Predictive And Descriptive Analysis Using Bayesian Networks

Project Report submitted in partial fulfillment of the requirement for the degree of

Bachelor of Technology

in

Computer Science & Engineering

under the Supervision of

Dr.Sakshi Babbar

By

Shubham Awasthi(111466)

to

Jaypee University of Information and Technology

Waknaghat, Solan – 173234, Himachal Pradesh

Certificate

This is to certify that project report entitled "Predictive Analysis Using Bayesian Networks", submitted by Shubham Awasthi in partial fulfillment for the award of degree of Bachelor of Technology in Information Technology to Jaypee University of Information Technology, Waknaghat, Solan has been carried out under my supervision.

This work has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Supervisor's Name

Dr. Sakshi Babbar

Designation

Signature

Date:

Acknowledgement

There are many people who are associated with this project directly or indirectly whose help and timely suggestions are highly appreciable for completion of this project. First of all, I would like to thank Prof. Dr. SP Ghrera, Head, Department of Computer Science Engineering for his kind support and constant encouragements, valuable discussions which is highly commendable.

I would like to express my sincere gratitude to my supervisor Dr.Sakshi Babbar , for her super vision, encouragement, and support which has been instrumental for the success of this project. It was an invaluable experience for me to be one of her students. Because of her, I have gained a careful research attitude.

Lastly, I would also like to thank my parents for their love and affection and especially their courage which inspired me and made me to believe in myself.

Date: Shubham Awasthi

Roll No. 111466

Table of Content

Abbreviations

IND- India

AUS- Australia

RSA-Republic of South Africa

WI- West Indies

NZ-New Zealand

SL-Sri Lanka

PAK- Pakistan

BAN-Bangladesh

ZIM-Zimbabwe

SCO-Scotland

IRE- Ireland

AFG- Afghanistan

ENG-England

UAE- United Arab Emirates

SP- Slightly Polluted

E-Excellent

A-Acceptable

HP- Highly Polluted

P- Polluted

DO-Dissolved Oxygen

BOD- Biological Oxygen Demand

ARN- Association Rule Network

List of Figures

List of Tables

Abstract

This project aims at predicting real life applications and stresses the importance of data mining in real world. Data mining is an important part of knowledge discovery process that we can analyze an enormous set of data and get hidden and useful knowledge. Keeping the fundamental value of data mining in mind this project concentrates on 2 real world applications. Firstly this project can be used in prediction of cricket world cup 2015 using Bayesian Networks technique. Secondly this project can be used in predicting the quality of river water sample from the given data sets.

For the first part using Bayesian model we forecast winners of groupstages, quarterfinals, semifinals and final. Our model predicts Australia to be the new cricket champion of the world. To prove strength of our approach we show predictive results on past world cup 2011.

Later in this project we make use of the classification technique of data mining to accomplish predictive task in water quality. Classification technique of data mining is basically the task of assigning objects to one of several predefined categories, is a pervasive problem that encompasses many diverse applications. In this project, we have used naïveBayes classification technique for prediction of water quality in the river Yamuna. To assess the quality of water we have considered 13 parameters some of them are dissolved oxygen (DO), biochemical oxygen demand (BOD) ,pH value etc, that are necessary in order to correctly label the pollution level in the given water sample. Lastly we compare our results with another known classifier decision tree which proves that naïve bayes produces better results than it. In addition to this we have also performed the descriptive task using Association Rule Networks that gives us a clear idea of the role of a specific feature in contributing to a particular class.

CHAPTER 1

1.Introduction

The whole idea of this project is to do predictive analysis of an event which in my case is the ICC Cricket World Cup 2015 .The predictive analysis will be accomplished with the help of Bayesian Model. First of all a Bayesian network is modeled that is capable of predicting the outcome of a single cricket match. This network is composed of 16 data points which will be discussed in detail later on. The data points used in this network are classified from my knowledge of the game as well as are learn't from expert interviews/discussions that are broadcasted before and after match on match days.

Since world cup consist of 49 matches and the bayesian model can simulate only 1 match at a time hence 49 iterations are required to get the outcome of the analysis. To track the progress of each team in the tournament the points table is stored in the database that keeps on updating itself as the matches are simulated. There will be 2 points table one for each pool after first round of tournament is simulated i.e group stage some sorting technique is used to extract top 4 teams from each pool based on the points gained by them during group stage simulation.

Following the tournament format next stage of simulation occurs now every time a simulation occurs it's outcome gives us the semi-finalists team. In this stage we will have 4 iterations that would eventually give us the top 4 semi finalists of the tournament. Based on the team's performance in the last year and half their attributes are stored in the database.

Scaled on 100 ,scaling is relative i.e a team that has been best in the last year or half is assigned highest points and then the second best team and so on. That is purely relative ranking and no two teams can have same value for a particular attribute. Now whenever a match is simulated the nodes whose prior probabilities need to be known are fetched from the database for both the teams & now relative comparisons for each require node is made using some algorithm, and then based on the results obtained a team gets high value for a particular node and later on further analysis takes place. The result of the analysis gives us the winner of the match.

Later on when we are done with predictive analysis we can use the same model for introspection of teams that couldn't do well at the tournament i.e what all factors contributed to their dismal performance and what all improvements they need to do in order to do well in the future .

1.1 Challenges

- Fixed sized hypothesis space
- May underfit or overfit the data
- May not contain any good classifiers if prior knowledge is wrong
- Harder to handle continuous features

1.2 Why Bayesian Network For Prediction?

Decision theory

As Bayesian networks are models of the problem domain probability distribution, they can be used for computing the predictive distribution on the outcomes of possible actions. This means that it is possible to use decision theory for risk analysis, and choose in each situation the action, which maximizes the expected utility. It can be shown that in a very natural sense, this is the optimal procedure for making decisions.

Consistent, theoretically solid mechanism for processing uncertain information

Probability theory provides a consistent calculus for uncertain inference, meaning that the output of the system is always unambiguous. Given the input, all the alternative mechanisms for computing the output with the help of a Bayesian network model produce exactly the same answer.

Flexible applicability

Bayesian networks model the problem domain as a whole by constructing a joint probability distribution over different combinations of the domain variables. This means

that the same Bayesian network model can be used for solving both discriminative tasks (classification) and regression problems (configuration problems and prediction). Besides predictive purposes, Bayesian networks can also be used for explorative data mining tasks by examining the conditional distributions, dependencies and correlations found by the modeling process.

CHAPTER 2

2. Statistical Background

2.1 Probabilistic Approach

The conditional probability of an event *B* is the probability that the event will occur given the knowledge that an event *A* has already occurred.

P(B | A) = P(A U B)/ P(A)

The probability of event A conditioned on knowing event B (or more shortly,the probability of A given B) is defined as

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

P(A|B) = P(A) P(B) (A,B are independent of each other)

2.2 Law of Total Probability

P(a) = S_b P(a, b)

= S_b P(a | b) P(b) where B is any random variable

Why is this useful?

given a joint distribution (e.g., P(a,b,c,d)) we can obtain any "marginal" probability (e.g., P(b)) by summing out the other variables, e.g.,

P(b) = S_a S_c S_d P(a, b, c, d)

Less obvious: we can also compute any conditional probability of interest given a joint distribution, e.g.,

$P(c \mid b) = S_a S_d P(a, c, d \mid b) = 1 / P(b) \; S_a S_d P(a, c, d, b)$

where $1 / P(b)$ is just a normalization constant

Thus, the joint distribution contains the information we need to compute any probability of interest.

We can always write

$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z) P(b, c, \ldots z)$(by definition of joint probability)

Repeatedly applying this idea, we can write

$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z) P(b \mid c,.. z) P(c| .. z)..P(z)$

This factorization holds for any ordering of the variables

This is the chain rule for probabilities

2.3 Conditional Independence Example

Conditional Independence example

Based on a survey of households in which the husband and wife each own a car, it is found that:

wife's car type⊥⊥husband's car type|family income

There are 4 car types, the first two being 'cheap' and the last two being 'expensive'. Using w for the wife's car type and h for the husband's:

$$p(w|inc = low) = \begin{pmatrix} 0.7 \\ 0.3 \\ 0 \\ 0 \end{pmatrix}$$

$$p(w|inc = high) = \begin{pmatrix} 0.2 \\ 0.1 \\ 0.4 \\ 0.3 \end{pmatrix}$$

$$p(h|inc = low) = \begin{pmatrix} 0.2 \\ 0.8 \\ 0.0 \\ 0.0 \end{pmatrix}$$

$$p(h|inc = high) = \begin{pmatrix} 0 \\ 0 \\ 0.3 \\ 0.7 \end{pmatrix}$$

$p(inc = low) = 0.9$

Then the marginal distribution p(w, h) is

p(w, h) = Xincp(w|inc)p(h|inc)p(inc)

giving

$$p(w, h) = \begin{pmatrix} 0.126 & 0.504 & 0.006 & 0.014 \\ 0.054 & 0.216 & 0.003 & 0.007 \\ 0 & 0 & 0.012 & 0.028 \\ 0 & 0 & 0.009 & 0.021 \end{pmatrix}$$

From this we can find the marginals and calculate

$$p(w)p(h) = \begin{pmatrix} 0.117 & 0.468 & 0.0195 & 0.0455 \\ 0.0504 & 0.2016 & 0.0084 & 0.0196 \\ 0.0072 & 0.0288 & 0.0012 & 0.0028 \\ 0.0054 & 0.0216 & 0.0009 & 0.0021 \end{pmatrix}$$

This shows that whilst w⊥⊥h|inc, it is not true that w⊥⊥h. For example, even if we don’tknow the family income, if we know that the husband has a cheap car

then his wife must also have a cheap car – these variables are therefore dependent.

CHAPTER 3

3 Bayesian Network

Bayesian networks (BNs), also known as belief networks(or Bayes nets for short), belong to the family of probabilistic graphical models(GMs). These graphical structures are used to represent knowledge about an uncertain domain. In particular, each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods. Hence, BNs combine principles from graph theory, probability theory, computer science, and statistics. GMs with undirected edges are generally called Markov random fields or Markov networks.

These networks provide a simple definition of independence between any two distinct nodes based on the concept of a Markov blanket. Markov networks are popular in fields such as statistical physics and computer vision. BNs correspond to another GM structure known as a directed acyclic graph (DAG) that is popular in the statistics, the machine learning, and the artificial intelligence societies. BNs are both mathematically rigorous and intuitively understandable. They enable an effective representation and computation of the joint probability distribution (JPD) over a set of random variables.

The structure of a DAG is defined by two sets: the set of nodes (vertices) and the set of directed edges. The nodes represent random variables and are drawn as circles labeled by the variable names. The edges represent direct dependence among the variables and are drawn by arrows between nodes. In particular, an edge from node Xi to node Xj represents a statistical dependence between the corresponding variables. Thus, the arrow indicates that a value taken by variable Xj depends on the value taken by variable Xi, or roughly speaking that variable Xi “influences” Xj. Node Xi is then referred to as a parent of Xj and, similarly,Xj is referred to as the child of Xi. An extension of these genealogical terms is often used to define the sets of “descendants” – the set of nodes that can be reached on a direct path from the node, or “ancestor” nodes – the set of nodes from which the node can be reached on a direct path . The structure of the acyclic graph guarantees that there is no node that can be its own ancestor or its own descendent. Such a condition is of vital importance to the factorization of the joint

probability of a collection of nodes as seen below. Note that although the arrows represent direct causal connection between the variables, the reasoning process can operate on BNs by propagating information in any direction.

A BN reflects a simple conditional independence statement. Namely that each variable is independent of its nondescendents in the graph given the state of its parents. This property is used to reduce,sometimes significantly, the number of parameters that are required to characterize the JPD of the variables. This reduction provides an efficient way to compute the posterior probabilities given the evidence. Such Graphical Models are now used as a standard framework in Engineering,Statistics and Computer Science.

A Bayesian network specifies a joint distribution in a structured form

Represent dependence/independence via a directed graph

- Nodes = random variables
- Edges = direct dependence

Structure of the graph ⇔ Conditional independence relations

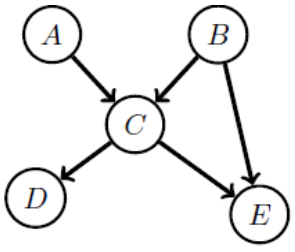
In general,

$p(X_1, X_2,....X_N) = P\ p(X_i \mid parents(X_i))$

Requires that graph is acyclic (no directed cycles)

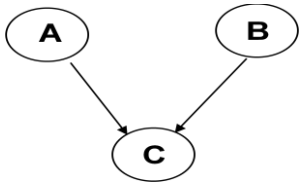
2 components to a Bayesian network

- The graph structure (conditional independence assumptions)
- The numerical probabilities (for each variable given its parents)



P(E|B, C)

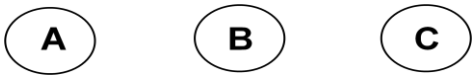
3.1 Example of a simple Bayesian network



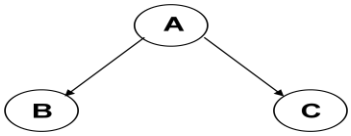
p(A,B,C) = p(C|A,B)p(A)p(B)

- Probability model has simple factored form
- Directed edges => direct dependence
- Absence of an edge => conditional independence
- Also known as belief networks, graphical models, causal networks
- Other formulations, e.g., undirected graphical models

3.2Dependence Independence



Marginal Independence:
p(A,B,C) = p(A) p(B) p(C)



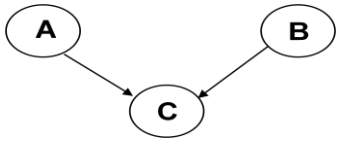
Conditionally independent effects:

p(A,B,C) = p(B|A)p(C|A)p(A)

B and C are conditionally independent

Given A

e.g., A is a disease, and we model B and C as conditionally independent symptoms given A



Independent Causes:

p(A,B,C) = p(C|A,B)p(A)p(B)

"Explaining away" effect:

Given C, observing A makes B less likely e.g., earthquake/burglary/alarm example

A and B are (marginally) independent but become dependent once C is known

Markov dependence:

p(A,B,C) = p(C|B) p(B|A)p(A)

3.3 Inference

It is the process of updating probabilities of outcomes based upon the relationships in the model and the evidence known about the situation at hand.

- Causal inference
- Diagnostic inference

3.4 Inference Example

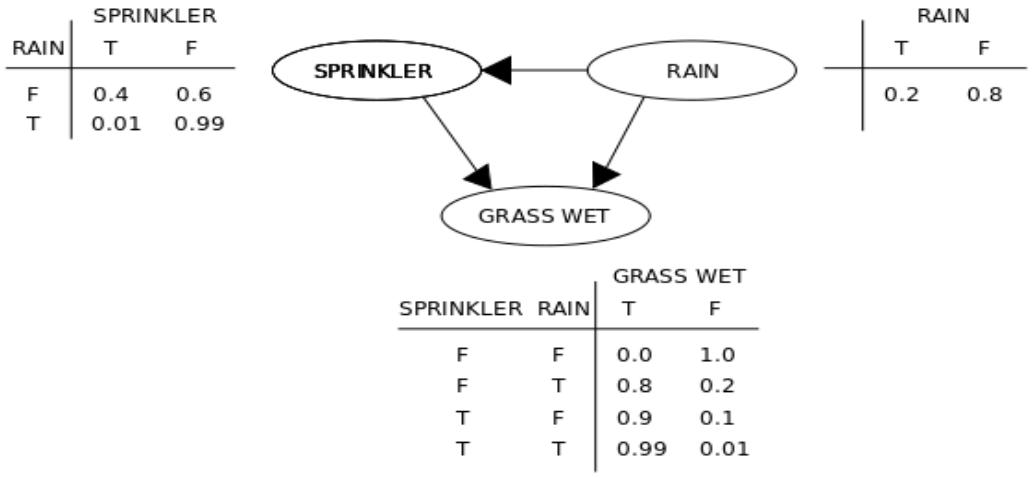


Fig 2.1

Joint probability of network

P (S, R, G) = P(G |S, R) * P (S | R) * P(R)

P(s= t,r = , g =t) = {(g= | s= t , r=t) * P (s=t | r =t) * p (r=t) = .00198

To compute:

$$P(R = T \mid G = T) = \frac{P(R = T, G = T)}{P(G = T)}$$

P (R = T, G= T) = ∑ ∑ P(R = T, G = T , S) = P(R = T, G = T, S = **T**) + P(R = T, G = T, S = **F**)

= P(G= T | S = T , R = T) * P(S = T | R = T) * P(R = T) + P(G= T | S = F , R = T) * P(S = F| R = T) * P(R = T)= 0.16038

P (G = T) = ∑ ∑ P (R , S, G = T)

= P (R = T, S = T , G =T) + P(R = F, S = T, G= T) + P (R = T, S = F , G =T) + P(R = F, S = F , G =T) =0 .44838

$$P(R = T \mid G = T) = \frac{0.16038}{0 .44838} = 35.77\ \%$$

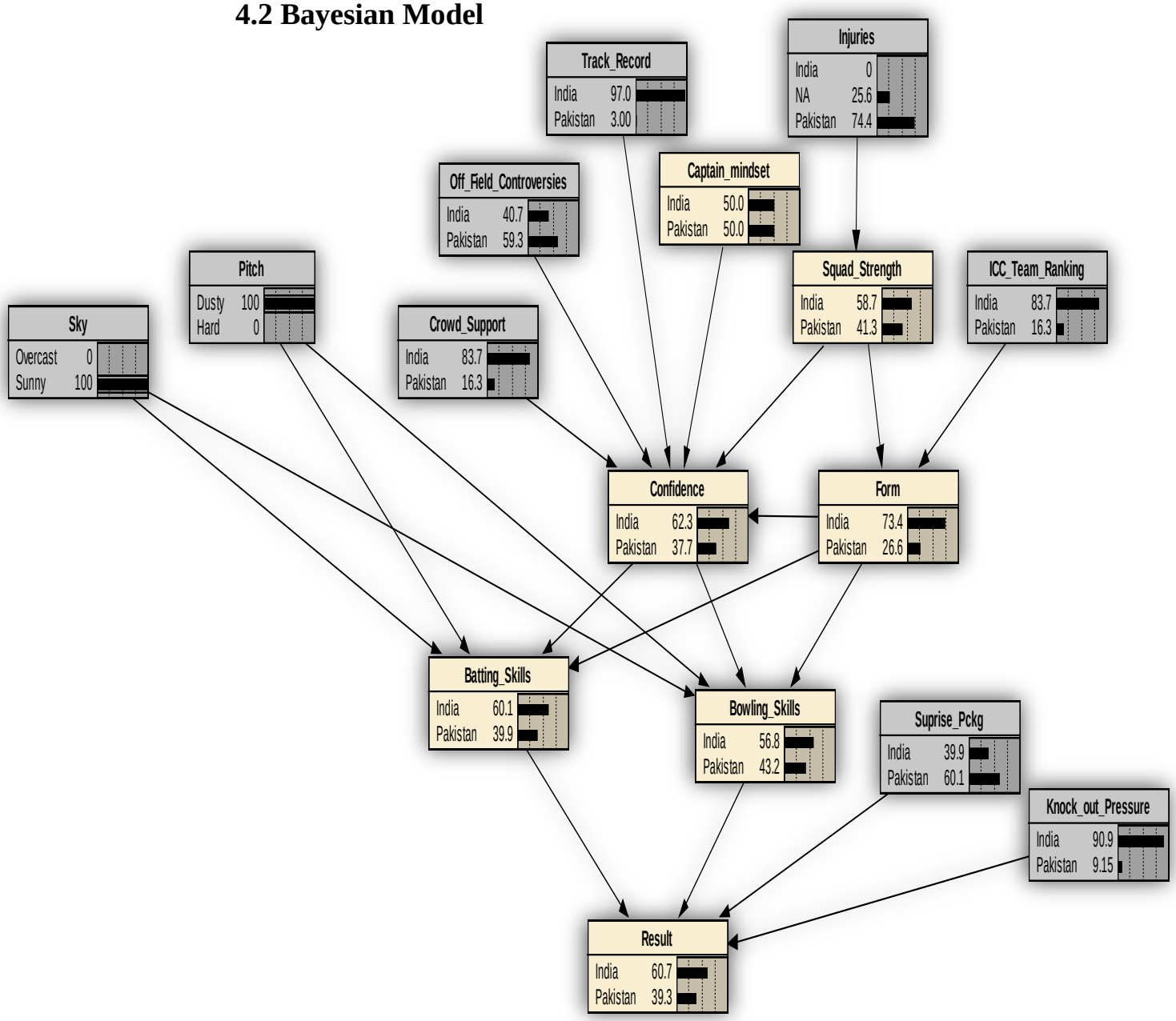
Chapter 4

4. Prediction Of A Cricket Match Using Bayesian Network

4.1Data Points

- Track Record-That is no. of times the teams have played against each other in the world cup.
- Injuries-Denotes players missing from the squad for the match due to injuriy issues.
- Off Field Controversies-If a team is in the news for the wrong reasons.
- Squad Strength-On paper which team stands better.
- ICC Team Rankings-Based on ranking both the teams are scaled on 100.
- Crowd Support-Which team is more likely to get crowd support keeping mind the home advantage.
- Confidence-Cricket is a game of confidence hence a very important factor in determining the result of the match
- Form-A current run of success or the purple patch
- Batting And Bowling Skills-Shows how good a team is in bowling and batting department
- Knock out pressure-Shows which team has better temperament of playing knock out games
- Result-Ouptut of the analysis

4.2 Bayesian Model



The above model was tested on ICC Cricket World Cup 2011 semifinal match between India And Pakistan held at Mohali ,India.

For that from the available data these were the values of particular node at that point of time

1)Track Record @CWC(1992-2011)

India-4

Pakistan-0

Since India hai 100 % track record against Pakistan at the world cups hence probability of India in Track Record(97) goes high as compare to Pakistan.

Source-Wisden Stats

2)ICC Ranking during 2011 WC

India-2nd out of 13 cricket playing nations

Pakistan-6th out of 13 cricket playing nations

Source(Reliance ICC Rankings)

Hence by applying the concept of ratios to calculate relative ranking between 2 as

India:Pakistan=2:6=1:3

India=(3/4)*100=75 %

Pakistan=(1/4)*100=25%

=>Relative wise India leads Ranking as compare to Pakistan

3)Crowd Support

Since the match took place in India the crowd support for India is high as compare to Pakistan

Out of 45000 people approx 40000 were India supporters while the rest were Pakistani supporters.

Hence prob for India in Crowd Support Goes high to 88%

4)Injuries

At the time of Match against Pakistani key fast bowler Shoaib Akthar announced his retirement from cricket due to injuries issues while India has full fit squad hence Injury parameter for Pakistan state goes high

Source(ICC news website)

Pakistan-70%

India-30%

5)Off Field Controversies

Shoaib Akthar's retirement in the middle of the tournament takes a beating on the confidence of the team admitted by his team members at the press conference,being the premiere fast bowler on which the entire team banks upon.Hence Pakistan state for Off field Controversies goes high.

Pakistan-60%

India-40%

Source(ICC news website)

6)Form

Before this match India & Pakistan have played 7 matches each

Pakistan had won 6 out of 7

While India had won 5 out of 7

Source(ICC WorldCup Points Table)

Again using the concept of ratios to calculate relation between two

Pakistan:India=6:5

Pakistan=(5/13)*100=54%

India=46%

7)Surprise Package

Since Pakistan possesses some players that can surprise any opposition on the given day be it some mystery spinner or some flashy batsman that can tear apart any opposition bowler on a given day hence Pakistan state is high for Surprise Package as compare to india

Pakistan-60.01

India-39.99

4.3 Condtional Probabilty Table For Nodes

Crowd_Support	Off_Field_Controversies	Captain_mindset	Squad_Strength	Track_Record	Form	India	Pakistan
India	India	India	Pakistan	Pakistan	Pakistan	34	66
India	India	Pakistan	India	India	India	34	66
India	India	Pakistan	India	India	Pakistan	50	50
India	India	Pakistan	India	Pakistan	India	50	50
India	India	Pakistan	India	Pakistan	Pakistan	34	66
India	India	Pakistan	Pakistan	India	India	50	50
India	India	Pakistan	Pakistan	India	Pakistan	34	66
India	India	Pakistan	Pakistan	Pakistan	India	34	66
India	India	Pakistan	Pakistan	Pakistan	Pakistan	17	83
India	Pakistan	India	India	India	India	95	5
India	Pakistan	India	India	India	Pakistan	83	17
India	Pakistan	India	India	Pakistan	India	83	17
India	Pakistan	India	India	Pakistan	Pakistan	66	34
India	Pakistan	India	Pakistan	India	India	83	17
India	Pakistan	India	Pakistan	India	Pakistan	66	34
India	Pakistan	India	Pakistan	Pakistan	India	66	34
India	Pakistan	India	Pakistan	Pakistan	Pakistan	50	50
India	Pakistan	Pakistan	India	India	India	66	34
India	Pakistan	Pakistan	India	India	Pakistan	66	34
India	Pakistan	Pakistan	India	Pakistan	India	66	34
India	Pakistan	Pakistan	India	Pakistan	Pakistan	50	50
India	Pakistan	Pakistan	Pakistan	India	India	66	34
India	Pakistan	Pakistan	Pakistan	India	Pakistan	50	50
India	Pakistan	Pakistan	Pakistan	Pakistan	India	50	50
India	Pakistan	Pakistan	Pakistan	Pakistan	Pakistan	34	66
Pakistan	India	India	India	India	India	66	34
Pakistan	India	India	India	India	Pakistan	50	50

Table 4.1
Confidence Node

Injuries	India	Pakistan
India	48	52
NA	55	45
Pakistan	60	40

Table 5.2
Squad Strength node

Pitch	Form	Confidence	Sky	India	Pakistan
Dusty	India	India	Overcast	70	30
Dusty	India	India	Sunny	75	25
Dusty	India	Pakistan	Overcast	55	45
Dusty	India	Pakistan	Sunny	55	45
Dusty	Pakistan	India	Overcast	45	55
Dusty	Pakistan	India	Sunny	45	55
Dusty	Pakistan	Pakistan	Overcast	35	65
Dusty	Pakistan	Pakistan	Sunny	30	70
Hard	India	India	Overcast	70	30
Hard	India	India	Sunny	75	25
Hard	India	Pakistan	Overcast	55	45
Hard	India	Pakistan	Sunny	50	50
Hard	Pakistan	India	Overcast	45	55
Hard	Pakistan	India	Sunny	45	55
Hard	Pakistan	Pakistan	Overcast	35	65
Hard	Pakistan	Pakistan	Sunny	30	70

Table 4.2
Batting Skills Node

Pitch	Form	Confidence	Sky	India	Pakistan
Dusty	India	India	Overcast	60	40
Dusty	India	India	Sunny	63	37
Dusty	India	Pakistan	Overcast	58	42
Dusty	India	Pakistan	Sunny	60	40
Dusty	Pakistan	India	Overcast	40	60
Dusty	Pakistan	India	Sunny	45	55
Dusty	Pakistan	Pakistan	Overcast	35	65
Dusty	Pakistan	Pakistan	Sunny	40	60
Hard	India	India	Overcast	55	45
Hard	India	India	Sunny	50	50
Hard	India	Pakistan	Overcast	50	50
Hard	India	Pakistan	Sunny	55	45
Hard	Pakistan	India	Overcast	45	55
Hard	Pakistan	India	Sunny	42	58
Hard	Pakistan	Pakistan	Overcast	30	70
Hard	Pakistan	Pakistan	Sunny	35	65

Table 4.3
Bowling Skills Node

Batting_Skills	Bowling_Skills	Suprise_Pckg	Knock_out_Pressure	India	Pakistan
India	India	India	India	90	10
India	India	India	Pakistan	75	25
India	India	Pakistan	India	75	25
India	India	Pakistan	Pakistan	50	50
India	Pakistan	India	India	75	25
India	Pakistan	India	Pakistan	50	50
India	Pakistan	Pakistan	India	50	50
India	Pakistan	Pakistan	Pakistan	25	75
Pakistan	India	India	India	75	25
Pakistan	India	India	Pakistan	50	50
Pakistan	India	Pakistan	India	50	50
Pakistan	India	Pakistan	Pakistan	25	75
Pakistan	Pakistan	India	India	50	50
Pakistan	Pakistan	India	Pakistan	25	75
Pakistan	Pakistan	Pakistan	India	25	75
Pakistan	Pakistan	Pakistan	Pakistan	10	90

Table 4.4
Result Node

4.3 Results of outcome

By simulating India -Pakistan world cup 2011 semifinal match we get the following results in favour of India the values of different nodes are kept keeping in mind the cricketing status of both the teams at that point of time

India-60.01%

Pakistan-39.99 %

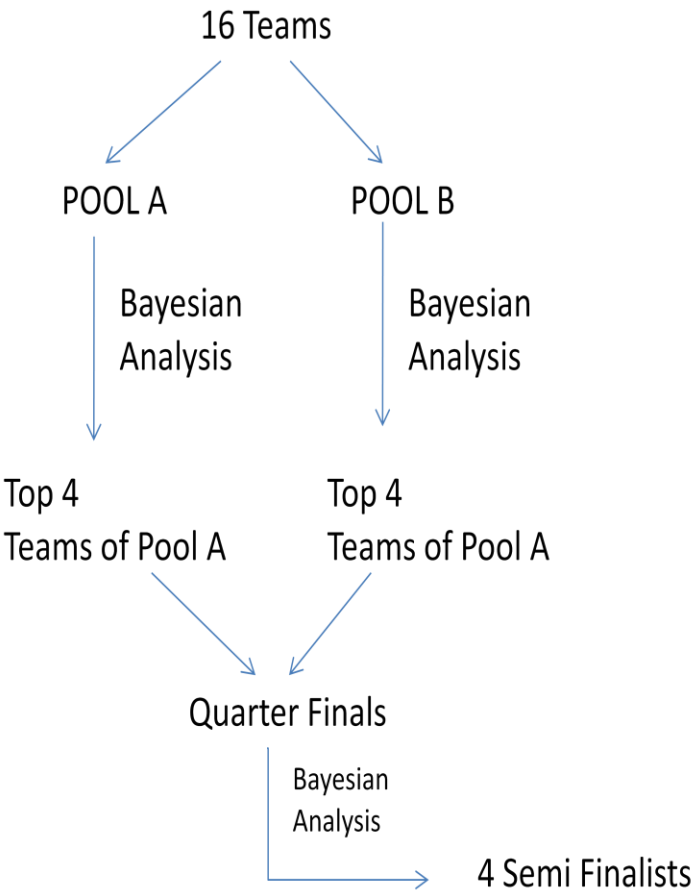
That also verifies the actual result that india has indeed defeated Pakistan in 2011 world cup semifinal match

CHAPTER 5

5.Predicting ICC Cricket 2015 winners

A Slight modification in the previous network can be used in prediction of the winner of ICC cricket world cup 2015

5.1Tournament Format



POOL A
ENGLAND
AUSTRALIA
NEW ZEALAND
SRI LANKA
BANGLADESH
AFGHANISTAN
SCOTLAND

TABLE 5.1

POOL B
SOUTH AFRICA
INDIA
PAKISTAN
WEST INDIES
ZIMBABWE
IRELAND
UNITED ARAB EMIRATES

TABLE 5.2

HOST NATION-AUSTRALIA AND NEW ZEALAND

Top 4 teams from each pool proceeds to quarterfinals

In this section, we detail Bayesian approach for predicting winners of 2015 Cricket World Cup quarterfinals, semifinals and finals. We first review known facts about 2015 Cricket World Cup, and later detail what we added to these facts in order to make predictive model. It is given that 14 teams will participate in the event . These 14 teams are further divided into two pools namely, pool A and pool B. Each pool consists of 7 teams. In each pool, every team will play against each other, resulting in total of 42 matches. This phase of matches is called as the group stage of the tournament. The top 4 teams from each pool progresses to quarterfinals based on the highest number of points they gain on account of victory against teams of their pools. The format for quaterfinals describing which team from Pool A will play against which team Pool B is also given.

In addition to this information, we also know that all matches concerning 2015 Cricket World Cup will be held in Australia and New Zealand.

We now present information which is not directly available about 2015 Cricket World Cup, but required for prediction purposes. First and foremost information required for

making prediction is cricket statistics. By cricket statistics we mean factors or features that can be used to judge teams on their performances during the event. We divide these statistics in two categories. We name, category one as independent features while, second category as dependent features. Independent features simply mean that they are not influenced by any external factors. De-pendent features on the other hand depends upon on one or more features. We collected in total 15 features, out of which 9 are independent and rest dependent.

After selecting these features, we then collected corresponding data. However, such data is not readily available to be used. So, we follow few data pre-processing steps in order to produce required facts. These pre-processing steps are only re-quired for the independent features. We can infer the values of dependent variables from independent variables, which we will discuss in detail shortly. For now we concentrate only on pre-processing steps involve in 9 independent features. team used in this work.

In order to find values for variable crowd support for each country, we use information available on Wikipedia . Since most of the matches are in Australia so, Australia is benefited from the maximum support of the crowd. To calculate crowd support for other countries, we studied their population in Australia on Wikipedia. According to Wikipedia, for example, 1.6% of the Australian population are Indians. A similar study is done for remaining countries. Based on these statistics we compute crowd support for each country on a scale of 100.

Finding ranking of each country on cricket, sport is easy. ICC maintains the rank of each country on the bases of their performances in cricket. According to the ICC top four best cricket teams are: Australia, India, South Africa and Sri Lanka. To decide on batting and bowling forms of teams, we follow a number of pre-processing steps. Since form refers to the performances in the recent time, so, we have concentrated only on the last 25 matches of the squad selected, i.e. performance in the last year and half for the world cup with matches considered till January 9, 2015. In case of batting form, we have considered three parameters namely, matches played, runs Scored and batting average. As per general knowledge, a good batsman must have good experience, i.e.,decent number of matches under his belt (in our case maximum matches can only be 25). He should have

scored decent number of runs during this time, and at the same time must maintain a good batting average throughout, i.e., consistency. Hence we can conclude that

$$\text{Good Batsman} \propto \text{Experience}$$
$$\propto \text{Runs Scored}$$
$$\propto \text{Batting Average}$$

$$\text{Good Batsman} = K(\text{Experience}) \times (\text{Runs Scored}) \times (\text{Batting Average})$$

Likewise for bowling Form we have considered three parameters such as Matches played, Wickets taken, bowling average.

$$\text{Good Bowler} \propto \text{Experience}$$
$$\propto \text{Wickets taken}$$
$$1/\propto \text{Bowling Average}$$

$$\text{Good Bowler} = K\frac{(\text{Experience}) \times (\text{Wickets taken})}{\text{Bowling Average}}$$

We used this formula to consider young cricket players participating in the competition for the very first time. These fresh players have not gained enough experience, hence ranking them just on the basis of runs scored or average during their short stint in one-day cricket may not give a true reflection of good batsman. Suppose a player has just played 5 matches and holds a batting average of 85 would top our list of batsmen if the list would be sorted or ranked solely on the basis of batting average, that would overshadow the performance of other batsmen that has performed consistently over a period of time.

Now the list of batsmen and bowlers based on the above formulae are sorted and then are ranked accordingly. Lastly, each player is given percentile value, i.e. on a scale of 100 how much better a particular player is than other players based on the rank so obtained. Bowling and batting form of a team is the average of all the percentiles of batsmen and bowlers of a particular squad. Team ranking, track record, and injury list, are extracted from ICC and espncricinfo website.All these features for different teams are scaled in 100.

After collecting the data, we establish relationship between independent and de-pendent features based on common knowledge of cricket domain and modeled Bayesian graphical structure. For each match scheduled in the tournament, we extract corresponding information related to 9 independent features in Bayesian network from table for teams participating in the match. Each independent feature is a binary state variable. Each state represents a team participating in the match. For example, if a match is between Australia and India, then each independent variable will have two states namely, Australia and India. Each state is given a probability indicating strength of the team. Since dependent features are modeled in Bayesian network so, sum of probability of states should sum one. So we used the simple ratio proportional theory to evaluate probability of one team over the other on a given independent feature.For example, let crowd support for Team A and Team B be x: y where, x and y are extracted from table 4 for teams A and B. Then, value of Team A in crowd support will be ((x / (x + y)) * 100), and ((y/ (x +y)) * 100) for Team B. This method is followed for each independent variable. To calculate cond probability of teams at independent node, we simply calculate the number of times respective teams are supported by parent variables of dependent node. Consider a dependent node with two parents. Let for each parent, Team A is given high probability. This causes the child node with Team A to go with higher probability than Team B.

Based on tournament schedule, we simulate 42 predictions for group-stage, 4 predictions for quarterfinals, 2 predictions for semi-finals and final winner of the competition.

5.3 Bayesian Model For Cricket World Cup 2015 Matches

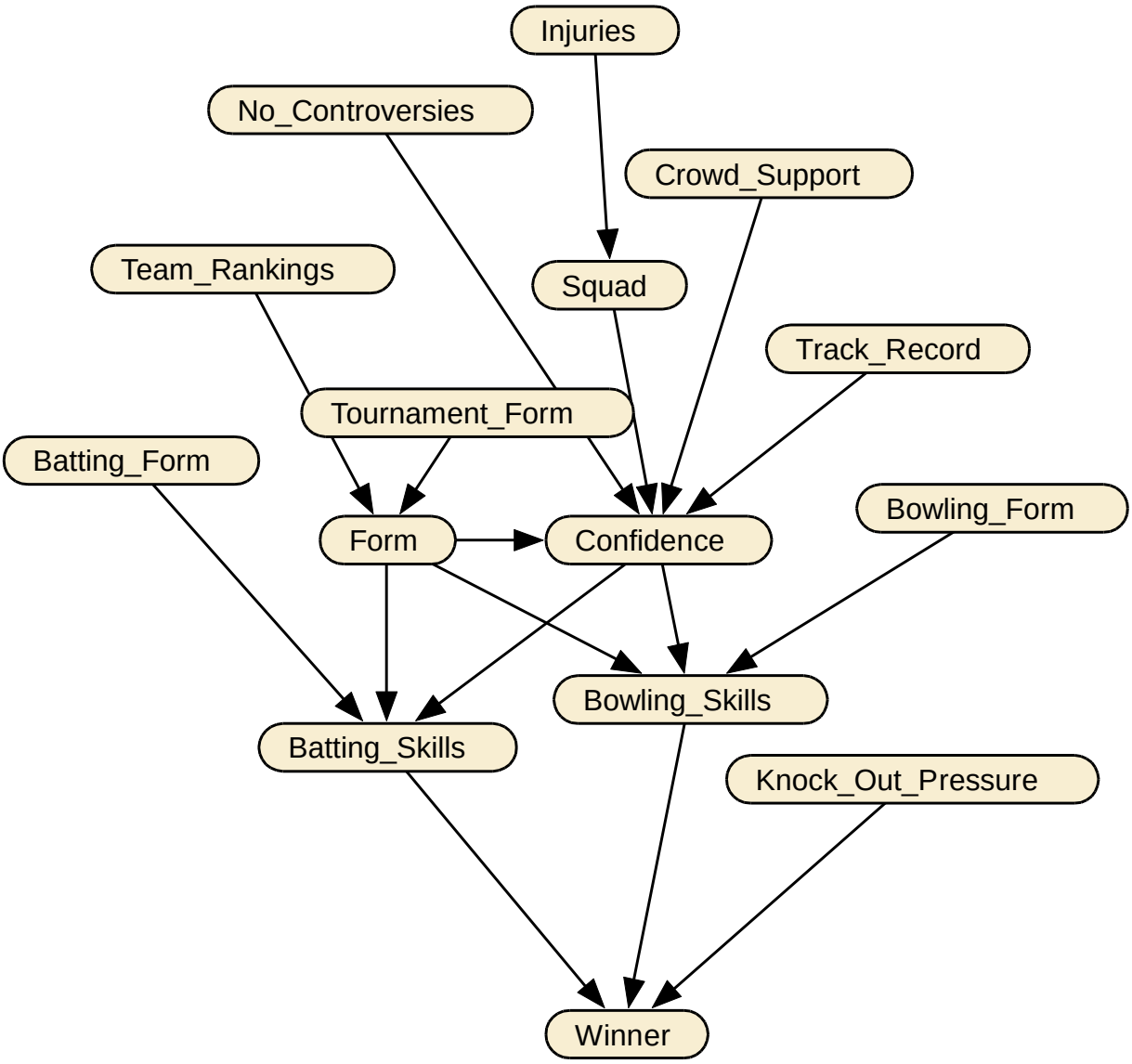


Fig 5.2 Revised Model

5.4 Changes from previous model

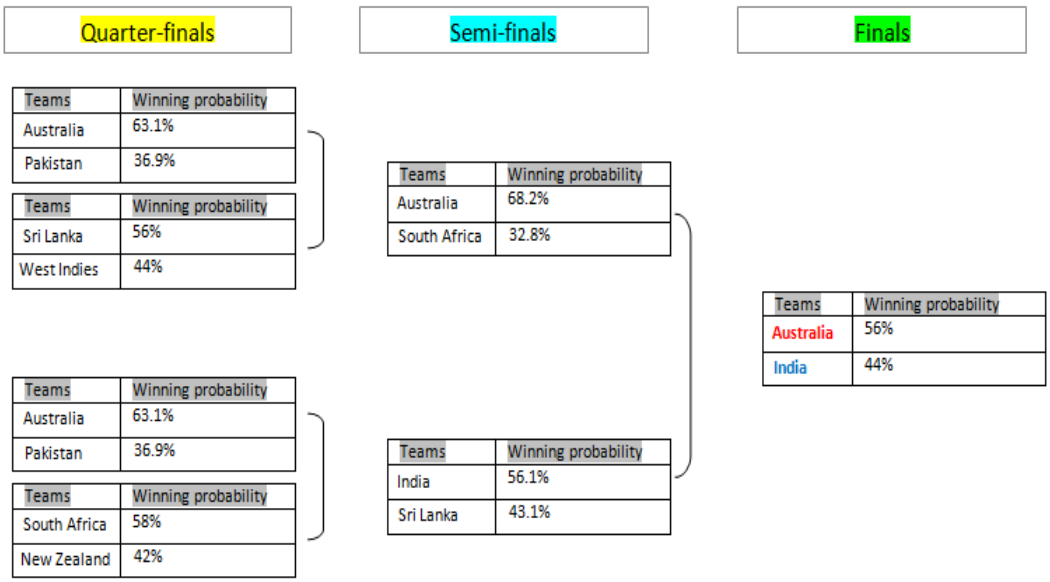
Introduction of two new nodes called batting and bowling form , these nodes will take take care of team's recent performance in the batting and bowling department.I have considered past 25 matches of players and based on their performances players have been rated the most successful batsman or bowler during that time is awarded maximum points while the least performer is awarded lowest points.

Also data points like pitch and ground are removed from this model due to lack of availability of data for the tournament.

5.5 Results

A)Blind prediction

Prediction made prior to world cup



B)Prediction based on current world cup status

The actual result of the group stage of current world cup is taken into considerations and predictions for further round of the tournament is made

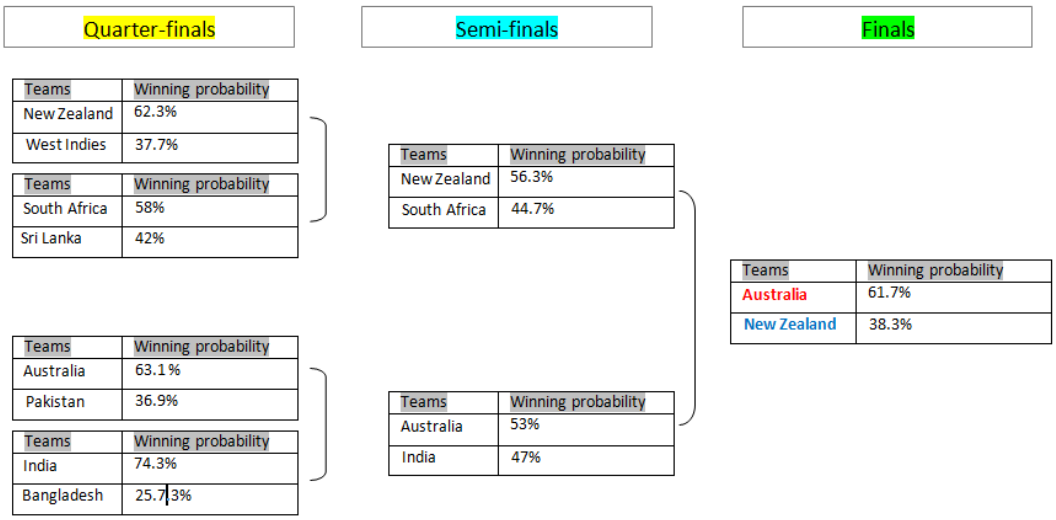


Fig 5.4

Chapter 6

6 Illustration

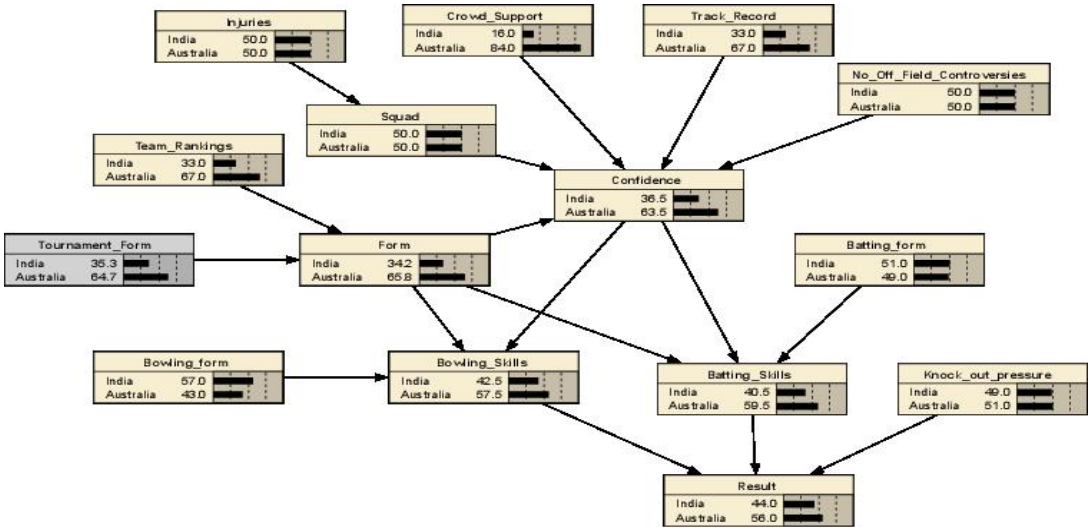


Fig 6.1 Final Model

Above is the Bayesian model for the world cup final featuring top two cricketing giants Australia and India. With the available data for the final the analysis shows Australia winning the final with a margin of 56 -44 against the defending champions India.
We will now focus on factors that would enable teams like India to do well in this event.

As a first example, let us see how a high proportion of Confidence in favour of India affected the final outcome

As shown in the chart below, we select the 'India' option in the node for Confidence. The node turns gray in colour and the probabilities in the Result node changes**.**

For example, the result node for the Australia comes down from 56% to 41.70% and for India it goes up from 44% to 58.3%

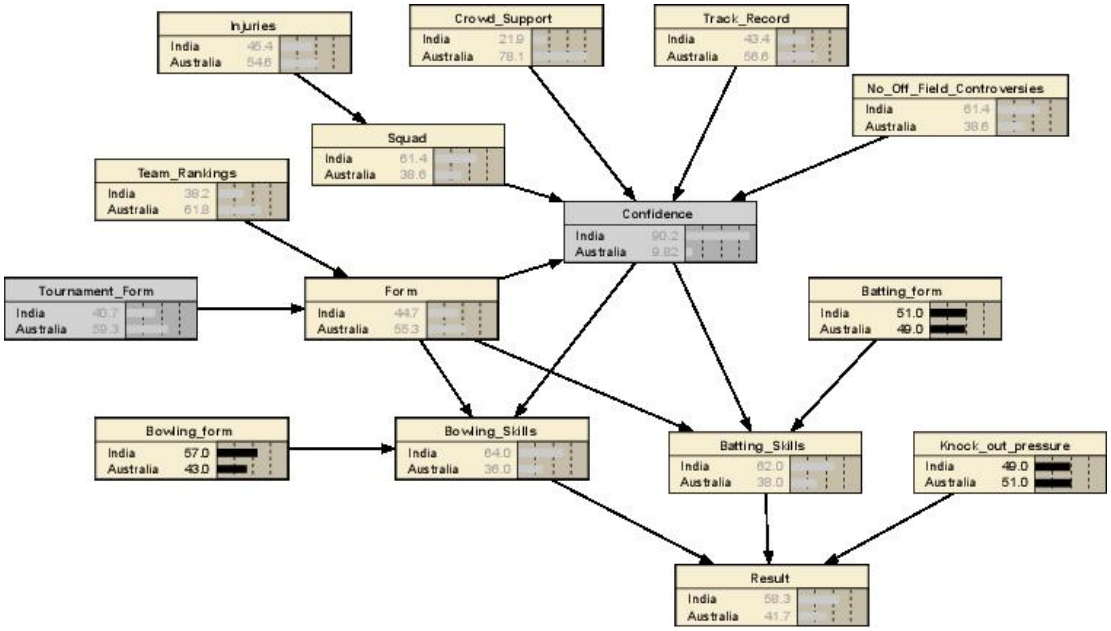


Fig 6.2 Role of Confidence

This is a clear indication that confidence factor plays a key role in the night of the final. But for teams to have high confidence on their side teams they need to have injuries free squad, decent support from crowd, good form throughout the competition, mustn't involve with controversies, and a good track record against opposition playing against. All these factors favouring a team will result in high confidence.

How did 'form' affect the outcome of final?

When we select 'India' option in the form node in the graph below . Australia's share in the result node goes down from 56 % to 41.7% while India's share rises from 44% to 59.3%.

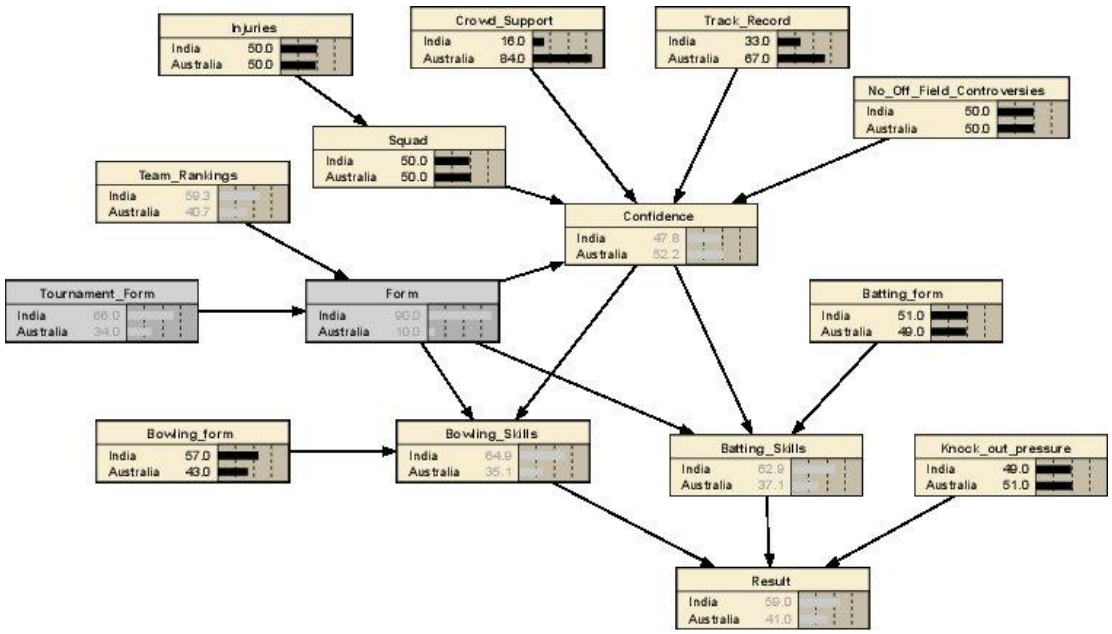
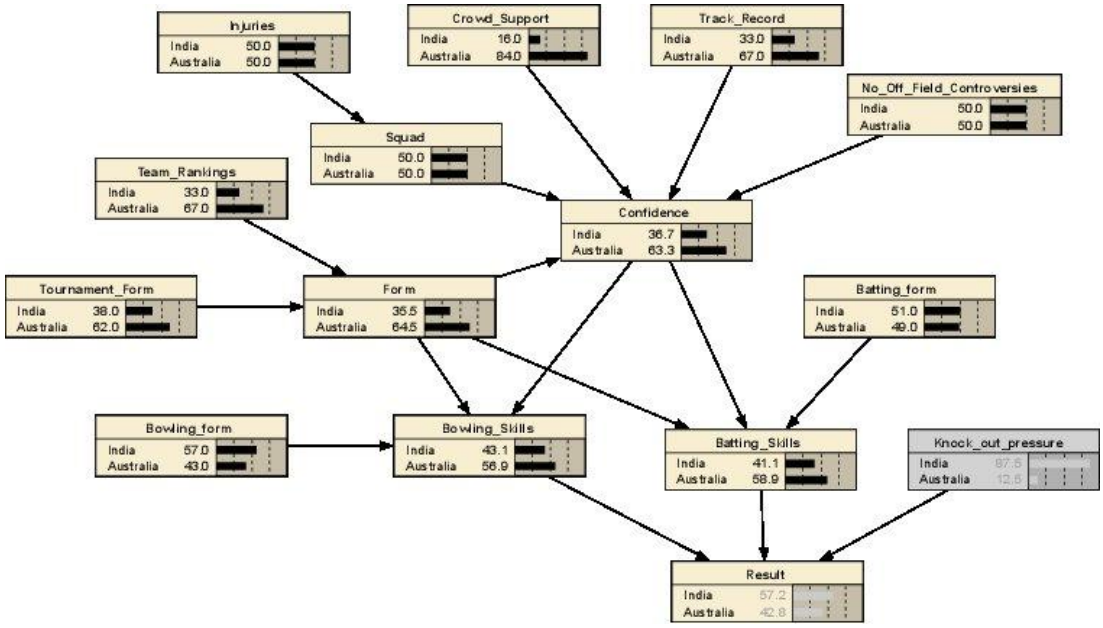


Fig 6.3 Role of Form

This provides a clear evidence that form plays a crucial factor in determining outcome of the final. The better momentum a team carries into the final, accordingly it's batting and bowling skills gets amplified and result goes in their favour.

Role of knock out pressure in the final

When we select 'India' option in knock out pressure node, the outcome i.e result node rises to 57.2% in favour of India while Australia's share drops to 42.8%



This shows along with batting and bowling skills a team needs to have ability to absorb pressure in tense final that is the ability to perform up to their true potential rather than succumbing to pressure even with good batting and bowling skills. A team that has good temperament to play knock out games as compare to it's opponent which generally struggles in such situations is most likely win the big final.

6.2 Sample of data used

Batsmen	Runs	Avg	Matches	Rank	Points
R sharma	1026	51.23	25	1	1314050
H Amla	1093	45.14	25	2	1233451
Q De Cock	1084	45.16	25	3	1223836
V kohli	964	45.9	25	4	1106190
Shehzad	1034	41.69	25	5	1077687
AB De Villiers	901	47.42	25	6	1068136
A Finch	1006	40.24	25	7	1012036
K Sangakkara	986	41	25	8	1010650
K willamson	893	44.65	25	9	996811.3
S Dhawan	949	41.2	25	10	977470
Faf Du Plesis	982	39.28	25	11	964324
T Dilshan	940	39.1	25	10	918850
S Watson	886	36.5	25	13	808475
george Bailey	866	36.7	25	15	794555
A Raydu	685	45	24	23	739800
R Taylor	751	35.61	25	18	668577.8
M Clarke	737	35.05	25	20	645796.3
I Bell	785	32	25	16	628000
J Root	741	33.6	25	19	622440
M Jayawardene	869	28	25	14	608300
A Rahane	776	31.04	25	17	602176
A Matthews	625	36	25	31	562500
D Warner	732	30.5	25	21	558150
Darren Bravo	727	30.29	25	22	550520.8

Sheet1 Sheet2 Sheet3

Table6.1 List of Batsmen

Bowlers	Matches	Wickets	Avg	Points	Rank
D Steyn	25	49	18.8	65.15957	1
M McCLENAGH	27	55	25.43	58.3956	2
J Anderson	25	41	21.19	48.37187	3
M starc	25	43	23.04	46.65799	4
I Tahir	25	40	23.37	42.7899	5
M Morkel	25	42	26.5	39.62264	6
S lakmal	25	40	25.6	39.0625	7
M Johnson	25	39	26.64	36.5991	8
J Taylor	25	37	27.62	33.49022	9
J khan	38	25	28.6	33.21678	10
K Roach	25	36	29.2	30.82192	11
herath	25	27	22.08	30.57065	12
Ishant	25	35	32.77	26.70125	13
shammi	25	35	32.77	26.70125	14
R Jadeja	25	36	34.08	26.40845	15
U Yadav	25	33	32.03	25.7571	16
Sennayke	25	31	30.87	25.10528	17
C woakes	24	34	32.55	25.06912	18
J Faulknar	25	33	33.64	24.52438	19
S Broad	25	33	33.75	24.44444	20
M Irfan	32	25	32.96	24.27184	21
V Philander	21	27	23.66	23.9645	22
W Parnell	25	28	30.03	23.31002	23
R Ashwin	25	32	34.48	23.20186	24

Table6.2 List of Bowlers

Table 6.3 Squad

Batting	Squad									Bowling	Squad					
Nz	Runs	Avg	Matches	Ranking	Percentile						wicks	match	Avg	Ranking	Percentile	
Guptill	508	24.2	25	38	43.28358209					K Mills	21	25	26.78	32	42.85714286	
Tom Latham	389	24.31	20	51	23.88059701					M McCLENAGH	55	27	25.43	2	96.42857143	
B Mcullum	503	24	25	39	41.79104478					Vettori	25	25	34.96	36	35.71428571	
K willamson	893	44.65	25	9	86.56716418					N Mcullum	19	25	61.57	42	25	
R Taylor	751	35.61	25	16	76.11940299					C Anderson	25	18	28.32	38	32.14285714	
C Anderson	261	20.04	18	57	14.92537313					A Milne	5	10	69.2	54	3.571428571	
G Elliot	448	20.61	25	41	38.80597015					T Boult	12	10	49.4	51	8.928571429	
L.Ronchi	496	27.77	25	37	44.7761194										Bow Form avg	35
						46										
Pak																
Misbah	559	26.61	25	34	49.25373134											
Shehzad	1034	41.69	25	5	92.53731343					J khan	25	38	28.6	10	82.14285714	
Y Khan	450	19.56	25	42	37.31343284					W Riaz	25	23	46.16	39	30.35714286	
Hafeez	676	3.72	25	25	62.68656716					S Afridi	25	27	37.07	35	37.5	
Afridi	410	20.5	25	47	29.85074627					M Irfan	25	32	32.96	21	62.5	
U akmal	402	21.15	25	46	31.34328358					S Khan	5	6	33.16	53	5.357142857	
Sarfraz	274	19.57	25	54	19.40298507											
S Maqsood	572	35.75	18	37	44.7761194					Y Shah	1	2	25.5	55	1.785714286	
					batt form avg	46									Bow Form avg	37

Track Record

India	Matches	Won		Pakistan	Matches	Won		South Africa	Matches	Won		West Indies	Matches	Won
vs				vs				vs				vs		
pak	5	5		ind	5	0		ind	3	3		ind	7	2
saf	3	0		saf	3	0		pak	3	3		pak	9	6
win	7	5		win	9	3		win	5	3		saf	5	2
zim	8	7		zim	2	2		zim	2	1		zim	3	3
re	1	1		ire	1	0		ire	1	1		ire	1	1
uae	0	0		uae	0	0		uae	1	1		uae		
eng	6	3		eng	6	4		eng	6	3		eng	5	1
aus	9	3		aus	7	4		aus	6	3		aus	9	5
sri	7	4		sri	4	4		sri	3	2		sri	6	4
nz	4	1		nz	7	5		nz	5	4		nz	6	3
bang	1	1		bang	1	1		bang	3	2		bang	3	3
scot	0	0		scot	1	1		scot	0	0		scot	0	0

Table 6.4 Track Record

Teams	Crowd_ Support	Rank	Batting_ Form	Bowling_ Form	No_ Controversies	Injuries	Knock _Out_Pressure
IND	1.6	2	68	57	50	50	70
PAK	0.2	6	46	37	30	70	50
RSA	0.7	3	54	66	50	50	0
SL	0.48	4	50	53	50	50	68
WIN	0.02	7	39	53	30	50	60
ENG	5.1	5	41	55	50	50	27
BAN	0.1	9	15	15	50	50	0
AUS	91.98	1	65	43	50	50	0
NZ	85	8	46	45	50	50	14
ZIM	0.1	10	10	10	50	50	0
SCOT	1 0.1	13	10	10	50	50	0
IRE	0.3	12	10	10	50	50	0
UAE	0.1	14	10	10	50	50	0
AFGH	0.1	11	10	10	50	50	0

Table 6.5 Independent features for various teams

6.3 Implementation snapshot

```
import norsys.netica.*;
public class DoInference {

  public static void main (String[] args){
    try {

        System.out.println("Enter 2 teams");


        Scanner sc=new Scanner(System.in);
        String a=sc.nextLine();
        String b=sc.nextLine();

    Node.setConstructorClass ("norsys.neticaEx.aliases.Node");
    Environ env = new Environ (null);

    Net net = new Net();
    net.setName("CWC");

    Node track    = new Node ("Track_Record",   "TeamA,TeamB",  net);
    Node squad = new Node ("Squad",   "TeamA,TeamB",  net);
    Node inj      = new Node ("Injuries",   "TeamA,TeamB",  net);
    Node rank       = new Node ("Team_Rankings",   "TeamA,TeamB",  net);
    Node contro       = new Node ("Off_Field_Controversies",   "TeamA,TeamB",  net);
    Node  crowd      = new Node ("Crowd_Support",    "TeamA,TeamB",  net);
    //Node cap      = new Node ("Captain_Mindset",   "TeamA,TeamB",  net);
    Node form   = new Node ("Form",   "TeamA,TeamB",  net);
    Node conf   = new Node ("Confidence",   "TeamA,TeamB",  net);
    Node bat   = new Node ("Batting_Skills",   "TeamA,TeamB",  net);
    Node bow   = new Node ("Bowling_Skills",   "TeamA,TeamB",  net);
    Node sup   = new Node ("Suprise_Element",   "TeamA,TeamB",  net);
    Node result   = new Node ("Result",   "TeamA,TeamB",  net);
```

```
    Node  ground   = new Node ("Ground",   "Large , Small",  net);
    Node pitch   = new Node ("Pitch",   "Green , Normal",  net);

    squad.state("TeamA").setTitle (a);
    squad.state("TeamB").setTitle (b);
    inj.state("TeamA").setTitle (a);
    inj.state("TeamB").setTitle (b);
    rank.state("TeamA").setTitle (a);
    rank.state("TeamB").setTitle (b);
    contro.state("TeamA").setTitle (a);
    contro.state("TeamB").setTitle (b);
    crowd.state("TeamA").setTitle (a);
    crowd.state("TeamB").setTitle (b);
    track.state("TeamA").setTitle (a);
    track.state("TeamB").setTitle (b);
    form.state("TeamA").setTitle (a);
    form.state("TeamB").setTitle (b);
    conf.state("TeamA").setTitle (a);
    conf.state("TeamB").setTitle (b);
    bat.state("TeamA").setTitle (a);
    bat.state("TeamB").setTitle (b);
    bow.state("TeamA").setTitle (a);
    bow.state("TeamB").setTitle (b);
    sup.state("TeamA").setTitle (a);
    sup.state("TeamB").setTitle (b);
    result.state("TeamA").setTitle (a);
    result.state("TeamB").setTitle (b);
:onf.addLink(crowd);
:onf.addLink(form);
:onf.addLink(contro);
:onf.addLink(track);
```

Fig 6.4

```
bow.addLink(pitch);
bow.addLink(ground);
bow.addLink(sky);
bow.addLink(conf);
bow.addLink(form);
result.addLink(bat);
result.addLink(bow);
result.addLink(sup);

track.setCPTable (0.70, 0.30);
inj.setCPTable(0.70,0.30);

contro.setCPTable (0.70, 0.30);
crowd.setCPTable(0.80,0.20);
pitch.setCPTable(0.50,0.50);
rank.setCPTable(0.83,0.17);
sky.setCPTable(0.50,0.50);
ground.setCPTable(0.50,0.50);

squad.setCPTable ("TeamA",0.7, 0.3);
squad.setCPTable ("TeamB" , 0.3, 0.7);

form.setCPTable ("TeamA",0.7, 0.3);
form.setCPTable ("TeamB" , 0.3, 0.7);

int []ps=new int[5];
ps[0]=1;
ps[1]=1;
ps[2]=1;
ps[3]=1;
ps[4]=1;
float []prob=new float[2];
prob[0]=0.80f;prob[1]=0.20f;
```

Fig 6.5

```
float []prob=new float[2];
prob[0]=0.80f;prob[1]=0.20f;
conf.setCPTable(ps,prob);

ps[0]=1;
ps[1]=0;
ps[2]=0;
ps[3]=1;
ps[4]=1;


prob[0]=0.48f;prob[1]=0.52f;
conf.setCPTable(ps,prob);


net.compile();
Streamer stream = new Streamer ("cwc.dne");
    net.write (stream);
    crowd.enterFinding(0);
     System.out.println(conf.getBelief("TeamA"));
    net.finalize();
    }
    catch (Exception e) {
    e.printStackTrace();
    }
  }
}
```

Fig 6.6

6.3 ROC Curve

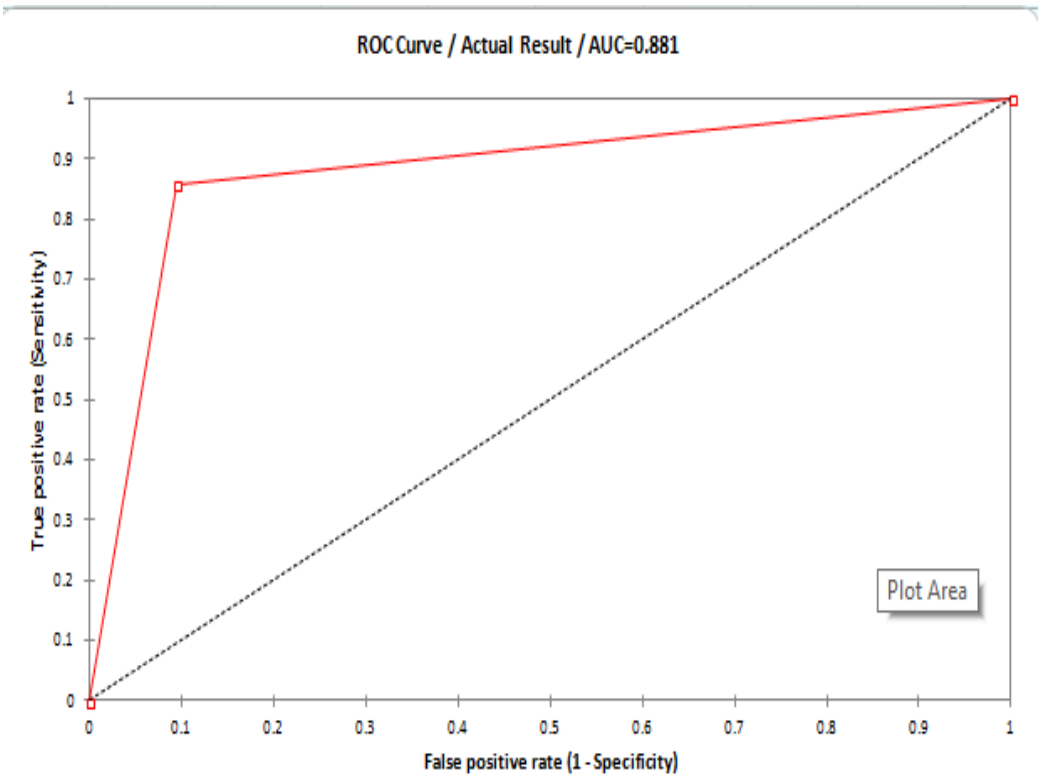


Fig 6.7 ROC plot for predictions

Fig 6.6 shows ROC curve for the prediction of group stage matches of CWC 15. Out of 42 matches our model predicted winning team correctly on 37 occasions making it a success rate of 88.1%

CHAPTER 7

Predicting Water Quality

7.1 Introduction to water quality

Water Quality is a major concern around the world. Water quality is affected by a wide range of natural and human influences. The most important of the natural influences are geological, hydrological and climatic, since these affect the quantity and quality of water available. Excellent detail on water quality can be found on [2] . Their influence is generally greatest when available water quantities are low and maximum use must be made of the limited resources; for example, high salinity is a frequent problem in arid and coastal areas. If the financial and technical resources are available, seawater or saline groundwater can be desalinated but in many circumstances this is not feasible. Thus, although water may be available in adequate quantities, its unsuitable quality limits the uses that can be made of it. Although the natural ecosystem is in harmony with natural water quality, any significant changes to water quality will usually be disruptive to the ecosystem.

The effects of human activities on water quality are both widespread and varied in the degree to which they disrupt the ecosystem and/or restrict water use. Pollution of water by human faeces, for example, is attributable to only one source, but the reasons for this type of pollution, its impact on water quality and the necessary remedial or preventive measures are varied. Faecal pollution may occur because there are n community facilities for waste disposal, because collection and treatment facilities are in adequate or improperly operated, or because on-site sanitation facilities (such as latrines) drain directly into aquifiers. The effects of faecal pollution vary. In developing countries intestinal disease is the main problem, while organic load and eutrophication may be of greater concern in developed countries (in the rivers into which the sewage or effluent is discharged and in the sea into which the rivers flow or sewage sludge is dumped). A single influence may, therefore, give rise to a number of water quality problemS, just as a problem may have a number of contributing influences. Eutrophication results not only from point sources, such as wastewater discharges with high nutrient loads, but also from diffuse sources such as

run-off from livestock feedlots or agricultural land fertilized with organic and inorganic fertilizers. Pollution from diffuse sources, such as agriculture run-off, or from numerous small inputs over a wide area, such as faecal pollution from unsewered settlements, is particularly difficult to control.

The quality of water may be described in terms of the concentration and state (dissolved or particulate) of some or all the organic and inorganic material present in the water, together with certain physical characteristics of the water. It is determined by *in situ* measurements and by examination of water samples on site or in the laboratory The main elements of water quality monitoring are, therefore, on-site measurements, the collection and analysis of water samples, the study and evaluation of the analytical results, and the reporting of the particular location and time at which that sample was taken. Once purpose of a monitoring programme is, therefore, to gather sufficient data (by means of regular or intensive sampling and analysis) to assess spatial and/or temporal variations in water quality.

7.2 Contribution

We summarize our contributions below:

- We propose the use of NaïveBayes classifier to predict the pollution level in the given water sample from the river Yamuna. The objective is to use existing approach for making predictions on a real life application.
- In addition to this we also propose the use of Association Rule Networks that will be useful in performing the descriptive task.
- The same model can be used to predict water quality of any river given the necessary datasets are available for that particular river.

7.3 Predicting Water Quality Of The River Yamuna

In this section we detail naiveBayes classification for classifying a given water sample into five classes namely HP, P, A, E, SP. We will discuss in this section why only naivebayes classifier is chosen for this project, the benefits of naiveBayes over other classifier etc.

In the starting we only have with us large data sets of river Yamuna. The datasets consists of 13 features that are necessary in order to classify future samples to a appropriate class. The quality of water from any river in the world depends on these 14 parameters. And we have approximately 1050 instances of these features in the complete datasets. The snapshot of the datasets showing few instances of these features is shown in figure below.

Fig 7.1 Sample instances of dataset

Initially we have this large data sets with us in order to gain some meaningful

Turbidity	pH	color	DO sat	BOD	TDS	hardness	chlorides	nitrates	sulphates	TC	As	fluoride	class
A	B	C	D	E	F	G	H	I	J	L	M	K	S
E	S	A	S	E	E	S	H	A	S	P	H	E	S
H	P	E	A	S	A	S	E	P	A	E	E	S	S
S	H	S	S	S	S	E	H	H	S	A	S	H	S
A	A	H	P	P	A	H	A	A	S	E	P	A	S
S	P	S	E	S	S	E	H	A	A	P	E	E	S
E	E	H	E	P	E	H	S	S	P	P	E	P	S
S	E	A	E	S	S	P	P	S	H	H	P	E	P
A	S	A	A	P	S	A	S	S	A	H	E	E	S
H	H	A	P	A	P	A	E	E	P	S	H	E	P
E	A	E	S	H	H	S	H	A	P	H	S	E	P
H	S	S	P	A	P	H	E	E	E	H	P	P	P
S	E	S	H	H	S	H	S	A	P	H	E	A	P
S	H	S	P	P	A	E	S	P	A	P	S	E	S
A	A	A	P	A	H	H	S	A	P	H	H	S	P
S	H	H	S	S	H	A	E	H	S	H	E	A	P

information out from this we need to apply some series of techniques which will give us a set direction to proceed further and achieve our sole motive.

First technique we used in this data set is association rule mining so that we can observe and study the hidden patterns found in this given set.

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness.

Let $I=\{i_1,i_{2,}i_3,\ldots\ldots i_n\}$ be a set of n binary attributes called items. Let be a set of $D=\{D_1,D_{2,}D_3,\ldots\ldots D_n\}$ transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form X=>Y where X,Y ⊃ I . The sets of items (for short itemsets) X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively.

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence.

Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times a mined rule have been found to be true.

We have kept minimum threshold value for support and confidence as 10% and 50% respectively. We get 37 significant rules using this min threshold criteria that can summarize the complete data set.

To get different view point from this we additionally plotted Association Rules Networks for the mined rules.

An ARN is a (hyper) graphical model to represent certain classes ,namely rules whose consequents(right-hand sides) are singletons.

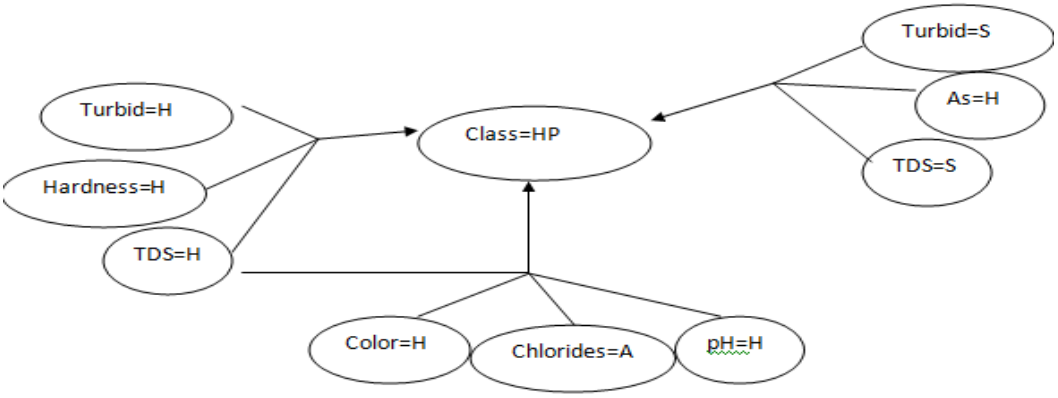


Fig 7.2 ARN for class HP

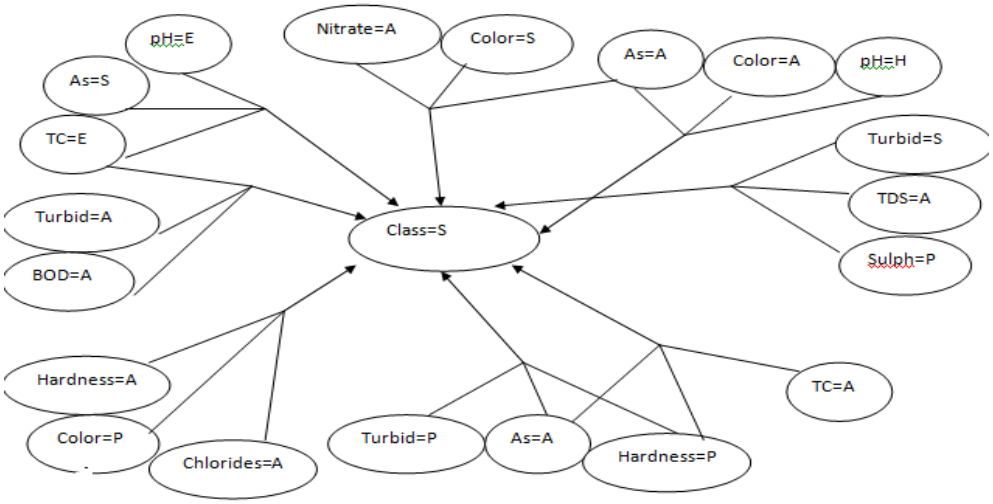


Fig 7.3 ARN for class S

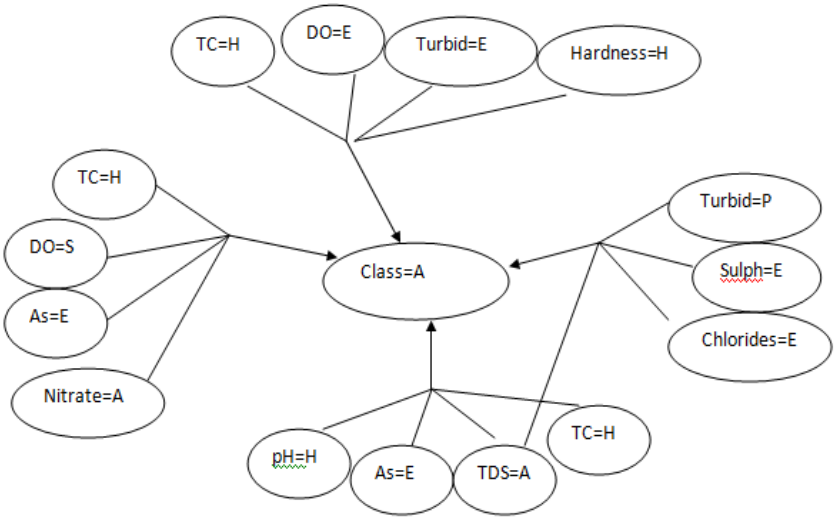


Fig 7.4 ARN for class A

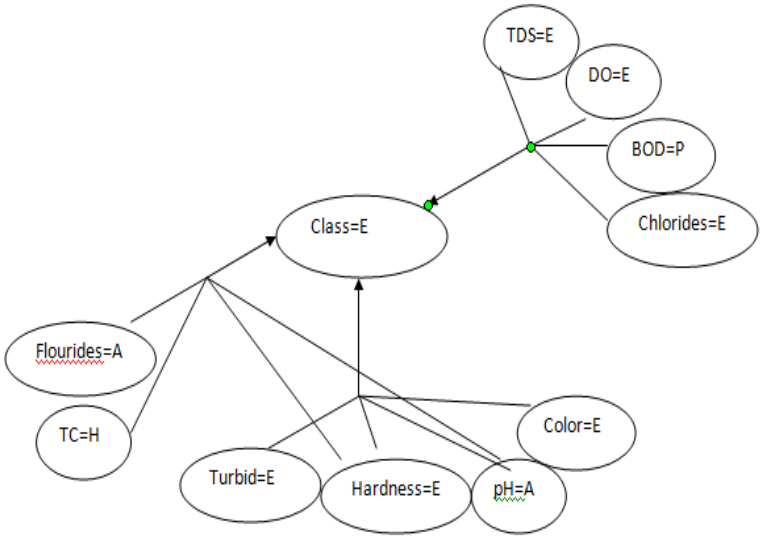


Fig 7.5 ARN for class E

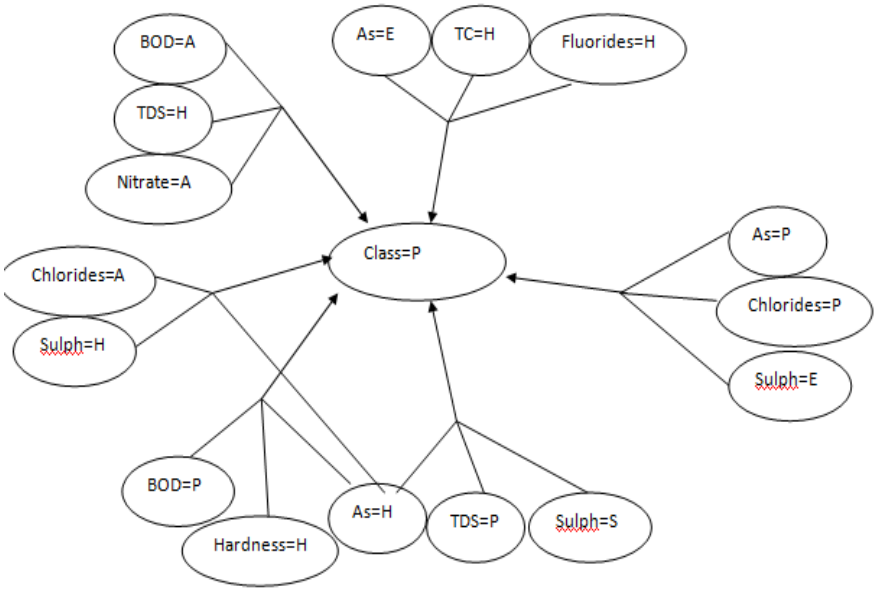


Fig 7.6 ARN for class P

7.2 Application of ARN on describing water quality of river Yamuna

After seeing the Rule hyper-edge graph for each class we saw very important results. These ARN's will be useful in performing the descriptive task. Say if a person wants to know what factors contribute in making the quality of water Excellent (E), by studying the ARN for class E a person gets a fair idea of what should be value of specific features that would make quality of water to E. Similarly ARN's for different classes can be interpreted to get the desired results.

Next we used Structure learning algorithm and see how different features are related in the datasets. We have used Hill-Climbing algorithm to learn the structure from the given data sets.

The learned structure so obtained also shows no edge pointing to any other feature except the class or target feature which signifies that each feature is independent of one another. In addition to that we have also calculated correlation for each feature and found out approximately 0 for every feature.

Hence combining the result of correlation and Learned Structure we can confidently say that each feature is independent.

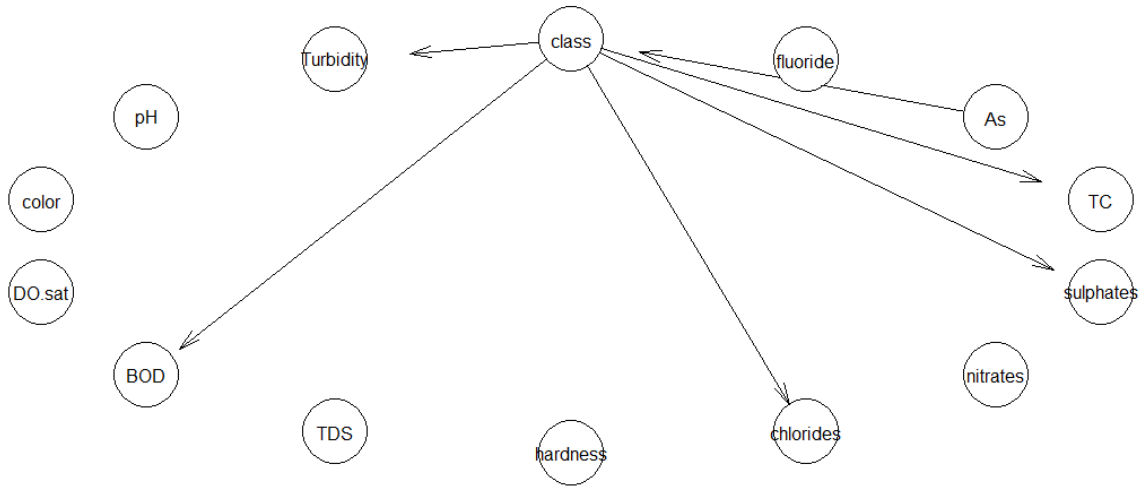


Fig 7.7 Bayesian Structure using Hill Climbing Algorithm

Now we need to choose the best possible classifier keeping in mind the above result.

Chapter 8

8.1 Predictive Analysis Using Naïve Bayes

In chapter 2 we concluded that each feature in the given data sets is independent. Hence it gives us a clear direction which classifier we should pick for prediction. Since naïve Bayes classifer assumes that all attributes are independent of each other given the class label. Hence it is best suited for our data sets.

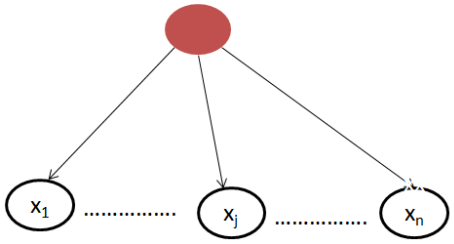
Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness and diameter features.

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers. Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests.

An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification.

$$P(x_i \mid x_j, C_i) = P (x_i \mid C_i)$$



Thus,

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) * P(x_2 \mid C_i) *..... * P(x_n \mid C_i)$$

Fig. 8.1 Generic Naïve Bayes Model

Now we have the best classification model for our datasets that can correctly label our class variable and we can get the quality of water through this class variable.

But in order to judge performance of naïve bayes classifier we have also used another classifier Decision tree and comparison between the two of them is made.

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

The benefits of having a decision tree are as follows −

- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast

Chapter 9

9. Experimental Results

Fig 9.1 shows number of instances of each class type in the target variable, data is critical in order to study precision and recall for each class

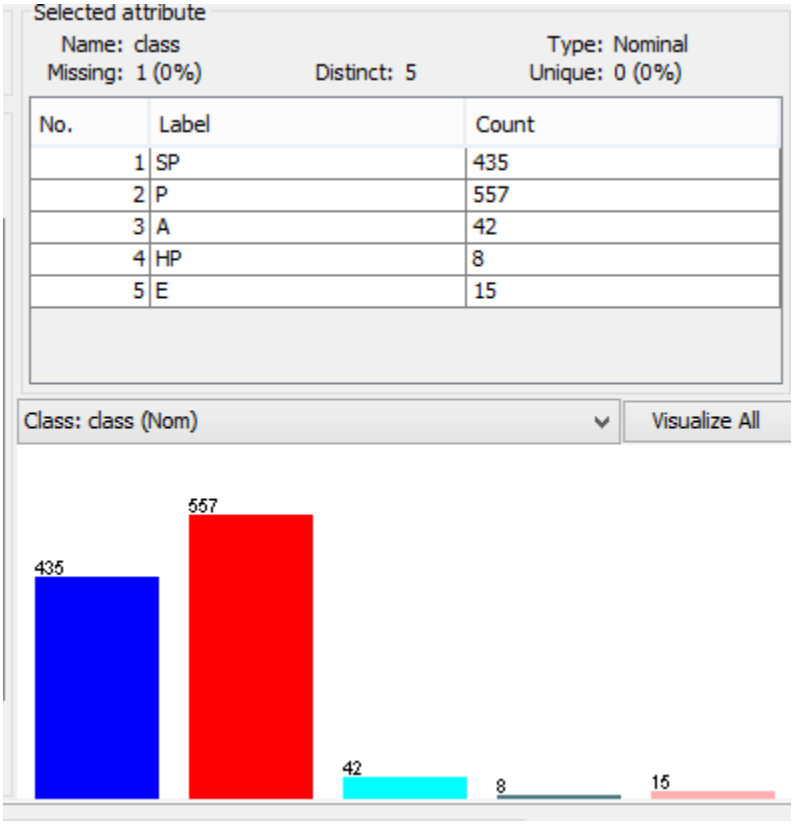


Fig 9.1 Count of each class variable

Next step in our process is to calculate precision and recall value for each class instances using naïve bayes classifier and evaluate the result so obtained. In order to get correct value for precision and recall we have created 10 instances of training(80%) and test cases(20%) and then average of precision and recall is taken for each iteration for every class. Results of such iterations are listed in table 9.1 .

9.1 Precision and Recall for naïve Bayes

In pattern recognition and information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant, while high recall means that an algorithm returned most of the relevant results

Precision and recall are then defined as:

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

tp= No. of true positives

fn= No. of false negatives

fp= No. of false positives

A **confusion matrix**, also known as a contingency table or an error matrix , is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

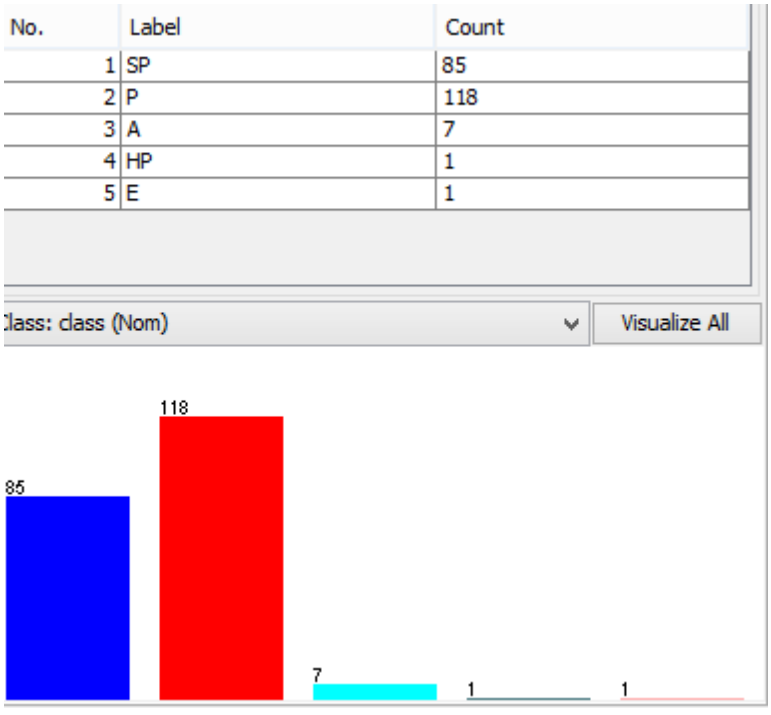
Class	Precision(in %)	Recall(in %)
SP	91.11	81.00
P	85.52	93.00
A	67.79	67.00
HP	100.00	89.00
E	98.88	95.60

Table 9.1

These are the average value of Precision and Recall obtained from successfully iterating 10 random instances of training and test data sets. Here high value of precision of class SP and P suggests that every item labeled as belonging to class SP and P does indeed belong to class SP and P.

However same cannot be said about the class HP that has very low average value(10%) of precision , It is mainly due to less instances of class HP we have such low precision.

While for class A and E we have decent value of precision that suggests that in most cases an item is labeled correctly. Similarly in case of recall , high values in SP and P clearly suggest that almost every item from class SP and P was labeled as belonging to class SP and P. Fig shown below clearly shows extremely low count for class HP and E.



9.2 Confusion Matrix

```
 a  b  c  d  e   <-- classified as
 11  0  0  0  0 |  a = SP
  1 14  0  0  0 |  b = P
  0  0  1  0  0 |  c = A
  0   0  0  0  0 |  d = HP
  0   0  0  0  1 |  e = E
```

9.3 Decision Tree

Next we plotted decision tree for the given training test with our root node as the As feature different values of As leads to different paths in the decision tree and possibly a different classified class.

Ex in the decision tree if a tuple has value of As as N or HP it is straight away classified as belonging to HP class while If value of As= SP in order to classify this tuple we need to look the value of feature TC that is what the path in decision tree leads to whose value will ultimately decide the class of that tuple and so on.

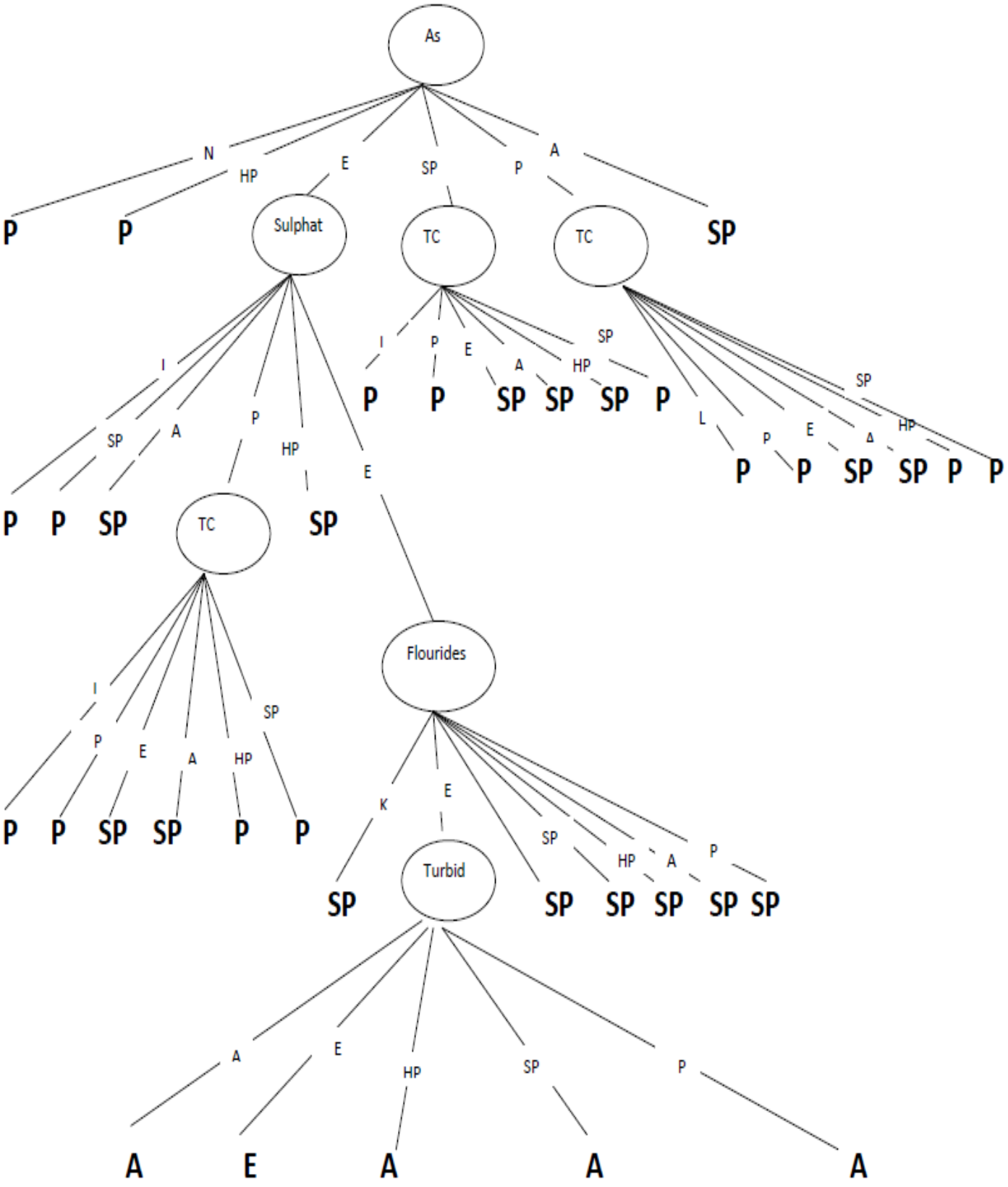


Fig 9.1 Decision tree for given data set

9.4 Precision and Recall data for Decision Tree

Class	Precision(in %)	Recall(in %)
SP	75.1	77.3
P	84.4	85.5
A	88.1	42.8
HP	98.3	97.9
E	98.8	98.7

Table 9.2

On comparing the results with that of naïve bayes classifier we conclude that naïve bayes performs slightly better than decision tree as it has consistent high value of precision and recall for all the classes while decision tree has fluctuating values for some classes.

In next section we will discuss about the ROC curves for our predictions.

9.5 ROC Curve

Receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings. (The true-positive rate is also known as sensitivity in biomedical informatics, or recall in machine learning. The false-positive rate is also known as the fall-out and can be calculated as 1 - specificity). The ROC curve is thus the sensitivity as a function of fall-out.

To validate our prediction we draw ROC curve for each instance of class variable.

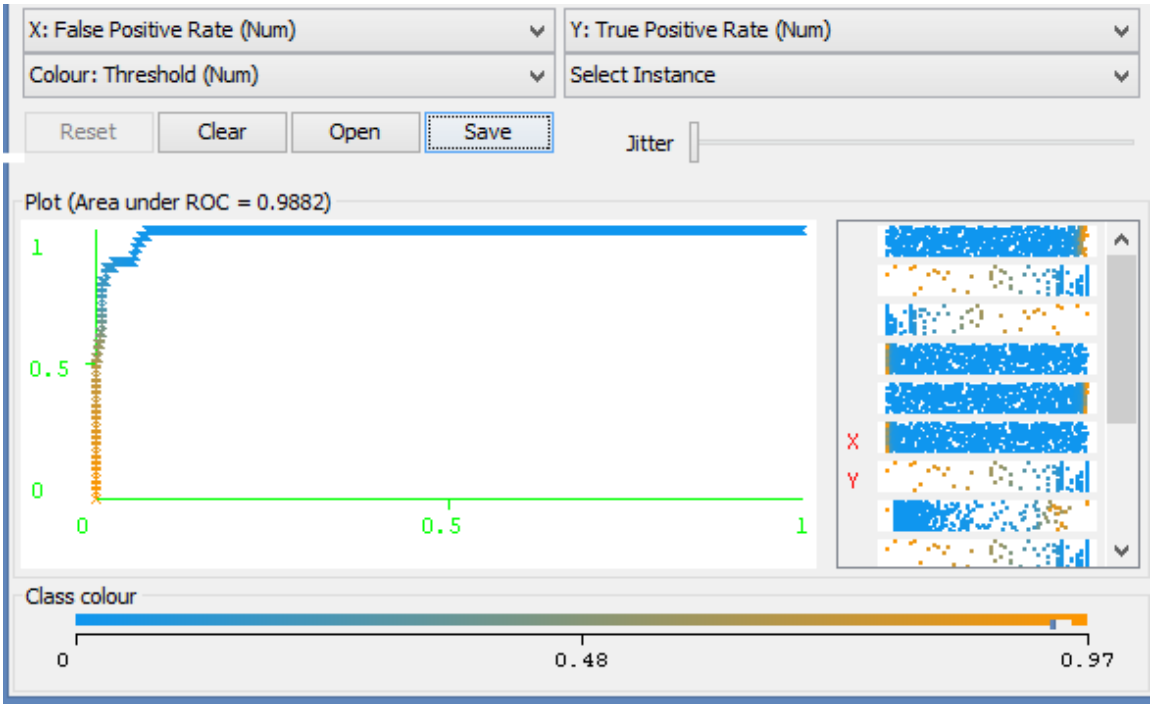


Fig 9.2 Roc for class A

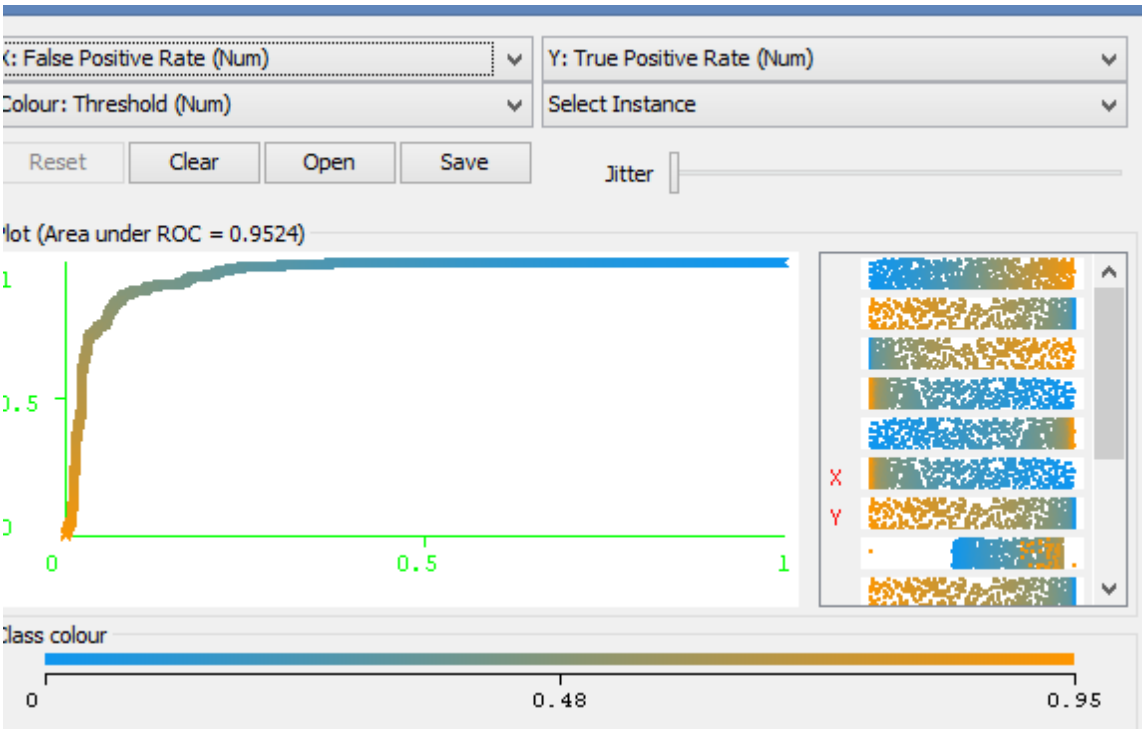


Fig 9.3 Roc for class SP

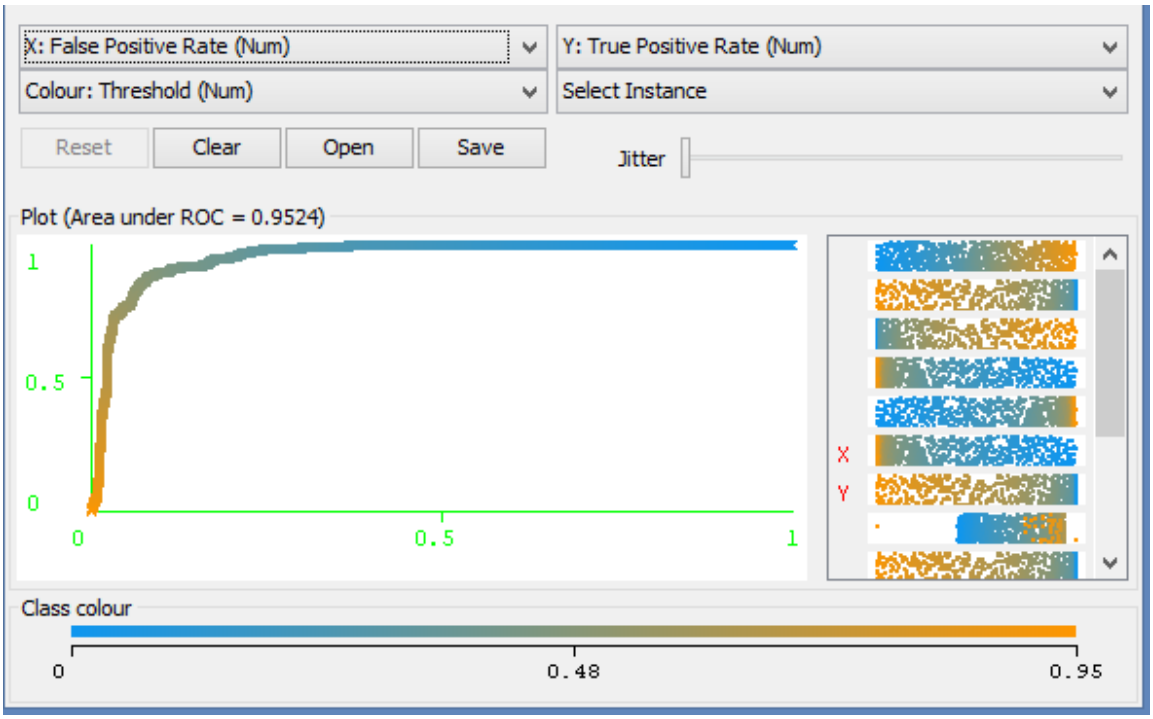


Fig 9.4 Roc for class P

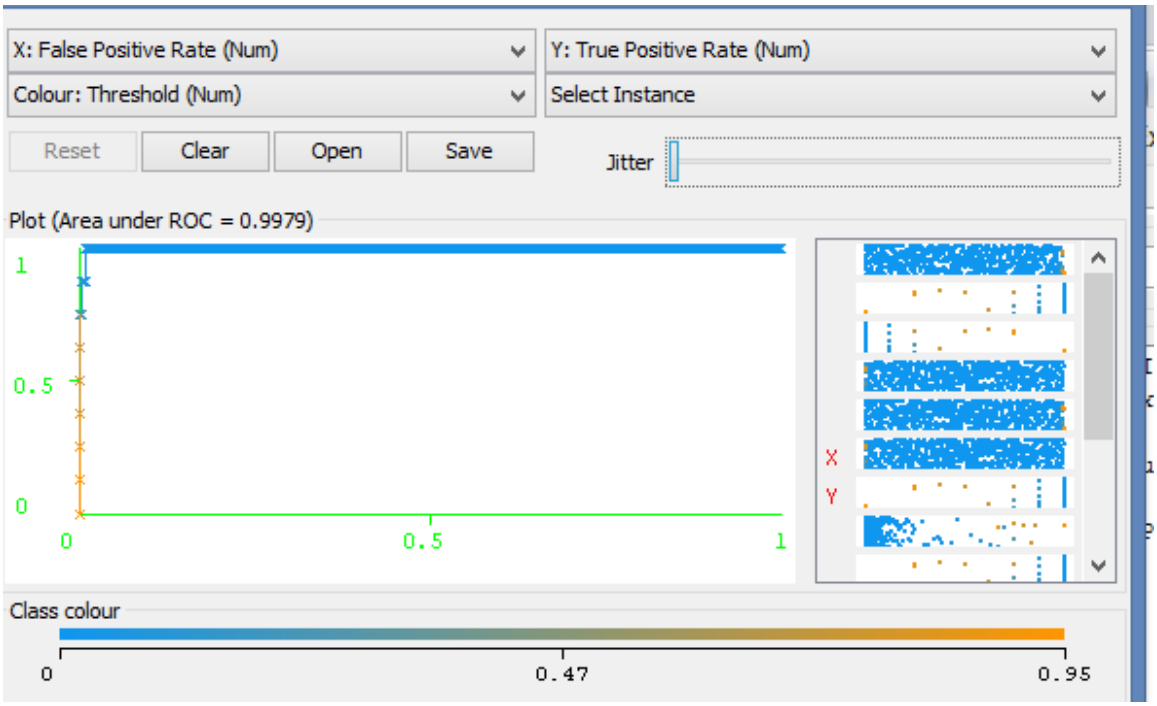


Fig 9.5 Roc for class HP

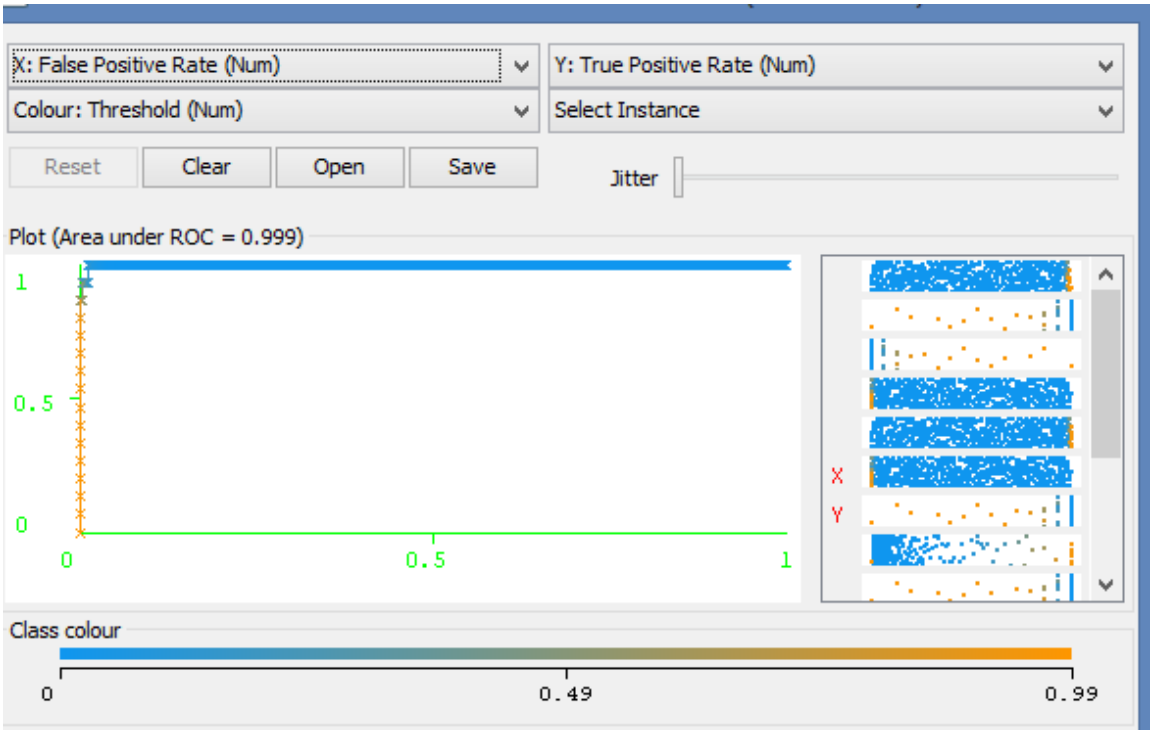


Fig 9.6 Roc for class E

Higher value of area under the curve for each variable clearly shows that our classifier has done exceedingly well in predicting the value of target variable.

Chapter 10

Discussion

In this project we did predictive analysis using naive bayes classifer but what made us follow the naïve bayes path is the nature of data sets given to us. From the given data sets after doing some initial analysis we came to know that each feature in our data sets is independent of one another. We confirmed this fact by calculating correlation of each feature which was close to zero for every feature confirms this fact. In addition to that we also draw Bayesian structure of the data sets using structure learning algorithm and found that each feature has direct edge terminating at class variable which further strengthen our fact. Since Naïve Bayes assumes each feature to be independent hence we are satisfying the very basic condition of Naïve bayes classifer hence we chose naïve bayes for predictive analysis. Later we also compare our results with another very efficient classifer called Decision tree and we found that numbers produced by decision tree weren't as convincing as produced by naïve bayes. Recall for decision tree was quite low as compare to that of naïve bayes. Hence given the nature of data sets naïve bayes proves to be the best classifier for this data set.

In addition to the predicting task we also performed the descriptive task using Association Rule Networks that gives us very useful results. It describes what all features at specific value contribute to a particular class. If someone needs to know what all features can make the quality of water polluted, it can be easily known from the ARN of class P. Hence fulfilling the descriptive task too.

Conclusion and Future Scope

In this project we successfully predicted real life applications and proves the importance of data mining in real world. Firstly we predicted results of cricket world cup 2015 using Bayesian Model with a success rate of 88.1% and predicted Australia to be the new champion of the world. We also presented a deep analysis on parameters, namely, crowd support and confidence which proved key factors in deciding the outcome of the match. Our proposed model can be used in upcoming Cricket World Cup tournaments for making predictions. Using Bayesian analysis teams can prepare on factors on which they may find themselves weak. This work is a real life example that motivates, and explicates usefulness of data mining approaches in predicting a big event. Secondly we predicted quality of water from a given sample using naïve bayes classifier. Our target variable gives us the quality of water with approximately 95-98% accuracy which is indeed a very good result. Our same model can be used for prediction of water quality for any river in India that can be useful to many organizations that are particularly monitoring pollution level in rivers by helping them keeping a track of level of particular parameter that could result in increase/decrease of pollution level in river. We also did descriptive analysis using Association Rule Mining that gives us a clear idea of the role of a specific feature in contributing to a particular class.

References

[1] Chen Q, Mynett AE (2003). Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu Lake. Ecol. Modell. 162 (1/2), 55-67

.

[2] Shoba G, Dr. Shobha G.(2014). Water Quality Prediction Using Data Mining techniques: International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 3 Issue 6 June, 2014 Page No. 6299-6306

[2] David Barber, " Introduction to Graphical Model", MLTA, Vol.1 , pages 70

[3] Adnan Darwiche. (2014). Making Sense of the Aam Aadmi Party Win in the Delhi Elections. Available: http://www.rediff.com/news/report/making-sense-of-the-aam-aadmi-party-win-in-the-delhi-elections/20131223.htm. Last accessed 17th Aug 2014.

[4] Cricinfo Team. (2014). Cricket Statistics. Available: http://www.espncricinfo.com/ci/content/stats/index.html. Last accessed 25th Jan 2015.

[5] Wikipedia. (2014). Demographics of Australia. Available: http://en.wikipedia.org/wiki/Demographics_of_Australia. Last accessed 25th Jan 2015

.

[6] ICC. (2015). World cup results. Available: http://www.icc-cricket.com/cricket-world-cup/results. Last accessed 22th Mar 2015.