history of diabetes.

Therefore, it is found that in total 18% of women in pregnancy have diabetes. So, in the older age there is risk of emerging the gestational diabetes in pregnancy.

The obesity is one of the main reasons for type-2 diabetes.

The type-2 polygenic disease is under control by proper workout and taking appropriate regime. When the aldohexose level isn't reduced by the higher strategies then medications are often recommended.

The polygenic disease static report says that 29.1 million people of the United States inhabitants has diabetes.

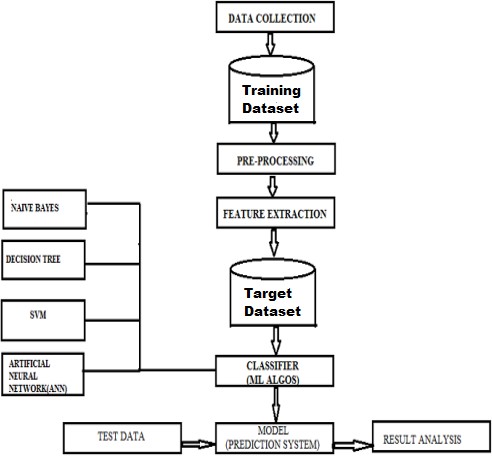
### 2.EXISTING SYSTEM AND PROPOSED SYSTEM

### Existing System:

The prior diabetes isn't therefore great than the traditional worth. Diabetes is appreciations to the exocrine gland not manufacturing plentiful hypoglycemic agent not responding properly to the hypoglycemic agent created. Various information mining algorithms presents different decision support systems for assisting health specialists. The effectiveness of the decision support system is recognized by its accuracy. Therefore, the objective is to build a decision support system to predict and diagnose a certain disease with extreme amount of precision. The AI consists of ML which is its subfield that resolves the real-world difficulties by "providing learning capability to workstation without supplementary program writing. 1.1 Types of Diabetes 1) Type one diabetes outcomes due to the failure of pancreas to supply enough hypoglycemic agent. This type was spoken as "insulin-dependent polygenic disease mellitus" (IDDM) or "juvenile diabetes". The reason is unidentified. The type one polygenic disease found in children beneath twenty years old. People suffer throughout their life because of the type one diabetic and rest on insulin vaccinations. The diabetic patients must often follow workouts and fit regime which are recommended by doctors. 2)The type two diabetes starts with hypoglycemic agent resistance, a situation inside which cells fail to response the hypoglycemic agents efficiently. The sickness develops due to the absence of hypoglycemic agent that additionally built. This type was spoken as "non-insulin- dependent polygenic disease mellitus". The usual cause is extreme weight. The quantity of people affected by type two will be enlarged by 2025. The existences of diabetes mellitus are condensed by 3% in rural zone as compared to urban zone. The prehypertension is joined with bulkiness, fatness and diabetes mellitus. The study found that an individual United Nations agency has traditional vital sign. 3) Type 3 Gestational diabetes occurs when a woman is pregnant and develops the high blood sugar levels without a previous history of diabetes.

### Proposed System:

The proposed system focuses using algorithms combinations shown above in the block diagram. The base classification algorithms are: Decision tree, Support Vector Machine, Naive Bayes and ANN for accuracy authentication.



#### Fig 2.1 Block diagram of diabetes prediction system

The dataset contains seven sixty eight instances and nine features. The dataset features are:

* + Total number of times pregnant
  + Glucose/sugar level
  + Diastolic Blood Pressure
  + Body Mass Index (BMI)
  + Skin fold thickness in mm
  + Insulin value in 2 hour
  + Hereditary factor- Pedigree function &Age of patient in year.

### 3.Literature Survey

#### Literature Survey

1. **Yilmaz N., Inan O., Uzer M.S., “A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases,” JMed Syst, vol. 38, no. 5 2014.**
   * The most important factors that prevent pattern recognition from functioning rapidly and effectively are the noisy and inconsistent data in databases.
   * A new data preparation method based on clustering algorithms for diagnosis of heart and diabetes diseases.
   * A new modified K-means Algorithm is used for clustering-based data preparation system for the elimination of noisy and inconsistent data and Support Vector Machines is used for classification.
   * A developed approach was tested in the diagnosis of heart diseases and diabetes, which are prevalent within society and figure among the leading causes of death.
   * The data sets used in the diagnosis of these diseases are the Stat log (Heart), the SPECT images and the Pima Indians Diabetes data sets obtained from the UCI database.
   * The proposed system achieved 97.87 %, 98.18 %, 96.71 % classification success rates from these data sets.
   * Classification accuracies for these data sets were obtained through using 10-fold cross- validation method.
   * According to the results, the proposed method of performance is highly successful compared to other results attained and seems very promising for pattern recognition applications.
   * The proposed work is based on the integration of Random Forest, mean, class’s mean, interquartile range (IQR), and Deep Learning to produce a clean and complete dataset. Which can enhance any machine learning model performance.
   * Moreover, the outliers repair technique is proposed based on dataset class detection, then repair it by Deep Learning (DL).
   * The final model accuracy with the two steps of imputation and outliers repair is 97.41% and 99.71% Area Under Curve(AUC).

#### Lowongtrakool C., Hiransakolwong N., “Noise filtering in unsupervised clustering using computation intelligence,” International Journal of Math, vol. 6, no. 59, pp. 2911– 2920, 2012.

* + The present data used for clustering contains data record not related to processing. This may be spam or noise causing error and delay of processing.
  + However, it is necessary to process this kind of data together, but the results may be incorrect, depending on the quantity of noise.
  + Therefore, it will be much better if data cleaning is conducted before processing system data.
  + The new method to develop unsupervised clustering intelligence to reduce the quantity of spam.
  + The new computational intelligence system applies the first layer of radial basis function network as an input layer of the system for incremental work.
  + The results of system can provide level of accuracy of least related membership in any cluster which is called the data not relevant to the dataset or the data not associated with the dataset.
  + According to an experiment, the data from UCI machine learning repository was used to test the system efficiency while classification algorithm in a program called weka3.6 was also utilized to test the system accuracy.
  + The results from noise filtering help the data processing more precise, compared to the processing without the noise filtering.
  + CMTNN is a technique based on ANNs. It concentrates on class noise or misclassification error
  + ANR is based on multi-layer neural network trained by Standard Backpropagation. In addition, a class of answer is applied to adjust the value of learning to be like the network.

#### Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil “Diabetes Disease Prediction Using Data Mining”. International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) 2016.

* + Medical professionals need a reliable prediction methodology to diagnose Diabetes. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information.
  + The main goal of data mining is to discover new patterns for the users and to interpret the data patterns to provide meaningful and useful information for the users.
  + Data mining is applied to find useful patterns to help in the important tasks of medical diagnosis and treatment.
  + The aim for mining the relationship in Diabetes data for efficient classification. The data mining methods and techniques will be explored to identify the suitable methods and techniques for efficient classification of Diabetes dataset and in mining useful patterns.
  + Diabetes, known also as Diabetes mellitus, is a prolonged disease that distresses a way body uses to take up sugar or glucose.
  + It can be described as a condition occurs when the blood sugar of a body is too high. The cases of diabetes are rapidly raising in low- and middle-income countries as compared to high income countries. Advancement in the field of Information and Communication Technology (ICT) are revolutionizing many sectors including health care and medical technologies.
  + It plays a critical role in improvising services offered to patients and hospitals. Machine learning is a field of Artificial Intelligence that makes computer system to learn from data, identify the patterns and take decisions without human intervention.
  + Machine learning can be applied on medical datasets to detect and diagnose diseases in an effective and accurate manner.

#### Anuja and R.Chitra., “Classification Of Diabetes Disease Using Support Vector Machine”, International Journal of Engineering Research and Applications (IJERA), vol.3,Issue 2, pp. 1797-1801, 2013.

* + Diabetes mellitus is one of the most serious health challenges in both developing and developed countries.
  + According to the International Diabetes Federation, there are 285 million diabetic people worldwide.
  + The total is expected to rise to 380 million within 20 years. Due to its importance, a design of classifier for the detection of Diabetes disease with optimal cost and better performance is the need of the age.
  + The Pima Indian diabetic database at the UCI machine learning laboratory has become a standard for testing data mining algorithms to see their prediction accuracy in diabetes data classification.
  + The proposed method uses Support Vector Machine (SVM), a machine learning method as the classifier for diagnosis of diabetes.
  + There are two general reasons for diabetes:
  + (1) the pancreas does not make enough insulin, or the body does not produce enough insulin. Only 5-10 % of people with diabetes have this form of the disease (Type-1).
  + (2) Cells do not respond to the insulin that is produced (Type-2). Insulin is the principal hormone that regulates uptake of glucose from the blood into most cells (muscle and fat cells).
  + If the amount of insulin available is insufficient, then glucose will not have its usual effect so that glucose will not be absorbed by the body cells that require it.
  + Diabetes mellitus being one of the major contributors to the mortality rate.
  + Detection and diagnosis of diabetes at an early stage is the need of the day. Diabetes disease diagnosis and interpretation of the diabetes data is an important classification problem.

#### Aiswarya I., S. Jeyalatha and Ronak S., “Diagnosis Of Diabetes Using Classification Mining Techniques”, International Journal of Data Mining & Knowledge Management Process (IJDKP), vol.5, ,No. 1, pp. 1-14, 2015.

* + Diabetes has affected over 246 million people worldwide with a majority of them being women. According to the WHO report, by 2025 this number is expected to rise to over 380 million.
  + The disease has been named the fifth deadliest disease in the United States with no imminent cure in sight.
  + With the rise of information technology and its continued advent into the medical and healthcare sector, the cases of diabetes as well as their symptoms are well documented.
  + The aims at finding solutions to diagnose the disease by analyzing the patterns found in the data through classification analysis by employing Decision Tree and Naïve Bayes algorithms.
  + The hopes to propose a quicker and more efficient technique of diagnosing the disease, leading to timely treatment of the patients.
  + The present work is intended to meet the following objectives:
  + 1. Present a Decision Tree and Naïve Bayes model for diabetes prediction in pregnant women.
  + 2. Summarize Diabetes – types, risk factors, symptoms, and diagnosis.
  + 3. Identify and discuss the field’s benefits to the society along with effective application.
  + The employed three techniques namely:
  + EM algorithm, H-means+ clustering and Genetic Algorithm (GA), for the classification of the diabetic patients.
  + The performance for H-means+ proved to be better than others when all the similar symptoms were grouped into clusters using these algorithm.

#### K.Rajesh and V.Sangeetha,”Application of Data Mining Methods and Techniques for Diabetes Diagnosis,” in proceedings of International journal of Engineering and Innovative Technology, vol.2, Issue 3, pp. 43-46, 2012.

* + Diabetes is a disease where the predominant finding is high blood sugar. The high blood sugar may either be because of deficient insulin production (Type 1) or insulin resistance in peripheral tissue cells (Type 2).
  + Many problems occur if diabetes remains untreated and unidentified. It is additional inventor of various varieties of disorders for example: coronary failure, blindness, urinary organ diseases etc. Nine different machine learning techniques are used in this research work for prediction of diabetes.
  + A dataset of diabetic patient’s is taken, and nine different machine learning techniques are applied on the dataset. Positive likelihood ratio, Negative likelihood ratio, Positive predictive value, Negative predictive value, Disease prevalence, Specificity, Precision, Recall, F1-Score, True positive rate, False positive rate of the applied algorithms is discussed and compared.
  + Diabetes is growing at an increasing in the world and it requires continuous monitoring. To check this we use Logical regression, Random Forest, Logical regression CV, Support Vector Machine, Artificial Neural Network (ANN), Decision Tree, k-nearest neighbors (KNN), XGB classifier.
  + The annual report of World Health Organization, add up to the number of individuals experiencing diabetes is estimated to be 9.3% (463 million people), rising to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045.

#### Prof. Dhomse Kanchan B. and Mr. Mahale Kishor M. “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis”. International Conference on Global Trends in Signal Processing, Information Computing and Communication 2016.

* + The worldwide study on causes of death due to heart disease/syndrome has been observed that it is the major cause of death.
  + If recent trends are allowed to continue, 23.6 million people will die from heart disease in coming 2030.
  + The healthcare industry collects large amounts of heart disease data which unfortunately are not “mined” to discover hidden information for effective decision making.
  + In study of PCA has been done which finds the minimum number of attributes required to enhance the precision of various supervised machine learning algorithms.
  + Data mining has number of important techniques like categorization, preprocessing.
  + Diabetic is a life-threatening disease which prevent in several urbanized as well as emergent countries like India.
  + The data categorization is diabetic patient’s datasets which is developed by collecting data from hospital repository consists of 1865 instances with dissimilar attributes.
  + The dataset are two categories of blood tests, urine tests. In this research paper we discuss a variety of algorithm approaches of data mining that have been utilized for diabetic disease prediction.
  + Data mining is a well-known practice used by health organizations for classification of diseases such as diabetes and cancer in bioinformatics research.
  + The genomes of thousands of people each of which has a length of 3.8x109 base pairs have been sequences by a variety of labs.
  + Walmart handles more than 1Milion transactions every hour and has databases containing larger than 2.5 x 1015 of information Cukier 2010.

#### Veena Vijayan V. And Anjali C, Prediction and Diagnosis of Diabetes Mellitus, “A Machine Learning Approach” ,2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10- 12 December 2015 | Trivandrum.

* Diabetes is a disease caused due of the expanded level of sugar fixation in the blood.
* Various computerized information systems were outlined utilizing diverse classifiers for anticipating and diagnosing diabetes.
  + Selecting legitimate classifiers clearly expands the exactness and proficiency of the system.
  + The decision support system is proposed that uses AdaBoost algorithm with Decision Stump as base classifier for classification.
  + Additionally Support Vector Machine, Naive Bayes and Decision Tree are also implemented as base classifiers for AdaBoost algorithm for accuracy verification.
  + The accuracy obtained for AdaBoost algorithm with decision stump as base classifier is 80.72% which is greater compared to that of Support Vector Machine, Naive Bayes and Decision Tree.
  + The dataset has been acquired from UCI machine learning repository. It contains total 16 attributes describing the symptom and the class attributes specifies whether a person is positive or negative. It contains 520 instances approved by doctor

1. P. Suresh Kumar and V. Umatejaswi, “Diagnosing Diabetes using Data Mining Techniques”, International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017 705 ISSN 2250-3153**.**
   * Diabetes is a disease which is affecting many people now-a-days. Most of research is happening in this area.
   * The model to solve the problems in existing system in applying data mining techniques namely clustering and classifications which are applied to diagnose the type of diabetes and its severity level for every patient from the data collected.
   * To diagnose diabetes based on the 650 patient’s data with which we analyzed and identified severity of the diabetes.
   * The procedure Simple k-means algorithm is used for clustering the entire dataset into 3 clusters i.e., cluster-0 - for gestational diabetes, cluster-1 for type-1 diabetes (juvenile diabetes), cluster-2 for type-2 diabetes.
   * The clustered dataset was given as input to the classification model which further classifies each patient’s risk levels of diabetes as mild, moderate, and severe.
   * The number of diabetic persons is increasing in every country, 4 out of 5 people with diabetes live in low- and middle-income countries and half of diabetics don’t know they suffer from this disease.
   * The global epidemic could be largely attributed to the rapid increase in the rates of overweight, obesity and physical inactivity.
   * C4.5 is the best algorithm with less error rate of 0.0938 and more accuracy value of 91%.
   * The association rules for the class value “yes”, and for the class value “no”, the author presented ten association rules. For increasing the dataset quality, the preprocessing methods are applied.
2. 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.
   * Diabetes (Diabetes Mellitus) is a group of metabolic disorders and millions of people are affected. Detection of diabetes is of a great significance and serious complications should be concerned.
   * The Pima Indian diabetes data set is used as it considered reliable.
   * The Pima Indian data set is a data set of studies of women in India's population that began in 1965., and its onset rate is relatively high in diabetes.
   * The most popular techniques in Machine Learning (e.g., KNN algorithm) used to identify the diabetes and pre-processing of data methods.
   * The accuracy of the cross validation on the UCI ML repository data set.
   * The five classification ways are anti-miner, Ad boost, RBF network, CN2 and Bagging for the diabetes prediction.
   * Data is a collection global dataset. Pima Indian data set is used for training a model. Data set contain 21 parameters and around 1000 dataset.
   * The diabetes classification and cardiovascular diseases using BNs and ANN.
   * The diabetes system, such an often blood sugar level and BP measured and hypo glycaemia event.
   * The ML algorithms that are Support Vector machine, NB, KNN and DT to predict the diabetes.
   * Diagnosis of diabetes is made at least 200mg/dL over 2-hours post-load plasma glucose, and various studies regarding the diagnosis of diabetes
3. Berina Alic, Lejla Gurbeta and Almir Badnjevic, “Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases”, 2017 6th Mediterranean Conference on Embeded Computing (MECO), 11-15 JUNE 2017, BAR, MONTENEGRO.
   * Cardiovascular disease (CVD) is an all-encompassing term for situations affecting the heart or blood vessels. This is commonly associated with an accumulation of fatty deposits within the arteries (atherosclerosis) and an increased risk of developing blood clots.
   * Cardiovascular disease is considered one of the largest causes of morbidity and mortality in the world's population. Predicting and diagnosing the disease is a critical challenge in clinical data analysis and health care providers to prevent people from contracting such a disease and conserve lives.
   * Healthcare industries collect massive amounts of data that contain some information related to heart disease diagnosis, which is serviceable in making effective decisions.
   * Furthermore, AI algorithms and deep neural networks can be used to analyze and diagnose heart disease.
   * The project intends to automatically detect cardiovascular disease using two datasets through a deep learning network and a variety of machine learning classification models.
   * The performance evaluated based on the accuracy, precision, recall, and f-score for each of the models. Hence, the Random Forest model achieved the highest performance at 94% accuracy in the heart diseases dataset, while Gradient Boosting model achieved the highest performance at 73% accuracy, 73% Recall, 73% F1-score, and 74% Precision in Cardiovascular Disease Dataset.
   * Each four out of 5 CVD cases are died due to heart attacks and strokes. According to the World Health Organization (WHO), [1] over 17.9 million patients have died from heart diseases around the world, which makes cardiovascular diseases (CVDs) the number one cause of death globally. Cardiovascular diseases are a group of diseases that causes a failure in a human’s heart and blood vessels.
4. Dr. M. Renuka Devi and J. Maria Shyla, “Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus”, International Journal of Applied Engineering Research ISSN 0973- 4562 Volume 11, Number 1 (2016) pp 727-730 © Research India Publications.
   * Based on the test report values, diagnose a potential problem. The patient's report can be entered as feedback by the doctors (Sugar level, Age, Blood pressure, etc.).
   * The diabetes is one of lethal diseases in the world. It is additional a inventor of various varieties of disorders foe example: coronary failure, blindness, urinary organ diseases etc. In such case the patient is required to visit a diagnostic center, to get their reports after consultation.
   * The advanced system mistreatment information processing that has the ability to forecast whether the patient has polygenic illness or not. Furthermore, forecasting the sickness initially ends up in providing the patients before it begins vital.
   * Information withdrawal has the flexibility to remove unseen data from a large quantity of diabetes associated information. The aim of analysis is to develop a system which might predict the diabetic risk level of a patient with a better accuracy.
   * Model development is based on categorization methods as Decision Tree, ANN, Naive Bayes and SVM algorithms.
   * For Decision Tree, the models give precisions of 85%, for Naive Bayes 77% and 77.3% for Support Vector Machine.
   * Outcomes show a significant accuracy of the methods like Decision Tree, ANN, Naive Bayes and SVM algorithms.
5. Rahul Joshi and Minyechil Alehegn, “Analysis and prediction of diabetes diseases using machine learning algorithm”: Ensemble approach, International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct -2017
   * Medical professionals need a reliable prediction methodology to diagnose Diabetes. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information.
   * The main goal of data mining is to discover new patterns for the users and to interpret the data patterns to provide meaningful and useful information for the users.
   * Data mining is applied to find useful patterns to help in the important tasks of medical diagnosis and treatment.
   * This project aims for mining the relationship in Diabetes data for efficient classification. The data mining methods and techniques will be explored to identify the suitable methods and techniques for efficient classification of Diabetes dataset and in mining useful patterns.
   * Machine learning techniques (MLT) are used to predict the medical datasets at an early stage of safe human life**.**
   * Huge medical datasets are accessible in different data repositories which used to in the real-world application.
   * Machine learning (ML) can answer questions. One of the missions is a prediction on disease data.
   * Diabetes Diseases (DD) are among the leading cause of death in the world. To group and predict symptoms in medical data, various data mining techniques were used by different researchers in different time.
   * A total of 768 instances, data set fromPIDD (Pima Indian Diabetes Data Set). In this system the most known predictive algorithms apply KNN, Naïve Bayes, Randomforest, and J48.
   * By using these algorithms make an ensemble hybrid model by combining individual techniques/methods into one in order to increase the performance and accuracy.
6. Zhilbert Tafa and Nerxhivan Pervetica, “An Intelligent System for Diabetes Prediction”, 4th Mediterranean Conference on Embedded Computing MECO – 2015 Budva, Montenegro.
   * Diabetes is a chronic disease and one of deadliest diseases and also a major public health challenge worldwide.
   * Diabetes diseases commonly stated by health professionals or doctors as diabetes mellitus (DM), which describes a set of
   * metabolic diseases in which the person has blood sugar, either insulin production inefficient, or because of the body cell do not
   * return correctly to insulin, or by both reasons. The day is now to prevent and diagnose diabetes in the early stages.
   * It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc.
   * The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports.
   * The discovery of knowledge from medical datasets is important in order to make effective medical diagnosis.
   * Furthermore, predicting the disease early leads to treating the patients before it becomes critical.
   * The primary aim of this project is to analyse the Diabetes Dataset and use Logistic Regression, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors algorithms for prediction and to develop a prediction engine.
   * The secondary aim is to develop a web application with following feature. Allow users to predict diabetes utilizing the prediction engine.
   * The objective is set to achieve the aims of the project through a Research on statistical models in machine learning and to understand how the algorithms works.

#### Eman Marzban and Ayush Patel, “Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay Among Diabetic Patients,13th International Conference on Machine Learning and Applications”,2014.

* + Diabetes has affected over 246 million people worldwide with most of them being women. According to the WHO report, by 2025 this number is expected to rise to over 380 million.
  + The disease has been named the fifth deadliest disease in the United States imminent cure in sight.
  + With the rise of information technology and its continued advent into the medical and healthcare sector, the cases of diabetes as well as their symptoms are well documented.
  + This paper aims at finding solutions to diagnose the disease by analyzing the patterns found in the data through classification analysis by employing Decision Tree and Naïve Bayes algorithms.
  + The research hopes to propose a quicker and more efficient technique of diagnosing the disease, leading to timely treatment of the patients.
  + classification accuracies for these data sets were obtained through using 10-fold cross- validation method.
  + According to the results, the proposed method of performance is highly successful compared to other results attained and seems very promising for pattern recognition applications.
  + According to the WHO (world health organization) report in Nov 14, 2016 in the world diabetes day Eye on diabetes reported 422 million adults are with diabetes, 1.6 million deaths, as the report indicates it is not difficult to guess how much diabetes is very serious and chronic.
  + In 2014, 8.5% of adults whose ages are 18 and older than 18 had diabetes.
  + In 2012 HBG (high blood glucose was the cause of 2.2 million people deaths.

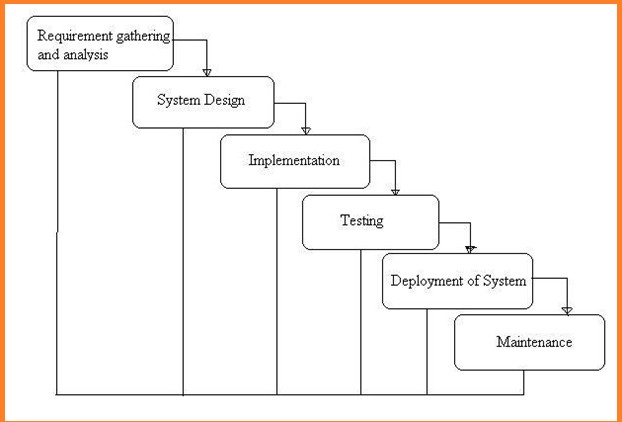
#### Rathore and Simran Chauhan, “Detecting and Predicting Diabetes Using Supervised Learning”. International Journal of Advanced Research in Computer Science, Volume: 08, MayJune 2017.

* + Diabetes has affected over 246 million people worldwide with most of them being women.
  + According to the WHO report, by 2025 this number is expected to rise to over 380 million. The disease has been named the fifth deadliest disease in the United States imminent cure in sight.
  + With the rise of information technology and its continued advent into the medical and healthcare sector, the cases of diabetes as well as their symptoms are well documented.
  + The aim at Detecting Diabetes with PIMA Indian Diabetes Dataset. PIMA India is concerned with women's health.
  + The risk of developing diabetes in Women is quite high due to various factors. Hence, the idea is to Detect and Predict this Disorder with the help of Machine Learning techniques- Support Vector Machine and Decision Trees respectively.
  + The advantage of using these techniques is that it helps in automation of process and makes tasks like Classification, Clustering simpler.
  + The Paper begins with the introduction and emphasize on the worst effect of the Diabetes by explaining various disorders associated with it.
  + According to the centers for disease control and prevention (CDCP) they give information for the duration of 9 ensuing years that is between 2001 and 2009 type II diabetes increased 23% in the United States (US).
  + There are different countries, organization, and different health sectors worry about this chronic disease control and prevent before the person death.

1. Sumi Alice Saji and Balachandran K, “Performance Analysis of Training Algorithms in Diabetes Prediction”, International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India 2015.
   * Diabetes has affected over 246 million people worldwide with most of them being women. According to the WHO report, by 2025 this number is expected to rise to over 380 million.
   * The disease has been named the fifth deadliest disease in the United States imminent cure in sight.
   * With the rise of information technology and its continued advent into the medical and healthcare sector, the cases of diabetes as well as their symptoms are well documented.
   * This paper aims at finding solutions to diagnose the disease by analyzing the patterns found in the data through classification analysis by employing Decision Tree and Naïve Bayes algorithms.
   * The research hopes to propose a quicker and more efficient technique of diagnosing the disease, leading to timely treatment of the patients.
   * No single technique gives highest accuracy or accuracy for all diseases, whereas one classifier provides or shows better performance in a given dataset, another method or approach outdoes the others for other diseases.
   * The new study or the proposed study concentrates on a novel combination of different classifiers for diabetes disease classification and prediction, thus overcoming the problem of individual or single classifiers.

## SYSTEM ANALYSIS AND ARCHITECTURE

### SYSTEM ARCHITECTURE



#### Fig 4.1 Diagram of waterfall model.

**What is Waterfall Model?**

Waterfall Model is a sequential model that divides software development into different phases. Each phase is designed for performing specific activity during SDLC phase.

#### Requirements:

The first phase involves understanding what needs to design and what is its function, purpose, etc. Here, the specifications of the input and output or the final product are studied and marked.

#### System Design:

The requirement specifications from the first phase are studied in this phase & design is prepared. System Design helps in specifying hardware & system requirements helps in defining overall system architecture.

The software code to be written in the next stage is created now.

#### Implementation:

With inputs from system design, the system is first developed in small programs called units, which are integrated into the next phase. Each unit is developed and tested for its functionality which is referred to as Unit Testing.

#### Integration and Testing:

All the units developed in the implementation phase are integrated into a system after testing of each unit. The software designed, needs to go through constant software testing to find out if there are any flaws or errors.

Testing is done so that the client does not face any problem during the installation of the software.

#### Deployment of System:

Once the functional and non-functional testing is done, the product is deployed in the customer environment or released into the market.

#### Maintenance:

This step occurs after installation and involves making modifications to the system or an individual component to alter attributes or improve performance.

These modifications arise either due to change requests initiated by the customer, or defects uncovered during live use of the system.

The client is provided with regular maintenance and support for the developed software.

#### REQUIREMENTS FUNCTIONAL REQUIREMENTS

* + Admin Login
  + Dataset Upload
  + Prediction
  + Classification

#### NON-FUNCTIONAL REQUIREMENTS

**What is Non-Functional Requirement?**

**NON-FUNCTIONAL REQUIREMENT** (NFR) specifies the quality attribute of a software system. They judge the software system based on Responsiveness, Usability, Security, Portability and other non-functional standards that are critical to the success of the software system. Example of nonfunctional requirement, *“how fast does the website load?”* Failing to meet non-functional requirements can result in systems that fail to satisfy user needs.

Non-functional Requirements allows you to impose constraints or restrictions on the design of the system across the various agile backlogs. Example, the site should load in 3 seconds when the number of simultaneous users are > 10000. Description of non-functional requirements is just as critical as a functional requirement.

* + Usability requirement
  + Serviceability requirement
  + Manageability requirement
  + Recoverability requirement
  + Security requirement
  + Data Integrity requirement
  + Capacity requirement
  + Availability requirement
  + Scalability requirement
  + Interoperability requirement
  + Reliability requirement
  + Maintainability requirement
  + Regulatory requirement
  + Environmental requirement

**Examples of Non-functional requirements**

Here, are some examples of non-functional requirement:

1. Users must change the initially assigned login password immediately after the first successful login. Moreover, the initial should never be reused.
2. Employees never allowed to update their salary information. Such attempt should be reported to the security administrator.
3. Every unsuccessful attempt by a user to access an item of data shall be recorded on an audit trail.
4. A website should be capable enough to handle 20 million users with affecting its performance
5. The software should be portable. So moving from one OS to other OS does not create any problem.
6. Privacy of information, the export of restricted technologies, intellectual property rights, etc. should be audited.

**Advantages of Non-Functional Requirement**

Benefits/pros of Non-functional testing are:

* + The nonfunctional requirements ensure the software system follow legal and compliance rules.
  + They ensure the reliability, availability, and performance of the software system
  + They ensure good user experience and ease of operating the software.
  + They help in formulating security policy of the software system.

**Disadvantages of Non-functional requirement**

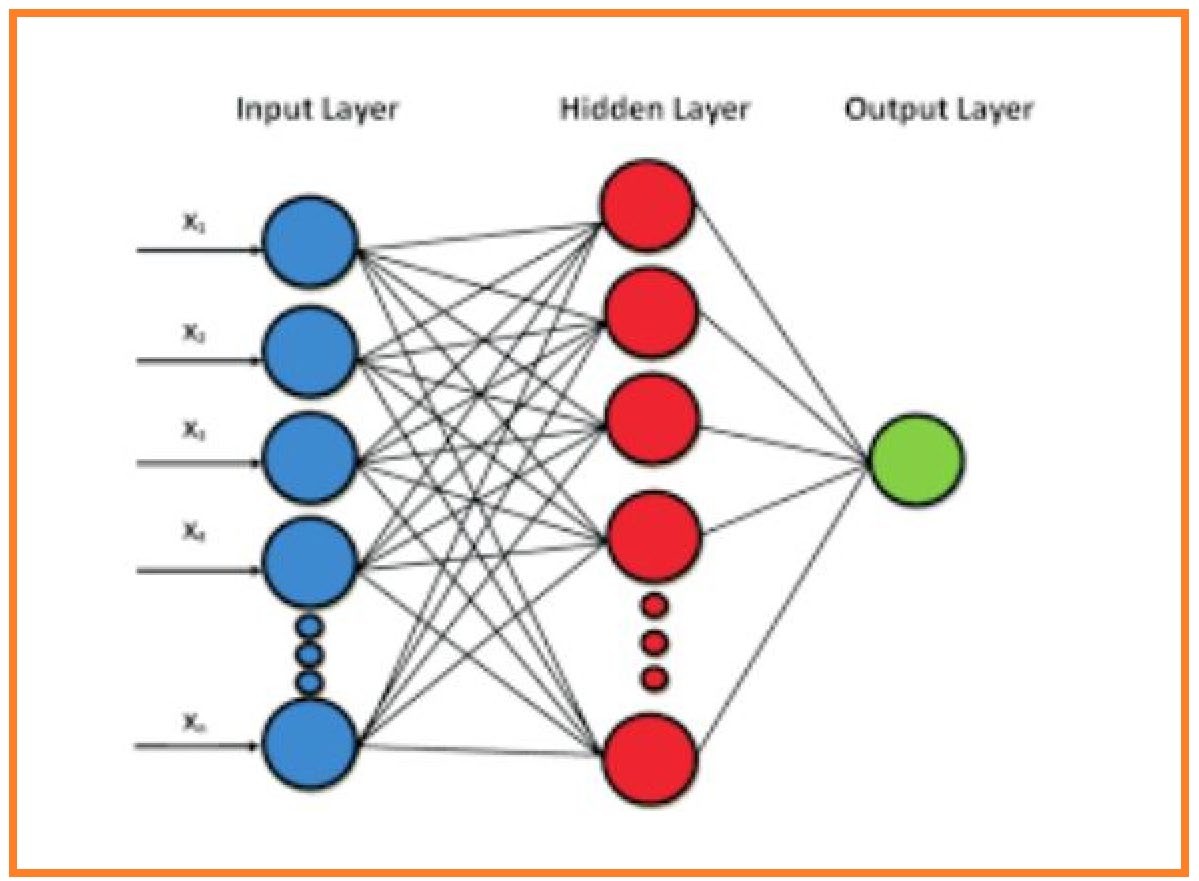
Cons/drawbacks of Non-function requirement are:

* + None functional requirement may affect the various high-level software subsystem
  + They require special consideration during the software architecture/high-level design phase which increases costs.
  + Their implementation does not usually map to the specific software sub-system,
  + It is tough to modify non-functional once you pass the architecture phase.

**KEY LEARNING**

* + A non-functional requirement defines the performance attribute of a software system.
  + Types of Non-functional requirement are Scalability Capacity, Availability, Reliability, Recoverability, Data Integrity, etc.
  + Example of Non-Functional Requirement is Employees never allowed to update their salary information. Such attempt should be reported to the security administrator.
  + Functional Requirement is a verb while Non-Functional Requirement is an attribute
  + The advantage of Non-functional requirement is that it helps you to ensure good user experience and ease of operating the software
  + The biggest disadvantage of Non-functional requirement is that it may affect the various high-level software subsystems.

#### IMPLEMENTATION



**Fig 4.2 Working method of neural network**

Therefore, we can define neural network as information flows from inputs through hidden layers towards the output. For a 3-layers neural network, the learned function would be:

f(x) = f\_3(f\_2(f\_1(x))) where:

f\_1(x): Function learned on first hidden layer

f\_2(x): Function learned on second hidden layer

f\_3(x): Function learned on output layer

Next, we’ll write down the dimensions of a multi-layer neural network in the general form to help us in matrix multiplication because one of the major challenges in implementing a neural network is getting the dimensions right.

#### Multi-Layer Perceptions

The field of artificial neural networks is often just called neural networks or multi-layer perceptron after perhaps the most useful type of neural network.

A perceptron is a single neuron model that was a precursor to larger neural networks.

It is a field that investigates how simple models of biological brains can be used to solve difficult computational tasks like the predictive modeling tasks we see in machine learning.

The goal is not to create realistic models of the brain, but instead to develop robust algorithms and data structures that we can use to model difficult problems.

The power of neural networks come from their ability to learn the representation in your training data and how to best relate it to the output variable that you want to predict.

In this sense neural networks learn a mapping. Mathematically, they can learn any mapping function and have been proven to be a universal approximation algorithm.

The predictive capability of neural networks comes from the hierarchical or multi-layered structure of the networks.

The data structure can pick out (learn to represent) features at different scales or resolutions and combine them into higher-order features.

#### Decision Tree:

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too.

The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data).

**Random Forest: In** our experiment, we use random forest as a classiﬁer. The popularity of decision tree models in data mining is owed to their simpliﬁcation in algorithm and ﬂexibility in handling different data attribute types. However, single-tree model is possibly sensitive to speciﬁc training data and easy to overﬁt.

Ensemble methods can solve these problems by combine a group of individual decisions in some way and are more accurate than single classiﬁers.

Random forest, one of ensemble methods, is a combination of multiple tree predictors such that each tree depends on a random independent dataset and all trees in the forest are of the same distribution.

The capacity of random forest not only depends on the strength of individual tree but also the correlation between different trees. The stronger the strength of single tree and the less the correlation of different tress, the better the performance of random forest.

The variation of trees comes from their randomness which involves bootstrapped samples and randomly selects a subset of data attributes. Below is the step-by-step

step 1:Python implementation. ...

Step 2: Import and print the dataset.

Step 3: Select all rows and column 1 from dataset to x and all rows and column 2 as y.

Step 4: Fit Random Forest regressor to the dataset.

Step 5: Predicting a new result.

Step 6: Visualizing the result.

#### NAÏVE BAYES ALGORITHM

Naïve Bayes classifier is based on Bayes theorem. It has strong independence assumption. It is also known as independent feature model.

It assumes the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature in the given class.

Naïve bayes classifier can be trained in supervised learning setting.

It uses the method of maximum similarity. It has been worked in complex real-world situation. It requires small amount of training data.

It estimates parameters for classification. Only the variance of variable need to be determined for each class not the entire matrix.

Naïve bayes is mainly used when the inputs are high. It gives output in more sophisticated form. The probability of each input attribute is shown from the predictable state.

Machine learning and data mining methods are based on naïve bayes classification.

Bayes theorem:- P(H|X) = P(X|H) P(H)

P(X)

* + Where P(H|X ) is posterior probability of H conditioned on X
  + P(X|H) is posterior probability of X conditioned on H
  + P(H)is prior probability of H P(X) is prior probability of X

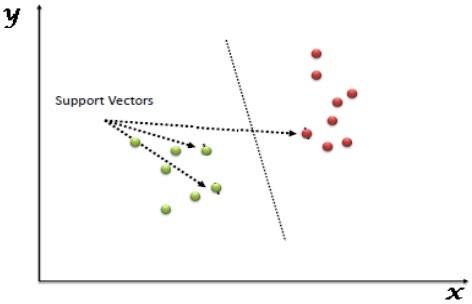
#### Support Vector Machine (SVM):

Support Vector Machine is an extremely popular supervised machine learning technique (having a pre-defined target variable) which can be used as a classifier as well as a predictor.

For classification, it finds a hyper-plane in the feature space that differentiates between the classes.

An SVM model represents the training data points as points in the feature space, mapped in such a way that points belonging to separate classes are segregated by a margin as wide as possible.

The test data points are then mapped into that same space and are classified based on which side of the margin they fall.



#### Fig: 4.3 Support Vector Machine diagram

**4.4 SOFTWARE OVERVIEW:**

**History of Python**

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, and Unix shell and other scripting languages.

Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL).

Python is now maintained by a core development team at the institute, although Guido van Rossum still holds a vital role in directing its progress.

#### Input as CSV File

Reading data from CSV (comma separated values) is a fundamental necessity in Data Science. Often, we get data from various sources which can get exported to CSV format so that they can be used by other systems. The Panadas library provides features using which we can read the CSV file in full as well as in parts for only a selected group of columns and rows.

The CSV file is a text file in which the values in the columns are separated by a comma. Let's consider the following data present in the file named input.csv.You can create this file using windows notepad by copying and pasting this data. Save the file as input.csv using the save As All files (\*.\*) option in notepad.

import pandas aspd

data=pd.read\_csv('path/input.csv') print(data)

#### Operations using NumPy

NumPy is a Python package which stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of array.

Using NumPy, a developer can perform the following operations −

* + Mathematical and logical operations on arrays.
  + Fourier transforms and routines for shape manipulation.
  + Operations related to linear algebra. NumPy has in-built functions for linear algebra and random number generation.

**Key Features of Pandas**

* + Fast and efficient Data Frame object with default and customized indexing.
  + Range of scale is higher in many dimensional plottings.
  + Tools for loading data into in-memory data objects from different file formats.
  + Data alignment and integrated handling of missing data.
  + Reshaping and pivoting of date sets.
  + Label-based slicing, indexing and subsetting of large data sets.
  + Columns from a data structure can be deleted or inserted.
  + Group by data for aggregation and transformations.
  + High performance merging and joining of data.
  + Time Series functionality.
  + Handling of data. The Pandas library provides a really fast and efficient way to manage and explore data.
* Alignment and indexing.
* Handling missing data.
* Cleaning up data.
* Merging and joining of datasets.
  + Multiple file formats supported.
  + Input and output tools.

**Sample Code DBConnection.py** importmysql.connector classDBConnection:

@staticmethod defgetConnection():

database = mysql.connector.connect(host="localhost", user="root", passwd="root", db='diabetes')

return database

if name ==" main\_\_":

print(DBConnection.getConnection())

**K\_Means.py** import random import math

NUM\_CLUSTERS = 2

LOWEST\_SAMPLE\_POINT = 1 # element 0 of SAMPLES. HIGHEST\_SAMPLE\_POINT = 5 # element 5 of SAMPLES.

mid\_value=7 # element 7 of SAMPLES. BIG\_NUMBER = math.pow(10, 10)

classK\_Means: TOTAL\_DATA=0 SAMPLES=[]

data = [] centroids = []

defsetSamples(self,samples):

self.SAMPLES=samples print("samm=", self.SAMPLES)

defgetSamples(self):

returnself.SAMPLES

defsetTotalData(self,total):

self.TOTAL\_DATA=total

defgetTotalData(self):

returnself.TOTAL\_DATA

definitialize\_centroids(self):

self.centroids.clear() self.SAMPLES=self.getSamples();

self.centroids.append(Centroid(self.SAMPLES[LOWEST\_SAMPLE\_POINT])) self.centroids.append(Centroid(self.SAMPLES[HIGHEST\_SAMPLE\_POINT])) print("Centroids initialized at:")

print("(", self.centroids[0].get\_x(), ")")

print("(", self.centroids[1].get\_x(), ")")

#print("(", self.centroids[2].get\_x(), ")") print()

return

definitialize\_datapoints(self):

self.SAMPLES = self.getSamples(); self.TOTAL\_DATA=self.getTotalData() self.data.clear()

# DataPoint objects' x and y values are taken from the SAMPLE array. # The DataPoints associated with LOWEST\_SAMPLE\_POINT and

HIGHEST\_SAMPLE\_POINT are initially

# assigned to the clusters matching the LOWEST\_SAMPLE\_POINT and HIGHEST\_SAMPLE\_POINT centroids.

for i in range(self.TOTAL\_DATA):

newPoint = DataPoint(self.SAMPLES[i]) newPoint.set\_x(self.SAMPLES[i])

if (i ==LOWEST\_SAMPLE\_POINT):

newPoint.set\_cluster(0)

elif (i ==HIGHEST\_SAMPLE\_POINT):

newPoint.set\_cluster(1) elif (i == mid\_value): newPoint.set\_cluster(2) else:

newPoint.set\_cluster(None) self.data.append(newPoint) return self.data

defget\_distance(self,dataPointX, centroidX): # Calculate Euclidean distance.

returnmath.sqrt(math.pow((centroidX - dataPointX), 2))

defrecalculate\_centroids(self,data): totalX = 0

totalY = 0

totalInCluster = 0

for j in range(NUM\_CLUSTERS):

for k in range(len(self.data)):

if (self.data[k].get\_cluster() == j):

totalX += self.data[k].get\_x() totalInCluster += 1

if (totalInCluster> 0):

self.centroids[j].set\_x(totalX / totalInCluster) return

defupdate\_clusters(self):

isStillMoving = 0

self.TOTAL\_DATA = self.getTotalData() for i in range(self.TOTAL\_DATA):

bestMinimum = BIG\_NUMBER currentCluster = 0

for j in range(NUM\_CLUSTERS):

distance =self.get\_distance(self.data[i].get\_x(), self.centroids[j].get\_x()) if (distance <bestMinimum):

bestMinimum = distance currentCluster = j

self.data[i].set\_cluster(currentCluster)

if (self.data[i].get\_cluster() is None or self.data[i].get\_cluster() != currentCluster): self.data[i].set\_cluster(currentCluster)

isStillMoving = 1

returnisStillMoving

defperform\_kmeans(self): isStillMoving = 1

self.initialize\_centroids()

self.data=self.initialize\_datapoints()

while (isStillMoving): self.recalculate\_centroids(self.data)

isStillMoving = self.update\_clusters() return

defprint\_results(self): self.TOTAL\_DATA = self.getTotalData() for i in range(NUM\_CLUSTERS): print("Cluster ", i, " includes:")

for j in range(self.TOTAL\_DATA): if (self.data[j].get\_cluster() == i): print("(", self.data[j].get\_x(), ")") print()

returnself.data

classDataPoint:

x1=0

def init (self, x): self.x1 = x

defset\_x(self, x): self.x1 = x

defget\_x(self):

return self.x1

defset\_cluster(self, clusterNumber): self.clusterNumber = clusterNumber

defget\_cluster(self): returnself.clusterNumber

class Centroid:

def init (self, x): self.x = x

defset\_x(self, x): self.x = x

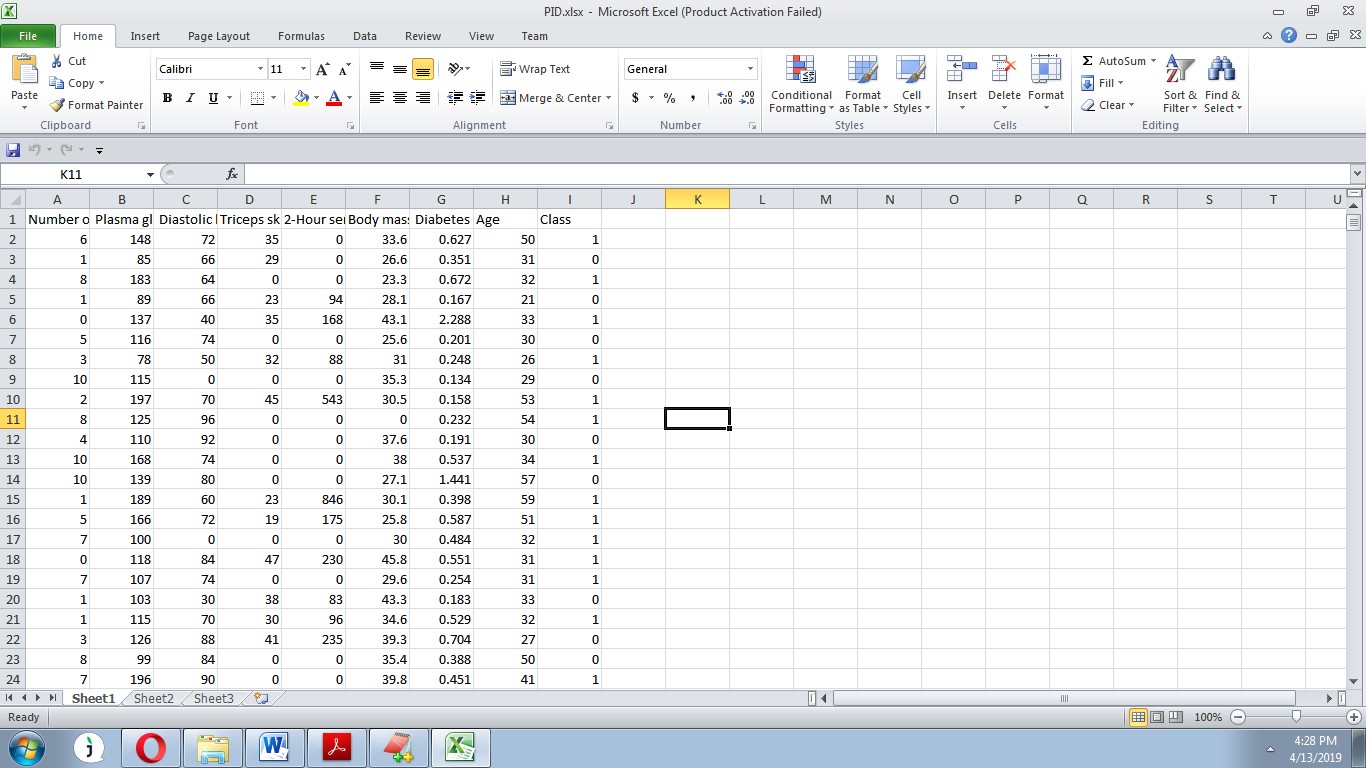
defget\_x(self): returnself.x

if name ==" main\_\_":

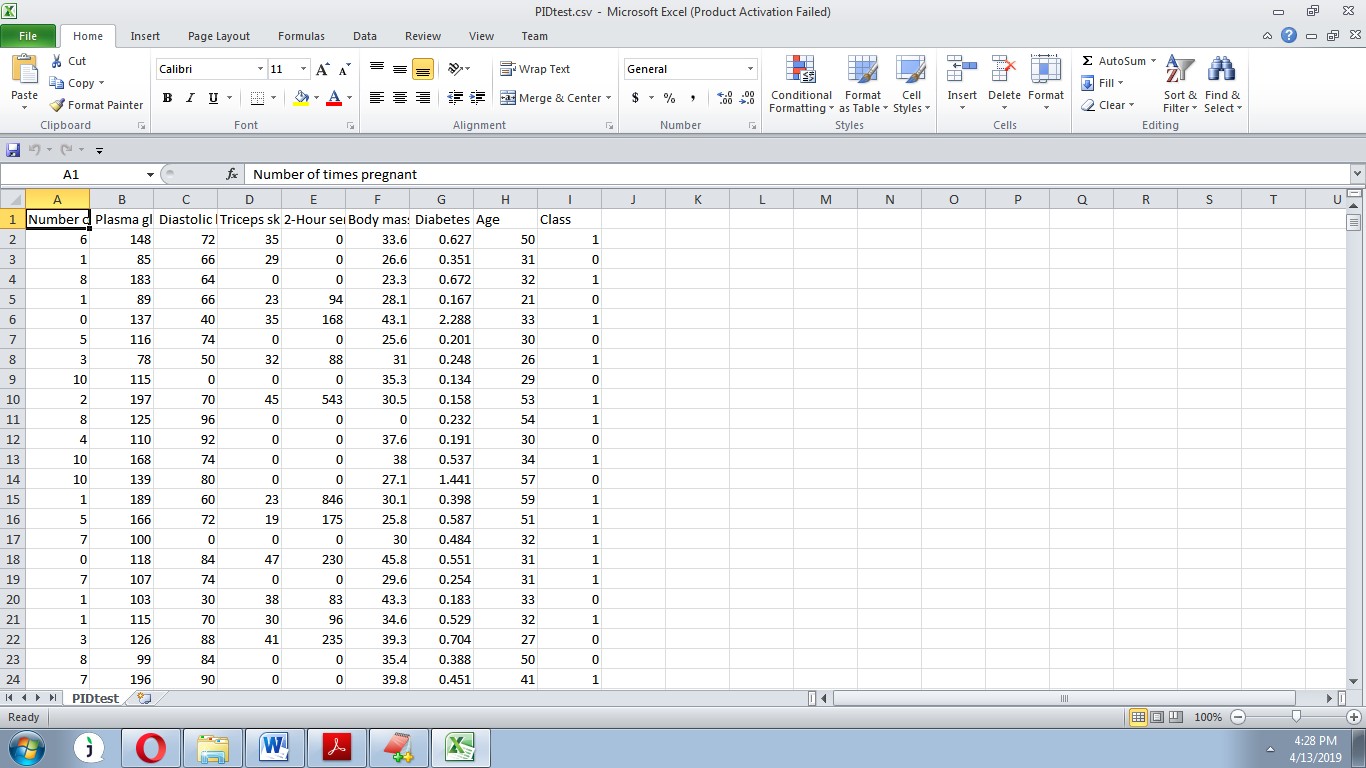
km=K\_Means()

km.perform\_kmeans() km.print\_results()

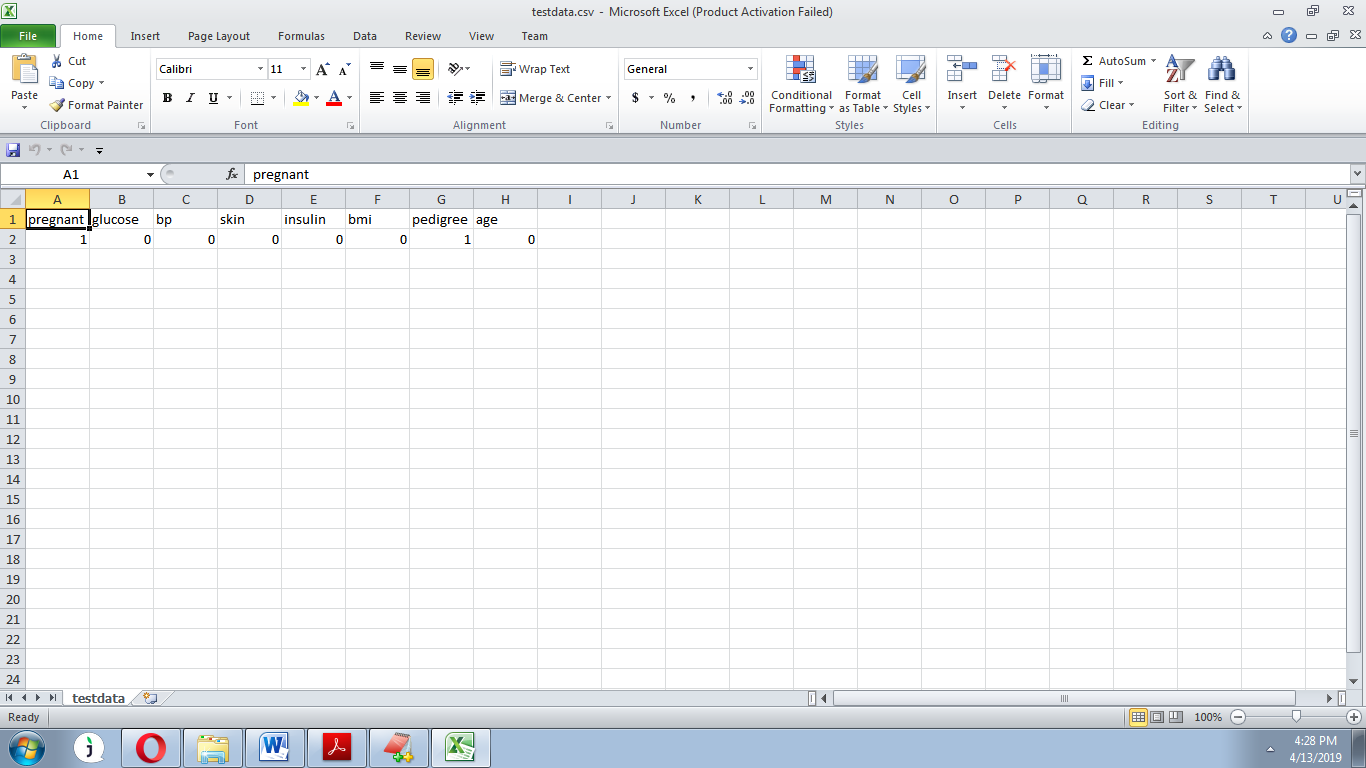
#### Data Sets

**Fig:4.4 PID.xlsx**

**Fig:4.5 PIDtest.csv**



**Fig:4.6 testdata.csv**



**5. Software Design**

* 1. **Intro to software designs:**

**Software design is the process to transform the user requirements into some suitable form, which helps the programmer in software coding and implementation.**

**During the software design phase, the design document is produced, based on the customer requirements as document. Hence the aim of this phase is to transform the SRS document into design document.**

**Software Design concept:**

**Concepts are defined as a principal idea or invention that comes into our minds or in though to understand something.**

**The software design concept simply means the idea or principle behind the design. It describes how you plan to solve the problem of designing software, the logic or thinking behind how you will design software.**

**It allows the system or software or product that is to be developed or built.**

**The software design concept provides a supporting structure or model for developing the right software.**

* 1. **Intro to UML DIAGRAMS:**

The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces.

#### Global Use Case Diagrams:

Identification of actors:

**Actor:** Actor represents the role a user plays with respect to the system. An actor interacts with, but has no control over the use cases.

Graphical representation:

An actor is someone or something that:

Interacts with or uses the system.

* + - Provides input to and receives information from the system.
    - Is external to the system and has no control over the use cases. Actors are discovered by examining:
    - Who directly uses the system?
    - Who is responsible for maintaining the system?
    - External hardware used by the system.
    - Other systems that need to interact with the system. Questions to identify actors:
      * Who is using the system? Or, who is affected by the system? Or, which groups need help from the system to perform a task?
      * Who affects the system? Or, which user groups are needed by the system to perform its functions? These functions can be both main functions and secondary functions such as administration.
      * Which external hardware or systems (if any) use the system to perform tasks?
      * What problems does this application solve (that is, for whom)?
      * And, finally, how do users use the system (use case)? What are they doing with the system?

5.3UML&ClASS Diagram:

The actors identified in this system are:

#### System Administrator

1. **Customer**
2. **Customer Care**
   * 1. **use cases**:

**Use case:** A use case can be described as a specific way of using the system from a user’s (actor’s) perspective.

#### Graphical representation:

A more detailed description might characterize a use case as:

* + - * Pattern of behavior the system exhibits
      * A sequence of related transactions performed by an actor and the system
      * Delivering something of value to the actor Use cases provide a means to:
      * capture system requirements
      * communicate with the end users and domain experts
      * test the system

Use cases are best discovered by examining the actors and defining what the actor will be able to do with the system.

Guidelines for identifying use cases:

* For each actor, find the tasks and functions that the actor should be able to perform or that the system needs the actor to perform. The use case should represent a course of events that leads to clear goal
* Name the use cases.
* Describe the use cases briefly by applying terms with which the user is familiar. This makes the description less ambiguous

Questions to identify use cases:

* What are the tasks of each actor?
* Will any actor create, store, change, remove or read information in the system?
* What use case will store, change, remove or read this information?
* Will any actor need to inform the system about sudden external changes?
* Does any actor need to inform about certain occurrences in the system?
* What usecases will support and maintains the system?

#### Flow of Events

A flow of events is a sequence of transactions (or events) performed by the system. They typically contain very detailed information, written in terms of what the system should do, not how the system accomplishes the task. Flow of events are created as separate files or documents in your favorite text editor and then attached or linked to a use case using the Files tab of a model element. A flow of events should include:

* When and how the use case starts and ends
* Use case/actor interactions
* Data needed by the use case
* Normal sequence of events for the use case
* Alternate or exceptional flows
* Construction of Use case diagrams:Use-case diagrams graphically depict system behavior (use cases). These diagrams present a high level view of how the system is used as viewed from an outsider’s (actor’s) perspective. A use-case diagram may depict all or some of the use cases of a system.

A use-case diagram can contain:

* + actors ("things" outside the system)
  + use cases (system boundaries identifying what the system should do)
  + Interactions or relationships between actors and use cases in the system including the associations, dependencies, and generalizations.

Relationships in use cases:

#### Communication:

The communication relationship of an actor in a usecase is shown by connecting the actor symbol to the usecase symbol with a solid path. The actor is said to communicate with the usecase.

#### Uses:

A Uses relationship between the usecases is shown by generalization arrow from the usecase.

1. **Extends:**The extend relationship is used when we have one usecase that is similar to another usecase but does a bit more. In essence it is like subclass.

#### 5.3.2 SEQUENCE DIAGRAMS

A sequence diagram is a graphical view of a scenario that shows object interaction in a time-based sequence what happens first, what happens next. Sequence diagrams establish the roles of objects and help provide essential information to determine class responsibilities and interfaces.

There are two main differences between sequence and collaboration diagrams: sequence diagrams show time-based object interaction while collaboration diagrams show how objects associate with each other. A sequence diagram has two dimensions: typically, vertical placement represents time and horizontal placement represents different objects.

#### Object:

An object has state, behavior, and identity. The structure and behavior of similar objects are defined in their common class. Each object in a diagram indicates some instance of a class. An object that is not named is referred to as a class instance.

The object icon is similar to a class icon except that the name is underlined: An object's concurrency is defined by the concurrency of its class.

#### Message:

A message is the communication carried between two objects that trigger an event. A message carries information from the source focus of control to the destination focus of control. The synchronization of a message can be modified through the message specification. Synchronization means a message where the sending object pauses to wait for results.

#### Link:

A link should exist between two objects, including class utilities, only if there is a relationship between their corresponding classes. The existence of a relationship between two classes symbolizes a path of communication between instances of the classes: one object may send messages to another. The link is depicted as a straight line between objects or objects and class instances in a collaboration diagram. If an object links to itself, use the loop version of the icon.

#### 5.3.3 CLASS DIAGRAM:

Identification of analysis classes:

A class is a set of objects that share a common structure and common behavior (the same attributes, operations, relationships and semantics). A class is an abstraction of real-world items.

There are 4 approaches for identifying classes:

1. Noun phrase approach:
2. Common class pattern approach.
3. Use case Driven Sequence or Collaboration approach.
4. Classes, Responsibilities and collaborators Approach

#### Noun Phrase Approach:

The guidelines for identifying the classes:

* + Look for nouns and noun phrases in the usecases.
  + Some classes are implicit or taken from general knowledge.
  + All classes must make sense in the application domain; Avoid computer implementation classes – defer them to the design stage.
  + Carefully choose and define the class names After identifying the classes we have to eliminate the following types of classes:
  + Adjective classes.

#### Common class pattern approach:

The following are the patterns for finding the candidate classes:

* + Concept class.
  + Events class.
  + Organization class
  + Peoples class
  + Places class
  + Tangible things and devices class.

#### Use case driven approach:

We have to draw the sequence diagram or collaboration diagram. If there is need for some classes to represent some functionality then add new classes which perform those functionalities.

#### CRC approach:

The process consists of the following steps:

* + Identify classes’ responsibilities ( and identify the classes )
  + Assign the responsibilities
  + Identify the collaborators.

Identification of responsibilities of each class:

The questions that should be answered to identify the attributes and methods of a class respectively are:

1. What information about an object should we keep track of?
2. What services must a class provide? Identification of relationships among the classes:

Three types of relationships among the objects are:

Association: How objects are associated?

Super-sub structure: How are objects organized into super classes and sub classes? Aggregation: What is the composition of the complex classes?

Association:

The **questions** that will help us to identify the associations are:

* 1. Is the class capable of fulfilling the required task by itself?
  2. If not, what does it need?
  3. From what other classes can it acquire what it needs? Guidelines for identifying the tentative associations:
* A dependency between two or more classes may be an association. Association often corresponds to a verb or prepositional phrase.
* A reference from one class to another is an association. Some associations are implicit or taken from general knowledge.

Some common association patterns are:

Location association like part of, next to, contained in….. Communication association like talk to, order to ……

We have to eliminate the unnecessary association like implementation associations, ternary or n- ary associations and derived associations.

Super-sub class relationships:

Super-sub class hierarchy is a relationship between classes where one class is the parent class of another class (derived class).This is based on inheritance.

Guidelines for identifying the super-sub relationship, a generalization are

1. **Top-down*:***Look for noun phrases composed of various adjectives in a class name. Avoid excessive refinement. Specialize only when the sub classes have significant behavior.

#### Bottom-up*:*

Look for classes with similar attributes or methods. Group them by moving the common attributes and methods to an abstract class. You may have to alter the definitions a bit.

#### Reusability*:*

Move the attributes and methods as high as possible in the hierarchy.

#### Multiple inheritances*:*

Avoid excessive use of multiple inheritances. One way of getting benefits of multiple inheritances is to inherit from the most appropriate class and add an object of another class as an attribute.

#### Aggregation or a-part-of relationship:

It represents the situation where a class consists of several component classes. A class that is composed of other classes doesn’t behave like its parts. It behaves very difficultly. The major properties of this relationship are transitivity and anti symmetry.

The **questions** whose answers will determine the distinction between the part and whole relationships are:

* + Does the part class belong to the problem domain?
  + Is the part class within the system’s responsibilities?
  + Does the part class capture more than a single value?( If not then simply include it as an attribute of the whole class)
  + Does it provide a useful abstraction in dealing with the problem domain? There are three types of aggregation relationships. They are:

#### Assembly:

It is constructed from its parts and an assembly-partsituation physically exists.

#### Container:

A physical whole encompasses but is not constructed from physical parts.

#### Collection member:

A conceptual whole encompasses parts that may be physical or conceptual. The container and collection are represented by hollow diamonds but composition is represented by solid diamond.

**Admin User**

System

**Login**

**Prediction System**

**Upload Data Set**

**K-Means**

**No of Clustures**

**DecisionTree**

**Prediction Test**

**Result**

**Logout**

#### Figure:5.1 use case Diagram



Login

Login

Upload Data Set

Prediction Test

K-Means

Result

No of Clustures

Logout

DecisionTree

Logout

User

Admin

**Fig:5.2 state chart**



Login

Upload Data Set

K-Means

No of Clustures

DecisionTree

Prediction Test

Result

Exit

: Admin

1 : UserName()

: User

2 : Password()

3 : File()

4 : Apply()

5 : Enter()

6 : Apply()

7 : Exit()

8 : UserName()

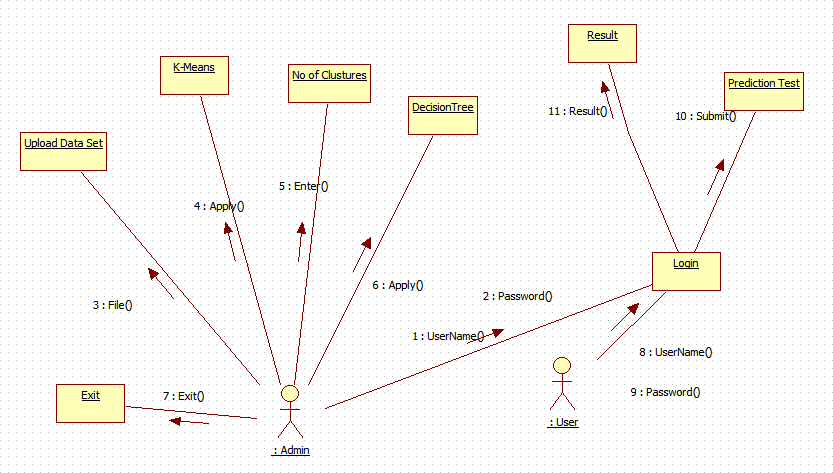
9 : Password()

10 : Submit()

11 : Result()

12 : Exit()

**Fig:5.3Sequence diagram**



**Fig:5.4 Activity Diagram**

+Submit()

+Admin

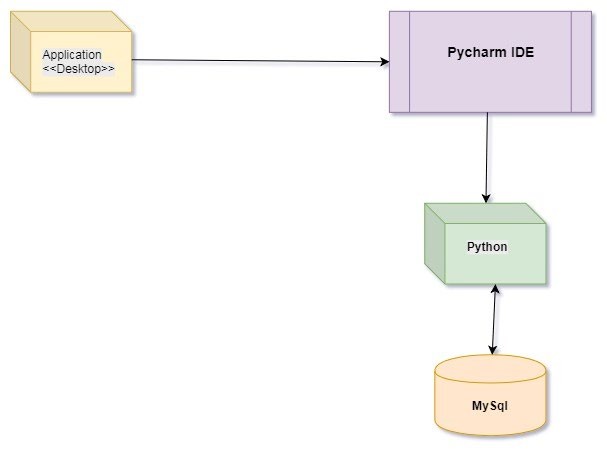
+User

**Login.**

|  |
| --- |
| **DataSet** |
| +Number of times pregnant  +Plasma glucose  +Diastolic blood pressure  +Triceps skin fold thickness  +2-Hour serum insulin  +Body mass index  +Diabetes pedigree function  +Age  +Class |
| +Upload()  +K-Means()  +DecisionTree() |

|  |
| --- |
| **Prediction** |
|  |
| +Prediction Test()  +Result() |

#### Fig:5.5 class diagram



**Fig:5.6 Diagram of deployment**

1. **Software Testing**
   1. **Software Testing**

Software testing is one of the main stages of project development life cycle to provide our cessation utilizer with information about the quality of the application and ours, in our Project we have under gone some stages of testing like unit testing where it’s done in development stage of the project when we are in implementation of the application after the Project is yare we have done manual testing with different Case of all the different modules in the application we have even done browser compatibility testing in different web browsers in market, even we have done Client side validation testing on our application

#### Unit testing

The unit testing is done in the stage of implementation of the project only the error are solved in development stage some of the error we come across in development are given below **TESTING**

Testing is the debugging program is one of the most critical aspects of the computer programming triggers, without programming that works, the system would never produce an output of which it was designed. Testing is best performed when user development is asked to assist in identifying all errors and bugs. The sample data are used for testing. It is not quantity but quality of the data used the matters of testing. Testing is aimed at ensuring that the system was accurately an efficiently before live operation commands.

#### Testing objectives:

The main objective of testing is to uncover a host of errors, systematically and with minimum effort and time. Stating formally, we can say, testing is a process of executing a program with intent of finding an error.

A successful test is one that uncovers an as yet undiscovered error.

A good test case is one that has probability of finding an error, if it exists. The test is inadequate to detect possibly present errors.

The software more or less confirms to the quality and reliable standards.**Levels of Testing:**

In order to uncover present in different phases we have the concept of levels of testing.

#### Levels of Testing:

**The basic levels of Testing:**

~~Client nee~~ds acceptance testing

~~Requirem~~ents system testing

~~Design~~  integration testing

Code unit testing

#### Figure:6.1 Levels of Testing

**Code testing:**

This examines the logic of the program. For example, the logic for updating various sample data and with the sample files and directories were tested and verified.

#### Specification Testing:

Executing this specification starting what the program should do and how it should perform under various conditions. Test cases for various situation and combination of conditions in all the modules are tested.

#### Unit testing:

In the unit testing we test each module individually and integrate with the overall system. Unit testing focuses verification efforts on the smallest unit of software design in the module. This is also known as module testing. The module of the system is tested separately. This testing is carried out during programming stage itself. In the testing step each module is found to work satisfactorily as regard to expected output from the module. There are some validation checks for fields also. For example the validation check is done for varying the user input given by the user which validity of the data entered. It is very easy to find error debut the system.

Each Module can be tested using the following two Strategies:

1. Black Box Testing
2. White Box Testing **BLACK BOX TESTING** **What is Black Box Testing?**

Black box testing is a software testing technique in which **functionality of the software under test (SUT) is tested without looking at the internal code structure**, implementation details and knowledge of internal paths of the software. This type of testing is based entirely on the software requirements and specifications.

**In Black Box Testing we just focus on inputs and output of the software system** without bothering about internal knowledge of the software program.



#### Fig 6.2 black box diagram

The above Black Box can be any software system you want to test. For example : an operating system like Windows, a website like Google ,a database like Oracle or even your own custom application. Under Black Box Testing, you can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation.

#### Black box testing - Steps

Here are the generic steps followed to carry out any type of Black Box Testing.

* Initially requirements and specifications of the system are examined.
* Tester chooses valid inputs (positive test scenario) to check whether SUT processes them correctly. Also some invalid inputs (negative test scenario) are chosen to verify that the SUT is able to detect them.
* Tester determines expected outputs for all those inputs.
* The test cases are executed.
* Software tester compares the actual outputs with the expected outputs.
* Defects if any are fixed and re-tested.

#### Types of Black Box Testing

There are many types of Black Box Testing but following are the prominent ones -

* **Functional testing** – This black box testing type is related to functional requirements of a system; it is done by software testers.
* **Non-functional testing** – This type of black box testing is not related to testing of a specific functionality, but non-functional requirements such as performance, scalability, usability.
* **Regression testing** – Regression testing is done after code fixes , upgrades or any other system maintenance to check the new code has not affected the existing code.

#### WHITE BOX TESTING

White Box Testing is the testing of a software solution's internal coding and infrastructure.It focuses primarily on strengthening security, the flow of inputs and outputs through the application, and improving design and usability.White box testing is also known as **clear, open, structural, and glass box testing**.

It is one of two parts of the **"box testing" approach** of software testing. Its counterpart, Blackbox testing, involves testing from an external or end-user type perspective. On the other hand, Whitebox testing is based on the inner workings of an application and revolves around internal testing. The term "whitebox" was used because of the see-through box concept. The clear box or whitebox name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "black box testing" symbolizes not being able to see the inner workings of the software so that only the end-user experience can be tested

#### What do you verify in White Box Testing ?

White box testing involves the testing of the software code for the following:

* Internal security holes
* Broken or poorly structured paths in the coding processes
* The flow of specific inputs through the code
* Expected output
* The functionality of conditional loops
* Testing of each statement, object and function on an individual basis

The testing can be done at system, integration and unit levels of software development. One of the basic goals of whitebox testing is to verify a working flow for an application. It involves testing a series of predefined inputs against expected or desired outputs so that when a specific input does not result in the expected output, you have encountered a bug.

#### How do you perform White Box Testing?

To give you a simplified explanation of white box testing, we have divided it into **two basic steps**. This is what testers do when test an application using the white box testing technique:

#### STEP 1) UNDERSTAND THE SOURCE CODE

The first thing a tester will often do is learn and understand the source code of the application. Since white box testing involves the testing of the inner workings of an application, the tester must be very knowledgeable in the programming languages used in the applications they are testing. Also, the testing person must be highly aware of secure coding practices. Security is often one of the primary objectives of testing software. The tester should be able to find security issues and prevent attacks from hackers and naive users who might inject malicious code into the application either knowingly or unknowingly.

#### Step 2) CREATE TEST CASES AND EXECUTE

The second basic step to white box testing involves testing the application’s source code for proper flow and structure. One way is by writing more code to test the application’s source code. The tester will develop little tests for each process or series of processes in the application. This method requires that the tester must have intimate knowledge of the code and is often done by the developer. Other methods include manual testing, trial and error testing and the use of testing tools as we will explain further on in this article.

#### System testing:

Once the individual module testing is completed, modules are assembled and integrated to perform as a system. The top down testing, which began from upper level to lower level module, was carried out to check whether the entire system is performing satisfactorily.

There are three main kinds of System testing:

1. Alpha Testing
2. Beta Testing
3. Acceptance Testing

#### Alpha Testing:

This refers to the system testing that is carried out by the test team with the Organization.

#### Beta Testing:

This refers to the system testing that is performed by a selected group of friendly customers

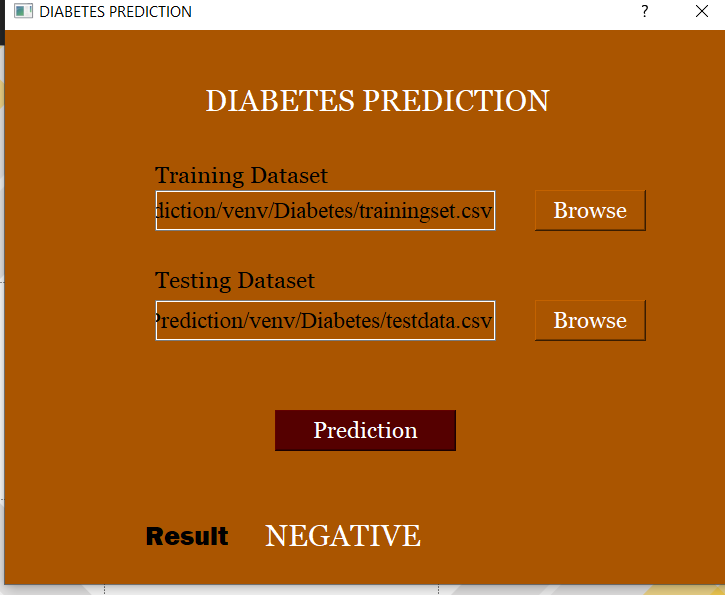
#### Acceptance Testing:

This refers to the system testing that is performed by the customer to determine whether or not to accept the delivery of the system.

#### Integration Testing:

Data can be lost across an interface, one module can have an adverse effort on the other sub functions, when combined, may not produce the desired major functions. Integrated testing is the systematic testing for constructing the uncover errors within the interface. The testing was done with sample data. The developed system has run successfully for this sample data. The need for integrated test is to find the overall system performance.

**Output testing:**After performance of the validation testing, the next step is output testing. The output displayed or generated by the system under consideration is tested by asking the user about the format required by system.



#### Fig 6.3 Diagram of result display

**7.CONCLUSION**

**7.1 Conclusion:**

* A fast and accurate diabetes prediction system is proposed in the paper.
* The proposed system used 768 instances within 8 attributes for each one of PID dataset.
* The used data is preprocessed in order to remove the unwanted data, and lead to faster processing time.
* Moreover, the dividing technique of the dataset into subset, made an optimal classification result. The proposed system focused on the features analysis and classification parts.
* The propositions of these parts lead to an optimal achievement.
* The results of experiments illustrated the effects of using the algorithms of the proposed system through achieving a higher classification rate that the other systems.

### 8.REFERENCES

#### 8.1References:

1. Yilmaz N., Inan O., Uzer M.S., “ A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases,” J Med Syst, vol. 38, no. 5 2014.
2. Lowongtrakool C., Hiransakolwong N., “Noise filtering in unsupervised clustering using computation intelligence,” International Journal of Math, vol. 6, no. 59, pp. 2911–2920, 2012.
3. V. Anuja and R.Chitra., “Classification Of Diabetes Disease Using Support Vector Machine”, International Journal of Engineering Research and Applications (IJERA), vol.3,Issue 2, pp. 1797- 1801, 2013.
4. Aiswarya I., S. Jeyalatha and Ronak S., “Diagnosis Of Diabetes Using Classification Mining Techniques”, International Journal of Data Mining & Knowledge Management Process (IJDKP), vol.5, ,No. 1, pp. 1-14, 2015.
5. K.Rajesh and V.Sangeetha,”Application of Data Mining Methods and Techniques for Diabetes Diagnosis,” in proceedings of International journal of Engineering and Innovative Technology, vol.2, Issue 3, pp. 43-46, 2012.
6. Harleen and Dr. Pankaj B.,”A Prediction Technique in Data Mining for Diabetes Mellitus,” Journal of Management Sciences and Technology, vol. 4, Issue 1, pp. 1-12, 2016.
7. Ravi S. and Smt T., ”Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine,” International Journal of Computer Trends and Technology (IJCTT), vol. 11, No. 2, pp. 94-98, 2014.
8. G. Krishnaveni\*, T. Sudha,” A Novel Technique To Predict Diabetic Disease Using Data Mining Classification Techniques” in International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT- 2017), vol. 3, Issue 1, pp. 5-11, 2017.
9. Raj A., Vishnu P., and Kavita B.,”K-Fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data Set Using Higher Order Neural Network and PCA”, International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-6, pp. 436-438, January 2013.
10. Vrushali B., and Rakhi W., “Review on Prediction of Diabetes using Data Mining Technique”, International Journal of Research and Scientific Innovation (IJRSI), Volume IV, Issue IA, pp. 43-46, January 2017.