CHAPTER 1: INTRODUCTION

1.1 INTRODUCTION TO THE PROBLEM DOMAIN

The donation of blood is important because most often people requiring blood do not receive it on time causing loss of life. Examples include severe accidents, patients suffering from dengue or malaria, or organ transplants.

Extreme health conditions such as Leukemia and bone marrow cancer, where affected individuals experience sudden high blood loss and need an urgent supply of blood and do not have it can also lead to loss of life.

Sound data-driven systems for tracking and predicting donations and supply needs can improve the entire supply chain, making sure that more patients get the blood transfusions they need, which can reduce mortality risk.

One of the interesting aspects about blood is that it is not a typical commodity. First, there is the perishable nature of blood. Grocery stores face the dilemma of perishable products such as milk, which can be challenging to predict accurately so as to not lose sales due to an expired product.

Blood has a shelf life of approximately 42 days according to the American Red Cross. However, what makes this problem more challenging than milk is the stochastic behavior of blood supply to the system as compared to the more deterministic nature of milk supply. Whole blood is often split into platelets, red blood cells, and plasma, each having their own storage requirements and shelf life.

For example, platelets must be stored around 22 degrees Celsius, while red blood cells 4 degree Celsius, and plasma at -25 degrees Celsius. Moreover, platelets can often be stored for at most 5 days, red blood cells up to 42 days, and plasma up to a one calendar year.

The requirement for blood is steadily increasing as a result of increasing population and also, advancement in clinical medicine. On the contrary, the number of voluntary donors is decreasing over the last few years. Recent findings bring to light the fact that a majority of the new donors fail to donate a second time.

It is not economically viable to continually check for donors who fail to return for a subsequent donation. Retention of donors who pass the initial health checks is vital to the strategy of increasing blood donation.

However, blood donation services can alter their strategies to rope in donors only if they can understand the factors that influence their donation behavior. Finding and understanding these factors would help the blood donation services to come up with newer and more efficient strategies that would increase the return rate of donors.

An analysis conducted in shows that the factors influencing the decision to donate blood are complex. This creates an impending need for an intelligent automation mechanism that can, with minimalistic human intervention, monitor blood stock, select potential blood donors and gather them based on blood requirement.

The incorporation of information and computer technology can help in decreasing the workload of blood centers and in the analysis of the contributing factors. In this paper, we review and contrast various existing implementations and previous research done on this aspect. Further, we identify the root components of the proposed intelligent system and offer a study on the potential methods of its implementation.

Problems in blood donation are mainly about insufficiency to collect blood to meet uncertain demands. In general, there is no plan to manage potential blood donors. With a good preparation, by classifying potential donors in such a way that intentions of donors to donate blood in the future can be determined, would greatly facilitate the blood acquisition to perform better.

Response to blood requisition can be conducted efficiently and effectively, leading to lifesaving of patients and reduction of relevant operating costs. Moreover, it will be beneficial for managing blood requisition on emergency requests.

Hence, information from blood donor classification would suggest the behavior of blood donors whether or not they will donate again in the future. Blood collection from voluntary donors is a primary activity to acquire blood as raw material into blood supply. chain.

Generally, Blood Centers have a standard procedure to screen blood effectively by incorporating collection costs, and amounts of time and units, to obtain blood properly into consideration.

The goal is to obtain blood that is safe for use in the subsequent activities. However, the main problem in blood collection is an inability to obtain sufficient blood to meet the patients’ needs or a difficulty to balance blood demand and supply in the blood supply chain.

Blood is always in great demand from the past, but nowadays as the population grows the ratio of road accidents, disease and medical surgeries are also growing in same amount. The blood and there donors are very much important as it cannot be manufactured and only come from generous donors.

This gives us the importance of looking for blood and providing conditions for donating blood . Blood donation and its transfusion service is an indispensable part of medicine and health-care sectors.

The volume of blood collected could be increased in two ways: by encouraging new donors to start donating, or by encouraging existing donors to donate more often, or both.

Blood donation is a series of interdependent operations in blood transfusion organization that consists of donor registration, donors' health evaluation, donors' availability information, blood collection, blood screening, supply when needed, inventory management, and proper dissemination of blood.

The challenge for blood collection services is to devise strategies that encourage non-donors to make their first donation, to devise further strategies to reduce donor dropout, and to motivate behavior change that will lead to committed regular donation.

Nowadays information and computer technology has gaining more importance in medicine and healthcare sectors, as it is needed to make it efficient by using computer Technology (CT) to database systems.

Despite the fact that the volume of collected blood has fallen, there has been an increase in the demand for whole blood and blood products as a result of greater use of blood products to treat medical conditions such as cancer, and because new uses have been found for blood products.

The volume of blood collected could be increased in two ways: by encouraging new donors to start donating, or by encouraging existing donors to donate more often, or both. The challenge for blood collection services is to devise strategies that encourage non-donors to make their first donation, to devise further strategies to reduce donor dropout, and to motivate behaviour change that will lead to committed regular donation.

Establishing a reliable method of predicting who is most likely to donate blood would improve the likelihood of such strategies succeeding. Nevertheless, since providing a dependable supply of blood is a primary mission for most blood centres, it is logical to target donors who are most likely to donate blood.

1.2 STATEMENT OF THE PROBLEM

To predict the blood donors who are willing to donate blood, by studying their knowledge, intention and attitude towards blood donation using machine learning techniques.

1.3 OBJECTIVES

The overall aim of this research is to examine the predictive ability of two approaches to predict the willingness to donate blood. More specifically the purpose of this study is to consider whether as a general principle when attempting to predict behaviour, attitude questions should be replaced with questions that have verifiable answers about respondent's environment, knowledge and past behaviour. The broad aim of this study was further divided into several more specific research objectives outlined below:

- To measure the level of knowledge regarding blood donation and its importance.
- To find out positive and negative attitudes towards blood donation
- Provide guidance on the measures needed to develop and implement effective systems for assessing the suitability of individuals to donate blood.
- Review the available evidence base and provide recommendations on criteria for blood donor selection.

1.4 APPLICATIONS

- It can be used by the hospitals and blood bank in emergency, who require blood and want to contact a person who is ready to donate blood.

- In a short span of time, the person who is willing to donate blood is identified and can be contacted.

- In case of emergency, where every second is valuable, one cannot contact all the people who are eligible to donate blood, so we help them by predicting the donors who are willing to donate blood and thus saving life of person in danger.

- Donating blood prevents Hemochromatosis: Health benefits of blood donation include reduced risk of hemochromatosis. Hemochromatosis is a health condition that arises due to excess absorption of iron by the body. This may be inherited or may be caused due to alcoholism, anaemia or other disorders. Regular blood donation may help in reducing iron overload. Make sure that the donor meets the standard blood donation eligibility criteria.

- Blood donation helps in lowering the risk of cancer. By donating blood the iron stores in the body are maintained at healthy levels. A reduction in the iron level in the body is linked with low cancer risk.

- Blood donation is beneficial in reducing the risk of heart and liver ailments caused by the iron overload in the body. Intake of iron-rich diet may increase the iron levels in the body, and since only limited proportions can be absorbed, excess iron gets stored in heart, liver, and pancreas.

- Regular blood donation reduces the weight of the donors. This is helpful for those who are obese and are at a higher risk of cardiovascular diseases and other health disorders. However, blood donation should not be very frequent and you may consult your doctor before donating blood to avoid any health issues.

- After donating blood, the body works to replenish the blood loss. This stimulates the production of new blood cells and, in turn, helps in maintaining good health.

1.5 EXISTING SOLUTION METHODS

Authors	Methods	Data	Drivers
(Godin, Conner et al. 2007)	Logistic Regression	Survey (2070 experience donors, 161 new donors)	Experienced donors:intention, perceived control, anticipated regret, moral norm, age, and past donation frequency, New donors: intention and age
(Sojka and Sojka 2008)	Descriptive statistics	Survey (531 participants)	General motivators: friend influence (47.2%), media requests (23.5%). Continued donations: altruism (40.3%), social Responsibility (19.7%), friend influence (17.9%)
(Masser, White et al. 2009)	Structural equation modeling	Survey 1 (263 participants): Follow-up survey (182 donors)	Moral norm, donation anxiety and donor identity indirectly predicted intention through attitude.

(Masser, Bednall et al. 2012)	Path analysis	Survey 1 (256 participants)	Their extended TPB model showed intention was predicted by attitudes, perceived control and self identity.

Table 1: Social and psychological studies investigating drivers of blood donations

Authors	Methods	Data	Results
(Mostafa 2009)	ANN (MLP), ANN (PNN), LDA	Survey (430 records, 8 features)	ANN (MLP): Test accuracy (98%) ANN (PNN): Test accuracy (100%) LDA: Test accuracy (83.3%)
(Santhanam and Sundaram 2010) (Sundaram 2011)	CART CART vs. DB2K7	UCI ML blood transfusion data (748 donors, 5 features)	Precision/ PPV (99%), Recall/ Sensitivity (94%)
(Darwiche, Feuilloy et al 2010)	PCA for feature reduction ANN (MLP) vs SVM (RBF)	UCI ML blood transfusion data (748 donors, 5 features)	SVM (RBF) using PCA: Test sensitivity (65.8%): Test Specificity (78.2%): AUC (77.5 %).

(Ramachandran, Girija et al. 2011)	J48 algorithm in Weka (aka C4.5)	Indian Red Cross Society (IRCS) Blood Bank Hospital (2387 records, 5 features)	Recall/Sensitivity (95.2%), Precision/ PPV (58.9%), Specificity (4.3%)
(Lee and Cheng 2011)	k-Means clustering, J48, Naïve Bayes, Naïve Bayes Tree, Bagged ensembles of (CART, NB, NBT)	Blood transfusion service centre data set(748 records/ donors, 5 features)	Bagged (50 times) Naïve Bayes: Accuracy (77.1 %), Sensitivity (59.5 %), Specificity (78.1%), AUC (72.2%)
(Zabihi, Ramezan et al. 2011)	Fuzzy sequential pattern mining	Blood transfusion service centre data set (748 records/ donors, 5 features)	Precision/PPV (Frequency feature 88%, Recency feature 72%, Time feature 94%)
(Sharma and Gupta 2012)	J48 algorithm in Weka (aka C4.5)	Blood bank of Kota, Rajasthan, India (3010 records, 7 features)	Accuracy (89.9%)
(Boonyanusith and Jittamai 2012)	Artificial Neural Network (ANN), J48 algorithm (aka C4.5)	Survey (400 records, 5 features)	ANN: Accuracy (76.3%); Recall/Sensitivity (81.7%); Specificity (53.8%) J48: Recall/ Sensitivity (81.2%);
(Testik, Ozkaya et	Two-Step	Blood donation	-

al. 2012)	Clustering with CART. This is fed into a serial queuing network model	center (1095 donors, 3 clusters)	
(Ashoori, Alizade et al. 2015)	C5.0, CART, CHAID, QUEST	Blood transfusion center in Birjand City in North East Iran (9231 donors, 6 features)	Model accuracy (train/test): C5.0 (57.49/ 56.4%), CART(55.9/56.4%). CHAID (55.56/55.61%), QUEST (55.34/ 56.11%)
(Ashoori, Mohammadi et al. 2017)	Two-step clustering, C5.0, CART, CHAID, QUEST	Census survey from a blood transfusion centers from Birjand, Khordad, & Shahrivar (1392 participants)	Important features: Blood pressure level, blood donation status, Temperature Model accuracy: C5.0 (99.98%), CART (99.60%), CHAID(99.30%), QUEST (89.13%)

Table 2: Predicting blood donation with a focus on machine learning techniques

We have chosen this system because

- We are predicting donors' who are willing to donate the blood, from the list of eligible donors.
- Using this, we can save a lot of time by directly contacting the donors who are readily available to donate the blood and thus helping in to save a life.

1.6 PROPOSED SOLUTION METHODS

i. Online Questionnaire:

At the first step of this work, questionnaire is be used to collect data about feelings or opinions of individuals who become blood donors. Online questionnaire is used to conduct data survey from different people i.e. people of SJCE, friends, family members, acquaintances etc.

The questionnaire consists of two parts: the first part is about personal information including sex, age, level of education, weight and height, contact address, phone numbers, and the other part consists of 15 questions about their feelings or opinions why they decide to become blood donors. The measurement format in this questionnaire will follow the Likert- type scale, which is commonly used with queries individual attitude or feelings. This format is the opinion of respondents with 1 to 5 (strongly disagree, disagree, not sure, agree, and strongly agree). It also includes questions on both positive and negative combinations.

The factors used in this study aiming to analyse the intensity of the feelings or opinions of the individuals comprise of the following five factors:

Altruistic values (ALT): These are the questions asking for attitudes about reduction in blood donation, benefit to the society**,** social ethics and social belief.

Knowledge in blood donation (KNL): These questions are related to the knowledge asking for individual's age and body mass index (BMI) in blood donation along with the understanding screening and recruiting first-time blood donors.

Perceived risks (RSK): These are related to an individual's perception of standard procedure and safety control in blood donation along with the risk issues in donating blood.

Attitudes in blood donation (ATT): In this category, questions relating to attitude in the satisfaction and the aspect of blood donation.

Intention to donate blood (INT): The questions about blood donation in future will be asked.

The preliminary statistics are analysed by calculating the median, the standard deviation and reliability testing (cronbach's alpha) for each factor. Finally, reliability of overall questionnaire is (α=0.65), thus it is in acceptable level.

ii. Classification using Naive Bayes:

To classify blood donor groups, it needs to prepare input data to the learning process by adjusted data obtained from questionnaires and made them as input attributes, according to the technique of Naive Bayes, in order to classify output as blood donor groups. A sample of input is shown in the following image.

	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	X
1	How obviou	For me to, do	Most peopl	Generally s	How often d	If I felt ill,	For me to	How ofter	How often	If I encou	I would a	I'm sca	My pe	If you cor	Which med	Class
2	4	5	4	4	2	2	5	4	3	4	2	1	4	3	Whatsapp,	YES
3	4	5	5	1	1	3	5	3	5	5	3	3	5	5	Whatsapp,	YES
4	5	3	3	5	1	1	5	3	5	5	5	1	5	5	Whatsapp	YES
5	1	4	5	2	1	1	5	3	2	2	2	2	5	4	Whatsapp	NO
6	1	5	3	1	3	5	5	4	2	1	4	1	5	5	Whatsapp,	NO
7	4	5	4	4	1	5	5	3	2	3	2	1	4	5	Whatsapp,	YES
8	4	5	1	3	2	5	5	1	3	2	3	1	5	5	Whatsapp,	YES
9	3	5	2	3	2	5	5	3	3	4	1	3	4	5	Whatsapp,	YES
10	4	5	4	3	1	5	5	3	3	4	3	2	2	4	Whatsapp,	YES
11	3	5	3	5	2	3	5	4	4	3	1	1	5	4	Whatsapp,	YES
12	3	5	4	2	1	5	5	4	4	4	4	2	4	5	Whatsapp	YES
13	3	5	2	2	3	4	5	4	3	4	1	4	5	4	Whatsapp,	YES
14	3	5	3	5	1	1	5	4	3	3	1	1	5	5	Whatsapp	YES
15	4	5	3	3	1	2	4	3	3	3	3	2	5	4	Whatsapp,	YES
16	3	5	5	5	4	5	5	3	4	3	5	1	5	5	Whatsapp,	YES
17	3	5	4	3	3	4	4	4	3	3	2	2	4	4	Mail	YES
18	5	5	5	5	1	5	5	4	3	4	5	1	5	5	Whatsapp,	YES
19	3	3	4	4	3	4	4	5	3	3	2	1	5	3	Whatsapp,	NO
20	3	5	5	2	2	4	5	3	3	2	2	1	5	5	Mail, Phone	YES
21	5	5	5	5	3	3	4	3	4	2	5	3	5	4	Whatsapp,	YES
22	4	5	1	1	3	5	5	4	5	5	1	1	5	5	Whatsapp,	YES
23	2	5	3	1	2	3	5	2	3	3	2	3	4	4	Whatsapp,	YES
24	3	5	4	2	1	3	5	4	3	3	2	3	5	4	Whatsapp	YES
25	5	5	1	1	3	4	5	4	5	5	1	1	5	5	Whatsapp,	YES
26	1	1	3	1	2	2	2	1	1	1	1	1	1	5	Phone Call	NO
27	1	3	4	1	1	3	5	3	3	4	1	1	5	5	Whatsapp,	NO
28	4	4	2	4	1	5	5	2	3	3	2	3	5	4	Whatsapp,	YES
29	4	4	4	4	2	4	5	5	2	1	1	1	5	5	Whatsapp,	YES
30	5	5	3	3	4	3	5	4	3	3	3	1	5	3	Whatsapp	YES
31	5	5	4	5	4	3	5	4	4	4	4	3	4	4	Whatsapp,	YES
32	4	5	2	2	2	3	2	1	2	2	2	1	2	4	Whatsapp	YES

Fig 1: Classification using Naïve Bayes

Overall data includes 246 records using the five factors in learning process by Naive Bayes technique. This study focuses only on information of feelings or opinions that influence on blood donation behaviours only. However, basic personal information can be used for tracking blood donors for future blood donation. It is also used to develop the donor database system for better donor management.

The data collected from donors can be used to develop the blood donor classification model appropriately. Although the sample size chosen in this study is only in the provincial level, the result shows that the Naive Bayes model has a relatively high accuracy value. Following figure outlines the flow of our study.

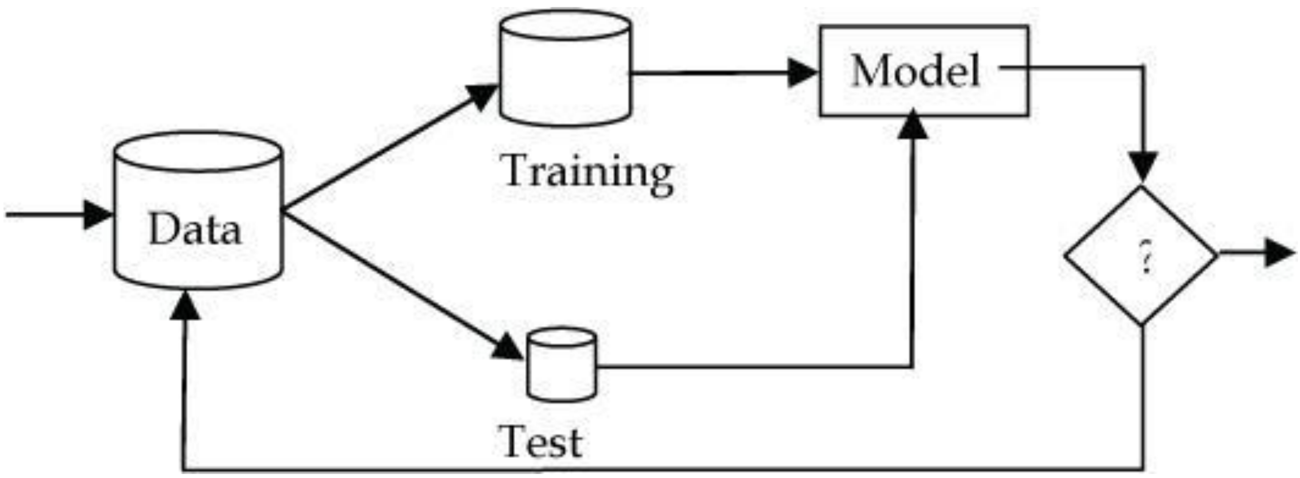


Fig 2: Work-flow model of proposed methods

1.7 GANTT CHART

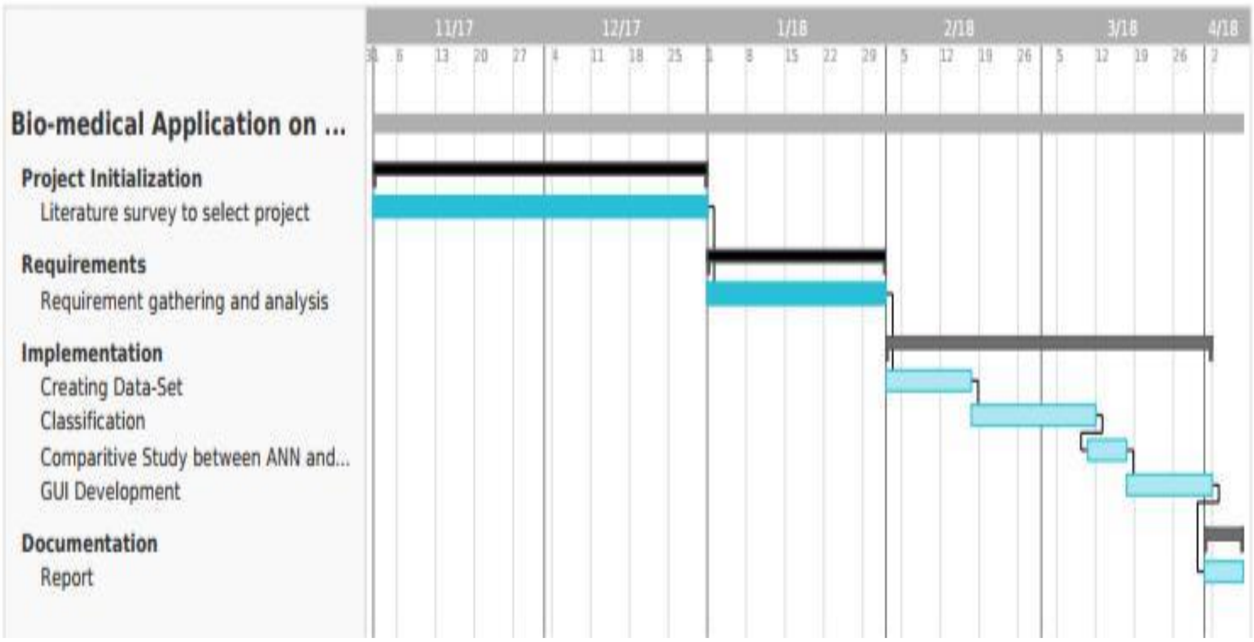


Fig 3: Schedule of the project

CHAPTER 2: LITERATURE SURVEY

2.1 Predicting blood donations using machine learning techniques

Authors: Deepti Bahel, Prerana Ghosh, Arundhyoti Sarkar, Matthew A. Lanham [2].

Among the algorithms examined, the non-clustered 5-fold cross -validated logistic regression model performed the best based on the test set AUC. However, AUC alone may not be best performance measure with respect to likelihood to predict blood. That is because AUC considers the area determined by True Positive Rate (TPR)/sensitivity and False Positive Rate (FPR)/(1- Specificity). This model could be used for targeted advertisement. In such a case, they are more interested in the TPR which would be to target the actual donors who would be interested in donating blood regularly. Hence, their performance would focus more on sensitivity leading them to recommend a clustered LDA model. Some authors used algorithms that are specific to a certain software package. For example, CHAID and QUEST are decision tree algorithms specific to SPSS Modeler. Two-Step clustering is also specific to SPSS. They compared these algorithms to have a more complete comparison to what other authors have done. Secondly, there are many other machine learning algorithms that could be tested where one (or a combination of several) might yield to significantly better results. Thirdly, building models on a smaller number of clusters might improve performance because there will be more observations in each cluster.

2.2 A study on automation of blood donor classification and notification techniques

Authors: Arunkumar Chinnnaswamy, Amrita Vishwa Vidyapeetham [1].

They presented a survey of papers relating to automation in the area of transfusion medicine for aiding and enhancing donation and distribution. Machine Learning techniques are the most viable candidates for implementing ranking and selection of donors, given their efficiency, accuracy and ease of implementation. The development of a stable, error free communication mechanism is very vital in the system proposed. SMS is the favoured option in this case, given the widespread use of cellular phones and hence ease of access. The basic components of the system proposed in this paper are thus, an intelligent Machine Learning algorithm with a fair accuracy and a sophisticated notification system. By the integration of the above components, we can achieve a machine that would aptly satisfy the needs of blood donation services.

2.3 Blood Donor Classification Using Neural Network and Decision Tree Techniques

Authors: Wijai Boonyanusith and Phongchai Jittamai [16].

Five factors influencing blood donor behaviors suggested by Mostafa [2] are used as a framework to construct a model to classify donors. Those five factors include altruistic values, knowledge in blood donation, perceived risks, attitudes towards blood donation, and intention to donate blood. Online questionnaires were used in the survey among students in four universities in Nakhon Ratchasima province. A sample size of 400 responses has yielded a reliability scale in an acceptable significant level (greater than 0.65 of Cronbach's Alpha). The accuracy test of donor group classification was done using the ANN technique with Multi-layer Perceptron function and back propagation algorithm in order to predict the answer from a series of donors. As a result, it has been found that the model is able to classify donors into the blood donor group with the precision and recall values of 81.7% and 88%, respectively, and classify donors into the non donor group with the precision and recall values of 53.8% and 41.6% respectively. The accuracy of this model to classify blood donor group is 76.25%. Decision Tree technique is also used to compare the model with the ANN in order to analyze the accuracy of the classification of the model. The result has shown that the ANN model has higher accuracy value to classify blood donor groups than that of the Decision Tree model. It can be concluded that the ANN technique is a reliable method to learn data series of the five factors influencing the individual behavior of blood donation used for blood donor classification with high level of accuracy value. This research is a pilot study to incorporate the blood donor classification model to manage blood donors in Thailand. Machine learning techniques are used to classify a group of donors by testing with a data from potential donors in the northeastern province of the country. According to the preliminary statistical results of the factors affecting blood donor behavior, it can be claimed that reliability of the questionnaire and the sample is in an acceptable significant level.

2.4 Predicting Willingness to Donate Blood

Authors: Judith Holdershaw, Philip Gendall & Malcolm Wright [10].

Attitude-based models such as the theory of planned behaviour, which underpin much of the traditional thinking in social marketing, rely on the assumption that behavioural intentions are highly correlated with actual behaviour. The blood donation study reported here merely confirms what has been observed many times before, that attitudes are not good predictors of behaviour. Labaw's approach to predicting behaviour eschews the use of attitudes, proposing instead variables that measure the environment, knowledge and behaviour of the people concerned. In our blood donation study, Labaw's approach was better at predicting reported behaviour than the theory of planned behaviour. In absolute terms the Labaw approach was not particularly successful, nevertheless, its results were promising enough to suggest that this approach merits further study. The significance of the study reported lies in its operationalization of Labaw's framework and its implicit assumption that there is a better way of predicting behaviour than traditional cognitive approaches. This suggestion has important managerial implications. Instead of devoting effort to measures designed to change potential donors' attitudes to blood donation (on the assumption that this will lead to more donors or donations), emphasis should instead be placed on aspects of potential donors' knowledge, environment and behaviour that are known to be related to donor behaviour. For example, if the fact that having a family member who has been a donor is a significant predictor of blood donation, donors could be explicitly encouraged to "recruit" family members as new donors.

2.5 Predicting a donor's likelihood of donating blood

Authors: W. A. Flegel, W. Besenfelder[2] and F. F. Wagner [15].

Donors give their blood without remuneration, and it is the finest responsibility of the physicians in transfusion medicine to put this gift of life to its optimal use in the interest of the donors and for the benefit of the patients. Driven by this motivation, the authors developed a procedure to predict a donor's likelihood of donating within a preselected time interval. They showed that the application of their technical procedure could increase the available quarantine plasma by about 30000 units per year without incurring additional costs or reduce the quarantine storage expenses by about 20% without limiting the current supply. These figures, if widely applicable, may result in a welcome relief to the stringent cost-containment efforts of the health care systems in Germany and elsewhere. They thought that the whole demand of the German health care system for single donor plasma units could be met by quarantine plasma and devised means to improve their cost efficiency. Previously, donor demographic factors such as age, age at first donation, sex, location of the blood drive and the donation history were reported as being relevant for predicting subsequent donation behaviour. A surprising finding, however, was the consistency of return across age groups, and therefore the limited impact of what has until now been considered a key predictive variable. As first time donors were rightfully perceived less likely to return than the average repeat donor (Burnett & Leigh, 1986), their units were often not stored for quarantine purposes. The analysis showed that this decision could improve the overall success rate by about 2´5%. Some repeat donors who are often found among those with small numbers of previous donations were even less likely to return than first time donors.

CHAPTER 3: SYSTEM REQUIREMENTS AND ANALYSIS

Software requirement specification

Software Requirement Specification (SRS) basically explains about, a customer or potential clients system requirements and dependencies at a particular point in time prior to any actual design or development work. It involves specific requirements, software and hardware requirements, interfaces.

The system requirements have been divided into modules like :

- Input Requirements
- Output Requirements
- Specific Requirements
- Functional Requirements
- Non-Functional Requirements
- Hardware Requirements
- Software Requirements

3.1 INPUT REQUIREMENTS

Dataset: Set of donor data i.e. around 250 samples consisting of personal information about the donor and his opinion to the questions are collected using Google forms.

Input: The user has to give city and blood group as the input, where he wants to find the donor.

3.2 OUTPUT REQUIREMENTS

Once the input is given, the donors matching the city and blood group are listed along with personal details.

3.3 SPECIFIC REQUIREMENTS

- **Google form :** It is a free service for data collection provided by Google. Users with a google account can create forms through an interface that can be shared via email, social media, or filled out in person.

- **csv :** The so-called CSV (Comma Separated Values) format is the most common import and export format for spreadsheets and databases.

- **MySQLdb :** MySQLdb is an interface for connecting to a MySQL database server from Python.

- **random :** This module implements pseudo-random number generators for various distributions.

- **math :** It provides access to the mathematical functions defined by the C standard.

- **phpMyAdmin :** It is a free and open source administration tool for MySQL and MariaDB.

3.4 FUNCTIONAL REQUIREMENTS

The functional requirements for a system describe what the system should do. It describes the system function in detail, its inputs and outputs, exceptions, and so on.

- **Collection of data :** The data collected from Google form are stored in file in .csv format.

- **Split data-set :** The data-set is divided into training and testing set using split ratio of 0.67.

- **Separate By Class :** The records are separated according to class i.e Donors and Non-Donors.

- **Classification :** Naive Bayes algorithm is used for this purpose.

- **User input :** It is necessary to know city where, and type of blood group required.

- **List the donors :** Based on the inputs given by the user, the donors who are willing to donate blood are listed.

3.5 Non-Functional Requirements

- **Portability :** The application should be portable to different computer systems running Linux.
- **Accuracy :** Since we will give the priority to the accuracy of the system, the performance will be based on the accuracy on classification.
- **Openness :** The system should be extensible to guarantee that it is useful for a reasonable period of time

3.6 Hardware Specifications

- Processor : Intel i3
- Speed : 2.2 GHz
- RAM : 4 GB
- Hard Disk : 1 TB

3.7 Software Requirements

- Operating System : Linux
- Programming Language : Python 2.7.12 (Classification), MySQL (Data-base), HTML, CSS, PHP, JavaScript (Front-End).

CHAPTER 4: TOOLS AND TECHNOLOGY USED

4.1 Linux

It is one of popular version of UNIX operating System. It is open source as its source code is freely available. It is free to use. Linux was designed considering UNIX compatibility. Its functionality list is quite similar to that of UNIX. Some important features of Linux operating system are:

- Portable
- Open source
- Multi User
- Multiprogramming
- Shell
- Security

Our prototype is written in Python which is platform independent. For the front end we have used HTML, CSS, PHP, JavaScript. Our development environment is Linux.

4.2 Python

It is a widely used high-level programming language for general-purpose programming. Python features a dynamic type system and automatic memory management and supports multiple programming paradigms, including object-oriented, imperative, functional programming, and procedural styles. It has a large and comprehensive standard library. The Python Package Index (PyPI) hosts thousands of third-party modules for Python. Both Python’s standard library and the community-contributed modules allow for endless possibilities.

- Web and Internet Development
- Database Access and Desktop GUIs
- Scientific and Numeric

- Education
- Network Programming
- Software and Game Development

4.3 Google forms

Google Forms is a tool that allows collecting information from users via a personalized survey or quiz. The information is then collected and automatically connected to a spreadsheet. The spreadsheet is populated with the survey and quiz responses. The Forms service has also received updates over the years. New features include, but are not limited to, menu search, shuffle of questions for randomized order, limiting responses to once per person, shorter URLs, custom themes, automatically generating answer suggestions when creating forms, and an "Upload file" option for users answering to share content through. Some Features of Google Forms are :

- See Responses, Live!
- Insert video/image.
- Multiple Choice Grids.
- Publish with pre-filled responses.
- Insert add-ons

4.4 CSV

In computing, a comma-separated values(CSV) files stores tabular data (numbers and text) in plain text. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format.

The CSV file format is not standardized. The basic idea of separating fields with a comma is clear, but that idea gets complicated when the field data may also contain commas or even embedded line-breaks. CSV implementations may not handle such field data, or they may use quotation marks to surround the field. Quotation does not solve everything: some fields may need embedded quotation marks, so a CSV implementation may include escape characters or escape sequences.

4.5 MySQLdb

MySQL is an open-source relational database management system(RDBMS). The MySQL development project has made its source code available under the terms of the GNU General Public License. MySQL is a central component of the LAMP open-source web application stack. Applications that use MySQL database include :

- TYPO3
- MODx
- WordPress
- Simple Machines Forum
- phpBB

- MyBB

MySQL is also used in many high profile, large-scale websites including Google, Facebook, Twitter, Flickr and Youtube.

4.6 phpMyAdmin

phpMyAdmin is a free and open-source administration tool for MySQL and MariaDB. As a portable web-application written primarily in PHP, it has become one of the most popular MySQL administration tools, especially for web hosting services. Features provided by phpMyAdmin include :

- Web interface
- MySQL and MariaDB database management
- Import data from CSV and SQL
- Export data to various formats: CSV, SQL, XML, PDF(via the TCPDF library), ISO/IEC 26300 - OpenDocument Text and Spreadsheet, Word, Excel, LaTeX and etc
- Administering multiple servers
- Creating PDF graphics of the database layout
- Creating complex queries using query-by-example (QBE)
- Searching globally in a database or a subset of it
- Transforming stored data into any format using a set of predefined functions, like displaying BLOB-data as image or download-link

- Live charts to monitor MySQL server activity like connections, processes, CPU/memory usage, etc.
- Working with different operating systems.

CHAPTER 5: SYSTEM DESIGN

Steps:

- **Collection of data :** For our project we used Google form to create data-set. The form consisted of two parts, first part was used to collect personal information of the donor and in the latter part, questionnaire consisting of 15 questions were present. In total we had created data-set of 246 records, which were stored in a file with .csv extension.

- **Split data-set :** The data-set is divided into training and testing set using split ratio of 0.67.

- **Separate By Class :** The records are separated according to class i.e Donors and Non-Donors.

- **Classification :** Naive Bayes algorithm is used for this purpose.

- **User input :** It is necessary to know city where, and type of blood group required.

- **List the donors :** Based on the inputs given by the user, the donors who are willing to donate blood are listed.

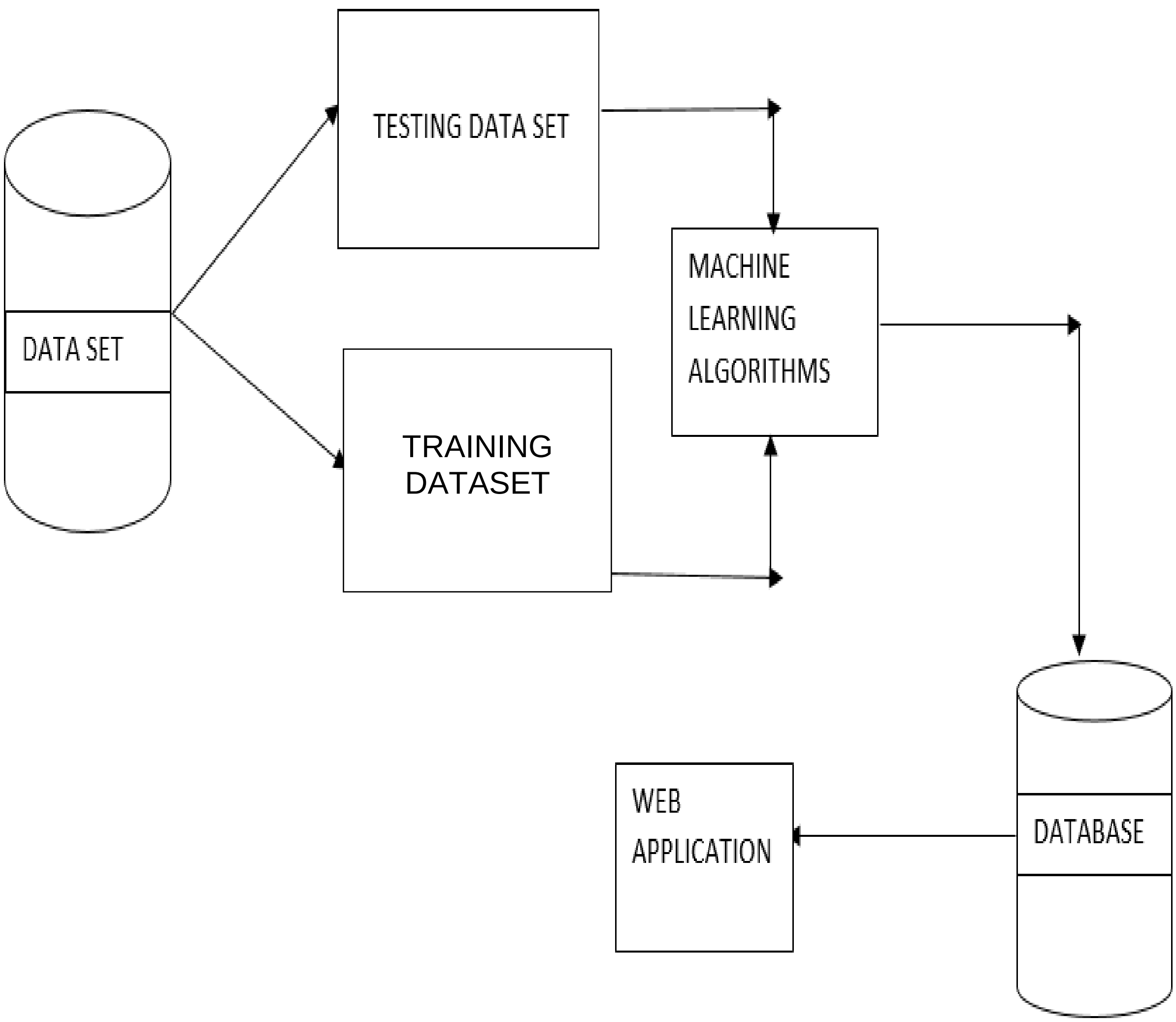


Fig 4 : Work Flow of the System

5.1: SOFTWARE DESIGN:

5.1.1: NAIVE BAYES:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

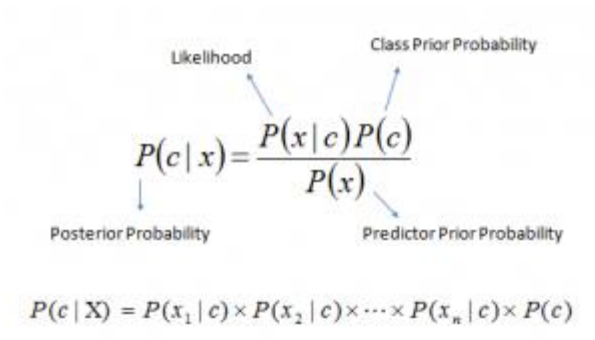


Fig 5: Bayes theorem formula.

Above,

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Advantages of Naïve Bayes:

- It is easy and fast to predict class of test data set. It also perform well in multi class prediction
- When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.

- It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

Disadvantages of Naïve Bayes:

- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
- Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

Applications of Naive Bayes Algorithms:

- **Real time Prediction:** Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time.
- **Multi class Prediction:** This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.
- **Text classification/ Spam Filtering/ Sentiment Analysis:** Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)

Improving the power of Naive Bayes Model:

- If continuous features do not have normal distribution, we should use transformation or different methods to convert it in normal distribution.

- If test data set has zero frequency issue, apply smoothing techniques “Laplace Correction” to predict the class of test data set.
- Remove correlated features, as the highly correlated features are voted twice in the model and it can lead to over inflating importance.
- Naive Bayes has no variance to minimize.

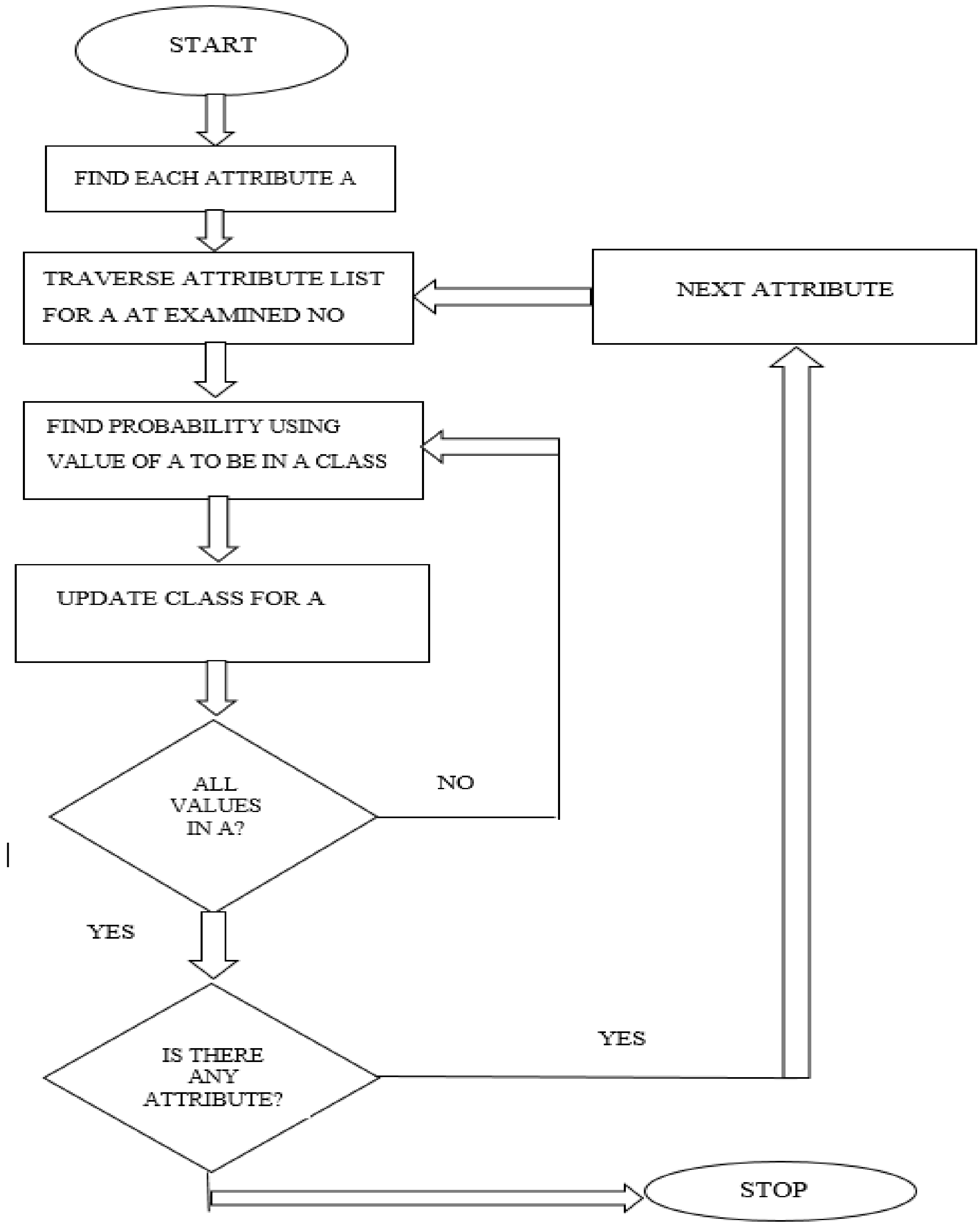


Fig 6: Flowchart for Naïve Bayes

5.1.2: KNN ALGORITHM:

K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition as a non-parametric technique.

Algorithm:

A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbour.

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3 and 10.

Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

Fig7: Distance formula for KNN.

Advantages of KNN:

- Can be applied to the data from any distribution for example, data does not have to be separable with a linear boundary.
- Very simple and intuitive.
- Good classification if the number of samples is large enough.

Disadvantages of KNN:

- Dependent on K Value.
- Test stage is computationally expensive.
- No training stage, all the work is done during the test stage.
- This is actually the opposite of what we want. Usually we can afford training step to take a long time, but we want fast test step.
- Need large number of samples for accuracy.

Improvise the data for KNN:

- **Rescale Data**: KNN performs much better if all of the data has the same scale. Normalizing your data to the range [0, 1] is a good idea. It may also be a good idea to standardize your data if it has a Gaussian distribution.
- **Address Missing Data**: Missing data will mean that the distance between samples cannot be calculated. These samples could be excluded or the missing values could be imputed.
- **Lower Dimensionality**: KNN is suited for lower dimensional data. You can try it on high dimensional data (hundreds or thousands of input variables) but be aware that it may not perform as well as other techniques. KNN can benefit from feature selection that reduces the dimensionality of the input feature space.

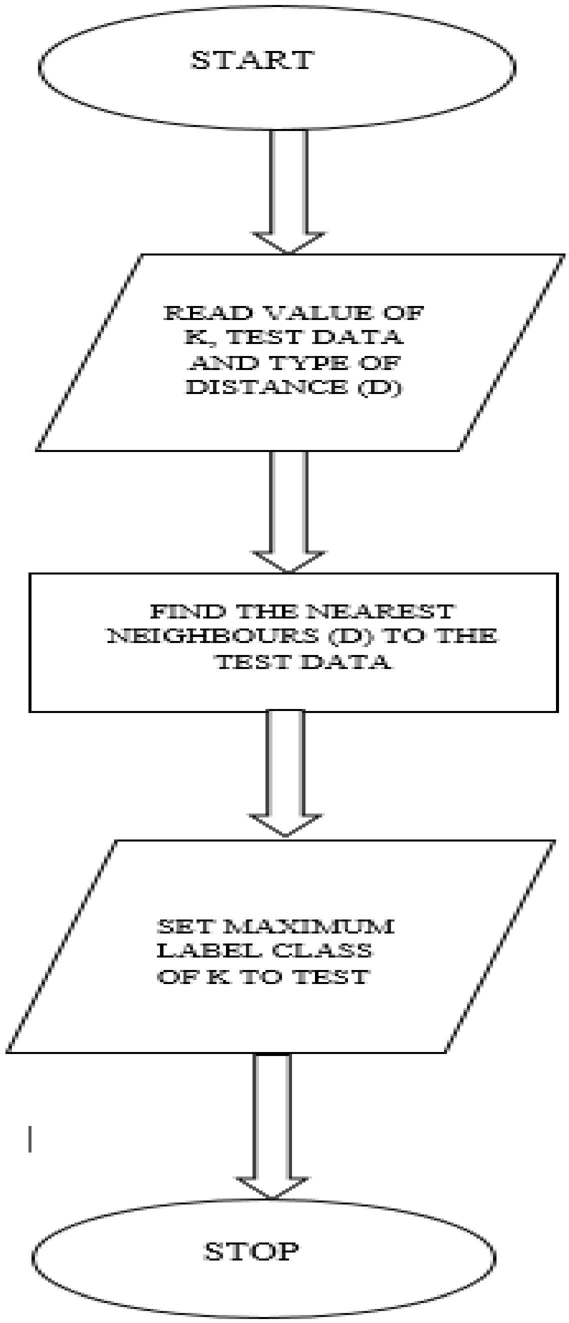


Fig 8: Flowchart for KNN

CHAPTER 6: SYSTEM IMPLEMENTATION

6.1 Creating Data-Set

The data-set is created using Google Forms. It comprises of two parts, the first part is used to collect personal details of the donors and the latter part consists of questionnaire. The questionnaire consists of 15 questions, 14 of which are likert-type scale based having inputs from 1 to 5, and one question asks opinion of the person on how he should be contacted.

The data- set comprises of 246 records. The questions that are present in the questionnaire are listed below :

- How obvious is it to you, that you will donate blood in the coming six months?
- For me to, donate blood is?
- Most people who are important to me think that I (should/should not) donate blood?
- Generally speaking, how much do you care what people think you should do?
- How often do you feel ill, tired, or listless? 6. For me to save life of people, by donating blood is?
- How often do you go out of station? 8. I would avoid aiding someone in a medical emergency if I could.
- I’m scared of being sick after donating blood?
- How often do other courses places heavy demands on your time?
- If I encountered unanticipated events that placed demands on my time, it would make it more difficult for me to donate blood?
- If I felt ill, tired, or listless, it would make it more difficult for me to attend the meetings of this class on a regular basis.
- Which media do you prefer as a means of getting information?

- If you come across information regarding need for blood, how likely are you to recommend anyone you know with same blood group?
- My personal values encourage me to donate blood.

The link to the google form is https://goo.gl/forms/6AhjxXOzDasgaut53

The following figures shows google form :

Fig 9: Google form to enter the donor’s personal information (Part 1)

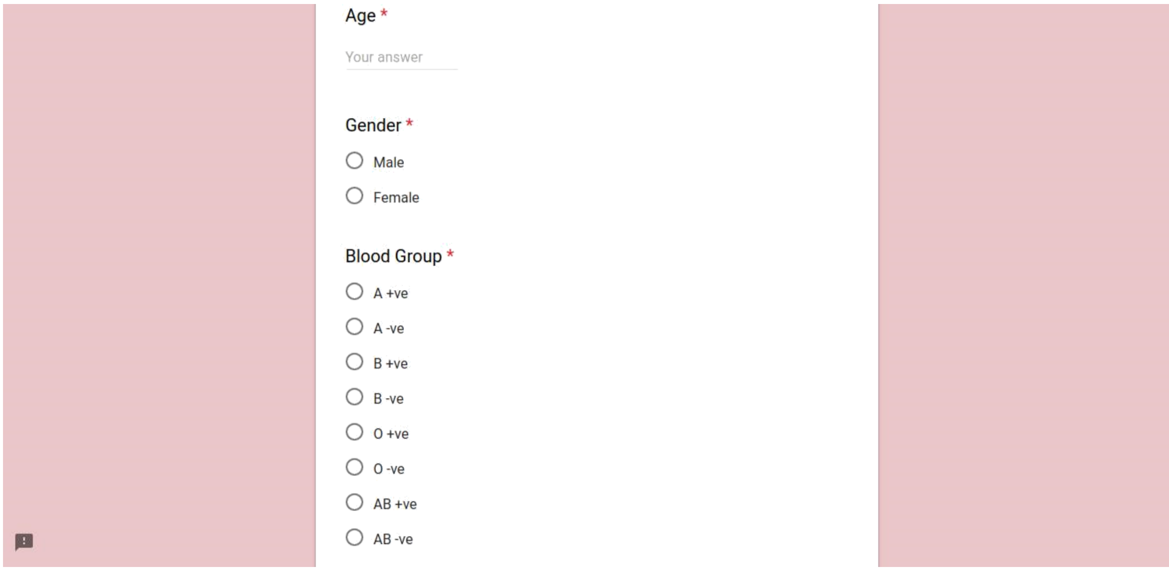


Fig 10: Google form to enter the donor's personal information (Part 2)

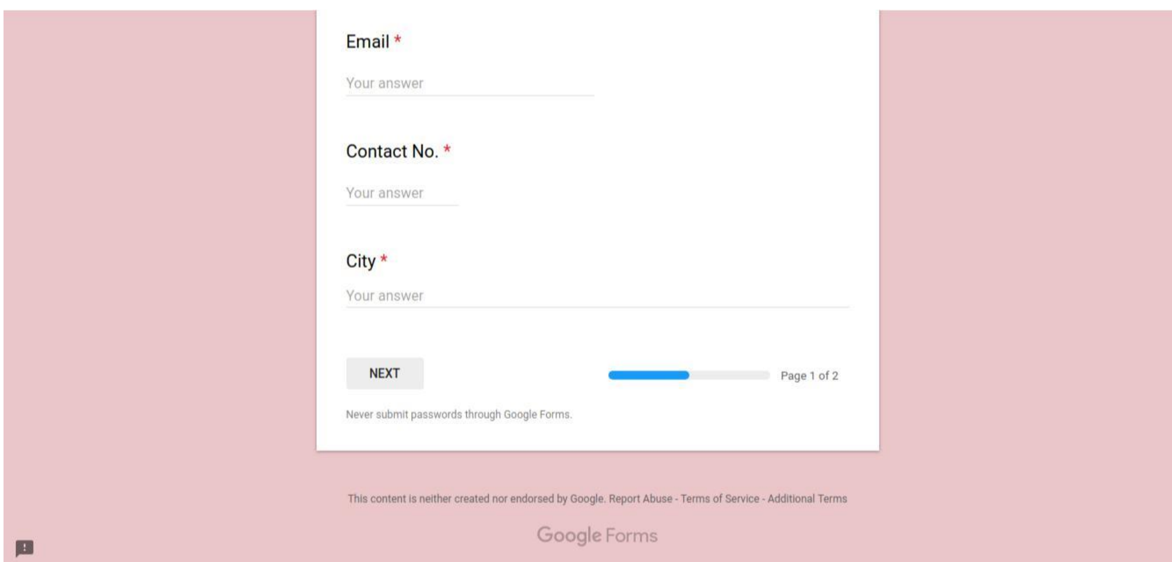
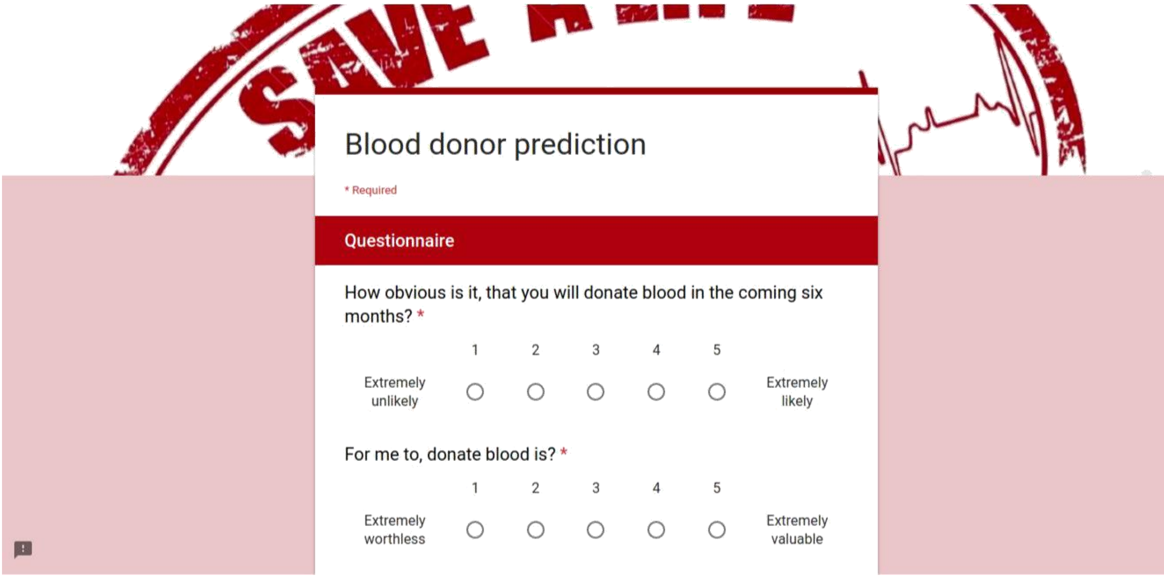


Fig 11: Google form to enter the donor's personal information (Part 3)



Blood donor prediction

* Required

Questionnaire

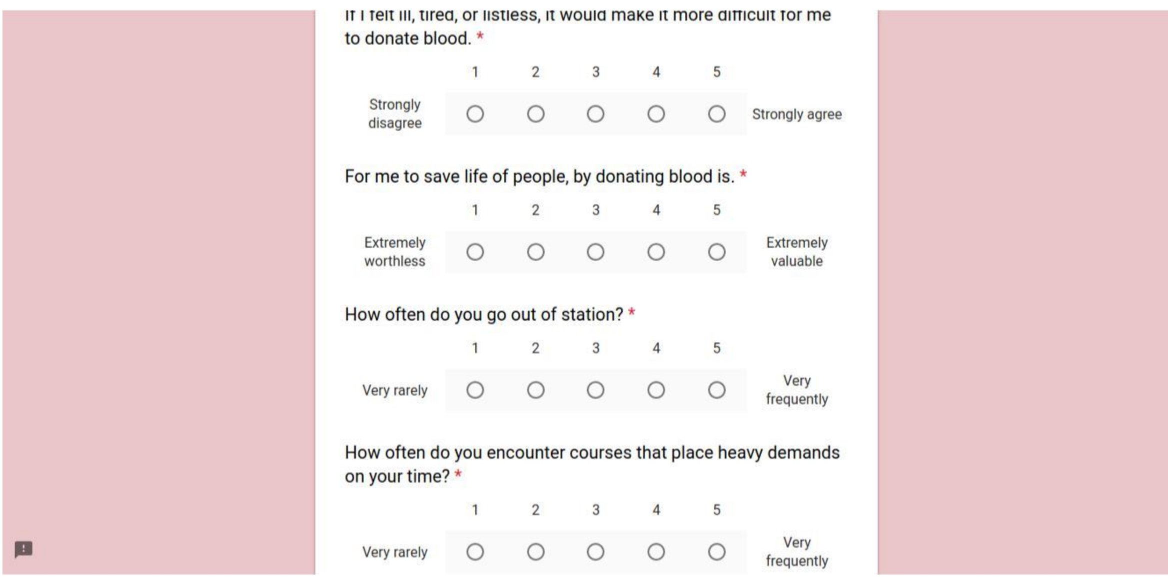
How obvious is it, that you will donate blood in the coming six months? *

	1	2	3	4	5	
Extremely unlikely	○	○	○	○	○	Extremely likely

For me to, donate blood is? *

	1	2	3	4	5	
Extremely worthless	○	○	○	○	○	Extremely valuable

Fig 12: Questionnaire (Part 1)



If I felt ill, tired, or listless, it would make it more difficult for me to donate blood. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

For me to save life of people, by donating blood is. *

	1	2	3	4	5	
Extremely worthless	○	○	○	○	○	Extremely valuable

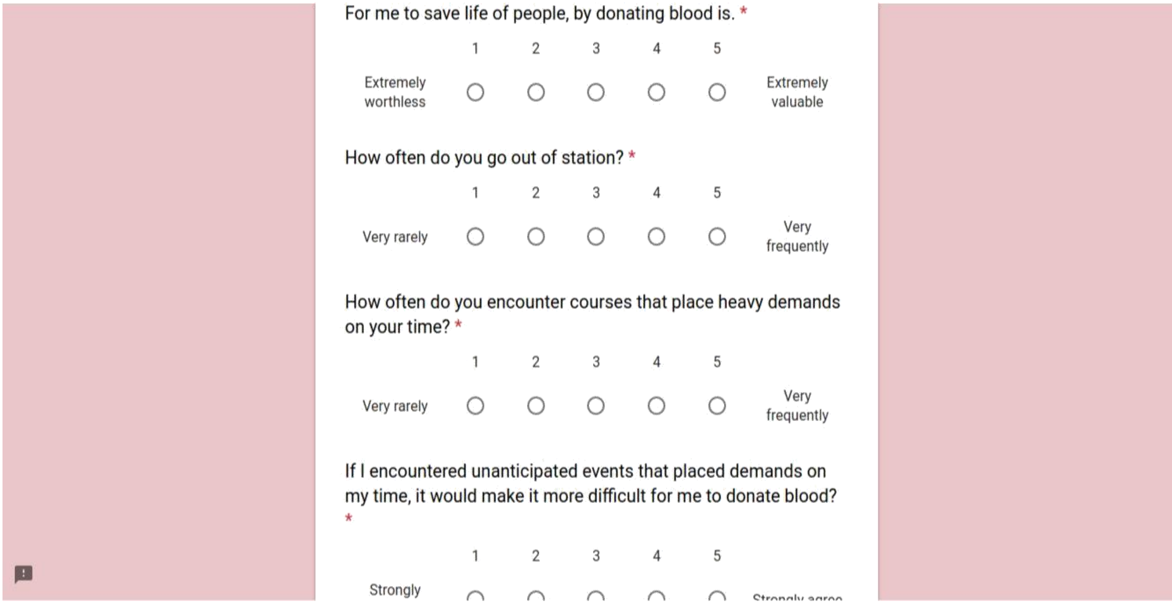
How often do you go out of station? *

	1	2	3	4	5	
Very rarely	○	○	○	○	○	Very frequently

How often do you encounter courses that place heavy demands on your time? *

	1	2	3	4	5	
Very rarely	○	○	○	○	○	Very frequently

Fig 13: Questionnaire (Part 2)



For me to save life of people, by donating blood is. *

	1	2	3	4	5	
Extremely worthless	○	○	○	○	○	Extremely valuable

How often do you go out of station? *

	1	2	3	4	5	
Very rarely	○	○	○	○	○	Very frequently

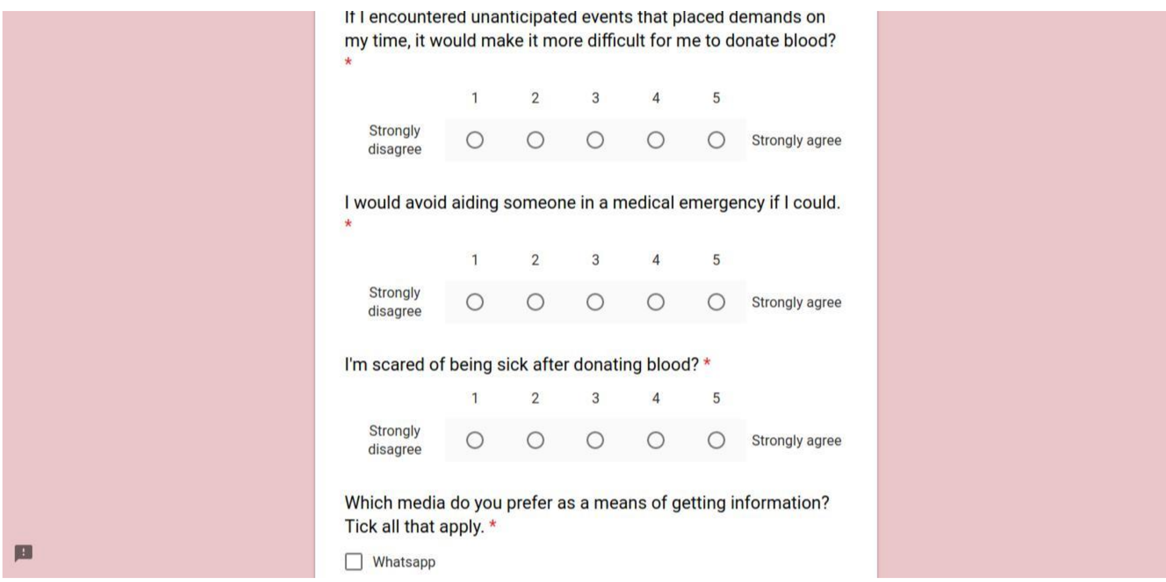
How often do you encounter courses that place heavy demands on your time? *

	1	2	3	4	5	
Very rarely	○	○	○	○	○	Very frequently

If I encountered unanticipated events that placed demands on my time, it would make it more difficult for me to donate blood? *

	1	2	3	4	5	
Strongly						Strongly agree

Fig 14: Questionnaire (Part 3)



If I encountered unanticipated events that placed demands on my time, it would make it more difficult for me to donate blood? *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I would avoid aiding someone in a medical emergency if I could. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

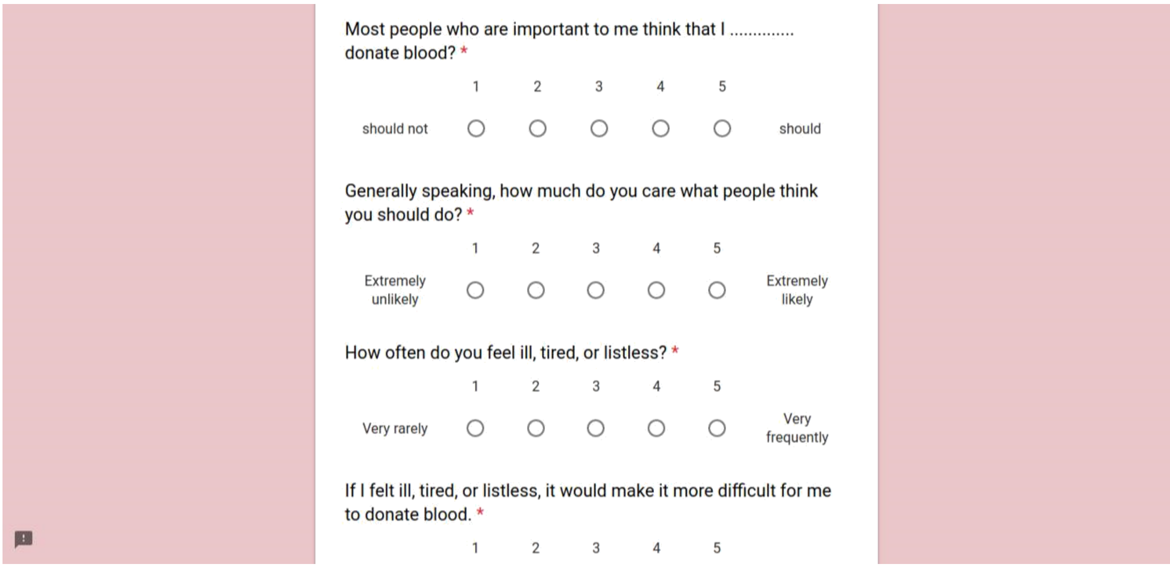
I'm scared of being sick after donating blood? *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

Which media do you prefer as a means of getting information? Tick all that apply. *

☐ Whatsapp

Fig 15: Questionnaire (Part 4)



Most people who are important to me think that I .... donate blood? *

	1	2	3	4	5	
should not	○	○	○	○	○	should

Generally speaking, how much do you care what people think you should do? *

	1	2	3	4	5	
Extremely unlikely	○	○	○	○	○	Extremely likely

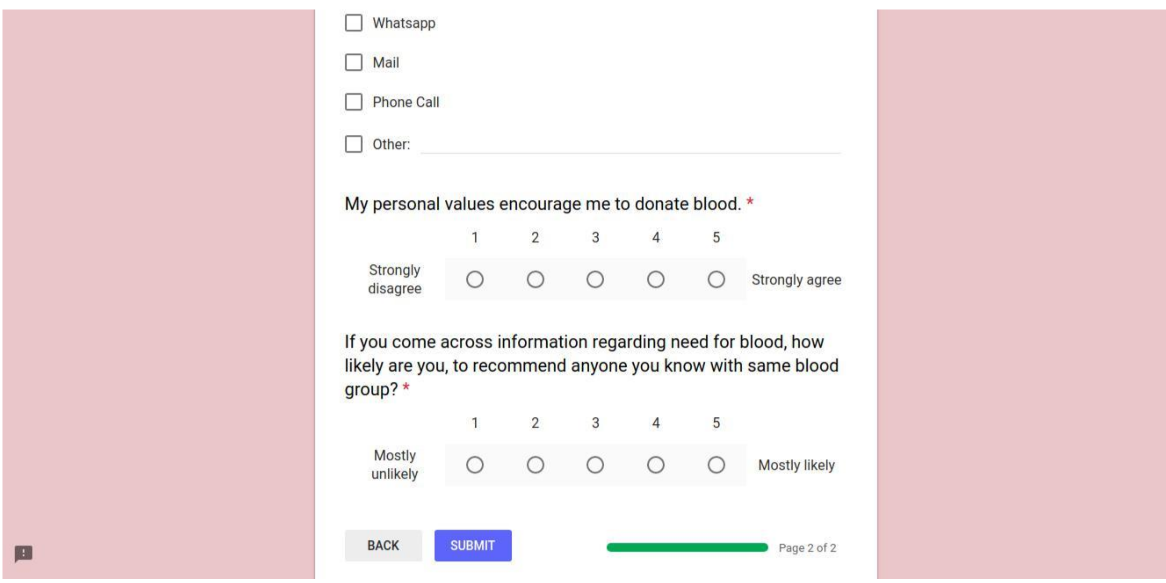
How often do you feel ill, tired, or listless? *

	1	2	3	4	5	
Very rarely	○	○	○	○	○	Very frequently

If I felt ill, tired, or listless, it would make it more difficult for me to donate blood. *

1 2 3 4 5

Fig 16: Questionnaire (Part 5)



- ☐ Whatsapp
- ☐ Mail
- ☐ Phone Call
- ☐ Other: ______

My personal values encourage me to donate blood. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

If you come across information regarding need for blood, how likely are you, to recommend anyone you know with same blood group? *

	1	2	3	4	5	
Mostly unlikely	○	○	○	○	○	Mostly likely

BACK SUBMIT Page 2 of 2

Fig 17: Questionnaire (Part 6)

6.2 Classification

For Classification we have used Naive Bayes classification algorithm. The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modelling problem probabilistically.

Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.

The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class.

To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.

Naive bases is often described using categorical data because it is easy to describe and calculate using ratios. A more useful version of the algorithm for our purposes supports numeric attributes and assumes the values of each numerical attribute are normally distributed (fall somewhere on a bell curve). Again, this is a strong assumption, but still gives robust results.

The classification problem is divided into the following steps:

- **Handle Data :** Load data from CSV file and split it into training and test datasets.
- **Summarize Data :** Summarize the properties in the training dataset so that we can calculate probabilities and make predictions.
- **Make a Prediction**: Use the summaries of the dataset to generate a single prediction.
- **Make Predictions :** Generate predictions given a test dataset and a summarized training dataset.

- **Evaluate Accuracy :** Evaluate the accuracy of predictions made for a test dataset as the percentage correct out of all predictions made.

6.2.1 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.

We also need to convert the attributes that were loaded as strings into numbers that we can work with them. Below is the **loadCsv()** function for loading the dataset.

```
import csv

def loadCsv(filename):

    lines = csv.reader(open(filename, "rb"))

    dataset = list(lines)

    for i in range(len(dataset)):

        dataset[i] = [float(x) for x in dataset[i]] return dataset
```

Next we need to split the data into a training dataset that Naive Bayes can use to make predictions and a test dataset that we can use to evaluate the accuracy of the model. We need to split the data set randomly into train and datasets with a ratio of 67% train and 33% test (this is a common ratio for testing an algorithm on a dataset).
Below is the **splitDataset()** function that will split a given dataset into a given split ratio.

```
import random
```

```
def splitDataset(dataset, splitRatio): trainSize = int(len(dataset) *
        splitRatio) trainSet = []
        copy = list(dataset)
        while len(trainSet) < trainSize:
                index = random.randrange(len(copy))
                trainSet.append(copy.pop(index))
        return [trainSet, copy]
```

6.2.2 Summarize Data :

The naive bayes model is comprised of a summary of the data in the training dataset. This summary is then used when making predictions.

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.

We can break the preparation of this summary data down into the following sub-tasks:

- Separate Data by Class
- Calculate Mean
- Summarize Dataset
- Summarize Attributes by Class

Separate Data By Class

The first task is to separate the training dataset instances by class value so that we can calculate statistics for each class. We can do that by creating a map of each class value to a list of instances that belong to that class and sort the entire dataset of instances into the appropriate lists.

The **separateByClass()** function below does just this.

```
def separateByClass(dataset):
        separated = {}
                 for i in range(len(dataset)):
                vector = dataset[i]
                if (vector[-1] not in separated):
                        separated[vector[-1]] = []
                separated[vector[-1]].append(vector)
        return separated
```

You can see that the function assumes that the last attribute (-1) is the class value. The function returns a map of class values to lists of data instances.

Calculate Mean

We need to calculate the mean of each attribute for a class value. The mean is the central middle or central tendency of the data, and we will use it as the middle of our gaussian distribution when calculating probabilities.
We also need to calculate the standard deviation of each attribute for a class value. The standard deviation describes the variation of spread of the data, and we will use it to characterize the expected spread of each attribute in our Gaussian distribution when calculating probabilities.

The standard deviation is calculated as the square root of the variance. The variance is calculated as the average of the squared differences for each attribute value from the mean. Note we are using the N-1 method, which subtracts 1 from the number of attribute values when calculating the variance.

```
import math
def mean(numbers):
	return sum(numbers)/float(len(numbers))

def stdev(numbers):
	avg = mean(numbers)
	variance = sum([pow(x-avg,2) for x in
numbers])/float(len(numbers)-1)
	return math.sqrt(variance)
```

Summarize Dataset

Now we have the tools to summarize a dataset. For a given list of instances (for a class value) we can calculate the mean and the standard deviation for each attribute.
The zip function groups the values for each attribute across our data instances into their own lists so that we can compute the mean and standard deviation values for the attribute.

```
def summarize(dataset):
	summaries = [(mean(attribute), stdev(attribute)) for
attribute in zip(*dataset)]
	del summaries[-1]
	return summaries
```

Summarize Attributes By Class

We can pull it all together by first separating our training dataset into instances grouped by class. Then calculate the summaries for each attribute.

```
def summarizeByClass(dataset):
	separated = separateByClass(dataset)
	summaries = {}
	for classValue, instances in separated.iteritems():
		summaries[classValue] = summarize(instances)
	return summaries
```

6.2.3 Make a Prediction

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. We can divide this part into the following tasks:

- Calculate Gaussian Probability Density Function
- Calculate Class Probabilities
- Make a Prediction
- Estmate Accuracies

Calculate Gaussian Probability Density Function

We can use a Gaussian function to estimate the probability of a given attribute value, given the known mean and standard deviation for the attribute estimated from the training data. Given that the attribute summaries where prepared for each attribute and class value, the result is the conditional probability of a given attribute value given a class value.

See the references for the details of this equation for the Gaussian probability density function. In summary we are plugging our known details into the Gaussian (attribute value, mean and standard deviation) and reading off the likelihood that our attribute value belongs to the class.

In the **calculateProbability()** function we calculate the exponent first, then calculate the main division. This lets us fit the equation nicely on two lines.

```
import math
def calculateProbability(x, mean, stdev):
	exponent = math.exp(-(math.pow(x-mean,2)/
(2*math.pow(stdev,2))))
	return (1 / (math.sqrt(2*math.pi) * stdev)) * exponent
```

Calculate Class Probabilities

Now that we can calculate the probability of an attribute belonging to a class, we can combine the probabilities of all of the attribute values for a data instance and come up with a probability of the entire data instance belonging to the class.
We combine probabilities together by multiplying them. In the **calculateClassProbabilities()** below, the probability of a given data instance is calculated by multiplying together the attribute probabilities for each class, the result is a map of class values to probabilities.

```
def calculateClassProbabilities(summaries, inputVector):
	probabilities = {}
	for classValue, classSummaries in summaries.iteritems():
		probabilities[classValue] = 1
		for i in range(len(classSummaries)):
			mean, stdev = classSummaries[i]
```

```
                x = inputVector[i]
                probabilities[classValue] *=
calculateProbability(x, mean, stdev)
    return probabilities
```

6.2.4 Make Predictions

Finally, we can estimate the accuracy of the model by making predictions for each data instance in our test dataset. The **getPredictions()** will do this and return a list of predictions for each test instance.

```
def predict(summaries, inputVector):
    probabilities = calculateClassProbabilities(summaries, inputVector)
    bestLabel, bestProb = None, -1
    for classValue, probability in probabilities.iteritems():
        if bestLabel is None or probability > bestProb:
            bestProb = probability
            bestLabel = classValue
    return bestLabel

def getPredictions(summaries, testSet):
    predictions = []
    for i in range(len(testSet)):
        result = predict(summaries, testSet[i])
        predictions.append(result)
    return predictions
```

6.2.5 Get Accuracy

The predictions can be compared to the class values in the test dataset and a classification accuracy can be calculated as an accuracy ratio between 0 and 100%. The **getAccuracy()** will calculate this accuracy ratio.

```
def getAccuracy(testSet, predictions):
    correct = 0
    for x in range(len(testSet)):
        if testSet[x][-1] == predictions[x]:
            correct += 1
    return (correct/float(len(testSet))) * 100.0
```

CHAPTER 7: SYSTEM TESTING AND RESULT ANALYSIS

7.1 RESULT ANALYSIS AND DISCUSSION

7.1.1 NAÏVE BAYES

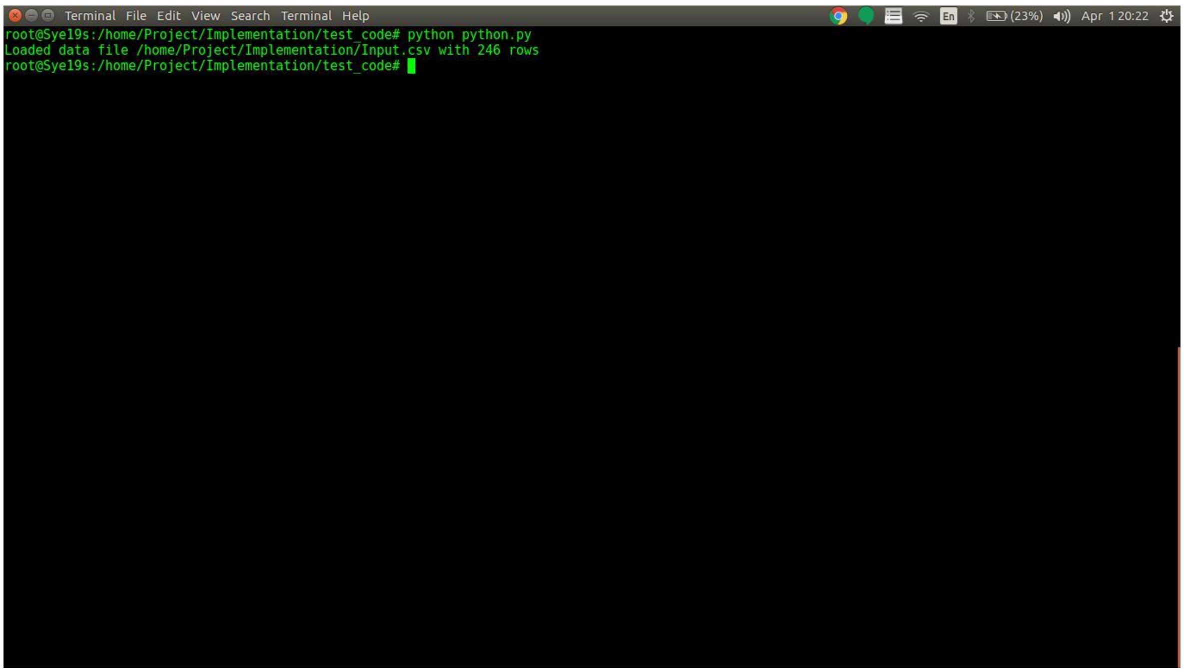


Fig 18: Loading Dataset

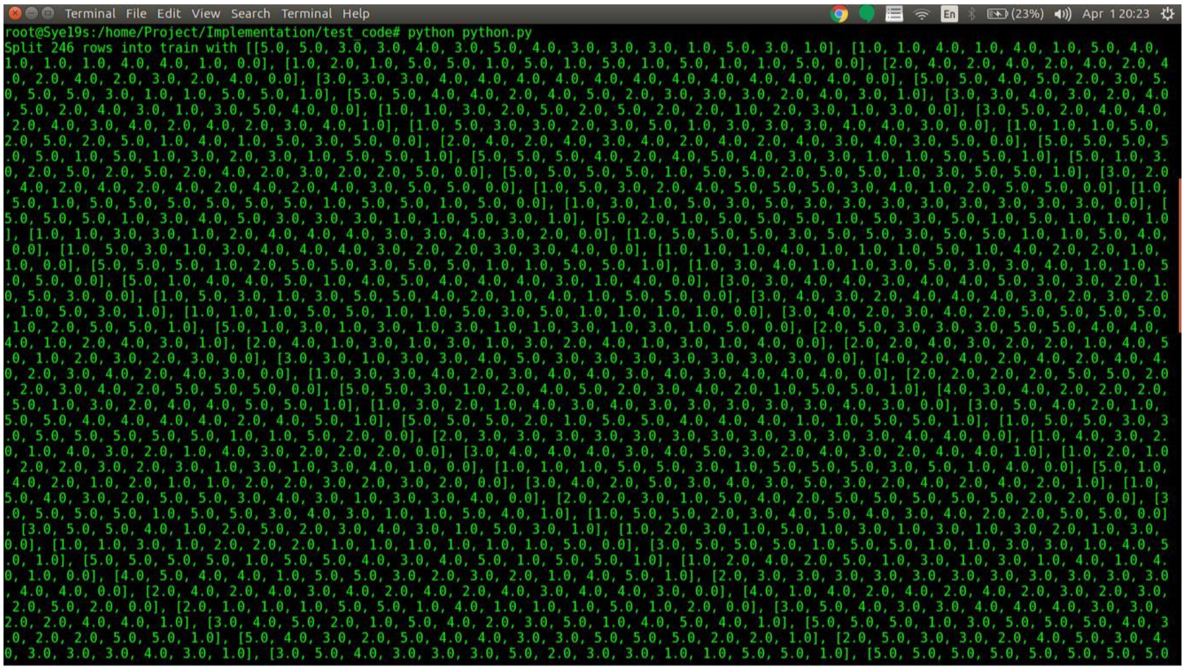


Fig 19: Splitting Dataset

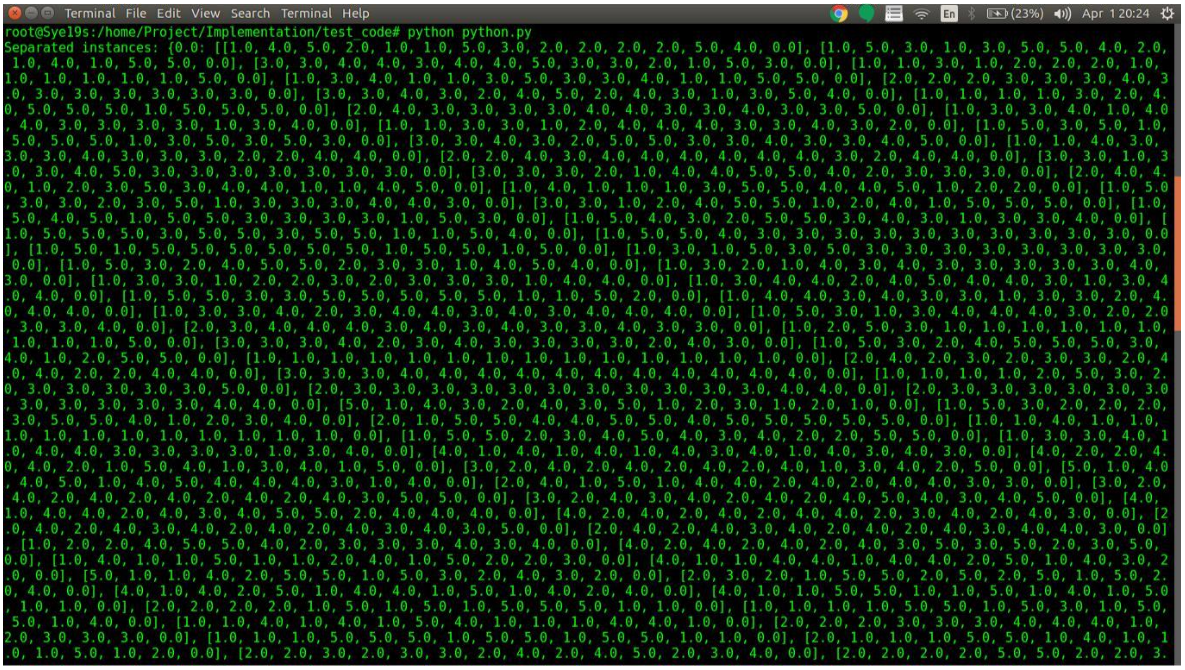


Fig 20: Separating instances

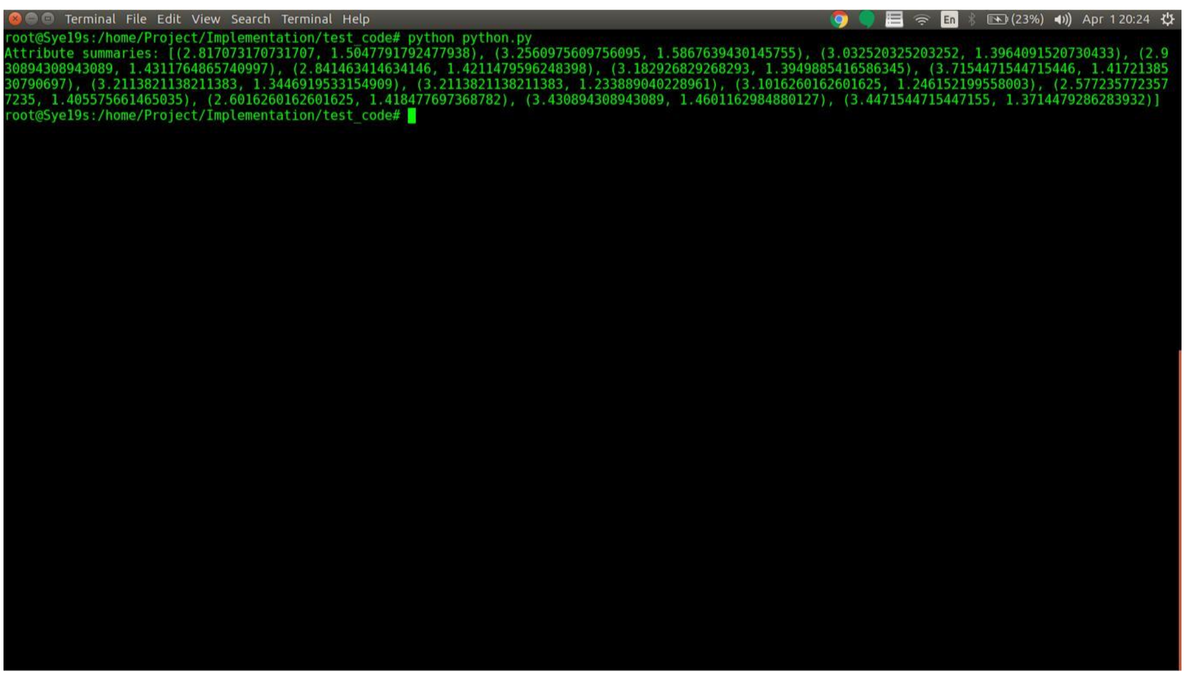


Fig 21: Summarizing attributes

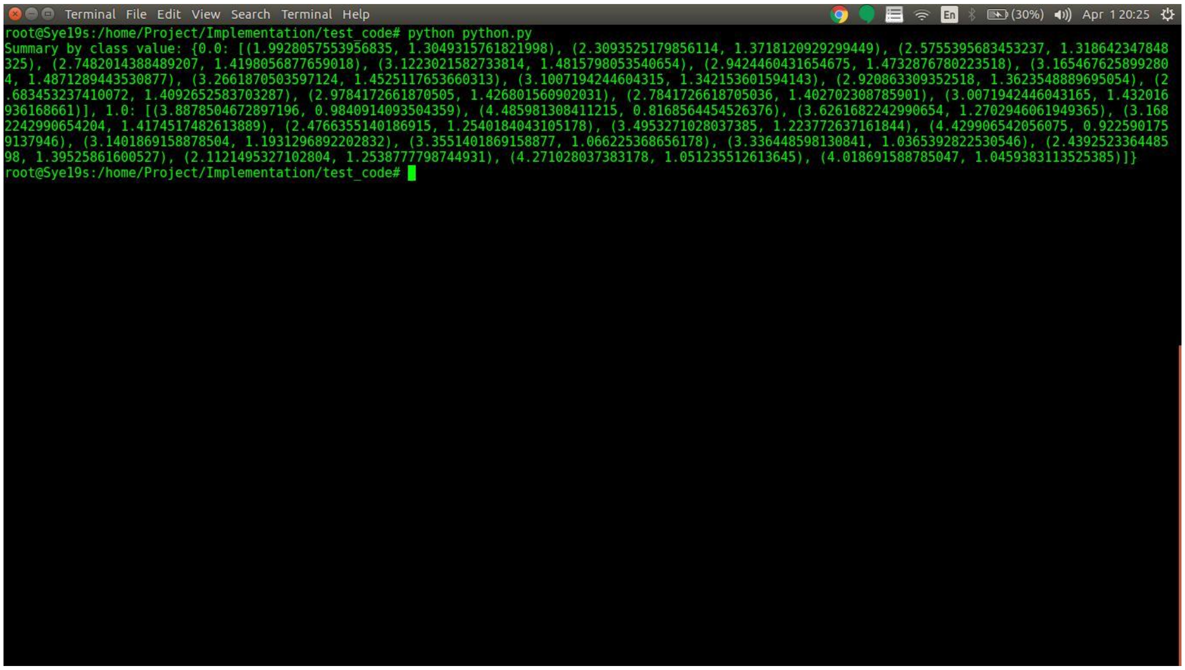


Fig 22: Summarizing by class values

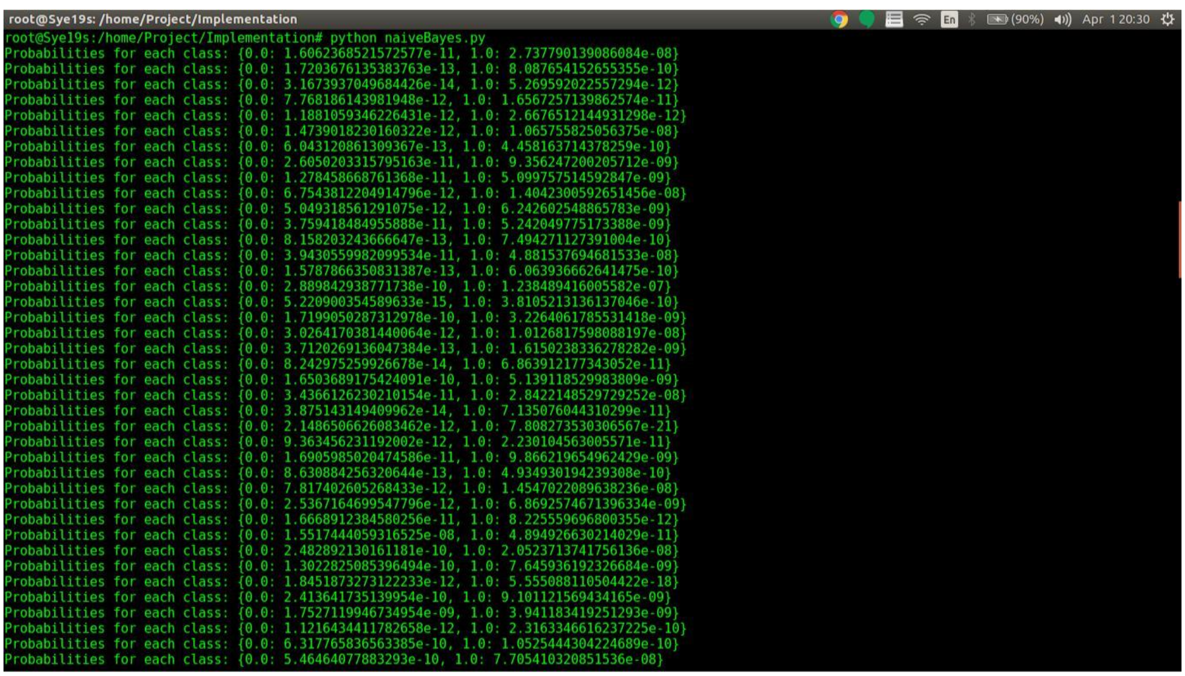


Fig 23: Calculating probabilities for each class

Fig 24: Predictions

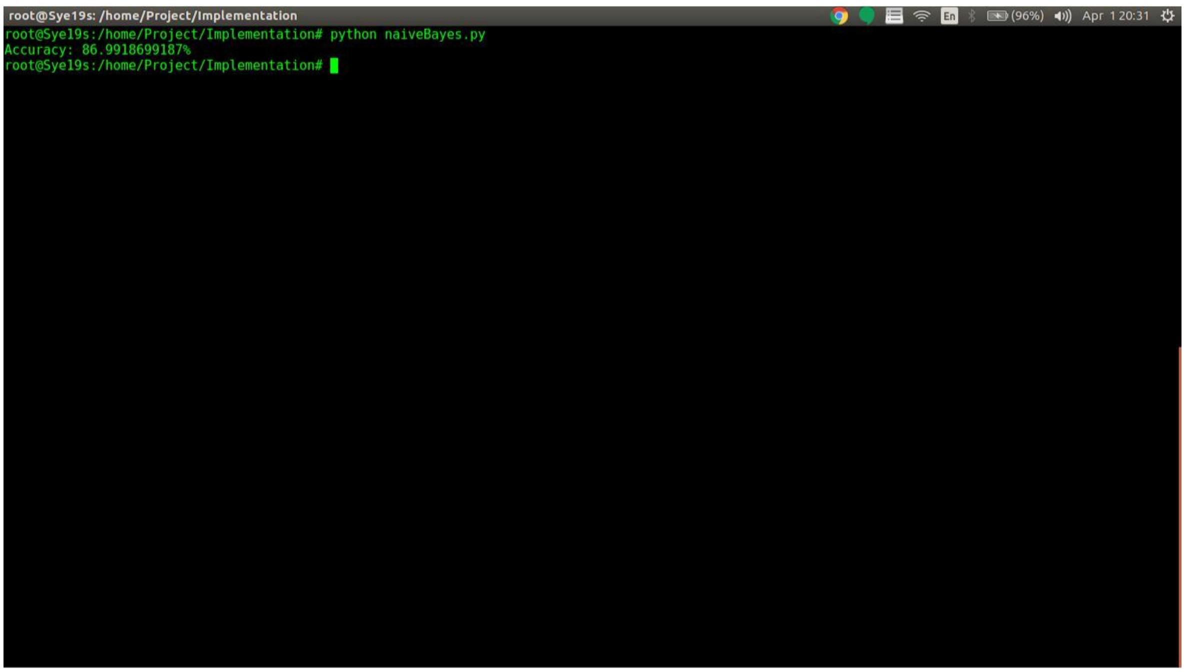


Fig 25: Calculating accuracy

7.1.2 KNN ALGORITHM

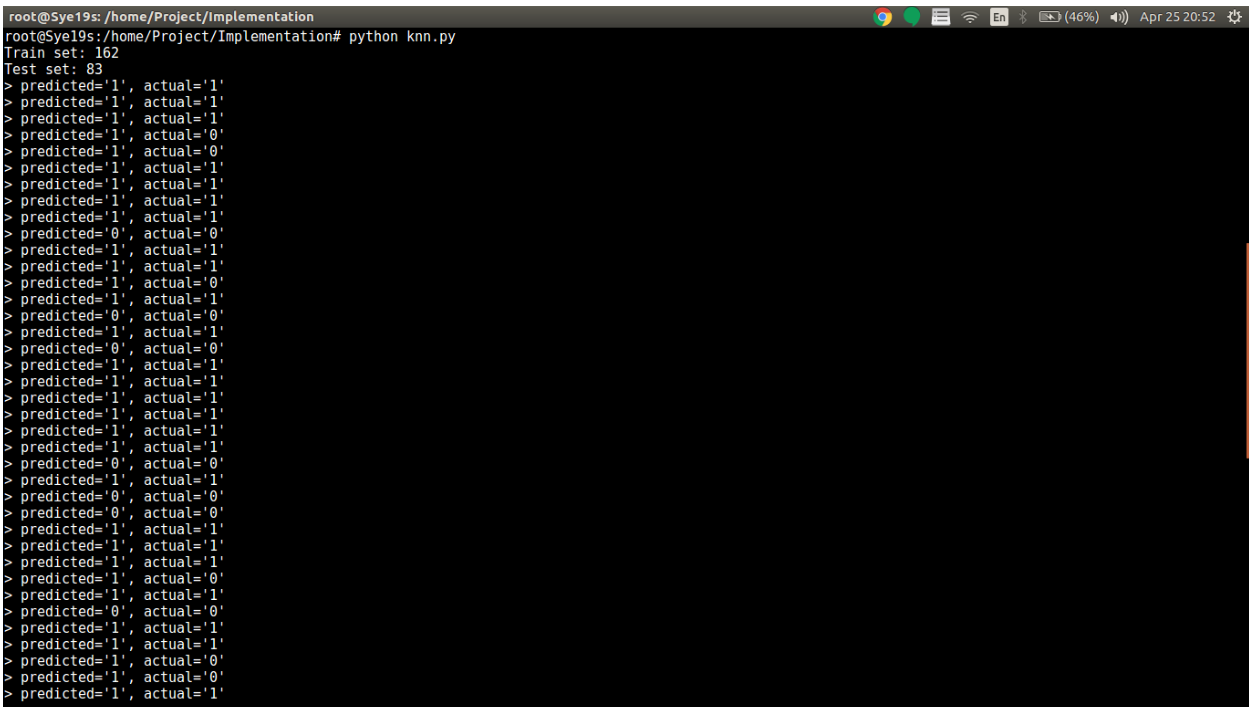


Fig 26 : Splitting dataset into training and testing set

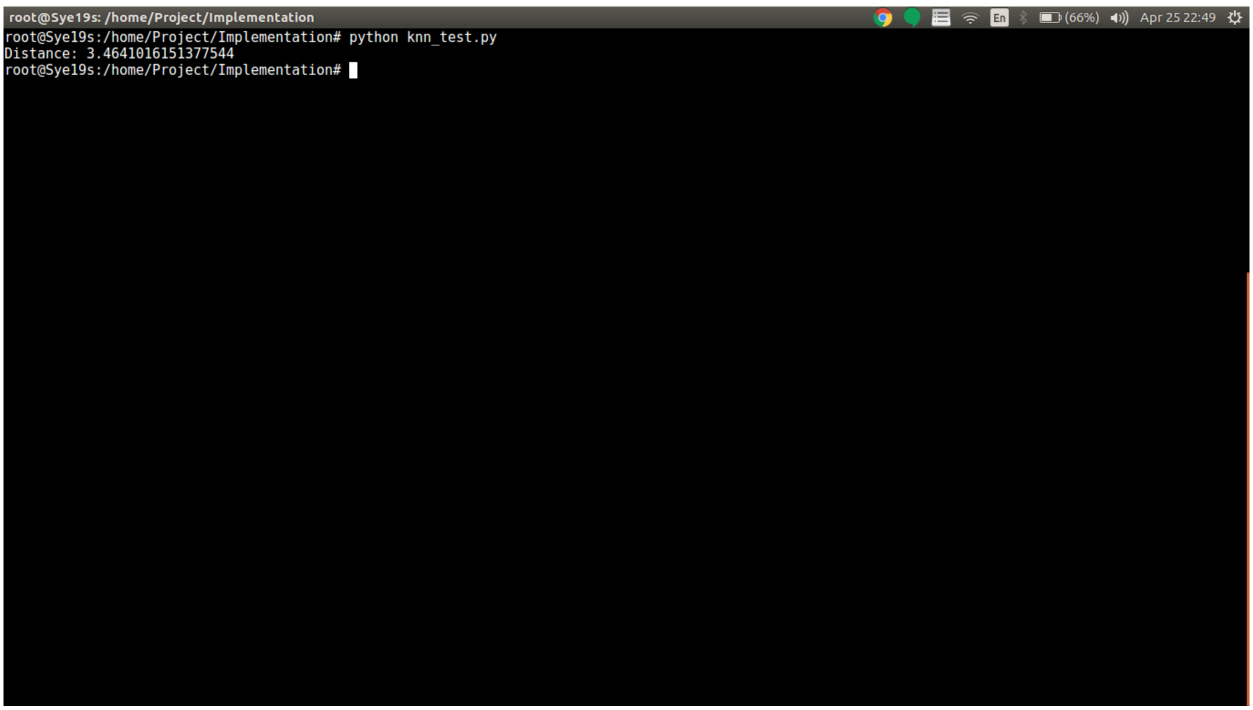


Fig 27 : Calculating distance

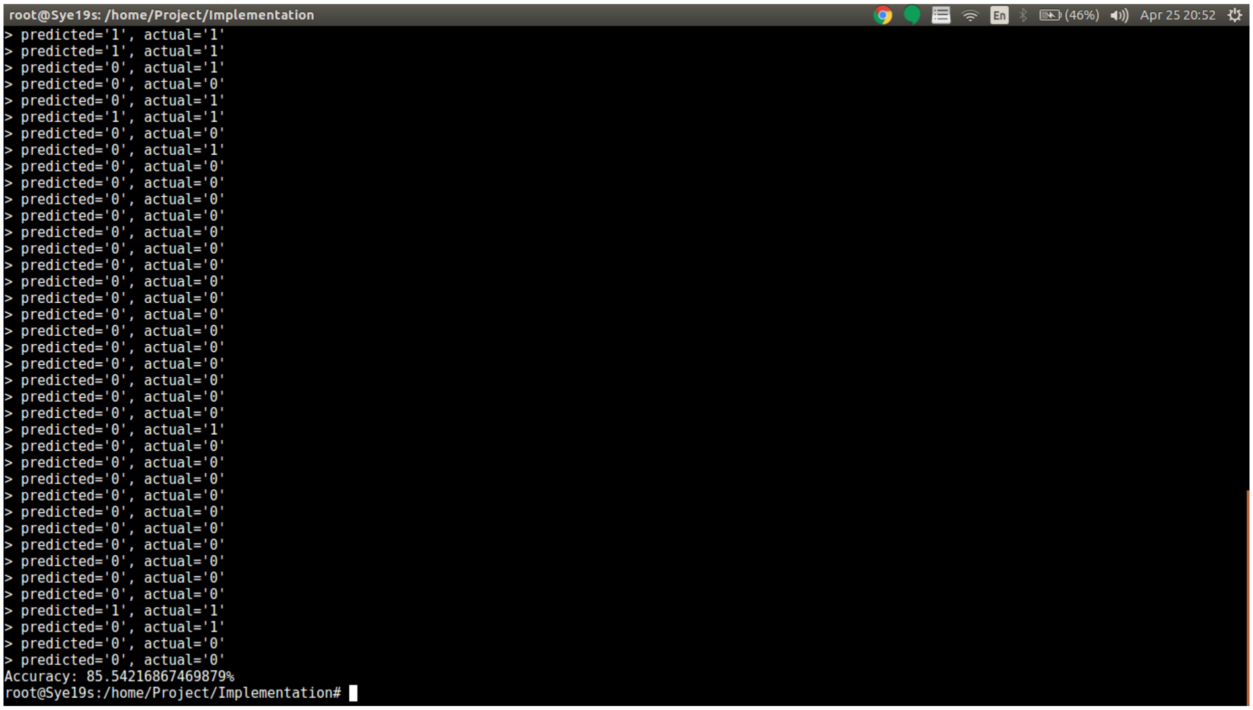


Fig 28 : Calculating Accuracy

It is clear that KNN algorithm provides more efficiency that the Naïve Bayes algorithm. As KNN algorithm's efficiency ranges from 83% to 93% and that of Naïve Bayes is from 78% to 88%.

It can be therefore concluded that the KNN algorithm is the more preferable choice for the prediction of blood donor compared to Naïve Bayes algorithm, even though KNN is complex than the naive bayes algorithm. It can also be seen that naive bayes provides an accuracy of nearly 96% for our experimental database.

7.2 SCREENSHOTS OF THE OUTPUT

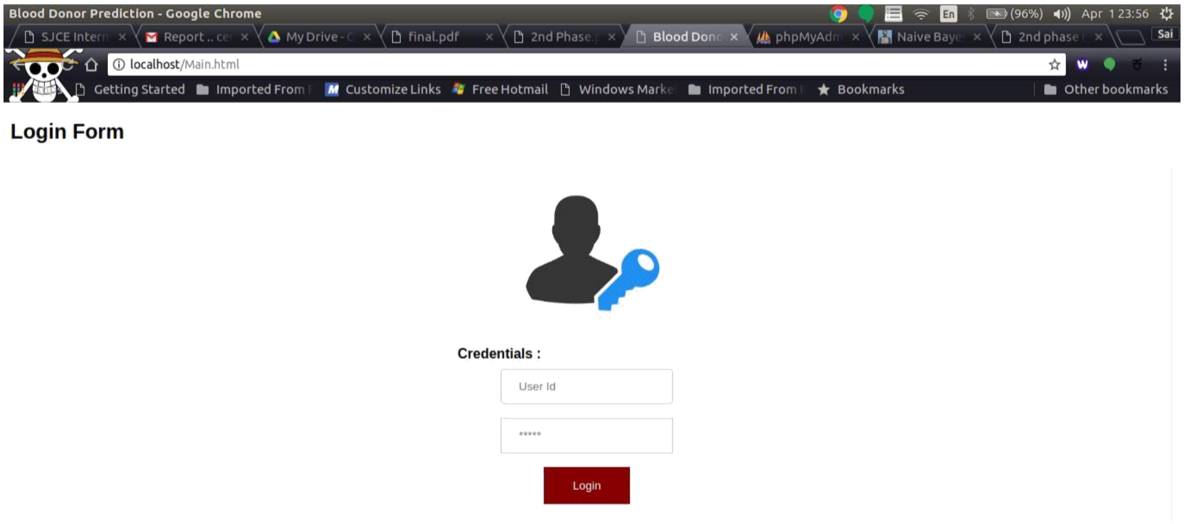


Fig 29: Login Page

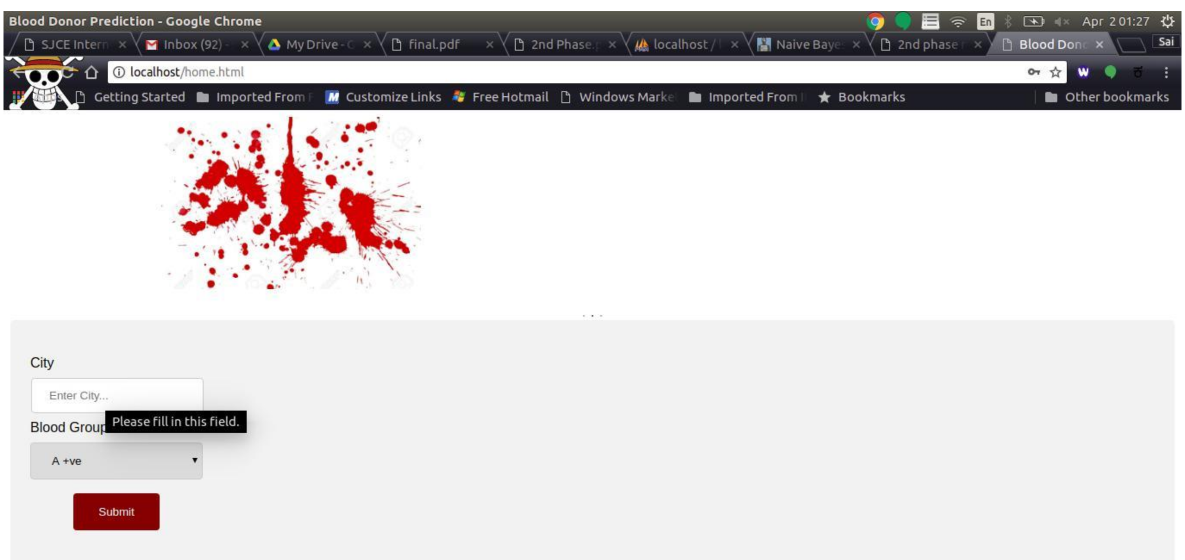
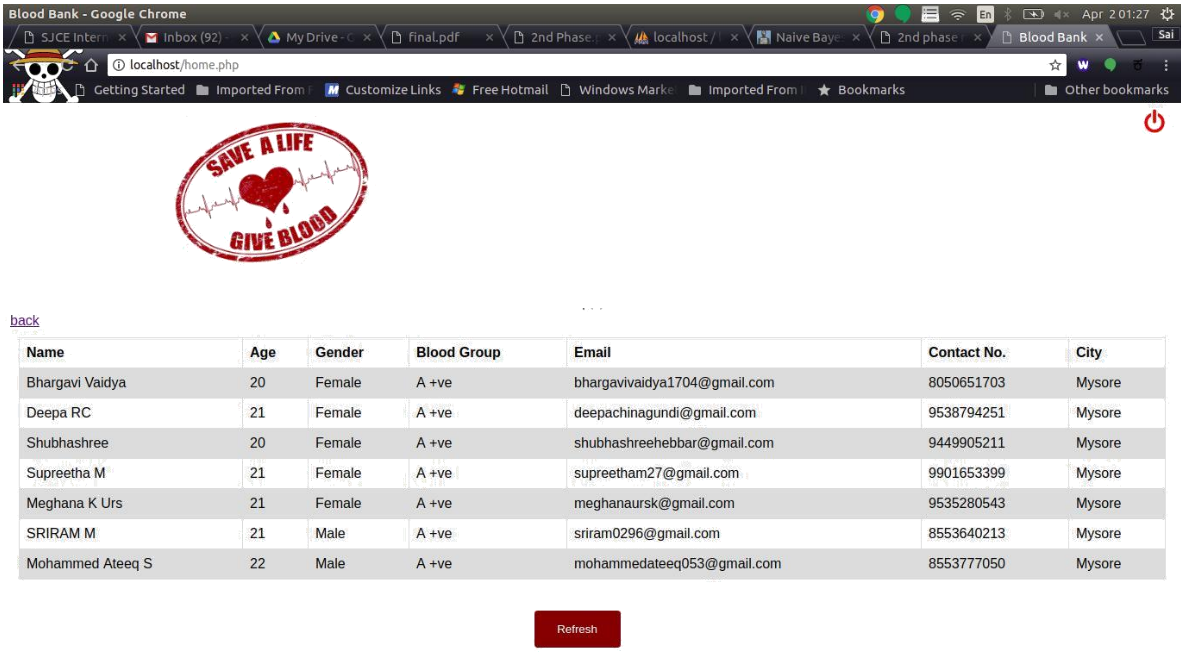
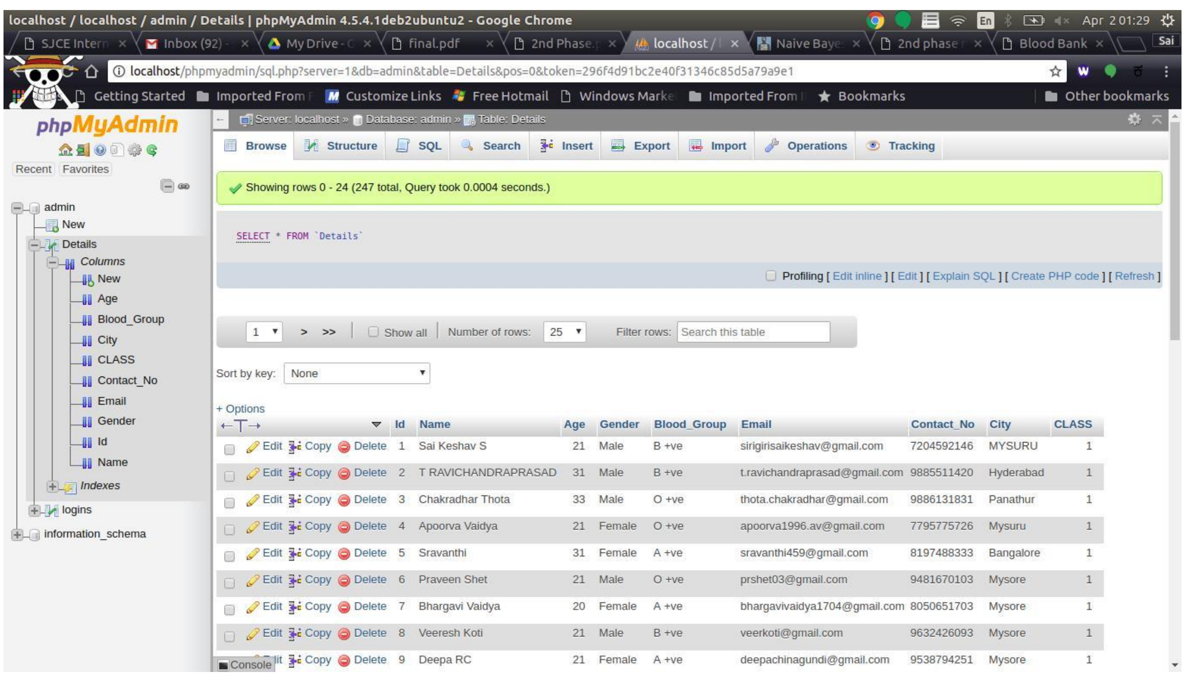


Fig 30: Home Page



Name	Age	Gender	Blood Group	Email	Contact No.	City
Bhargavi Vaidya	20	Female	A +ve	bhargavivaidya1704@gmail.com	8050651703	Mysore
Deepa RC	21	Female	A +ve	deepachinagundi@gmail.com	9538794251	Mysore
Shubhashree	20	Female	A +ve	shubhashreehebbar@gmail.com	9449905211	Mysore
Supreetha M	21	Female	A +ve	supreetham27@gmail.com	9901653399	Mysore
Meghana K Urs	21	Female	A +ve	meghanaursk@gmail.com	9535280543	Mysore
SRIRAM M	21	Male	A +ve	sriram0296@gmail.com	8553640213	Mysore
Mohammed Ateeq S	22	Male	A +ve	mohammedateeq053@gmail.com	8553777050	Mysore

Fig 31: Sample Results



Id	Name	Age	Gender	Blood_Group	Email	Contact_No	City	CLASS
1	Sai Keshav S	21	Male	B +ve	sirigirisaikeshav@gmail.com	7204592146	MYSURU	1
2	T RAVICHANDRAPRASAD	31	Male	B +ve	t.ravichandraprasad@gmail.com	9885511420	Hyderabad	1
3	Chakradhar Thota	33	Male	O +ve	thota.chakradhar@gmail.com	9886131831	Panathur	1
4	Apoorva Vaidya	21	Female	O +ve	apoorva1996.av@gmail.com	7795775726	Mysuru	1
5	Sravanthi	31	Female	A +ve	sravanthi459@gmail.com	8197488333	Bangalore	1
6	Praveen Shet	21	Male	O +ve	prshet03@gmail.com	9481670103	Mysore	1
7	Bhargavi Vaidya	20	Female	A +ve	bhargavivaidya1704@gmail.com	8050651703	Mysore	1
8	Veeresh Koti	21	Male	B +ve	veerkoti@gmail.com	9632426093	Mysore	1
9	Deepa RC	21	Female	A +ve	deepachinagundi@gmail.com	9538794251	Mysore	1

Fig 32: Database

7.3 SYSTEM TESTING

7.3.1: White-box testing:

It is a method of testing software that tests internal structures or workings of an application, as opposed to its functionality (i.e. black-box testing).

Test Scenario 1: Enter proper credentials, login as Admin and check for successful login.

Test Case 1:

Step #	Description	Input	Expected result	Actual result	Status
1	Open the application	N/A	‘Blood Donor Prediction’ login page must be displayed	‘Blood Donor Prediction’ login page is displayed	Pass
2	Enter user Id and password, Click on login	user Id: admin Password: pswd	Home page must be displayed	Home page is displayed	Pass

Test Scenario 2: Enter improper credentials, login as admin and check for unsuccessful login.

Test Case 2:

Step#	Description	Input	Expected result	Actual result	Status
1	Open the application	N/A	‘Blood Donor Prediction’ login page must be displayed	‘Blood Donor Prediction’ login page is displayed	Pass
2	Select admin, Enter improper user Id and password, Click on login	user Id: 20 Password: 123	Login form is displayed with an error message ‘Your Login	Login form is displayed with an error message ‘Your Login	Pass

			Name or Password is invalid. Retry!!!’	Name or Password is invalid. Retry!!!’	

7.3.2: Black box testing :

It is a method of software testing that examines the functionality of an application without peering into its internal structures or workings. This method of test can be applied virtually to every level of software testing.
Admin, when opens the application gets to the ‘Login page’, the admin puts his user ID and password. And we he gives correct credentials he is sent to the home page,where he gives city and blood group where the blood is required. Then system displays all the willing donors of that particular blood group is that city.

CHAPTER 8: CONCLUSION AND FUTURE WORK

8.1 CONCLUSION

Applying information of consumer survey such as behaviors, feelings and opinions of the donors in blood donation can enhance the analysis of the feasibility of blood donation of each individual. Five factors influencing blood donor behaviors suggested by Mostafa [2] are used as a framework to construct a model to classify donors. Those five factors include altruistic values, knowledge in blood donation, perceived risks, attitudes towards blood donation, and intention to donate blood Online questionnaire is used to conduct data survey from different people i.e. people of SJCE, friends, family members, acquaintances etc. A sample size of 250 responses has yielded a reliability scale in an acceptable significant level (greater than 0.65 of Cronbach's Alpha). Although the sample size chosen in this study is only in the provincial level, the result shows that the ANN model has a relatively high accuracy value. This is an indication that the model is able to learn the pattern of blood donors from questionnaire with satisfactory results. Moreover, this study can be used as a prototype and expand the sample group in order to develop blood donor classification model in both the regional and the national levels, which will be beneficial for developing blood donor database system in different levels. This database system is advantageous for classifying potential blood donors and tracking them to donate blood again in the future. The obtained information can be used to determine potential blood donors and manage blood on the supply chain more effectively. Furthermore, this classification model and donor database system will contribute greatly for blood acquisition especially when there are emergency needs for blood for uses in the life saving treatments.

8.2 FUTURE WORK

Other factors affecting blood donation should be explored in order to better manage blood donors' behaviors. The work can be extended by enlarging the sample size of the study. Moreover, other machine learning techniques can be used comprehensively in analyzing the model and comparing the results. The person willing to donate blood will be tracked by his phone number through GPS. People not willing to donate blood can be made to donate through motivational methods.

Appendix A: Project Team Details

Project Title	*Bio Medical Application on Predicting Blood Donors Using Machine Learning Techniques*		
USN	**Team Members Name & CGPA**	**E-Mail Id**	**Mobile Number**
4JC14CS001	ABHISHEK SAJJAN 8.31	aasajjan44@gmail.com	7204132625
4JC14CS085	SAI KESHAV S 7.57	sirigirisaikeshav@gmail.com	7204592146
4JC15CS416	SANGANNA HALLAD 6.99	sanguhallad371@gmail.com	8861824493
-	Dr. M.P PUSHPALATHA (GUIDE)	mppvin@gmail.com	9141410120

Appendix B: COs, POs and PSOs Mapping for the Project Work (CS84P)

Course Outcomes:

CO1: Formulate the problem definition, conduct literature review and apply requirements analysis.

CO2: Develop and implement algorithms for solving the problem formulated.

CO3: Comprehend, present and defend the results of exhaustive testing and explain the major findings.

Program Outcomes:

PO1: Apply knowledge of computing, mathematics, science, and foundational engineering concepts to solve the computer engineering problems.

PO2: Identify, formulate and analyze complex engineering problems.

PO3: Plan, implement and evaluate a computer-based system to meet desired societal needs such as economic, environmental, political, healthcare and safety within realistic constraints.

PO4: Incorporate research methods to design and conduct experiments to investigate real-time problems, to analyze, interpret and provide feasible conclusion.

PO5: Propose innovative ideas and solutions using modern tools.

PO6: Apply computing knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to professional engineering practice.

PO7: Analyze the local and global impact of computing on individuals and organizations for sustainable development.

PO8: Adopt ethical principles and uphold the responsibilities and norms of computer engineering practice.

PO9: Work effectively as an individual and as a member or leader in diverse teams and in multidisciplinary domains.

PO10: Effectively communicate and comprehend.

PO11: Demonstrate and apply engineering knowledge and management principles to manage projects in multidisciplinary environments.

PO12: Recognize contemporary issues and adapt to technological changes for lifelong learning.

Program Specific Outcomes:

PSO1: Problem Solving Skills: Ability to apply standard practices and mathematical methodologies to solve computational tasks, model real world problems in the areas of database systems, system software, web technologies and Networking solutions with an appropriate knowledge of Data structures and Algorithms.

PSO2: Knowledge of Computer Systems: An understanding of the structure and working of the computer systems with performance study of various computing architectures.

PSO3: Successful Career and Entrepreneurship: The ability to get acquaintance with the state of the art software technologies leading to entrepreneurship and higher studies.

PSO4: Computing and Research Ability: Ability to use knowledge in various domains to identify research gaps and to provide solution to new ideas leading to innovations.

Write justification for the mapping. (Write around 4 to 5 paragraphs)

Note:

1. Scale 1 – Low relevance
2. Scale 2 – Medium relevance
3. Scale 3 – High relevance

SEM	SUBJECT	CODE	CO	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO1	PSO2	PSO3	PSO4
VIII	Project Work	CS84P	CO1	3	3	2	3	2	2	2	2	3	3	2	2	3	3	2	3
			CO2	3	3	2	3	2	2	2	2	3	3	2	2	3	3	2	2
			CO3	3	2	2	3	3	3	3	3	3	2	2	2	3	3	2	2

JUSTIFICATION FOR CO AND PO MAPPING:

CO1:

Our project is software based application and mainly deals with Prediction of the Blood donor through machine learning. There are currently many researches going on to address this issue. So we have come across a variety of existing solutions to better understand the problem in depth and drawback of current solutions. To develop efficient approaches applying knowledge of computer, complex engineering problems and proper planning is must so we have rated 3 for PO1, and PO2. We gave PO4 as 3 because as our project uses two important machine learning algorithms, it is possible to conduct experiments to investigate real-time problems. We have used the existing applications such as phpMyAdmin technology so we have rated PO5 and PO6 as 2. We haven't used any patented algorithms or confidential data so we rank PO8 as 2. For PO9 and PO10, as we work on modules separately we assigned 3. For PO11 and PO12 we have rated 2, since we have to employ the knowledge gained in many of the subjects we learned during engineering such as Programming, Web technology, Data mining and Pattern recognition. For PSO1, PS02, PSO4 we assigned 3 because we come across the core subjects such as DBMS, algorithms and data mining in order to solve the problem.

CO2:

To develop and implement algorithm basic concept of computer engineering is must. As our project based on machine learning we have designed in such a way that it should be economic, reliable and usable for real world systems etc. so we rated PO1, PO2, PO4 as 3 and PO3 as 2. We designed the system by taking help of existing methods so we assigned 2 for PO5, PO6 and PO8. Our project has several modules corresponding to the different roles that exist in our problem system, so for PO9 and PO10, we assigned 3. For PO11 and PO12 we assigned 2, because we have used many subjects we learnt during the engineering such as DBMS, Data mining, web development, Pattern Recognition. All core subjects of engineering is necessary to implement the project so we rated PSO1 and PS02 as 3.

CO3:

This project work shows that the new primitive is cost effective and predicts the donor efficiently in most of the cases due to the improved machine learning algorithms. So we rated PO1 as 3. We have done several testing and trials of training to ensure that application is working properly and gives better classification performance so it should be available all the time and also efficiently predicts the donor, so we have assigned 2 for PO2 and PO3. We gave 3 for PO4, PO5 and PO6 because we have real world based application which is tested in real time environments. Also we concentrated more on confidentiality of user data and on the efficiency of the algorithms in predicting the donor as it is the main goal of our project, so we rated PO7 and PO8 as 3. For PO11 and PO12 we assigned 2, because we have used core subjects like DBMS, Data mining, programming etc. and PS01 is assigned 3 because this project applies mathematical methodologies to solve computational tasks, model real world problems with an appropriate knowledge of Data mining, pattern recognition and machine learning algorithms etc.

REFERENCES

1. Arunkumar Chinnaswamy, 2015 “A study on automation of blood donor classification and notification techniques”, 2015 in International Journal Of Applied Engineering Research.

2. Deepti Bahel, Prerna Ghosh, Arundhyoti Sarkar, Matthew A. Lanham, “ Predicting blood donations using machine learning techniques ”

3. https://medlineplus.gov/bloodtransfusionanddonation.html

4. https://www.ncbi.nlm.nih.gov/pubmed/19040598

5. http://people.umass.edu/aizen/tpb.html

6. https://www.czp.cuni.cz/czp/images/stories/Vystupy/Seminare/2010%20Theory%20of%20Planned%20Behavior/Prezentace/Lukacovska_Motivation%20towards%20blood%20donation.pdf

7. Icek Ajzen, “ Behavioral Intentions Based on the Theory of Planned Behavior ”

8. Icek Ajzen, “ Sample Theory of Planned Behavior Questionnaire ”

9. Imran Ahmed Khan, 2013 “Knowledge extraction from Survey data using Neural Networks”, 2013, Scholar works at UT Tyler.

10. Judith Holdershaw, Philip Gendall and Malcolm Wright, “ Predicting Willingness to Donate Blood ”

11. Kathi L. Rose, “ The Blood Donor of the Future : Demographic Predictions ”

12. Mehrdad Jalalian, Latiffah Latiff, Syed Tajuddin Syed Hassan, Parichehr Hanachi and Mohamed Othman, May 2010 “ Development of a Questionnaire for assessing factors predicting Blood Donation among University Students : A pilot study ”

13. Mustafa .M, 2009 “Profiling blood donors in Egypt: a neural network analysis”, Expert systems with applications.
14. Nur Zainie Abd Hamid, Rohaida Basiruddin and Narehan Hassan, July 2013 “ Factors influencing the Intention to Donate Blood : The Application of the Theory of Planned Behavior ” , July 2013, International Journal of Social Science and Humanity.

15. W. A. Flegel, W. Besenfelder and F.F. Wagner, “ Predicting a donor’s likelihood of donating blood ”

16. Wijai Boonyanusith and Phongchai Jittamai, 2012 “Blood donor classification using neural networks and decision tree techniques”, 2012, IAENG.