ReadME

CS 677 Project

Tennis Data Analysis : Winner Predictor

Project Description :

This is a Machine Learning project focused on analyzing data of tennis matches (from years 2015-2019) and creating ML classifiers to predict the winner based on the match statistics and analyse a player's performance over the years.

Dataset description :

I wrote a script to download raw data from the data source. The initial raw data consisted of yearly data from years 2015-2019 having approx 4000 rows and 94 columns for each year.

After cleaning and preprocessing the data, the final shape of the dataset was : (19969, 35).

Data analysis and Machine learning :

After data preprocessing I ran some ML classifiers on the dataset to find the model that gives me the best accuracy.

Data was split into 70:30 for training and testing
Accuracy of Various classifiers I used

Models that required Scaling

1. KNN (best N=13) : 97.58
2. Linear SVM : 98.73
3. Gaussian SVM : 98.61
4. Polynomial SVM (deg=2) : 93.72
5. Linear Discriminant Analysis : 98.75
6. Quadratic Discriminant Analysis : 96.46

Models without Scaling

1. Logistic Regression : 98.81
2. Gaussian Naive Bayes : 96.68
3. Decision tree : 97.76
4. Random Forest (n=30) : 98.05

Logistic regression was the best model for the winner predictor.

So using the Pickle package I created a model.pkl file of Logistic Regression and using Python Flask I created a Basic web application (User Interface) for the winner predictor.

Model Testing :

If you want to test the model for predicting the winner, just run the ‘Final code.ipynb’ code in jupyter notebook and run line no. 14 - ‘In [14]’ with the names of two players from the list provided in the cell right above 14 (In [13]).