OZNAL - Dátová analýza dát hráčov z hry FIFA 19

Timotej Zaťko and Tomáš Hoffer

Fakulta informatiky a informačných technológií STU v Bratislave Ilkovičova 2, 842 16 Bratislava 4
xzatkot1@stuba.sk xhoffer@stuba.sk
https://www.fiit.stuba.sk/

Abstract. Táto práca obsahuje analýzu dát hráčov z hry FIFA 19, opis dát a ich charakteristiky. V tejto práci skúmame vzťahy atribútov na predikovaný atribút trhovej ceny hráča a jeho hernú pozíciu na ihrisku.

Keywords: Analýza dát, FIFA 19, Strojové Učenie, Regresia, Klasifikácia

1 Opis problému a motivácia

Rozhodli sme sa pre analýzu dát hráčov z hry FIFA 19. V hre sa nachádza veľké množstvo hráčov z rôznych krajín, hrajúcich v rôznych súťažiach. Ich schopnosti v hre by mali odzrkadlovať ich schopnosti z reálneho sveta. Tvorcovia hry sa o to snažia vytvorením herných atribútov, akými sú napríklad rýchlosť šprintu, sila strely, zakončovanie alebo hlavičkovanie. Keďže dokázali vyjadriť tieto schopnosti hráčov číselne (na určitej škále), zaujíma nás či, a ako herná pozícia hráča závisí od týchto atribútov. Ďalšou zaujímavou otázkou je ako tieto atribúty (a ktoré najviac) vplývajú na trhovú cenu hráča v hre.

2 Opis dát s charakteristika dát

Naša dátová sada obsahuje 18207 záznamov - hráčov, ktorý každý z nich má 87 atribútov. Z toho je 42 numerických a 45 kategorických.

2.1 Očistenie dát

Kvôli analýze bolo treba dátovú sadu, očistiť a urobiť predspracovanie niektorých atribútov. Bolo nutné spraviť nasledovné úpravy – konverzia a očistenie finančných hodnôt (napr. '$77.5M', '$1K'), konverzia mier (napr. '159lbs', '5'11'), konverzia časových údajov (napr. 'Jan 25, 2019', '2018') a rozdelenie niektorých atribútov na viac atribútov - niektoré atribúty obsahovali dva numerické atribúty. Taktiež atribút určujúci pozíciu hráča obsahoval až 38 rôznych herných pozícii, preto sme sa rozhodli tento atribút rozšíriť do ďalších dvoch atribútov v ktorých

sme podobné pozície spojili čím vznikli dva atribúty s 13 resp. 4 hodnotami. Po týchto úpravách sa nám počet atribútov zmenil na 129 – 89 numerických a 40 kategorických atribútov.

Toto veľké množstvo atribútov sme zaradili do nasledovných kategórií, uvádzame k nim aj niektoré atribúty. Atribúty sme rozdelili aj na základe toho či jeho hodnotu musela hra nejakým spôsobom odvodiť z reálneho sveta.

- **človek** - meno, národnosť, výška, hmotnosť, vek...
- **futbalový hráč (reálny svet)** - názov klubu, číslo dresu, dátum príchodu do klubu, dĺžka kontraktu, plat hráča, hosťovský klub, herná pozícia, preferovaná noha
- **futbalový hráč (hra)** - trhová hodnota hráča, triky, pracovitosť v útoku/obrane, vhodnosť na určitú špecifickú pozíciu a potenciálny rast (78 atribútov), medzinárodná reputácia
- **futbalové schopnosti hráča určené hrou** (celkovo 34 atribútov) - krátke prihrávky, hlavičkovanie, zakončovanie...
- **iné atribúty hry** - logo klubu, vlajka (na základe národnosti), typ postavy (tj. typ herného modelu), typ tváre (tj. kvôli herného modelu)

2.2 Analýza chýbajúcich hodnôt

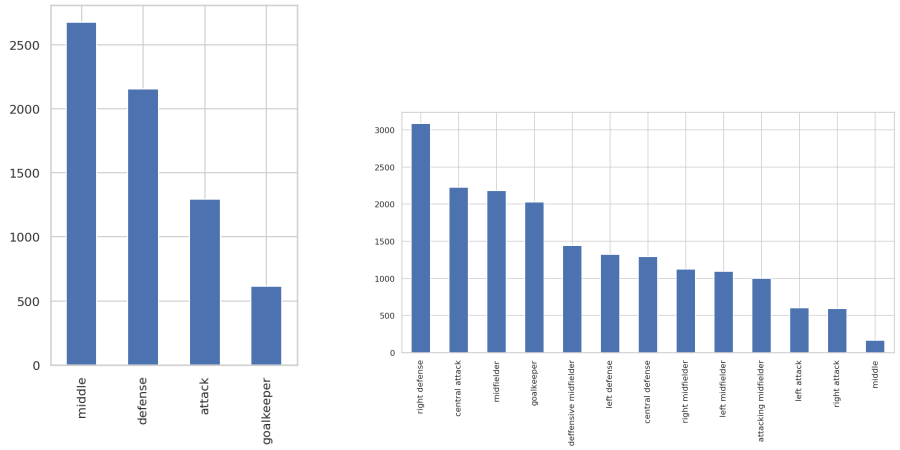
Ako prvé sme sa pozreli na chýbajúce hodnoty v našej dátovej sade. Početnosti sme vizualizovali stĺpcovým diagramom a vzťahy tepelnou mapou a dendrogramom. Zistili sme, že niektorým hráčom chýbajú hodnoty v atribútoch ako sú klub, plat, dĺžka kontraktu, číslo dresu či výkupná klauzula. Tieto hodnoty chýbajú väčšinou spoločne, o čom sme sa presvedčili v dendrograme a tepelnej mape. Je to aj logické, keďže hráč, ktorý nemá klub nemôže poberať plat alebo mať výkupnú klauzulu v kontrakte. Z našich zistení môžeme konštatovať, že ak hráčovi chýbajú hodnoty v atribúte, väčšinou sa nejedná o chybu v úplnosti dátovej sady.

2.3 Analýza z pohľadu pozície hráča

Analýzu sme ďalej realizovali z pohľadu dvoch atribútov, ktoré sa budeme snažiť predikovať - pozíciu hráča (kategorický atribút) a jeho trhovú hodnotu (numerický atribút). Dátová sada celkovo obsahuje 36 herných pozícií.

Na základe našich futbalových znalostí sme sa rozhodli zoskupiť podobné pozície do skupín, čím sme znížili počet rôznych hodnôt v atribúte ‘Position‘ (Fig. 1). Môžeme pozorovať, že triedy nie sú vyvážené.

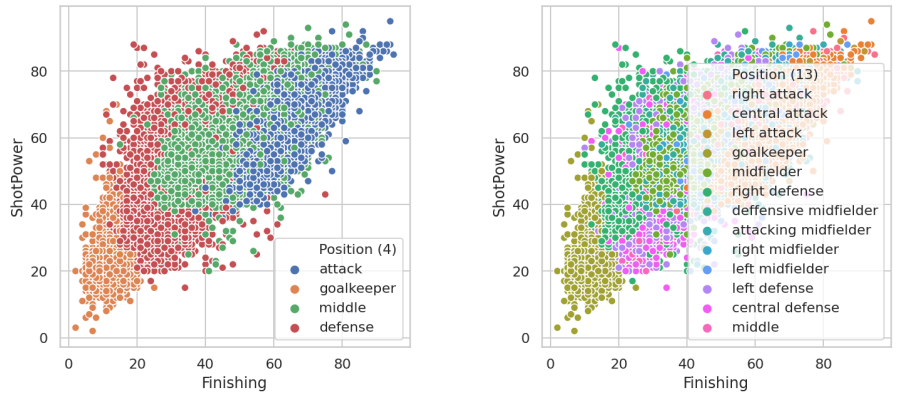
Pozíciu sa môžeme pokúsiť predikovať na základe 34 atribútov definujúcich futbalové schopnosti hráča určené hrou. Vizualizovali sme všetky dvojice týchto atribútov a pomocou ”scatter plot” -ov sme hľadali zhluky jednotlivých pozícií. Ako príklad uvádzame vzťah medzi atribútmi ‘Shot Power‘ (sila strely) a ‘Finishing‘ (zakončovanie), kde môžeme pozorovať jednotlivé zhluky podľa pozície hráča (Fig. 2).



(a) 4 herné pozície(‘Position (4)‘)

(b) 13 herných pozícií (‘Position (13)‘)

Fig. 1: Početnosti nových atribútov ‘Position (4)‘ a ‘Position (13)‘ po zoskupení podobných hodnôt z atribútu ‘Position‘.



(a) 4 herné pozície (‘Position (4)‘)

(b) 13 henrých pozícií (‘Position (13)‘)

Fig. 2: Vzťah medzi atribútmi ‘Shot Power‘ a ‘Finishing‘. Pre 4 herné pozície sú zhluky zreteľnejšie ako pre 13 herných pozícií.

Z prieskumnej analýzy tiež vyplýva, že pre konkrétne pozície hráčov sú typické určité čísla dresov. Pre brankárov (angl. goalkeeper) je typickým číslom dresu číslo *1* pričom toto číslo nemá priredený žiaden hráč na inej pozícii. Pre útočníkov (angl. attack) to je *9*, pre obrancov (angl. defense) sú to čísla *2* - *6* a pre stredopoloarov *7*, *8* a *10*. Tento atribút môže byť veľmi dobrý na klasifikáciu pozície hráča, avšak my sa v prvom rade zamieriame na predikciu z herných atribútov. Atribúty, ako napríklad čislo dresu nám môžu úlohu príliš zjednodušiť.

Ďalšou zaujímavosťou je, že v našej dátovej sade sa nachádzajú prevažne hráči ktorí preferujú pravú nohu, avšak na pozícii ľavého obrancu výrazne prevládajú hráči s preferovanou ľavou nohou. (Fig. 3).

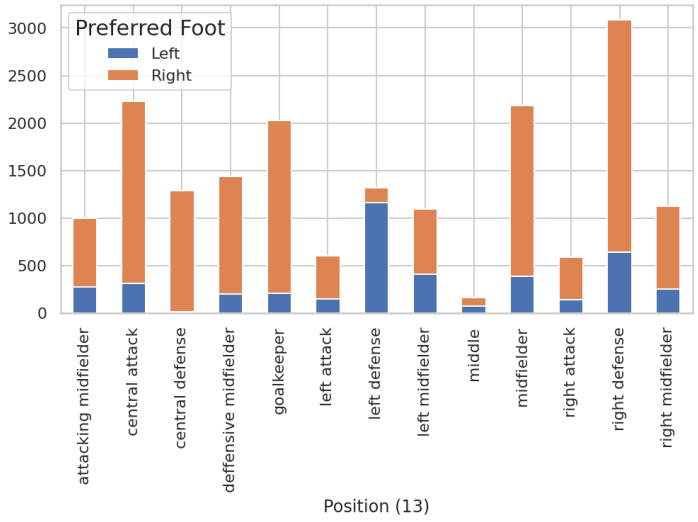


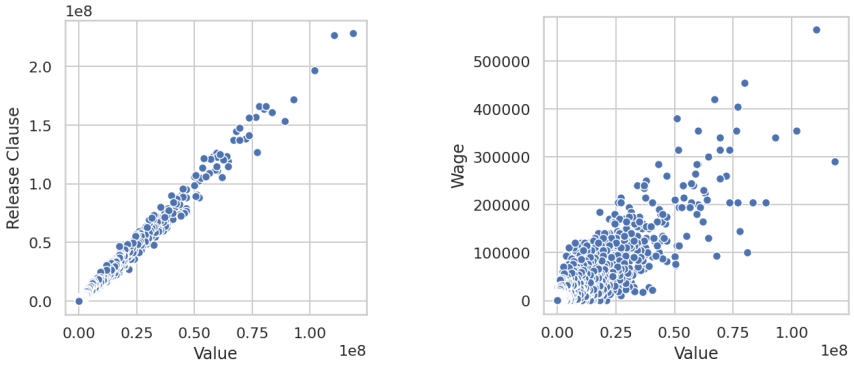
Fig. 3: Preferovaná noha hráčov podľa hernej pozície.

2.4 Analýza z pohľadu trhovej hodnoty hráča

Pomocou Pearsonovho korelačného koeficientu sme hľadali korelácie medzi atribútom ‘Value‘ a ostatnými numerickými atribútmi. Takmer lineárnu koreláciu (0.99) vykazuje atribút ‘Release Cause‘ (Fig. 4). Hráči s vysokou trhovou hodnotou majú v datasete podpísanú zmluvu s vyššou výkupnou klauzulou. Vysokú koreláciu taktiež vykazujú atribúty ‘Overall‘ (0.631), ‘Wage‘ (0.850) a ‘International Reputation‘ (0.656).

3 Definovanie úlohy objavovania znalostí

Rozhodli sme sa, že budeme vykonávať nasledujúce úlohy:



(a) Vzťah 'Release Cause' voči 'Value'.

(b) Vzťah 'Wage' voči 'Value'.

Fig. 4: Niektoré atribúty vykazujú vysokú mieru korelácie s atribútom 'Value'.

– predikcia hernej pozície hráča – všeobecnej (4 triedy), rozšírenej (13 tried)
– predikcia hodnoty hráča (atribút s názvom 'Value')

Obe tieto úlohy budeme realizovať z atribútov určujúcich herné shopnosti hráča a následne zo všetkých atribútov. Výsledky porovnáme a očakávame, že model natrénovaný zo všetkých atribútov bude výrazne lepší, keďže niektoré atribúty výrazne ovplyvňujú predikovanú premennú a to – Release Clause - Value; Jersey Number - Position.

4 Predpokladaný scenár riešenia (problémy)

Predpokladáme, že bude potrebné vykonať nasledovné úlohy:

– predspracovanie kategorických hodnôt (tj. one-hot encoding)
– normalizácia dát
– odstránenie odľahlých pozorovaní
– trénovanie modelu

Trénovanie modelu zahŕňa výber atribútov (angl. feature selection) a výber a trénovanie modelu. Na úlohu predikcie hodnoty hráča budeme pravdepodobne používať lineárnu regresiu/neurónovú sieť a na určenie hernej pozície hráča rozhodovací strom / náhodný les / SVM / neurónovú sieť.

5 Predspracovanie dát

Prvotnému predspracovaniu dát na účely analýzy sme sa venovali v kapitole 2.1 Jedná sa o nasledujúce transformácie dát:

- Transformácia hodnôt peňažných atribútov (napr. hodnoty 1.2M, 200K sme transformovali na jednotný numerický tvar)
- Dátumové atribúty sme transformovali na jednotný tvar, reprezentovaný UNIX časovou pečiatkou
- Atribúty miery (napr. výška a hmotnosť hráča) sme transformovali tak, aby boli v spoločných jednotkách
- Špeciálne pozície hráča boli udávané v tvare reťazca *Value+Grow*. Atribúty špeciálnych pozícií sme rozdelili na dvojice aby nám vznikli numerické atribúty.
- Atribút *Work Rate* bol udávaný reťazcom v tvare *Attack/Defense*. Tento atribút sme tiež rozdelili na 2 numerické.
- Boolean atribúty udávané reťazcami sme transformovali na numerické 0/1
- Vytvorili sme 2 nové atribúty Position(4) a Position(13), ktoré reprezentujú pozíciu hráča po zoskúpení pozícií, opísanom v kapitole 2.3
- Vytvorili sme atribút *Contract length*, ktorý reprezentuje dĺžku aktuálne podpísanej zmluvy hráča

Pre účely klasifikačnej a regresnej úlohy sme sa rozhodli kategorické atribúty kódovať pomocou "One-Hot encoding" -u. Numerické atribúty sa nachádzali v rôznych rozsahoch a jednotkách (napríklad hmotnosť v lb, výška v cm, rôzne schopnosti hráča v rozsahu [0,100]). Rozhodli sme sa ich normalizovať na jednotný interval [0,1].

5.1 Chýbajúce hodnoty

V kapitole 2.2 sme opísali dôvod výskytu chýbajúcich hodnôt. Chýbajúce hodnoty opísané v kapitole 2.2 považujeme za opodstatnené a nebudeme ich nahrádzať.

5.2 Vychýlené hodnoty

Rozdelenie číselných hodnôt v datasete sme podrobne analyzovali a vyzualizovali. Na základe podrobnej analýzy sme dospeli k záveru, že dataset neobsahuje žiadne vychýlené hodnoty.

5.3 Problém nevyváženosti tried

Pri zoskúpení všetkých pozícií do 13 tried sme odhalili nevyváženosť jednotlivých tried (Fig.1), čo by mohlo negatívne ovplyvniť presnosť predikcie. Tento problém sme sa rozhodli riešiť algoritmom SMOTE, ktorý využíva tzv. "oversampling" a vytvára syntetické inštancie minoritnej triedy pomocou lineárnej kombinácie reálnych inštancií.[1]

6 Výber atribútov

Predikciu pri oboch typoch úloh sa budeme snažiť realizovať pomocou 34 atribútov definujúcich futbalové schopnosti hráča určené hrou. Využijeme tiež demografické

údaje, napr. výšku, hmotnosť a vek. K tejto základnej množine atribútov sme sa dopracovali na základe doménovej znalosti a rozsiahlej prieskumnej analýzy. V častiach 2.3 a 2.4 sa venujeme výberu týchto atribútov a opisujeme dôvody, prečo sme niektoré atribúty do tejto množiny nezahrnuli. V druhej etape riešenia vyskúšame úlohy predikcie realizovať s využitím všetkých atribútov a úspešnosti modelov porovnáme. Očakávame mierny nárast úspešnosti pri použití všetkých atribútov.

Pri riešení klasifikačnej úlohy rozšírime základnú sadu atribútov o atribúty *Preferred Foot*, *Work Rate Attack* a *Work Rate Defense*. Pri regresnej úlohe využijeme dopočítaný atribút *Contract Length* a tiež medzinárodnú reputáciu hráča.

7 Výber metrík pre evaluáciu úspešnosti TODO

8 Predikcia pozície hráča (klasifikácia)

Úlohu sme sa v prvej etape rozhodli riešiť pomocou viacerých jednoduchých modelov. Klasifikácia hráča do 4 pozícií už pri základných nastaveniach modelov vykazovala vysokú úspešnosť. V nasledujúcej tabuľke uvádzame modely a ich úspešnosti pri predikcii pozície. Ako metriku sme sa rozhodli sledovať f1 skóre.

model	f1-skóre (pozícia)			
	attack	defense	goalkeeper	middle
Rozhodovací strom	0,74	0,86	1,0	0,76
Logistická regresia (10-násobná krížová validácia)	0,82	0,93	1,0	0,86
Metóda podporných vektorov	0,80	0,93	1,0	0,85

Table 1: Úspešnosti jednoduchých modelov pri predikcii 4 tried pozície

Pri klasifikácii do 13 tried pozície je úspešnosť jednoduchých modelov významne nižia. F1-skóre je pre niektoré pozície vysoké (right defense = 0,92) ale niektoré pozície model vôbec nepredikoval (middle = 0,0) Úlohu budeme ďalej riešiť pomocou neurónových sietí.

9 Predikcia trhovej hodnoty hráča (regresia)

10 Existujúce práce TODO dokoncit

Soto-Valero, C. na klasifikáciu pozície hráča do štyroch tried v rovnakom datasete najprv využil PCA na extrakciu čŕt a redukciu dimenzionality. Následne využil zhlukovanie na klasifikáciu hráčov do 4 pozícií a pomocou "gradient tree boosting" algoritmu ohodnotil dôležitosť jendotlivých atribútov. Medzi najdôležitejšie atribúty na predukciu patria "dribbling", "standing tackle" a "goalkeeper reflexes". [2]

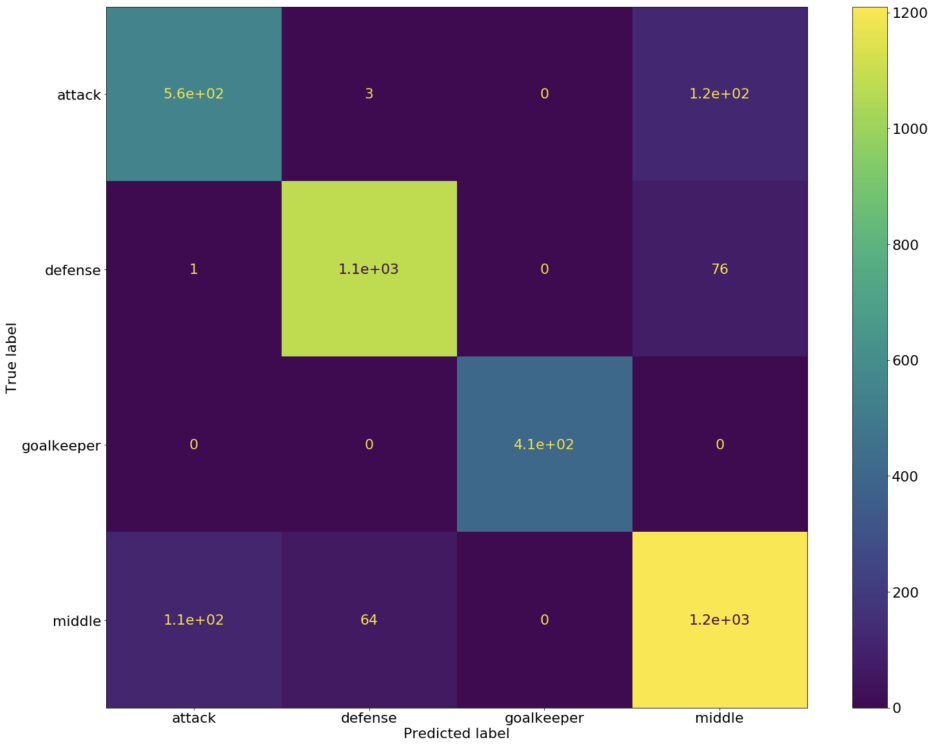


Fig. 5: Chybová matica pre logistickú regresiu (4 pozície).

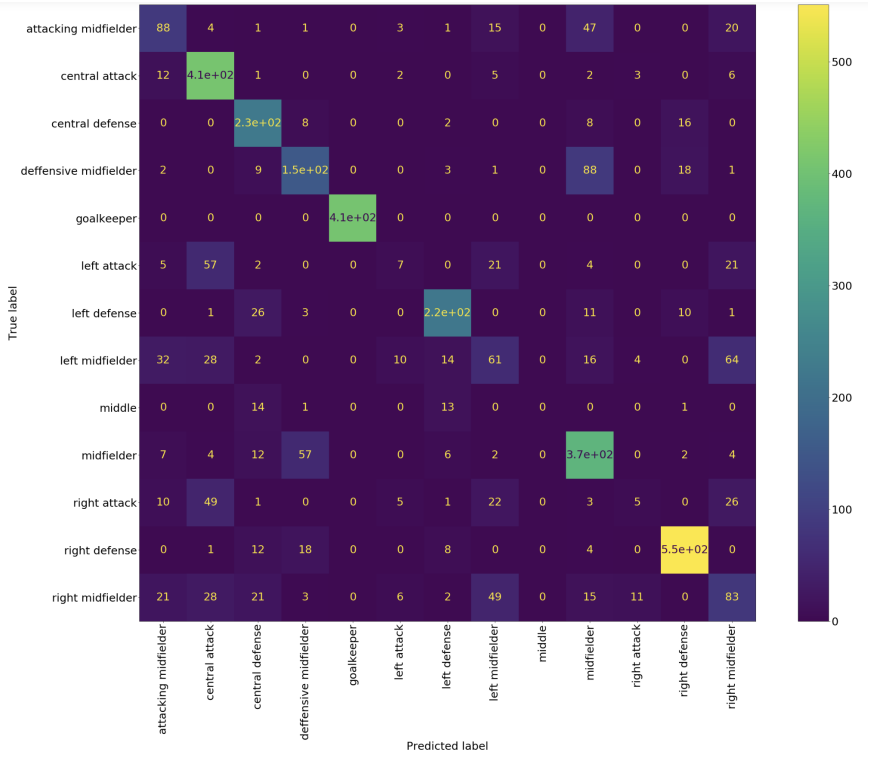


Fig. 6: Chybová matica pre logistickú regresiu (13 pozícií).

References

1. Alberto Fernandez, Salvador Garcia, Francisco Herrera, Nitesh V. Chawla: SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research **61**, 863–905 (2018), https://www.jair.org/index.php/jair/article/view/11192
2. Soto-Valero, C.: A gaussian mixture clustering model for characterizing football players using the ea sports' fifa video game system. [modelo basado en agrupamiento de mixturas gaussianas para caracterizar futbolistas utilizando el sistema de videojuegos fifa de ea sports]. RICYDE. Revista Internacional de Ciencias del Deporte. doi:10.5232/ricyde **13**(49) (2017), https://www.cafyd.com/REVISTA/ojs/index.php/ricyde/article/view/1165