INTRODUCTION
1. There is a huge amount of statistical data available in Cricket. There is a growing interest in using this historical data to understand the performance of the team and thereby do some performance prediction for the upcoming matches.
2. The goal of my capstone project is to predict the results of the IPL cricket matches based on the past data about a team’s performance using the various data mining tools and algorithms.
3. IPL is the Indian Premier League domestic 20 overs cricket tournament held in India every year. There has been a total of 8 seasons of IPL held so far starting from 2008 to 2016.

REFERENCES

METHODOLOGY
Dataset Description: The data is collected from the http://cricsheet.org website. This Cricket dataset consists of data about 574 different IPL matches comprising of all the seasons from 2008 to 2016. The dataset is divided into 21 different attributes as shown below.

- **Original List of Attributes**
  - Season
  - Match Number
  - Team 1
  - Team 2
  - Venue
  - Home Team
  - Toss Winner
  - Toss Decision
  - Player of Match
  - Team Batting First
  - Team Batting Second

- **Reduced List of Attributes**
  - Team 1
  - Team 2
  - Venue
  - Home Team
  - Toss Winner
  - Toss Decision
  - Player of Match
  - Team Batting First
  - Team Batting Second

RESULTS
The dataset was divided into Training and Test Set and then the accuracy rate on the Test Set was used as the evaluation criteria for the classification algorithms. The accuracy results are:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Percentage Split</th>
<th>10 Fold Cross Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree (J48)</td>
<td>31.07%</td>
<td>63.24%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>48.02%</td>
<td>71.08%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>57.06%</td>
<td>60.97%</td>
</tr>
<tr>
<td>K-Nearest Neighbor</td>
<td>52.54%</td>
<td>51.39%</td>
</tr>
</tbody>
</table>

CONCLUSION
1. The prediction results in WEKA were significantly better with 10 Fold Cross Validation as compared to the Percentage Split.
2. Datasets with more instances would help in improving the accuracy rate of the model.
3. The prediction results were good with Decision Tree algorithm (63.24%) and Random Forest algorithm (71.08%).

FUTURE WORK
A Web Application which would take in the dataset as input and display the prediction result for each instance as output in a PDF or Text file.

ACKNOWLEDGEMENTS
1. Dr. Carol Romanowski for being my Project Advisor and providing me her valuable advise in completing the project.
2. Dr. Joe Geigel for being my Colloquium Advisor.