University of Liège

PROJ0016 - Big Data Project

Review 6 - Final improvements and prediction

Master 1 in Data Science & Engineering

Authors:
Antoine Louis
Tom Crasset
Maxime Lamborelle

Professors :
G. Louppe
P. Geurts
B. Cornelusse

Academic year 2018-2019

1 Reminder of previous milestone

In order to train a machine learning model, we first aim to construct new statistical features for both players of a match, reflecting their recent performances. To do so, our method consists in making a weighted mean of their statistics over their previous matches using "time discounting", meaning that each match enrolled in the mean is given a weight that depends on the elapsed time between the match time and the computation time of the new feature. Table 1 shows the basic statistical features on which we perform the weighted averages. It concerns eight features of the Jeff Sackmann Dataset [1], that is, the dataset that we use as a starting point, containing data on approximately 70000 tennis matches between 1993 and 2018.

Players details	**Match details**	**Player statistics for the match**
Name	Tournament name	Number of service points
Age	Tournament type	Number of first serves in
Nationality	Date	Number of first serve points won
Handedness	Draw size	Number of second serve points won
ATP rank	Best of	Number of aces
ATP points	Surface	Number of double faults
	Round	Number of break points faced
	Score	Number of break points saved
	Duration	

Table 1: Features of Jeff Sackman's dataset

2 New features construction

When looking at the features concerning the players statistics, one can noticed a drawback. All the features are expressed in terms of "number of", which means that the new average statistics will also be expressed that way. This might be an inconvenient when putting these statistical means face-to-face as inputs of a machine learning model, as they do depend on another feature : the average match duration of a player. Hence, more representative features can be computed in terms of percentage. We define :

1. Service points % $= \frac{\text{Number of service points of Player A}}{\text{Number of service points of Player A + Number of service points of Player B}}$
2. 1st serve % $= \frac{\text{Number of first serves in}}{\text{Number of service points}}$
3. 1st serve points won % $= \frac{\text{Number of first serve points won}}{\text{Number of first serves in}}$
4. 2nd serve points won % $= \frac{\text{Number of second serve points won}}{\text{Number of service points - Number of first serves in}}$
5. Aces % $= \frac{\text{Number of aces}}{\text{Number of service points + Number of aces + Number of double faults}}$
6. Double faults % $= \frac{\text{Number of double faults}}{\text{Number of service points + Number of aces + Number of double faults}}$
7. Break points faced % $= \frac{\text{Number of break points faced}}{\text{Number of service points + Number of aces + Number of double faults}}$
8. Break points saved % $= \frac{\text{Number of break points saved}}{\text{Number of break points faced}}$

These eight new features will replace the ones expressing player statistics in the initial Jeff Sackmann Dataset [1] (see Table 1). It is these features that will be considered in the weighted means.

3 Surface weighting

In addition to time discounting, we wanted to take into consideration the surface on which each match enrolled in the weighted averages was played, giving a greater weight to matches played on clay. That way, we will compute representative statistics about a player current performance, by putting an emphasis on its recent performances on clay. The question is knowing which weight to give to each one of them.

Obviously, matches on clay will be given a maximal weight, that is a weight of 1. Concerning the other surfaces, we compute correlation coefficients between clay and each surface. These coefficients express the winning correlation between surfaces, that is, the correlation between winning a match on clay and winning a match on another surface. This approach, inspired by Sipko [2], can easily be executed using our dataset. Indeed, for each pair of surfaces (A, B), we can calculate their correlation by applying the following formula :

$$\rho_{A,B} = \frac{cov(A,B)}{\sigma_A \sigma_B} = \frac{1}{N} \frac{\sum_{k=0}^{N} (x_{A,k} - \mu_A)(x_{B,k} - \mu_B)}{\sigma_A \sigma_B} \tag{1}$$

where

$$\begin{cases} N & \text{is the number of considered players} \\ x_{S,k} & \text{is the percentage of matches won by player } k \text{ on surface } S \\ \mu_S & \text{is the mean percentage of matches won on surface } S \\ \sigma_S & \text{is the standard deviation of percentage of matches won on surface } S \end{cases}$$

Computing Equation (1) between clay and the other surfaces, namely Hard, Grass and Carpet, yields the coefficients presented in Table 2. These coefficients will be used directly as the surface weights when computing the weighted means.

	Clay	**Carpet**	**Grass**	**Hard**
Clay	1	0.257	0.277	0.491

Table 2: Correlation coefficients between clay and other court surfaces

Now, the problem is knowing how to weight each one of our computed weights, namely the timing weight and the surface weight, in a final weight that will be attributed to the matches. After having tested different combinations and analysed the results on the further predictions, described later, the right compromise seems to give each past match $i = 1, ..., n$ of a player a total weight described by the following equation :

$$W_i = 0.95 * w_{i,time} + 0.05 * w_{i,surface} \tag{2}$$

It seems indeed logical to give more importance to the timing weight than to the surface weight, as what we really want to put forward is the current performance of a player. The surface weighting only makes our model more accurate for the specific problem of predicting outcomes of matches on clay.

Then, by looping over each match of our dataset, we compute for both players new statistics representing a weighted average of the players' match statistics over all their past matches before the time of their meeting. Formally, for a given statistical feature x, we compute the new feature:

$$\overline{x} = \frac{\sum_{i=1}^{n} W_i x_i}{\sum_{i=1}^{n} W_i} \tag{3}$$

where x_i is the value of the feature x in the i^{th} match, and W_i represent the weight given to the i^{th} match enrolled in the average and is given by Equation (2).

4 Summary of the features

Table 3 shows the final features that will be taken into account in the training step. All these features were standardized in order to have them all in the same order of magnitude.

Final Features
Difference in ages
Difference in ATP rankings
Difference in ATP points
Same handedness
Difference in average match duration
Difference in average winning percentage
Difference in best-of average
Difference in average percentage of aces
Difference in average percentage of double faults
Difference in average percentage of service points
Difference in average percentage of first serves
Difference in average percentage of first serves won
Difference in average percentage of second serves won
Difference in average percentage of break points faced
Difference in average percentage of break points saved

Table 3: Final features used to train our model

5 Model testing

In order to get the best model as possible, multiple supervised learning techniques were tested, namely the Random Forest algorithm, Logistic Regression, Support Vector Machine (SVM) and Neural Networks. The results given by the tuned algorithms are described in Table 4.

At the end, we decided to keep the Random Forest model as this was the one achieving the best prediction outcomes among all the models. Moreover, this classification algorithm also gives some measure of the certainty of an instance belonging to a class, which can be used as the match-winning probability.

Model	**Hyperparameters**	**Considered features**	**Test accuracy**
Random Forest	n_estimators = *2000* max_depth = *10* min_samples_split = *5* min_samples_leaf = *1*	All 15 features	66.84%
Logistic Regression	solver = *liblinear* penalty = *l2*	All 15 features	66.56%
MLP Classifier	solver = *sgd* hidden_layer_sizes = *(20,)* activation = *tanh* learning_rate = *0.05* momentum = *0.6*	All 15 features	66.47%
SVC	kernel = *rbf* kernel = *linear* kernel = *poly*	All 15 features	66.38% 66.36% 65.38%

Table 4: Test accuracy comparison between different machine learning algorithms

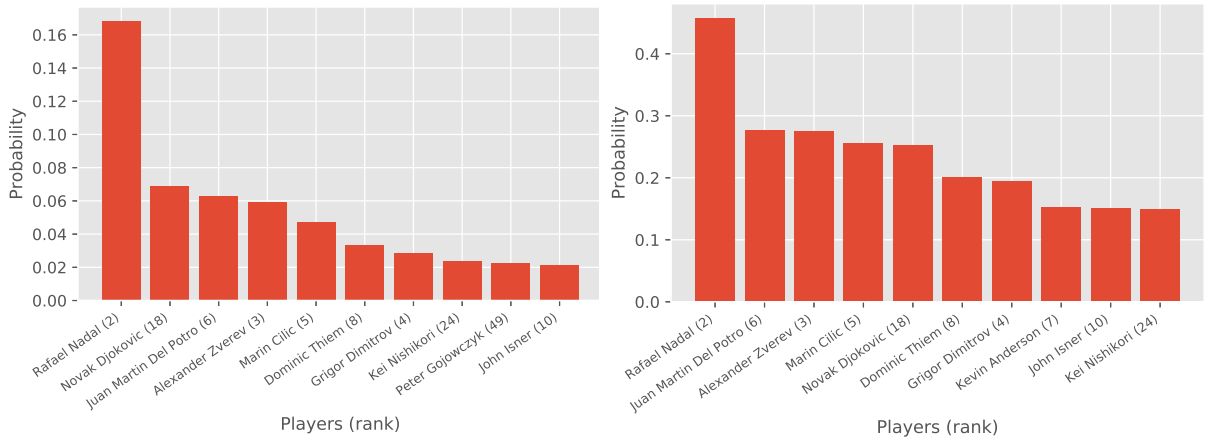
6 Predicting on past French Opens

In order to evaluate the accuracy of our model, it was tested on the three last French Opens. As a reminder, we used Monte-Carlo simulations, where we randomly sample a certain number of possible draws and simulate them according to our predictions. Eventually, each simulation will output a unique final winner. By computing, for each player of the tournament, the number of times that he comes out as a winner from some simulations on the total number of simulations, we can estimate which players will have the highest probabilities to win the tournament. To be even more precise, a similar process

can be applied in order to compute the probability for each player to reach a certain round. These precise probabilities will allow to analyse in depth the accuracy of our model when tested on different past tournaments.

6.1 Testing on the French Open 2018

By applying our final model on the French Open 2018, we obtained very satisfactory results. First, as can be seen in Figure 1a, we predict Rafael Nadal, the actual winner of the tournament, as the player who had the highest probability to win the Grand Slam. Moreover, when looking at the actual quarterfinals draw in Figure 2, and comparing it to the eight players that we predicted to have the highest probabilities to reach the quarterfinals (QF), shown in Figure 1b, we notice that six out of the eight players actually did reach the QF.



(a) Probability to win the tournament

(b) Probability to reach the Quarterfinals

Figure 1: Top 10 players having the highest predicted probabilities to (a) win the tournament and (b) reach the QF of the French Open 2018. The number in parenthesis corresponds to the player rank.

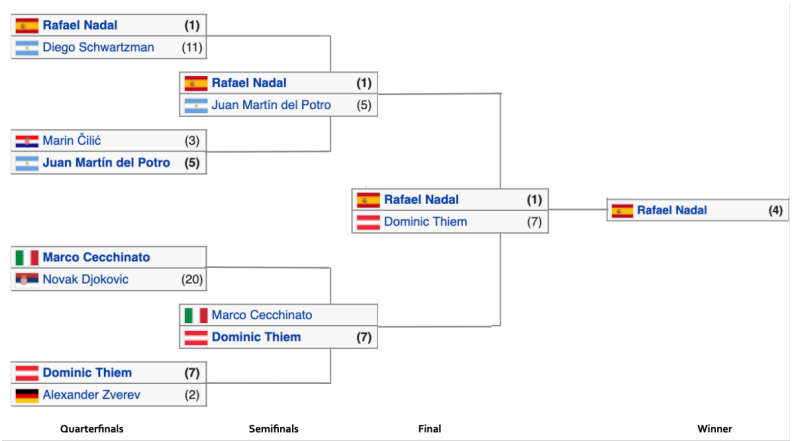


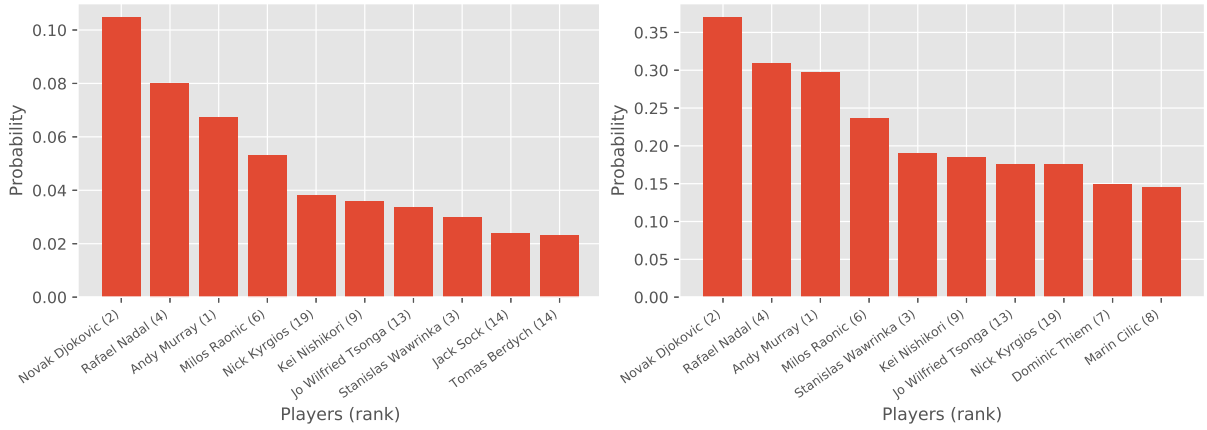
Figure 2: Draw from the Quarterfinals of the French Open 2018. The number in parenthesis corresponds to the seed number.

6.2 Testing on the French Open 2017

Here, we predicted Rafael Nadal, once again actual winner of the tournament, as the player with the second highest probability to win the Grand Slam, as can be seen in Figure 3a. Before him, Novak Djokovic, eliminated in quarterfinals by the Austrian Dominic Thiem, had 2% more chance of winning than Nadal, according to our model. However, this difference in winning probability between the two best

predicted players is quite small compared to the one of 2018. Indeed, even though Novak Djokovic has here 2% more chance of winning the tournament than Rafael Nadal, this difference in winning probability between Nadal and Djokovic amounted to more than 10% in 2018, which dips the gap more clearly here.

Then, by analysing the quarterfinals draw, shown in Figure 4, and comparing it with our predicted players having the most chances to reach the QF, five out of eight predicted players actually reached the QF. Notice that the two next players that had the biggest probabilities to win were Dominic Thiem and Marin Cilic, who both reached the quarterfinals.



(a) Probability to win the tournament

(b) Probability to reach the Quarterfinals

Figure 3: Top 10 players having the highest predicted probabilities to (a) win the tournament and (b) reach the QF of the French Open 2017. The number in parenthesis corresponds to the player rank.

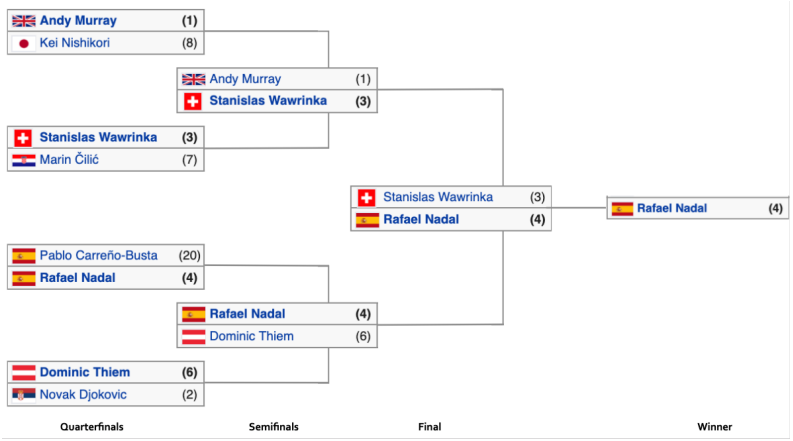


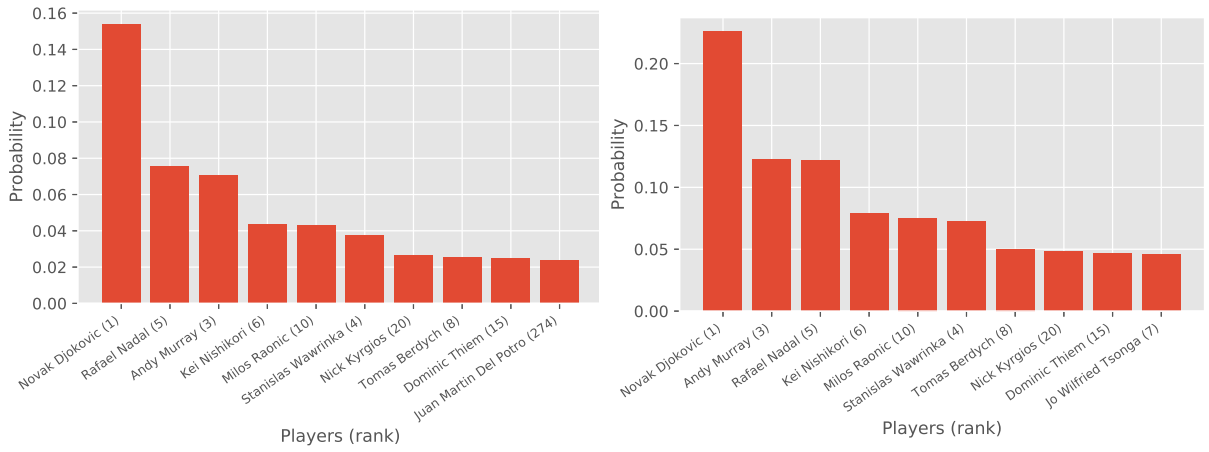
Figure 4: Draw from the Quarterfinals of the French Open 2017. The number in parenthesis corresponds to the seed number.

6.3 Testing on the French Open 2016

Finally, we made one last test on the 2016 French Open. As can be observed in Figure 5a, we predicted Novak Djokovic, the actual winner of the Grand Slam, as the player with the highest probability to win, outperforming the others with a difference in winning probability of about 8% on the second player with highest probability, Rafael Nadal. Moreover, what is interesting to notice in Figure 5b is that, even though Nadal has the second highest probability to win the tournament, the two players that have the highest probabilities to reach the final are Novak Djokovic, obviously, and Andy Murray, which actually represents the real final. This means that, on all the simulation runs, Murray reached the final more often than Nadal did, though the latter won the final a greater number of times than Murray.

Also note that Rafael Nadal, who appears in the top 3 of all our predicted probabilities for the different rounds, retired during the third round due to a wrist injury. This abandonment obviously affects the

progress of the tournament a lot, as it actually concerns the best tennis player on clay, winner of 11 different French Opens. This misfortune factor is something that is difficult to control and can sometimes distort a lot the predictions, even with the best prediction model in the world in hand.



(a) Probability to win the tournament

(b) Probability to reach the Final

Figure 5: Top 10 players having the highest predicted probabilities to (a) win the tournament and (b) reach the Final of the French Open 2016. The number in parenthesis corresponds to the player rank.

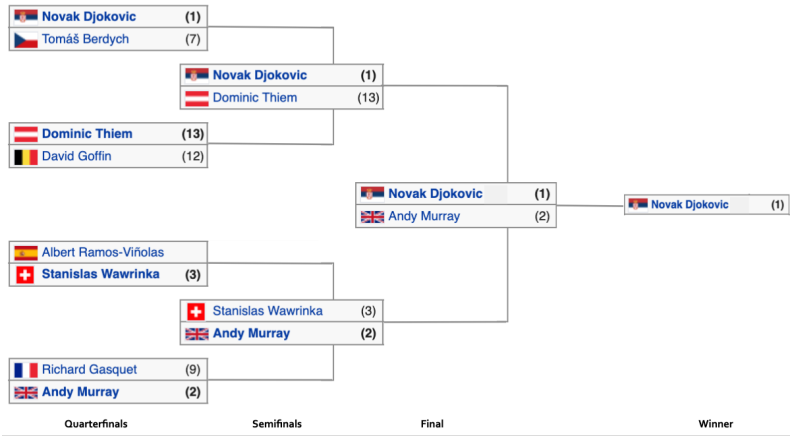


Figure 6: Draw from the Quarterfinals of the French Open 2016. The number in parenthesis corresponds to the seed number.

7 Predicting the winner of the 2019 French Open

7.1 Players selection

Before finally predicting the winner of the 2019 French Open, we first need to determine the 128 players that will participate in the future tournament.

7.1.1 Wild-cards

The first difficulty concerns the wild-cards, eight special invitations offered to eight players currently ranked outside the TOP 104, and allowing them to access the Grand Slam without playing qualifying matches. Since this year, a new process set up by the French Tennis Federatio (FTF) will allow two French players to receive wild-cards for the French Open. The first card will be offered to the French player who has won the most ATP points after a "Race", comprising ten specific French tournaments, if

this player is not already selected for the final draw. Otherwise, the wild-card will be given to the first player in the "Race" ranking who can not access it. The second invitation will be awarded to the highest ranked French player of the ATP ranking Race to London, stopped on May 13. The ATP Race is another ranking established by the ATP which, unlike the technical ATP ranking which is updated every week by taking into account the points earned over the previous 52 weeks, only takes into account the points earned during the current season.

Then, the FTF grants each year a wild card to an Australian player and to an American player. In return, the Australian Tennis Federation and the American Tennis Federation respectively gives each year a wild-car to a French player at the Australian Open and the US Open. To determine these players, we made the simple hypothesis that the players that are the most likely to be chosen are probably the best ranked Australian and American players outside the TOP 104. This concerns the Australian Alexei Popyrin, 19 years old and currently ranked #120, and the American Ryan Harrison, 26 years old and ranked #109.

With some exceptions, the remaining wild-cards are most of the time attributed to French players. By considering that the FTF gives a lot of importance on very young players demonstrating good performances, the players that are the most likely to receive these cards according to us are Antoine Hoang, Quentin Halys, Maxime Janvier and Elliot Benchetrit, as they seem to correspond to the ranking-age ratio to which the FTF values so much. Note that these predictions are personal estimations resulting from the analysis of the previous selected players on past French Opens. The ideal solution would be to train a machine learning model using data about the players who have received a wild-card in the past French Opens, in order to predict which players will be nominated this year. However, the amount of training samples would be extremely limited, as there have only been 46 French Open since the creation of the Association of Tennis Professionals (ATP) in 1972, and very limited data was available before the 2000's. For these reasons, intuition seems to be the only viable option.

To summarise, Table 5 shows the eight players that we estimate to receive a wild-card for this edition of the French Open.

Players	**Nationality**
Gregoire Barrere	FR
Corentin Moutet	FR
Antoine Hoang	FR
Quentin Halys	FR
Maxime Janvier	FR
Elliot Benchetrit	FR
Alexei Popyrin	AUS
Ryan Harrison	US

Table 5: Players prone to receive a wild-card for the 2019 French Open according to our estimations.

7.1.2 Qualifications

The second difficulty is to predict the sixteen players who will pass the qualifications. The qualifications are a pre-tournament that begins five days before the real tournament and also welcomes 128 players, ranked outside the TOP 104. The sixteen players who will win the final round of the qualifications, that is the round of 32, will access the final tournament. As in the final Grand Slam, there are also "seeds" in the qualifications, where the same distribution applies when creating the draw. To predict the sixteen players, we can use our tournament-predictor model on the 128 best players outside the TOP 104. We deliberately choose not to exclude the players that we estimate to receive a wild-card, shown in Table 5, if they are part of the TOP 104-232. This way, we ensure to take them into account if our estimations on the wild-card attributions turn out to be wrong. After applying our model on the considered players, we were able to predict the sixteen players having the highest probabilities to pass the qualifications. These players are shown in Figure 7. One can notice that Alexei Popyrin and Ryan Harrison stand among them, which strengthens our choice to consider these two players, whether through the attribution of wild-cards or through the qualifications. Hence, we will take into account the two next players having the highest probabilities to reach the final draw, namely Marcos Baghdatis and Yasutaka Uchiyama.

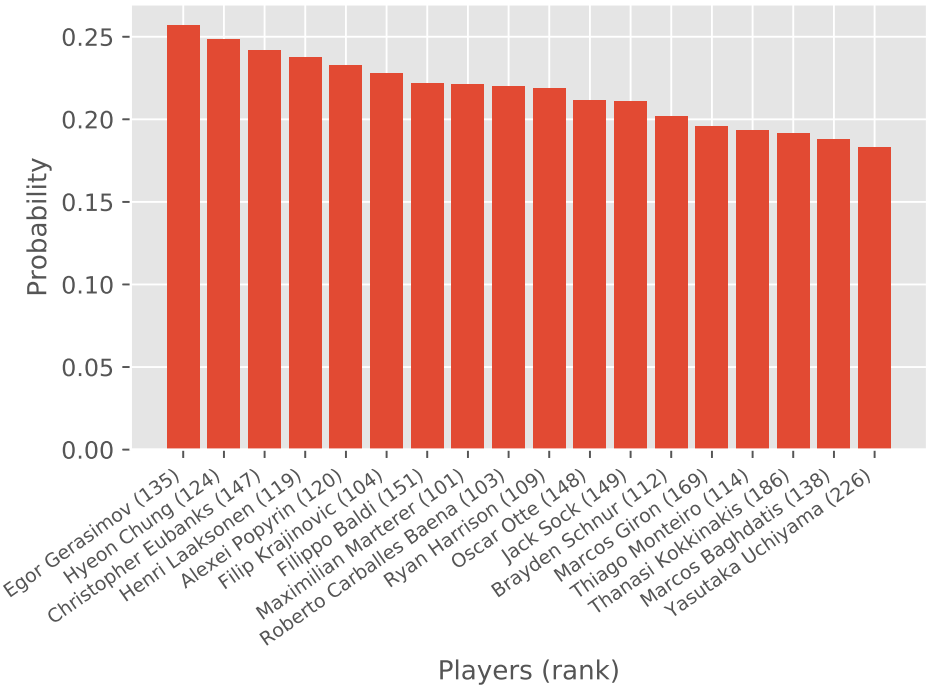


Figure 7: Top 18 players having the highest predicted probabilities to pass the qualifications of the French Open 2019. The number in parenthesis corresponds to the player rank.

7.1.3 TOP 104 players

Lastly, the 104 remaining players are the easiest to consider, as they simply are the TOP 104 players according to the ATP ranking six weeks before the beginning of the French Open. Therefore, the 2019 French Open starting on Sunday, May 26, the first 104 selected players will be the best 104 players according to the ATP ranking of the week of April 14.

7.1.4 Injured players

When finally having our 128 players, one needs to consider the ones that could possibly not be able to play the French Open due to an injury. From early April to the end of May, there are a total of 12 different ATP tournaments, all happening on clay, including the Monte-Carlo Masters, the Barcelona Open, the Madrid Open and the Italian Open, which are four ATP Masters 1000 attracting a lot of good players due to the high prize money involved. Therefore, we must consider the fact that some of the selected players may suffer from an injury in the meantime. The website `tennisexplorer.com` [4] helps doing so by giving a complete list of all the tennis players currently injured. It turns out that, among the 128 players, seven of them currently suffer from an injury at the time of writing, leaving some doubts about their ability to play the incoming French Open. The names of these injured players are given in Table 6.

If some of these players were to withdraw before the first round of the tournament, their substitutes would be picked among the "Lucky losers". This is the status given to a player coming from the qualifications who lost just before accessing the final draw. Among the 16 players concerned, the top eight players must stay in Paris and a draw is made to determine the order in which they would have to replace the forfeit player. Hence, determining which players might replace the injured ones comes down to determining which players will lose the final round of the qualifications. This can be retrieved from our predicted results concerning the qualifications. Indeed, we only have to consider the next sixteen players having the highest probabilities to win the final round of the qualifications and, among them, the eight best ranked ones will represent the possible substitutes. The names of these predicted substitutes are given in Table 6.

At this writing, there are three weeks left until the beginning of the tournament, which makes it completely possible for a player to recover from a minor injury. Therefore, we will consider two scenarios : one where

all the selected participants are able to play the tournament, and the other where the currently injured players are replaced by their predicted substitutes.

Injured players	Substitute players
Gilles Simon (26)	Denis Istomin (105)
Jo-Wilfried Tsonga (102)	Marcel Granollers (107)
Damir Dzumhur (54)	Tennys Sandgren (111)
Jiri Vesely (95)	Sergiy Stakhovsky (115)
Pablo Carreno Busta (27)	Matthias Bachinger (127)
Gael Monfils (19)	Bjorn Fratangelo (132)
Matthew Ebden (53)	Marco Trungelliti (139)
	Stefano Travaglia (154)

Table 6: Injured players among the official participants three weeks before the beginning of the Grand Slam. The number in parenthesis corresponds to the player rank.

7.2 Predictions

To compute recent statistics for the 128 players of this edition, we had to collect new data about recent tennis matches. Indeed, the Jeff Sackman Dataset [1] that we used until here only lists matches from 1993 to 2018. It neither contained the end of year matches in 2018 nor the 2019 matches. Thus, we scraped the official ATP website using the Python scripts from [3] with some modifications to fit our personal needs, and were able to complete our dataset with the most recent tennis matches.

Now, after considering the 128 players that are the most likely to play this edition of the French Open, we can finally use our tournament-predictor model on these particular players to predict the winner of the French Open 2019. And the predicted winner is... Rafael Nadal ! He would then pick up his 12^{th} title for the Grand Slam in Paris.

However, this year, competition seems fierce. Indeed, Novak Djokovic, the current best player in the world, has the second highest winning probability by just 0.0013% beneath Nadal's one, as shown in Figure 8. When analysing their probabilities to progress in the tournament, shown in Figure 9, we would be tempted to say that these two players probably represent the most likely Final.

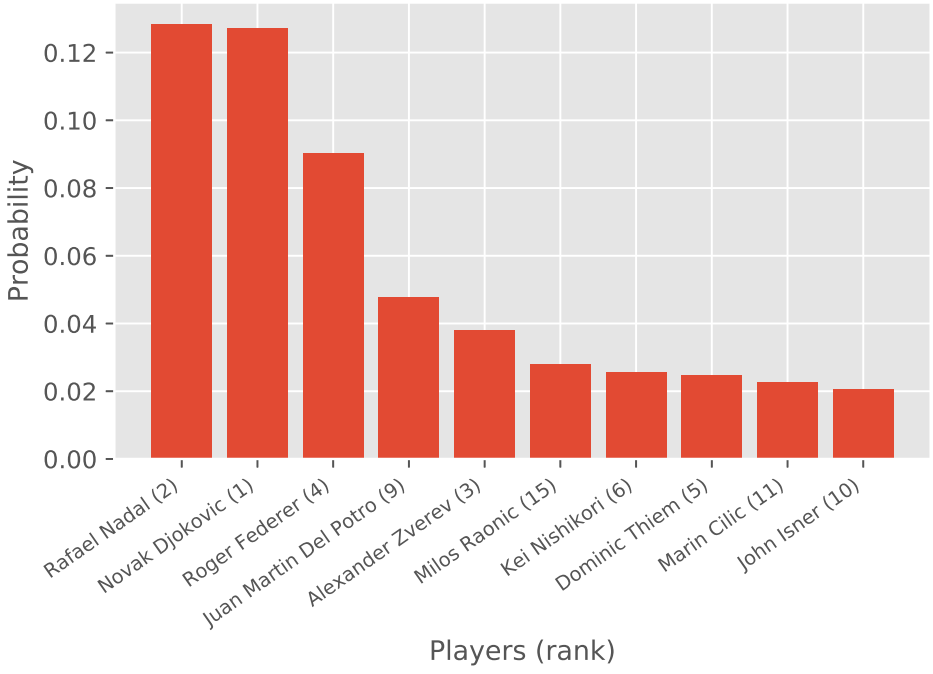
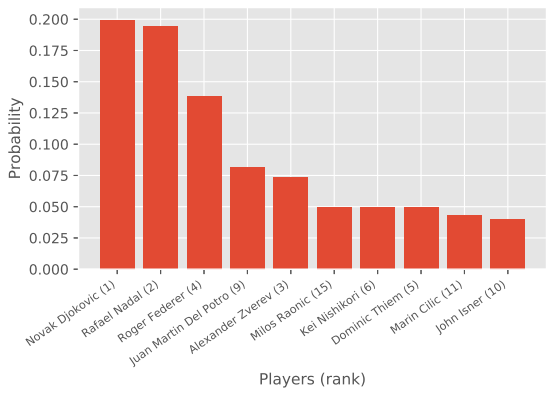


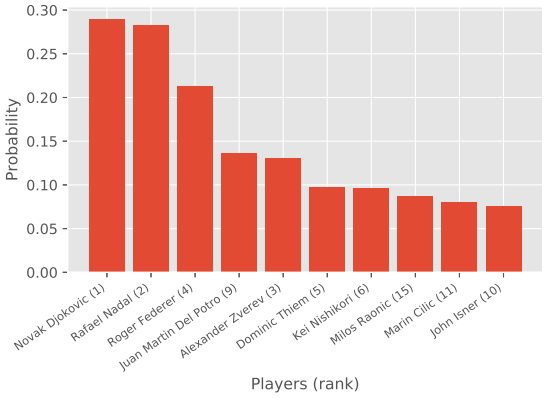
Figure 8: Top 10 players having the highest predicted probabilities to win the French Open 2019. The number in parenthesis corresponds to the player rank.

The Semifinals would concern N. Djokovic and R. Nadal facing either R. Federer or J.M. Del Potro, depending on the draw, as shown in Figure 9b. Note that A. Zverev stands very close to J.M. Del Potro in terms of probability to reach the Semifinals.

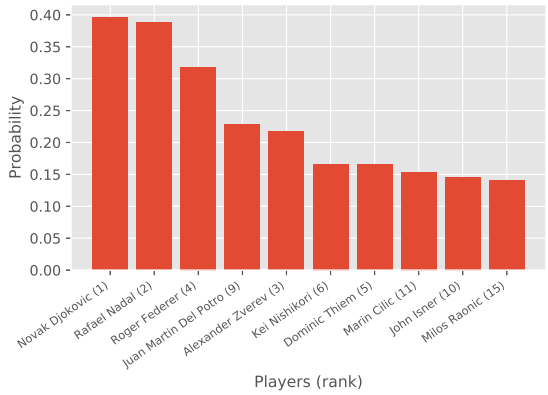
Concerning the Quarterfinals, the eight players that have the most chances to reach them according to our predictions are N. Djokovic, R. Nadal, R. Federer, J.M. Del Potro, A. Zverev, K. Nishikori, D. Thiem and M. Cilic, as shown in Figure 9c. Hence, a predicted draw from the Quarterfinals is presented in Figure 10.



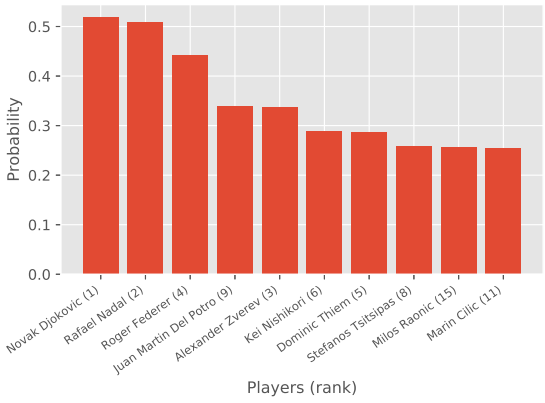
(a) Probability to reach the Final



(b) Probability to reach the Semifinals



(c) Probability to reach the Quarterfinals



(d) Probability to reach the Round of 16

Figure 9: Top 10 players having the highest predicted probabilities to reach the last rounds of the French Open 2019. The number in parenthesis corresponds to the player rank.

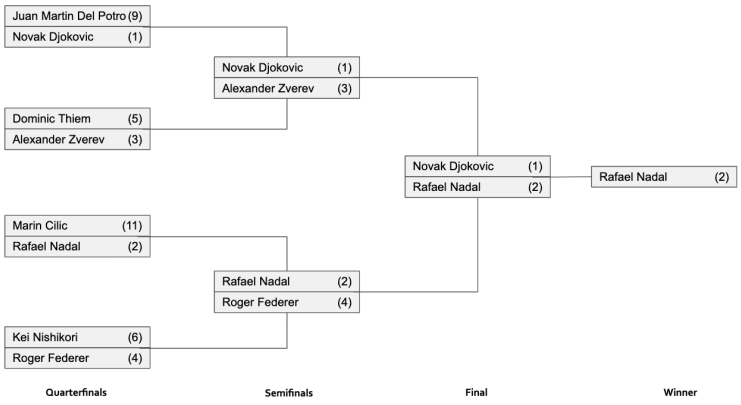


Figure 10: Predicted draw from the Quarterfinals of the French Open 2019. The number in parenthesis corresponds to the player rank.

References

[1] Sackmann J. (2015-2018). tennis_atp, *GitHub repository.* Retrieved from https://github.com/JeffSackmann/tennis_atp.

[2] Sipko M. (2015, June). Machine learning for the prediction of professional tennis matches. *MEng computing - Final Year Project, Imperial College London.*

[3] Lin K. (2017). atp-world-tour-tennis-data, *GitHub repository.* Retrieved from https://github.com/serve-and-volley/atp-world-tour-tennis-data.

[4] Recently injured players. Retrieved from https://www.tennisexplorer.com/list-players/injured/.