LOVELY PROFESSIONAL UNIVERSITY

Strategize Framework for Contesting Plan with Known Variable Learning Technique

A Dissertation

submitted

By

Jaskirat Singh

(11307712)

To

Department of Computer Science

In Partial fulfilment of the Requirement for the

Award of the Degree of

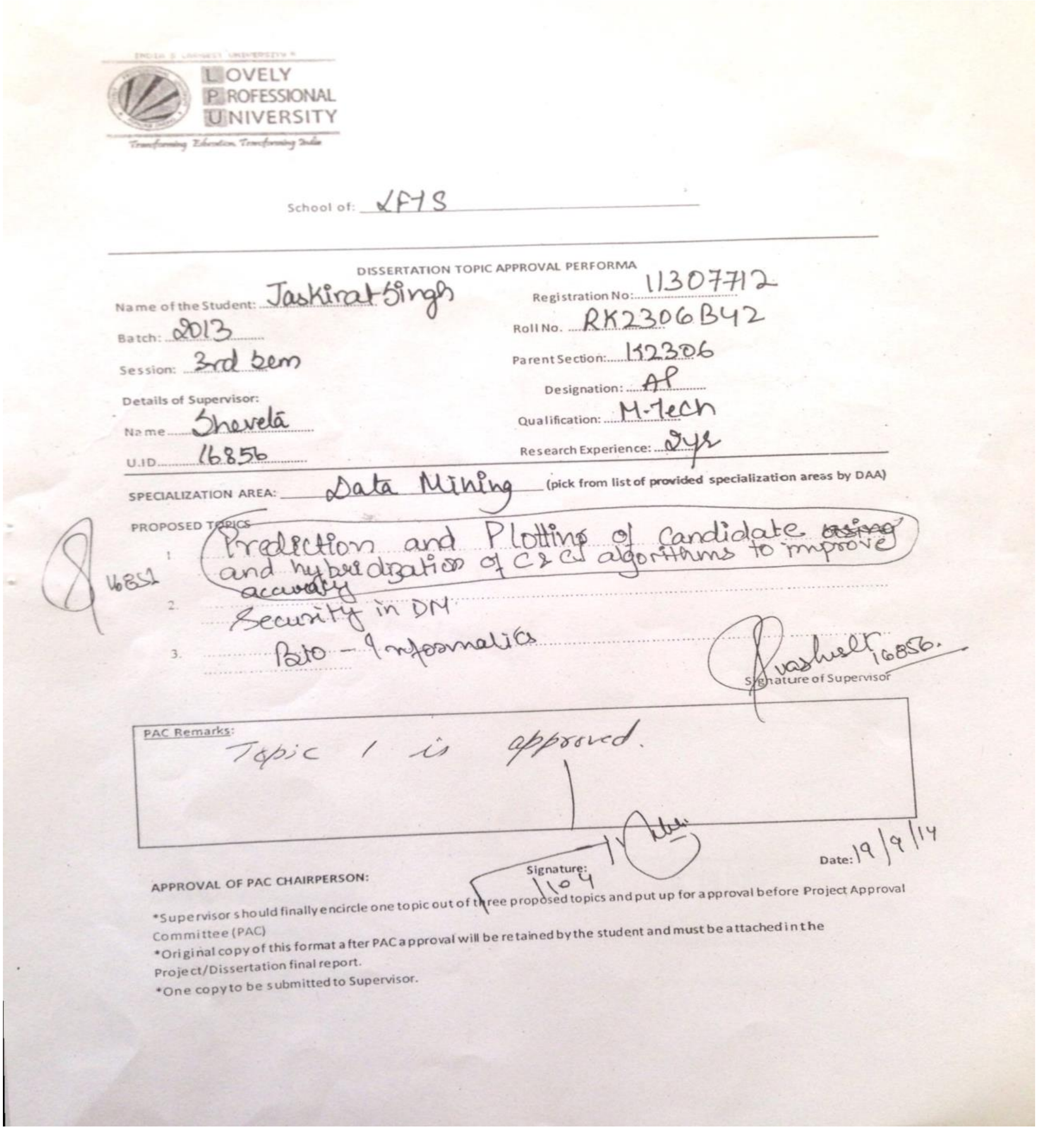
Master of Technology in Computer Science

Under the guidance of

Sheveta Vashisht

(May 2015)

Approval of PAC



INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education Transforming India

School of: LFTS

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Jaskirat Singh

Registration No: 11307712

Batch: 2013

Roll No. RK2306B42

Session: 3rd sem

Parent Section: K2306

Details of Supervisor:

Designation: AP

Name Shevetā

Qualification: M-Tech

U.ID 16856

Research Experience: 2yr

SPECIALIZATION AREA: Data Mining (pick from list of provided specialization areas by DAA)

PROPOSED TOPICS

16851

1. Prediction and Plotting of candidate ~~using~~ and hybridization of C & C4 algorithms to improve accuracy

2. Security in DM

3. Bio - Informatics

vashist 16856.

Signature of Supervisor

PAC Remarks:

Topic 1 is approved.

APPROVAL OF PAC CHAIRPERSON:

Signature: 1104

Date: 19/9/14

*Supervisor should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.

*One copy to be submitted to Supervisor.

ABSTRACT

Data mining has gained great attention in the information industry and in society due to the existence of large amounts of data and the crucial need for changing such data into useful information and knowledge. This paper will discuss use of data mining process to help out the election contesting party to decide the candidates and strategize on contesting plan as per the results will be out of the system. From large amount of voter data useful information will be extracted and used to take effective decisions. All the castes will be listed while selecting the data for voter caste and when it's time to display results, the castes can be aggregated to General, SC, BC and OBC etc. As the state has been divided into many levels, the data can be viewed at any level according to the access rights and level. The results can be segregated to the level of a booth as per selection. Visualizations will be an important part of the system. As the voter data for the state will be in crores, the visualization of the data will be an important part .Data can be displayed in form of Line Charts & Pie Charts and Graphs .It will be easily understandable to the users. The results will be viewed according to the parameters. Bayesian Classification is applied on the parameters of research. After getting results by Bayesian Classification we then apply clustering algorithm k –means .Then we will get two clusters .One cluster containing values to be processed further and values in the other cluster will be discarded.

ACKNOWLEDGEMENT

First I offer my sincerest gratitude to my supervisor, Sheveta Vashisht, who has supported me throughout my thesis .Without her this thesis would not have been completed or written. I am thankful for her aspiring guidance, invaluably constructive criticism and friendly advice during the work am sincerely grateful to her for sharing their truthful and illuminating views on a number of issues related to the research. Finally, I thank my parents for supporting me throughout all my studies at University.

I would also like to thanks my family and friends who have been a source of encouragement and inspiration throughout the duration of the research.

DECLARATION

I hereby declare that the dissertation entitled, "**Strategize Framework for Contesting Plan with Known Variable Learning Technique**." submitted for M-Tech degree is entirely my original work and all ideas and references have been duly acknowledged .It does not contain any work for award of any other degree or diploma.

Date:

Jaskirat Singh Sandhu
11307712

CERTIFICATE

This is to certify that Jaskirat Singh has completed M.TECH dissertation titled **“Strategize Framework for Contesting Plan with Known Variable Learning Technique”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original investigation and study. No part of the dissertation has ever been submitted for any other degree or diploma. The dissertation is fit for the submission and the partial fulfillment of the conditions for the award of M-Tech Computer Science & Engineering.

Date:

Sheveta Vashisht
(16856)

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

With the rapid development of information technology endless amount of data and information is available and more emphasis is paid on its analysis .Now world is flooded with data i.e. Big Data. Data can be comprised of videos, images etc. We analyze this data to get useful information which further helps in taking good decisions. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, science exploration, medical fields, real estate, and web mining [2]. Globalization increases the development of information world, high capacity data appear everywhere and wit this there is a different prospective for everyone to do the analysis and summarize the data into useful information and this analysis and summarization is not easy task to do, to obtain the optimized result the concept of data mining is used. Data mining is extraction of hidden predictive information from large databases and using its various techniques the hidden pattern from large amount of data can be discovered and it can forecast the behavior and future trends to support the people, organization or companies to take decisions.

As datasets have grown in size and complexity, direct hands on data analysis has increasingly been augmented with indirect, automatic data processing. Data mining now has great attention in the information industry and society due to existence of large amount of data and crucial need for changing such data into useful information and knowledge which can be used further and this interest has inspired a rapidly maturing research field with developments both on a theoretical as well as on practical level with the help of a range of commercial tools As we data mining is process of applying various methods such as clustering, decision tree etc. to data with the intention of uncovering hidden patterns. A primary reason for using data mining is assist in the analysis of collections of observations of behavior.

We can define Data Mining as extracting knowledge from data. In other words mining or extracting useful information from huge amount of data (data warehouse) in visual forms like pie charts, graphs etc. Other terms that emulates or different meaning to data mining, such as mining knowledge from data, knowledge acquisition, pattern analysis, data archaeology, and data dredging. Term that emulates data mining is KDD (Knowledge Discovery Process) [2].

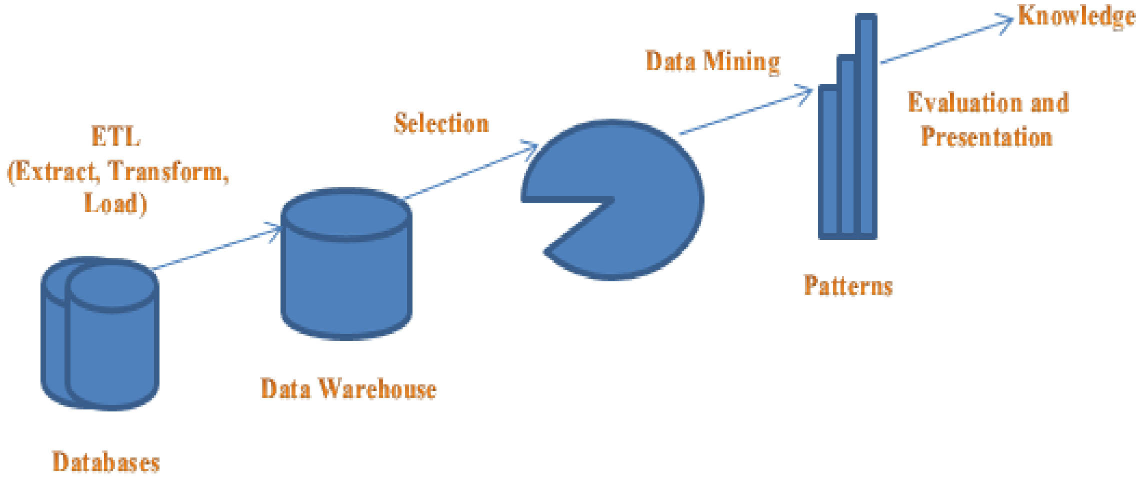


Figure 1.1: Data Mining Process

Data from different sources like operational databases, files etc. is extracted ,transformed and loaded (ETL) into database called data warehouse .Now data is clean and can be used for analysis purpose . Knowledge is extracted from the data and which is then shown to user in way it is easily understandable i.e. in graphical forms. Data mining now has great attention in the information industry and society due to existence of large amount of data and crucial need for changing such data into useful information and knowledge which can be used further for analysis. Given below explain every step with step name.

Data mining steps are:

I. **Data Cleaning**: In this we remove noise, inconsistent, incomplete data. There may be duplicate values, missing values etc. which are removed in this.

II. **Data Integration**: In this we integrate data from multiple data sources. It can operational databases, files, disks etc.

III. **Data Selection**: We select relevant data from the database for analysis.

IV. **Data Transformation**: Data are transformed into appropriate or standard forms for mining.

V. **Data Mining:** In this intelligent methods are applied to extract data patterns.

VI. **Pattern Evaluation**: In this we identify interesting patterns representing knowledge.

VII. **Knowledge Presentation**: In this extracted knowledge is presented in visual forms to user i.e. in pie charts, graphs etc.

1.1Classification

Classification is one of the most abrupt used technique by data mining and machine learning (ML) researchers. There are different classification methods like Bayesian Classification etc. In Classification classifier is framed to predict categorical labels. Classification method makes use of mathematical techniques such as decision trees, neural network, linear programming and statistics. In classification, software can be developing which can learn how to classify the data items into groups. For e.g. a marketing manager of a company want to analyze whether a customer will buy a new computer .Here labels are yes and No. In above example Prediction can be how much a customer will spend during sale [2]. Output is a numeric value. Supervised learning has a classifier that is well known and it has sufficient categorical labels such as yes and no for above example. For unsupervised learning either the classifier is unknown or appeared for small number of cases. Classification can be done by using decision tree .Decision tree is used for attaining information for decision making purpose. Internal nodes are represented by rectangles and leaf nodes by ovals. Internal nodes can have two or more child nodes [2].Final result is tree which represents the decision and its outcome. Decision tree algorithms are ID3, C4.5 etc. ID3 uses Information gain for attribute selection measure.

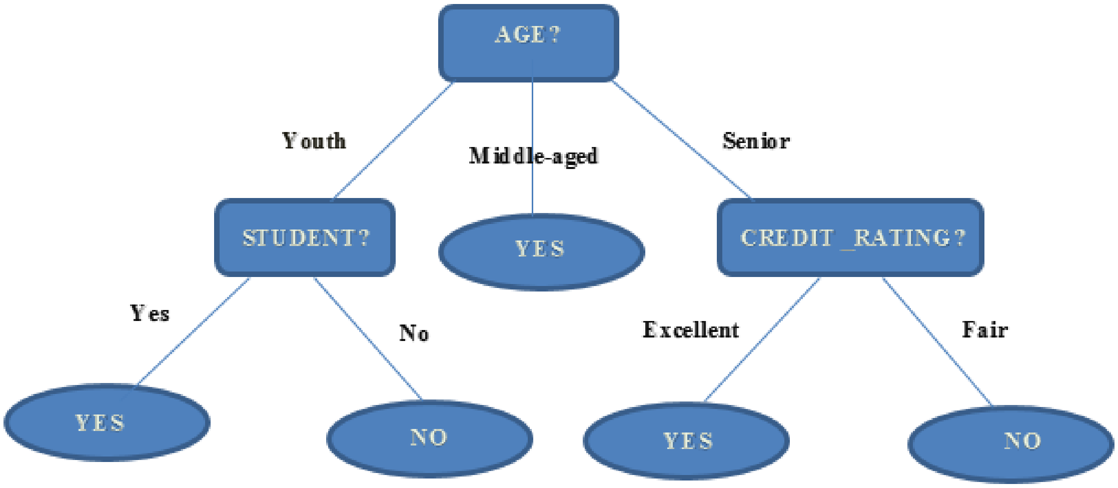
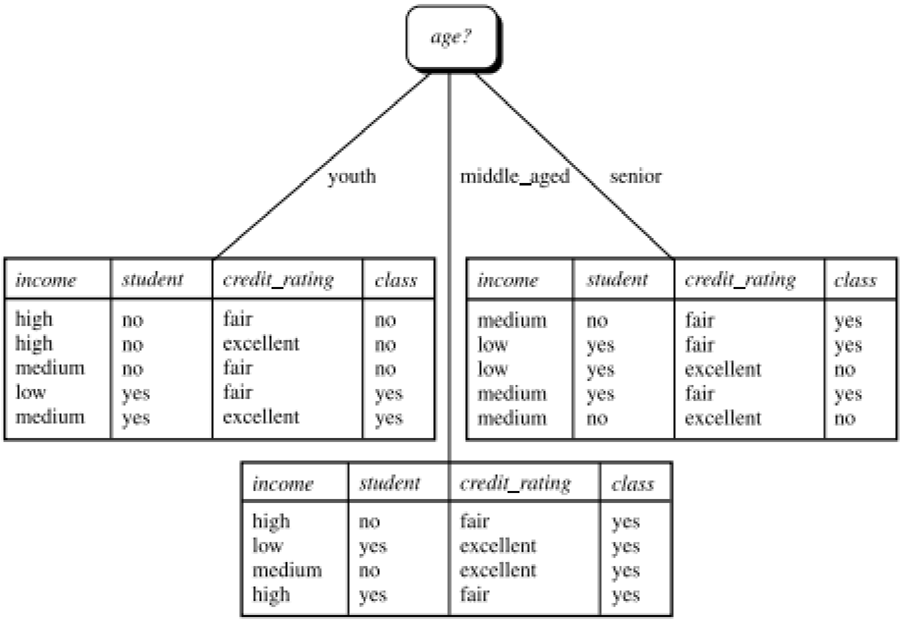


Figure 1.2: Decision Tree

In above figure age is taken as classifier .Why only age is chosen? There is attribute selection measures .ID3 uses Information gain for attribute selection measure. It is used to select the root i.e. the node from where splitting begins. Formula to calculate the information gain is given below.

$$\text{Info (D)} = -\sum_{i=1}^{m} \text{pi} \log2(\text{pi})$$

We calculate the information gain of every attribute in data set .The attribute having highest information gain is chosen as splitting attribute. Figure below showing the splitting attribute.Attribute having highest information gain become the root of the tree.



income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

Figure 1.3: Splitting attribute [2]

Another classification methods Bayesian classification that uses concept of Bayes theorem.

$$P\ (H|X) = \frac{P(X|H)\ P(H)}{P(X)}$$

Rule based classification which is represented in form of IF-THEN rules.

Classification and Prediction Issues

Before preparing the data for classification and prediction involves the following activities:

Data Cleaning: Data cleaning includes eradication of noise and filling of missing values. We can get rid of noise in data by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value or with simply NULL.

Relevance Analysis: Database can have the irrelevant attributes means redundant attributes. Correlation analysis is used to know whether any two given attributes are related with each other. Then one of the attribute is removed.

1.2Clustering

Clustering is a technique in which objects of similar category is placed in one group and other are in different group. Objects having similar properties into one cluster and others into different group. It is an unsupervised learning technique. So there are no labels as in classification (yes or no for example in classification mentioned above in 1.1). Cluster analysis can be used in the areas such as image processing, analysis of data, market research (buying patterns) etc. Using clustering we can do outlier detection where outliers are values lying outside the cluster. We can see the outliers in below figure. Dots lying outside the clusters are the outliers.

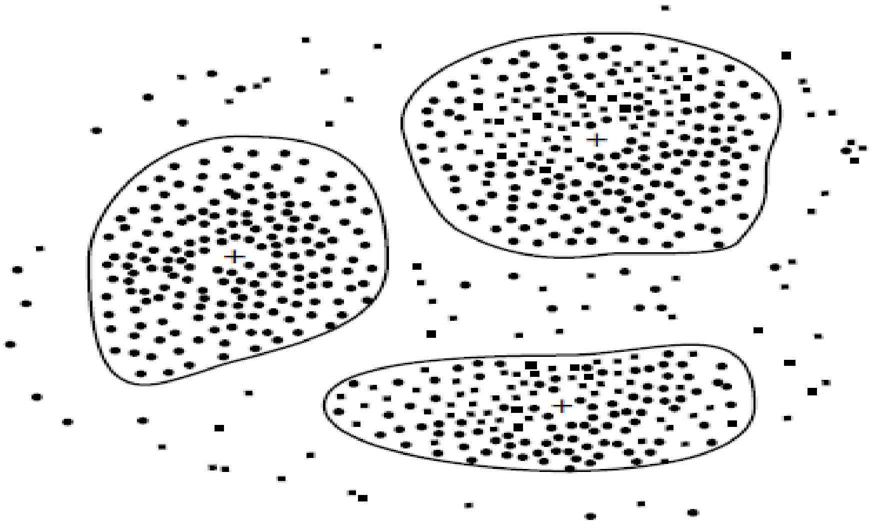
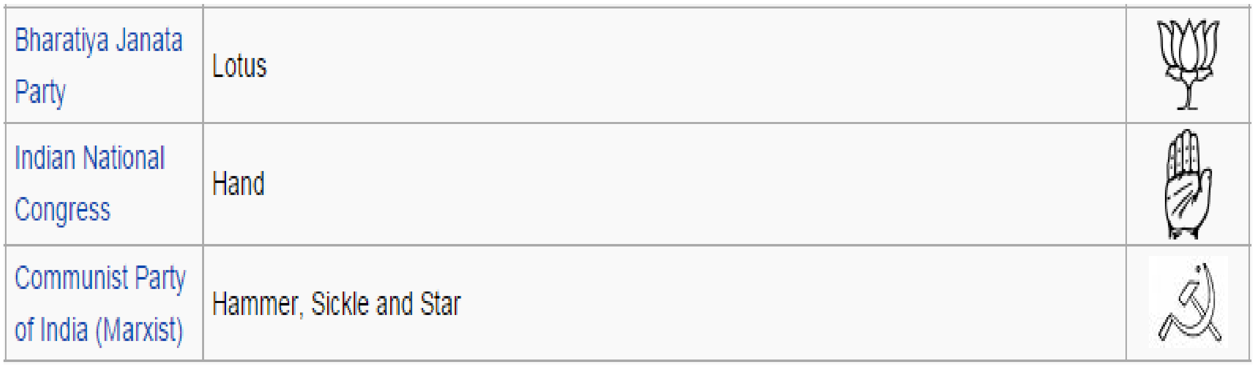


Figure 1.4: Clustering [2]

1.3Political Parties

India is called largest democracy. India contains a multi-party system that mean multiple parties can form government like BJP and SAD (Shiromani Akali Dal) each running government in Punjab, with recognition accorded to national and state level parties [14]. The status is reviewed periodically by the Election Commission of India. Since the last three decades the two dominant national parties (INC and BJP). Political parties can be local, state or national and should be registered by the Election Commission of India (ECI). The ECI (Election Commission of India) is an independent, established authority accountable for administering all the processes associated with elections in India. The Election Commission of India do oversight of elections in order that election may be honest and free. The necessary feature of democratic rule is elections at regular intervals, thus free and honest elections periodically are the essentials of democratic nation. The ECI is seen as guardian of free and honest elections The Election Commission of India categorizes parties as National, State, and Unregistered parties. In India total number of registered parties are 1766 from which 3 are National Parties, 57 State parties and rest are unrecognized parties.



Bharatiya Janata Party	Lotus	
Indian National Congress	Hand	
Communist Party of India (Marxist)	Hammer, Sickle and Star	

Figure 1.5: Political Parties with symbols [15]

Every Party has its symbol like Lotus of Bharatiya Janata Party as shown in Figure 1.7 .With the help of symbols party ca easily be identified during time of voting.

Every party frame plans, strategies before election with motive to win the election. For this they first generate list of people who will contest election with their constituency names. This work can be done by analyzing the data. This paper will discuss the use of data mining process to help out the election contesting party to decide the candidates and strategize on the contesting plan as per result will be out of the system. Party high rank people can do candidate plotting very easily and efficiently. Analyzing data can help the election contesting party to decide the candidates and strategize on contesting plan as per the results will be out of the system. If number of votes of women is more as compared to men then contestant can be female such prediction can be made. If more percentage of OBC caste people so contestant can be from OBC caste .So this how we can do candidate plotting and frame strategies .Visualizations will be an important part of the system. As the voter data for the state will be in crores, the visualization of the data will be an important part. The data can be displayed in form of charts and Graphs. As Fig 1.8 shows pie chart showing the party wise results of Delhi Assembly Elections 2015.

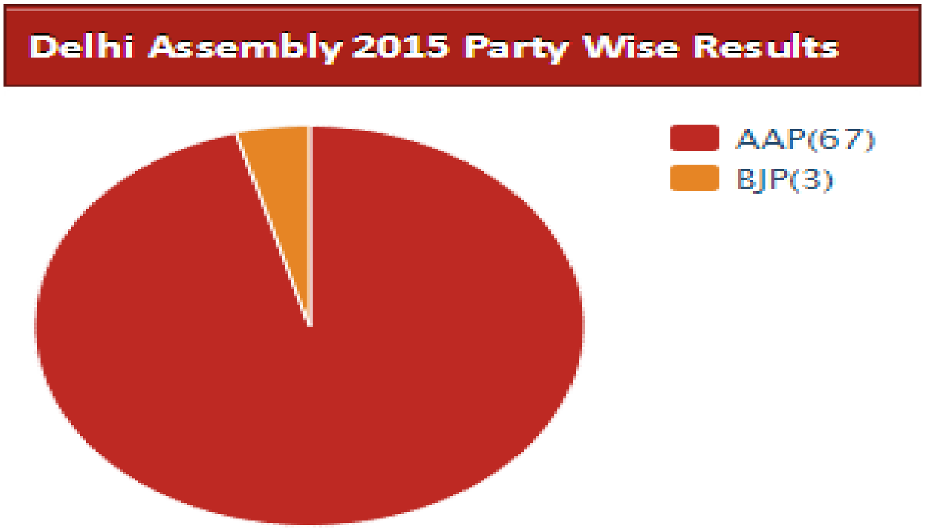


Figure 1.6: Delhi Elections and Results 2015 [16]

CHAPTER 2
REVIEW OF LITERATURE

Vasile Paul Bresfelean (2007) used classification technique to predict whether students will continue their studies after graduation [1].He used J48 algorithm on data received from surveys to predict whether students will continue their studies after graduation (M-Tech, Ph.D.).Data was collected from the final year students at Faculty of Economics and Business Administration in Cluj-Napoca. To gather the data he used on line and written surveys .Weka data mining tool was used to extract the data.

Brijesh Kumar Bhardwaj *et al* (2011) used classification technique to predict the improvement of performance of students [3].For this he used Bayesian Classification technique .Data set was collected from different colleges on sampling method of Computer Applications department of course BCA of session 2009-10.Variables or we can say parameters were selected such as Gender , Category (obc,general etc.),student food habit ,student other habit(drinking ,smoking),living location (village ,city) etc.

S.R.Pande *et al* (2012) provides the data mining techniques of clustering. Cluster analysis divides data into the groups having similar properties [4]. Clustering is unsupervised classification technique. Clustering is divided into two classes, first is hierarchical clustering techniques and other is partitioning technique. Partitioning clustering techniques include K-means, K-mediods, and CLARA etc. The hierarchical method forms tree like structure. It includes agglomerative and divisive technique. They also density based methods like DBSCAN, DENCLUE. In this paper they process of clustering from the point of view of the data mining.

Mahendra Pratap Yadav *et al* (2012) explain relationship between data mining and e-commerce with the continuously increasing growth of data in World Wide Web is discussed. The user wants to extract desirable information and resources. The main idea of this research is to find the behavior of customer that what they want or what are their requirements. For e-commerce conventional methods are no longer useful to find customers behavior [5]. With the advanced technologies, large amount of data is stored in servers about thousands number of customers profiles and from they can search the data about customers' requirements. K-Means algorithm in cluster customer is used for mining the input data coming from various e-commerce websites. To increase customer's behavior in online shopping strategy of attracting customers with good offers and combos is done by seeing their profiles. Age, gender and behavior are main attributes for analyzing the customers marketing in e-commerce.

Neelamadhab Padhy *et al* (2012) gave an overview of data mining and areas where it can be used .They told data mining can be used to extract information from very large amount of data .They mentioned the data mining techniques : Decision tree and rules ,classification methods and nonlinear regression etc. [6] .They told areas where data mining can be done to get information which can be used for making decisions .Areas are Healthcare ,Education Systems ,CRM ,Web Education ,Sports data mining ,E-Commerce etc. The various data mining techniques are used to extract the useful patterns.

Kamal Bunkar *et al* (2012) uses data mining process to help in improving the performance of graduate students. They used classification for prediction .They used the student data of Vikram University, Ujjain of course B.A first year students [7]. Classification methods used were :Decision Tree , ID3 (Iterative Dichotomiser 3) ,C4.5 .They used the examination data of the students and stored in MS Excel .Result indicated was if prediction is student tends to fail the examination before the examination then students need to work hard to score passing marks in the courses .

S.Anupama Kumar *et al* (2012) uses data mining process to predict the result of final semester of student based on marks got in previous semester. They predict the result of fifth semester based on the marks obtained in previous four semester .Rule based Classification techniques were used in prediction process [8] .Classification techniques they used are Decision table and One R rule algorithms .The accuracy of the algorithm used by them was analyzed by comparing the prediction of miner and actual result obtained by students in the fifth semester .Data set was taken from master degree course .They considered the student ID , their total number of marks and their result obtained in previous four semester .This paper shows how efficiently Rule Based algorithm can be used in predicting student's result in higher semester using previous data .

Y. Ramamohan *et al* (2012) gave an overview on data mining tools for KDD i.e. Knowledge Discovery Process .Data mining tools are used to predict future trends and behaviors, which allows to take effective business decisions to increase productivity. Data mining tools can answer business questions that traditionally were too time consuming to resolve [9]. This paper gives a brief introduction about data mining tools like Weka, Witness Miner Tool, Tanagra, Rapid Miner, and Orange. Small overview of these tools are given in this paper .The Weka tool provides data mining tasks such as data preprocessing, feature selection, classification ,clustering and visualization .

Dr Tariq Mahmood *et al* (2013) uses data mining on the Twitter Big Data to predict 2013 Pakistan Election Winner .Large amount of data is generated on the twitter daily , to do prediction data of relevant twitter users were taken , pre-processing (replacing double and triple spaces with single space ,removed all URLs etc. from the data) of data and then framed predictive models for three political parties PTI (Pakistan Tahreek-e-Insaaf) , PMLN (Pakistan Muslim League Nawaz) and MQM (Muttahida Qaumi Movement). Data was collected from the website called Twimemachine .Prediction labels were Pro and Anti .Pro means favoring the party and Anti represents against .For this Rapid Miner Tool was used with three predictive models CHAID decision tree, Naïve Bayes and Support Vector Machine .Accuracies of these three models were compared and no significant difference in accuracies was there .CHAID was selected as it gave better accuracy for PMLN and MQM. Predictions revels that PTI will be the winner of 2013 election but actual winner was PMLN .The study did not include the opinion of rural people of Pakistan which is in large number [10].

Ajay Kumar Pal *et al* (2013) represented how to evaluate and investigate the performance of teacher .The aim was to predict quality ,potential and productivity of teacher based on the student's and organizational feedback ,research activity etc. to understand the faculty's motivation ,growth ,decline .Data was gathered from post graduate studies at department of engineering college over three years of same student with aim of investigating if there is any modification in lectures evaluation [11] . To evaluate the performance four classification algorithms were used based on the Weka .The algorithms used for classification were Naïve Bayes, ID3, CART and LAD. On comparison, Naïve Bayes was the best as it had lowest average error. Survey used numbers for feedback for ex 5 for Strongly Agree for presentation of teacher in the class .Many attributes were tested and some were effective on the prediction of the performance.

Malhar Anjaria *et al* (2014) introduced the new method how public opinion can help to predict the outcome of the election, for this they used twitter data .This paper tried to prove the results for the US Presidential Elections 2012 and state elections in Karnataka 2013 and added sentiment analysis approach. Twitter provides strong and efficient ways of monitoring public opinion about any movie , electoral events etc. [12] .Classifications techniques used are SVM (Support Vector Machines),Naïve Bayes ,Maximum Entropy and Artificial Neural Networks .SVM gave highest prediction accuracy of 88% in case of elections in USA and 58% for Karnataka elections .They used unigram ,bigram and both unigram +bigram for sentiment analysis .Data taken for mining was twitter tweets for the events .

M.Durairaj *et al* (2014) used clustering algorithms for prediction of student performance [13]. The main aim of the research is to make a model using data mining techniques which mines the information, so that current education system can adopt it as a strategic management tool. Weka tool is used for the data mining. Data set contain student name and semester wise marks. Research used naïve Bayes algorithm to analyze the data.

CHAPTER 3
PRESENT WORK

3.1SCOPE OF STUDY

As datasets have grown in size and complexity, direct hands on data analysis has increasingly been augmented with indirect, automatic data processing. Data mining now has great attention in the information industry and society due to existence of large amount of data and crucial need for changing such data into useful information and knowledge which can be used further for taking effective decisions. Data mining is process of applying various methods such as clustering, decision tree etc. to data with the intention of uncovering hidden patterns. Data mining is extraction of hidden predictive information from large databases and using its various techniques the hidden pattern from large amount of data can be discovered and it can forecast the behavior and future trends to support the people, organization or companies to take decisions

Analyzing data can help the election contesting party to decide the candidates and strategize on contesting plan as per the results will be out of the system .Then candidate plotting can be done very easily and effectively . The system will digitize all the booths which will be aggregated to a constituency, constituencies will be aggregated to districts, districts can be aggregated to the MP constituencies and lastly, the whole data can be aggregated at state level. All the castes will be listed while selecting the data for voter caste and when it’s time to display results, the castes can be aggregated to General, SC, ST and OBC etc. As the state has been divided into many levels, the data can be viewed at any level according to the access rights and level. The results can be segregated to the level of a booth as per selection. We can also see this which party is leading from particular booth, which caste is in majority at that booth such things can be seen. Visualizations will be an important part of the system. As the voter data for the state will be in crores, the visualization of the data will be an important part. The data can be displayed in form of Line Charts & Pie Charts and Graphs. It will be easily understandable to the users. The results will depend on the selection of parameters.

Parameters can be gender, occupation, caste, education, age of the voter. If number of votes of women is more as compared to men then contestant can be female such prediction can be made. If more percentage of OBC caste people so contestant can be of OBC caste .So this how we can do candidate plotting and frame strategies .Party preference data of everyone can tell who is in majority in particular area .So there will be viable use of resources in terms of money and human resources. This means if people favor one party most out of three, then other two parties will come to know the public's favorite party and they will plan accordingly .So less emphasis on that constituency can be the decision of less favorite parties.

Output of system will be user friendly and easily understandable .Then crucial decisions can be taken on the basis of the output obtained .Manually processing large data set is time consuming and can produce inaccurate results and also cumbersome task. So using computerized systems data processing will be fast from huge data set as mentioned above crores of voters.

3.2OBJECTIVES

Objectives will give emphasis on the aim of the research work .There will be parameters like gender, occupation, caste, education of the voter. The results will depend on the selection of parameters .If number of votes of women is more as compared to men then contestant can be female (plotting). Results will be in the visual form, so it will be very user friendly and effective, easy and accurate decision can be taken. Party preference data of everyone can tell who is in majority in that particular area. If people favor one party most out of three, then other two parties will come to know the public's favorite party and they will plan accordingly. So less emphasis on that constituency can be the decision of less favorite parties .Results will be in visual form i.e. in form of pie-charts ,histogram etc. Output of system will be user friendly and easily understandable. After getting results then we can take decisions very easily and effectively.

I. To construct an efficient system for framing contesting plan.
II. To take easy and effective decisions after analysing the data.
III. To analyse accurate results and fast processing in mining of data.

All above mentioned objectives are obtained. An efficient system has been constructed that can be used for framing contesting plan. After analyzing the large data easy and effective decisions can be taken. System is showing results in the form of pie chart .From that result one can easily frame contesting plan .For e.g. one pie chart is representing results of Party preference data. It clearly represent which party is in the majority, whether it is one sided (only one party is in large majority) or 2-3 parties having nearly same party preference. So less emphasis on that constituency can be the decision of less favorite parties. If parties having nearly same party preference then there can be need of more campaigning to win the election etc. such decisions can be framed.

There is one result representing caste of voters .If percentage of general caste favoring any party is more, than in that area contestant will be from General caste. Such decisions can be taken after analyzing of data.

3.3RESEARCH METHODOLOGY

3.3.1Research Design

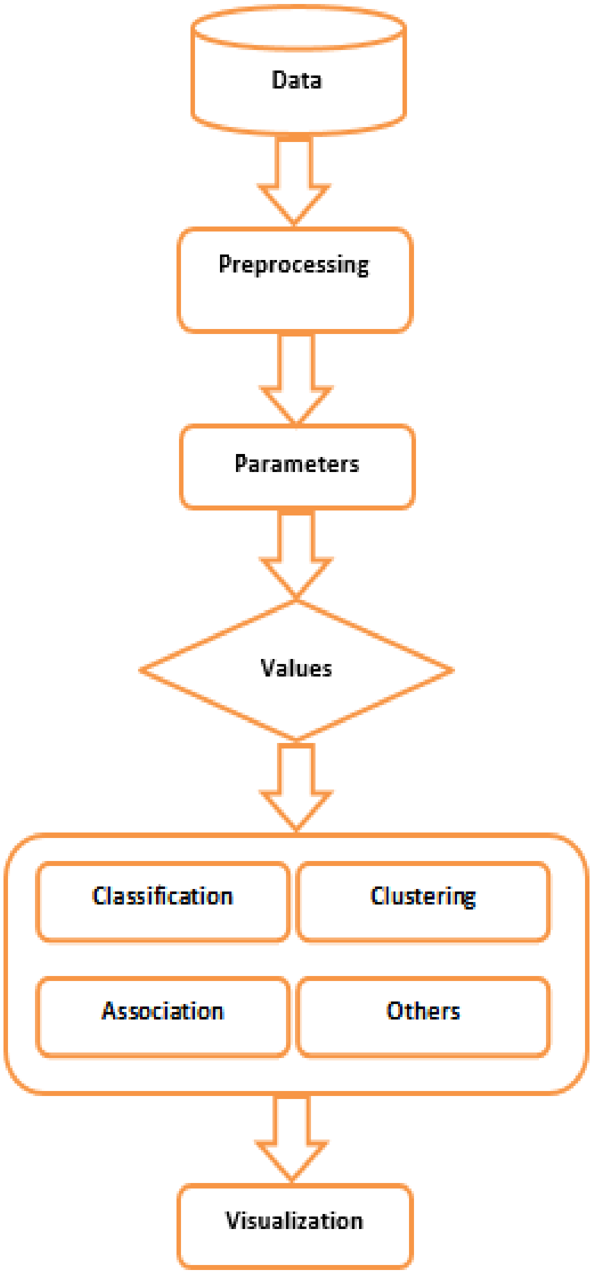


Figure 3.1: System Flow Diagram

Figure 5.1 shows the research design. There is data of voters which is placed in the database. Data is present in the tables in SQL server. Then pre-processing will be applied on the collected data. It includes making data noise free, consistent, and removing null and duplicate values from the data. In this data set missing values were replaced by NA values. NA means no information for that attribute. Then comes parameters and values for these parameters would be extracted from the database. For example party preference values of all parties will be extracted from the database which are supported by any particular caste one by one. Then we apply data mining techniques on the data and we will get two clusters Cluster 1 and Cluster 2 as shown in the Figure 3.1 .Research will not consider the values of Cluster 2 .So only Cluster 1 will used further. At last but not the least, visualization part. Results will be shown using pie charts. Visualizations will be an important part of the system. As the voter data for the state will be in crores, the visualization of the data will be an important part. The data can be displayed in form of Line Charts & Pie Charts and Graphs. It will be easily understandable to the users and help the user in taking effective, easy and efficient decisions. The results will depend on the selection of parameters.

3.3.2Attributes

The main aim of research is candidate plotting .This depends on the parameters .Parameters in this research are Party Preference , Education , Age ,Employment and Caste of voters. All the castes will be listed while selecting the data for voter caste and when it's time to display results, the castes can be aggregated to General, SC, BC and OBC etc.

- **Caste:** Voters are divided into 5 categories: General, OBC, SC, ST and N/A. N/A is assigned to those voters whose caste is not known (for removing missing value). These caste are further divided into sub castes.
  In **General** sub caste names are Jaat, Arora, Baniya, Ahluwalia, Ramgarhia, Rajput etc.
  In **OBC** sub caste names are Aheri, Mirasi, Nai, Julaha, Chimba, Gorkhas, Dhobi etc.
  In **SC** sub caste names are Bangali, Chamar, Mazhabi, Ad Dharmi, Mochi, Mahatam etc.
  In **ST** sub caste names are Barad, Gadhile, and Nat.

- **Party Preference** : Contesting parties are INLD ,Congress ,BJP , AAP ,Other and N/A (assigned to those voters whose party preference data is not known just for removing missing value)

- **Employment** : A Voter can be Contractor , Shopkeeper , Farmer , Doctor , Dealers, Private Job , Govt Job , Police , Unemployed and N/A (assigned to those voters whose Employment data is not known just for removing missing value) .

- **Education:** Qualification of voter can be Primary, Secondary, Higher Secondary, Diploma, Graduation, Post-Graduation, Doctorate, Illiterate and N/A (assigned to those voters whose qualification data is not known just for removing missing value).

  Education is further divided into sub education name as mentioned below:

  Primary is further be divided into 1, 2,3,4,5 sub education names.
  Secondary is further divided into 6, 7, 8,9,10 sub education names.
  Higher Secondary is further divided into 10+1, 10+2 sub education names.
  Diploma is further divided into ITI, Polytechnic sub education names.
  Graduation can be further divided into B.Tech, BA, BCA sub education names.
  Post-Graduation is further divided into M.Tech, MA, M.Phil. MCA sub education names.
  Doctorate is further divided into Ph.D. sub education name.
  Sub education name for **Illiterate** is Illiterate.

Data set contains data of voters of Dabwali Constituency of Haryana. As data set contains data of one Constituency Dabwali so prediction will be for only one constituency. Data is placed in the tables in SQL server. Data is in the normalized form. For example attributes of table stateParties in the data set are PartID, PartyName and Updated. Data set contain many tables some names are stateCastesMain (which contain main castes of voters), stateCastesSub (which contains subcatename for castes of voters), stateEmployment (which contains employment names of voters) etc.

3.3.3Proposed System

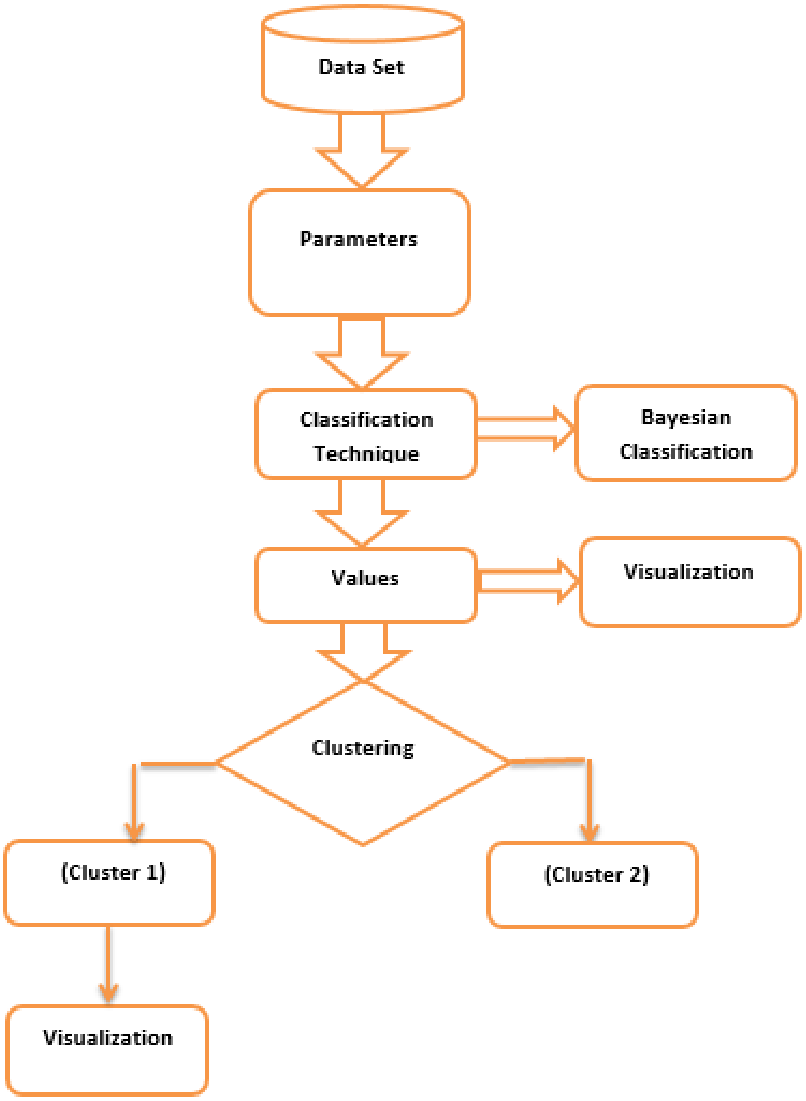


Figure 3.2: Proposed System

Data set contains data of voters of Dabwali Constituency of Haryana. The main aim of research is candidate plotting .This depend on the parameters .Parameters in this research are Party Preference, Education, Age, Employment and Caste of voters. As data set contains data of one Constituency Dabwali so prediction will be for only one constituency.

Data mining technique of classification is used in this research .Bayesian Classification is used to get the results and then results are binded to chart series. For this research Visual Studio 2012 is used as front end and for database Microsoft SQL Server Management Studio 2012 is used as backend.

Bayesian Classification is applied on the parameters of research. For e.g. for caste, value for general caste out of total is extracted and count of voters who are general and party preference let's say INLD is extracted .Then count of voters who are preferring INLD are extracted from database and then divided by total voters we will get a value . The value of count of voters who are general and party preference is INLD is divided by count of voters who are preferring INLD.Then these values are put in Bayes theorem formula and we will get a value. Same is done for other castes and other parties and then these values are binded to chart series, showing us result in the pie chart form. Then we come to know which caste has more value .Then any threshold value will be set .Then the caste having value above the threshold will be selected and only their sub caste name will be plotted on the pie chart.

For e.g. for caste, value for general caste out of total is extracted and count of voters who are general and party preference let's say INLD is extracted .Then count of voters who are preferring INLD are extracted from database and then divided by total voters we will get a value . The value of count of voters who are general and party preference is INLD is divided by count of voters who are preferring INLD.Then these values are put in Bayes theorem formula and we will get a value. Like this we can calculate values for all parties by putting in Bayes theorem formula and result will be shown by the pie chart. On these values we then apply clustering algorithm k –means .Then we will get two clusters .One cluster containing values to be processed further and values in the other cluster will be discarded. For e.g. as shown in Figure 6.1 presenting the percentages of parties preferred by people of constituency Dabwali. Clustering result can be seen on the point chart .Points in Cluster1 will be represented by blue dots and Cluster2 with red dots.Cluster2 will not be considered in research.

3.3.3.1 k-means Clustering

Clustering means dividing large data sets into smaller data set of some similarity. In this we create clusters and elements of cluster will have same properties. For instance the things in the grocery area unit are clustered in classes (butter, cheese and milk are grouped in dairy products). K-means is one of the easiest unsupervised learning algorithm that gives the solution of the well-known clustering problem .The Lloyd's algorithm is known as k-means algorithm.

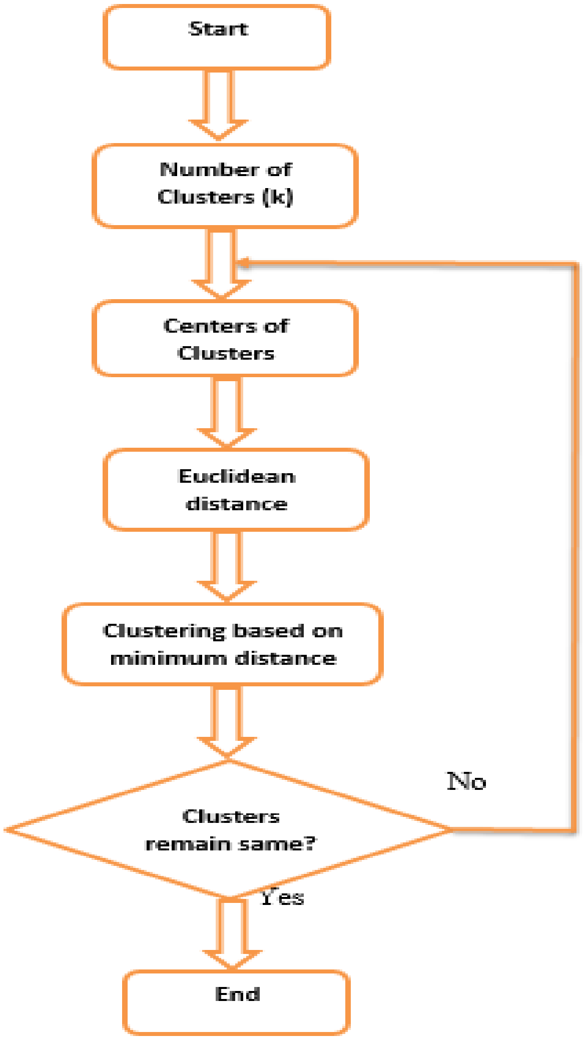


Figure 3.3: Flow chart of k-means

K-means algorithm works as follows:

Step 1: k represent number of clusters .First we have to decide number of clusters we want to make.

Step 2: Initialize the centers of clusters.

Step 3: Calculate the distance between each point and centers of clusters initialize above.

$$\| x_i-v_j \|^2$$

$\| x_i-v_j \|^2$ is the Euclidean distance between x_i and $v_{j.}$

Step 4: Assign the points to the clusters whose distance from the cluster center is minimum of all the cluster centers**.**

Step 5: Recalculate the cluster centers and repeat steps 3 and 4.

Step 6: If clusters data remains same then finish, otherwise repeat steps 3 and 4.

3.3.3.2Bayesian Classification

Classification is one of the most abrupt used technique by data mining and machine learning (ML) researchers. There are different classification methods like Bayesian Classification etc. In Classification classifier is framed to predict categorical labels. Naïve Bayes algorithm is prediction algorithm .Based on past facts we can predict future activities. Bayes Theorem has three parameters that we provide to calculate the theorem.

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

Figure 3.4: Byes theorem

- **Prior** includes all the information from day to day past experience. For instance my friend calls up in the morning and tells me he is dinking something? Ask me to guess what he is drinking? We know in the morning we drink either coffee or tea .So from past experience I can guess it can be either coffee or tea.

- **Likelihood** in the previous example if my friend tells me that whatever he is drinking is very cold .Then there is possibility he is drinking cold coffee .So this tells us about the possibility of the information across the scenario.

- **Posterior** predicts the particular information based on the given information. In the previous example if my friend is telling he is drinking tea or coffee, then based on this information I can predict that tomorrow morning he will drink tea or coffee.

- **Evidence** is the total number of cases when event occurs alone.

Now taking an example .Suppose there are 100 people and from these we want to calculate number of people who are above 60 and prone to heart disease .Given information is 25 people having heart disease and 75 are above 60. So posterior will be

$$\text{Posterior} = (25 * 100)/75$$

$$= 33.3$$

This shows that 33.3 percent people are 60 above and prone to heart disease out of 100 people.

3.3.3.3SQL server 2012 and Visual Studio 2012

Microsoft SQL Server is a relational database management system (RDBMS) created by Microsoft. As a database, it is a product whose main perform is to store and retrieve information as requested by different software applications, applications can be on the same laptop or those running on another computer across a network (including the Internet).SQL (Structured Query Language) is employed to manage the RDBMS. SQL Server 2014 permits high-performance across OLTP (Online Transaction Processing), data warehousing, BI and analytics for up to 30x quicker transactions. We can create databases, drop databases, create tables, update tables, and drop tables using sql queries in studio. We can also send e-mails via sql server.

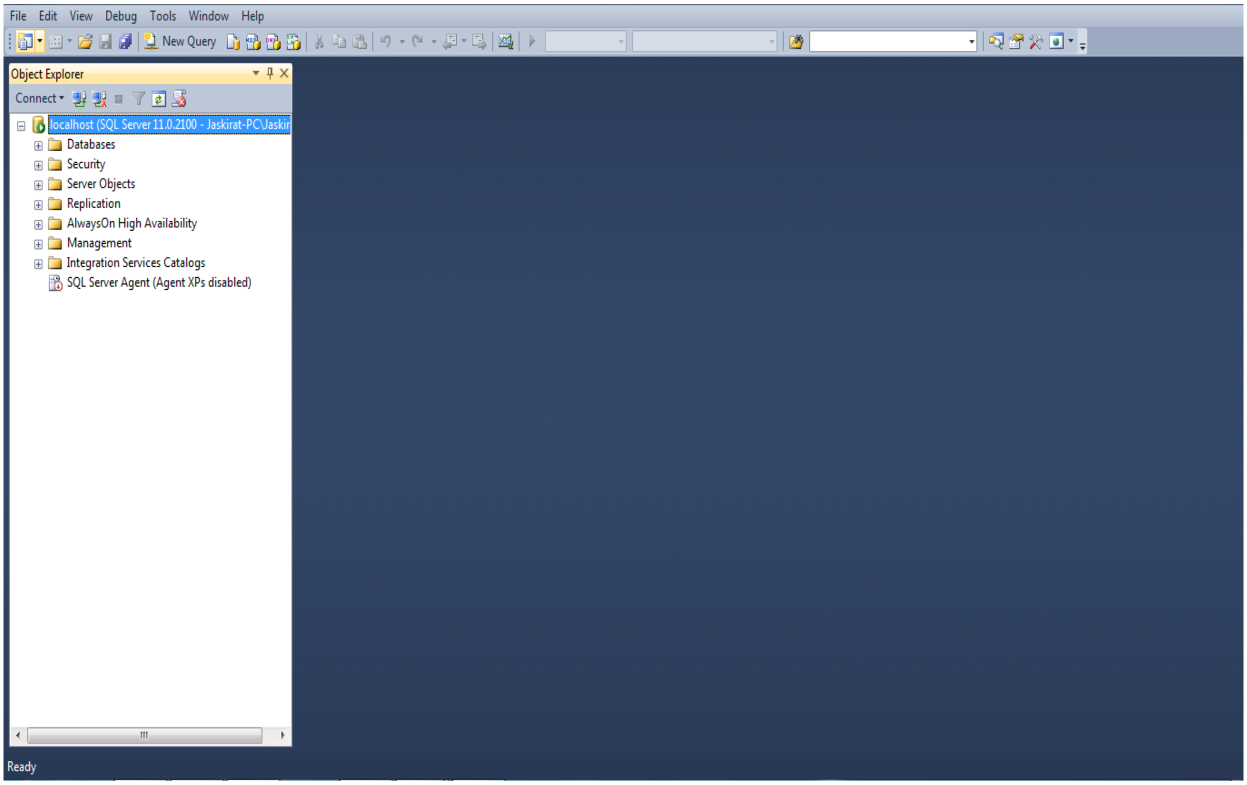
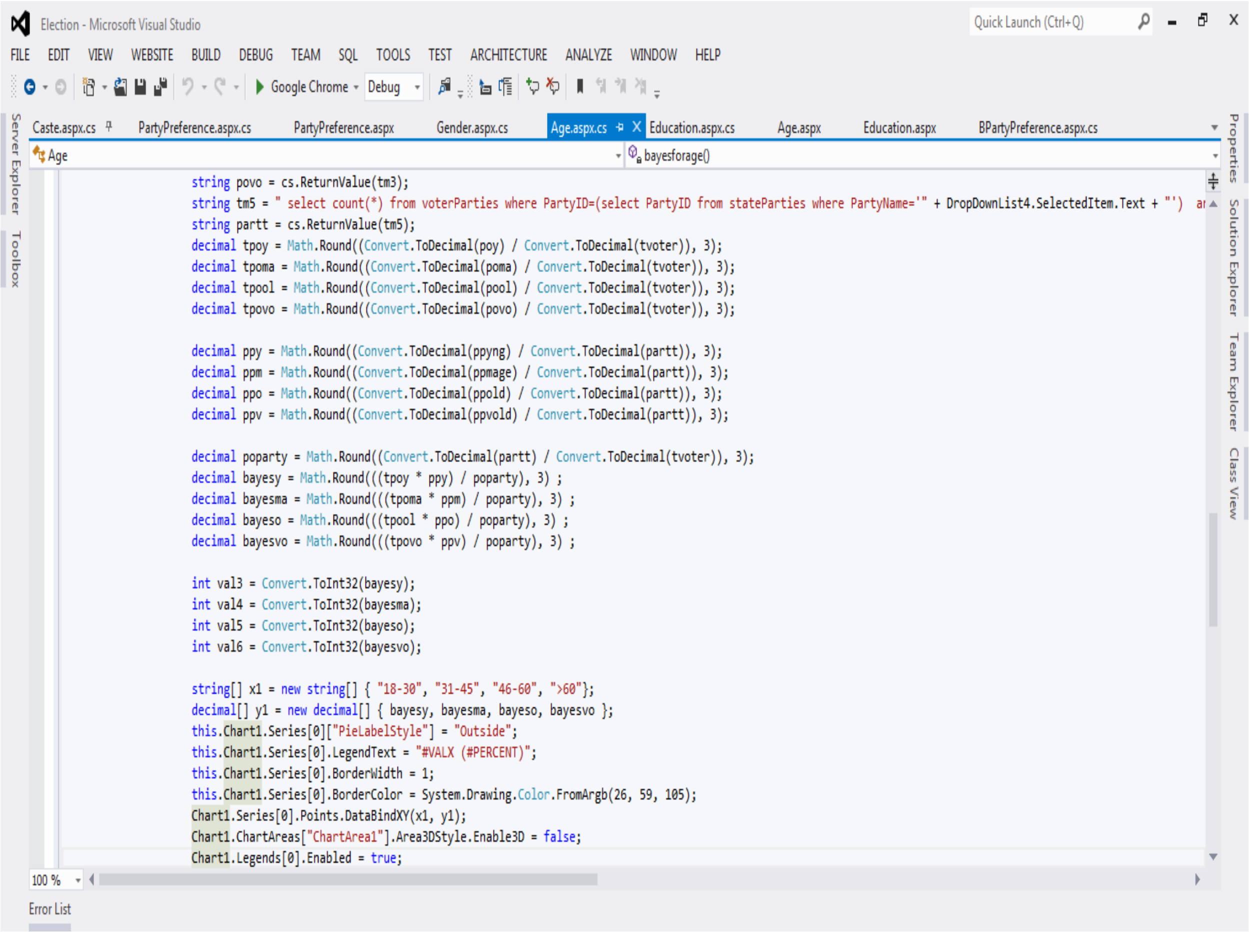

File Edit View Debug Tools Window Help
New Query
Object Explorer
Connect
localhost (SQL Server 11.0.2100 - Jaskirat-PC\Jaskir
Databases
Security
Server Objects
Replication
AlwaysOn High Availability
Management
Integration Services Catalogs
SQL Server Agent (Agent XPs disabled)
Ready

Figure 3.5: Microsoft Management Studio 2012

Microsoft Visual Studio is IDE (Integrated Development Environment) .It is developed by Microsoft .IDE provides full facilities to the programmers to develop the software’s. It contains code editor, tools (which can be dragged from toolbox for use, no need to code it) and debugger. It can be used to construct the web sites, web applications, windows applications, web services (communication of two electronic devices can be either phones, pc over a network).Its features are designer , debugger and last but not least code editor .



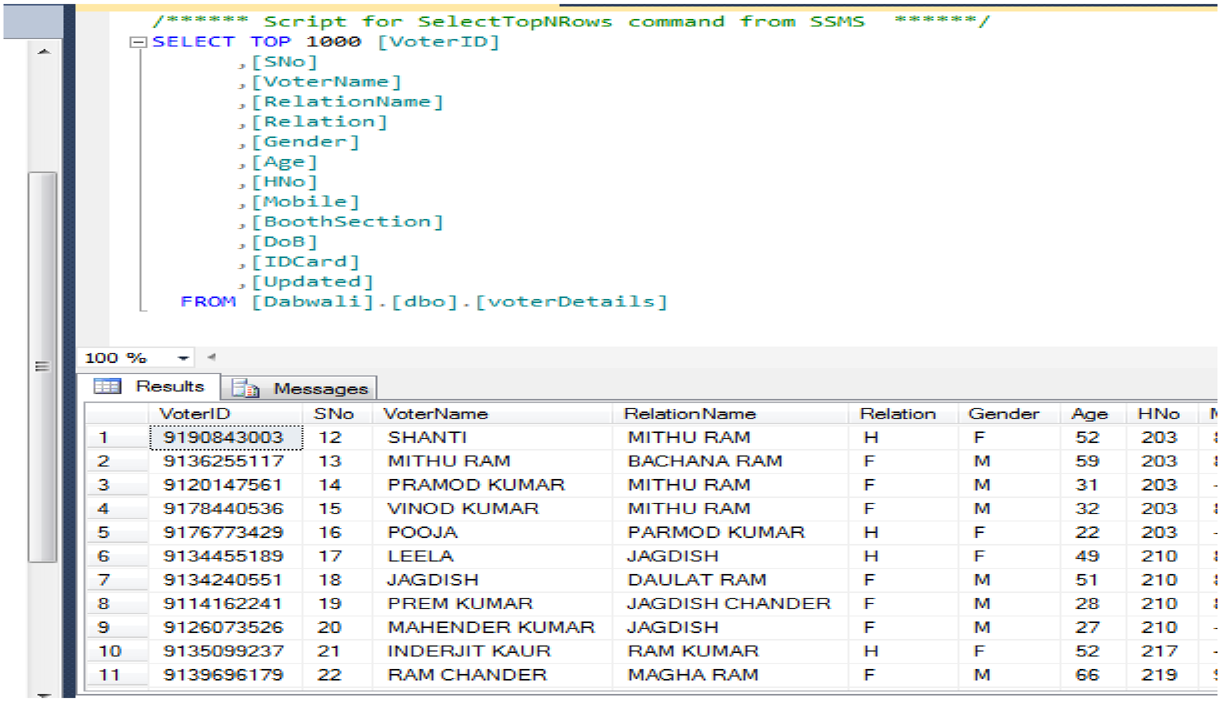
```
string povo = cs.ReturnValue(tm3);
string tm5 = " select count(*) from voterParties where PartyID=(select PartyID from stateParties where PartyName='" + DropDownList4.SelectedItem.Text + "')  ar
string partt = cs.ReturnValue(tm5);
decimal tpoy = Math.Round((Convert.ToDecimal(poy) / Convert.ToDecimal(tvoter)), 3);
decimal tpoma = Math.Round((Convert.ToDecimal(poma) / Convert.ToDecimal(tvoter)), 3);
decimal tpool = Math.Round((Convert.ToDecimal(pool) / Convert.ToDecimal(tvoter)), 3);
decimal tpovo = Math.Round((Convert.ToDecimal(povo) / Convert.ToDecimal(tvoter)), 3);

decimal ppy = Math.Round((Convert.ToDecimal(ppyng) / Convert.ToDecimal(partt)), 3);
decimal ppm = Math.Round((Convert.ToDecimal(ppmage) / Convert.ToDecimal(partt)), 3);
decimal ppo = Math.Round((Convert.ToDecimal(ppold) / Convert.ToDecimal(partt)), 3);
decimal ppv = Math.Round((Convert.ToDecimal(ppvold) / Convert.ToDecimal(partt)), 3);

decimal poparty = Math.Round((Convert.ToDecimal(partt) / Convert.ToDecimal(tvoter)), 3);
decimal bayesy = Math.Round(((tpoy * ppy) / poparty), 3) ;
decimal bayesma = Math.Round(((tpoma * ppm) / poparty), 3) ;
decimal bayeso = Math.Round(((tpool * ppo) / poparty), 3) ;
decimal bayesvo = Math.Round(((tpovo * ppv) / poparty), 3) ;

int val3 = Convert.ToInt32(bayesy);
int val4 = Convert.ToInt32(bayesma);
int val5 = Convert.ToInt32(bayeso);
int val6 = Convert.ToInt32(bayesvo);

string[] x1 = new string[] { "18-30", "31-45", "46-60", ">60"};
decimal[] y1 = new decimal[] { bayesy, bayesma, bayeso, bayesvo };
this.Chart1.Series[0]["PieLabelStyle"] = "Outside";
this.Chart1.Series[0].LegendText = "#VALX (#PERCENT)";
this.Chart1.Series[0].BorderWidth = 1;
this.Chart1.Series[0].BorderColor = System.Drawing.Color.FromArgb(26, 59, 105);
Chart1.Series[0].Points.DataBindXY(x1, y1);
Chart1.ChartAreas["ChartArea1"].Area3DStyle.Enable3D = false;
Chart1.Legends[0].Enabled = true;
```

Figure 3.6: Microsoft Visual Studio 2012

CHAPTER 4

IMPLEMENTATION

Data mining technique of classification is used in this research .Bayesian Classification is used to get the results and then results are binded to chart series. For this research Visual Studio 2012 is used as front end and for database Microsoft SQL Server Management Studio 2012 is used as backend. Visualizations is an important part of the system. As the voter data for the state will be in crores, the visualization of the data will be an important part. The data can be displayed in form of Line Charts & Pie Charts and Graphs. It will be easily understandable to the users. Data set contains data of voters of Dabwali Constituency of Haryana. As data set contains data of one Constituency Dabwali so prediction will be for only one constituency. Data is placed in the tables in SQL server.

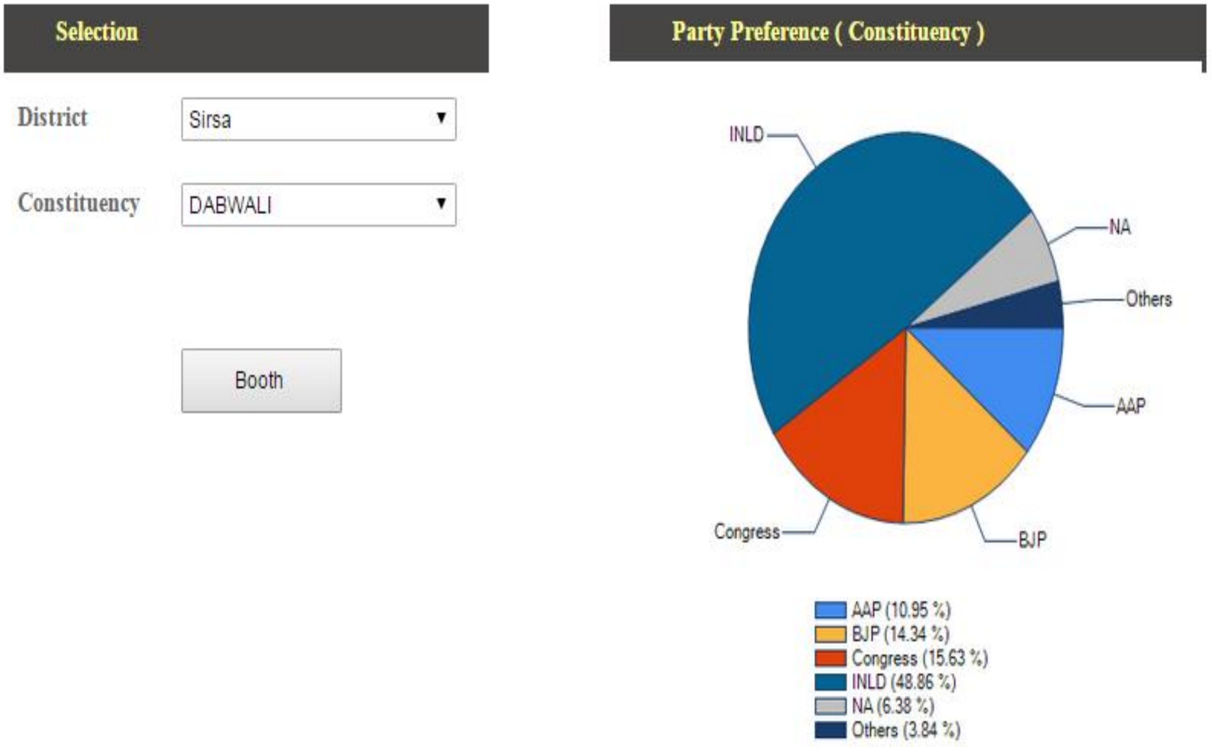


```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP 1000 [VoterID]
      ,[SNo]
      ,[VoterName]
      ,[RelationName]
      ,[Relation]
      ,[Gender]
      ,[Age]
      ,[HNo]
      ,[Mobile]
      ,[BoothSection]
      ,[DoB]
      ,[IDCard]
      ,[Updated]
  FROM [Dabwali].[dbo].[voterDetails]
```

	VoterID	SNo	VoterName	RelationName	Relation	Gender	Age	HNo	M
1	9190843003	12	SHANTI	MITHU RAM	H	F	52	203	[illegible]
2	9136255117	13	MITHU RAM	BACHANA RAM	F	M	59	203	[illegible]
3	9120147561	14	PRAMOD KUMAR	MITHU RAM	F	M	31	203	-
4	9178440536	15	VINOD KUMAR	MITHU RAM	F	M	32	203	[illegible]
5	9176773429	16	POOJA	PARMOD KUMAR	H	F	22	203	-
6	9134455189	17	LEELA	JAGDISH	H	F	49	210	[illegible]
7	9134240551	18	JAGDISH	DAULAT RAM	F	M	51	210	[illegible]
8	9114162241	19	PREM KUMAR	JAGDISH CHANDER	F	M	28	210	[illegible]
9	9126073526	20	MAHENDER KUMAR	JAGDISH	F	M	27	210	-
10	9135099237	21	INDERJIT KAUR	RAM KUMAR	H	F	52	217	-
11	9139696179	22	RAM CHANDER	MAGHA RAM	F	M	66	219	[illegible]

Figure 4.1: Database Table

Figure 4.1 shows the table from database .Tables are stored in the SQL server. Above figure shows the table of voterdetails.This table contains only general information of voter like Voter ID, Voter Name, Relation name whether voter is husband, wife, and child like this and gender of voter.



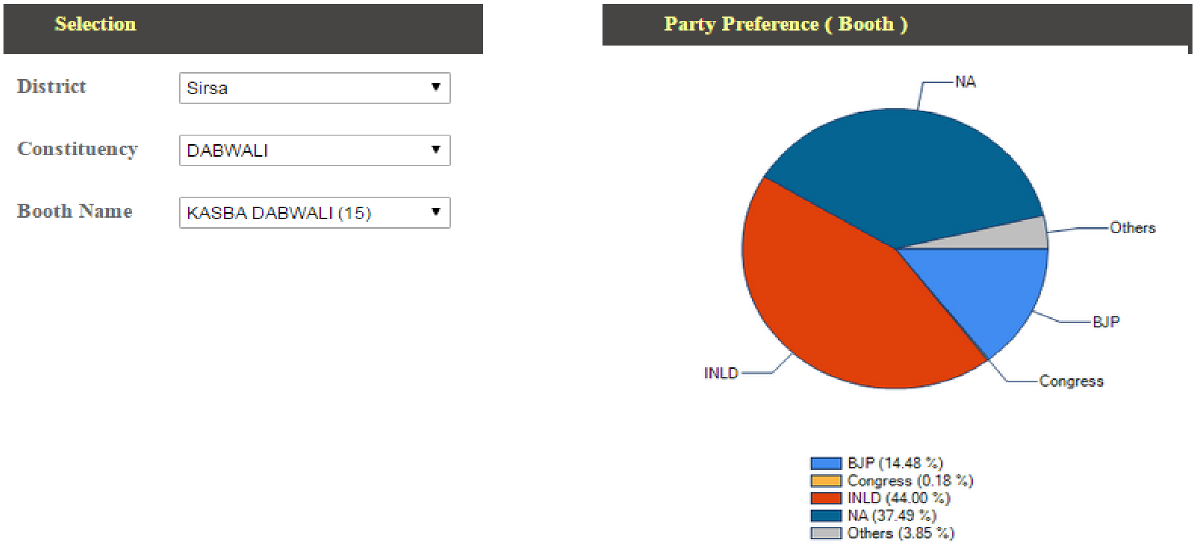
Selection

District: Sirsa

Constituency: DABWALI

Figure 4.2: Party Preference

Figure 4.2 showing Party Preference of voters in Dabwali constituency. Figure 4.2 clearly showing INLD is in majority with 48.86% and then is congress with 15.63%.There can be fluctuation in the results because people can change their preference at time of the elections.



Selection

District: Sirsa

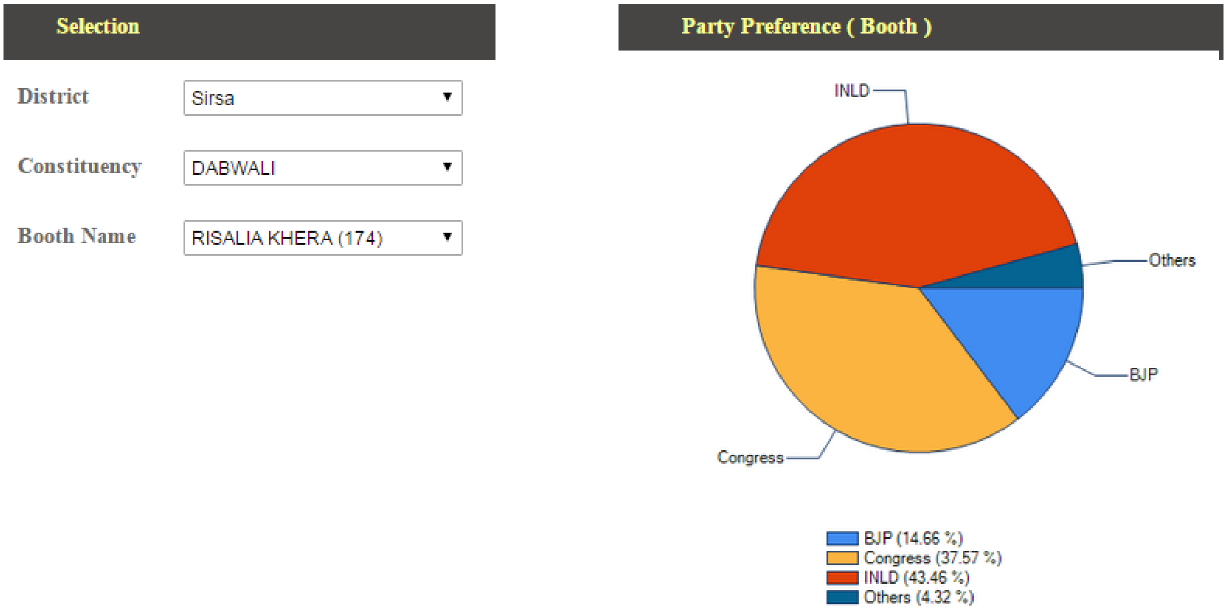
Constituency: DABWALI

Booth Name: KASBA DABWALI (15)

Party Preference (Booth)

Figure 4.3: Party Preference Booth wise

Figure 4.3 showing Party Preference Booth wise. In this figure INLD is preferred by 44% voters in KASBA DABWALI (15) booth of Constituency Dabwali. NA is used for those voters whose Party Preference is not known. Others can be for voters preferring Independent candidates.



Selection

District: Sirsa

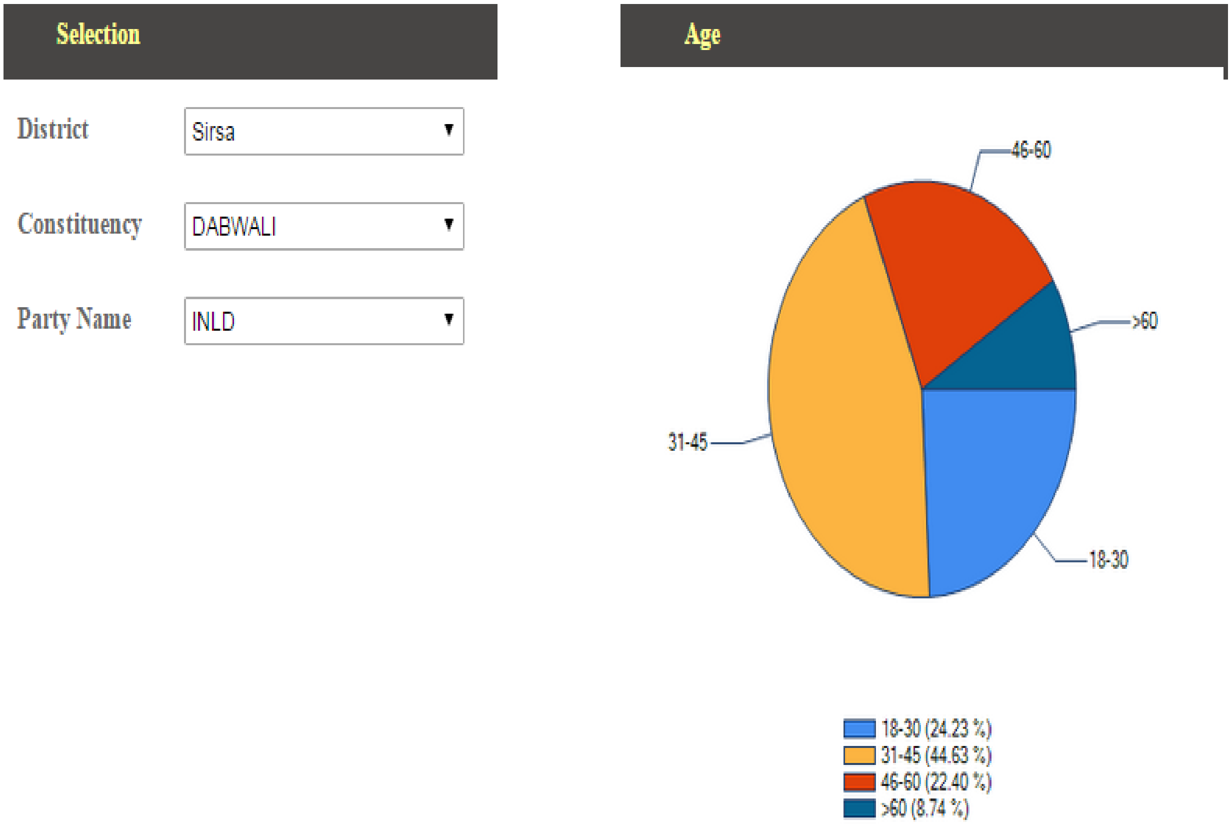
Constituency: DABWALI

Booth Name: RISALIA KHERA (174)

Party Preference (Booth)

Figure 4.4: Party Preference Booth wise

Figure 4.4 showing Party Preference Booth wise. In this figure INLD is preferred by 43.46% voters in RISALIA KHERA (174) booth of Constituency Dabwali. Others can be for voters preferring Independent candidates.



Selection

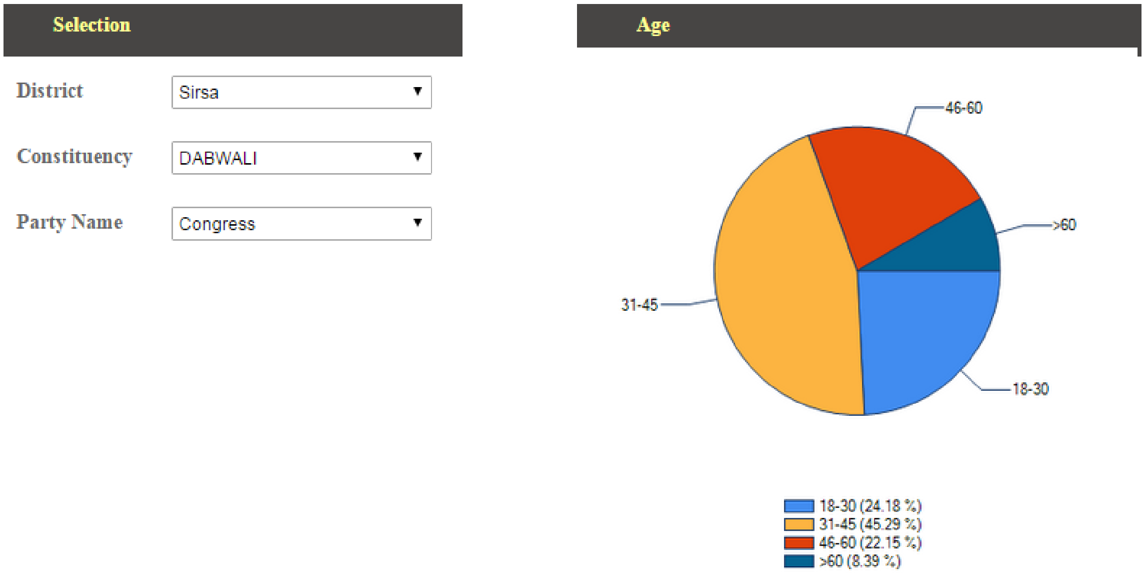
District: Sirsa

Constituency: DABWALI

Party Name: INLD

Figure 4.5: Age Groups

Figure 4.5 showing Age group of voters of Constituency Dabwali preferring INLD.This shows the percentage that what can be the age of the candidate for INLD. Above figure shows age group from 31-45 are preferring INLD most. From this we can predict that INLD will plot that candidate whose age is between 31 – 45 years. From above figure such prediction can be made. At backend Bayesian classification is used. From this we can come to know percentage of all age groups favoring the particular party.



Selection

District: Sirsa

Constituency: DABWALI

Party Name: Congress

Figure 4.6: Age Groups

Figure 4.6 showing Age group of voters of Constituency Dabwali preferring Congress. This shows the percentage that what can be the age of the candidate for Congress. Above figure shows age group from 31-45 are preferring Congress most. From this we can predict that Congress will plot that candidate whose age is between 31 – 45 years.

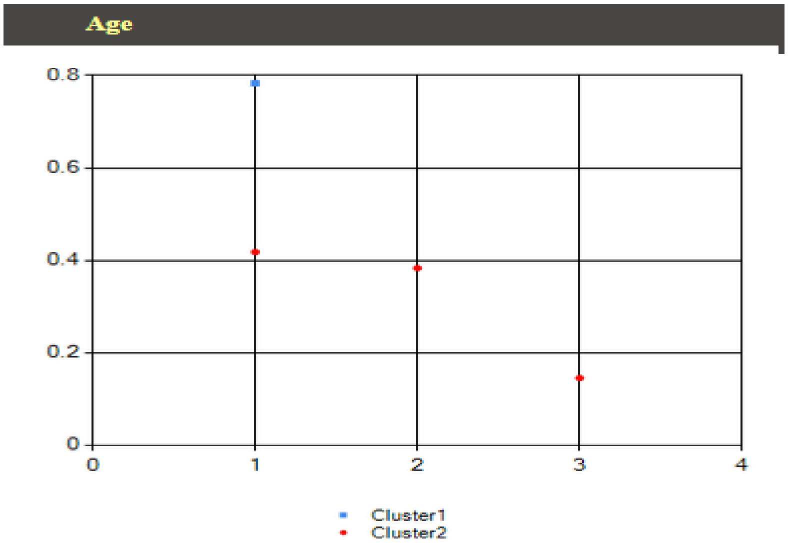
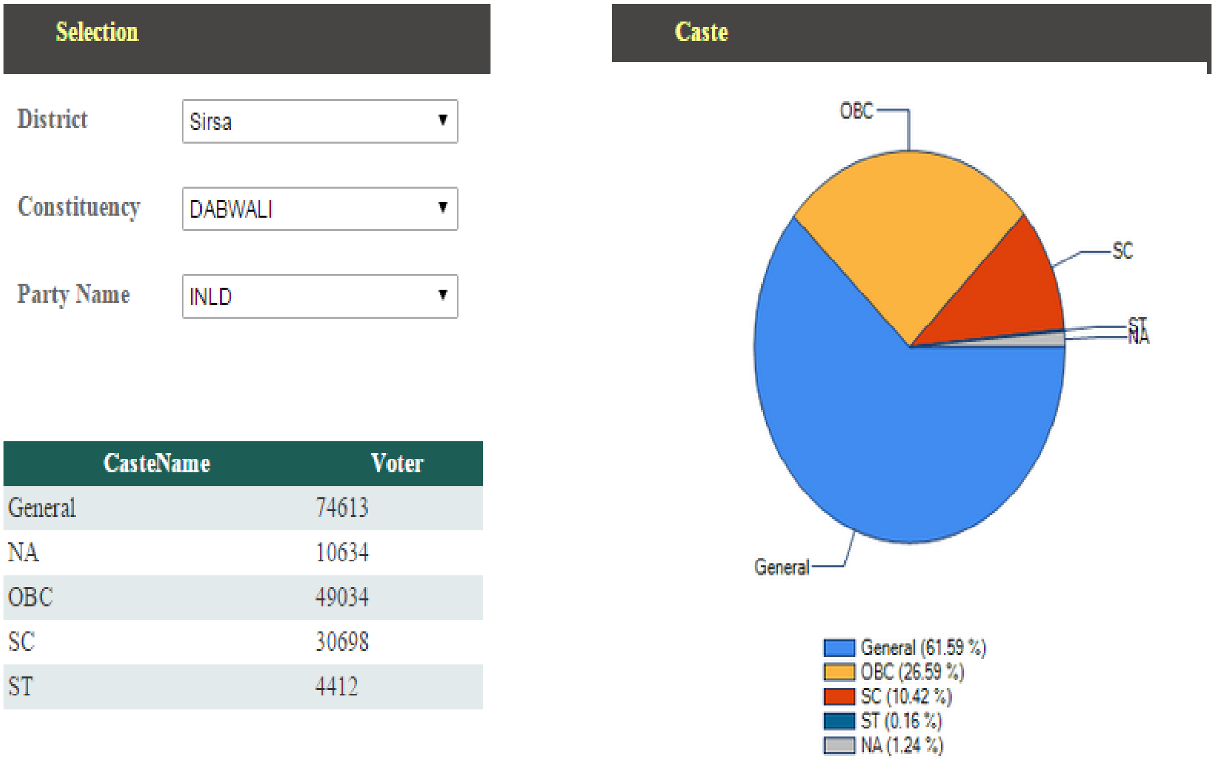


Figure 4.7: Age Groups Clustering

Figure 4.7 shows the clustering of age groups. Cluster1 will be carried further and Cluster2 will be discarded.Cluster1 contains only one data point (which is age group 31-45), rest of age groups will not be considered. This shows that only Cluster1 is favorable and age group for candidate can be 31-45. For this k-means is used at backend.



Selection

District: Sirsa

Constituency: DABWALI

Party Name: INLD

CasteName	Voter
General	74613
NA	10634
OBC	49034
SC	30698
ST	4412

Figure 4.8: Caste of Voters

Figure 4.8 shows which caste is preferring which party most. As above figures clearly shows that INLD is supported by General category most (61.59%).From this we can predict that caste of candidate can be General in that constituency as it is preferred by General category most. Such prediction can be made. Bayesian classification is used to get above results.

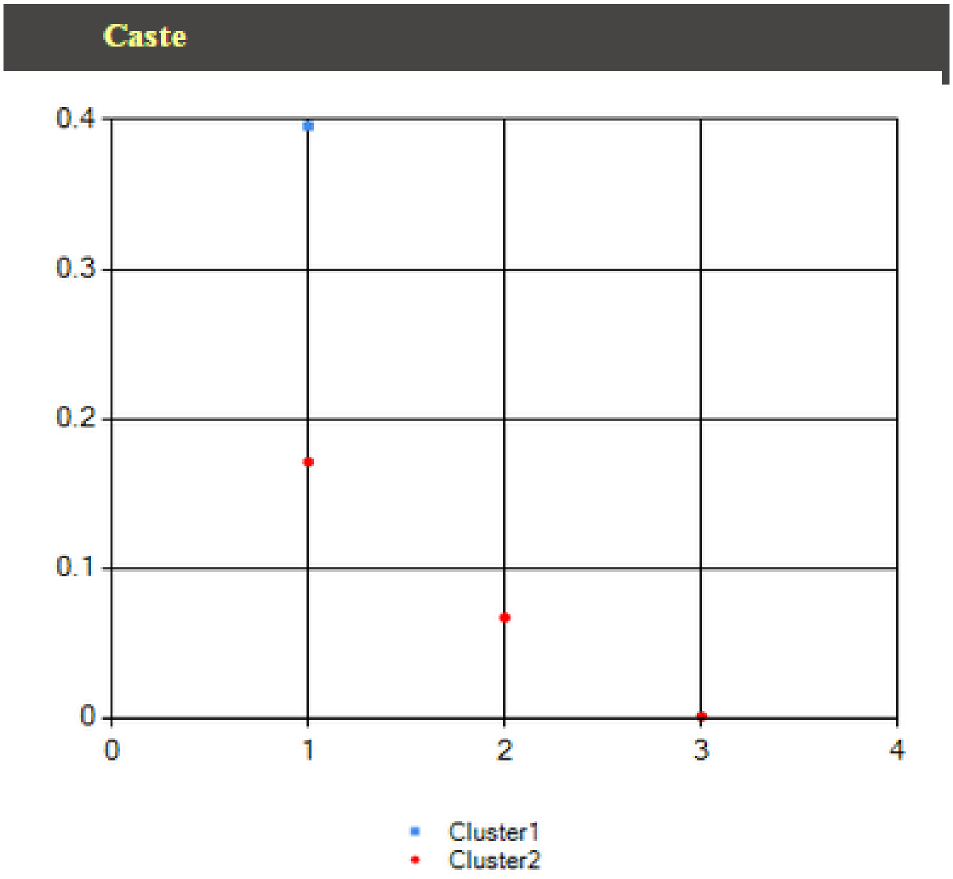
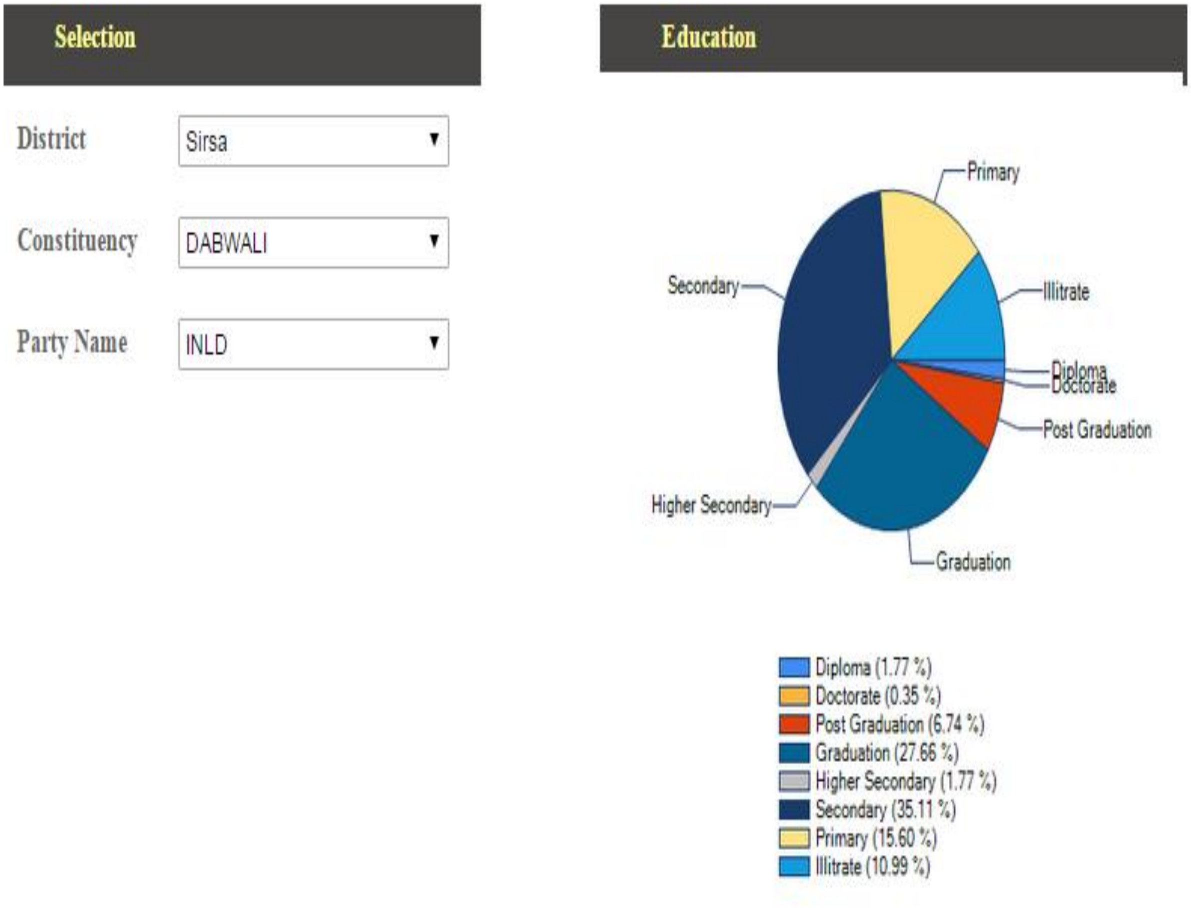


Figure 4.9: Clustering of castes

Figure 4.9 shows the clustering of castes of voters in Dabwali constituency. Cluster1 will be carried further and Cluster2 will be discarded which is represented by red dots(NA is not included).Cluster2 contains caste names such as OBC , SC, and ST.Cluster1 contains only one data point (which is General category), rest of castes will not be considered. This shows that only Cluster1 which is represented by Blue dots is favorable and caste of candidate can be General category in that constituency. Such result can be concluded from clustering shown in above figure. For this k-means is used at backend.



Selection

District: Sirsa

Constituency: DABWALI

Party Name: INLD

Figure 4.10: Education of Voters

Figure 4.10 represents the qualification of voters of constituency Dabwali who are supporting INLD. In above figure percentage of voters possessing secondary education (35.11%) is more and they are supporting INLD as well. From this we cannot predict this that candidate should possess only secondary education in that constituency. The prediction can be candidate should be well educated as after secondary education, the percentage of voters who are supporting INLD and graduated as well is more i.e. 27.66%. Only 11% illiterate people are supporting INLD.So only educated candidate can be plotted in Dabwali constituency.

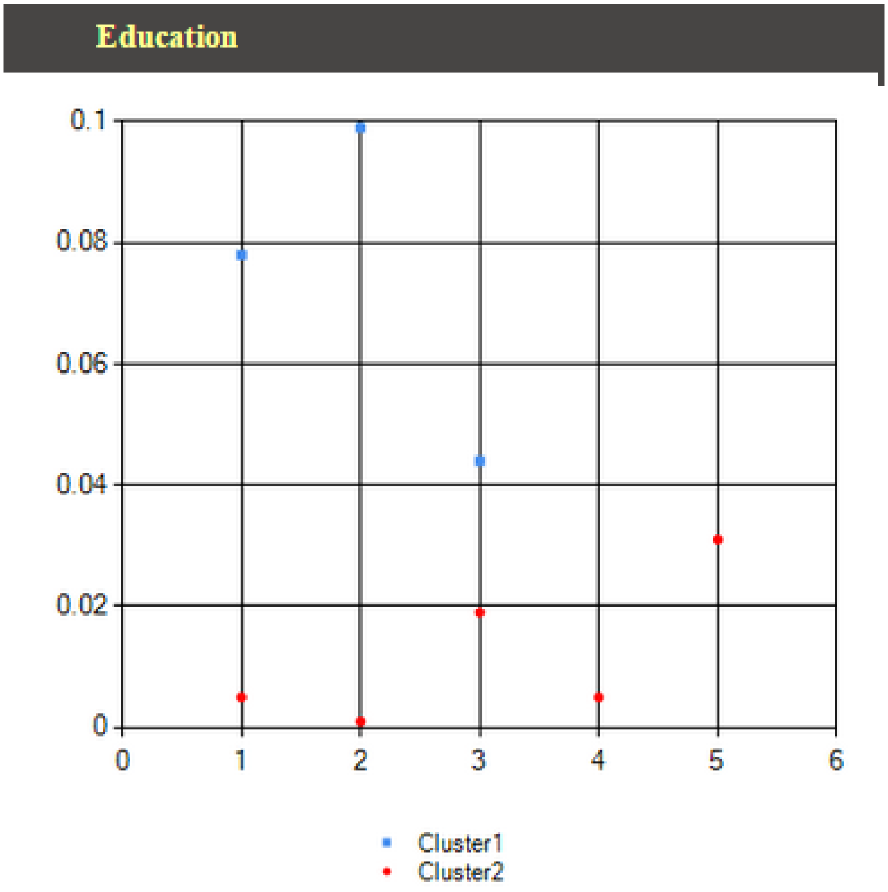


Figure 4.11: Clustering of Education

Figure 4.11 represents the clustering of Education of voters in Dabwali constituency. Cluster1 will be carried further and Cluster2 will be discarded which is represented by red dots.Cluster2 contains Education names such as Diploma, Doctorate, illiterate and Post-graduation and higher secondary because their percentage is less.Cluster1 contains only three data point (which are Graduation, Primary, Secondary), rest of Education will not be considered. This shows that only Cluster1 which is represented by Blue dots is favorable and candidate should be well educated in that constituency. It should not be illiterate as favorable cluster contains only data of literate voters who supports INLD. Such result can be concluded from clustering shown in above figure. For this k-means is used at backend.

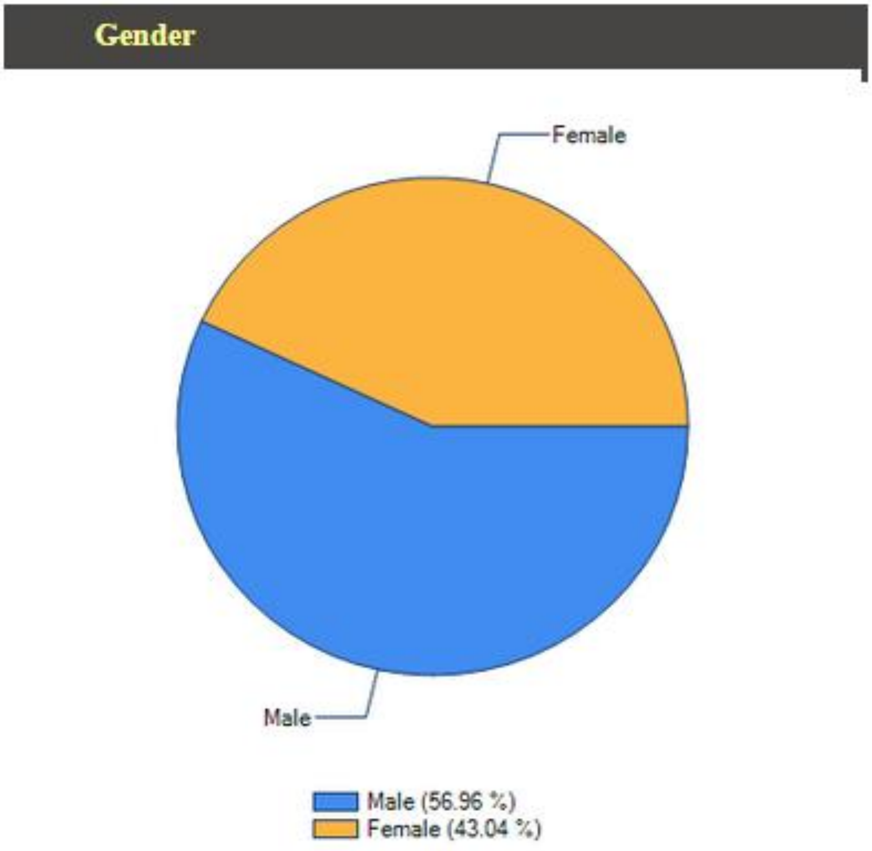


Figure 4.12: Gender

Figure 4.12 represents the gender of voters. It shows whether females are supporting particular party more or males, from that we can predict what can be the gender of the candidate. Figure 4.12 represents male voters are preferring INLD more than the voters whose gender is female. It shows possibility of plotting male or female candidate who supports INLD.So from here we can predict that gender of candidate can be male or we can plot male candidate in the Dabwali constituency of Haryana for INLD. We can predict the gender of other parties also in the particular constituency.

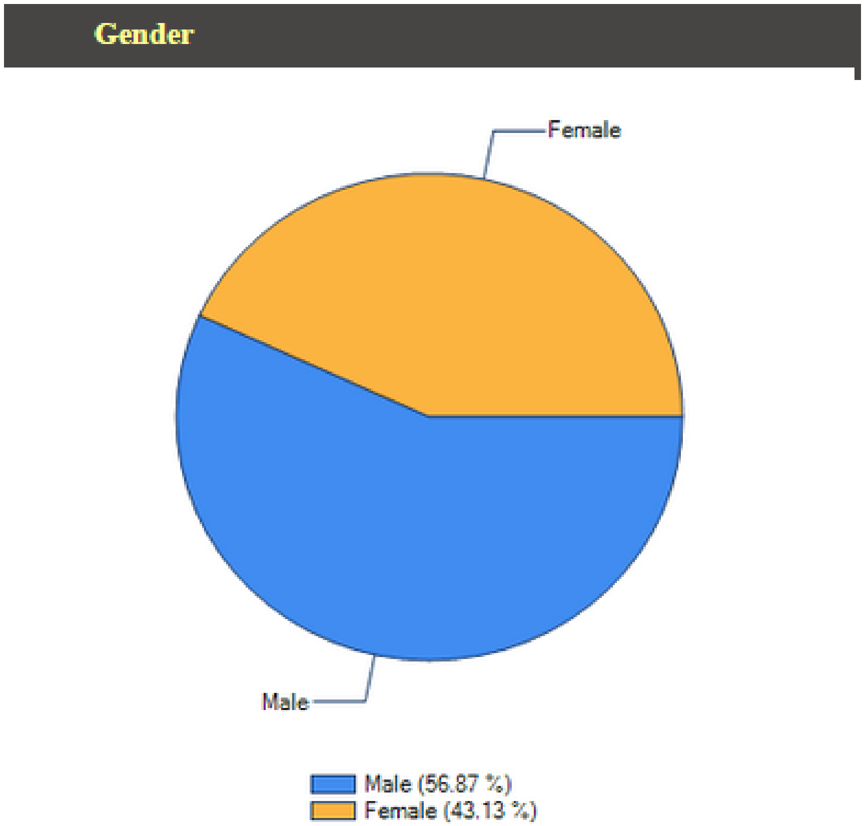


Figure 4.13: Gender

Figure 4.13 represents male voters (56.87%) are preferring AAP more than the voters whose gender is female (43.13%). It shows possibility of plotting male or female candidate who supports AAP.So from here we can predict that gender of candidate can be male or we can plot male candidate in the Dabwali constituency of Haryana for AAP (Aam Aadmi Party). We can predict the gender of other parties also in the particular constituency.

CHAPTER 5

CONCLUSIONS AND FUTURE SCOPE

To extract useful information from large data sets data mining techniques like Classification, Clustering and Association are used. KDD (Knowledge Discovery Process) is data mining method to extract information from large amount of data. Knowledge can be extracted which can be used for predictions and making important decisions .Bayesian Classification and k-means clustering is used for mining purpose in this research. The proposed system has emphasis on the helping the election contesting party to decide the candidates and strategize on contesting plan. There will be a system which will help the contesting parties to frame their contesting plan, so that easy and efficient decisions can be taken. After analyzing the data easy and effective decisions are made .Decisions are like who will be the contestant.

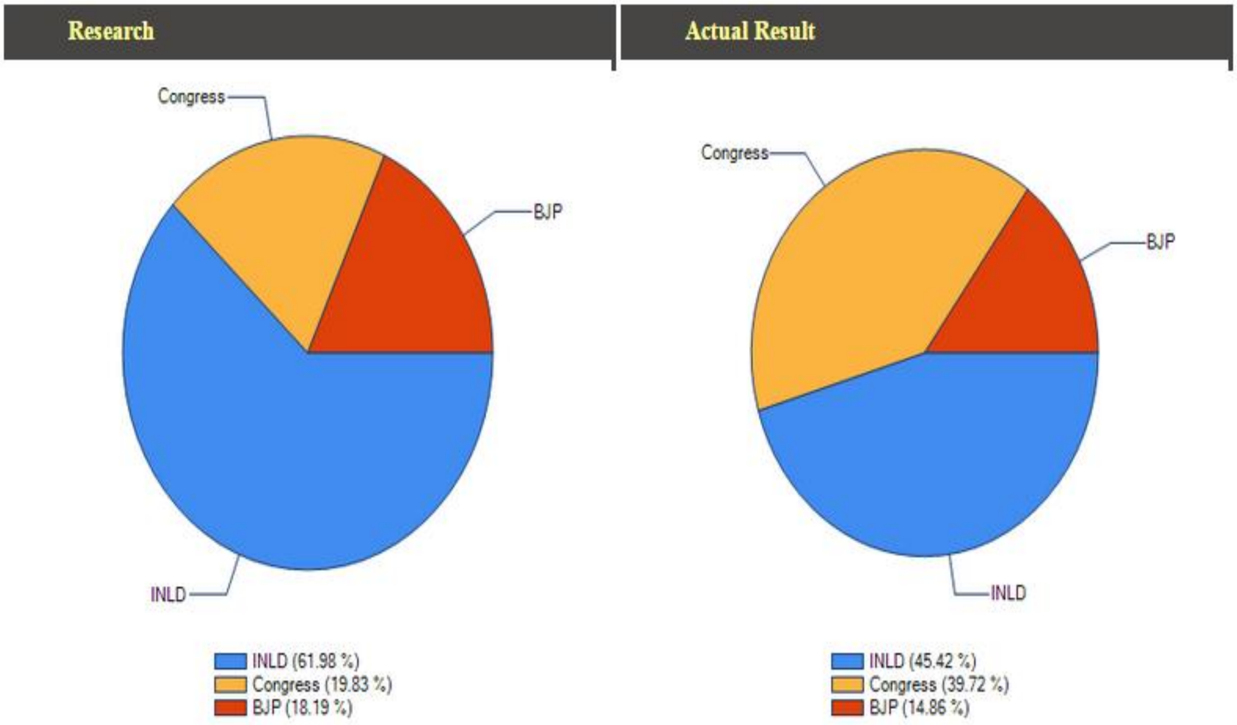


Figure 5.1: Result Comparison

Actual results were 68029 votes were casted in the favor of INLD and 59484 in favor of Congress and 22261 for BJP. According to the research 82758 votes are in the favor of INLD and 26480 in favor of Congress and 24294 for BJP.Difference in result may be due to change of preference of voters at the time of election. As election were held in October 2014 and data set was created at the last of February 2014 .So 7-8 months duration is more than enough to change the preference. But winning party is INLD and gender of candidate is female (Naina Singh Chautala). Another reason can be the due to records which were added manually without any survey.

This research has done prediction for only one constituency, in future research will be done on whole state and predictions can be done for all MLA constituencies and MP constituencies as well.Reserach can be done using another techniques of classification and clustering other than Bayesian classification and k-means clustering.

APPENDIX

Glossary of Terms

Classification
In Classification classifier is framed to predict categorical labels .It is supervised learning technique. . In classification, software can be developing which can learn how to classify the data items into groups. For e.g. a marketing manager of a company want to analyze whether a customer will buy a new computer .Here labels are yes and No.

Clustering
Clustering is a technique in which objects of similar category is placed in one group and other are in different group. It is unsupervised learning technique.

Data mining
Extraction of hidden information from large databases. Term that emulates data mining is KDD (Knowledge Discovery Process).

Decision tree
Tree shaped structure used for taking decisions.

ID3
ID3 uses Information gain for attribute selection measure. It used to select splitting attribute.

Data cleaning
Removing noise, duplicate values and missing values from the database used for data mining.

visualization
Showing output in graphical form .It can be in form of pie charts ,line charts, bar charts etc.

noise
Noise in data means data is not complete, there are missing values duplicate values. Noisy data cannot be used for analysis purpose.

outlier
Using clustering we can do outlier detection where outliers are values lying

outside the cluster.

ETL Data from different sources like operational databases, files etc.is extracted, transformed and loaded (ETL) into database called data warehouse

Data Selection Selection of relevant data from the database for analysis. For e.g. using SQL query “select * from table1 where partyname=’INLD’”

Data warehouse It is database used to store the data. It contains noise free data. Data warehouse data is used for analysis purposes as it is noise free.

CLARA Clustering for Large Applications (CLARA) is technique of data mining for clustering of larger datasets.

Extraction Retrieval of results from the database. For e.g. “select * from table1”.

Data Transformation Data are transformed into appropriate or standard forms for mining.

LIST OF REFERENCES

[1] Vasile Paul Brefelean “*Analysis and Predictions on Students’ Behavior Using Decision Trees in Weka Environment* “, International Conference on Technology Interfaces (ITI), June 25-28, 2007.

[2] Jiawei Han J and Kamber M, Data Mining: Concepts and Techniques (2rd Edition). *Morgan Kaufmann, San Francisco*, CA, 2012.

[3] Brijesh Kumar Bhardwaj, Saurabh Pal Kamal “*Data Mining: A prediction for performance improvement using classification*” (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011.

[4] S.R.Pande, Ms.S.S.Sambare, V.M.Thakre,”*Data Clustering Using Data Mining Techniqes*”,
IJARCCE Vol. 1, issue 8, October 2012.

[5]] Mahendra Pratap Yadav, Mhd Feeroz and Vinod Kumar Yadav (2012*) “Mining the customer behavior using web usage mining In e-commerce”* Coimbatore, India. IEEE-201S0.

[6] Neelamadhab Padhy, Dr. Pragnyaban Mishra and and Rasmita Panigrahi (2012) “*The Survey of Data Mining Applications and Feature Scope*” International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[7] Bunkar, Rajesh Bunkar, Umesh Kumar Singh (2012) “*Data Mining: Prediction for Performance Improvement of Graduate Students using Classification*”.

[8] S.Anupama Kumar, Vijayalakshmi M.N “*Mining Of Student Academic Evaluation Records in Higher Education* “, IEEE, 2012.

[9] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam “*A Study of Data Mining Tools in Knowledge Discovery Process* “International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-3, July 2012.

[10] Dr.Tariq Mahmood, Tasmiyah Iqbal, Farnaz Amin, Wajeeta Lohanna, Atika Mustafa”*Mining Twitter Big Data to Predict 2013 Pakistan Election Winner*” , IEEE, 19-20 Dec. 2013.

[11] Ajay Kumar Pal, Saurabh Pal "*Evaluation of Teacher's Performance: A Data Mining Approach* "International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 12, December 2013, pg.359 – 369.

[12] Malhar Anjaria, Ram Mahana Reddy Guddeti "*Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning* ", IEEE (2014).

[13] M.Durairaj, C.Vijitha "*Educational Data mining for Prediction of Student Performance Using Clustering Algorithms*", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, 5987-5991.

[14] Election Commission of India, http://en.wikipedia.org/wiki/Election_Commission_of_India,1 April 2015.

[15] List of political parties in India, http://en.wikipedia.org/wiki/List_of_political_parties_in_India,1 April 2015.

[16] Elections, http://www.elections.in/delhi/,1 April 2015.