The Indian Premier League (IPL) is a professional cricket league in India based on the Twenty20 format. It was constituted in 2008 and has been conducted every April-May since then. The tournament is conducted in a Double round-robin league and Playoffs fashion, where each team encounters the remaining teams twice, each played at their home venues respectively. It was initially founded by BCCI.

The exhaustive dataset associated with IPL calls for segregation of data in keeping with various parameters such as home ground, toss result, etc. In this particular report, we will be focusing on the winning/losing team based on the teams that won/lost the toss and the location of the pitch it's being played on. This is to ensure that the dimensionality of the data is kept at a minimum, including only the most essential parameters. Since IPL follows super over methodology in case of a tie, we have no class labels pertaining to ties.

Having extracted the data, it will be subjected to the various steps pertaining to data preprocessing, to ensure that there are no anomalies in our data set such as missing data, redundant data, etc. We will then subject a subset of the dataset to 2 Data Mining techniques that we deemed fit, Classification and Association. Using both these methods we will later use test records to test the model/rules that we have built/selected.

The dataset that has been used for this particular report has been sourced from kaggle. The data has been collected during the span of 2008-2016 (9 seasons).

Section 1 will focus on application of a classification method (Bayes theorem) and Section 2 will focus on deriving rules using association mining (Apriori algorithm).

Section 1

Bayes theorem is an approach for modeling probabilistic relationships

between the attribute set and the class variable. The section begins with an introduction to the Bayes theorem, a statistical principle for combining prior knowledge of the classes with new evidence gathered from data.
Let X denote the attribute set (which in this case is pitch, time, toss) and Y denote the class variable (which in this case is Win/Lose). During the training phase, we need to learn the posterior probabilities P(YlX) for every combination of X and Y based on information gathered from the training data. By knowing these probabilities, a test record X' can be classified by finding the class Y' that maximizes the posterior probability, P(Y'/X').

Bayes theorem is given by

P(Y/X)= [P(X/Y)xP(Y)] / P(X)

When comparing the posterior probabilities for different values of Y, the denominator term, P(X), is always constant, and thus, can be ignored. Prior probability P(f) can be easily estimated from the training set by computing the fraction of training records that belong to each class.

Consider a subset of the data as displayed below:
Table 1 shows us the winner for each match, along with the parameters we have taken into consideration. This includes analysing the possibility that the team which won the toss won the match, the possibility of the home team reaping benefits of the home pitch and lastly Table 2 shows us the home ground of each team, enabling us to make deductions about the degree to which the location of the match influences the result in favour of the home team. SInce Bayes theorem assumes conditional independence of each variable, we can calculate the posterior probability to classify a test record.

	A	B	C	D	E	F	G	H
1	Team_Name_Id	Opponent_Team_Id	Season_Id	Venue_Name	Toss_Winner_Id	Toss_Decision	Match_Winner_Id	City_Name
2	2	1	1	M Chinnaswamy Stadium	2	field	1	Bangalore
3	4	3	1	Punjab Cricket Association Stadium, Mohali	3	bat	3	Chandigarh
4	6	5	1	Feroz Shah Kotla	5	bat	6	Delhi
5	7	2	1	Wankhede Stadium	7	bat	2	Mumbai
6	1	8	1	Eden Gardens	8	bat	1	Kolkata
7	5	4	1	Sawai Mansingh Stadium	4	bat	5	Jaipur
8	8	6	1	Rajiv Gandhi International Stadium, Uppal	8	bat	6	Hyderabad
9	3	7	1	MA Chidambaram Stadium, Chepauk	7	field	3	Chennai
10	8	5	1	Rajiv Gandhi International Stadium, Uppal	5	field	5	Hyderabad
11	4	7	1	Punjab Cricket Association Stadium, Mohali	7	field	4	Chandigarh
12	2	5	1	M Chinnaswamy Stadium	5	field	5	Bangalore
13	3	1	1	MA Chidambaram Stadium, Chepauk	1	bat	3	Chennai
14	7	8	1	Dr DY Patil Sports Academy	8	field	8	Mumbai
15	4	6	1	Punjab Cricket Association Stadium, Mohali	6	bat	4	Chandigarh
16	2	3	1	M Chinnaswamy Stadium	3	bat	3	Bangalore

Table 1: Match data table

	A	B	C
1	Team_Id	Team_Name	Team_Short_Code
2	1	Kolkata Knight Riders	KKR
3	2	Royal Challengers Bangalore	RCB
4	3	Chennai Super Kings	CSK
5	4	Kings XI Punjab	KXIP
6	5	Rajasthan Royals	RR
7	6	Delhi Daredevils	DD
8	7	Mumbai Indians	MI
9	8	Deccan Chargers	DC
10	9	Kochi Tuskers Kerala	KTK
11	10	Pune Warriors	PW
12	11	Sunrisers Hyderabad	SRH
13	12	Rising Pune Supergiants	RPS
14	13	Gujarat Lions	GL

Table 2: Team data table

	A	B	C	D	E
1	Venue_Name	Host_team_Id_1	Host_team_Id_2	Host_team_Id_3	Country
2	M Chinnaswamy Stadium	2			India
3	Punjab Cricket Association Stadium, Mohali	4			India
4	Feroz Shah Kotla	6			India
5	Wankhede Stadium	7			India
6	Eden Gardens	1			India
7	Sawai Mansingh Stadium	5			India
8	Rajiv Gandhi International Stadium, Uppal	8	11		India
9	MA Chidambaram Stadium, Chepauk	3			India
10	Dr DY Patil Sports Academy	10			India
11	Newlands				South Africa
12	St George's Park				South Africa
13	Kingsmead				South Africa
14	SuperSport Park				South Africa
15	Buffalo Park				South Africa
16	New Wanderers Stadium				South Africa
17	De Beers Diamond Oval				South Africa
18	OUTsurance Oval				South Africa
19	Brabourne Stadium	5	7		India
20	Sardar Patel Stadium, Motera	5			India
21	Barabati Stadium	8	4	1	India
22	Vidarbha Cricket Association Stadium, Jamtha				India
23	Himachal Pradesh Cricket Association Stadium	4			India
24	Nehru Stadium	9			India
25	Holkar Cricket Stadium	4			India
26	Dr. Y.S. Rajasekhara Reddy ACA-VDCA Cricket Stadium	7	12		India
27	Subrata Roy Sahara Stadium	10	4	12	India
28	Shaheed Veer Narayan Singh International Stadium	6			India
29	JSCA International Stadium Complex	3			India

Table 3: Venue data table

Section 2

Association rule mining (Apriori algorithm) is used to discover hidden relationships among the data objects. In this case we again consider the same subset of data as considered in Section 1. We proceed via the following steps:

1. Data preprocessing
    a. Feature Extraction : Since we are focusing only on a certain number of factors influencing the outcome of a match, we need to clean the data in order to keep only the features that is applicable to our application. We perform this directly on the source of the data (Removal of variables)
    b. Feature Evaluation : The dataset did not have explicit mention of which was the home team. This had to be done by identifying the stadium as being the home ground of the particular team.

The rules as mentioned below are the ones extracted using association rule mining. This is done in 2 steps, first being frequent itemset generation using a support threshold, say 50% and the second being rule mining using confidence threshold, which is chosen in a range of 60-80% to avoid ending up with too few rules or too many rules. The rules generated below can be strengthened by taking into consideration parameters such as bowler capacity (average number of wickets/over, speed for pace bowlers, etc), batsman capacity (strike rate, number of 100's, 50's, etc.), history of the team's performance (in terms of ranking, Net run rate, etc.). Including these variables as a part of our analysis will not only strengthen the model but also give a more accurate and predictable set of rules. For the time being, for demonstration purpose, we have considered only the 3 parameters as specified before.

The rules that were mined using the 3 factors considered were:

1. Did Team 1 win the toss?
2. Did Team 2 win the toss?
3. Did Team 1 choose to bat first?
4. Did Team 2 choose to bat first?
5. Did Team 1 choose to ball first?
6. Did Team 2 choose to ball first?
7. Did Team 1 play at home?
8. Did Team 2 play at home?
9. Did Team 1 play away?
10. Did Team 2 play away?
11. Did the Team Batting First Win?
12. Did the Team Batting Second Win?

Manual calculation of best fit association rules :

DATE

SEASON 1 data summary, SEASON 3 data summary is considered.

total number of matches = 53 total number of matches: 52

1. Team won toss → Team wins match.

support = $\frac{\sigma(X \cup Y)}{N} = \frac{28+26}{105}$ = ~~50%~~ 51.4%.

confidence = $\frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{28+26}{105}$ = ~~50%~~ 51.4%.

since the confidence does not cross the threshold, we ignore this rule.

2. Host team → Team won

Support = $\frac{27+27}{105}$ = ~~27%~~ ~~54%~~ 51.4%.

confidence = $\frac{27+27}{105}$ = 54% 51.4%.

we repeat the same procedure as above for elimination of this rule.

3. Team wins Toss and Team Bats first → Team won

support = $\frac{13+13}{105}$ = ~~26%~~ 24.7%.

confidence = $\frac{13+13}{65}$ = 40.0%.

Home Pitch —Yes,No→ Team won toss

Yes / No

Team chose to bat: Yes / No

Team was assigned batting first: Yes / No.

From the above tree we have mined the rules using apriori algorithm. Out of the 12 rules we have analysed 3 for representation purpose.

Terminal Screenshot of Results, Calculations and Conclusions:

For cross verification of the data points taken into consideration, we ran a python script, to get a summary of the data we desired. The code and the output have been attached.

```
--MATCH NO: 56 --
[7, -1, -1]
[6, 5]
AWAY !
[]


--MATCH NO: 57 --
[7, -1, -1]
[3, 4]
AWAY !
[]


--MATCH NO: 58 --
[10, -1, -1]
[3, 5]
AWAY !
[]




SEASON 1 SUMMARY
total matches: 59
total home matches played on either of the teams pitch: 53
total away matches,neither of their pitches: 6
toss won + home teams: 22
toss won + team won:  28
host team + team won:  27
toss won +team won + bat first team:  13
toss lost by team 1, match won  0
```

Result and Conclusion

Since we have constricted ourselves to a small subset of data, we will truly not end up with the rules which we intend to use since we are yet to scale the data considered. We can also approach the problem in another manner, i.e, consider a training set of 3-4 seasons and test the model built on the latest season. This not only builds a model that is robust but is also more accurate.
We can also consider team wise data and present results for the same.