**Tennis Winner Predictor**

**Ana Rudić, Milan Šalić, Sreten Stokić**

1. Motivation

Tennis is very popular sport and millions of people watch matches regularly trying to predict outcome of the match. We wanted to create a program, that is good to predicting who the winner will be, based on various parameters.

2. Research questions

Goal is to predict which player will win the match based on various parameters. There are parameters about tournament, such as name, surface, level, number of players, current tournament stage, maximum number of sets and number of match. There are also parameters about players, such as id, seed on this tournament, way he qualified for the tournament, stronger hand, height, atp rank with atp rank points, his origin country and age. Those parameters are known for both players of the match.

The other goal was to make similar calculations about how match ended, based on parameters from above and some others which represent statistics of the match. Those parameters are match duration in minutes, number of aces, double faults, number of successfully first service, number of won points on first service, number of points won on second service, number of saved break points and number of faced break points. Those statistics are included for both players.

There are nearly 45000 matches in the dataset dating from 2000 to today.

3. Related work

This problem is researched by many people. Most of them used neural network to access to t this problem. But actually there is no related work for this project and this access.

4. Methodology

At the very beginning, we were planning to solve the problem using multiply linear regression, but after some discussions, we realized that there are probably more efficient ways to do that.

We tried neural network, but it gave us 50% of successfully predicted matches, which is equal to lucky guess.

Finally we have chosen classification approach and it gave us way better results.

5. Discussion

First of all, we needed to modify dataset in the way our approach demands. In original dataset, there are winner fields and loser fields and that would mean that first player won every single match. We doubled all matches so we could exchange places of two players so it would be irrelevant on which position who stands.

There were many categorical variables that needed to be converted to numbers. Then we normalized all data to interval between 0 and 1 which really improved results.

We used Support Vector Machine from scikit-learn library. We chosen ‘rbf’ as kernel type, 0.002 as kernel coefficient (gamma), 0 as independent term and 9 as random state.

We fit model with 2/3 of data and tested it with rest.

Metrics.accuracy\_score is used to measure percentage of successfully predicted matches.

For predicting future matches, without knowing any match’s statistics, we successfully predicted 72% of winners.

For guessing winners of the matches, with knowing match’s statistics, there is 84% of successfully predicted winners.

Knowing the fact that tennis is unpredictable sport and there are many surprises at tournaments and retired matches, 72% sound pretty solid accuracy.

Also tennis matches are known as really intensive where one hit or one mistake can decide the winner, so 84% looks like good accuracy too, even that match’s statistics are used to predict winner. On the other hand, statistics are not always on the winner side.

6. References

No referenced works.

<The maximum number of pages is 4>