FM_21_FCD

December 5, 2021

1 Projeto - Fundamentos de Ciência dos Dados: Football Manager 2021

Elaborado por: **Tiago Alves - 96144**

1.1 Introdução

No decorrer deste projeto irei aplicar os conteúdos lecionados durante ao longo da unidade curricular Fundamentos de Ciência dos Dados numa base de dados retirada do jogo Football Manager 2021, de modo a conseguir construir, através de machine learning, a melhor equipa de futebol possivel, para os proximos 5 anos.

1.2 Definição do problema

Football manager é um jogo de simulação desenvolvido pela Sports Interactive, cujo objetivo passa por criar um treinador e comandar uma clube de futebol enquanto treinador principal.

Um treinador principal é responsável por varios aspetos do clube, nomeadamente, criar esquemas taticos, comandar a equipa durante os jogos, dar entrevistas, realizar transferências, entre outros.

De todas as responsabilidades do treinador a mais complicada é sem duvida a construção da equipa em si. Não só para o presente mas também para o futuro. Uma equipa de futebol precisa de ser equilibrada em varios aspetos. Precisa de lideres, jovens e de jogadores capazes de fazer a diferença. Deste modo, existe um mercado de transferênçias, baseado no mercado real, em que o treinador é capaz de contratar os jogadores que mais se adequam ao sistema tactico e à equipa no geral.

Neste contexto, com este projeto, a ideia passa pela resolução deste problema, através da aquisição da base de dados do jogo, em que estão presentes (quase)todos os jogadores do mundo(profissionais e semi-profissionais) e da consequente previsão do valor de mercado de um jogador baseado na habilidade atual do jogador. Ou seja, este sistema de recomendação seria capaz de recomendar os melhores jogadores para cada posição, dado um determinado limite de dinheiro.

Neste projeto será utilizada uma base de dados retirada de um jogo de Football Manager 2021, que se encontra em Maio de 2021, ou seja, já foram gerados alguns jogadores, que não existem na vida real.

1.3 Aquisição dos dados

Os dados utilizados neste projeto foram retirados do jogo Football Manager 2021 através do software Genie Scout 21g em https://www.fmscout.com/c-fm_genie_scout.html e a base de dados é

referente a um jogo com um ano de simulação, ou seja, o jogo começa em Julho de 2020 e os dados retirados são referentes a Maio de 2021.

1.4 Data Wrangling

Remove-se as colunas que não serão utilizadas como, jadeness, condition,hapiness level

```
                    Name             Nation    Position  \
0          Mbappé, Kylian             France   AM RL, ST
1             Kane, Harry  England / Ireland      AM/F C
2                  Neymar             Brazil  AM LC, F C
3        De Bruyne, Kevin  Belgium / England  DM, AM RLC
4             Mané, Sadio  Senegal / England   AM RL, ST
…                     …                …         …
405012     Zytko, Mateusz             Poland         D C
405013  Zyumbulev, Evgeni           Bulgaria         D C
405014     Zyuzin, Khalid             Russia          GK
405015      Zyuzin, Maxim             Russia          DM
405016   Zyznowski, Jakub             Poland        AM R

                          Club   Age  Int Caps  Int Goals     Wage  \
0          Paris Saint-Germain  22.0      48.0       17.0  476,150
1                    Tottenham  28.0      61.0       33.0  200,000
2          Paris Saint-Germain  29.0     115.0       65.0  858,680
3                     Man City  30.0      90.0       22.0  236,900
4                    Liverpool  29.0      76.0       24.0  162,000
…                        …   …       …         …       …
405012                       -  38.0       0.0        0.0      0.0
405013               Kyustendil  32.0       0.0        0.0      0.0
405014  Metallurg Magnitogorsk  16.0       0.0        0.0      0.0
405015                       -  34.0       0.0        0.0      0.0
405016               Stal Brzeg  20.0       0.0        0.0     80.0

              Value   Sale Value Best Rating Best Pot Rating   PoD
0       87,084,440  300,000,000  94.4% (FS)      95.3% (FS)  100%
1       84,171,600  282,088,380  95.3% (FS)      95.3% (FS)  100%
2       82,065,160  270,284,830  94.0% (FS)      96.1% (FS)   15%
3       78,798,970  267,164,000   90.6% (W)       90.6% (W)  100%
4       80,825,750  266,202,800   91.4% (W)       92.1% (W)  100%
…              …           …         …              …   …
405012         0.0         0.0  60.8% (CB)      60.8% (CB)    0%
405013         0.0         0.0  64.1% (CB)      64.8% (CB)  100%
405014         0.0         0.0  39.9% (GK)      53.0% (GK)   24%
405015         0.0         0.0   58.9% (M)       58.9% (M)    0%
405016         0.0         0.0   51.2% (W)       57.8% (W)   54%

[405017 rows x 13 columns]
```

1.5 EDA

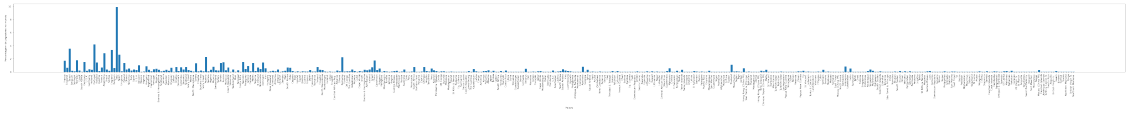
EDA vai ser maioritariamente através de vizualiações de gráficos e tabelas:

- Graficos / Tabela por pais inclui:
  - Gráficos de Nation x Percentagem de jogadores no mundo, para verificar qual os paises
  - Gráficos de Nation x Best rating, para verificar quais os paises que atualmente possu
  - Gráficos de Nation x Best potencial rating, para verificar que paises são melhores em
- Escolha dos melhores 11 por posição
- Graficos / Tabela por posição, para verificar quais a posições mais populares a nivel mundial e fazer comparações com premissas comuns

1.5.1 Nações

```
Número de Paises:  445
Número maximo de jogadores pretencentes a um pais:  40172
Número minimo de jogadores pretencentes a um pais:  1
```

temos 445 paises, alguns deles com 0 jogadores e por isso vamos retirar todos os paises com menos de 11 jogadores já que 11 é o minimo para se jogar um jogo



	0	1
0	Italy	9.919
1	Portugal	4.184
2	Brazil	3.542
3	England	3.370
4	Argentina	2.865
..	…	…
403	Yemen	0.003
404	Bahrain	0.003
405	Cayman Islands	0.003
406	US Virgin Is.	0.003
407	British Virgin Is.	0.003

[408 rows x 2 columns]

Conseguimos verificar que grande parte dos jogadores de futebol são maioritariamente pretencentes a paises europeus, sul americanos e africanos. Sendo que existem varios paises cuja representação a nivel mundial é extremamente pequena, no entanto, vamos considerar todos os paises que conseguem ter pelo menos 11 jogadores, sendo que todos aqueles que não atendem a essa condição foram

retirados num passo anterior. Verificamos ainda que o top 5 de paises com mais representação a nivel mundial são, por ordem: Itália, Portugal, Brazil, Inglaterra e Argentina

1.5.2 Posições

A partir da coluna de posições

	0	1
0	ST	47488
1	GK	44753
2	D C	40668
3	DM	26584
4	M C	25166
..	…	…
496	D RC, DM, AM/F C	1
497	D C, AM RLC, F C	1
498	WB RL, ST	1
499	D L, DM, AM R	1
500	D R, DM, ST	1

[501 rows x 2 columns]

Na tabela de posições verifica-se que existem jogadores que quando são capazes de jogar mais que uma posição essa posição aparece na coluna da posição e por isso é considerada uma posição diferente. Por exemplo o jogador Bruno Fernandes e o jogador Kevin de Bruyne apesar de jogarem na mesma posição, a sua posição é categorizada de forma diferente. O Kevin De Bruyne como joga todas as posições no meio campo a posição dele é representada como Defensive Midfileder, Attacking Midfilder (Right, Left, Center). Deste modo é necessário fazer a categorização das posições deste tipo de jogadores

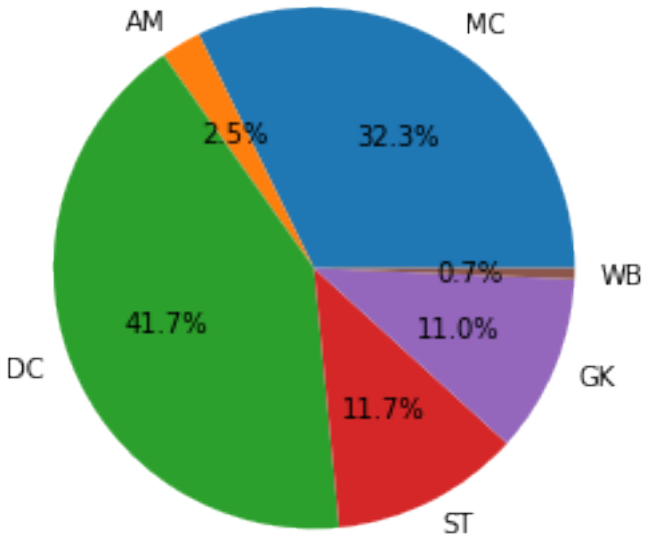
Para reduzir o número de posições assume-se que:

- Qualquer medio, seja defensivo ou centro,esquerdo ou direito é considerado médio, neste caso médio centro (MC)
- Qulquer médio que joga somente posições ofensivas(AM, AMR, AML, etc.) a sua posição será considerada como AM visto que a maior parte destes jogadores pode jogar em qualquer posição do ataque sendo que geralmente mudam a forma como atacam.(Um extremo do lado direito que utilize o pé esquerdo costuma fazer diagonais para centro, enquanto que se jogar no lado oposto costuma arragar mais a linha).
- Qualquer Defesa lateral, seja ele, direito ou esquerdo, ou ala esquerdo ou direito(DL,DR,WBR,WBL, respetivamente), é considerado como defesa lateral, WB.
- Para jogadores que jogam multiplas posições em varias partes do terreno, será considerada sempre a opção defensiva pela natureza unica das posições comparativamente ao resto do campo.(É provável que um médio centro possa jogar do lado direito ou esquerdo ou mesmo a falso 9, mas se este jogador for posto numa posição defensiva o seu rendimento baixará significativamente. Muitas vezes o que acontece é que quando falta jogadores para posições defensivas estes jogadores são muitas vezes quem preenche essas faltas. por exemplo o Joshua Kimmich)

- Jogadores que jogam multiplas posições ofensivas e no meio campo, será considerada a posição do meio campo pela polivalência da mesma

Os seguintes filtros serão aplicados, por ordem:

- Se o jogador joga a defesa centro a sua posição será DC
- Se o jogador joga a lateral, a sua posição será WB
- Se o jogador joga a médio centro a sua posição será MC
- Se o jogador joga a medio atacante, a sua posição será AM



Se as posições forem agrupadas em médios, atacantes e defesas, verifica-se que a area do terreno mais popular é a defensiva seguida do meio campo e da ofensiva sendo que em ultimo o espaço do terreno menos popular é o de guarda-redes.

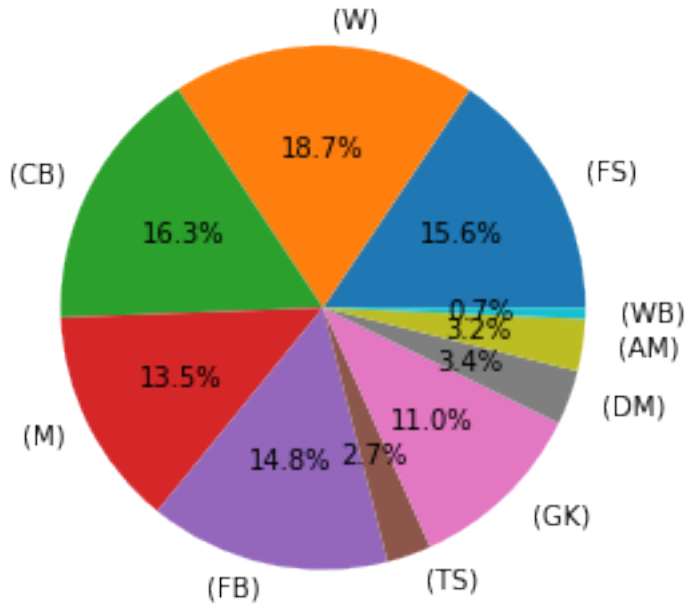
Após observação é possivel verificar que na coluna de Best Rating está documentada a melhor posição para cada jogador, deste modo um grafico de posições mais preciso seria o seguinte:

	Name	Nation	Position	\
0	Mbappé, Kylian	France	AM RL, ST	
1	Kane, Harry	England / Ireland	AM/F C	
2	Neymar	Brazil	AM LC, F C	
3	De Bruyne, Kevin	Belgium / England	DM, AM RLC	
4	Mané, Sadio	Senegal / England	AM RL, ST	
…	…	…	…	
405012	Zytko, Mateusz	Poland	D C	
405013	Zyumbulev, Evgeni	Bulgaria	D C	
405014	Zyuzin, Khalid	Russia	GK	
405015	Zyuzin, Maxim	Russia	DM	
405016	Zyznowski, Jakub	Poland	AM R	

	Club	Age	Int Caps	Int Goals	Wage	\
0	Paris Saint-Germain	22.0	48.0	17.0	476,150	
1	Tottenham	28.0	61.0	33.0	200,000	
2	Paris Saint-Germain	29.0	115.0	65.0	858,680	
3	Man City	30.0	90.0	22.0	236,900	
4	Liverpool	29.0	76.0	24.0	162,000	
…	…	…	…	…		
405012	-	38.0	0.0	0.0	0.0	
405013	Kyustendil	32.0	0.0	0.0	0.0	
405014	Metallurg Magnitogorsk	16.0	0.0	0.0	0.0	
405015	-	34.0	0.0	0.0	0.0	
405016	Stal Brzeg	20.0	0.0	0.0	80.0	

	Value	Sale Value	Best Rating	Best Pot Rating	PoD
0	87,084,440	300,000,000	94.4% (FS)	95.3% (FS)	100%
1	84,171,600	282,088,380	95.3% (FS)	95.3% (FS)	100%
2	82,065,160	270,284,830	94.0% (FS)	96.1% (FS)	15%
3	78,798,970	267,164,000	90.6% (W)	90.6% (W)	100%
4	80,825,750	266,202,800	91.4% (W)	92.1% (W)	100%
…	…	…	…	…	
405012	0.0	0.0	60.8% (CB)	60.8% (CB)	0%
405013	0.0	0.0	64.1% (CB)	64.8% (CB)	100%
405014	0.0	0.0	39.9% (GK)	53.0% (GK)	24%
405015	0.0	0.0	58.9% (M)	58.9% (M)	0%
405016	0.0	0.0	51.2% (W)	57.8% (W)	54%

[405017 rows x 13 columns]



Se for tida em consideração somente a melhor posição de cada jogador, verifica-se que a posição mais popular passa a ser a posição lateral em terreno ofensivo, W ou Winger, seguida de defesa centro, CB, Ponta de lança, FS, Medio, M e lateral, FB. Categorizando posições pela area do terreno, verifica-se:

- Avançados com 37% de popularidade (FS + TS + W)
- Médios com 20.1% de popularidade (AM + DM + M)
- Defesas com 31.8% de popularidade (CB + FB + WB)
- Guarda-redes com 11% de popularidade

Deste modo e tendo em conta as duas formas de obter este gráfico de posições, ir-se-á considerar o ultimo como a forma correta de proceder à obtenção de posições para casos futuros

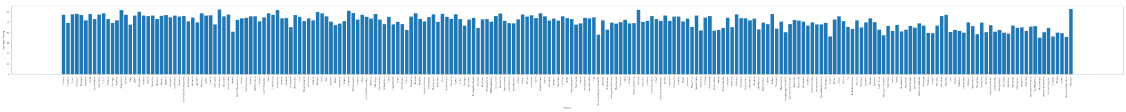
1.5.3 Best Rating x nation

	Name	Nation	Position	Club	Age	\
0	Mbappé, Kylian	France	AM RL, ST	Paris Saint-Germain	22.0	
1	Kane, Harry	England	AM/F C	Tottenham	28.0	
2	Neymar	Brazil	AM LC, F C	Paris Saint-Germain	29.0	
3	De Bruyne, Kevin	Belgium	DM, AM RLC	Man City	30.0	
4	Mané, Sadio	Senegal	AM RL, ST	Liverpool	29.0	
…	…	…	…	…	…	
405012	Zytko, Mateusz	Poland	D C	-	38.0	
405013	Zyumbulev, Evgeni	Bulgaria	D C	Kyustendil	32.0	
405014	Zyuzin, Khalid	Russia	GK	Metallurg Magnitogorsk	16.0	
405015	Zyuzin, Maxim	Russia	DM	-	34.0	
405016	Zyznowski, Jakub	Poland	AM R	Stal Brzeg	20.0	

	Int Caps	Int Goals	Wage	Value	Sale Value	Best Rating	\
0	48.0	17.0	476,150	87,084,440	300,000,000	94.4	
1	61.0	33.0	200,000	84,171,600	282,088,380	95.3	
2	115.0	65.0	858,680	82,065,160	270,284,830	94.0	
3	90.0	22.0	236,900	78,798,970	267,164,000	90.6	
4	76.0	24.0	162,000	80,825,750	266,202,800	91.4	
…	…	…	…	…	…	…	
405012	0.0	0.0	0.0	0.0	0.0	60.8	
405013	0.0	0.0	0.0	0.0	0.0	64.1	
405014	0.0	0.0	0.0	0.0	0.0	39.9	
405015	0.0	0.0	0.0	0.0	0.0	58.9	
405016	0.0	0.0	80.0	0.0	0.0	51.2	

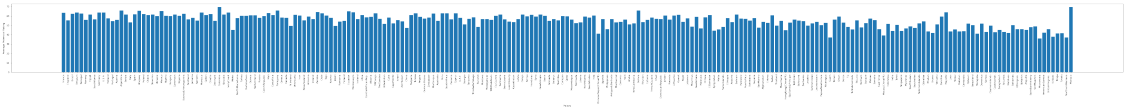
	Best Pot Rating	PoD
0	95.3% (FS)	100%
1	95.3% (FS)	100%
2	96.1% (FS)	15%
3	90.6% (W)	100%

```
4             92.1% (W)  100%
…                   …    …
405012       60.8% (CB)    0%
405013       64.8% (CB)  100%
405014       53.0% (GK)   24%
405015        58.9% (M)    0%
405016        57.8% (W)   54%

[405017 rows x 13 columns]

'Monaco'

'Micronesia'
```



Deste gráfico não é possivel retirar grande conclusão à parte do facto de que o average best rating varia entre 30 e 60 sendo que na maior parte dos paises onde o futebol é um desporto relativamente popular, está entre 40 e 55. Será possivel ver melhor quais são as nações com melhores jogadores se selecionarmos uma amostra mais pequena somente com os melhores 10%, por exemplo.

1.5.4 Best Potencial Rating x nation



Os melhores cinco paises em termos de potencial

```
['Colombia', 'Monaco', 'Argentina', 'Paraguay', 'Bolivia']
```

Novamente não é possivel retirar grandes informações deste gráfico à parte do facto de que os ratings potenciais são claramente superiores aos atuais, como era de esperar. Existe um claro dominio da região sul-americana em termos de jogadores com mais potencial

Best 11 players per position Agora, sabendo a melhor forma de obter as posições dos jogadores, vai-se limpar a dataframe de modo a que todos os valores possam ser utilizados (remover simbolos, etc)

```
                  Name    Nation Position                   Club   Age  \
0       Mbappé,Kylian    France      (FS)    Paris Saint-Germain  22.0
1          Kane,Harry  England      (FS)              Tottenham  28.0
```

```
2                 Neymar    Brazil     (FS)     Paris Saint-Germain  29.0
3         DeBruyne,Kevin  Belgium      (W)                 Man City  30.0
4             Mané,Sadio  Senegal      (W)                Liverpool  29.0
…                    …       …      …                        …   …
405012     Zytko,Mateusz    Poland     (CB)                        -  38.0
405013  Zyumbulev,Evgeni  Bulgaria     (CB)              Kyustendil  32.0
405014     Zyuzin,Khalid    Russia     (GK)  Metallurg Magnitogorsk  16.0
405015      Zyuzin,Maxim    Russia      (M)                        -  34.0
405016   Zyznowski,Jakub    Poland      (W)              Stal Brzeg  20.0

            Wage       Value   Sale_Value Best_Rating Best_Pot_Rating  PoD
0       476,150  87,084,440  300,000,000        94.4            95.3  100
1       200,000  84,171,600  282,088,380        95.3            95.3  100
2       858,680  82,065,160  270,284,830        94.0            96.1   15
3       236,900  78,798,970  267,164,000        90.6            90.6  100
4       162,000  80,825,750  266,202,800        91.4            92.1  100
…           …           …           …         …              …   …
405012      0.0         0.0         0.0        60.8            60.8    0
405013      0.0         0.0         0.0        64.1            64.8  100
405014      0.0         0.0         0.0        39.9            53.0   24
405015      0.0         0.0         0.0        58.9            58.9    0
405016     80.0         0.0         0.0        51.2            57.8   54

[405017 rows x 11 columns]

                 (FS)                    (W)             (CB)  \
0         Messi,Lionel         Silva,Bernardo   vanDijk,Virgil
1           Kane,Harry            Hazard,Eden  deLigt,Matthijs
2           Félix,João             Mané,Sadio  Laporte,Aymeric
3        Mbappé,Kylian        Sterling,Raheem  Acerbi,Francesco
4        Agüero,Sergio         DeBruyne,Kevin     Ramos,Sergio
5               Neymar             Foden,Phil    Demiral,Merih
6   Lewandowski,Robert         RobertoFirmino     Hummels,Mats
7     CristianoRonaldo  Alexander-Arnold,Trent    deVrij,Stefan
8       Haaland,Erling       Kulusevski,Dejan     Piqué,Gerard
9        Benzema,Karim           Sancho,Jadon       LuizFelipe
10   Griezmann,Antoine        Ødegaard,Martin       Dias,Rúben

                 (M)             (FB)                    (TS)  \
0      Barella,Nicolò   Kimmich,Joshua      Ibrahimovic,Zlatan
1      deJong,Frenkie  Robertson,Andrew           Lukaku,Romelu
2         Modric,Luka       Marquinhos             D eko,Edin
3          Kroos,Toni      Alaba,David  Milinkovic-Savic,Sergej
4   Bentancur,Rodrigo     Chilwell,Ben          Giroud,Olivier
5     Tielemans,Youri          Fabinho            Zapata,Duván
6        Kanté,N'Golo    Mendy,Ferland       Llorente,Fernando
7          Keïta,Naby      MarioGaspar              MaxiGómez
```

8	Valverde,Federico	Gayà,José	Mand ukic,Mario
9	Brozovic,Marcelo	Müldür,Mert	Belotti,Andrea
10	Parejo,Dani	Reguilón,Sergio	DiegoCosta

	(GK)	(DM)	(AM)	\
0	Courtois,Thibaut	Busquets,Sergio	Silva,David	
1	Alisson	Casemiro	LuisAlberto	
2	Oblak,Jan	Witsel,Axel	Rodríguez,James	
3	terStegen,Marc-André	Kessié,Franck	Özil,Mesut	
4	Szczesny,Wojciech	Henderson,Jordan	Canales,Sergio	
5	Neuer,Manuel	Carvalho,William	Eriksen,Christian	
6	Handanovic,Samir	Rice,Declan	Mata,Juan	
7	Ederson	deRoon,Marten	Pedri	
8	Arrizabalaga,Kepa	Ademi,Arijan	PabloHernandez	
9	DeGea,David	Iborra,Vicente	Payet,Dimitri	
10	Donnarumma,Gianluigi	Parolo,Marco	Isco	

	(WB)
0	Ayman,Ahmed
1	Hamza,AbdulrazackMohamed
2	Hernández,Bryan
3	Colella,Manuel
4	Arikan,Onur
5	Begishev,Ramis
6	Jiménez,Jonathan
7	Lizama,Blas
8	AhmedAyad
9	Alfaro,Fabricio
10	Kashken,Dinmukhamed

Verifica-se a partir desta tabela que a maioria dos jogadores, tal como esperado, pertencem a clubes que atuam nas principais ligas europeias e que a posição de lateral é a posição em que a maior parte dos jogadores foram gerados pelo jogo(verifica-se a falta de jogadores reais nesta coluna) enquanto que todas as outras são dominadas por jogadores existentes.

Agora, compara-se a tabela anterior com a mesma tabela, mas por Potencial Rating

	(FS)	(W)	(CB)	\
0	Messi,Lionel	Silva,Bernardo	vanDijk,Virgil	
1	Wolfe,Barry	Hazard,Eden	deLigt,Matthijs	
2	Neymar	González,Sebastián	Barrero,Jorge	
3	Griezmann,Antoine	Kulusevski,Dejan	Rae,Scott	
4	Kane,Harry	Mané,Sadio	Floiras,Didier	
5	Mbappé,Kylian	RobertoFirmino	Demiral,Merih	
6	Félix,João	Foden,Phil	LuizFelipe	
7	Agüero,Sergio	Ødegaard,Martin	Leclercq,Jean-Jacques	
8	Lewandowski,Robert	Williams,Nico	Laporte,Aymeric	
9	Kolesnyk,Vadym	Trincão,Francisco	Benkovic,Filip	

10	Dybala,Paulo	García,Hugo	Dias,Rúben	

	(M)	(FB)	(TS)	\
0	Pardo,Stefano	Derrien,Stéphan	Ibrahimovic,Zlatan	
1	vanKouwen,René	Simmes,Louis	Lukaku,Romelu	
2	Camavinga,Eduardo	Paredes,AlbertoLeon	D eko,Edin	
3	deJong,Frenkie	Lipcsei,Roland	Milinkovic-Savic,Sergej	
4	Barella,Nicolò	Garbutt,Carl	MaxiGómez	
5	Göksel,Cemal	Ruiz,Javier	Cornelius,Andreas	
6	Blé,Ladji	Gvardiol,Jo ko	Belotti,Andrea	
7	Kroos,Toni	Paiva,João	Zapata,Duván	
8	Cook,Lewis	Dettmann,Jan	Muriqi,Vedat	
9	Villar,Gonzalo	Robertson,Andrew	Giroud,Olivier	
10	Tielemans,Youri	Kimmich,Joshua	Llorente,Fernando	

	(GK)	(DM)	(AM)	\
0	Tsokos,Evangelos	Busquets,Sergio	Pedri	
1	Courtois,Thibaut	Casemiro	Silva,David	
2	Alisson	Malpeli,Giorgio	Rodríguez,James	
3	Donnarumma,Gianluigi	Rice,Declan	Eriksen,Christian	
4	Oblak,Jan	Kessié,Franck	Reinier	
5	Woller,Yves	Luís,Florentino	Pipi	
6	Vandevoordt,Maarten	Majchrzak,Grzegorz	LuisAlberto	
7	terStegen,Marc-André	Carvalho,William	Isco	
8	Domingo	Vasiu,Alex	Carvalho,João	
9	Simón,Unai	Witsel,Axel	Hannibal	
10	Lowe,Ed	Romeu,Oriol	Canales,Sergio	

	(WB)
0	Colella,Manuel
1	BenAhmed,Hamza
2	Ancheta,Maximiliano
3	Pasqualetti,Tom
4	Zanfir,Antonio
5	Weaver,Kenny
6	Hamza,AbdulrazackMohamed
7	irdum,Leon
8	Gundelach,Moritz
9	Kardaris,Thodoros
10	Ayman,Ahmed

Comparando esta tabela com a anterior consegue-se verificar que, tendo em conta o Best Potencial rating, a tabela inclui bastantes mais jovens, tal como esperado. Ha uma clara dominançia em algumas posições, em que o melhor jogador atual, é também o melhor futuro jogador, sendo incontestavelmente o melhor jogador da posição como é o caso do Lionel Messi ou do Bernardo Silva. Verifica-se ainda varias entradas de jogadores, *Computer generated*, como é o caso do Evangelos Tsokos na posição de guarda redes, como o futuro melhor da posição.

11 Most valuable players per position Será considerado o value em vez do sale value, já que, o sale value é influenciado pela condição económica do pais e do clube em que o jogador se insere. De modo a fazer a comparação entre a primeira tabela e esta, para verificar se existe correlação entre os valores de mercado e os ratings dos jogadores

	(FS)	(W)	(CB)	(M)	\
0	Mbappé,Kylian	Mané,Sadio	Laporte,Aymeric	Valverde,Federico	
1	Kane,Harry	DeBruyne,Kevin	Dias,Rúben	deJong,Frenkie	
2	Neymar	Sterling,Raheem	Varane,Raphaël	Ndombele,Tanguy	
3	Son,Heung-Min	Silva,Bernardo	vanDijk,Virgil	Rodri	
4	Salah,Mohamed	Havertz,Kai	deLigt,Matthijs	Mount,Mason	
5	Haaland,Erling	RobertoFirmino	Süle,Niklas	Goretzka,Leon	
6	Griezmann,Antoine	Foden,Phil	Torres,Pau	Bentancur,Rodrigo	
7	Werner,Timo	Sané,Leroy	Zouma,Kurt	McKennie,Weston	
8	Kean,Moise	Gnabry,Serge	Kimpembe,Presnel	Partey,Thomas	
9	GabrielJesus	Fernandes,Bruno	Demiral,Merih	Tolisso,Corentin	
10	Félix,João	Coutinho	Giménez,José	Bennacer,Ismaël	

	(FB)	(TS)	(GK)	\
0	Alaba,David	Lukaku,Romelu	Ederson	
1	Fabinho	Milinkovic-Savic,Sergej	Courtois,Thibaut	
2	Marquinhos	Jiménez,Raúl	Alisson	
3	Kimmich,Joshua	MaxiGómez	terStegen,Marc-André	
4	Roberto,Sergi	Sørloth,Alexander	Oblak,Jan	
5	Robertson,Andrew	Delort,Andy	DeGea,David	
6	Gayà,José	Calvert-Lewin,Dominic	Szczesny,Wojciech	
7	Aké,Nathan	Haller,Sébastien	Neto	
8	ÉderMilitão	Belotti,Andrea	Onana,André	
9	Mendy,Benjamin	Zapata,Duván	Pickford,Jordan	
10	Gosens,Robin	Nsamé,Jean-Pierre	Donnarumma,Gianluigi	

	(DM)	(AM)	(WB)
0	Casemiro	LuisAlberto	Ayman,Ahmed
1	Henderson,Jordan	Isco	Villa,Adrián
2	Allan	Pedri	Jiménez,Jonathan
3	Rodríguez,Guido	Canales,Sergio	Lizama,Blas
4	Kessié,Franck	Rodríguez,James	Alfaro,Fabricio
5	Højbjerg,Pierre-Emile	Quintero,JuanFernando	Komarets,Maxym
6	Carvalho,William	Eriksen,Christian	Mbong,Joseph
7	Llorente,Marcos	OmarAbdulrahman	Kybyrai,Eskendir
8	Soucek,Tomas	Malinovskyi,Ruslan	Sarsengaliev,Salamat
9	Rice,Declan	Swift,John	Kashken,Dinmukhamed
10	Romeu,Oriol	Vanaken,Hans	Volkovytskyi,Eugene

A partir desta tabela posso confirmar que nem sempre os jogadores mais valiosos são os melhores. Por exemplo a idade tem uma grande influência no valor de mercado, ou seja, um jogador mais velho, com o mesmo potencial que um jogador mais novo, tem sempre um valor de mercado inferior. Deste modo, nem sempre é benéfico para uma equipa contratar o melhor jogador, se ao fim de alguns anos,

o retorno financeiro desse jogador vai ser muito inferior ao atual, mesmo que o retorno desportivo seja alto.

Highest paid players per position Para verificar a relação entre salários e habilidade do jogador

	(FS)	(W)	(CB)	(M)	\
0	CristianoRonaldo	Oscar	Fellaini,Marouane	Paulinho	
1	Neymar	Bale,Gareth	Ramos,Sergio	Kroos,Toni	
2	Messi,Lionel	Hazard,Eden	Bonucci,Leonardo	Kanté,N'Golo	
3	Griezmann,Antoine	Talisca	deLigt,Matthijs	Pjanic,Miralem	
4	Mbappé,Kylian	Coutinho	Varane,Raphaël	Verratti,Marco	
5	Bakambu,Cédric	Sané,Leroy	Koulibaly,Kalidou	Modric,Luka	
6	Lewandowski,Robert	Hamsik,Marek	Maguire,Harry	Thiago	
7	Suárez,Luis	Pogba,Paul	vanDijk,Virgil	Rabiot,Adrien	
8	Arnautovic,Marko	Koke	Kimpembe,Presnel	Banega,Éver	
9	Werner,Timo	Marcelo	Martínez,Iñigo	deJong,Frenkie	
10	Benzema,Karim	Müller,Thomas	Alderweireld,Toby	Goretzka,Leon	

	(FB)	(TS)	(GK)	\
0	Marquinhos	D eko,Edin	DeGea,David	
1	Hernández,Lucas	Lukaku,Romelu	Neuer,Manuel	
2	Umtiti,Samuel	Cavani,Edinson	Oblak,Jan	
3	AlexSandro	Ibrahimovic,Zlatan	Courtois,Thibaut	
4	Chilwell,Ben	Mary,John	Szczesny,Wojciech	
5	Kimmich,Joshua	Al-Soma,Omar	Navas,Keylor	
6	Roberto,Sergi	Mariano	Donnarumma,Gianluigi	
7	Meunier,Thomas	Prijovic,Aleksandar	Ederson	
8	ÉderMilitão	Deeney,Troy	Arrizabalaga,Kepa	
9	Carvajal,Dani	Milinkovic-Savic,Sergej	terStegen,Marc-André	
10	Mendy,Ferland	Jiménez,Raúl	Areola,Alphonse	

	(DM)	(AM)	(WB)
0	Busquets,Sergio	Eriksen,Christian	Al-Abbas,Essa
1	Casemiro	Isco	Ayman,Ahmed
2	Henderson,Jordan	Pastore,Javier	Al-Sulaiman,Abdulkareem
3	Witsel,Axel	Fàbregas,Cesc	Al-Dakhil,Abdulrahman
4	Matic,Nemanja	Rodríguez,James	Hazazi,Adel
5	Gueye,Idrissa	Özil,Mesut	Al-Maimouni,Manawir
6	Strootman,Kevin	Silva,David	Al-Shammari,Bander
7	Bakayoko,Tiémoué	LuisAlberto	Al-Dossari,TurkiHamad
8	Nzonzi,Steven	Cazorla,Santi	Traoré,Anas
9	Højbjerg,Pierre-Emile	KooJa-Cheol	Al-Jabeeri,Hassan
10	DavidLópez	Aleñá,Carles	Al-Harbi,Sultan

Olhando para esta tabela, em comparação à anterior consegue-se verificar que salário não estará necessáriamente ligado a habilidade. Por exemplo, o Cristiano Ronaldo, é o jogador mais bem pago na posição FS, no entanto está longe de ser o melhor jogador a nivel de habilidade potencial e atual.

Verifica-se ainda que jogadores a atuar na liga chinesa também são pagos bem acima do valor que seriam pagos num liga europeia, por exemplo.

Base de dados final, com tipos atualizados. Deste modo a base de dados, após o **EDA** é a seguinte:

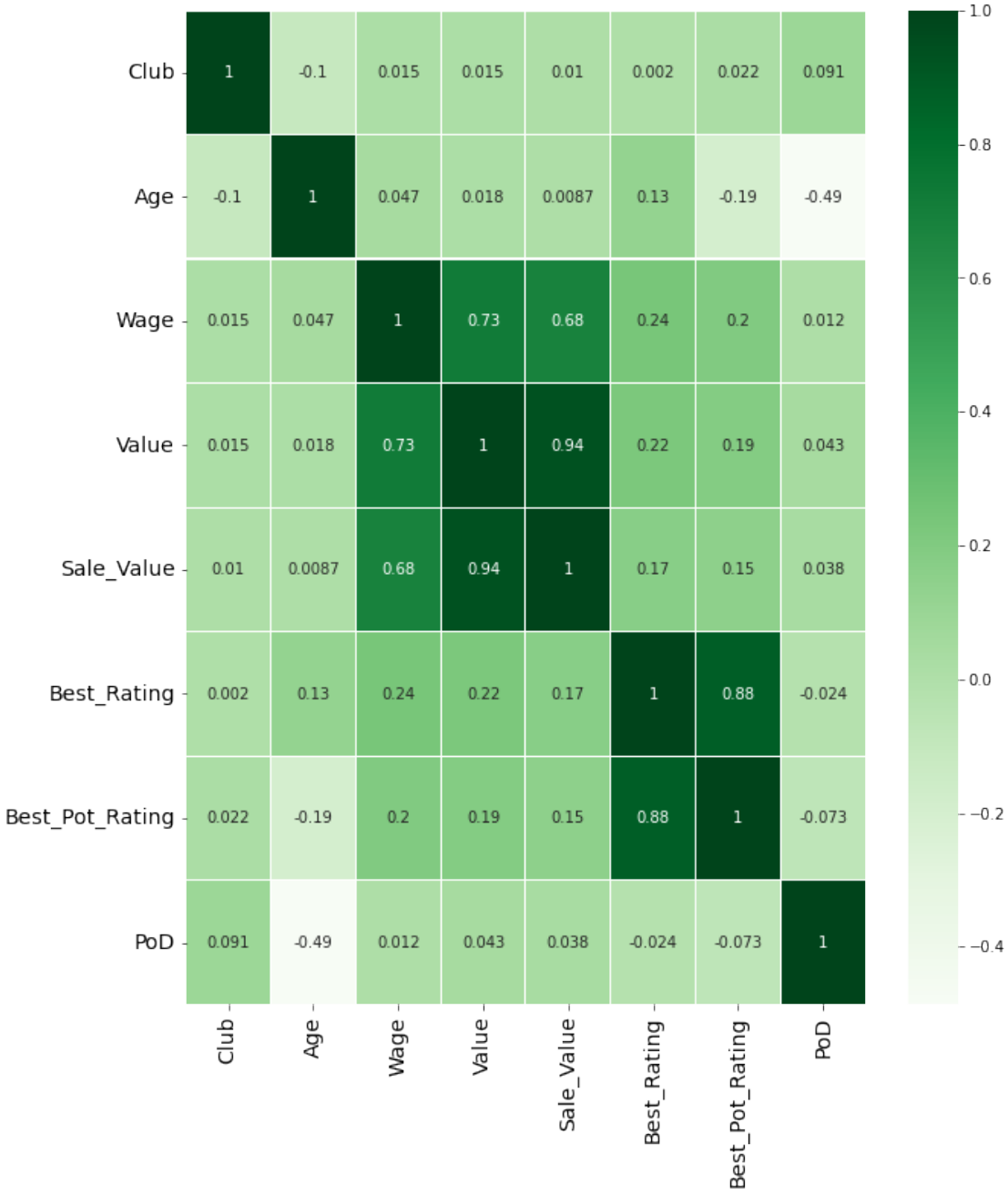
```
                    Name    Nation Position                    Club   Age  \
0          Mbappé,Kylian    France     (FS)     Paris Saint-Germain  22.0
1             Kane,Harry  England     (FS)               Tottenham  28.0
2                 Neymar    Brazil     (FS)     Paris Saint-Germain  29.0
3         DeBruyne,Kevin  Belgium      (W)                Man City  30.0
4             Mané,Sadio  Senegal      (W)               Liverpool  29.0
…                    …       …      …                      …   …
405012     Zytko,Mateusz    Poland     (CB)                       -  38.0
405013  Zyumbulev,Evgeni  Bulgaria     (CB)             Kyustendil  32.0
405014     Zyuzin,Khalid    Russia     (GK)  Metallurg Magnitogorsk  16.0
405015      Zyuzin,Maxim    Russia      (M)                       -  34.0
405016   Zyznowski,Jakub    Poland      (W)              Stal Brzeg  20.0

          Wage     Value  Sale_Value  Best_Rating  Best_Pot_Rating  PoD
0       476150  87084440   300000000         94.4             95.3  100
1       200000  84171600   282088380         95.3             95.3  100
2       858680  82065160   270284830         94.0             96.1   15
3       236900  78798970   267164000         90.6             90.6  100
4       162000  80825750   266202800         91.4             92.1  100
…          …       …          …           …               …  …
405012       0         0           0         60.8             60.8    0
405013       0         0           0         64.1             64.8  100
405014       0         0           0         39.9             53.0   24
405015       0         0           0         58.9             58.9    0
405016       0         0           0         51.2             57.8   54

[405017 rows x 11 columns]

Name               object
Nation             object
Position           object
Club               object
Age               float64
Wage                int64
Value               int64
Sale_Value          int64
Best_Rating       float64
Best_Pot_Rating   float64
PoD                 int64
dtype: object
```

Categorização dos dados Passagem de todos os tipos para int

```
Name               object
Nation             object
Position           object
Club                int16
Age                 int64
Wage                int64
Value               int64
Sale_Value          int64
Best_Rating         int64
Best_Pot_Rating     int64
PoD                 int64
dtype: object
```

1.6 Machine Learning

Heatmap para perceber a relação entre as features



Verifica-se um clara relação entre valor de mercado, idade, e a habilidade do jogador. O potencial de desenvolvimento do jogador não parece ter grande influência no valor do jogador. De notar que este valor é bastante influenciado pelo clube que o jogador representa bem como equipa tecnica e mentores. Sendo que pode variar durante a carreira do jogador. Todos os outros valores são basicamente constantes durante toda a carreira do jogador(idade, habilidade atual e valor, variam durante a carreira obviamente). Verifica-se ainda uma clara relação entre o valor de mercado do jogador e o salário que este aufere.

Vou tentar prever o valor de um jogador de modo a conseguir, através das habilidades atuais e potenciais dos jogadores e das outras features depictadas anteriormente.

Retirada amostra dos dados, 20000 dados retirados aleatoriamente

```
                    Name                Nation Position   Club  Age  Wage  \
249755       ChungManTsun  Hong Kong (China PR)     (CB)   6712   25     0
273272      Fortes,Danilo            Cape Verde     (AM)    948   22     0
285133          Hall,Adam                 Wales      (M)  16579   34     0
143429   Williamson,Ambry               Bahamas     (DM)  17202   16    30
64537     Rifan,M.Chairul              Indonesia     (CB)  15318   31   370
…                   …                    …      …     …  …   …
329578     Mihailovic,Nik              Slovenia     (TS)  11459   31     0
25812       Kayode,Joshua              Nigeria      (TS)  17541   20   700
111861  Matenciuc,Tiberiu               Romania     (GK)  23001   20   100
256026    Dandanell,Elias                Sweden     (CB)   9893   19     0
73227           Faye,Pape                 Spain      (FS)  15073   16   110

        Value  Sale_Value  Best_Rating  Best_Pot_Rating  PoD
249755      0           0          411              442   38
273272      0           0          445              463  100
285133      0           0          429              429    0
143429    430         730          454              471  100
64537   10890       11460          495              508    2
…       …          …           …                …  …
329578      0           0          604              627    1
25812   50090       70440          639              717   50
111861   2140        2250          498              525   82
256026      0           0          439              454  100
73227    4110        8450          599              753   39

[20000 rows x 11 columns]
```

Y ->Value X -> Age, Wage, Best_Rating, Best_Pot_Rating, PoD

Devidi o dataset em 70/30, 70 de teste e 30 de treino

SVM (Support Vectors Machines)

`Accuracy: 0.4678333333333333`

K-Nearest Neighbours

`Accuracy: 0.44683333333333336`

Decision Trees

`Accuracy: 0.46016666666666667`

Logistic Regression

`Accuracy: 0.4676666666666667`

Random Forest

```
Accuracy:  0.4811666666666667
```

Naive Bayes

```
Accuracy:  0.471
```

Neural network - MLP Classifier

```
Score:  0.462
```

Com uma accuracy de cerca de 46% verifico que nenhum dos modelos é capaz de prever, com alguma regularidade o valor do jogador. Assumo que, isto se deva, para além das escolhas feitas na parte deste projeto referente ao EDA(retirou-se colunas como o clube que representa, jadeness, hapiness, jogos pelas seleções, etc.), devido aos varios valores, não presentes na base de dados, como liga onde o jogador atua, lesões, o facto de o jogador estar listado como transferivel ou não, performances da epoca transata, etc. Estes valores impactam também o valor de mercado e por isso é dificil para qualquer modelo prever valores de mercado sem todos os valores necessários para tal. Dito isto e considerando a falta destes valores, concluo que estes valores (cerca de 46%), apesar de espaço para melhorar, são aceitaveis neste contexto.

Agora vou tentar melhorar estes valores a partir do grid search

1.6.1 Grid Search

SVM

```
Fitting 2 folds for each of 16 candidates, totalling 32 fits
[CV 1/2] END …C=0.1, gamma=1, kernel=rbf;, score=0.468 total time= 6.7min
[CV 2/2] END …C=0.1, gamma=1, kernel=rbf;, score=0.468 total time= 6.4min
[CV 1/2] END …C=0.1, gamma=0.1, kernel=rbf;, score=0.468 total time= 6.4min
[CV 2/2] END …C=0.1, gamma=0.1, kernel=rbf;, score=0.468 total time= 6.4min
[CV 1/2] END …C=0.1, gamma=0.01, kernel=rbf;, score=0.468 total time= 6.3min
[CV 2/2] END …C=0.1, gamma=0.01, kernel=rbf;, score=0.468 total time= 6.2min
[CV 1/2] END …C=0.1, gamma=0.001, kernel=rbf;, score=0.468 total time= 5.5min
[CV 2/2] END …C=0.1, gamma=0.001, kernel=rbf;, score=0.468 total time= 6.0min
[CV 1/2] END …C=1, gamma=1, kernel=rbf;, score=0.468 total time= 6.8min
[CV 2/2] END …C=1, gamma=1, kernel=rbf;, score=0.468 total time= 6.8min
[CV 1/2] END …C=1, gamma=0.1, kernel=rbf;, score=0.468 total time= 6.5min
[CV 2/2] END …C=1, gamma=0.1, kernel=rbf;, score=0.468 total time= 6.6min
[CV 1/2] END …C=1, gamma=0.01, kernel=rbf;, score=0.468 total time= 6.7min
[CV 2/2] END …C=1, gamma=0.01, kernel=rbf;, score=0.468 total time= 6.6min
[CV 1/2] END …C=1, gamma=0.001, kernel=rbf;, score=0.468 total time= 6.0min
[CV 2/2] END …C=1, gamma=0.001, kernel=rbf;, score=0.468 total time= 6.0min
[CV 1/2] END …C=10, gamma=1, kernel=rbf;, score=0.453 total time= 7.2min
[CV 2/2] END …C=10, gamma=1, kernel=rbf;, score=0.453 total time= 8.3min
[CV 1/2] END …C=10, gamma=0.1, kernel=rbf;, score=0.468 total time= 6.7min
[CV 2/2] END …C=10, gamma=0.1, kernel=rbf;, score=0.467 total time= 6.4min
[CV 1/2] END …C=10, gamma=0.01, kernel=rbf;, score=0.468 total time= 6.8min
[CV 2/2] END …C=10, gamma=0.01, kernel=rbf;, score=0.468 total time= 6.8min
[CV 1/2] END …C=10, gamma=0.001, kernel=rbf;, score=0.468 total time= 6.6min
[CV 2/2] END …C=10, gamma=0.001, kernel=rbf;, score=0.468 total time= 6.7min
```

```
[CV 1/2] END …C=100, gamma=1, kernel=rbf;, score=0.376 total time= 6.9min
[CV 2/2] END …C=100, gamma=1, kernel=rbf;, score=0.372 total time= 7.1min
[CV 1/2] END …C=100, gamma=0.1, kernel=rbf;, score=0.466 total time= 6.6min
[CV 2/2] END …C=100, gamma=0.1, kernel=rbf;, score=0.465 total time= 6.6min
[CV 1/2] END …C=100, gamma=0.01, kernel=rbf;, score=0.467 total time= 7.0min
[CV 2/2] END …C=100, gamma=0.01, kernel=rbf;, score=0.467 total time= 6.9min
[CV 1/2] END …C=100, gamma=0.001, kernel=rbf;, score=0.468 total time= 7.0min
[CV 2/2] END …C=100, gamma=0.001, kernel=rbf;, score=0.468 total time= 6.9min
0.4675714285714286
{'C': 0.1, 'gamma': 1, 'kernel': 'rbf'}
SVC(C=0.1, gamma=1)
```

KNN

```
Fitting 2 folds for each of 360 candidates, totalling 720 fits
0.4665714285714285
{'algorithm': 'auto', 'leaf_size': 1, 'n_neighbors': 30, 'weights': 'uniform'}
KNeighborsClassifier(leaf_size=1, n_neighbors=30)
```

Decision Trees

```
Fitting 2 folds for each of 750 candidates, totalling 1500 fits
0.4830714285714286
{'criterion': 'entropy', 'max_depth': 8, 'min_samples_split': 150}
DecisionTreeClassifier(criterion='entropy', max_depth=8, min_samples_split=150)
```

Logistic Regression

```
Fitting 2 folds for each of 20 candidates, totalling 40 fits
0.4675714285714286
{'C': 0.001, 'penalty': 'l2'}
LogisticRegression(C=0.001)
```

Random Forest

```
Fitting 2 folds for each of 108 candidates, totalling 216 fits
0.48314285714285715
{'criterion': 'gini', 'max_depth': 8, 'max_features': 'log2', 'n_estimators':
200}
RandomForestClassifier(max_depth=8, max_features='log2', n_estimators=200)
```

Naive Bayes

```
Fitting 2 folds for each of 100 candidates, totalling 200 fits
0.4675714285714286
{'var_smoothing': 1.0}
GaussianNB(var_smoothing=1.0)
```

MLP Classifier

```
Fitting 2 folds for each of 48 candidates, totalling 96 fits
0.46771428571428575
{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (50, 100, 50),
```

```
'learning_rate': 'constant', 'solver': 'sgd'}
MLPClassifier(hidden_layer_sizes=(50, 100, 50), random_state=1, solver='sgd')
```

Verifica-se que em alguns modelos se melhora o valor do passo anterior mas não significativamente.