Lovászi Zsuzsanna
VH57FE

Fizetések a Women’s Super League-ben

Statisztikai szoftverek esszé

2023. június 12.
Szeged

1. Helyzetfelmérés

Az elemzéshez a FIFA névre hallgató számítógépre és konzolra készült szimulációs játéksorozat adatbázisát használtam fel, azon belül is a női labdarúgók adatait tartalmazó táblát. Azért erre esett a választásom, mert ez a népszerű játék nagyon komoly kutatómunka után képez adatokat a labdarúgók képességeiről, melyek a valóságot kellően tükrözik.

	player_id	player_url	fifa_version	fifa_update	fifa_update_date	short_name	long_name	player_positions	overall	potential	...
0	227125	/player/227125/sam-kerr/230009	23	9	2023-01-13	S. Kerr	Samantha May Kerr	ST	91	91	...
1	227316	/player/227316/wendie-renard/230009	23	9	2023-01-13	W. Renard	Wéndèleine Thérèse Renard	CB	91	91	...
2	226301	/player/226301/alex-morgan/230009	23	9	2023-01-13	A. Morgan	Alexandra Morgan Carrasco	ST	90	90	...
3	227310	/player/227310/ada-hegerberg/230009	23	9	2023-01-13	A. Hegerberg	Ada Martine Stolsmo Hegerberg	ST	90	91	...
4	227246	/player/227246/lucy-bronze/230009	23	9	2023-01-13	L. Bronze	Lucia Roberta Tough Bronze	RB	89	89	...

1. Nyers adattábla részlete

A nyers adatsor 110 oszlopból és 1857 sorból áll. Az oszlopok típusait tekintve 47 nomenklatúra és 63 mutató szerepel benne. Ezek közül sokat el fogok dobni a későbbiekben, mert az elemzés szempontjából redundánsak.

Az adatbázis 2016-2023 közötti adatokat tárol, azonban csak a 2022-es kiadásában, tehát a FIFA23-ban jelentek meg a női csapatok klub szinten is, előtte csak válogatottak szerepeltek benne, sok hiányos oszlopokkal, köztük a fizetésekkel is, ezért csak a 23-as adatsorokat vettem figyelembe, azon belül is a legújabb frissítést. Három liga szerepelt az így kapott táblában: francia és angol első osztály, illetve a világbajnokság. Ha a feltüntetett liga világbajnokság volt, 91 esetben nem volt érték a fizetés oszlopban. A két bajnokság közül az angolt választottam, mert kevesebb NaN értéket tartalmazott, csupán 5%-a NaN.

A csak angol ligát tartalmazó adattábla 216 sorból áll, 15 oszlopa tartalmaz összesen 1750 hiányzó értéket, ami 5,7%-ot jelent.

Az adatokat két részre osztottam:

- df_skills: a játékos labdarúgáshoz szükséges képességeit tartalmazza, mint például sebessége, mennyire jól passzol, lő, cselez, szerel, összesen 44 mutató

- df_info: a játékos egyéb tulajdonságait tartalmazza, mint például kor, magassága, nemzetisége, klubja, 3 nomenklatúra és 9 mutató

Mindkét tábla tartalmazza a játékos fizetését és értékét is.

	overall	potential	value_eur	wage_eur	pace	shooting	passing	dribbling	defending	physic	...
0	91	91	134500000.0	4000.0	87.0	91.0	74.0	90.0	42.0	83.0	...
1	89	89	110000000.0	3000.0	81.0	88.0	86.0	88.0	67.0	73.0	...
2	88	89	111000000.0	3000.0	83.0	88.0	70.0	88.0	32.0	82.0	...
3	88	88	72000000.0	2000.0	79.0	81.0	84.0	90.0	65.0	67.0	...
4	87	87	85000000.0	2000.0	90.0	79.0	85.0	85.0	64.0	63.0	...

3. df_skills tábla részlete

	player_positions	value_eur	wage_eur	age	height_cm	weight_kg	club_team_id	club_contract_valid_until_year	nationality_id	international_reputation
0	ST	134500000.0	4000.0	28	168	66	116010.0	2024.0	195	5
1	RW, LW, RM	110000000.0	3000.0	27	163	58	116009.0	2025.0	14	4
2	ST, CAM	111000000.0	3000.0	25	178	65	116009.0	2023.0	34	5
3	CM, CAM, CDM	72000000.0	2000.0	32	162	57	116009.0	2023.0	42	4
4	LM, CAM, ST	85000000.0	2000.0	27	167	60	116010.0	2025.0	36	3

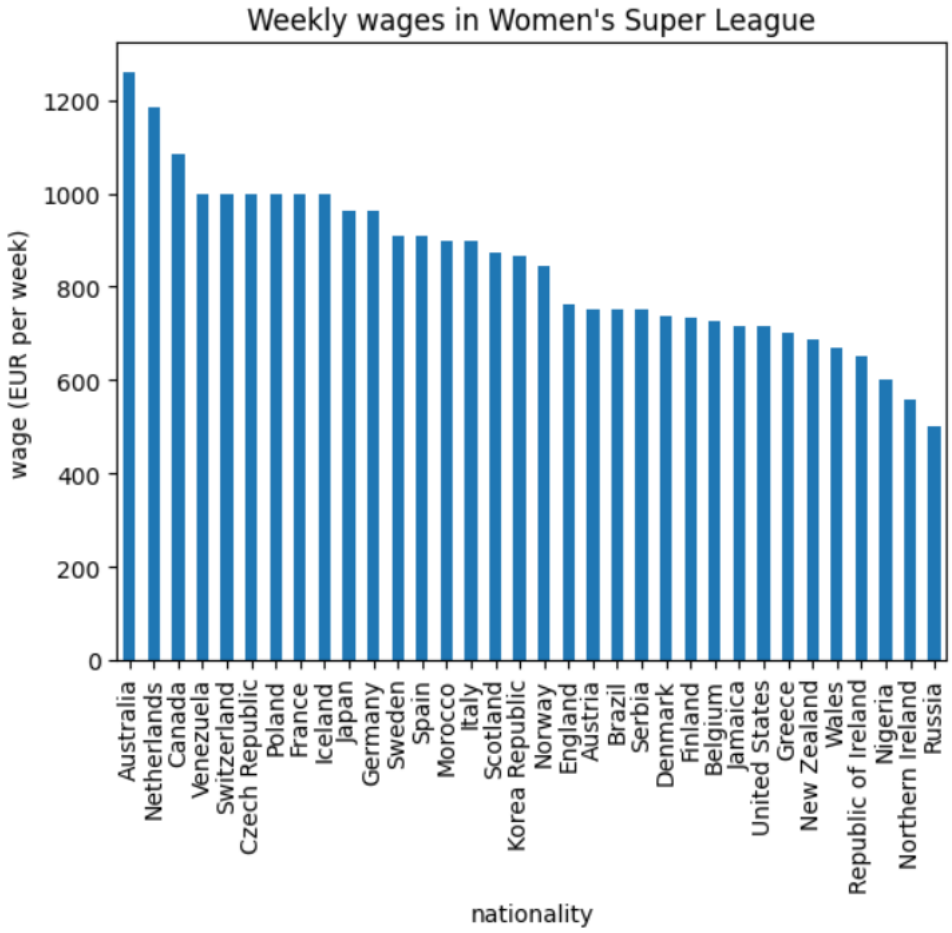
2. df_info tábla részlete

Az angol elsőosztályban, vagyis a Women's Super League-ben az átlagos heti fizetés 823€. Ez éves szinten 42796€. (Érdekesség, hogy ez körülbelül a kétszerese a női átlagfizetésnek az egész Egyesült Királyságot tekintve. A férfi angol elsőosztállyal összehasonlítva pedig hatalmas különbséget fedezhetünk fel: a Premier League-ben ennek közel másfélszeresét keresik meg, hetente.)

Azonban meg kell jegyezni, hogy a fizetések közel sem normális eloszlásúak, az átlag ugyanis a harmadik kvartilisba esik:

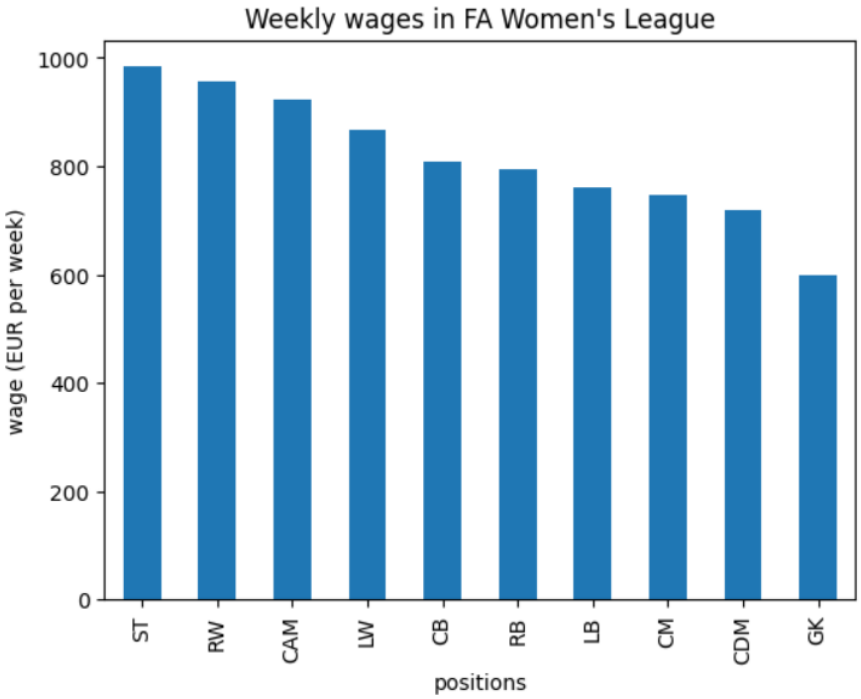
mean	810.326087
std	424.898151
min	500.000000
25%	500.000000
50%	750.000000
75%	1000.000000
max	4000.000000

4. Fizetések eloszlása



5. Átlagos fizetés nemzetiségekre lebontva

Ha nemzetekre lebontva nézzük az átlag fizetést, szembetűnő az első helyen Ausztrália. A bűnös: Sam Kerr, a Chelsea csatára, aki 4000 eurót visz haza hetente, ezzel bőven lekörözve mindenki mást. Elgondolkodtathat egyeseket, hogy vajon miért van olyan hátul Anglia? Nem becsülik meg a hazai játékosokat? Külföldieket előnyben részesítik? A válasz: nem. A magyarázat, hogy a külföldi játékosokat legtöbb (és legideálisabb) esetben csak olyanokat igazolnak a klubok, akik jobbak, mint a hazai választék, ezért a fizetési igényük is magasabb. Ez nem azt jelenti, hogy az angolok gyengébbek, hanem inkább úgy lehetne felfogni, hogy a gyengébb képességű játékosok zöme angol, a legmagasabb szintet viszont csak kevesen érik el, ezért külföldről hoznak elit játékosokat, akiknek a fizetési igénye is nagyobb. Alapvetően, ha van két nagyon hasonló képességű labdarúgó, a hazait részesítik előnyben (egy normális helyen), így itt is, minek hozzanak alacsonyabb szintű külföldi játékost, ha ott van helyben az angol, aki pont ugyanannyit tud, de nincs annyi macera az átigazolással.



6. Átlag fizetés posztok szerint

Előzetesen számítottam rá, hogy a támadóbb szellemű játékosok átlagosan többet keresnek, hiszen a legtöbb sztárjátékos – férfiaknál és nőknél egyaránt - csatárok és szélsők közül kerül ki, mint például az említett Sam Kerr, Beth Mead, Alexandra Popp, vagy éppen Lewandoski és Haaland. A népszerűség pedig együtt jár a magasabb fizetéssel. Az viszont meglepő, hogy a kapusok ennyire kilógnak. Pedig nélkülük sehol sem lenne egy csapat sem, egy jobbszélsőt lehet nélkülözni, pótolni, de egy kapust nagyon nehéz. Érdekes még az is, hogy a baloldali játékosok kevesebbet keresnek mint jobboldali megfelelőjük.

2. Követelményspecifikáció

Az elemzés célja, összefüggést találni a játékosok jellemzői és képességei, valamint a fizetésük nagysága között, meghatározni, hogy mely tulajdonságokkal kell rendelkeznie egy labdarúgónak, hogy magasabb fizetésre tegyen szert.

Mivel egy bajnokságra van leszűkítve a kutatás, ezért egyenlő feltételekkel indul minden sportoló, nem kell számításba venni a helyi szabályozásokat, illetve az adott országra egyébként jellemző jövedelmi rátákat. Ugyanakkor ez azt is jelenti, hogy az eredményt nem általánosíthatjuk minden bajnokságra világszerte.

3. Megvalósíthatósági tanulmány

Összefüggések keresésére első körben korrelációt fogok használni, hogy kiderüljön, van-e kapcsolat a fizetések és jellemzők között.

A korreláció két érték közötti lineáris kapcsolat nagyságát és irányát adja meg. Akkor használjuk, ha arra vagyunk kíváncsiak, hogy két jelenség összefügg-e egymással, és ha igen, mennyire és hogyan. Értéke -1 és 1 közé esik, a következőképp értelmezhető:

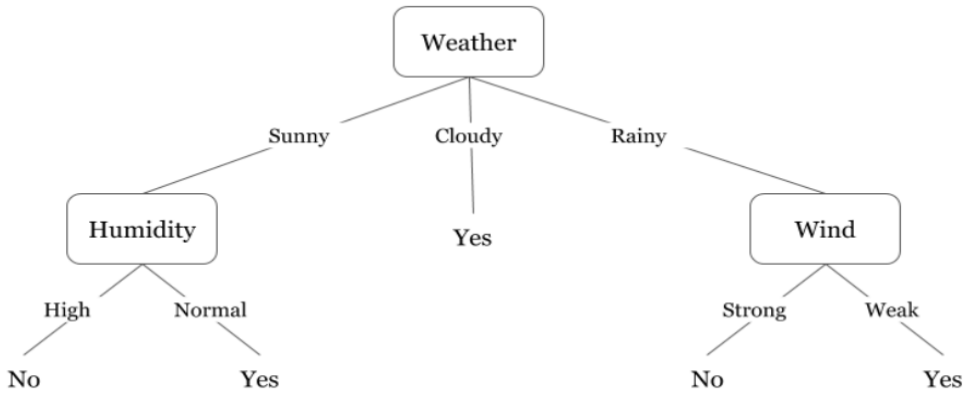
0 : nincs lineáris kapcsolat
0 - 0 .2 (-0.2 - 0) : gyenge, majdnem hanyagolható kapcsolat
0.2 - 0.4 (-0.4 - -0.2) : biztos, de gyenge kapcsolat
0.4 - 0.7 (-0.7 - -0.4) : közepes korreláció, jelentős kapcsolat
0.7 - 0.9 (-0.7 - -0.9) : magas korreláció, markáns kapcsolat
0.9 - 1 (-1 - -0.9) : nagyon magas korreláció, erős függő kapcsolat

Azonban a lineáris kapcsolat nem minden esetben jelent ok-okozati összefüggést is, így nem elég kiszámolni, értelmezni is kell a kapott értékeket.

Amennyiben találok legalább közepes korrelációt, döntési fa segítségével megpróbálom predikálni az egyes játékosok fizetését.

A döntési fa osztályozás egy tanuló algoritmus. A döntési fa egy olyan fa, melynek belső csúcsaiban egy-egy jellemző van, gyerekei pedig ennek lehetséges értékei. Példán keresztül egyszerűen megérthető, így egy nagyon egyszerű feladaton mutatom be működését. Azt szeretnénk eldönteni, hogy „jóidő” van-e, alkalmas-e focizásra, ha ma napos, párás az időjárás. Az alábbi döntési fán végig haladva megkapjuk a választ a következőképpen: a gyökércsúcsból

indulunk, ahonnan a napos gyerek felé haladunk. A páratartalom magas, ezért ma nincs jó idő focizáshoz.



7.Döntési fa példa

A döntési fák használhatók diszkrét és folytonos értékek osztályozására is. Folytonos értékek esetén vágási pontot keres, így egy diszkrét változót kapunk: kisebb vagy nagyobb a csúcs béli értéknél a keresett egyed értéke.

4. Megvalósítás

4.1. Adattisztítás

Első lépésként az korábban bemutatott két adattábla tisztítását, átalakítását végeztem.

A df_skills oszlopszámát lecsökkentettem 44-ről 17-re: az egy kategóriába tartozó oszlopokat összevontam az átlag értékük alapján. Például az ’attacking_crossing’, 'attacking_finishing', 'attacking_heading_accuracy', 'attacking_short_passing', 'attacking_volleys' értékeket egy darab attacking oszlopba mentettem. Így a következő oszlopok maradtak a df_skills táblában:

'overall', 'potential', 'value_eur', 'wage_eur', 'pace', 'shooting', 'passing', 'dribbling', 'defending', 'physic', 'attacking_attributes', 'skills_attributes', 'movements_attributes', 'power_attributes', 'mentality_attributes', 'defending_attributes', 'goalkeeping_attributes'

Ezután a df_info táblán kellett több átalakítást is végrehajtani, mivel abban object típus is volt, korrelációt pedig csak folytonos értékekkel lehet számolni. Rögtön az elején problémába ütköztem, mivel a játékos pozíciójánál egyrészt több poszt is szerepel. Itt egyszerűen a csak a legelsőt vettem figyelembe, a többit eldobtam, ezenkívül, a nagyon hasonló posztokat, mint például az LWB-LB, RW-RM, CF-CAM, összevontam. Másrészt numerikusan kódolni

nehézkes lenne, mivel nem csak simán számértékeket kell adni a stringeknek, hanem az azok közti összefüggést is kéne kódolni, mint például a csatár jellemzői közelebb állnak egy támadó középpályáséhoz, mint egy szélsővédőéhez. Első ötlet az volt, hogy vertikálisan szintekre osztom a pályát, és így a kapusok lesznek a 0. szinten, 1. szinten a védők, 2. szinten a védekező középpályások és így tovább, a 6. szinten a csatárok. Horizontálisan is hasonlóan jártam el, csak ott három részre osztottam: jobb, bal, közép.

A következő, amit szükséges volt egyedileg kódolni, az a work_rate, vagyis munkamorál oszlop volt. Három értéket vett fel: high, medium és low, de minden work_rate oszlopban kettő szerepel, / karakterrel elválasztva, egy támadó és egy védekező work rate. Mind a három stringnek adtam egy számértéket (high:3, medium: 2, low:1) és ezeket összeadtam.

A harmadik és egyben utolsó object típus, amit át kellett alakítani, az a body_type, vagyis testalkat oszlop volt. Ezek között nincs olyan kapcsolat, mint az előző két esetben, ezért ehhez egy sima OrdinalEncoder-t használtam az sklearn prepocessing csomagból.

```
pos_map={'ST':55, 'RW':54, 'CM':35, 'LW':56, 'CB':15, 'CAM':45,
         'GK':0, 'RB':14, 'LB':16, 'CDM':25}
workrate_map={'High/High':6, 'High/Medium': 4, 'Low/High':3,
'Medium/Medium':2,
       'Medium/High':4, 'High/Low':3, 'Low/Medium':1,
'Medium/Low':1}
def encode_pos(position):
  encoded=pos_map.get(position)
  return encoded
def encode_workrate(col):
    encoded = workrate_map.get(col)
    return encoded
df_info['player_positions']=df_info['player_positions']
     .apply(encode_pos)
df_info['work_rate'] = df_info['work_rate'].apply(encode_workrate)
```

8. Position és work_rate kódolása

```
from sklearn import preprocessing
e = preprocessing.OrdinalEncoder()
cat=df_info[['body_type']]
enc=e.fit_transform(cat)
enc=pd.DataFrame(enc, columns=cat.columns)
```

9. Body_type kódolása

4.2.Korreláció

Az elemzés első lépése a korreláció. Megnéztem, hogy a játékos képességei, illetve egyéb tulajdonságai közül melyek mutatnak lineáris kapcsolatot a a fizetéssel.

wage_eur	1.000000
value_eur	0.819639
overall	0.707248
power_attributes	0.574545
dribbling	0.521301
passing	0.470697
movements_attributes	0.453810
attacking_attributes	0.421609
mentality_attributes	0.417014
shooting	0.399652
potential	0.392572
skills_attributes	0.392519
physic	0.387154
pace	0.369403
goalkeeping_attributes	0.131406
defending_attributes	0.113133
defending	0.036268

10. df_skills korreláció

wage_eur	1.000000
value_eur	0.819639
international_reputation	0.774912
work_rate	0.533649
age	0.327306
player_positions	0.214618
nationality_id	0.159660
club_contract_valid_until_year	0.147131
body_type	0.049018
weight_kg	0.001415
height_cm	-0.063621
club_team_id	-0.149964

11. df_info korreláció

A két táblázatból kiolvasható, hogy bizony van kapcsolat, és ha végig gondoljuk, akkor ezek ok-okozati összefüggések, nem véletlen egybeesések. Számomra a legmeglepőbb, hogy a védekezés mennyire nem korrelál a fizetéssel, főleg, mivel általánosságban elmondható, hogy egy jó csapat egy jó védelemre épül. Illetve előzetesen arra számítottam, hogy a játékos nemzetisége nagyobb mértékben lehet befolyásoló tényező. Ez azonban valószínűleg azért ilyen alacsony, mert csak az angol bajnokságot vettük figyelembe, ahol a játékosok többsége angol, minden szintet lefedve, a legalacsonyabbtól a magasabb fizetési szintekig. Ha a két táblázatot összehasonlítjuk, látható, hogy a képességek összeségében meghatározóbbak, mint az egyéb jellemzők. Ez azért örvendetes, mert ezalapján tényleg azt díjazzák, amit egy illető letett az asztalra, amire képes, nem pedig „pofára”, egyéb diszkriminatív módszerekkel. (Magyarországon nem biztos hogy ezt kapnánk, sajnos)

4.3.Döntési fa

A döntési fa alkalmazásához egy új adattáblát hoztam létre, mely azokat az oszlopokat tartalmazza a df_skills és df_info táblákból, melyek legalább jelentős kapcsolatot, azaz |0.4|-nél jobban korrelálnak a fizetéssel. Az új, df_dt elnevezésű tábla tehát az alábbi oszlopokat tartalmazza:

'wage_eur', 'value_eur', 'overall', 'power_attributes', 'dribbling', 'passing', 'movements_attributes', 'attacking_attributes', 'mentality_attributes', 'international_reputation', 'work_rate'

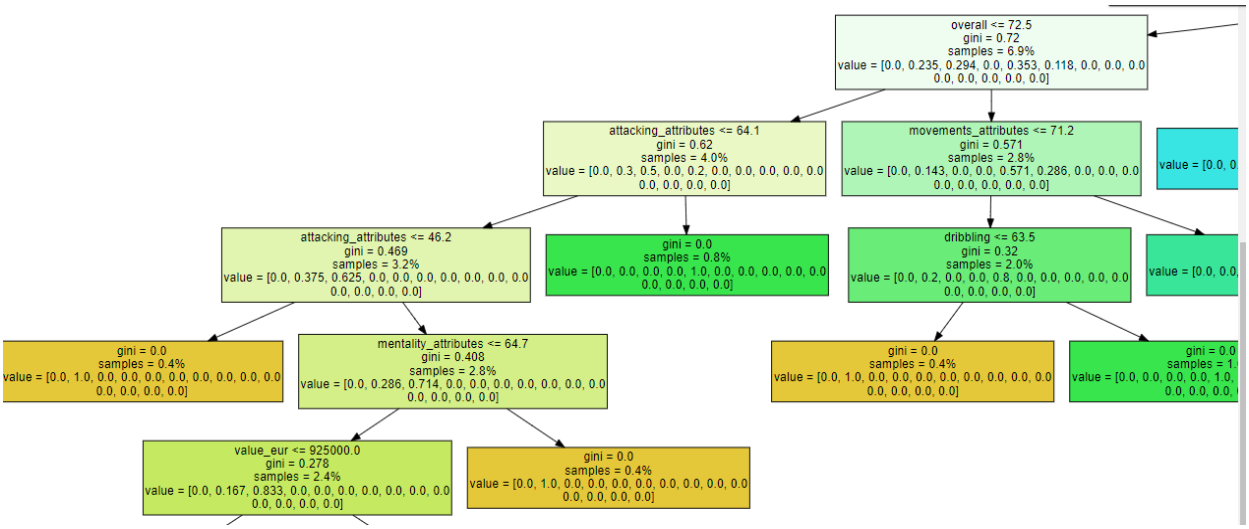
A kapusokat eldobtam, mert hozzájuk nem tartozott pace, shooting, passing, dribbling, defending és physics érték, melyek közül a dribbling és a passing fontosak számunkra, illetve korábban láttuk, hogy a kapus képességeknek és a fizikumnak, melyekben jellemzően jobbak szoktak lenni az átlagnál, nincs nagy jelentőségük a fizetésekben, így célravezetőbb ezektől a játékosoktól eltekinteni, mint a megfelelő helyettesítő értékek megtalálásával bíbelődni.

A scikit-learn függvény könyvtárnak köszönhetően a folyamat nagyon egyszerű, csupán néhány sor szükséges a döntési fa megalkotásához. Az adatokat két részre bontottam: egy tanító és egy teszt halmazra, fele-fele arányban. Majd a teszt halmaz elemein végrehajtottam a döntési fa osztályozást és a kapott értékeket összevetettem az eredeti értékekkel.

```
from sklearn.tree import DecisionTreeClassifier

from sklearn.model_selection import train_test_split

from sklearn.metrics import accuracy_score

classlabel=df_dt['wage_eur']

features=df_dt[['value_eur', 'overall', 'power_attributes',
'dribbling', 'passing', 'movements_attributes',
'attacking_attributes', 'mentality_attributes',
'international_reputation', 'work_rate']]

X_train, X_test, y_train,
y_test=train_test_split(features,classlabel,test_size=0.5)

dt=DecisionTreeClassifier()

dt.fit(features, classlabel)

prediction=dt.predict(X_test)

print(accuracy_score(y_test, prediction))
```

12. Döntési fa osztályozás

Az eredmény: accuracy_score=1.0, vagyis tökéletesen meghatározható egy játékos jövedelme képességei alapján.



10. Döntési fa részlete

5. Összegzés

Kimondható, hogy az angol női elsőosztályban kiszámítható a játékosok fizetése, viszonylag jól behatárolható, hogy milyen tulajdonságokkal kell rendelkezni annak, aki sokat szeretne keresni. A legjobban az következő képességekre érdemes nagyobb hangsúlyt fekteni: erő, cselezés, passzolás, munkamorál, de nagyon ajánlott a válogatottba is bekerülni, nemzetközileg is elismertté válni.

Azonban számos új kérdést is felvet, amelyekkel a jövőben érdemes lehet foglalkozni. A legsarkalatosabb pont a nők és férfiak közti különbségek, bár ez inkább társadalmi-gazdasági kérdés. Számomra elgondolkodtató viszont, hogy az egyes posztok közti hatalmas különbségeknek mi lehet az oka. Ezenkívül, ha több bajnokság adata is rendelkezésre állna, kíváncsi lennék, milyen különbségek lennének azok között, illetve összesítve milyen eredményre jutnánk.

6. Források

Adat:

https://www.kaggle.com/datasets/stefanoleone992/fifa-23-complete-player-dataset?select=female_players.csv

Korreláció értékek:

https://psycho.unideb.hu/munkatarsak/balazs_katalin/stat1/stat1ora3.pdf

Döntési fa:

https://www.inf.u-szeged.hu/~rfarkas/ML21/dontesi_fa.html

Python segédlet:

http://www.inf.u-szeged.hu/~rfarkas/DS/

A kódom:

https://github.com/lovajujo/statszoftverek/blob/main/players_salary.ipynb