**Q2) Report on the analysis performed on the statistics of ATP matches to predict the player that is going to win the upcoming matches in future.**

This is a summary of the project that aims to predict whether a player wins or loses a match based on his match statistics. The core idea here is to use the Naive Bayesian classification approach to calculate the likelihood of winning and the likelihood of losing a match.

**Data collection:**

The data for the project was collected using manual entry of the statistics into a spreadsheet. To ensure that we do not encounter human error in the process of data collection, auditing of the data was performed. The dataset contained statistics from different length of matches, i.e. the data was a combination of three, four and five set match statistics. The following features of the match were excluded from the data collection for the reasons mentioned against them.

* 1st serves in - This field does not signify if the player won the point or not even if the serve was in.
* Fastest serve – Even if the player has the fastest serve it is not granted the player won the point when he served the fastest.
* Average 1st serve speed – This feature does not undermine the player's performance on points he scored.
* Average 2nd serve speed – This number is not helpful in making a claim that the player with high average 2nd speed would have won most points to win the match.
* Total points won – This feature is a combination of the other parameters such as Net points won etc. Since we have considered the other parameters using this parameter would result in redundancy.
* Distance covered (M) – It provides the information on how much distance one player has covered in the course of the match. A player might cover a lot of court space through the match but it does not guarantee that he won most of the points. A player who can score more Aces would have less Distance covered.
* Distance Covered/Point (M) – This number is similar to the above feature but calculated per point won. This number does not weigh high in our algorithm.

**Algorithm:**

Using the training data collected above. We performed the following algorithm.

1. Divide the training data set into two subsets for “Win” and “Loss” respectively.
2. Exclude the “Result” column in each of the data set.
3. Calculate the Mean and Standard deviation for each of column in the data set.
4. The likelihood for each of the feature inputed was calculated for each of the class “Win” and “Loss”.
5. The likelihood of each feature was calculated using the Gaussian distribution function for that feature column in training data.
6. The prior was No.of.Wins(or Loss)/Total number of match statistics for “Win” and “Loss” class respectively.
7. The probability of the features “Double Faults” and “Unforced errors” was inversed as they had negative effect on the class that we were predicting.
8. Using the prior probability of known outcomes and the likelihood of each feature we calculate the Probability of Win and Probability of loss.
9. These values were compared to predict if the player would win or lose the match.

**Predicting to which set the given data belongs to:**

In order to predict whether the given data belongs to 3 set or 4 set or 5 set match, we used Multivariate Gaussian Distribution function.

**Final Prediction Results:**

For 3 set and 4 set matches we got prediction accuracy as 100 %.

For 5 set matches we got the prediction Accuracy as 60%. (This is because of the less training data for 5 set matches that we obtained from match statistics).