Data Analytics

- The data used is from https://github.com/JeffSackmann/tennis_atp and atp_matches from 2001 to 2020
- The dataset consists of 59789 rows x 50 columns
- We visualized the frequencies of **'winner_rank', 'loser_rank', 'winner_age', 'loser_age', 'winner_ht', 'loser_ht', 'w_svpt', 'l_svpt'** with the help of bar graphs.
- We visualized the box plot graph for surfaces against the aces.
- We also visualized the number of Grand Slam wins by countries, most number of aces by player, age of Grand Slam champions, total Grand Slam match wins & losses by countries.
- We found the player effectiveness for **Roger Federer, Rafael Nadal, Novak Djokovic**

Predictive Modeling

- We separated the winner player and loser player columns.
- We visualized the player rank points in both win & lose players case along with the players' age
- We visualized the violin plot graph for surfaces against the ace, and against their count, player hand frequency in case of wins & loses.
- We created the atp_clean dataset with non null ace values as our training data, and then selected the following columns **'player_hand_l', 'player_hand_u', 'player_ht_diff', 'player_age_diff', 'player_rank_diff', 'player_rank_points_diff', 'surface_carpet', 'surface_clay', 'surface_grass', 'surface_hard', 'tourney_level_A', 'tourney_level_D', 'tourney_level_F', 'tourney_level_G', 'tourney_level_M', 'result'** for correlation graph
- After finding the correlation graph, the final features used for training were **'player_hand_l', 'player_hand_u', 'surface_carpet', 'tourney_level_D', 'tourney_level_F'.**
- We applied Logistic Regression and found the cross val score to be **0.6518974**
- We plotted the ROC curves for it & found the AUC value to be **0.71274224**
- We then applied Decision Tree algorithm and found the cross val score as **0.56451658**
- We then applied Random Forest algorithm and found the best score to be **0.653569** along with best parameters set as **gini criterion, max_depth as 7 & number of estimators as 40.**
- We also calculated the confusion matrix for this and found the accuracy to be **64.94%**.