Stakeholder report

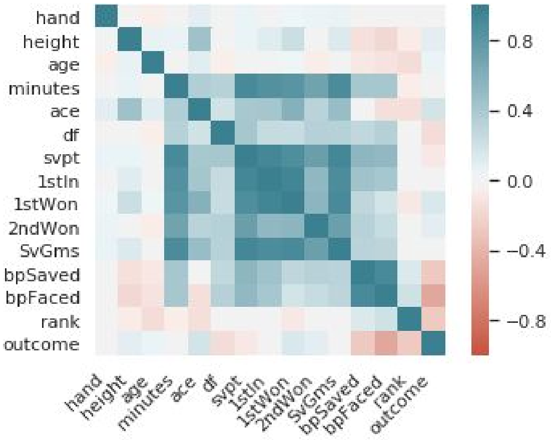
In this report we will summarize the analysis conducted on the tennis database from the year 2018. As mentioned in the Google colaboratory we decided to crop the database both by the columns and by the rows. From the columns we dropped the data that was not relevant to what we wanted to focus on. For example the data such as tournament ID, draw-size or how the player qualified for the tournament. From the rows, that represented each match of each tournament from the whole year, we decided to drop a lot of tournaments. The reason for this is, that in one data frame we had data from Grand Slams, which are the best tournaments in tennis world and the quality of the players and performances is the highest here. On the other hand we had some Davis Cup matches, which are matches between the countries or Challengers, which are one of the lowest tournaments on the ATP level. The players are not even the same on Grand Slams levels, as on the Challengers, therefore for this project it did not seem right to compare the performances of different players from different levels of quality.

After the pre-processing steps we are now ready to conduct the Exploratory Data Analysis.

From this we learned that the average height of the winner was 187.85 cm and the average height of loser was 186.1 cm but the standard deviations for these means were 8.5 and 7.5 cm, which means that any number within this range +/- 8 or 7 centimeters is normal and by this statistic we could say, that the height of the player did not matter in regards to the outcome of the match. This fact also confirmed a correlation we ran between the outcomes of height, where we got the number of only 8% correlation between the two variables, which is very low.

A very important statistic was the comparison of means of serve losing between winners and losers. As one of the main objectives in tennis is to hold your serve and try to take your opponent’s serve. If you lose your serve, it is called that the opponent “broke” you. In this sense, the winners were on average “broken” only 2 times per match, whereas the losers were broken 5 times on average. The standard deviation for the winners was 2, this means that it is very common, that the winners of the matches don’t lose their serve even once in the match. This is one of the key elements that determine the winners from losers in tennis world, which is the ability to hold your serve.

Later we ran the correlation matrix in order to see to what degree the variables influence each other.



The first high correlation is between height and aces - this is obvious, as it is easier for a taller player to hit an ace and this is what usually their game is based on - this is what they train the most.

Second high correlation is minutes and the stats of the points, serving and serves games. It is obvious that the longer the match takes, the more points they play and therefore both of the stats increase simultaneously.

One of the biggest negative correlation is between the outcome and break points Faced - as we know break point is when a player is facing loss of his serve and loss of his serve directly impacts the outcome of the match. The more BP you face the more likely you are to lose your serve and lose the match. The players who struggle with "big moments" - moments when they can not afford to lose the break point needs to train their mental part of the game, because as we see from the data, if they are not mentally endurant in big moments, they are very likely to lose the match.

We see almost zero correlation between the age and the length of the match, which might also be a bit surprising stats, as it is generally believed that older players don’t have as much energy as young ones and want to finish the match faster (usually play rallies under 5 points, go to net more often to finish the point there, or play serve and volley style of game) in order to save some energy. But according this statistics we see that this is absolutely not the case.

Then we decided to focus on three best players and create statistics based only for them on how they performed on the biggest stages of tennis during the last year. We take Djokovic, Nadal and Federer.

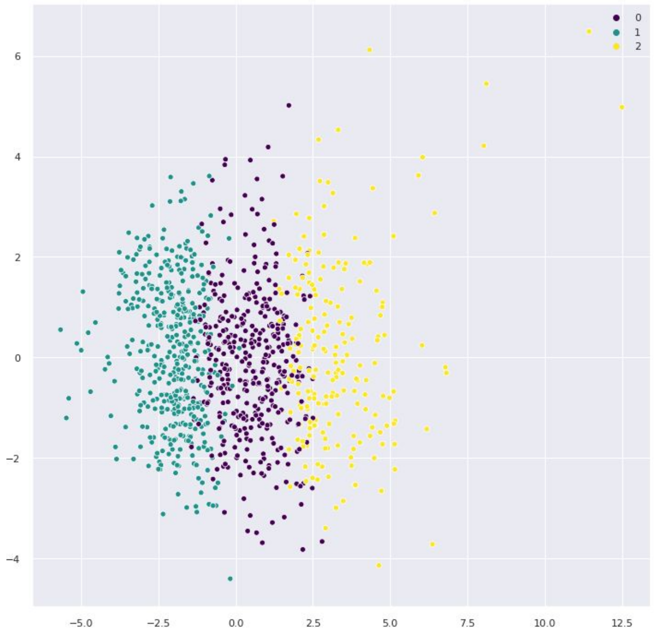
With all of the above-mentioned legends, we experienced the correlation in same type of data, data about the serving, as it is not a surprise, since serving is one of the most crucial parts of the game. Each time there was a very high correlation between the

- Number of 1st serves in and the points won on their serve
- Number of 1st serves in and number of points won after their 1st serves
- Number of points won after the 1st serve and number of serve games won in the whole match

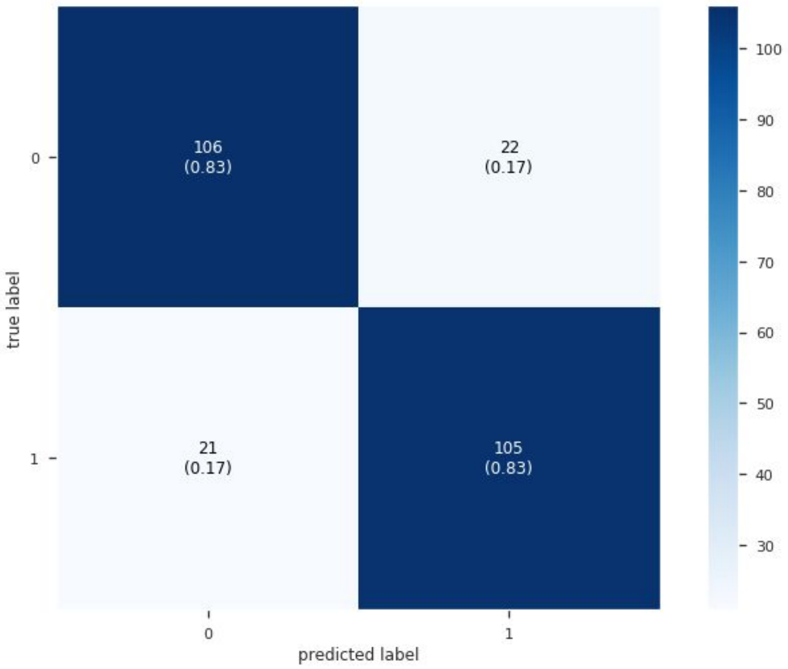
That’s why we decided to focus on the next couple of information about the players on their performance on the serve.

In tennis match it is very crucial, whether the player can manage to hit the first serve, the first serve is the fastest one he can hit and the player should have the advantage in the rally after a good first serve, therefore we wanted to see how many percent of the points the players won after they hit the first serve and by the result also confirm this prediction. Djokovic and Nadal won 75% of the rallies after the first serve in. This means that out of every 4 times, when they were serving and hit the first serve, 3 times they managed to win the point. For Federer the number was even higher and he managed to win 84% of the rallies after first serve.

After better understanding of the dataset we move to Machine Learning. In unsupervised ML we learn that the ideal number of clusters for our current data is three. Here we can visualise the clusters. We also found out, that the variable “first serve in” is one of the variables on the bases of which the clusters are divided, since each one of the data don't appear in more than two clusters, such as the players, who can serve a lot of first serves in, were only in 3rd cluster(we expect the “winning” cluster), and the players who struggle with first serve were mainly in 1st of 2nd cluster (we expect the “losing” or “average” cluster).



In order to predict the outcome of the match based on the independent variables, we chose the Logistic Regression model as it was the best model and explains with an accuracy of 83% if the outcome is a Loss or a Win. The model predicts with 83% precision that the outcome of the match is a Loss when it is indeed a Loss. This is the same for predicting that the outcome is a Win. In the case of recall we get 83% for both Win and Loss. This means for example that from all the instances in which the outcome was a loss, the model was correct 83% of the time and only 17% of the time the model is not correct.



Here we also attach the legend, explaining the tennis terms of column names

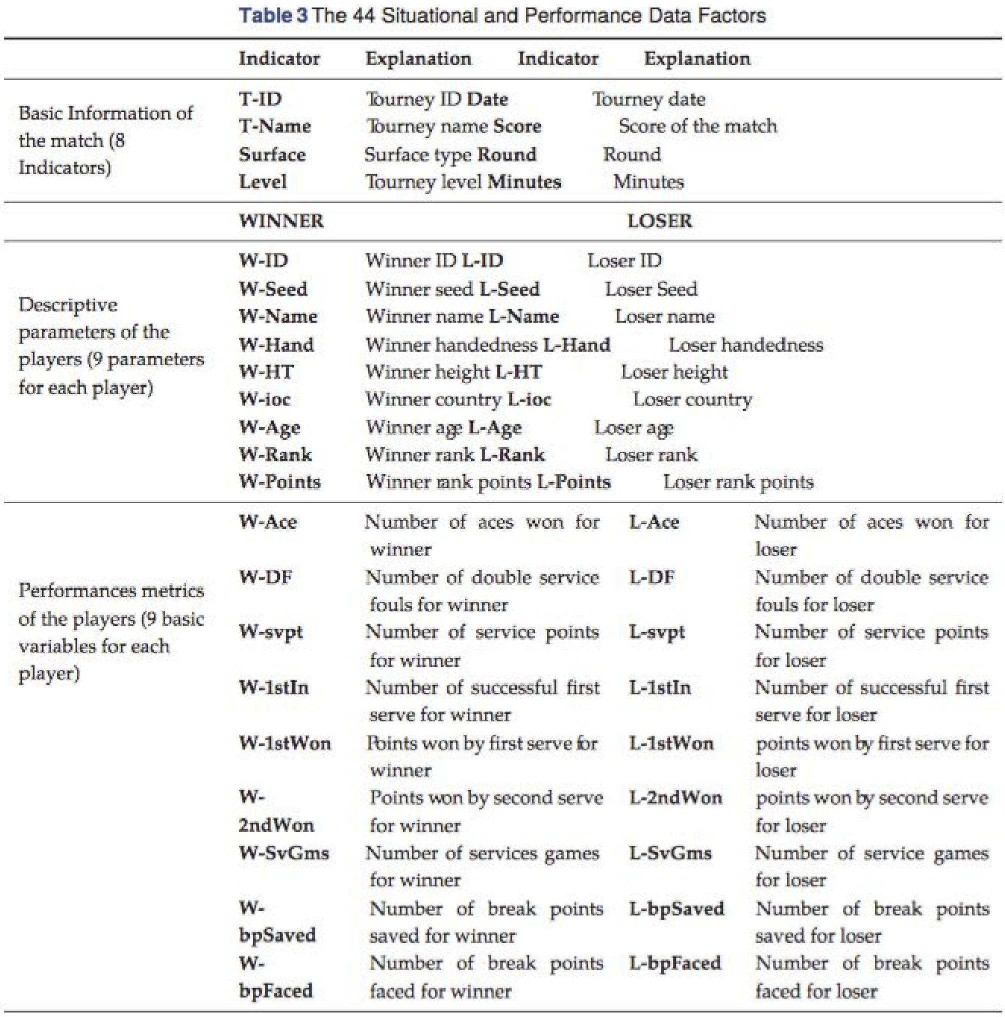


Table 3 The 44 Situational and Performance Data Factors

	Indicator	Explanation	Indicator	Explanation
Basic Information of the match (8 Indicators)	**T-ID**	Tourney ID	**Date**	Tourney date
	T-Name	Tourney name	**Score**	Score of the match
	Surface	Surface type	**Round**	Round
	Level	Tourney level	**Minutes**	Minutes
	WINNER		**LOSER**	
Descriptive parameters of the players (9 parameters for each player)	**W-ID**	Winner ID	**L-ID**	Loser ID
	W-Seed	Winner seed	**L-Seed**	Loser Seed
	W-Name	Winner name	**L-Name**	Loser name
	W-Hand	Winner handedness	**L-Hand**	Loser handedness
	W-HT	Winner height	**L-HT**	Loser height
	W-ioc	Winner country	**L-ioc**	Loser country
	W-Age	Winner age	**L-Age**	Loser age
	W-Rank	Winner rank	**L-Rank**	Loser rank
	W-Points	Winner rank points	**L-Points**	Loser rank points
Performances metrics of the players (9 basic variables for each player)	**W-Ace**	Number of aces won for winner	**L-Ace**	Number of aces won for loser
	W-DF	Number of double service fouls for winner	**L-DF**	Number of double service fouls for loser
	W-svpt	Number of service points for winner	**L-svpt**	Number of service points for loser
	W-1stIn	Number of successful first serve for winner	**L-1stIn**	Number of successful first serve for loser
	W-1stWon	Points won by first serve for winner	**L-1stWon**	points won by first serve for loser
	W-2ndWon	Points won by second serve for winner	**L-2ndWon**	points won by second serve for loser
	W-SvGms	Number of services games for winner	**L-SvGms**	Number of service games for loser
	W-bpSaved	Number of break points saved for winner	**L-bpSaved**	Number of break points saved for loser
	W-bpFaced	Number of break points faced for winner	**L-bpFaced**	Number of break points faced for loser