

**MULTIMEDIA UNIVERSITY OF KENYA**

FACULTY OF COMPUTING & INFORMATION TECHNOLOGY

PROJECT DOCUMENTATION

DISEASE PREDICTION SYSTEM

BY

**NAME: KIPYEGON JEMOSOP MERCY  
REG. No: CIT-227-011/2015**

**SUPERVISOR: Mr. James Adunya**

**4th April 2019**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

This documentation is submitted in partial fulfillment of the requirements of Forth Year Bachelor of Science in Software Engineering

# DECLARATION

I hereby declare that this Project proposal has my own work, to the best of my knowledge, and has not been submitted to any other institution of higher learning.

**Student: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Registration Number: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature: ............................................... Date: .....................................................**

This project proposal has been submitted as a partial fulfillment of the requirements for the Bachelor of Science in Software Engineering in Multimedia University of Kenya with my approval as the University supervisor.

**Supervisor: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature: ..................................................... Date: ..................................................**

# ACKNOWLEDGEMENT

This project has taken a lot efforts. However, it would not have been possible to complete this project without the kind support and help from my family, friends, classmates and lectures. Sincere thanks goes to you all.

We are highly indebted to Mr. James Adunya for his guidance and constant supervision as well as for providing necessary information regarding the project and for his support in completing the project.

Our thanks and appreciations go to our classmates. Above all, we would like to express our gratitude to Almighty God for providing us all these great people and the strength to complete this project.

# ABSTRACT

This system is a predictive model which predicts the disease of the user on the basis of the symptoms that user provides as an input to the system. The system analyzes the symptoms provided by the user and gives the probability of the disease as an output. It acts as an intelligent doctor.

The main objective of this research is to develop an Intelligent System using data mining modeling technique, namely, Naive Bayes. It is implemented as web based application where user answers the predefined questions. It then compares users data with the trained datasets and predicts the patient disease. The data is then send to the data to the database from where it will be retrieved when generating the user report. It can answer complex queries for diagnosing diseases and thus assist users who can’t afford the high cost of medical services to know the disease he/she is suffering from hence purchase the right medication. This helps reduce cases of drug abuse where people take drugs without knowing the specific disease he/she is suffering from. The present system lacks an effective analysis tools to discover hidden relationships and trends in data hence don’t take advantage of the available data. Advanced data mining techniques can help remedy this situation. It enables significant knowledge, e.g. patterns, relationships between medical factors related to the disease, to be established.

The cost of medication has increased tremendously which has deprived many patient access to medication. When a patient visits the hospital they are charged consultation to see a doctor which is also separate from diagnostic fee, all this charges makes patients shy away from hospital because they can’t afford the high medical charges and self-prescribe medication which is drug abuse and is not advice able.

The long queues in hospitals caused by lack of enough medical practitioners is also a major problem in Kenya. It might have happened so many times that you or someone need doctor’s help, but they are not available due to some reason and they cannot get to the hospital since people are always busy chasing money and forget about their health status hence the need of the intelligent system which will predict the disease you are likely to suffer from.

This system will help in the treatment process and allow faster diagnosis of patient since treatment can be done online and the doctor can only prescribe medication and refer you to the lab if necessary over the internet from any location. This will greatly reduce congestions in hospitals. Some life threatening diseases have similar symptoms with common diseases which if not handled on time may lead to death or paralysis of the body, this system will also help differentiate the diseases with similar symptoms.

This system is designed to also assist doctors and health professionals in determining the diagnosis of patient data. Doctors and health professionals use their knowledge and experience to make decision for the diagnosis of various disease for patients. Therefore, this system could help doctors and health professionals to determine the diagnosis and analysis of the patient health status. It can be used to support education for the undergraduate and postgraduate young physicians as a tool to improve the quality of care for the patients. This system can also be used as a reference for those student and new doctor. Presently, doctors have difficulties in determining disease in a new patient who does not have existing medical record. Therefore, those data can be used to diagnose a disease for new patients who do not have existing medical records.

# Table of Contents

[DECLARATION i](#_Toc5573159)

[ACKNOWLEDGEMENT ii](#_Toc5573160)

[ABSTRACT iii](#_Toc5573161)

[Table of Contents vi](#_Toc5573162)

[ABBREVIATION AND ACRONYMS ix](#_Toc5573163)

[List of Figures x](#_Toc5573164)

[List of Tables xi](#_Toc5573165)

[CHAPTER 1.0 INTRODUCTION 1](#_Toc5573166)

[1.1 Background of study 1](#_Toc5573167)

[1.2 Problem statement 1](#_Toc5573168)

[1.3 Aim of the study 1](#_Toc5573169)

[1.3.1 Research objectives 1](#_Toc5573170)

[1.4 Significance/Justification of the study 2](#_Toc5573171)

[1.5 Scope (i.e. defines the system boundary) 2](#_Toc5573172)

[1.6 Assumptions 2](#_Toc5573173)

[1.7 Limitations ( i.e. challenges and counter measures) 2](#_Toc5573174)

[2.0 Chapter 2 LITERATURE REVIEW 3](#_Toc5573175)

[2.1 Introduction 3](#_Toc5573176)

[2.2 Related systems (2-3) 3](#_Toc5573177)

[2.3 Limitations/Weaknesses of these systems 8](#_Toc5573178)

[2.4 How your proposed solution will handle these weaknesses*.* 8](#_Toc5573179)

[3.0 CHAPTER 3 METHODOLOGY 10](#_Toc5573180)

[3.1 Introduction 10](#_Toc5573181)

[3.2 The Methodology (stating the methodology, its description and justification of using this methodology.) 10](#_Toc5573182)

[3.3 Data collection (needs assessment) methods and tools 14](#_Toc5573183)

[Chapter 4 – System Analysis 16](#_Toc5573184)

[4.0 Introduction 16](#_Toc5573185)

[4.1 Detailed analysis of current system using flow charts, DFDs, UML, and Context diagrams, etc**.** 16](#_Toc5573186)

[Activity diagram 16](#_Toc5573187)

[4.2 System requirements 18](#_Toc5573188)

[4.2.1 Functional requirements 18](#_Toc5573189)

[4.2.2 Non-functional requirements 20](#_Toc5573190)

[Chapter 5 System Design 21](#_Toc5573191)

[Introduction 5.0 Detailed design of the proposed system using tools such as ERDs, DFDs, UML, etc. 21](#_Toc5573192)

[5.1 Architectural design 21](#_Toc5573193)

[**5.2 Database design** 27](#_Toc5573194)

[**5.3 User interface design** 28](#_Toc5573195)

[5.3.1 Signup 28](#_Toc5573196)

[5.3.2 Login 29](#_Toc5573197)

[5.3.3 Symptom form 30](#_Toc5573198)

[5.3.4 Results 31](#_Toc5573199)

[5.3.5 Final report 32](#_Toc5573200)

[Chapter 6 Implementation and testing 34](#_Toc5573201)

[6.0 Introduction 34](#_Toc5573202)

[6.1 Development environment 34](#_Toc5573203)

[**PHP (Hypertext Preprocessor)** 37](#_Toc5573204)

[ADVANTAGES OF PHP 37](#_Toc5573205)

[6.2 System components 39](#_Toc5573206)

[6.3 Test Plan (test data, test cases and test results) 40](#_Toc5573207)

[Chapter 7: Conclusion 42](#_Toc5573208)

[7.1 Achievements and lessons learnt 42](#_Toc5573209)

[7.2 Conclusions 42](#_Toc5573210)

[7.3 Recommendations 42](#_Toc5573211)

[Reference 43](#_Toc5573212)

[Appendix 44](#_Toc5573213)

[Appendix IUser manual 44](#_Toc5573214)

[Appendix IIData collections tools 44](#_Toc5573215)

[**Appendix III** Project schedule 46](#_Toc5573216)

[Appendix IV Project budget 46](#_Toc5573217)

[Appendix V Any other relevant document. 47](#_Toc5573218)

# 

# ABBREVIATION AND ACRONYMS

CHD- Coronary heart disease

CSS- Cascading Style Sheets

ECP- Extracorporeal Photo apheresis

ERD- Entity Relationship Diagram.

HTML- hypertext markup language.

ID3- Iterative Dichotomized 3

IHDS- Intelligent Heart Disease Prediction System

MDL- minimum description length

MMU-Multimedia University of Kenya.

PHP- Hypertext Preprocessor

# List of Figures

[Figure 1: Activity Diagram 26](#_Toc5561107)

[Figure 3: UML diagram 31](#_Toc5561108)

[Figure 4: block diagram 32](#_Toc5561109)

[Figure 5: State Diagram 33](#_Toc5561110)

[Figure 6: sequence diagram 34](#_Toc5561111)

[Figure 7: Database tables 35](#_Toc5561112)

[Figure 8: signup form 37](#_Toc5561113)

[Figure 9: Login figure 38](#_Toc5561114)

[Figure 10: symptom form 39](#_Toc5561115)

[Figure 11: Prediction Graph 40](#_Toc5561116)

[Figure 12: Final Report 40](#_Toc5561117)

# List of Tables

[Table 1: disease table 35](#_Toc5560687)

[Table 2: Users table 36](#_Toc5560688)

[Table 3: Relational table 36](#_Toc5560689)

[Table 4: Unit testing 47](#_Toc5560690)

[Table 5: Integration table 47](#_Toc5560691)

[Table 6: Project schedule 52](#_Toc5560692)

# CHAPTER 1.0 INTRODUCTION

## 1.1 Background of study

At present, when one suffers from particular disease, the person has to visit to doctor which is time consuming caused by the long queues experienced in hospitals and costly because of the high medical charges. If the user is out of reach of doctor and hospitals it may be difficult for the user to get professional advice on the disease they are suffering from hence patient may experience both physical and mental trauma. This system eliminates this problem by bringing the doctor closer to the patient by automating the diagnosis process and creating an intelligent doctor.

This system compares the visible symptoms and predict the likely hood of a disease occurring. As the use of internet is growing every day more people have access to internet than hospitals and doctors. So, this system can be helpful to the people as they have access to internet 24 hours.

Disease Prediction system has data sets collected from different health related sites. With the help of these system, the user will be able to know the probability of the disease with the given symptoms occurring and go for necessary test and scanning to confirm the results. This system will help in the treatment process and allow faster diagnosis of patient since treatment can be done online and the doctor can only prescribe medication and refer you to the lab if necessary over the internet from any location

## 1.2 Problem statement

Online consultation plays a big role in bringing the doctor closer to patients. It might have happened so many times that you or someone need doctor’s help immediately, but they are not available due to some reason. People cannot identify their symptoms and take medicines without consulting doctors. Some medicines are very much harmful to health. So user needs online consultation

## 1.3 Aim of the study

To design a health prediction system for medical data classification and early disease by using naïve Bayes algorithm.

### 1.3.1 Research objectives

To predict diseases given symptoms using naïve Bayes algorithm.

To save on time on the diagnostic process and reduce the cost of medication using a web interface platform for the prediction of the disease.

To increase the diagnostic accuracy and reduces human effort.

## 1.4 Significance/Justification of the study

To reduces cases of misdiagnosis and improve services offered to patients in hospitals.

## 1.5 Scope (i.e. defines the system boundary)

This project aims to provide a web platform to predict the occurrences of disease on the basis of various symptoms. The user can input various symptoms and can find the diseases which matches the symptoms based on their probabilistic figures.

## 1.6 Assumptions

The patients/ users visiting the system will be computer literate.

The users of the application will have access to internet.

## 1.7 Limitations ( i.e. challenges and counter measures)

The limitations of this project are:

a. Disease Predictor does not recommend medications of the disease.

b. Past history of the disease has not been considered

# 2

# .0 Chapter 2 LITERATURE REVIEW

## 2.1 Introduction

This chapter elaborates the literature survey of the project and the projects existing projects that are similar with the project being done. This systems uses data although data mining has been around for more than two decades, its potential is only being realized now. Fayyad defines data mining as “a process of non-trivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database”. It combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., k-means clustering is unsupervised). Each data mining technique serves a different purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms

## 2.2 Related systems (2-3)

Data mining is widely used for prediction in many fields, medical being one of them. The more the documented data the more it can be manipulated in predictions. Many diseases are being predicted by data mining techniques with high efficiency. A study on prediction of breast cancer shows that. It can be predicted with high accuracy using different techniques [8]. Imagers of cancer cells at different stages help to show the patients conditions and even see if any particular drug works on the person. Heart disease prediction have also been studied in many cases.

Jyoti Soni in her paper showed how the performance of decision tree and Bayesian classification can be improved after using genetic algorithm. The paper first shows the analysis before using genetic algorithm and the analysis after [9]. The method used was classification via clustering. The classification was performed based on clustering. Sellappan Palaniappan and Rafiah Awang built a model of Intelligent Heart Disease Prediction System (IHDS) with the use of data mining techniques namely, Neural Network, Naïve Bayes and Decision Tree [10]. Using medical profiles such as age, sex, blood pressure and blood sugar the system can predict the likelihood of patients getting a heart disease. IHDPS was capable of responding to “what if” queries that the usual decision support systems were not able to. It exploits the data using knowledge such as patterns, relationships amid medical factors connected with heart disease. The IHDS is a Web based, user-friendly, scalable, reliable and expandable system.

In 2013, another paper was published which shows the diagnosis of lung cancer prediction using data mining classification techniques One Dependency Augmented Naïve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2) which are extensions of Bayesian classifier improved to work on data set which are small or incomplete [11]. Using generic lung cancer symptoms such as age, sex, Wheezing, Shortness of breath, Pain in shoulder, chest, arm, it can predict the likelihood of patients getting a lung cancer disease. The paper mainly aims to make a model to provide early detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient.

In 2014, M.A.Nishara Banu B.Gomathy Professor, Department of Computer Science and Engineering has published a research paper “Disease Forecasting System Using Data Mining Methods” [12]. In this paper data was clustered using clustering algorithms like K-means to cluster relevant data in database. Maximal Frequent Itemset Algorithm (MAFIA) is used for mining maximal frequent patterns in heart disease database. Their result showed that the designed system was capable of predicting heart attacks successfully.

The many systems used in predicting diseases inspired this research. Some of the motivated works are stated above. These prove that diseases can be predicted successfully using various data mining techniques.

K. Srinivas et al.

Presented application of Data Mining Technique in Healthcare and Prediction of Heart Attacks. The potential use of classification based data mining techniques such as rule based Decision Tree, Naïve Bayes and Artificial Neural Network to the massive volume of healthcare data. Tanagra data mining tool was used for exploratory data analysis, machine learning and statistical learning algorithms. The training data set consisted of 3000 instances with 14 different attributes. The instances in the dataset were representing the results of different types of testing to predict the accuracy of heart disease. The performance of the classifiers was evaluated and their results were analyzed. The results of comparison were based on 10 tenfold cross-validations. The comparison made among these classification algorithms out of which the Naive Bayes algorithm showed better performance.

Jyoti Sonia, et.al.

In year 2011 presented three classifiers Decision Tree, Naïve Bayes and Classification via clustering to diagnose the presence of heart disease in patients. Classification via clustering: Clustering is the process of grouping similar elements. This technique may be used as a preprocessing step before feeding the data to the classifying model. Experiments were conducted with WEKA 3.6.0 tool. Data set of 909 records with 13 different attributes. All attributes were made categorical and inconsistencies were resolved for simplicity. To enhance the prediction of classifiers, genetic search was incorporated. Observations exhibit that the Decision Tree data mining technique outperforms other two data mining techniques after incorporating feature subset selection but with high model construction time.

Chaitrali S. Dangare and Sulabha S. Apte

They showed that Artificial Neural Network outperforms other data mining techniques such as Decision Tree and Naïve Bayes. In this research work, Heart disease prediction system was developed using 15 attributes. The research work included two extra attributes obesity and smoking for efficient diagnosis of heart disease in developing effective heart disease prediction system.

[1] The prediction of survival of Coronary heart disease (CHD) has been a challenging research problem for medical society. The goal of this paper is to develop data mining algorithms for predicting survival of CHD patients based on 1000 cases .We carry out a clinical observation and a 6-month follow up to include 1000 CHD cases. The survival information of each case is obtained via follow up. Based on the data, we employed three popular data mining algorithms to develop the prediction models using the 502 cases. We also used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. The results indicated that the SVM is the best predictor with 92.1 % accuracy on the holdout sample artificial neural networks came out to be the second with91.0% accuracy and the decision trees models came out to be the worst of the three with 89.6% accuracy. The comparative study of multiple prediction models for survival of CHD patients along with a 10-fold cross-validation provided us with an insight into the relative prediction ability of different data.

Decision tree is one kind of inductive learning algorithms that offers an efficient and practical method for generalizing classification rules from previous concrete cases that already solved by domain experts. It is considered attractive for many real-life applications, mostly due to its interpretability. Recently, many researchers have been reported to endow decision trees with incremental learning ability, which is able to address the learning task with a stream of training instances. However, there are few literatures discussing the algorithms with incremental learning ability regarding the new attributes. In this paper, I + Learning (Intelligent, Incremental and Interactive Learning) theory is proposed to complement the traditional incremental decision tree learning algorithms by concerning new available attributes in addition to the new incoming instances. The experimental results reveal that I + Learning method offers the promise of making decision trees a more powerful, flexible, accurate and valuable paradigm, especially in medical data mining community.

The data mining comprises of analysis of large data from various perspectives and obtaining summary of useful information. The information can be transferred into knowledge regarding future trends and history. Data mining has a very important role in the information technology domain. Huge amounts of complex data is generated by health care sector today. These data includes details about diseases, patients, diagnosis methods, electronic patients details hospitals resources etc,. The data mining methods are very helpful in making medicinal decisions in disease curing. The vast data collected by healthcare industry are not mined and hence information is hidden. And as a result the decision making is not effective. The knowledge discovered can be used by the healthcare administrators for enhancing the service quality. In this paper, a method for identifying frequency of diseases in particular geographical location for a given period of time using Apriori data mining technique based on association rules is proposed.

Medical data mining has been a popular data mining topic of late. Especially, diagnosing of the heart disease is one of the important issue and many researchers investigated to develop intelligent medical decision support systems to help the physicians. In this paper, we propose the use of decision tree C4.5 algorithm, bagging with decision tree C4.5 algorithm and bagging with Naïve Bayes algorithm to identify the heart disease of a patient and compare the effectiveness, correction rate among them. The data we study is collected from patients with coronary artery disease.

Naive Bayes classifier has gained wide popularity as a probability-based classification method despite its assumption that attributes are conditionally mutually independent given the class label. This paper makes a study into discretization techniques to improve the classification accuracy of Naive Bayes with respect to medical datasets. Our experimental results suggest that on an average, with minimum description length (MDL) discretization the Naive Bayes classifier seems to be the best performer compared to popular variants of Naive Bayes as well as some popular non-Naive Bayes statistical classifiers.

Clinical databases store large amounts of information about patients and their medical conditions. Data mining techniques can extract relationships and patterns holding in this wealth of data, and thus be helpful in understanding the progression of diseases and the efficacy of the associated therapies. A typical structure of medical data is a sequence of observations of clinical parameters taken at different time moments. In this kind of contexts, the temporal dimension of data is a fundamental variable that should be taken in account in the mining process and returned as part of the extracted knowledge. Therefore, the classical and well established framework of sequential pattern mining is not enough, because it only focuses on the sequentially of events, without extracting the typical time elapsing between two particular events. Time-annotated sequences (IAS), is a novel mining paradigm that solves this problem. Recently defined in our laboratory together with an efficient algorithm for extracting them, IAS are sequential patterns where each transition between two events is annotated with a typical transition time that is found frequent in the data. In this paper we report a real-world medical case study, in which the IAS mining paradigm is applied to clinical data regarding a set of patients in the follow-up of a liver transplantation. The aim of the data analysis is that of assessing the effectiveness of the extracorporeal photo apheresis (ECP) as a therapy to prevent rejection in solid organ transplantation. For each patient, a set of biochemical variables is recorded at different time moments after the transplantation. The IAS patterns extracted show the values of interleukins and other clinical parameters at specific dates, from which it is possible for the physician to assess the effectiveness of the ECP therapy. We believe that this case study does not only show the interestingness of extracting IAS patterns in this particular context but, more ambitiously, it suggests a general methodology for clinical data mining, whenever the time dimension is an important variable of the problem in analysis.

## 2.3 Limitations/Weaknesses of these systems

The accuracy level of the existing system was a little bit lower than that of naïve Bayes hence could lead to misdiagnosis.

The existing system predicted only one disease hence a patient had to go to different sites to search for different diseases which was tiresome.

These systems does not recommend a doctor to a patient hence the patient had to search for a doctor by themselves.

## 2.4 How your proposed solution will handle these weaknesses*.*

The proposed system is a web based system, this will enable patients to easily access medical services from the comfort of their home at a cheaper price. The patient won’t be charged with any consultation fee. Patient will be required to register and login into the system in order to access services provided. Patients will provide their symptoms to the system in order to predict their disease. The patient can then search for a freelancer doctor who will be suggested by the system for prescription and advanced care.

This system will help in the treatment process and allow faster diagnosis of patient since treatment can be done online and the doctor can only prescribe medication and refer you to the lab if necessary over the internet from any location. This will greatly reduce congestions in hospitals. This is system can also reduce cases of drug abuse where patients self-prescribe medication without knowing the real disease affecting them and worsen their condition though it is not advice able to prescribe medication without the doctor’s advice since some life threatening diseases have similar symptoms with common diseases which if not handled on time may lead to death or paralysis of the body. This system will also help detect here a patient.

This system can also be used to support education for the undergraduate and postgraduate young physicians as a tool to improve the quality of care for the patients. This system can also be used as a reference for those student and new doctor. Presently, doctors have difficulties in determining disease in a new patient who does not have existing medical record. Therefore, those data can be used to diagnose a disease for new patients who do not have existing medical records.

This system is also designed to assist doctor and health professionals in determining the diagnosis of patient data. Doctors and health professionals use their knowledge and experience to make decision for the diagnosis of various disease for patients. Therefore, this system could help doctors and health professionals to determine the diagnosis and analysis of the patient health status. This system will be helpful in identifying early signs of diseases which will help get rid of the disease from the infected person on time. Terminal illnesses like cancer can be treated and eliminated at the early stages, which will reduce cost of medication and fatality of the diseases.

People will be visiting the site and fill in their personal information and the system will determine if the symptoms match the disease the patient is searching, if the results are positive the patient can seek immediate medical attention.

# 3.0 CHAPTER 3 METHODOLOGY

## 3.1 Introduction

Disease Prediction has been already implemented using different techniques like Neural Network, decision tree and Naïve Byes algorithm. Particularly heart related disease is mostly analyzed. From the analysis it was found that Naïve Bayes is more accurate than other techniques. So, Disease Predictor also uses Naïve Bayes for the prediction of different diseases.

## 3.2 The Methodology (stating the methodology, its description and justification of using this methodology.)

1. Decision tree algorithm

Decision trees are one of the most regularly used techniques of data analysis. Decision trees are easy to visualize and understand and resistant to noise in data. Generally, decision trees are used to classify records to a proper class. Besides, they are applicable in both regression and associations tasks. In the medical field decision trees specify the sequence of attributes values and a decision that is based on these attributes. One of the most popularly used decision tree algorithm is Iterative Dichotomized 3 (ID3). Quinlan introduced ID3 algorithm. The algorithm is based on Occam’s razor, which means that the smaller trees are preferred. The Occam’s razor is formalized using information entropy concept. The construction of a tree is top-down and start with the appropriate attribute for the root node.

2. Neural networks

Artificial neural networks are analytical techniques that are formed on the basis of superior learning processes in the human brain. As the human brain is capable to, after the learning process, draw assumptions based on earlier observations, neural networks are also capable to predict changes and events in the system after the process of learning. Neural networks are groups of connected input/output units where each connection has its own weight. The learning process is performed by balancing the net on the basis of relations that exist between elements in the examples. Based on the significance of cause and effect between certain data, stronger or weaker connections between "neurons" are being formed. Network formed in this manner is ready for the unknown data and it will react based on previously acquired knowledge. One of the key advantages of Artificial Neural Networks is their high performance. The core function of Artificial Neural Networks is prediction. One of most popular algorithm of neural network is back propagation algorithm [6]. Rojas [2005] claimed that Back Propagation algorithm could be broken down to four main steps. After choosing the weights of the network randomly, the back propagation algorithm is used to compute the necessary corrections.

3.  SVM Algorithm

SVM has attracted a great deal of attention in the last decade. It also applied to various domains of applications. SVMs are used for learning classification, regression or ranking function. SVM is based on statistical learning theory and structural risk minimization principle. And have the aim of determining the location of decision boundaries. It is also known as a hyper plane. That produces the optimal separation of classes. Thereby creating the largest possible distance between the separating hyper plane. Further, the instances on either side of it have been proven. That is to reduce an upper bound on the expected generalization error. The efficiency of SVM based does not depend on the dimension of classified entities. Though, SVM is the most robust and accurate classification technique. Also, there are several problems. The data analysis in SVM is based on convex quadratic programming. Also, expensive, as solving quadratic programming methods. That need large matrix operations as well as time-consuming numerical computations. Training time for SVM scales in the number of examples. So researchers strive all the time for more efficient training algorithm. That resulting in several variant based algorithm.

SVM can also extend to learn non-linear decision functions. That is by first projecting the input data onto a high-dimensional feature space. As by using kernel functions and formulating a linear classification problem. The resulting feature space is much larger than the size of a dataset. That is not possible to store on popular computers.

Investigation of this issues leads to several decomposition based algorithms. The basic idea of decomposition method is to split the variables into two parts:

a set of free variables called as a working set. That can update in each iteration and set of fixed variables. That are fix during a particular. Now, this procedure have to repeat until the termination conditions are met the SVM was developed for binary classification. And it is not simple to extend it for multi-class classification problem. The basic idea to apply multi-classification to SVM. That is to decompose the multi-class problems into several two-class problems. That can address using several SVMs.

4. Sense Clusters (an adaptation of the K-means clustering algorithm)

We have made use of Sense Clusters to classify the email messages. Sense Cluster available package of Perl programs. As it was developed at the University of Minnesota Duluth. That we use for automatic text and document classification. The advantage of sense Clusters is that it does not need any training data;

It makes use of unsupervised learning methods to classify the available data. Now, particularly in this section will understand the K-means clustering algorithm. That has been used in sense Clusters. Clustering is the process in which we divide the available data. That instances of a given number of sub-groups. These sub-groups are clusters, and hence the name “Clustering”. To put it, the K-means algorithm outlines a method. That is to cluster a particular set of instances into K different clusters. Where K is a positive integer. It should notice K-means clustering algorithm requires a number of clusters from the user. It cannot identify the number of clusters by itself. However, sense Clusters has the facility of identifying the number of clusters. That the data may comprise of the K-means clustering algorithm starts by placing K centroids. Then each of the available data instances has to assign a particular centroid. That depends on a metric like Euclidian distance, Manhattan distance, Minkowski distance, etc.

The position of the centroid has to recalculate every time an instance is added to the cluster. This continues until all the instances are group into the final required clusters.

5. Naïve Bayes algorithm

Naive Bayes algorithm outperforms most of the sophisticated algorithms. It is a good tool in medical diagnosis. For example given a list of symptoms, it predicts occurrence of a disease. Naïve Bayes assumes its attributes to be conditionally independent. The classifier computes the probability of each attribute in a class. The result of the classification is the class with the highest posterior probability. Posterior probability is proportional to product of prior probability & like hood. Naïve Bayes main strength is its simplicity, efficiency and good classification performance. It combines efficiency with good accuracy. Due to its good accuracy it is used in medical diagnosis. A small amount of training data is required for the estimation of variable values necessary for classification. Naïve Bayes is a very powerful technique in diagnosing diseases. It is used to provide efficient output with attributes independent to each other. The Naïve Bayes classifier needs a very large training set to obtain good results. Bayes' theorem can be used to compute the probability that a proposed diagnosis is correct, given that observation.

The Bayesian classification is a statistical technique which follows supervised learning method. It determines the probabilities of the outputs. It solves the predictive problems. It provides a model for understanding many learning algorithms. And this algorithm computes the probabilities for the problems.

Steps:

Step 1: Data Preparation

The data set obtained in this study was obtained from kaggle.com website. Initially the size of data was 120.

Step 2: Handle Data

The first thing we need to do is load our data file. The data is in CSV format without a header line or any quotes. We can open the file with the open function and read the data lines using the reader function in the CSV module.

Step 3: Data Transformation

In this step, the dataset is explored and necessary data is selected and the dataset is converted into machine understandable form.

Step 4: Summarize the Data

The summary of the training data collected involves the mean and the standard deviation for each attribute, by class value. These are required when making predictions to calculate the probability of specific attribute values belonging to each class value.

Step 5: Feature Extraction

Feature extraction is the process to reduce the size of data so as to only take informative, non-redundant and relevant data, so as to facilitate subsequent learning and generalization step to acquire better human interpretation.

Step 6: Making Predictions

We are now ready to make predictions using the summaries prepared from our training data. Making predictions involves calculating the probability that a given data instance belongs to each class, then selecting the class with the largest probability as the prediction. Then we define our main function where we call all these methods we have defined, one by one to get the accuracy of the model we have created.

Advantages of Naïve Bayes

1. Very simple, easy to implement and fast.
2. If the NB conditional independence assumption holds, then it will converge quicker than discriminative models like logistic regression.
3. Even if the NB assumption doesn’t hold, it works great in practice.
4. Need less training data.
5. Highly scalable. It scales linearly with the number of predictors and data points.
6. Can be used for both binary and multiclass classification problems.
7. Can make probabilistic predictions.
8. Handles continuous and discrete data.
9. Not sensitive to irrelevant features.

## 3.3 Data collection (needs assessment) methods and tools

Internet, Data collection has been done from the internet to identify the disease, the real symptoms of the disease are collected i.e

. no dummy values are entered. The symptoms of the disease are collected from different health related websites. Example kaggle.com which provides a wide variety of data sets.

Questionnaires have advantages over some other types of medical symptoms that they are cheap, do not require as much effort from the questioner as verbal or telephone surveys, and often have standardized answers that make it simple to compile data. However, such standardized answers may frustrate users. Questionnaires are also sharply limited by the fact that respondents must be able to read the questions and respond to them.

# Chapter 4 – System Analysis

## 4.0 Introduction

It is the process of collecting and interpreting facts, identifying the problem and decomposing of a system into components. It is a problem solving technique that improves the system and ensures that all the components of the system work efficiently to accomplish their purpose. This chapter describes the system study, analysis, design, strength and weakness of the current system. In this study stage, requirements are categorized into user requirements, system and hardware requirements.

## 4.1 Detailed analysis of current system using flow charts, DFDs, UML, and Context diagrams, etc**.**

The existing system uses the traditional method of diagnosis, where the doctor diagnoses patient based on the knowledge he/she has acquired during the period of their profession. It is also difficult to get medical attention when a person is not in the hospital physically. This system is time consuming and also its accuracy level is very low since it is determined by the doctor in charge their judgment on a disease.

UML DIAGRAM

## Activity diagram

Activity diagram shows the state of an object and represent activities as arrows connecting the states. It highlights the activities taking place. Each activity is represented by a rounded rectangle and more oval shaped than the state icon. Arrow represents the transition from one activity to the next. The activity diagram starting point is represented by a filled rectangle and an end-point by bull’s eye.

In the existing system a user has to be physically present in the hospital for him/her to be diagnosed and long process carried out in the hospital discourages many people from going to the hospital when they have common symptoms like fever and they self-medicate themselves.

This traditional method is time consuming especially to business people who value money than their health.

Figure : Activity Diagram

Enter Hospital

Check in at the reception

Undergo diagnosis

Take treatment

Take medication

Lab Test

Print Report

Laboratory

## 4.2 System requirements

After analyzing the data collected, we formulated a number of requirements namely user requirement, system hardware software attribute. These were grouped as functional and non-functional systems requirements

**Hardware Requirements for the project**

The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware

Processor: Intel dual Core

i3RAM:4GB

HARDDISK: 750GB

**Software Requirement for the project**

Software Requirements deal with defining software resource requirements and pre-requisites that need to be installed on a computer to provide optimal functioning of an application.

Operating System: windows 10.

FRONT END: HTML, CSS, JAVASCRIPT.

SERVER SIDE SCRIPT: phpDataBase, MySQL, Jet Brains PyCharm Community Edition

## 4.2.1 Functional requirements

User Module

1. Registration module

This is done by all the system users when it is their first time to use the system. The user enters their personal details in a registration form which is then stored in the database. Information users provide include:

Name, age, gender, email, phone number etc.

1. Login module

After registration user can login to the system using the information they provided during registration.

1. Symptom Module

When a user logs into the system, they are provided with a form in which they are to enter the symptoms they are suffering from in order for the system to predict their disease.

1. Feedback report

After diagnosis is done the use is provided with feedback in form of a report indicating the disease they are suffering from.

1. Comment Module

User can give their views about their experience while using the system.

Administrator Module

1. Add training data

Administrator can add the training data to the dataset with time when data becomes available or when a disease is being search by many users.

1. View Details

Administrator can view user’s details both personal and the disease they are suffering from.

1. View Feedback

Administrator can view comments given by user’s in order to determine the strengths and weakness of the system.

1. View Training data

Administrator can view training data if the need arises in the system or for clarification purposes.

## 4.2.2 Non-functional requirements

These are requirements that are not directly concerned with the specific services delivered by the system to its users. The system must verify and validate all user input and users must be notified in case of errors detected in the course of using the system.

1. Performance requirement.

The system will give a fast response after checking the user’s information.

The system user interface screen will be reliable and give fast response.

1. Availability.

The system shall be readily all the time for use and not experience delays.

1. Ensure security.

* The system will be password protected : user enters name and password to be verified in order to access the system
* Each user of the system will have login id and password in order to prevent unauthorized users from accessing the system.
* Only the administrator can insert, delete, update, user’s information and system information.
* Administrator will have the highest level of access in the system.

1. Reliability

This system is reliable and produces fast and verified output of all its processes.

1. Maintainability

This system will be designed in a maintainable manner. It will be easy to in-cooperate new requirements in the individual module

# Chapter 5 System Design

## Introduction 5.0 Detailed design of the proposed system using tools such as ERDs, DFDs, UML, etc.

Our application will be at affordable cost. Naïve Bayes Machine Learning Algorithm predicts Diseases based on symptoms provided. This algorithm is implemented to increase operational efficiency, reduces Query retrieval time and improved Accuracy using Machine Learning algorithm. The proposed system will be a simple web interface where user will interact with the system by registering and logging into the system. The user will be provided with a form where he/she will answer some predefined questions which are to be filled with the symptoms experienced by the user. These user input will be used to predict the disease the user is suffering from.

## 5.1 Architectural design

The general disease prediction system predicts chance of presence of a disease present in a patient on the basis of their symptoms. It will also recommend necessary precautionary measures required to treat the predicted disease. The system will initially be fed data from different sources i.e. patients, the data will then be pre-processed before further process is carried out, this is done so as to get clean data from the raw initial data, as the raw data would be noisy, or flawed. This data will be processed using Data mining algorithms, the system, will be trained so as to predict the disease based on the input data given by the user.

UML DIAGRAM

This is a graphical representation of a set of elements, mostly represented by connected graph of vertices and arcs.

1. Flow Chart

It represent the flow from one activity to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another.

Standard symbols used in DFDs

Source link/external entity- A source of system input or sink of system output.

Process- performs some transformation to input data to output data.

Data store- a repository of data. Arrowheads indicate net inputs or net output to the store.

Figure : flow diagram

Specify symptoms

User

Enter input inform of symptom

Admin

Disease prediction System

Results

Login

Valid/invalid

Valid/invalid

Analysis

It consists of a three module i.e. user, system and Admin. Admin can maintains the flow of a system, He can add, remove training data as well as patients. Admin also handles the database of system. Patient is a prior user of system, They can specify the symptoms from which they are suffers and system sends this message to doctor site, Then doctor give the prescription on their health issue which can be received by the patient.

1. Block Diagram for General disease prediction system

It is a diagram of a system in which the principal parts or functions are represented by blocks connected by lines that show the relationships of the blocks. They are heavily used in engineering in hardware design, electronic design, software design, and process flow diagrams.

Datasets are trained and transformed into meaningful information which can be used for prediction. Processing of data involves cleaning and removing unnecessary information on the data and also removing incomplete data. Data mining involves training the dataset to learn certain patterns which will be used in the prediction process.

Figure : block diagram

Training Data

Data Transformation

Processing data

Data mining

Predicted disease

User input

Disease prediction Model

1. **States Diagram**

It explains different state of the system. First the user opens /logs into the Disease Predictor System. The user selects the symptoms they are suffering from. When finished selecting symptoms the user submits the symptoms into the system for analysis and prediction of the diease. Disease Predictor analyzes the symptoms and displays the result.

Figure : State Diagram

Login into disease predictor system

Enter form with symptoms

submit input to the system

Analyze the symptoms

Display Disease

1. Sequence diagram

It explains the sequence of the Disease Predictor. Initially system shows the symptoms to be selected. The user selects the symptoms and submits to the system .The Disease Predictor predicts and display the result.

Figure : sequence diagram

Select symptoms and submit

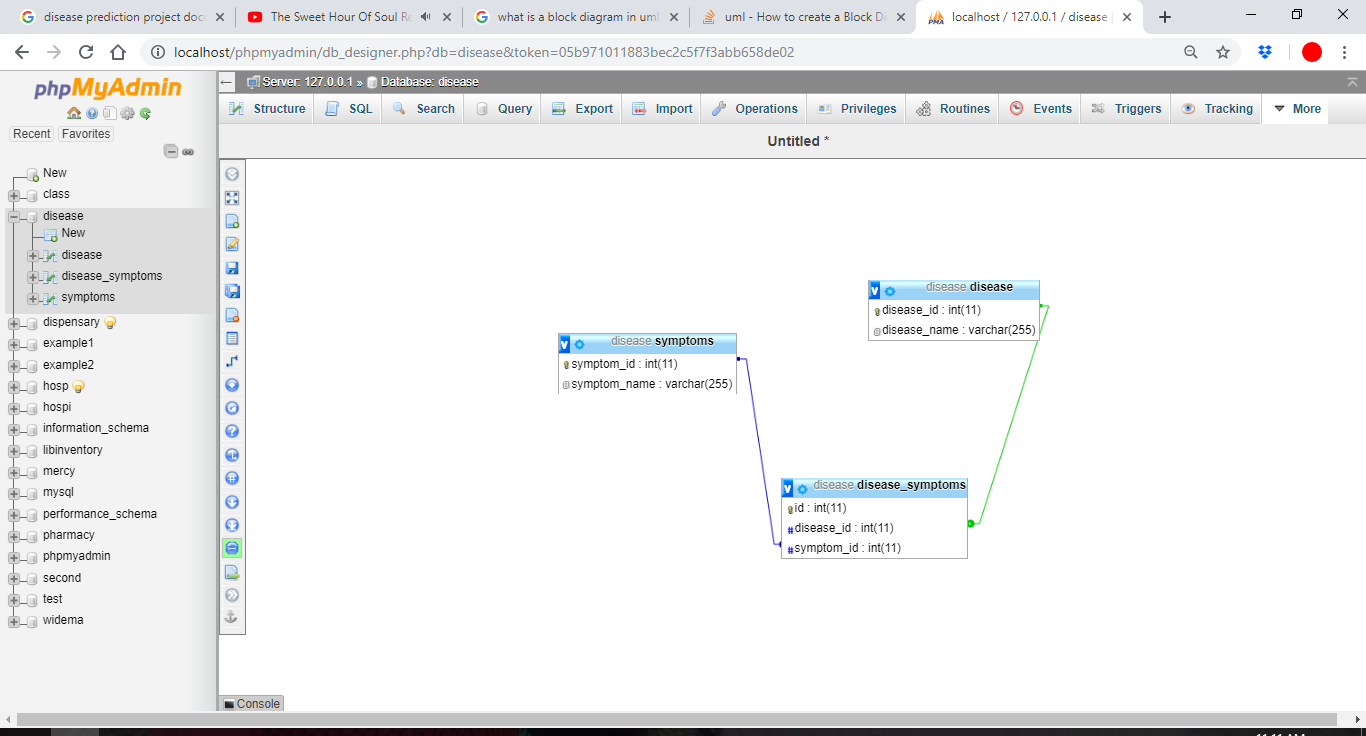
Display Disease

user

Disease predictor

## **5.2 Database design**

Figure : Database tables



Disease table

Table : disease table

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Type | Size | Description |
| disease\_id | int | 11 | Primary key of disease table |
| Disease name | varchar | 255 | Name of the disease |

Table : Users table

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Type | Size | Description |
| Symptom\_id | Int | 11 | Symptom primary key |
| Symptom\_name | Varchar | 255 | Symptom name |

Table : Relational table

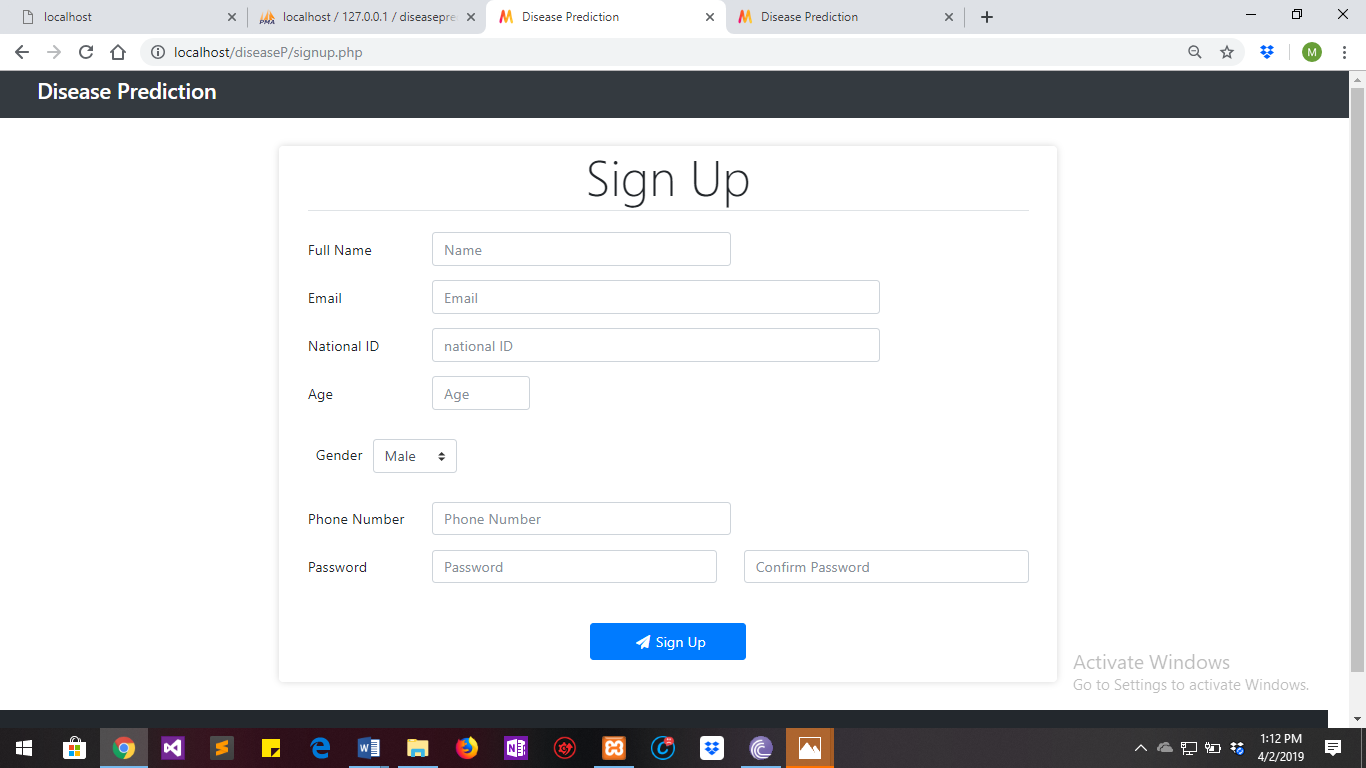
|  |  |  |  |
| --- | --- | --- | --- |
| Name | Type | Size | Description |
| Id | Int | 11 | Primary key of relational table |
| Symptom\_id | Varchar | 255 | Primary key of symptoms table |
| Disease\_id | Varchar | 255 | Primary key of disease table |

## **5.3 User interface design**

### 5.3.1 Signup

This is where the user enters their personal details into the system in order to make it easy to keep record of the user for reference.

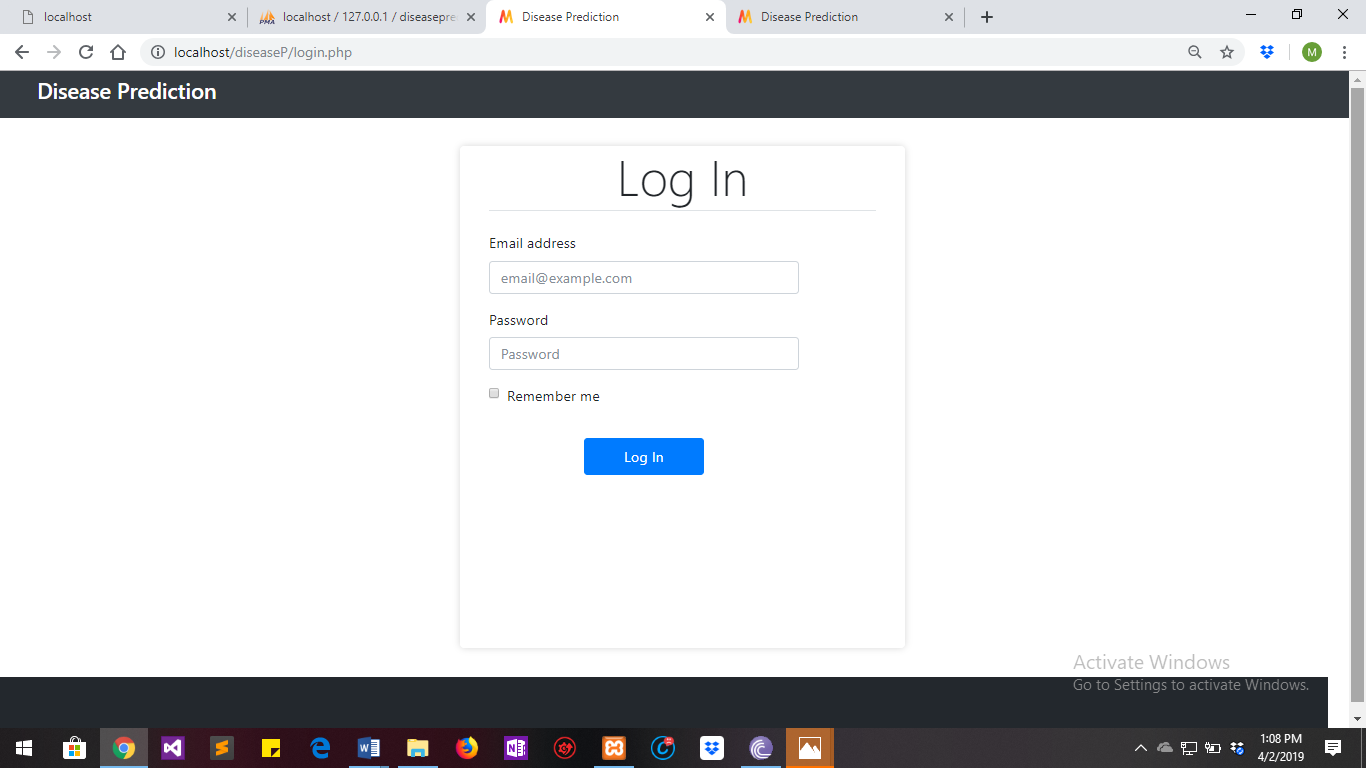
Figure : signup form



### 5.3.2 Login

User logs into the system with the details they provided during signup.

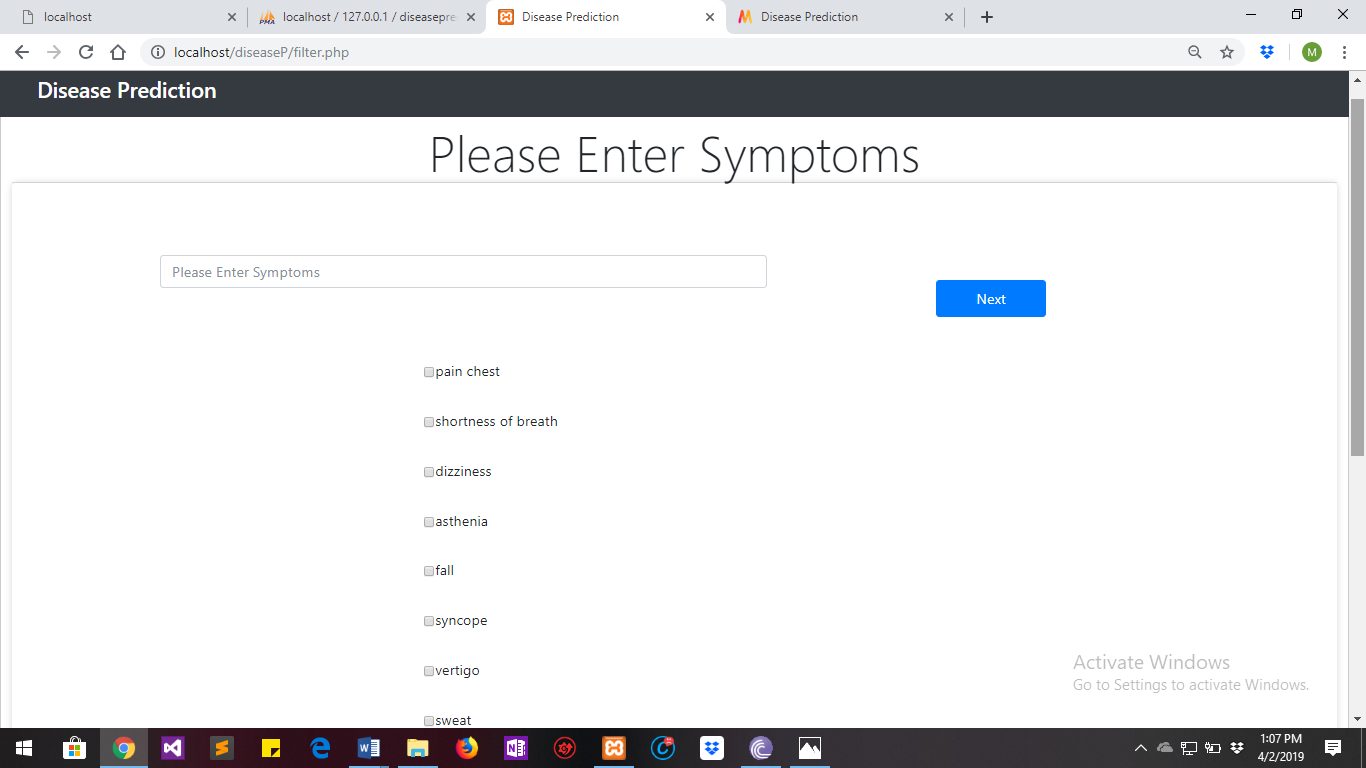
Figure : Login figure



### 5.3.3 Symptom form

User enters their symptoms and submits into the system for analysis of the symptom and to predict the occurrence of a disease based on the symptom provided.

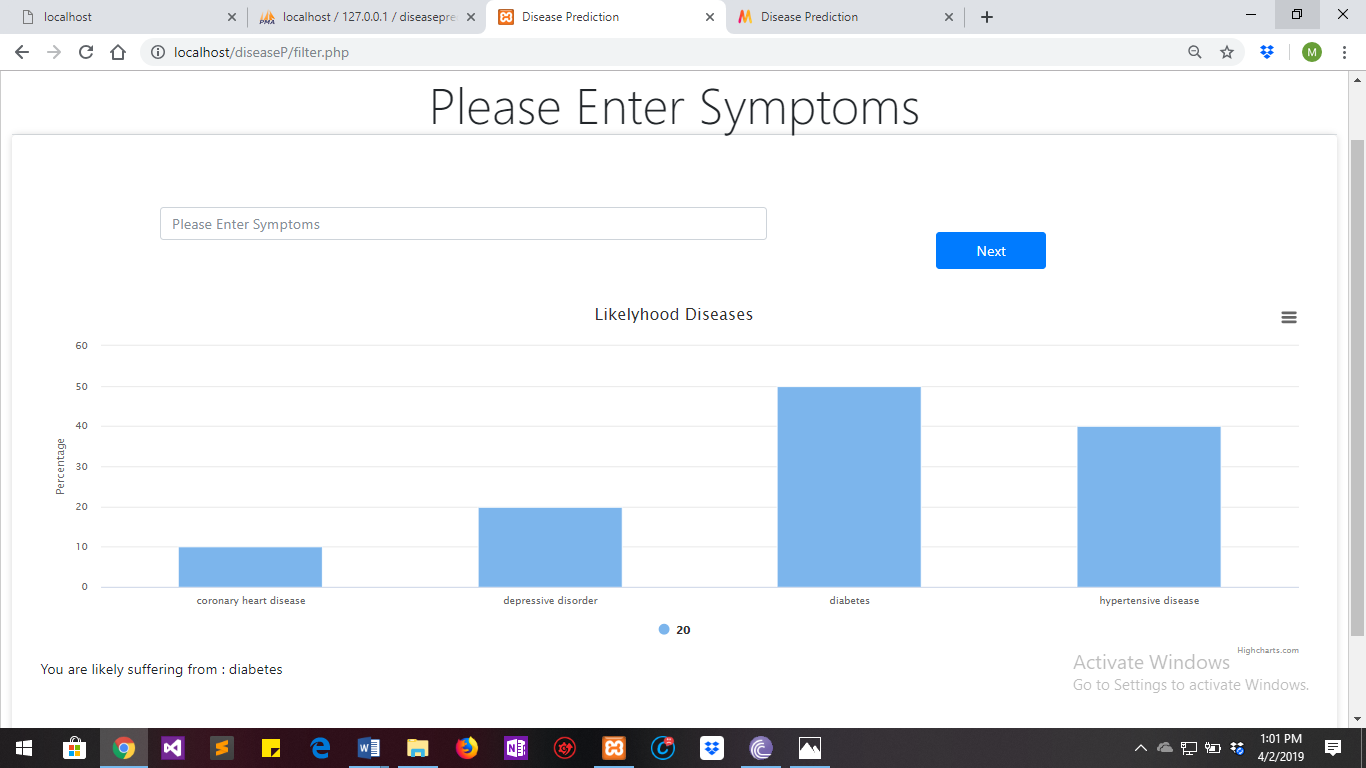
Figure : symptom form



### 5.3.4 Results

The results of the prediction are displayed in a bar graph with the predicted disease and other possible diseases with similar symptoms but with a higher probability.

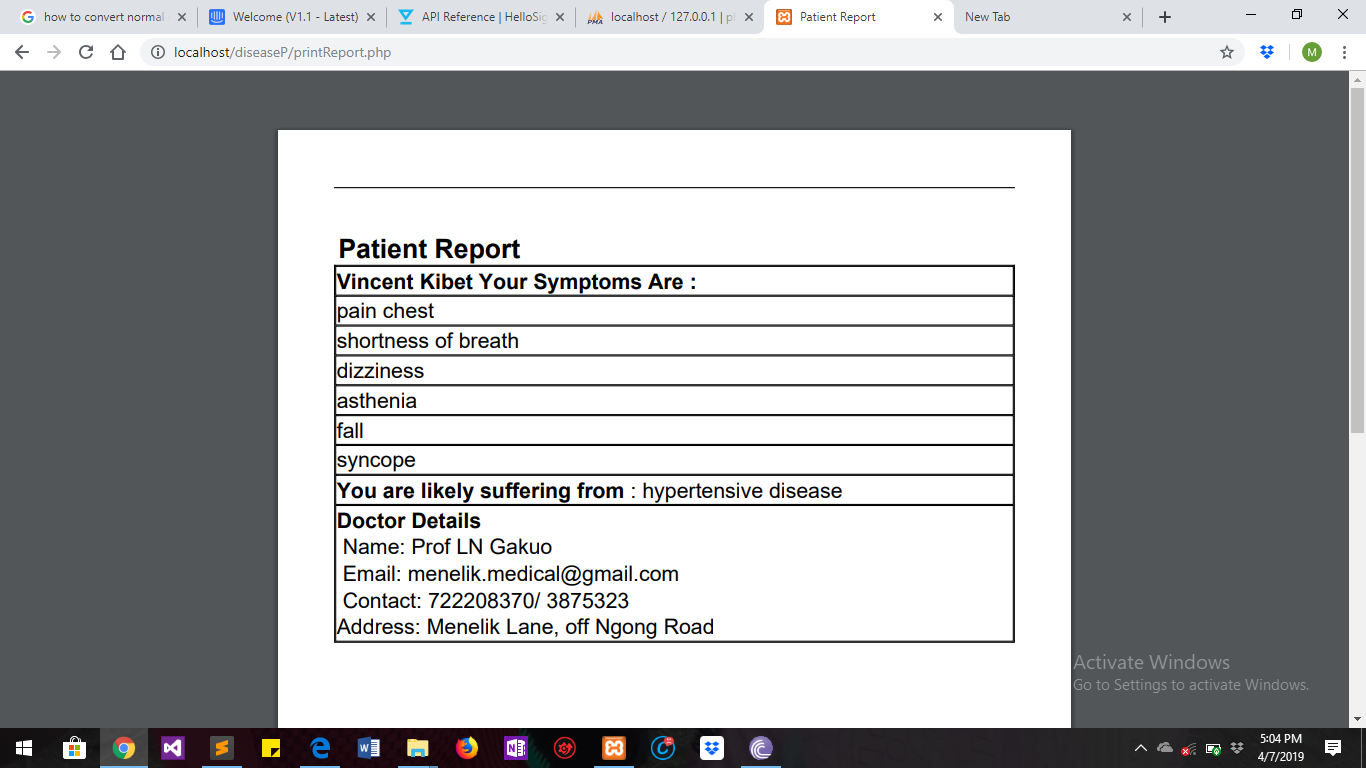
Figure : Prediction Graph



### 5.3.5 Final report

A report is generated with the user’s personal information and prediction details

Figure : Final Report



# Chapter 6 Implementation and testing

## 6.0 Introduction

This is the phase of the system where the theoretical design is turned into a working system hence it is considered to be the most critical stage in achieving a successful system and in giving the user, confidence that the new system will work and be effective. It involves careful planning, investigation of the existing system and it’s constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

## 6.1 Development environment

 HYPERTEXT MARKUP LANGUAGE (HTML):

This is the standard [markup language](http://en.wikipedia.org/wiki/Markup_language) used to create [web pages](http://en.wikipedia.org/wiki/Web_page).

HTML is written in the form of [HTML elements](http://en.wikipedia.org/wiki/HTML_element) consisting of tags enclosed in [angle brackets](http://en.wikipedia.org/wiki/Angle_brackets) (like <html>). HTML tags most commonly come in pairs like <h1> and </h1>, although some tags represent empty elements and so are unpaired, for example <img>. The first tag in a pair is the start tag, and the second tag is the end tag (they are also called opening tags and closing tags). Though not always necessary, it is best practice to append a slash to tags which are not paired with a closing tag.

The purpose of a [web browser](http://en.wikipedia.org/wiki/Web_browser) is to read HTML documents and compose them into visible or audible web pages. The browser does not display the HTML tags, but uses the tags to interpret the content of the page. HTML describes the structure of a website [semantically](http://en.wikipedia.org/wiki/Semantic) along with cues for presentation, making it a [markup language](http://en.wikipedia.org/wiki/Markup_language) rather than a [programming language](http://en.wikipedia.org/wiki/Programming_language).

HTML elements form the building blocks of all [websites](http://en.wikipedia.org/wiki/Website). HTML allows [images and objects](http://en.wikipedia.org/wiki/Img_(HTML_element)) to be embedded and can be used to create [interactive forms](http://en.wikipedia.org/wiki/Fieldset). It provides a means to create [structured documents](http://en.wikipedia.org/wiki/Structured_document) by denoting structural [semantics](http://en.wikipedia.org/wiki/Semantic) for text such as headings, paragraphs, lists, [links](http://en.wikipedia.org/wiki/Hyperlink), quotes and other items. It can embed [scripts](http://en.wikipedia.org/wiki/Scripting_language) written in languages such as [JavaScript](http://en.wikipedia.org/wiki/JavaScript) which affect the behavior of HTML web pages.

**CASCADING STYLE SHEETS** (**CSS**):

It is a [style sheet language](http://en.wikipedia.org/wiki/Style_sheet_language) used for describing the [look and formatting](http://en.wikipedia.org/wiki/Presentation_semantics) of a document written in a [markup language](http://en.wikipedia.org/wiki/Markup_language). While most often used to style [web pages](http://en.wikipedia.org/wiki/Web_page) and [interfaces](http://en.wikipedia.org/wiki/Interface_(computing)) written in [HTML](http://en.wikipedia.org/wiki/HTML) and [XHTML](http://en.wikipedia.org/wiki/XHTML), the language can be applied to any kind of [XML](http://en.wikipedia.org/wiki/XML) document, including [plain XML](http://en.wikipedia.org/wiki/Plain_Old_XML), [SVG](http://en.wikipedia.org/wiki/Scalable_Vector_Graphics) and [XUL](http://en.wikipedia.org/wiki/XUL). CSS is a cornerstone specification of [the web](http://en.wikipedia.org/wiki/The_web) and almost all web pages use CSS style sheets to describe their presentation.

CSS is designed primarily to enable the separation of document content from document presentation, including elements such as the [layout](http://en.wikipedia.org/wiki/Page_layout), [colors](http://en.wikipedia.org/wiki/Color), and [fonts](http://en.wikipedia.org/wiki/Typeface).[[1]](http://en.wikipedia.org/wiki/Cascading_Style_Sheets#cite_note-1) This separation can improve content [accessibility](http://en.wikipedia.org/wiki/Accessibility), provide more flexibility and control in the specification of presentation characteristics, enable multiple pages to share formatting, and reduce complexity and repetition in the structural content.

CSS can also allow the same markup page to be presented in different styles for different rendering methods, such as on-screen, in print, by voice (when read out by a speech-based browser or [screen reader](http://en.wikipedia.org/wiki/Screen_reader)) and on [Braille-based](http://en.wikipedia.org/wiki/Braille_display), tactile devices. It can also be used to allow the web page to display differently depending on the screen size or device on which it is being viewed. While the author of a document typically links that document to a CSS file, readers can use a different style sheet, perhaps one on their own computer, to override the one the author has specified. However if the author or the reader did not link the document to a specific style sheet the default style of the browser will be applied.

**MySQL:**

MySQL is developed, distributed, and supported by Oracle Corporation. MySQL is a database system used on the web it runs on a server. MySQL is ideal for both small and large applications. It is very fast, reliable, and easy to use. It supports standard SQL. MySQL can be compiled on a number of platforms. The data in MySQL is stored in tables. A table is a collection of related data, and it consists of columns and rows. Databases are useful when storing information categorically.

FEATURES OF MySQL**:**

1. Internals and portability:
2. Written in C and C++.
3. Tested with a broad range of different compilers.
4. Works on many different platforms.
5. Tested with Purify (a commercial memory leakage detector) as well as with Val grind, a GPL tool.
6. Uses multi-layered server design with independent modules.
7. SECURITY:
8. A privilege and password system that is very flexible and secure, and that enables host-based verification.
9. Password security by encryption of all password traffic when you connect to a server.
10. CONNECTIVITY:

Clients can connect to MySQL Server using several protocols:

Clients can connect using TCP/IP sockets on any platform.

1. LOCALIZATION:
2. The server can provide error messages to clients in many languages.
3. All data is saved in the chosen character set.

Advantages of MySQL:

1. Leading open source RDBMS
2. Ease of use – No frills
3. Fast
4. Robust
5. Security
6. Multiple OS support
7. Free
8. Technical support large database– up to 50 million rows, file size limit up to 8 Million TB

**PHP (Hypertext Preprocessor)**

PHP is an acronym for "PHP Hypertext Preprocessor". It is widely-used, open source scripting language. The scripts are executed on the server.

CONTENTS OF PHP FILE:

* PHP files can contain text, HTML, CSS, JavaScript, and PHP code
* PHP code are executed on the server, and the result is returned to the browser as plain HTML
* PHP files have extension ".php"

FEATURES OF PHP

1. PHP can generate dynamic page content
2. PHP can create, open, read, write, delete, and close files on the server
3. PHP can collect form data
4. PHP can send and receive cookies
5. PHP can add, delete, modify data in your database
6. PHP can restrict users to access some pages on your website
7. PHP can encrypt data

ADVANTAGES OF PHP

1. PHP runs on various platforms.
2. PHP is compatible with almost all servers used today.
3. PHP supports a wide range of databases
4. PHP is free.

PYTHON

It is a general purpose, dynamic, high level and interpreted programming language. It supports Object Oriented programming approach to develop applications. It is simple and easy to learn and provides lots of high-level data structures.

Features

1) Easy to Learn and Use: Python is easy to learn and use. It is developer-friendly and high level programming language.

2) Expressive Language: Python language is more expressive means that it is more understandable and readable.

3) Interpreted Language: Python is an interpreted language i.e. interpreter executes the code line by line at a time. This makes debugging easy and thus suitable for beginners.

4) Cross-platform Language: Python can run equally on different platforms such as Windows, Linux, UNIX and Macintosh etc. So, we can say that Python is a portable language.

5) Free and Open Source: Python language is freely available at address. The source-code is also available. Therefore it is open source

Advantages

* **Extensive Support Libraries**

It provides large standard libraries that include the areas like string operations, Internet, web service tools, operating system interfaces and protocols. Most of the highly used programming tasks are already scripted into it that limits the length of the codes to be written in Python.

1. **Integration Feature**

Python integrates the Enterprise Application Integration that makes it easy to develop Web services by invoking COM or COBRA components. It has powerful control capabilities as it calls directly through C, C++ or Java via python. Python also processes XML and other markup languages as it can run on all modern operating systems through same byte code.

1. **Improved Programmer’s Productivity**

The language has extensive support libraries and clean object-oriented designs that increase two to tenfold of programmer’s productivity while using the languages like Java, VB, Perl, C, C++ and C#.

1. **Productivity**

With its strong process integration features, unit testing framework and enhanced control capabilities contribute towards the increased speed for most applications and productivity of applications. It is a great option for building scalable multi-protocol network applications.

## 6.2 System components

User Registration

This is done by all the system users when it is their first time to use the system. The user enters their personal details in a registration form which is then stored in the database. Information users provide include: Name, age, gender, email, phone number etc.

User Login

After registration user can login to the system using the information they provided during registration.

Symptom Analysis

When a user logs into the system, they are provided with a form in which they are to enter the symptoms they are suffering from in order for the system to predict their disease based on the symptoms provided.

Feedback

After diagnosis is done the use is provided with feedback in form of a report indicating the disease they are suffering from.

Comment

User can give their views about their experience while using the system.

### 6.3 Test Plan (test data, test cases and test results)

The purpose of testing is to discover errors. Testing is the process of trying to discover every weakness in a project. It provides a way to check the functionality of components and the finished product. It is the process of exercising software with the intention of ensuring that the software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

**TYPES OF TESTING**:

**Unit testing:**

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs.

Table : Unit testing

|  |  |  |
| --- | --- | --- |
| **Test Data** | **Test Case** | **Test Result** |
| Module Testing | Display symptoms | The system should be able to display related symptoms to the one the user has entered. |
| Display disease | The system should be able to display predicted disease. |
| Display doctor | The system should be able to display a doctor referred to a patient. |

**Integration testing:**

Integration tests are designed to test integrated software components to determine if they actually run as one program. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Table : Integration table

|  |  |  |
| --- | --- | --- |
| **Test Data** | **Test Cases** | **Test Result** |
| Integration test | User Authentication | The system should allow valid users to log into the system. |
| Relationship among modules. | The established relationship should work as expected by its users. |

**Acceptance Testing**

Testing the system to determine user acceptance and the readiness of the system.

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Acceptance table

|  |  |  |
| --- | --- | --- |
| **Test Data** | **Test Cases** | **Test Result** |
| Acceptance testing | Interaction of users with the system | The user should be able to use the system with ease. |

# 

# Chapter 7: Conclusion

## 7.1 Achievements and lessons learnt

The major achievement of this project is the ability to filter symptoms of many diseases and provide those related to what the user of the system has searched, the system also provides other symptoms which the user of the system may be suffering from, these will help in the prediction process since the user is provided with all information necessary.

The understanding of the database design has greatly improved because in order to generate in order to generate the final report the database design has to be greatly followed.

## 7.2 Conclusions

This main objective of this system is to predict the disease on the basis of the symptoms provided by user. The system is designed in such a way that it takes symptoms from the user as input and produces output of the predicted disease. Disease Prediction system was successfully implemented using Naïve Bayes.

## 7.3 Recommendations

This system has not implemented prescription of medication to the user. Prescription implementation can be an advantage to this system since user will be getting all the services they need at the comfort of their homes/work places

# Reference

Noh, K., Lee, H. G., Shon, H. S., Lee, B. J., & Ryu, K. H. (2006). Associative classification approach for diagnosing cardiovascular disease. In *Intelligent computing in signal processing and pattern recognition* (pp. 721-727). Springer, Berlin, Heidelberg.

Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, *47*(10), 44-48.

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), 43-48.

Arora, R., & Sondhi, S. MobDoc: Mobile-based Medical Diagnostic Tool.

Thenmozhi, K., & Deepika, P. (2014). Heart disease prediction using classification with different decision tree techniques. *International Journal of Engineering Research and General Science*, *2*(6), 6-11.

Manikantan, V., & Latha, S. (2013). Predicting the analysis of heart disease symptoms using medicinal data mining methods. *International Journal of Advanced Computer Theory and Engineering*, *2*, 46-51.

# Appendix

## Appendix IUser manual

Step one:

Signup into the system in order to access more services. Fill in the form with your personal information.

Step two:

Login into the system with the information you provided during signup.

Step three:

Enter any symptom you are experiencing in order for the system to predict other possible symptoms which will help you in filling up the symptoms for prediction purposes.

If you don’t find a symptom you are experiencing you click “next button” for more symptoms and if you get what you are experiencing click submit.

Step four:

Print the report of the details the system has predicted for you.

Step five:

You can logout after you have finished the process.

## Appendix IIData collections tools

Questionnaire

a. Do patients pay the facility for medicines?

■ Yes, full payment ■ Yes, partial payment –> Proportion paid by patient: …….% ■ No, medicines are provided for free

b. If medicines are provided for free or for partial payment, who subsidizes it?

■ Central government ■ Local government ■ Private insurance ■ Social assistance plans ■ other (specify: .................) ■ don’t know

c. Do patients pay the facility for consultations?

■ Yes, full payment ■ Yes, partial payment –> Proportion paid by patient: ■ No, consultations are provided for free

d. If consultations are provided for free or for partial payment, who subsidizes it?

■ Central government ■ Local government ■ Private insurance ■ Social assistance plans ■ other (specify: .................) ■ don’t know

e. Do patients pay the facility for diagnostic tests?

■ Yes, full payment ■ Yes, partial payment –> Proportion paid by patient ■ No, diagnostic tests are provided for free

f. If diagnostic tests are provided for free or for partial payment, who subsidizes it? ■ Central government ■ Local government ■ Private insurance ■ Social assistance plans ■ other (specify: .................) ■ don’t know

## **Appendix III** Project schedule

Table : Project schedule

GANT CHART

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **DURATION** | **January**  **2019** | | | | **February**  **2019** | | | | **March**  **2019** | | | | **April**  **2019** |
| **ACTIVITY** | **WEEKS** | | | | WEEKS | | | | WEEKS | | | | WEEKS |
| Data Collection |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data Preparation |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Handle Data |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data Transformation |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Making Predictions |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Create User Interface |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Testing |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Maintenance |  |  |  |  |  |  |  |  |  |  |  |  |  |

## 

## Appendix IV Project budget

|  |  |  |
| --- | --- | --- |
| **Resource** | **Location** | **Amount(Ksh)** |
| Research resources | data bundles | 300.00 |
| Printing Project documentation and binding | MMU Printing Press | 600.00 |
| CD-ROM |  | 50.00 |
| Flash disk |  | 800.00 |

## Appendix V Any other relevant document.

# Import Dependencies and other pthon libraries

import csv

import pandas as pd

import numpy as np

from collections import defaultdict

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib inline

