GÉPI TANULÁS ÉS BESTSELLEREK

Szerző:

Uzonyi Noémi
Debreceni Egyetem

A szerző e-mail címe:
unomi95@gmail.com

Lektorok:

Mező Ferenc (PhD)
Eszterházy Károly Katolikus Egyetem

Simó Ferenc Zoltán (Dr. Jur.)
Eszterházy Károly Katolikus Egyetem

…és további két anonim lektor

Absztrakt

Jelen tanulmány tanulmány az Amazon weboldalon elérhető könyvek bestseller listájának adathalmazán tesztelt alábbi osztályozási módszereket hasonlítja össze: mesterséges neurális hálózat, támogató vektor gép, logisztikus regresszió, döntési fa, véletlen erdő, k legközelebbi szomszéd, és naiv Bayes-osztályozó. Bemutatjuk a különböző osztályozási módszerek pontosságának összehasonlító elemzését is. Az adatsor tizenegy év vonatkozásában tartalmaz adatokat az évi ötven legtöbbet eladott könyvről.

Kulcsszavak: gépi tanulás, bestseller

Diszciplinák: informatika, irodalomtudomány

Abstract

MACHINE LEARNING AND BESTSELLERS

The present study compares the below classification methods tested on the best-selling data set of books available on the Amazon website: Artificial Neural Network, Support Vector Machine, Logistic Regression, Decision Tree, Random Forest, k Nearest Neighbors, and Naive Bayes methods. A comparative analysis of the accuracy of different classification methods is presented too. The data set for eleven years contains data on the fifty best-selling books of the year.

Keywords: machine learning, bestseller

Disciplines: computer science, literary studies

Uzonyi Noémi (2021): Gépi tanulás és bestsellerek. *Mesterséges intelligencia – interdiszciplináris folyóirat, III. évf. 2021/2. szám.* 43-53. doi: 10.35406/MI.2021.2.43

E tanulmány az Amazon weboldalon elérhető könyvek bestseller listájának adathalmazán tesztelt különböző osztályozási módszerek összehasonlítását mutatja be. Az adatsor tizenegy év vonatkozásában tartalmaz adatokat az évi ötven legtöbbet eladott könyvről. Az adatelemzési és adatbányászati eszközök alkalmazásával kvalitatív és kvantitatív változók feldolgozására, valamint nagy dimenziós adat vizualizálására van lehetőségünk. A tanulmány fókuszában gépi tanulási módszerek alkalmazását mutatjuk be Python nyelven az adatkészlet fikcióra és valós történeten alapuló kötetekre való osztályozására. Hét osztályozási módszert tekintünk át, ezek: neurális hálózat, logisztikus regresszió, k legközelebbi szomszéd, tartóvektor-gép, döntési fa, véletlen erdő és naiv Bayes-osztályozó. Végezetül ismertetjük a különböző osztályozási módszerek pontosságának összehasonlító elemzését.

Bevezetés és statisztikai elemzés

A tanulmány különböző gépi tanulási eszközök osztályozási feladatra való alkalmazására koncentrál. Az elemzés során egy adathalmazból indulunk ki, amely valamilyen szempont alapján diszjunkt csoportokra osztható fel. Célunk, hogy az adatok mélyére ásva olyan új ismeretet detektáljunk, mely segítségével képesek lehetünk nagy valószínűséggel meghatározni egy ismeretlen kategóriába tartozó egyed hovatartozását. Ehhez különböző mesterséges intelligencián alapuló eljárásokkal modelleket építünk. A modellek illesztését követően az osztályozás teljesítményét mérő metrika összevetésével igyekszünk a modelleket bíráló megállapítás megtételére.

Egy adathalmazból indulunk ki, ami esetünkben 2009 és 2019 között tartalmazza az Amazon webáruház évi ötven legtöbbet eladott könyvét (kaggle.com). Az egyes könyveket az alábbi tulajdonságok jellemzik:

- Cím (Name): a könyv címe
- Szerző (Author): a könyv szerzője
- Felhasználói értékelés (User Rating): A-mazon felhasználók értékelése
- Vélemények (Reviews): az Amazon-on írt vélemények száma a könyvről
- Ár (Price): a könyv ára 2020. október 13-án
- Év (Year): az év, amikor a könyvet bestsellerként sorolták be
- Műfaj (Genre): a könyv műfaja (fikció vagy sem).

A tanulmány célja olyan mesterséges intelligencián alapuló modellek definiálása, melyek hatékonyan képesek a könyvek osztályozására a műfaj tekintetében. Különböző osztályozási eljárások összevetésével kívánjuk megtalálni azt az eljárást, amely a lehető legjobban képes eldönteni a tulajdonságok ismeretében azt, hogy egy adott mű fikció-e vagy sem.

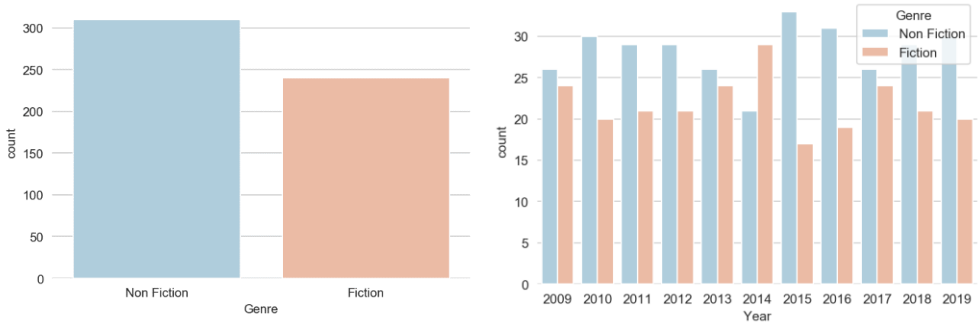
Első lépésként statisztikai szempontból vizsgáljuk az adatokat. Az adathalmaz 550 rekordot tartalmaz 351 egyedi könyvcímmel, tehát megállapítható, hogy számos könyv több évben is szerepelt a bestsellerek között. Az adatkészlet három nominális és négy numerikus változóból

áll. A nominális változók közül az osztályozási feladatban a műfaj kategóriaváltozót célváltozóként határozzuk meg. Az adatkészlet 240 fikciót és 310 non-fikciót tartalmaz. A célváltozó éves eloszlásait tekintve megállapíthatjuk, hogy a 2014. év kivételével minden évben több valós történeten alapuló könyv került be a bestsellerek listájába, mint fikció (1. ábra).

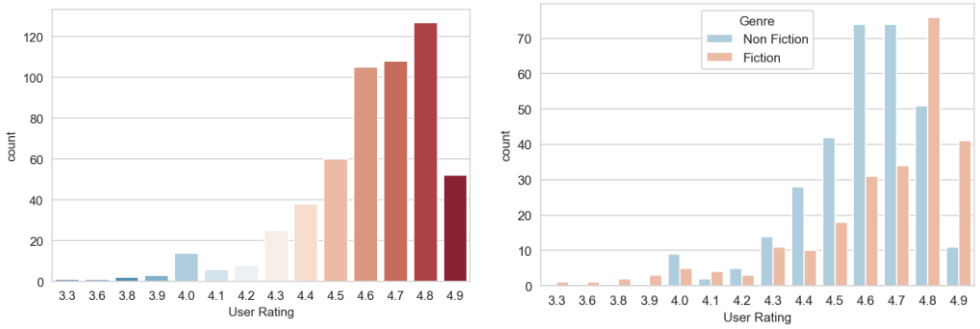
A felhasználói értékeléseket tekintve a könyvek átlagosan 4,62 pontot kaptak (2. ábra). Az értékelések eloszlása jobbra ferdült, balra elnyúló normális eloszlást mutat. A fikciók átlagos pontszáma 4,65, a non-fikciók áltagos pontszáma pedig 4,60. Megfigyelhetjük, hogy a non-fikciók értékelése kisebb, 0,19 szórást mutat a fikciók 0,27-es szórásához képest. A non fikciók pontszámai 4,0 és 4,9 közöttiek, így terjedelme jelentősen kisebb a fikciók értékelésének 3,3-tól 4,9-ig terjedő értékeivel szemben.

Az árak balra ferdült, jobbra elnyúló, 13,1 átlagú normális eloszlást mutatnak (3. ábra). A fickók átlagos ára 10,85 dollár, míg a non-fikcióké 14,84 dollár. Megfigyelhetjük, hogy a fikciók ára 0 és 82,0 dollár között mozog 8,57 szórással, míg a nem fikcióké 0 és 105 dollár közötti 12,04 szórással.

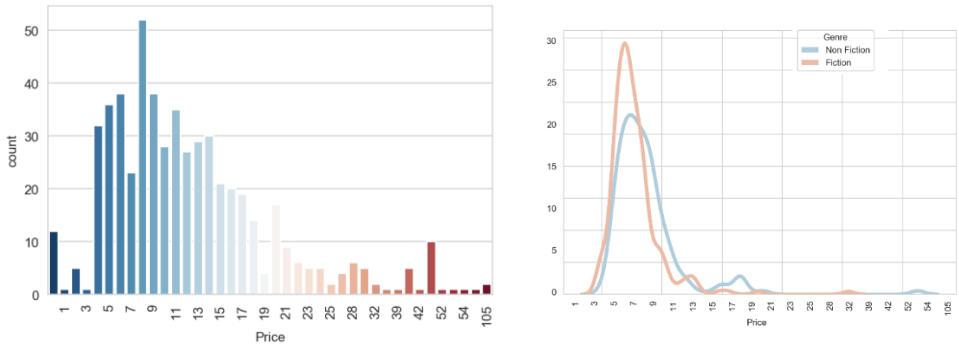
1. ábra: Az adathalmazban található fikciók és non-fikciók darabszáma. Forrás: a Szerző



2. ábra: A felhasználói értékelések gyakorisága. Forrás: a Szerző

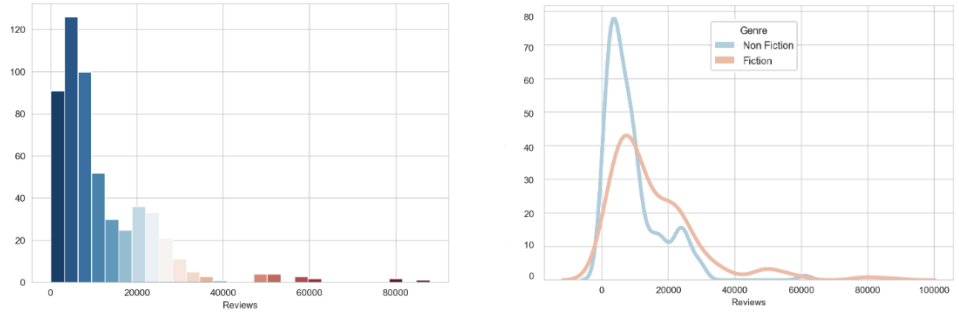


3. ábra: Az árak értékének gyakorisága és eloszlása. Forrás: a Szerző

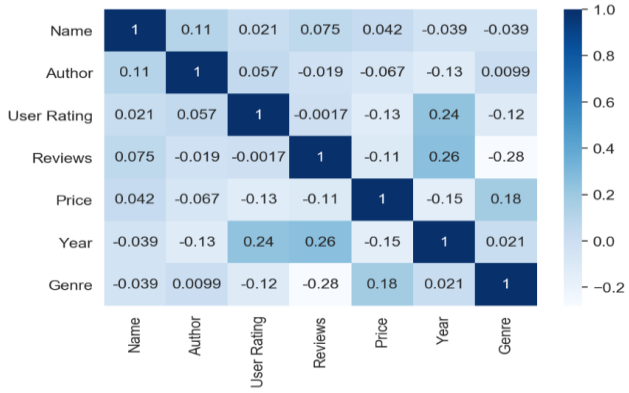


Az aggregált vélemények számának átlaga könyvenként 11953. A fikciók átlagosan 15684, míg a nem fikciók átlagosan 9065 véleményt kaptak. Megfigyelhetjük, hogy a fikciókra érkezett vélemények számának eloszlása laposabb, míg a non-fikcióké csúcsosabb (4. ábra). Az adatkészletben található változók korrelációs mátrixát (5. ábra)

4. ábra: A vélemények számának gyakorisága és eloszlása. Forrás: a Szerző



5. ábra: Az adatkészlet változóinak korrelációs mátrixa. Forrás: a Szerző



tekintve megállapíthatjuk, hogy az egyes attribútumok között nincs 0,26-nál erősebb kapcsolat, vagyis a változók korrelálatlanok. A legerősebb korreláció az év és a felhasználói értékelés, valamint az év és a vélemények száma között tapasztalható, azonban ezek is csak gyenge kapcsolatot mutatnak.

Adatelőkészítés

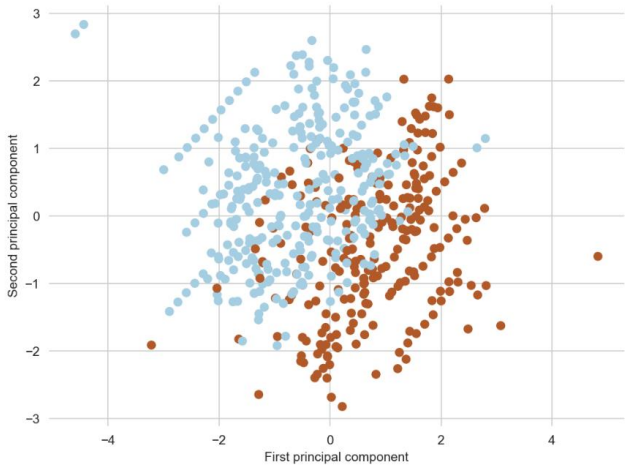
A tanulmány célja az adathalmaz fikciókra és nem fikciókra bontása. Az osztályozási feladat célváltozója a műfaj (Genre). Az osztályozás az adatkészlet többi attribútuma alapján történik.

Ahhoz, hogy a modellek kezelni tudják a nem numerikus változókat – így a könyvcím, szerző és műfaj változókat –, numerikussá kell alakítanunk őket. A tanulmány megvalósítása során a scikit-learn szoftver könyvtár preprocessing csomagjának LabelEncoder eszközét használjuk. Ezt követően a csomag StandardScaler eszközével standardizáljuk az adatokat. A standardizáció fontos előkövetelménye számos gépi tanulási metódusnak. Lényege, hogy az egyes változókat nulla átlagú, egy szórású normális eloszlású változókká skálázzuk annak érdekében, hogy egy-egy tulajdonság ne dominálhassa a gépi tanulási feladatot.

A skálázást követően vizualizáljuk az adathalmazt. Hét változóval dolgozunk, azaz hét dimenziós adatunk van, melyet két dimenzióban szeretnénk megjeleníteni. A vizualizáláshoz szükség van az adatkészlet két főkomponensére, melyek mentén ábrázolhatjuk az egyes adatrekordokat. A főkomponens analízist a scikit learn könyvtár decomposition csomagjának PCA, azaz Principal Component Analysis módszerének használatával valósítjuk meg.

A 6. ábrán látható, hogy a fikciók (barna) és a nem fikciók (kék) lineárisan nem szeparálható módon helyezkednek el a két főkomponens mentén.

6. ábra: Az adathalmaz vizualizálása a két főkomponens mentén. Forrás: a Szerző



Megfigyelhetjük, hogy a két csoport átfedően helyezkedik el, a fikciók a bal felső sarokhoz közel helyezkednek el, míg a non-fikciók a jobb alsó sarokhoz közel pozícionálódnak.

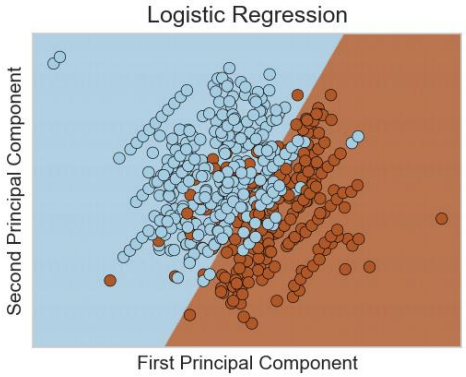
Osztályozás

Az adatelőkészítést követően az osztályozási feladat modellezését végezzük el. Az adathalmazt tanító és tesztadatokra osztjuk, melyek a scikit learn könyvtár model selection csomagjának train_test_split algoritmusával végzünk el. Az előfeldolgozott adathalmaz 70%-át tanító, 30%-át pedig tesztelő halmazba soroljuk.

Ezt követően definiálhatjuk az egyes modelleket, melyeket a tanító adattokra illesztünk, majd a betanított modelleket a tesztadatok segítségével teszteljük.

Elsőként *logisztikus regresszióval* dolgozunk. A modellt Python nyelven definiáljuk a scikit learn könyvtár linear_model csomagja segítségével. Az algoritmus statisztikai módszeren alapulva bináris adatok modellezésére jelez valószínűségeket (v.ö.: Bodon és Búza, 2014). Példánkban annak a valószínűségét számítja ki, hogy az ismérvek alapján az adott könyv fikció vagy sem. A 7. ábrán kék háttérrel láthatjuk a non-fikció kategóriába sorolt adatpontok halmazát, barnával pedig a fikció kategóriát. Az adatpontokat a valós kategóriának megfelelő színezéssel tüntettük fel. Több adatpont esetében is eltérést tapasztalunk a modell által jósolt és a valódi osztálycímke között. A modellezés során arra törekszünk, hogy ezen hibás osztályozások számát minimalizáljuk.

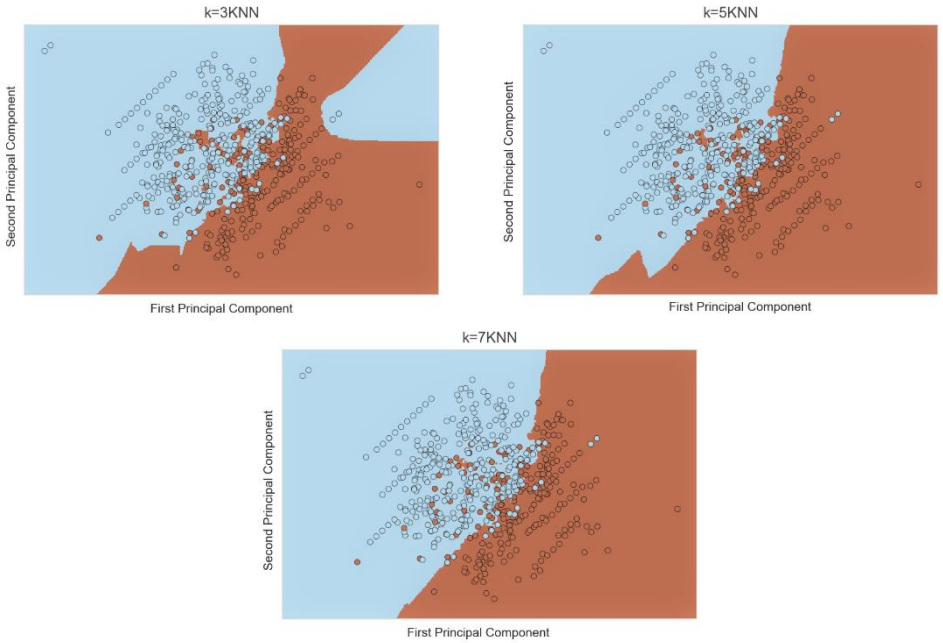
7. ábra: Osztályozás logisztikus regresszióval. Forrás: a Szerző



Második kísérletünkben a *k legközelebbi szomszéd*, azaz *kNN* osztályozót alkalmazzuk. A modell működése a következő. Adott az osztályozási feladat, az adatbázis, valamint egy függvény, ami számszerűsíti bármely két adatpont távolságát. Az algoritmus megkeresi a tanító adatbázisban az osztályozandó adatponthoz k darab legközelebbi adatpontot, azaz a k darab legközelebbi szomszédot, majd ezek osztálycímkéi közül kiválasztja a leggyakrabban előfordulót, és ehhez az osztályhoz sorolja az osztályozandó adatpontot (Kovács, 2013). Az optimális paraméter megtalálásához a könyök metódust (elbow method) használjuk. A 8. ábra értelmében több k paramétert is megvizsgáltunk. A tanulmány során k=3, k=5 és k=7 paraméterekkel dolgoztunk. A 9. ábra szemlélteti a különböző k paraméter hatásait az osztályozás során. Annak elbírálására, hogy a különböző paraméterekkel definiált modellek közül melyik teljesít legjobban, a dolgozat következő szakaszában kerül sor.

8. ábra: a könyök metódus. Forrás: a Szerző

9. ábra: Osztályozás a kNN medótussal k=3 (balra fent), k=5 (jobbra fent) és k=7 (lent) paraméter esetén. Forrás: a Szerző



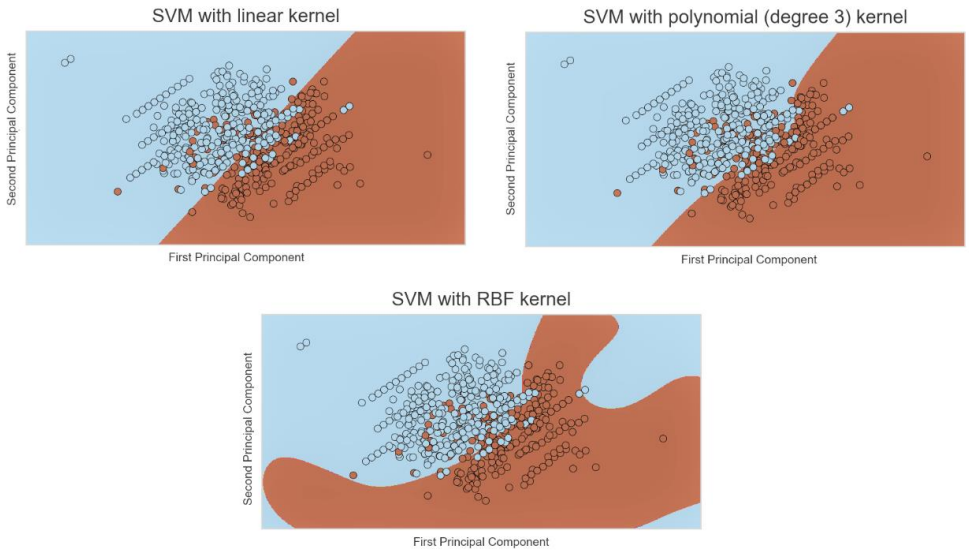
A következő vizsgálat során *Support Vector Machine*-t *(SVM)*, azaz tartóvektorgépeket illesztünk az osztályozási feladat megoldására. Az SVM alapvetően olyan hipersíkot keres, ami a lehető legjobban leválasztja a tanító adathalmaz különböző osztályokba tartozó pontjait (lásd: Tan, Steinbach és Kumar, 2011). A lineárisan nem szeparálható halmazok elválasztására különböző kernel függvényeket alkalmazhatunk a modellekben. A kísérletben több kernel függvény is összehasonlításra került (lásd: 10. ábra).

Az osztályozást *mesterséges neurális hálózat*tal folytatjuk. Mivel az adathalmaz nem lineárisan szeparálható, többrétegű perceptronnal dolgozunk. A hálózat bemeneti-, rejtett-, és kimeneti rétegből épül fel. A hálózat hiba-visszaáramoltatással minimalizálja az osztályozás hibáját (Fazekas, 2013).

A *Naive Bayes* osztályozó módszer feltételes valószínűségekre alapozva vizsgálja annak valószínűségét, hogy egy adatpont az egyes osztályokba tartozik, majd a maximális valószínűségű osztályba sorolja (v.ö.: Russell, Norvig, 2005).

Az utolsó vizsgálatunk során *döntési fát* és *véletlen erdőt* (Random Forest) alkalmazunk az osztályozási feladat elvégzésére. A döntési fa egy faszerkezet, melynek minden csúcsa egy értékre vonatkozó ellenőrzést, döntést jelöl. A csúcsokból kivezető élek a döntések lehetséges kimeneteleit reprezentálják. A döntési fa levél elemei határozzák meg az osztályba sorolást, (Tan, Steinbach, Kumar, 2011). A véletlen erdő több, különböző döntési fa által adott előrejelzésre alapozva végzi el az osztályozást.

10. ábra: Osztályozás SVM-mel lineáris (balra fent), ploinomiális (jobbra fent) és RBF (lent) kernellel. Forrás: a Szerző.



Az osztályozási módszerek pontosságának összehasonlítása

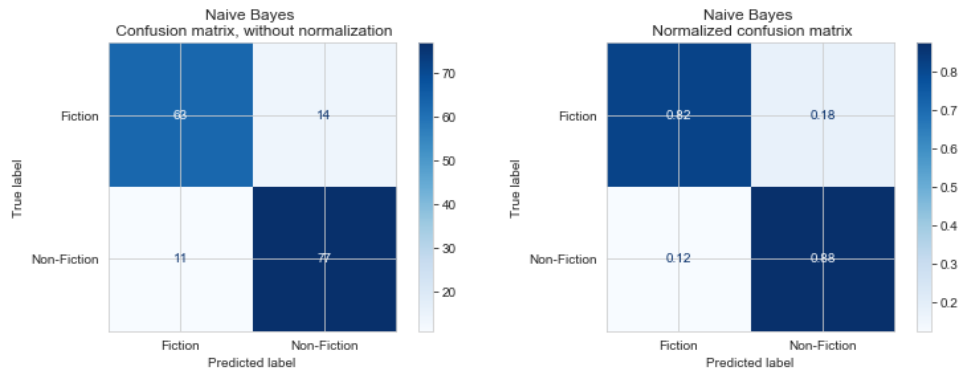
Az ismertetett algoritmusok által elvégzett osztályozások jellemzésére különböző metrikákat használhatunk, melyek közül a konfúziós mátrixot vizsgáljuk meg először. A mátrix az egyes osztályokba helyesen, illetve hibásan osztályozott egyedek számát reprezentálja. A 11. ábra segítségével összevethetjük a különböző osztályozási módszerek pontosságát. A kategória tengelyen a modellek által jósolt osztályozás, az értéktengelyen pedig a valódi osztályozás látható. A konfúziós mátrix bal felső eleme azon tesztadatok számát adja meg, melyeket a modell helyesen sorolt a fikciók közé. A jobb alsó elem azon tesztadatok számát mutatja, melyeket a modell helyesen osztályozott a non-fikció kategóriába.

A mátrix további két eleme a hibás osztályozások számát reprezentálja. A jobb felső elem azon osztályozások számát mutatja, melyeket a modell non-fikciónak osztályozott, viszont fikció volt. A bal alsó elem a modell által hibásan fikcióként osztályozott elemek számát adja meg.

A 11. ábrán a naiv Bayes osztályozó 63 esetben osztályozta helyesen a fikciókat, 77 esetben pedig a nem fikciókat. A modell 14 esetben osztályozta hibásan non-fikciónak a valójában fikciót, 11 esetben pedig non-fikció helyett fikció kategóriába sorolta a kiadványt.

A konfúziós mátrixban szerepeltetett adatok megadják a modell helyes, illetve hibás osztályozásainak számát. A mátrix alapján az osztályozást jellemezhetjük pontossággal, melyet az alábbiak szerint számolunk. Összeadjuk a modell helyes predikcióinak számát minden osztályra vonatkozóan, majd ezt az összeget elosztjuk az összes osztályozás számával.

11. ábra: Normalizálatlan (balra) és normalizált (jobbra) konfúziós mátrix a naiv Bayes osztályozó esetén. Forrás: a Szerző



A fent látható példa esetében a pontosság a következőképpen alakul. A tesztelhalmaz 165 adatot tartalmaz. A modell összesen 140 esetben osztályozott helyesen. A modell által végrehajtott osztályozás pontossága 84.84%.

A modellek teljesítményét nem célszerű egyetlen kísérlet alapján megítélni. Az osztályozási feladat során, a tanító- és tesztelő adatok meghatározásánál fontos szerepe van a véletlen faktornak. Hogy minél relevánsabb megállapítást tehessünk az osztályozó módszerek pontosságának értékelésénél, az osztályozási feladatokat a kísérlet során ezerszer hajtottuk végre.

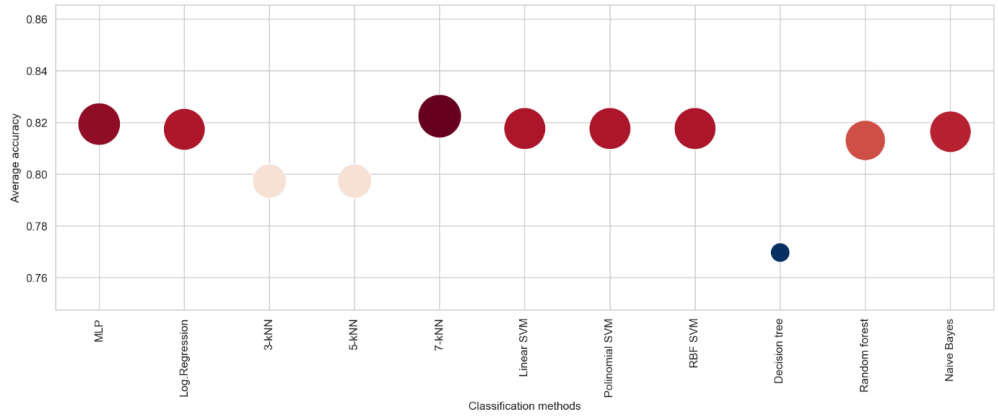
Eredmények összefoglalás

A 12. ábra a tanulmány során végrehajtott ezer kísérlet során tapasztalt átlagos pontosságot mutatja osztályozó eljárásonként. Megállapíthatjuk, hogy az általunk ismertetett gépi tanulási módszerek átlagosan 81% pontossággal képesek meghatározni az Amazon webáruház legnagyobb darabszámban eladott könyveit tartalmazó adatbázisban szereplő kiadványokról, hogy azok fikció vagy non-fikció kategóriába tartoznak-e.

Az ismertetett osztályozási módszerek közül a döntési fa szignifikánsan pontatlanabbul volt képes osztályozni az adatbázist, mint a többi eljárás. Továbbá a k=3 és k=5 paraméterekkel definiált kNN modellek teljesítménye is gyengébb a többi módszerhez képest.

A véletlen erdő pontossága 81.3%. A logisztikus regresszió, az SVM osztályozók és a Naive Bayes hálózat átlagos pontossága 81.7%. A mesterséges neurális hálózat 81.9% pontosságot mutatott. A legnagyobb átlagos pontosságot a k=7 paraméterrel definiált k legközelebbi szomszéd modell esetén tapasztaltuk, mely 82.2%-os pontosságot mutatott.

12. ábra: Az egyes osztályozási módszerek átlagos pontossága 1000 kísérlet alapján. Forrás: a Szerző



Javaslatként megfogalmazható a döntési fa használatának elkerülése, valamint a k=7 paraméterű legközelebbi szomszéd algoritmus használatának preferálása az adott adathalmaz osztályozására.

A tanulmány jövőbeli folytatásaként az osztályozási módszerek összehasonlító elemzésének további metrikák figyelembevételével való kiegészítését tervezzük megvalósítani.

Irodalom

Bodon Ferenc és Buza Krisztián (2014): *Adatbányászat*. Letöltés: 2021.10.21. Web: http://www.cs.bme.hu/nagyadat/bodon.pdf

Fazekas István (2013): *Neurális hálózatok*. Debreceni Egyetem. Letöltés: 2021.10.21. Web: https://gyires.inf.unideb.hu/GyBITT/19/index.html

kaggle.com: *Amazon Top 50 Bestselling Books 2009 – 2019*. Letöltés: 2020.10.13. Web: https://www.kaggle.com/sootersaalu/amazon-top-50-bestselling-books-2009-2019

Kovács György (2013): *Párhuzamos programozási eszközök és összetett alkalmazásaik*. Typotex Kiadó, Budapest

Tan, Pang-Ning; Steinbach, Michael és Kumar, Vipin (2011): *Bevezetés az adatbányászatba*. Panem Kft., Budapest. Letöltés: 2021.10.10. Web: https://gyires.inf.unideb.hu/KMITT/a04/

Russell, Stuart és Norvig, Peter (2005): *Mesterséges intelligencia*. Panem Kft., Budapest