Applying Statistics to Understand how using different balls can increase the Competitiveness of a Cricket Test Match

Presented By: Ayush Somani & Aditya Mudgil
<Serial No. 53>

Name of The Author:
Abhinav Agarwal
Aditya Mudgil
Ayush Somani

IIT (ISM) Dhanbad

Introduction

Test Match in Brief:

&#;All players must wear white dress.

&#;You have 3 session and breaks in a day, lunch, tea and post tea till end of days play.

&#;Test has to be played with red ball.

&#;Slightly wider delivery in ODI is given wide, but in Test it is a little relaxed and considered a legitimate delivery (unless it is very wide)

&#;No free hit for a "NO" ball in Test.

&#;Test Match has four innings (Two innings each for both teams)

Test Cricket Balls

At the Test match level, three basic balls used: the **Kookaburra** in Australia, New Zealand, West Indies, Sri Lanka, the **Duke** in the United Kingdom ,South Africa and the **SG** in India, Pakistan.

KOOKABURRA

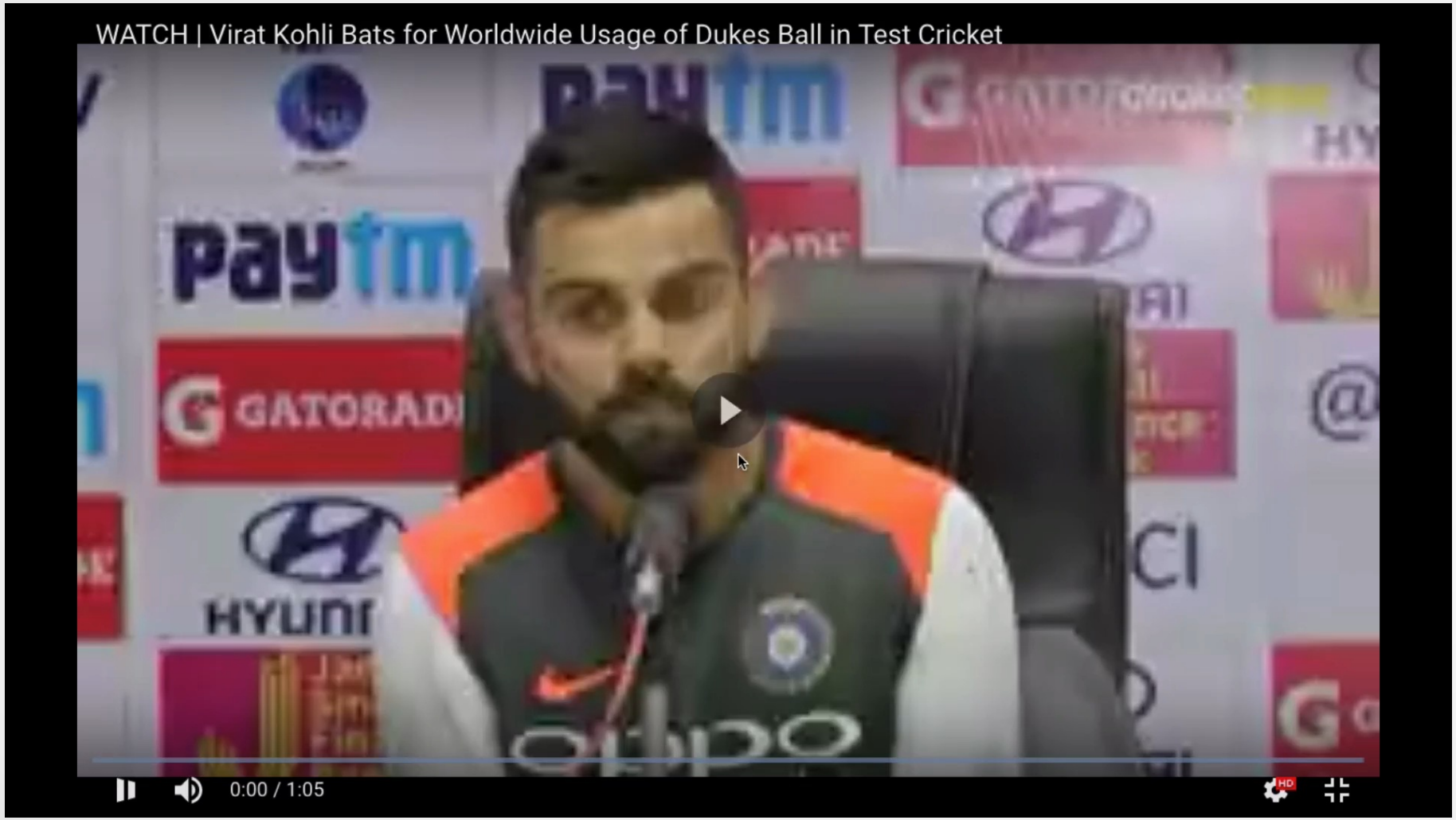
SG

DUKES

Comparison between the Kookaburra, the Duke and the SG

	Kookaburra	Dukes	SG
Manufacturing:	Machine-made	Hand-made	Hand-made
Helpful for Spinners:	Less Seam	Pronounced Seam	Pronounced Seam
Swing:	Swings more at initial 20-30 over.	Swings more at later stage of match.	Hardly Swings
Reverse Swing:	Ideally on hot and Humid Condition	Less Reverse Swing	Assist more
Leather Longevity:	Last long in Turf Condition.	Cannot last on Rougher condition	Last Satisfactorily

Virat Kohli’s Opinion on Test Cricket Ball



Prerequisite

Pitches

Pitches in different parts of the world have different characteristics. The nature of the pitch plays an important role in the actual game.

&#;A spin bowler may be preferred in the dry pitches of Indian subcontinent.
&#;All pace attack may be used in places like Australia where the pitches are bouncy.

Pitches across Countries:

01 Australia

- Good for fast bowlers because of the bounce
- &#; Some assist spinners for more dust cover
- &#; Tennis-ball bounce which can negate the potency

02 England

- &#; Green, swing promoting and humid conditions
- &#; Spinners prove effective in the second half
- &#; Reverse Swing with a 50-over old ball.

03 India

- Historically supported Spin Bowling.
- &#; Indian batsmen relishes home conditions
- &#; Surfaces are often tailor made to be flat tops or excessively batsmen-friendly.
- &#; Some contemporary pitches provide good support.

04 New Zealand

- Ball swings a lot due to the proximity of most grounds to the sea
- &#; bouncy and quick in nature due to the usual grass cover.
- &#; Batsman often take time to adjust

05 Pakistan & UAE

- The UAE features spin-friendly pitches.
- &#; Reverse Swing and spin with the older ball.
- Pitches in Pakistan are flat and favourable for batsmen in winter and spinners in summers.
- &#; Dry and windy conditions.

06 South Africa

- Added swing movement and lesser bounce.
- &#; Genuine fast bowlers can well do the most damage

Proposed Approach

&#; Data Collection
6 Teams on the Basis of ICC Ranking Average
For the Year Range 2005-2017
Total number of Matches- 294

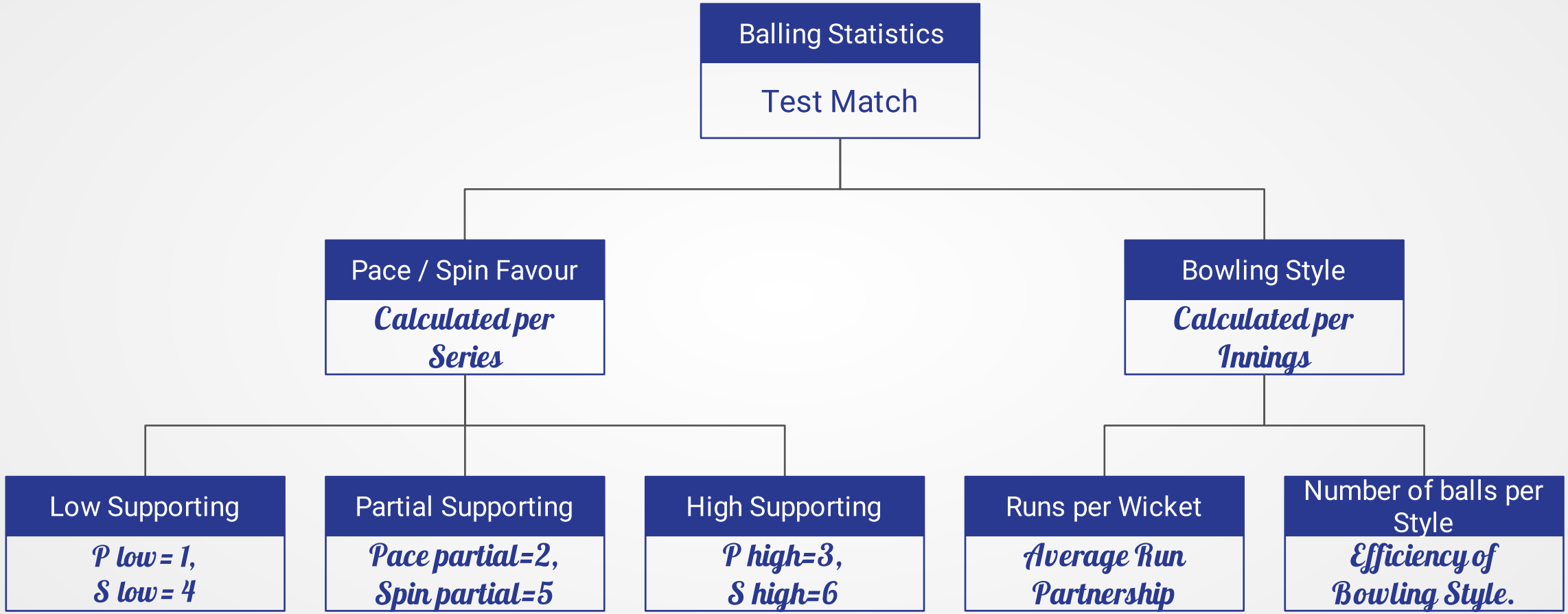
&#; Label Encoding

&#; Pitching a Competitive Factor Analysis

$$\text{Competitive Factor (Y Label)} = \sum A_i X_i^r$$

Ai= Feature Weightage
Xi= X Label Feature
r= If the feature is polynomial dependent

Balling Style Feature



Correlating Features

Most Significant Factor in Deciding the Type of Ball to be Used.

Ball is manufactured on the basis of Pitch and climate condition.

Relating the Competitiveness with Margin of Winning.

Relative Competitive level.

Stats of Pace and Spin Bowler to justify the Pitch Favoring Condition.

The Problem

Dataset

Unavailability of Dataset

&#; Ambiguous Feature Selection.
&#; Lack of Analytics Record in Past.
&#; Unclear Ideal Dataset at Present.
&#; Overfitting of Dataset.

Calculation

Time Period Constraint

- Climatic Effect
- Pitch Wearout
- Subjected to Mental Strength of Players.
- Roughness in Ball.

Prediction

Unpredictability Game Ending

- Subjected to External Factor and Deviation Rule Out.
- Abnormality in Game Ending.
- Missing Bowler Statistics.

Challenges deep-dive

Challenge 1

Designing a Competitive Factor

Designing a Competitive Factor Formula was initially difficult due to unavailability of proper dataset.

Challenge 2

Relating the Ball with Pitch.

Ball has varying specification but incorporating it directly was difficult.

- Venue-> Climate
- Pitch Condition
- Temperature

Challenge 3

Appropriate Feature Scaling

&#;Features to be invoked in the model to predict the class is still to be justified.

&#;Ambiguity in Draw Matches.

Steps of Overcoming Failure:

01 **Unbiased Balanced Data**

02 **Bowling Features**

03 **Competitive Factor (Y Label)**

04 **Ambiguous 3rd and 4th Inning Score**

Margin Point Calculation.

Each Winning team scores points based on the results of their Series Margin of Win. A series must include at least two Tests.

Finding the Average Score of Teams in Test Innings:

A. OVERALL SUMMARY OF RESULTS - BY PERIOD

Period	Tests	FI-AvgeRuns	Results	%	FB-Wins	%	SB-Wins	%	Draws	%
All Tests	2085	324.6	1365	65.5	690	33.1	675	32.4	720	34.5
1877 - 1948	307	307.8	217	70.7	125	40.7	92	30.0	90	29.3
1949 - 1979	560	320.8	328	58.6	184	32.9	144	25.7	232	41.4
1980 - 1999	613	316.7	366	59.7	173	28.2	193	31.5	247	40.3
2000 - 2013	605	344.6	454	75.0	208	34.4	246	40.7	151	25.0

Margin of Winning	Margin points Earned
W Runs	**'W'** Points
X Innings **Y** Runs	**(X*340)+ Y** Points
Z Wicket	**Z* (340/10)** Points
Draw	**0**

Experimental Details

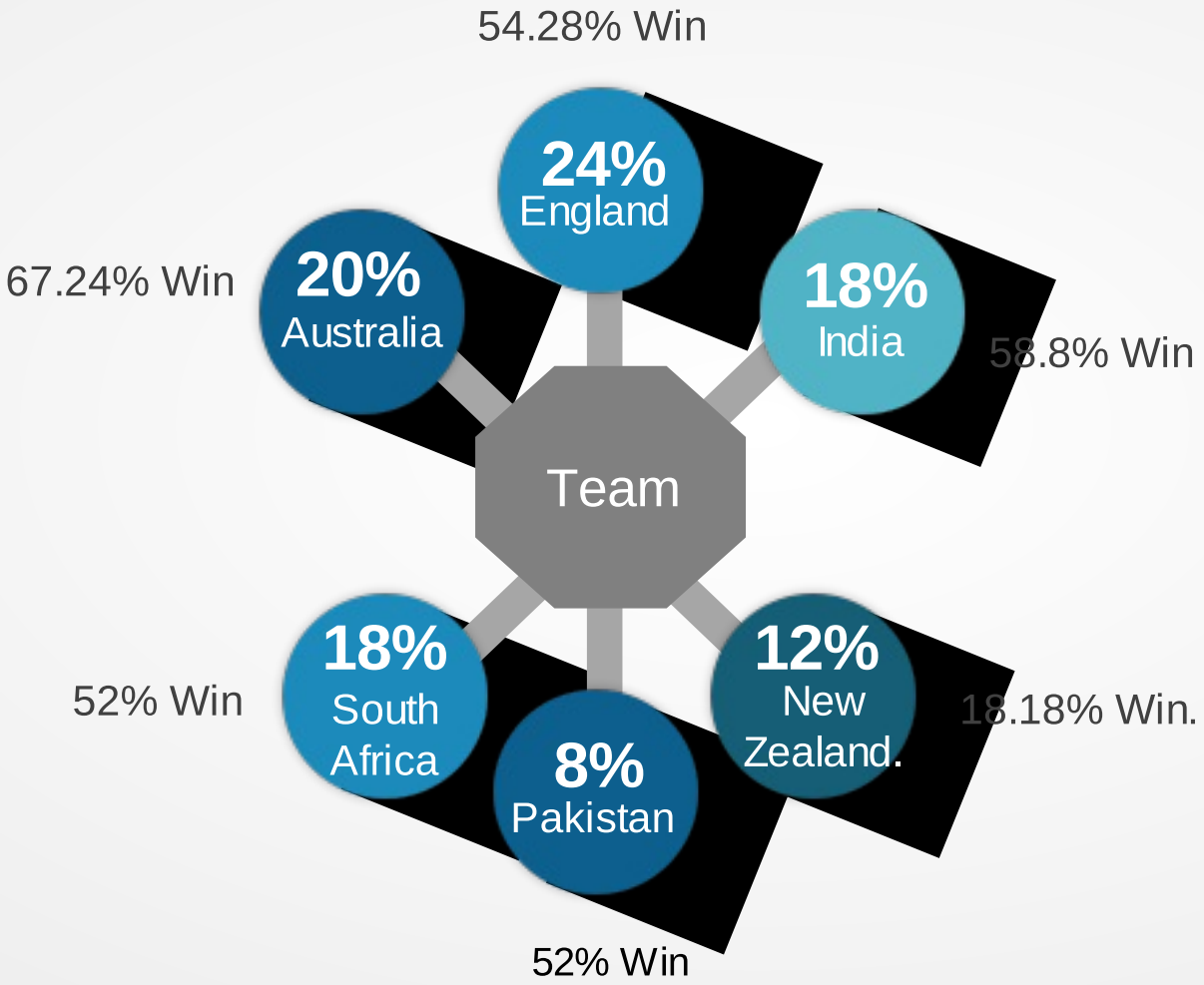
Idea:

- The more the Margin Point -> The Less is the Competitiveness in Match.
- The Less the Margin Point -> The more is the close Competition in the Match.
- We consider the draw of a match signifies that the match has the most close Competitiveness and goes till last day.

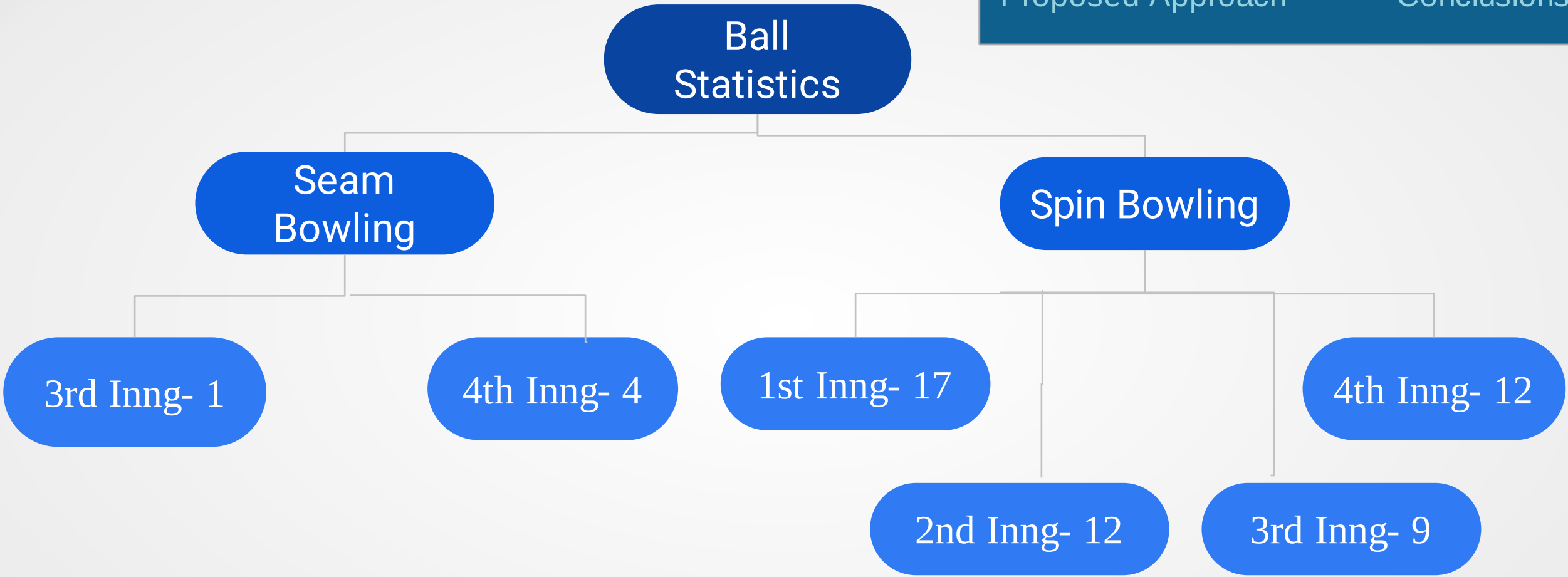
$$\text{Competitiveness} \Longleftrightarrow \frac{1}{\text{Win Margin Point}}$$

Margin Points	Competitive Class Classification
1-99 Points	**5**
100-199 Points	**4**
200-299 Points	**3**
300-399 Points	**2**
400-499 Points	**1**
500-599 Points	**0**
0 (Draw)	**6**

Hosting Countries



Missing Bowling Statistics- A Challenge.

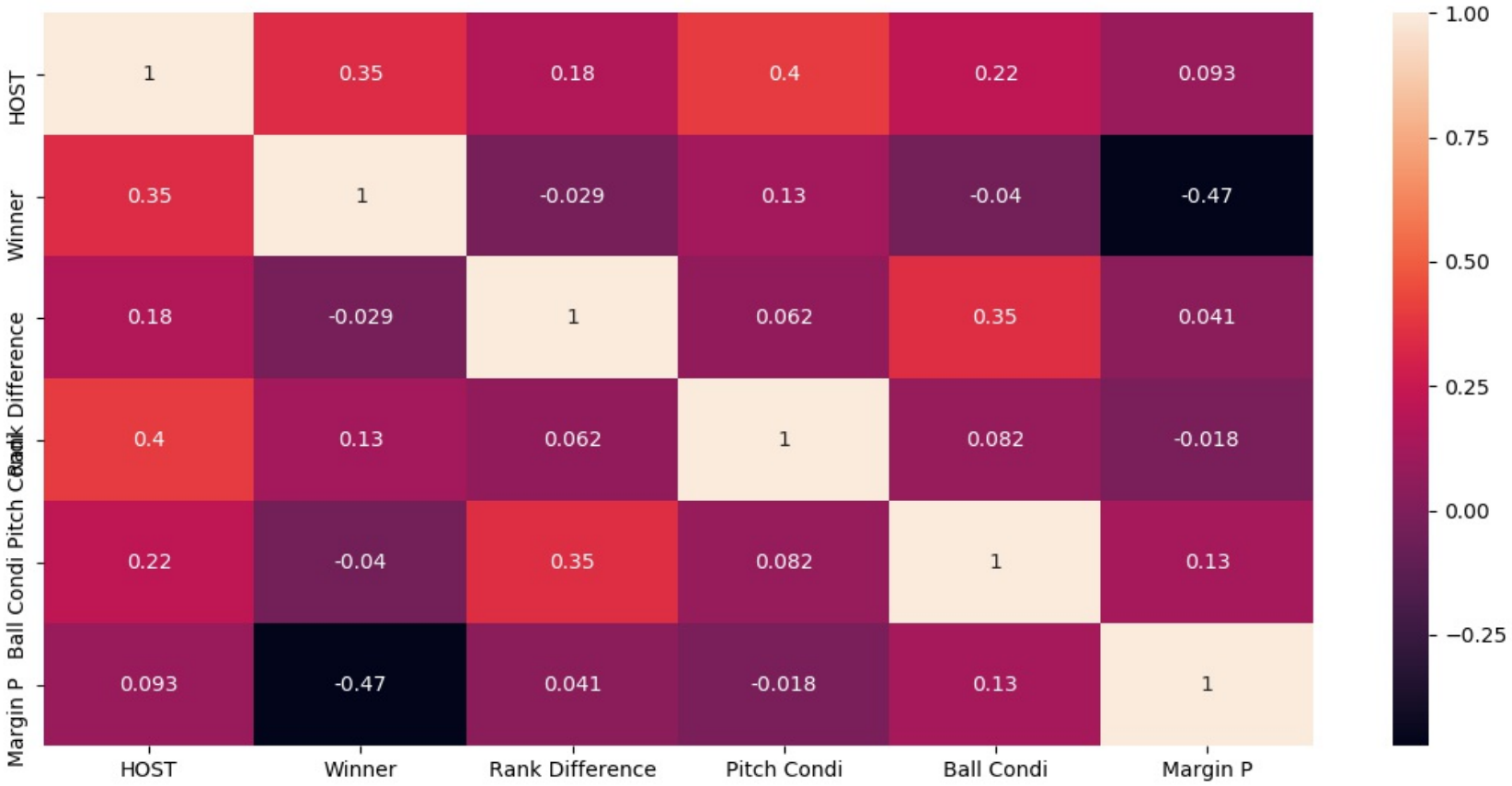


3 Matches Ends before 3rd Inning- 1.04%
67 Matches Ends before 4th Inning- 23.26%

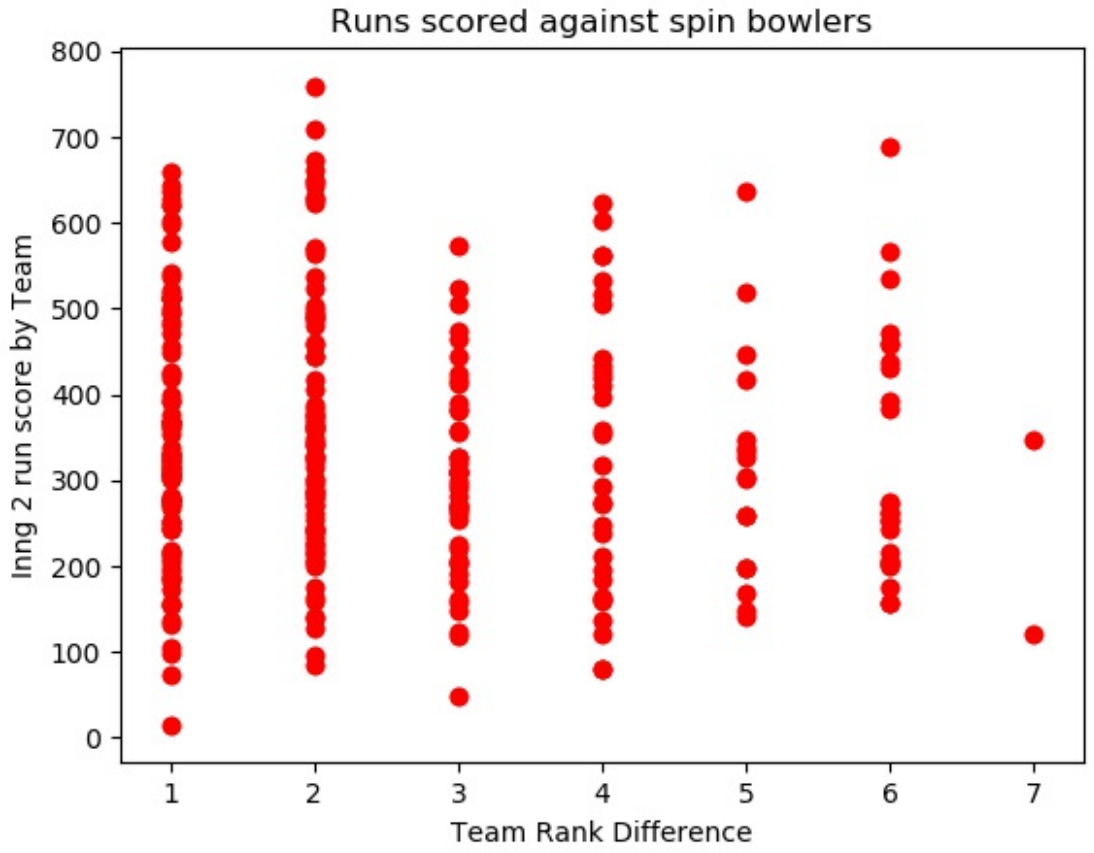
Implementation

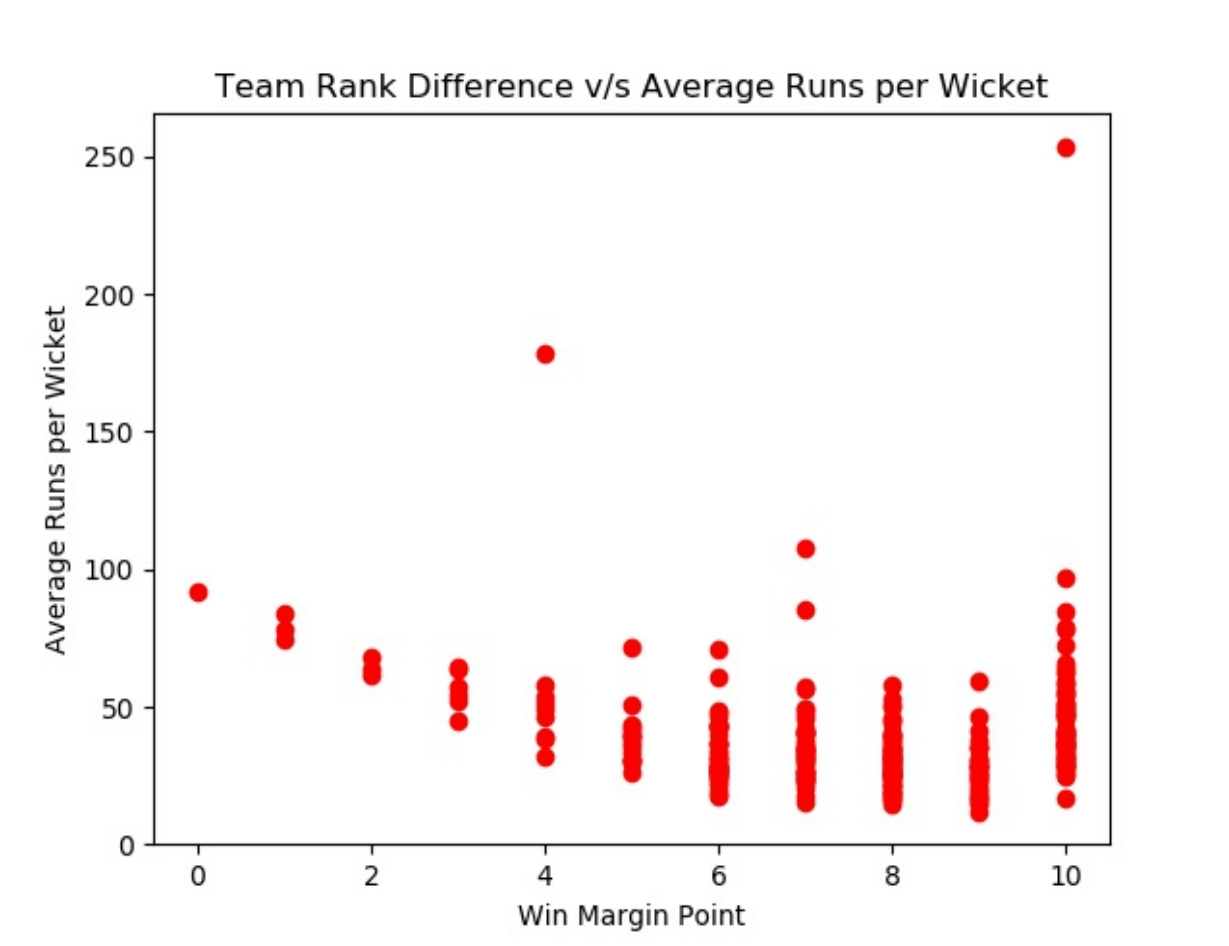
Features Selection

Correlation of Features

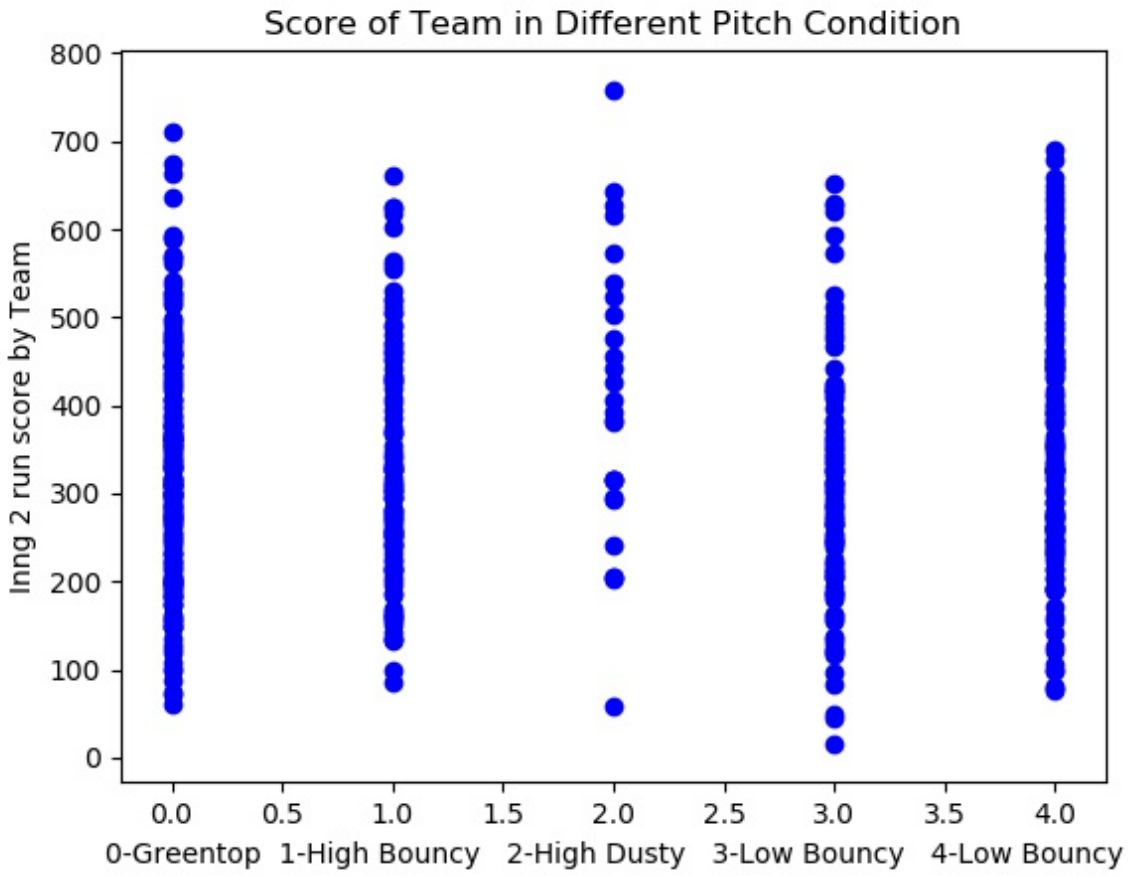


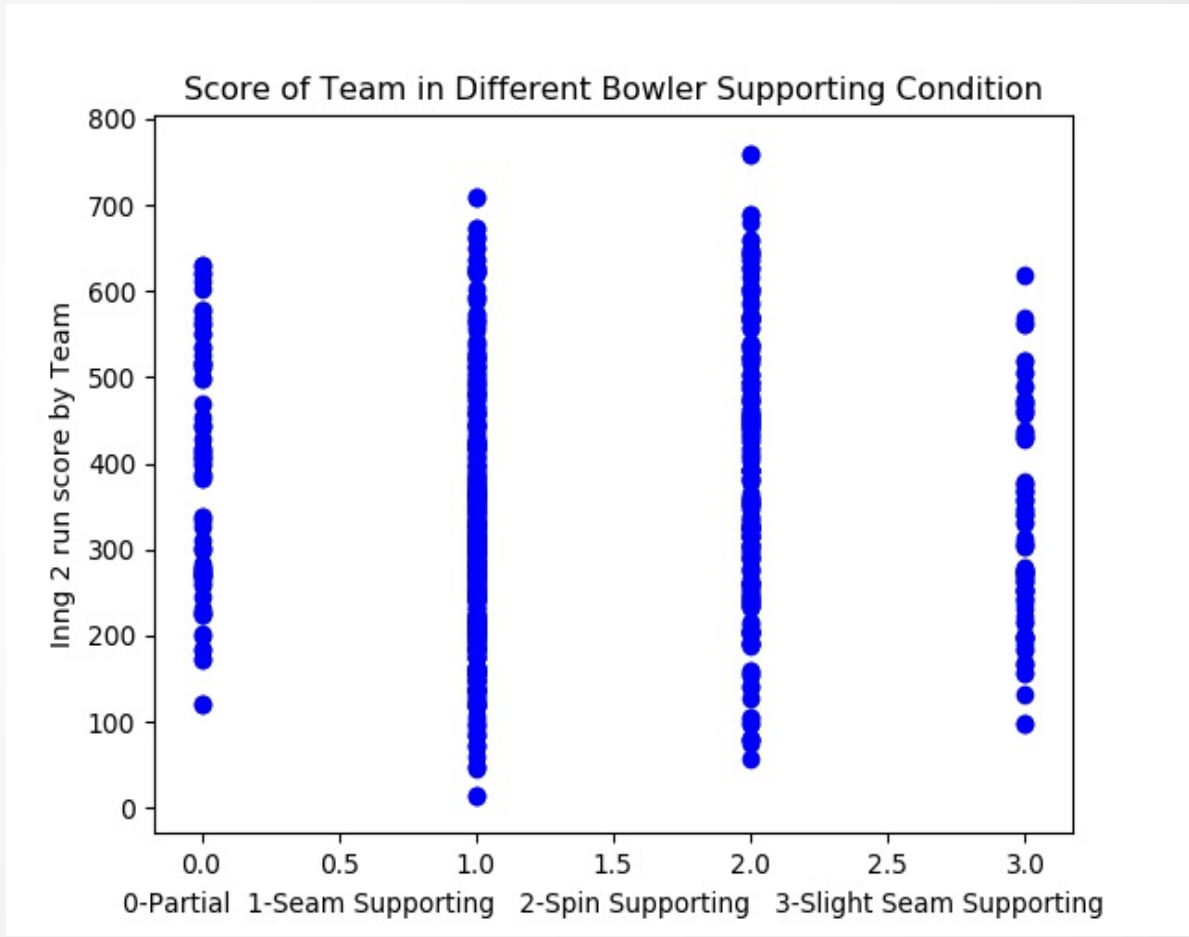
Intuition


Runs scored against spin bowlers
Inng 2 run score by Team
Team Rank Difference
0
100
200
300
400
500
600
700
800
1
2
3
4
5
6
7

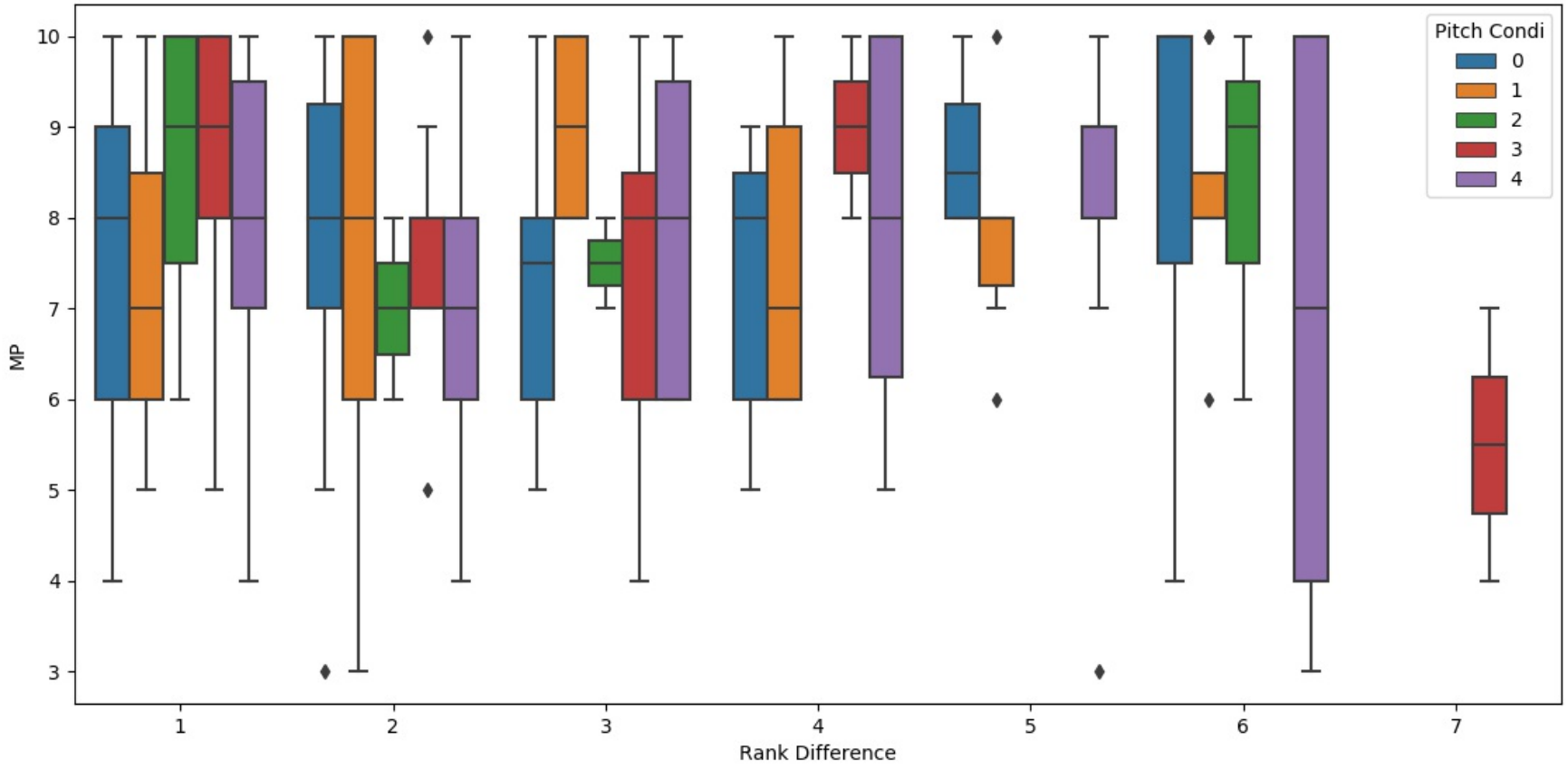

Team Rank Difference v/s Average Runs per Wicket
Average Runs per Wicket
Win Margin Point
0
50
100
150
200
250
0
2
4
6
8
10

Analysis of 2nd Inning

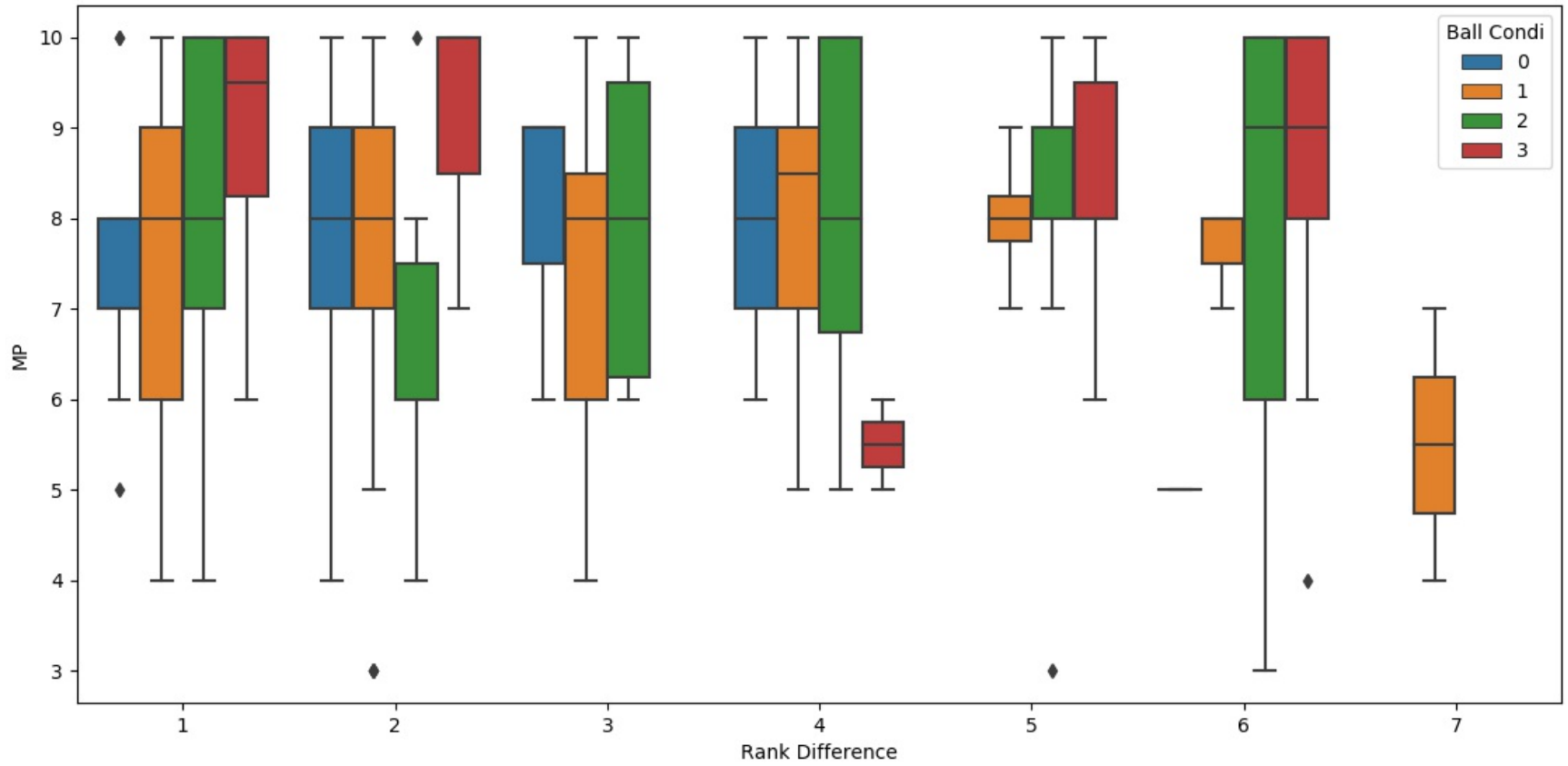




Pitch Based MP Classification



Ball Style based MP Classification



Results and Analysis


```
In [21]: logreg=OneVsOneClassifier(LogisticRegression())
         logreg.fit(X_train,y_train)
         y_predict=logreg.predict(X_test)
         logreg.score(X_test,y_test)
         #accuracy_score(y_predict,y_test)
Out[21]: 0.44827586206896552

In [22]: svc=OneVsOneClassifier(SVC())
         svc.fit(X_train,y_train)
         svc.score(X_test,y_test)
Out[22]: 0.29310344827586204

In [23]: rfc=OneVsOneClassifier(RandomForestClassifier())
         rfc.fit(X_train,y_train)
         #y_predict=logreg.predict(X_test)
         rfc.score(X_test,y_test)
Out[23]: 0.34482758620689657

In [24]: knn=OneVsOneClassifier(KNeighborsClassifier())
         knn.fit(X_train,y_train)
         #y_predict=logreg.predict(X_test)
         knn.score(X_test,y_test)
Out[24]: 0.34482758620689657

In [25]: logreg=OneVsRestClassifier(LogisticRegression())
         logreg.fit(X_train,y_train)
         #y_predict=logreg.predict(X_test)
         logreg.score(X_test,y_test)
         #accuracy_score(y_predict,y_test)
Out[25]: 0.41379310344827586
```

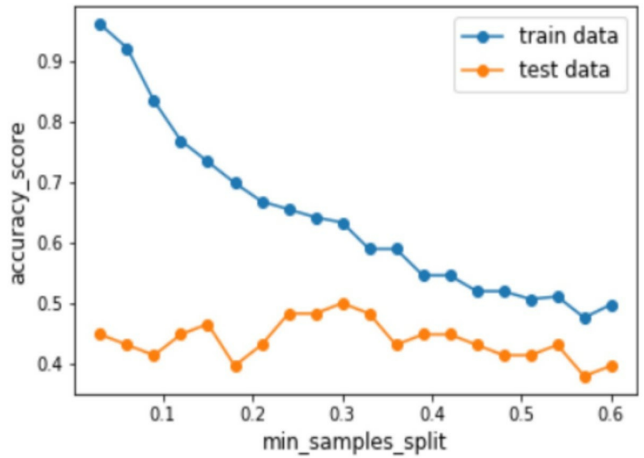
```
In [185]: mlpc=OneVsRestClassifier(MLPClassifier())
          mlpc.fit(X_train,y_train)
          #y_predict=logreg.predict(X_test)
          mlpc.score(X_test,y_test)
Out[185]: 0.20689655172413793

In [271]: df2=df1
          df_new=pd.read_excel(r'C:\Users\Insprion\Desktop\msa\NEW end.xlsx')
          df_toss=df_new.iloc[:,7:11]
          for i in df_toss.columns:
              df1[i]=df_toss[i].values
          X_train2,X_test2,y_train2,y_test2=train_test_split(df1,y,test_size=0.2,random_state=0)
          rfc=OneVsRestClassifier(RandomForestClassifier(random_state=6))
          rfc.fit(X_train2,y_train2)
          y_predict=logreg.predict(X_test)
          rfc.score(X_test2,y_test2)
Out[271]: 0.29310344827586204

In [273]: df_bowl=df_new.iloc[:,32:40]
          for i in df_bowl.columns:
              df2[i]=df_bowl[i].values
          X_train2,X_test2,y_train2,y_test2=train_test_split(df2,y,test_size=0.2,random_state=0)
          rfc=OneVsRestClassifier(RandomForestClassifier(random_state=6))
          rfc.fit(X_train2,y_train2)
          y_predict=logreg.predict(X_test)
          rfc.score(X_test2,y_test2)
          #df_new.iloc[:,32:40]
Out[273]: 0.46551724137931033
```

Random Forest Classifier
Accuracy- 60.34%

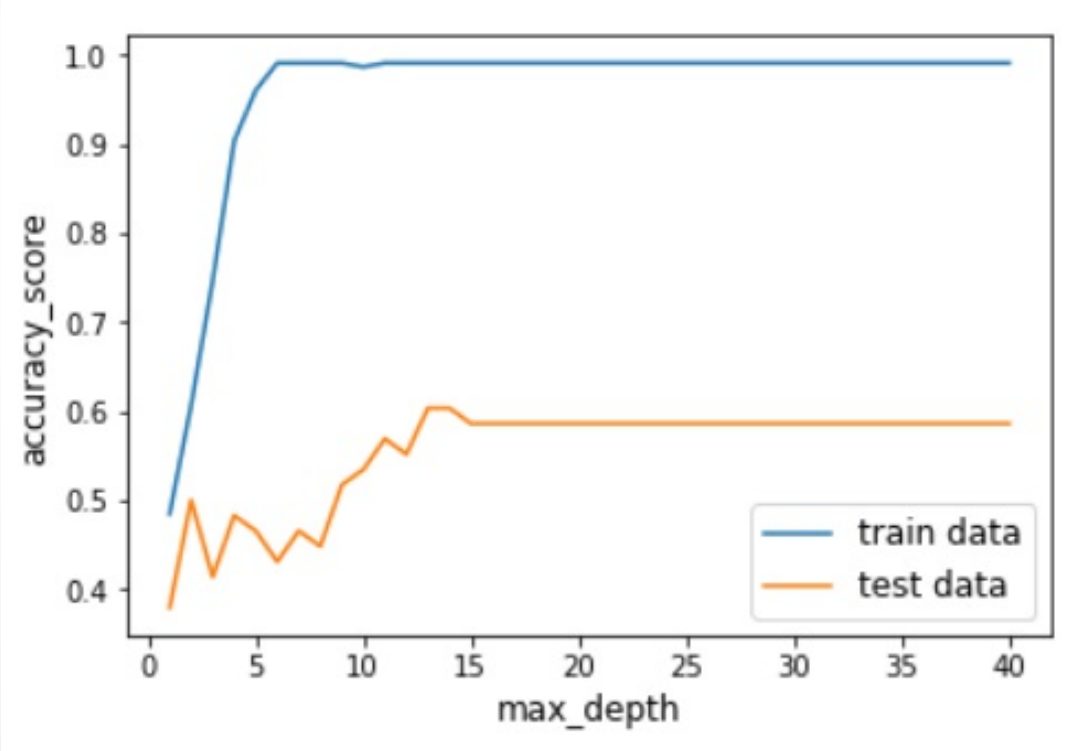

```
Out[67]: <matplotlib.legend.Legend at 0x2d9461320>
```




```
In [ ]: param={'estimator__n_estimators':[20,40,60,80,90,100,400,500],'estimator__max_features':[None,"auto"],
        'estimator__max_depth':np.linspace(4,15,12),'estimator__min_samples_split':np.linspace(0.03,0.21,7),
        'estimator__min_samples_leaf':np.linspace(0.021,0.048,10)}
gs=GridSearchCV(rfc,param_grid=param,scoring="accuracy",cv=5)
gs.fit(X_train2,y_train2)
Y_predict=gs.predict(X_test2)
accuracy_score(y_test2,Y_predict)
```

```
In [83]: rfc=OneVsRestClassifier(RandomForestClassifier(random_state=18,max_depth=13,min_samples_leaf=0.003,))
rfc.fit(X_train2,y_train2)
rfc.score(X_test2,y_test2)
```

```
Out[83]: 0.60344827586206895
```



Conclusion

As long as a steady balance is maintained, one can still be confident about Test cricket not going down the drain that easily and will be continued to be played fair and square.

Hypothesis we Initially Started with is Unique

But, is it a Dead End??

Impact of Increase Competitiveness

01 **Upliftment to Weak Team.**

02 **Engaging Audience.**

03 **Commercial Benefit to Organizer and Broadcasters.**

04 **Reduce the Probability of Biased/ One-Sided Match.**

05 **It proves that whether pitches are doctored or not, the contest between the bat and ball is the King.**

Future Scope of Improvement

Availability of useful Dataset

Finding Appropriate Dataset initially was tough.

01

Adding more Features

Feature scaling appropriately to ensure no overfitting.

02

Tuning Dataset

Enhancing the accuracy by repeatedly tuning the feature.

03

Advanced Algorithm

SOM
Boltzmann Machine
Auto Encoder
Generic Encoder
Recurrent NN
LSTM

04

References:

&#;www.espncricinfo.com

&#;www.cricsheet.com

&#;Research paper on "*Attributes Deciding Cricket Winning*" by Swetha, Professor Saravanan.KN.

&#;Research paper on "*Performance of Cricket Players using Factor Analysis Approach*" by Sricharan Shah, Partha Jyoti Hazarika and Jiten Hazarika .

&#;Research paper on "*Research Directions in Cricket*" by Tim B. Swartz.

&#;Research paper on "*Fairly Random: The Impact of Winning the Toss on the Probability of Winning*" by Gaurav Sood and Derek Willis.

&#;Special thanks to **Mr. Vishal Kumar** for his Assistance.

Xavier School of Management
For the greater good

Thank You

MathSport ASIA