**Stats 517 – Final Project Preliminary Report**

**Abhinav Prabhu Adarapuram**

**Problem Statement, Motivation, Research Goals**

For the final project, I am using the Tennis ATP dataset, which comprises of match details such as tournament date, location, player details, match results, scores, surface used, etc.

This dataset has data which were collected in the past 50 years (1968 – 2018\*). From this I am using the recent 30 years data to make my analysis.

The motive of my project is to analyze whether the left handed players won more matches or the right handed players. This will be the analysis for classification part.

In clustering part, I am going to analyze which country has consistent performance in the past three decades.

## Data Source and Description

I found this interesting dataset from Github – Awesome Public Datasets.

<https://github.com/awesomedata/awesome-public-datasets>

This dataset is updated with recent match’s data (Till September 2018).

## Modeling Process

Initially, this dataset was merged into a single csv file ranging from most recent to last. Few variables which were not required for analysis such asWinner\_seed, Loser\_seed, Tournament\_id, Winner\_entry, Winner\_rank, Loser\_entry, Loser\_rank, Score, surface, etc. The dataset comprised of many missing values. These missing values were filled with the most recurring value of that particular column using R script. Once the imputation is done, the target variable **‘winner\_hand’** was converted into continuous values using the ‘revalue’ function.

After the data cleaning, I imported the csv file to Jupyter Notebook to perform the analysis. To start with, I worked on the classification part, where the machine was trained with the data and various models such as Linear Regression, K nearest neighbors, Decision Tree, Navye Bayes and Random Forest were used.

**Preliminary Results**

On training the data with various models, it is observed that Decision Tree produced better results in both training as well as testing.

|  | **Types of Models** | **Training percentage** | **Testing percentage** |
| --- | --- | --- | --- |
| **0** | Logical Regression | 84.33 | 84.91 |
| **1** | K Nearest Neighbour | 87.10 | 85.55 |
| **2** | Decision Tree | 100.00 | 99.42 |
| **3** | Naive Bayes | 74.63 | 74.94 |
| **4** | Random Forest | 100.00 | 98.53 |

**Future Work**

Next, I have to proceed with the clustering part, where performance of every country in the last three decades will be analyzed.