**A Titanic katasztrófa túlélőinek becslése döntési fa alapú gépi tanuló eljárással**

**Estimation of survivors of the Titanic catastrophe by decision tree-based machine learning algorithm**

Barta Gergő, Pitlik László - MYX-team

**Kivonat:**

Vélelmezhetően nem sok olyan ember létezik, aki a híres RMS[[1]](#footnote-1) Titanic óceánjáróról, és annak tragikus katasztrófájáról ne hallott volna. A cikk céljául tűzte ki, hogy döntési fa alapú gépi tanuló módszer és lineáris regressziók felhasználásával mintákat keres a rendelkezésre álló utasadatokban, és előrejelzést kíván adni arra, vajon mely utasok él(het)ték túl a szerencsétlenséget.

**Kulcsszavak:**

Titanic, Gépi Tanulás, Döntési Fa, CHAID módszer

**Abtract:**

Most probably there are not many people who would have never heard of the famouse RMS Titanic ocean cruiser and its tragical catastrophe. The purpose of the article is to adopt decision tree-based machine learning algorithm and linear regressions to find patterns in the available passenger data and provide a forecast of which passengers might have been able to survive the misfortune.

**Keywords:**

Titanic, Machine Learning, Decision Tree, CHAID method

1. Bevezetés

Az RMS Titanic utasszálító óceánjárót korának legnagyobbikának tartották, melyről kinyílvánították, hogy képtelenség elsüllyeszteni. A hajó 1912. április 14-én jéghegynek ütközött, és alig 2 és fél óra elsüllyedt. Összesen a hajó 2200 főt szállított, melyből kb. 1500-an életüket vesztették a katasztrófa során. (rd.hu, é. n.) A cikk témája CHAID bináris-fa alapú algoritmus implementálása, mely a rendelkezésre álló utasadatok alapján kívánja megbecsülni a túlélőket, illetve arra adhat választ, hogy mely ismérvek járultak a legjobban ahhoz, hogy valaki túlélhette-e a katasztrófát vagy, sem. A feladatot a kaggle.com hirdette meg 2016-ben, és nyilvánosan elérhetővé tette a számítások alapját képező adathalmazt, mely összesen 1309 utasinformációt tartalmaz, amelyből 891 utas esetén számunkra ismert a túlélési információt tartalmazó bináris változó, a maradék esetén, pedig becslést adhatunk, melyet a kaggle.com rendszere kiértékel, és a találatról értesítést küld. A dokumentumban feldolgozott fájlok az 5. mellékletben érhetőek el. A kaggle.com bevonása egy modellezési feladatba garantálja a modell-minőség objektív feltárását.

1. Feldolgozott adathalmaz

Az adathalmaz a kaggle.com oldalán nyilvánosan elérhető[[2]](#footnote-2). Az alábbi attribútumokat tartalmazza minden az utasok esetében a 891 adatra:

* PassengerID: Utasok egyedi azonosítója. Típus: numerikus (mértékegység nélküli szám).
* Survived: Státuszváltozó, hogy adott utas túlélte-e a katasztrófát. Típus: bináris mértékegység nélküli szám (0 – nem élte túl, 1 – túlélte).
* Pclass: Státuszváltozó, hogy adott utas mely osztályon utazott az óceánjárón. Típus: numerikus (mértékegység nélküli száma: 1 – első osztály, 2 – másod osztály, 3 – harmad osztály).
* Name: Utas neve. Típus: szöveges.
* Sex: Utas neme. Típus: szöveges (male – férfi, female – nő).
* Age: Utas kora. Típus: numerikus (mértékegység = év).
* SibSp: Rokonok száma I., vagyis adott utasnak hány rokona (testvér, mostohatestvér és házastárs) utazott az óceánjárón. Típus: numerikus (mértékegység = fő).
* Parch: Rokonok száma II., vagyis adott utasnak hány rokona (szülő és gyermek) utazott az óceánjárón. Típus: numerikus (mértékegység = fő).
* Ticket: Jegyazonosító. Típus: szöveges.
* Fare: Jegy ára. Típus: numerikus (mértékegység = dollár).
* Cabin: Kabin sorszáma. Típus: szöveges.

Embarked: Felszállás helye. Típus szöveges (C = Cherbourg, Q = Queenstown, S = Southampton)

Hiányzó adatok

Az adathalmazban a következő hiányzó értékek szerepelnek, mely rekordok a továbbiakban kiszűrésre kerültek a teljes adattáblából, így összesen a szűkített adatbázis 714 rekordot tartalmaz:

* Age: 177 utas esetén hiányzik
* Cabin: 687 utas esetén hiányzik
* Embarked: 2 utas esetén hiányzik

Vélelmek és összefüggések

* Pclass (Osztály): vélhetően minél magasabb osztályon utazott az adott utas, annál nagyobb esélyel élte túl a szerencsétlenséget a „nemesi” származás előnyeivel élve. A független változó és célváltozó kategórikus változók, ezért a kettő közötti összefüggést célszerű kereszttábla elemzéssel kimutatni. A Pearson-féle Khi-négyzet értéke 92.9, 0.000-nál kissebb kétoldali szignifikanciaszinten, mely alapján azt feltételezzük, hogy az Osztály és Túlélő változó között összefüggés mutatkozik. A felhasznált szimmetrikus mutatók az összefüggés erősségének mérésére (Phi, Cramer V, kontingencia-együttható) értékei sorra 0.361, 0.361 és 0.339, mely azt jelenti, hogy mind a három mutató alapján a két változó között gyenge összefüggés található. Az 1. táblázat szemlélteti a két változóból képzett mátrixot, mely ránézésre is megerősíti a várakozásokat. Az 1. osztályon utazott utasok 65.6%-a, a 2. osztályon utazók 48%-a, a 3. osztályon utazók mindössze 23.9%-a élte túl a katasztrófát, tehát minél előkelőbb osztályon utazott az illető, annál nagyobb esélye volt a túlélésre.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Survived (Túlélte) és Pclass (Osztály) változók mátrixa | | | Survived (Túlélte) | | Összesen |
| 0 (nem) | 1 (igen) |
| Pclass (Osztály) | 1 | Összeg | 64 | 122 | 186 |
| Várt összeg | 110.5 | 75.5 | 186.0 |
| % Pclass változón belül | 34.4% | 65.6% | 100.0% |
| % Survived változón belül | 15.1% | 42.1% | 26.1% |
| % összesen | 9.0% | 17.1% | 26.1% |
| 2 | Összeg | 90 | 83 | 173 |
| Várt összeg | 102.7 | 70.3 | 173.0 |
| % Pclass változón belül | 52.0% | 48.0% | 100.0% |
| % Survived változón belül | 21.2% | 28.6% | 24.2% |
| % összesen | 12.6% | 11.6% | 24.2% |
| 3 | Összeg | 270 | 85 | 355 |
| Várt összeg | 210.8 | 144.2 | 355.0 |
| % Pclass változón belül | 76.1% | 23.9% | 100.0% |
| % Survived változón belül | 63.7% | 29.3% | 49.7% |
| % összesen | 37.8% | 11.9% | 49.7% |
| Összesen | Összeg | | 424 | 290 | 714 |
| Várt összeg | | 424.0 | 290.0 | 714.0 |
| % Pclass változón belül | | 59.4% | 40.6% | 100.0% |
| % Survived változón belül | | 100.0% | 100.0% | 100.0% |
| % összesen | | 59.4% | 40.6% | 100.0% |

**1. ábra: Pclass (Osztály) és Survived (Túlélte) változók összefüggését mutató mátrix**

**Forrás:** saját szerkesztés

* Name (Utas neve): a változó önmagában nem, azonban a nevet alkotó karakterek száma alkalmas lehet az elemzésre, minél hosszabb az utas neve, annál inkább „nemesi” származású (vö. keresztnevek száma), tehát annál inkább esélyes a mentőcsónakba jutásra: ennek megfizetése, és/vagy az általános társadalmi hierarchia erőterei alapján, ezért LenName váltózónévvel a név hosszúságára új változó került létrehozásra az adathalmazban. A független változó folytonos, míg a célváltozó kategórikus, ezért ésszerű Bináris logisztikus regresszióval az összefüggések megvizsgálása. Az eljárás során kiszámításra került a Wald-statisztika, mely szignifikáns volta jelzi, hogy az utas nevének hossza, hatással van-e a célváltozóra vagy sem. A Wald-statisztika értéke esetünkben 58.117, 0.000-nál kisebb szignifikancia szinten, így az eljárás alapján alapján azt feltételezhetjük, az utas nevének hossza hatással van arra, hogy az utas túlélte-e a szerencsétlenséget vagy nem.[[3]](#footnote-3)
* Sex (Utas neme): Feltételezésünk, hogy a nők túlélési esélyei magasabbak voltak a férfiakénál, az az általános elgondolás miatt, hogy egy szülőképes nő nagyobb értéket jelent egy közösség számára, ezért elsődlegesen a nők mentése élvezett prioritást az utazásra jelentkezők átlagos kulturális kötődéseinek földrajzi erőterei alapján. Változóink kategorikus változók, ezért a kereszttábla elemzés módszert biztosít az összefüggések kimutatására. A Pearson-féle Khi-négyzet értéke 207.3, 0.000-nál kissebb kétoldali szignifikanciaszinten, mely alapján azt feltételezzük, hogy a Nem és Túlélő változó között összefüggés mutatkozik. A felhasznált szimmetrikus mutatók az összefüggés erősségének mérésére (Phi, Cramer V, kontingencia-együttható) értékei sorra -0.539, 0.539 és 0.3474, mely azt jelenti, hogy mind a három mutató alapján a két változó között közepes összefüggés található. A 2. táblázat szemlélteti a két változóból képzett mátrixot, mely alapján beláthatjuk, hogy a nők nagyobb eséllyel élték túl a katasztrófát. Míg csupán a férfi utasok 20.5%-a, a nők 75.5% túlélő volt.

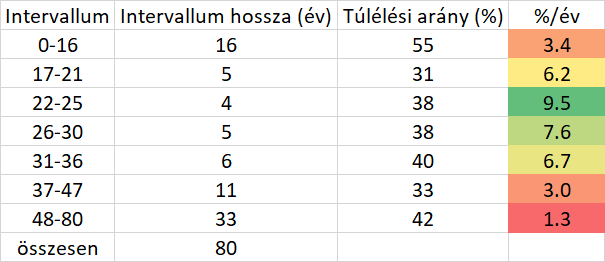
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Survived (Túlélte) és Sex (Nem) változók mátrixa | | | Survived (Túlélte) | | Összesen |
| 0 (nem) | 1 (igen) |
| Sex (Nem) | 0 (Nő) | Összeg | 64 | 197 | 261 |
| Várt összeg | 155.0 | 106.0 | 261.0 |
| % Sex változón belül | 24.5% | 75.5% | 100.0% |
| % Survived változón belül | 15.1% | 67.9% | 36.6% |
| % összesen | 9.0% | 27.6% | 36.6% |
| 1 (Férfi) | Összeg | 360 | 93 | 453 |
| Várt összeg | 269.0 | 184.0 | 453.0 |
| % Sex változón belül | 79.5% | 20.5% | 100.0% |
| % Survived változón belül | 84.9% | 32.1% | 63.4% |
| % összesen | 50.4% | 13.0% | 63.4% |
| Összesen | | Összeg | 424 | 290 | 714 |
| Várt összeg | 424.0 | 290.0 | 714.0 |
| % Sex változón belül | 59.4% | 40.6% | 100.0% |
| % Survived változón belül | 100.0% | 100.0% | 100.0% |
| % összesen | 59.4% | 40.6% | 100.0% |

**2. ábra: Sex (Nem) és Survived (Túlélte) változók összefüggését mutató mátrix**

**Forrás:** saját szerkesztés

* Age (Utas kora): az utas korával kapcsolatban több feltételezés is logikusnak tűnhet.
  + a relatív túl fiatal és a relatív túl idős személyek túlélési esélye fiziológiailag kisebb egy potenciális „tülekedés”, ill. a jeges víz és a hideg levegő hatásainak, az extrém pszichés stressznek tűrése esetén.
  + az általánosan vélelmezhető kulturális kötődés miatt a kor minél kisebb, annál nagyobb a túlelés esélye, mert a gyermekeket az idősebbek (kifejezetten pl. a szüleik) a saját életük árán is menteni akarták (s vélelmezhetően ritkán utazik gyermek szülő nélkül, ill. egy szülőre általában több gyermek is juthat a születésszámok alapján, valamint ha a felnőtt kísérő nem szülő, az erkölcsi kötelesség a gyermek mentését illetően tulajdonképpen nem kisebb, mint a szülő esetén).

Az Age változó folytonos, míg a célváltozó kategorikus, ezért Bináris logisztikus regresszió alkalmazható. Hasonlóan az „utas nevének hossza” változónál megvizsgálásra került a Wald-statisztika, melynek értéke 4.231, 0.04-es szignifikancia szint mellett, így látszólag a kor változónak hatása van a túlélésre[[4]](#footnote-4). A 3. ábra szemlélteti az utasok korából egyenlő eloszlásban képzett kategóriák alapján a túlélési statisztikákat. A táblázatból jól kivehető, hogy a legnagyobb túlélési aránnyal a 0 – 16 éves korosztály rendelkezett, míg kis különbséggel, de a 48 – 80 év közöttiek a második legnagyobb valoszínűséggel élték túl a szerencsétlenséget, melynek értéke 42%. Ez a kiindulási vélelmekhez képest egy fordított optimum-hatást enged értelmezni, mely azonban az intervallumhatárok aránytalansága miatt adódik. Ha az egy intervallumévre jutó túlélési számokat nézzük, akkor az elvárt optimum rajzolódik ki, vagyis a középkorúak túlélési esélye fizikailag, fiziológiailag, pszichológiailag egyaránt nagyobb:



**3. ábra: Age (Utas kora) és Survived (Túlélte) változók összefüggését mutató diagram**

**Forrás:** saját szerkesztés

* SibSp (Rokonok száma I): A rokonok száma pozitívan hathat a túlélési esélyekre a család segítő- és összetartó erejét feltételezve. Logisztikus regresszióval megvizsgálva a változó hatását a túlélésre a Wald-statisztika értéke 0.215, 0.643-as szignifikancia szinten, mely azt sugallja, azonban, hogy a rokonok száma nincs hatással a célváltozóra.
* Parch (Rokonok száma II): Hasonlóan az előző változóhoz pozitív összefüggést feltételezhetünk ezúttal is. A különbség a két változó között, hogy a Parch oszlop a szülők és gyerekek számát tartja számon a testvérek helyett. Logisztikus regresszióval megvizsgálva a változó hatását a túlélésre a Wald-statisztika értéke 6.015, 0.014-es szignifikancia szinten, mely azt jeletni, hogy az előző változóval ellentétben a rokonok száma (gyerekek és szülők) hatással van a célváltozóra (vö. kor és túlélés kapcsolata). A 4. ábra szemlélteti a kimutatást a két változóval kapcsolatban. A legnagyobb túlélési arány 1 és 2 szülő és gyerekkel való utazás esetén a legmagasabb, illetve 3 rokon esetén is, azonban túl kevés ilyen eset szerepel az adathalmazban a mélyebb konklúzió levonása érdekében.

|  |  |  |  |
| --- | --- | --- | --- |
| Parch (Rokonok II) | Survived (Túlélte) | Összesen | Túlélési arány |
| 0 | 186 | 521 | 35.70% |
| 1 | 61 | 110 | 55.45% |
| 2 | 39 | 68 | 57.35% |
| 3 | 3 | 5 | 60.00% |
| 4 | 0 | 4 | 0.00% |
| 5 | 1 | 5 | 20.00% |
| 6 | 0 | 1 | 0.00% |

**4. ábra: Parch (Rokonok II) és Survived (Túlélte) változók kimutatása**

**Forrás:** saját szerkesztés

* Ticket (Jegyazonosító): A jegyazonosító egy egyedi kulcs, mely jelölheti az osztályt, illetve a kabin elhelyezkedését is, mely releváns lehet a mentőcsónakok közelsége miatt. Mivel, azonban nincs további adat arra vonatkozóan, hogy adott jegy hova mutat így kizárására került az elemzésből. Amennyiben a jegyazonosító nem egy sorfolytonos szám, akkor egy fajta titkosítási/desifrírozási feladatként lenne értelmezhető a jegysorszámok esetleges beszédes kódként való értelmezéséig eljutni tudni.
* Fare (Jegy ára): A jegy ára a magasabban osztályban való utazást és magasabb státuszt jelenthet, így feltételeztük a korrelációt a jegy ára és a túlélési esélyek között. Logisztikus regresszióval számolva, a Wald-statisztika értéke 40.886, 0.000-nál kisebb szignifikancia szinten, így egyértelműen kimutatható a változó hatása a célváltozóra. Azokban az esetekben, amikor adott utas nem élte túl a szerencsétlenséget, az átlagos jegyár közel 23 dollár volt, míg a túlélők esetében ez kb. 52 dollár, ahol a jegyárak maximuma 512 dollár volt. Így kezdeti feltételezésünk beigazolódni látszik.
* Cabin (Kabin sorszáma): A kabin sorszáma változó egy egyedi kulcs, mely jelölheti a jegyazonosítóhóz hasonlóan a kabin elhelyezkedését. Mivel a változó 687 esetben hiányzik, így nem tartottuk érdemesnek bevenni az elemzésbe. A kabinok sorszáma kapcsán a fedélzetek (emeletek) számára való utalás, mint belső beszédességi tényező vélelmezhető. Emellett egy kabinban vélelmezhetően többen is el voltak szállásolva, így az egy kabinhoz tartozás is információértékkel bírhat még az elemzés sikerességének növelése érdekében.
* Embarked (Felszállás helye): Az oszlop további három változóra lett bontva „one-hot-encoding” eljárással, ahol adott utas adott felszállási helye 1 értéket kapott, a többi pedig 0-t. Mivel a változók kategorikus változók, így kézenfekvő kereszttábla elemzéssel megvizsgálni az összefüggés lehetőségét, melynek összesített eredményét az alábbi táblázat mutatja:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Embarked (Felszállás helye) | Khi-négyzet | Szignifikancia szint | Phi | Cramer V | Kontingencia-együttható | Konklúzió |
| Southamptoon | 19.259 | 0.000 < | -0.164 | 0.164 | 0.162 | Nagyon gyenge  összefüggés |
| Cherbourg | 26.763 | 0.000 < | 0.194 | 0.194 | 0.19 | Nagyon gyenge  összefüggés |
| Queenstown | 1.753 | 0.186 | -0.05 | 0.05 | 0.049 | Összefügés hiánya |

**5. ábra: Embarked (Felszállás helye) és Survived (Túlélte) változók összefüggés-vizsgálata**

**Forrás:** saját szerkesztés

A kezdeti összefüggés-vizsgálatok után az alábbi változók bizonyultak érdemesnek a modell építésbe, melyek a végrehajtott statisztikai számítások alapján hatással vannak a célváltozóra:

* Pclass (Osztály)
* LenName (Utas nevének hosza)
* Sex (Utas neme)
* Age (Utas kora)
* Parch (Rokonok száma II)
* Fare (Jegy ára)
* Embarked\_Southampton (Felszállás helye: Southampton)
* Embarked\_Cherbourg (Felszállás helye: Cherbourg)

1. Módszertani áttekintő

A becslés a CHAID (Chi-Squared Automatic Interaction Detector) automatikus klasszifikáló eljárással valósult meg, mely alkalmas folytonos és kategórikus független változók alkalmazására is. A független változóinkat a függő változó szempontjából csoportosítjuk azon matematikai gondolatmenetem mentén, hogy az egyes csoportokon belüli variancia minél kisebb, míg a csoportok közötti variancia minél nagyobb legyen. (Hámori, 2001) Az eredménytermék egy döntési fa lesz, mely alkalmazható egy betanított gépnek is, mely új minták esetén képes automatikusan kategorizálni egy utas adatai alapján, hogy az túlélte-e a katasztrófát vagy sem?

1. Megoldás
2. lépés: Az adatok transzformálása és szűrése

Az adathalmaz összesen 5 szöveges mezőt tartalmaz, mely nem alkalmas matematikai számítások elvégzésére. Az előzetes elemzések alapján az adatok transzformálásának és szűrésének összefoglalója az alábbi:

* Name (Utas neve): a nevet alkotó karakterek száma alkalmas lehet az elemzésre: minél hosszabb a neve, annál inkább nemesi származású, tehát annál inkább esélyes a mentőcsónakba jutásra: ennek megfizetése, vagy az általános társadalmi hierarchia erőterei alapján, ezért LenName váltózónévvel a név hosszúságára új változó került létrehozása az adathalmazban.
* Ticket (Jegyazonosító): az oszlop teljes kiszűrésre került az adathalmazból, mivel vélhetően nem járul hozzá a célváltozó pontosabb becsléséhez.
* Cabin[[5]](#footnote-5) (Kabin sorszáma): az oszlop teljes kiszűrésre került az adathalmazból, mivel vélhetően nem járul hozzá a célváltozó pontosabb becsléséhez.
* Sex (Utas neve): bináris kódolást kapott, ahol 1-el került feltüntetésre a férfiak, 0-val, pedig a nők.
* Embarked (Felszállás helye): az oszlop további három változóra lett bontva „one-hot-encoding” eljárással, ahol adott utas adott felszállási helye 1 értéket kapott, a többi pedig 0-t.

1. lépés: Tulajdonságok kiválasztása

Ezek után Anaconda környezetben Python 3.6 programozási nyelv segítségével, SckitLearn könyvtár RandomForest algoritmusával véletlen fákat generáltunk, mely összesítve képes egy statisztikát adni arra vonatkozóan, vajon mely attribútumok és hogyan járulnak hozzá a legjobban a célváltozóhoz (variancia módszer), melynek eredményét a következő ábra szemlélteti. A módszer alkalmazására az előzetes felmérések visszaellenőrzése miatt került sor. A RandomForest az összes attribútumot úgy súlyozza, hogy azok összege 1-el legyen egyenlő. A megírt programkód megtalálható a dokumentum 4. számú mellékletében.

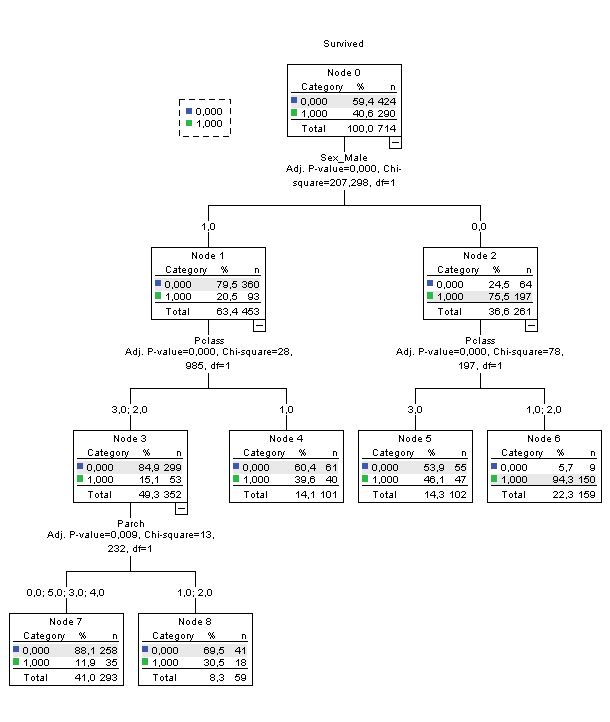
**6. ábra: Attribútum súlyok RandomForest alapján**

**Forrás:** saját szerkesztés

Az 6. ábra azt szemlélteti, hogy sorra a Sex (Utas neme), Age (Utas kora),, a Fare (Jegy ára) és a LenName (Utas nevének hossza) közel hasonló módon járul hozzá a célváltozó magyarázatához, sorra 0.2257, 0.2080, 0.1949 és 0.1769 súllyal. A maradék változó hozzáadott értéke elhanyagolható ezekhez az értékekhez képest. Összevetve az előzetes statisztikákkal, ez azt jelenti, hogy a Pclass (Osztály), Parch (Rokonok II.) és Embarked (Felszállás helye) változók csak kis mértékben járulnak hozzá a célváltozó magyarázatához. Érdemes a két különböző eredmény miatt két eltérő változószámú modell felépítése és azok eredményeinek összehasonlítása.

1. lépés: CHAID algoritmus futtatása SPSS-ben

Az algoritmus futtatása a kezdeti statisztikák eredményei alapján, az abban kiválasztott változó esetén a következő fát generálta automatikusan (Modell I). A modellezés minden rendelkezésre álló adatot feldolgozott tanulási mintaként. Éles tesztelésre a feladat kibocsátójával való későbbi kapcsolatfelvétel keretében kerül majd sor, így a cikk minden találati aránya tanulási találati arány.



**7. ábra: Túlélők becslése (Modell I)**

**Forrás:** saját szerkesztés

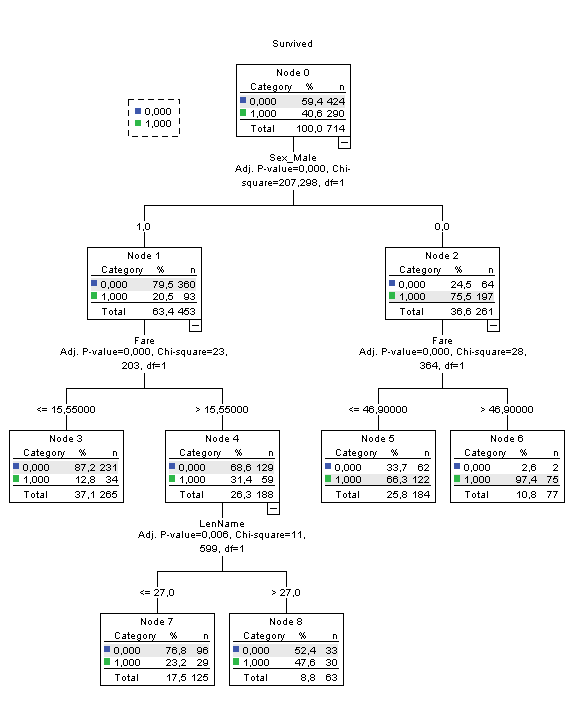
|  |  |  |  |
| --- | --- | --- | --- |
| Megfigyelt | Becsült | | |
| 0 (elhunyt) | 1 (túlélte) | Eltalált % |
| 0 (elhunyt) | 415 | 9 | 97.9% |
| 1 (túlélte) | 140 | 150 | 51.7% |
| Összesen % | 77.7% | 22.3% | 79.1% |

**8. ábra: Túlélők becslése (Modell I) – A modell eredményei, ahol a minimálisan realizálható találati arány egyetlen túlélőt sem feltételezve: 415/714 = 58%**

**Forrás:** saját szerkesztés

Az első modell összességében 79.1%-os találati eredményt produkált, melyben 97.9%-os pontossággal megbecsülhető azok aránya, akik nem élték túl a szerencsétlenséget, míg 51.7% pontossággal becsülhetők a túlélők. A modell első döntési pontja a nem (Sex) meghatározása, ahol a férfiak esetén a túlélők aránya 20.5%. Ezen az ágon a második döntési pont az utas osztálya (Pclass), mely 1. osztályra és 2.-3. osztályra bontja továbbá a döntési fát, melyet a Rokonok száma II. (Parch) változó követ két elágazással (1. elágazás: gyerekek és szülő száma 0, 3, 4, 5, 2. elágazás: gyerekek és szülők száma 1, 2). A másik ágon (női utasok) a szétbontás kizárólag az osztály szerint történt meg, az 1.-2. ostályon utazók és 3. osztályon utazók szétválasztásával. Az ábrából az is kiderül, hogy miután férfi utasról lévén szó, felesleges továbbmenni a döntési fán, mivel az összes eredmény 0-ra vezet (azaz nem élte túl a szerencsétlenséget), míg nők esetében az 1. és 2. osztályon utazók 94.3%-a túlélő, míg a harmadik osztályon utazók esetén a magasabb arány miatt a becslés a nem túlélést javasolja. A modell a felvitt változókból kizárólag 3-at alkalmazott: Nem (Sex), Osztály (Pclass) és Rokonok száma II. (Parch), míg a többi változót figyelmen kívül hagyta az alacsony célváltozóra gyakorolt hatás miatt.

A RandomForest által generált összefüggések alapján elkészített modell (Modell II) döntési fáját és becslési pontosságát a 9. és 10. ábrák szemléltetik. Összességében elmondható, hogy a modell hasonlóan produkált, összességében 78%-os pontosságal képes volt a helyes kategorizálásra a Nem (Sex), Jegyár (Fare) és a Névhossz (LenName) változók bevonásával és a kor figyelmenkívül hagyásával. Hasonlóan az első modellhez, a férfiak esetén a döntés ismételten 0 (azaz nem élte túl a katasztrófát, a nők esetében azonban a fa minden levele a túlélés választása mellett dönt.



**9. ábra: Túlélők becslése (Modell II)**

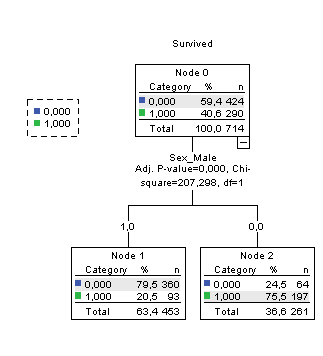
**Forrás:** saját szerkesztés

|  |  |  |  |
| --- | --- | --- | --- |
| Megfigyelt | Becsült | | |
| 0 (elhunyt) | 1 (túlélte) | Eltalált % |
| 0 (elhunyt) | 360 | 64 | 84.9% |
| 1 (túlélte) | 93 | 197 | 67.9% |
| Összesen % | 63.4% | 36.6% | 78.0% |

**10. ábra: Túlélők becslése (Modell II) – A modell eredményei**

**Forrás:** saját szerkesztés

A 11. ábra alapján: Amennyiben az utas férfi volt, abban az esetben a legjobb döntés a „nem túlélés” választása (79.5%), nők esetében, pedig a túlélésé (75.5%), ami megfelel a vélelmek/korrelációk fejezetben leírt irányultságnak: vagyis a nők mentése kulturális érték a férfiak neveltetése alapján. Ez a klasszifikáció összességében 78.0%-ban volt képes megfelelő döntést hozni a tanulás során, 84.9%-ban eltalálta azokat az utasok, akik nem élték túl a szerencsétlenséget, 67.9%-ban, pedig eltalálalta a túlélő utasokat, azaz egy változóval képes pontosan ugyanolyan becslést hozni, mint a Modell II esetén 3-mal, mely félrevezető lehet, ugyanis kétváltozó esetén a levelek becslési aránya összességében jobb eredményt produkált.



**11. ábra: Túlélők becslése (Modell III)**

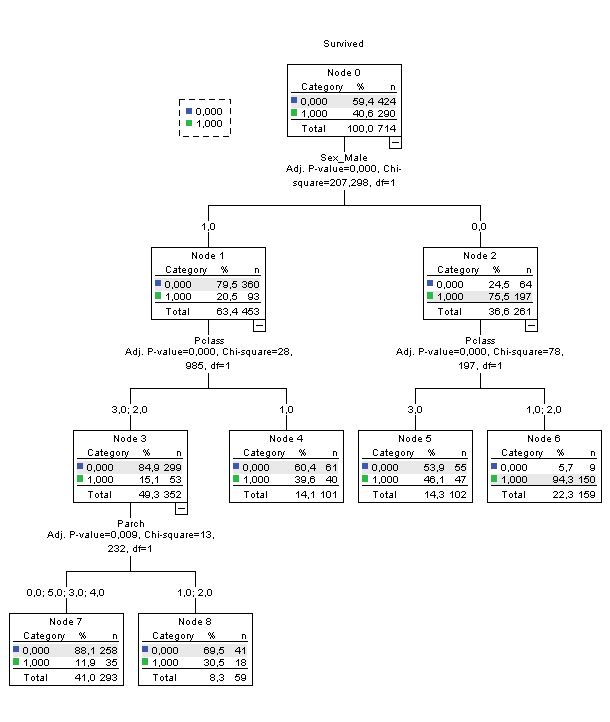
**Forrás:** saját szerkesztés

|  |  |  |  |
| --- | --- | --- | --- |
| Megfigyelt | Becsült | | |
| 0 (elhunyt) | 1 (túlélte) | Eltalált % |
| 0 (elhunyt) | 360 | 64 | 84.9% |
| 1 (túlélte) | 93 | 197 | 67.9% |
| Összesen % | 63.4% | 36.6% | 78.0% |

**12. ábra: Túlélők becslése (Modell III) – A modell eredményei**

**Forrás:** saját szerkesztés

A következő (13.) ábra szemlélteti azt az esetet, ahol nem hagyatkoztunk a kezdeti statisztikák és RandomForest eredményeire, és beválogattuk az összes változót a modellépítésbe (Modell IV).



**13. ábra: Túlélők becslése (Modell IV)**

**Forrás:** saját szerkesztés

A 14. ábra a 8., 10., és 12. ábra logikáját követve a IV. modell találati arányait mutatja be a modellek összehasonlítását támogatandó.

|  |  |  |  |
| --- | --- | --- | --- |
| Megfigyelt | Becsült | | |
| 0 (elhunyt) | 1 (túlélte) | Eltalált % |
| 0 (elhunyt) | 415 | 9 | 97.9% |
| 1 (túlélte) | 140 | 150 | 51.7% |
| Összesen % | 77.7% | 22.3% | 79.1% |

**14. ábra: Túlélők becslése (Modell IV) – A modell eredményei**

**Forrás:** saját szerkesztés

A 14. ábra szerint: Az eredmény hasonló az első modellhez, az algoritmus, összesen 3 változót vett be az elemzésbe, és összesen 79.1% pontosságú becslést volt képes produkálni, melyből 97.9%-ban a „nem túlélést” és 51.7%-ban a „túlélést” eltalálta a tanulás során.

1. A modellek összehasonlítása

Az eddig bemutatott értékelő táblázatok alapján a modellek összehasonlítására minimum 4 változó kínálkozik automatikusan:

* [I.] az összes találati arány (minél nagyobb, annál jobb a modell)
* [II.] az elhunytakra vonatkozó találati arány (minél nagyobb, annál jobb)
* [III.] a túlélőkre vonatkozó találati arány (minél nagyobb, annál jobb)
* [IV.] a túlélőkre és az elhunytakra vonatkozó találati arányok különbségének abszolút értéke (minél kisebb, annál jobb)

További értékelési tényező lehet az Occam-borotvája elv[[6]](#footnote-6) alapján a modellek egyszerűségének leírására alkalmas jelenségek köre – például:

* [V.] felhasznált változók száma (minél kisebb, annál jobb)
* [VI.] kialakított szabályok száma (minél kisebb, annál jobb)
* [VII.] szabályrendszer mélysége (minél kisebb, annál jobb)

A fenti logikát követve a modellek értékelésére tehát itt és most 7 mutatószám került kiválasztásra. A modellek száma 4. Így az OAM egy 4\*7-es mátrix:

**14. ábra: Modell-értékelő táblázat (erős zöld: oszlopnyertes, halványzöld: oszlop második)**

**Forrás:** saját szerkesztés

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Modell / Szempontrendszer | I. | II. | III. | IV. | V. | VI. | VII. |
| Modell I. | 565 | 415 | 150 | 265 | 3 | 4 | 3 |
| Modell II. | 557 | 360 | 197 | 163 | 3 | 4 | 3 |
| Modell III. | 557 | 360 | 197 | 163 | 1 | 1 | 1 |
| Modell IV. | 565 | 415 | 150 | 265 | 3 | 4 | 3 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Modell / Szempontrendszer | I. | II. | III. | IV. | V. | VI. | VII. | Összesen |
| Modell I. | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 12 |
| Modell II. | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 12 |
| Modell III. | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 9 |
| Modell IV. | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 12 |

**15. ábra:**  **A modell-értékelés nyers adatainak rangsor-nézete**

**Forrás:** saját szerkesztés

A 15. ábrából kitűnik, hogy a 3. modell érte el a megalkotott szempontrendszerek alapján a pontozásos versenyt, mely egyben a legegyszerűbb modellnek is számít az összes közül.

Regressziós modellek tapasztalatai

A döntési fák szabályképző logikájával ellentétben a regresszió az input-számértékek súlyozott összegén keresztül közelíti az Y értékét a korreláció, mint célérték maximalizálása mellett MS Excel Solver támogatással (restikciók nélkül, nem lineáris ÁRG keretek között, indulási értékként minden változóra 1-t adva a súly-pozícióknak):

1. Pclass (3 érték)
2. Sex\_Male (2 érték)
3. SibSp (2 érték)
4. Parch (2 érték)
5. Fare (8 érték)
6. Embarked\_Southhampton (2 érték)
7. Embarked\_Cherbourg (2 érték)
8. Embarked\_Queenstown (2 érték)
9. névhossz (8 érték)
10. Age-delta (8 érték)

Y=Survived (2 érték)

A kor adatai kapcsán a 3. ábrán feltárt optimum-jellegű összefüggés alapján a 23.5 évhez képesti abszolút eltérés került kiszámításra, rangsorolásra és 8 rangsorsávra arányosan szűkítésre. A névhossz és a jegyárak kapcsán a 714 nyers adat szintén sorszámozásra és 8 értékintervallumba sűrítésre került.

A regresszió súlyai a férfiak és a nők adatait egyszerre feldolgozó modellben:



**16. ábra:**  **A teljes utaskört egyszerre kezelő regressziós modell súlyai**

**Forrás:** saját szerkesztés

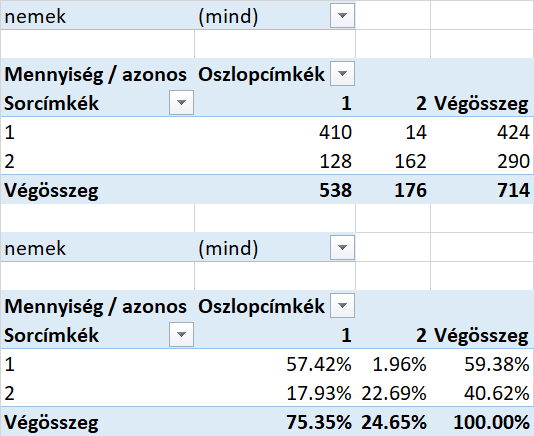
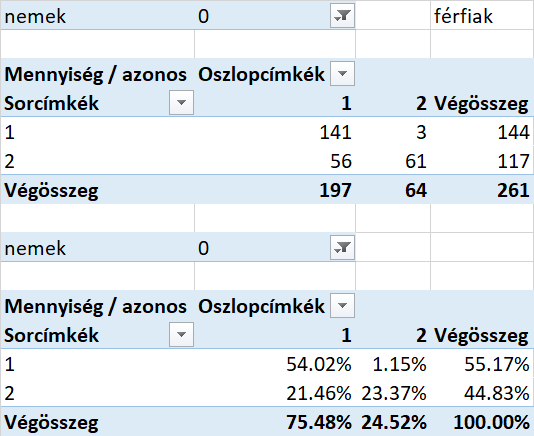
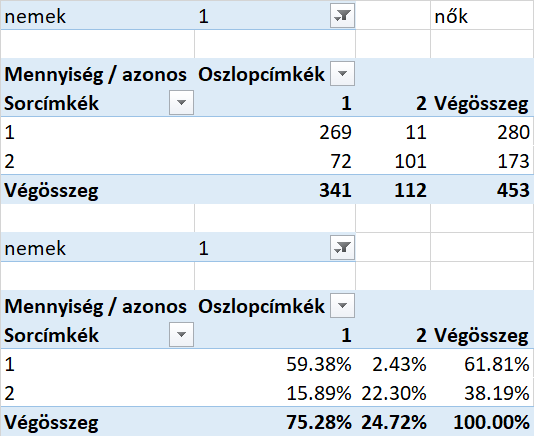
A regressziós modell hatásmechanizmusa a 714 becsült és a tényérték között -0.61 egységnyi korreláció elérése mellett:



**17. ábra:**  **A teljes utaskörre készített regresszió alapján feltárt hatásmértékek változónként**

**Forrás:** saját szerkesztés

Míg a 16. ábra önkényes nagyságrendű értékei lényegében semmilyen érdemi hermeneutikával nem láthatók el, addig a hatásmértékek alapján látható, hogy a névhossz és a kor-optimumtól való távolság mindösszesen csak 4%-ban hat, de hat. Az utasosztály és a nem a legerősebb hatású, de a Cherbourg-i beszállás is masszív hatást mutat.

**18. ábra:**  **A teljes utaskört feldolgozó regressziós modell találati arányai,a hol sor-oszlop-fejléceken: 1 = elhunyt, 2 = túlélte**

**Forrás:** saját szerkesztés

Ahogy az a 18. ábrán látható a teljes találati arány a két nemre egyszerre: 57.42+22.96=80.38%, ami nagyobb, mint a döntési fák esetén. A férfiak találati aránya a kisebb: 54.02+23.37=77.39%

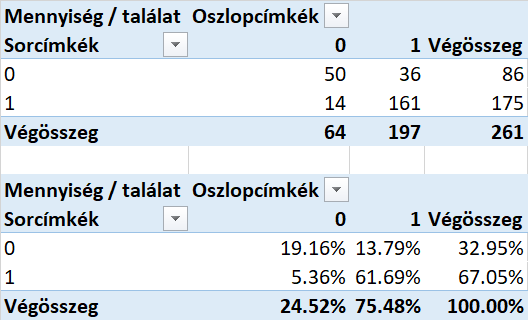
A férfiak és a nők kiemelésének oka az, hogy amennyiben a 261 férfire önálló regressziós modellt készítve a férfiakra jellemző találati arány felemelhető 80.84%-ra. A csak 453 nő adatsorral dolgozó regressziós modell nem volt képes javulást elérni a vegyes modellben a nőket jellemző 81.68%-hoz képest (ami további kutatási kérdésként egy önálló cikket enged vélelmezni). A súlyozott találati arány a csak férfi-adatokat feldolgozó modell integrálás után: 81.37%.



**19. ábra:**  **A csak férfiakra kialakított regressziós modell hatásmechanizmusa és korrelációja**

**Forrás:** saját szerkesztés

A 19. ábra kapcsán elmondható, hogy a kor-optimum hatása 1%-ról 4%-ra erősödött. Az utazási osztály lett a domináns változó, s a konstansként megtartott nem 7%-a azt jelzi, hogy a regressziós modellezés nem tud automatikusan mit kezdeni egy kontans érték megjelenésével. A csak férfi modellben már két kikötő is hatásmértékkel bír.



**20. ábra:**  **A csak férfiakról szóló regressziós modell találati arányai, ahol a sor-oszlop-fejléceken: 0 = túlélte, 1 = elhunyt**

**Forrás:** saját szerkesztés

A 20. ábrát összevetve a vegyes nemű találati arányok kontingencia-értékeivel látható, hogy az 54-59%-os és a 22-23%-os találati aránykomponensek 62:19 arányban szétnyíltak.

Mindezeken felül megjegyzendő:

* A döntési fák négy részeredménye (becslése) és ezek szórása alapján készített hibrid-regressziós modell és a 4 döntési fa becslései, a vegyes nemű regressziós modell becslése és ezen becslések szórása alapján készített regressziós modellek nem voltak képesek további találati aránynövekedést elérni sem additív, sem multiplikatív, sem vegyes becslőfüggvényekkel. Vagyis az egyes modellek által el nem talált személyek esetén érdemi mintázat nem volt felismerhető.
* A regresszióhoz felhasznált merev inputszerkezetet hasonlóságelemzéssel értelmezve kiderült, hogy a nyers inputok 8-sávos rétege nem tekinthető optimális paraméternek, hiszen nem sikerült a hasonlóság-alapú szabályképzés rugalmasságát találati előnyre váltani.

A regressziós modellépítési kísérletek döntési fákkal szembeni jobb eredményeinek okai:

* a kor, mint optimumhatást tükröző származtatott változó kialakítása,
* a férfi és vegyes modell hibridizálása,
* s önmagában a számarányok rugalmasabb erőtereinek kihasználása a tisztán kombinatorikai megoldás rel. merevségével szemben (ahol a max. 8 sávra sűrített input-rangsorszámok tudatosan ennek ellenében hatóan lettek kialakítva, egy fajta input-kombinatorikát kikényszerítve).

A nyers adatok és a modellezés lehetőségeinek egyelőre emberi intuíciót igénylő finomhangolása lehetséges. Esetlegesen modellgeneráló-robotok építésekor az emberileg előre elképzelt modell-variánsokat erőből egyesével tesztelve a robot illene, hogy átmenjen a Turing teszten, s akár az emberi intuíciós folyamatok munkaideje alatt jobb megoldásra is jusson.

1. Egyéb megjegyzések

Az eredeti verseny

Az eredeti a kaggle.com-on meghirdetett verseny célkitűzése, hogy a rendelkezésre bocsájtott adathalmaz alapján leszűrt következtetések szerint becslést adjunk még nem látott adathalmaz Survived (Túlélte) változó értékére, melynek tesztelése nem volt jelen cikk tárgya. Érdemes kiemelni, hogy a Leaderboard[[7]](#footnote-7) alapján 19 játékos is 100%-os tesztelési becslés-pontosságot ért el, és összesen ezzel együtt 46 játékos teljesített 90% felett. A játékban azonban nem csak a megadott adattáblát lehet felhasználni, bármilyen más egyéb statisztika vagy kutatási eredmény felhasználható, mely képes a felhasználókat közelebb vinni a helyes előrejelzés irányába.

Az adathalmaz elemzésbe bevont oszlopainak értékei és a kalkulált névhossz változó alapján elkészítettünk egy azonosítót minden egyes utasra és megvizsgáltuk, hogy lehetséges-e ugyanazon változók értékeiből képzett azonosító alapján különböző kimeneti értékeket kapni, azaz létezik-e olyan eset, hogy ugyanazon jellemzőkkel rendelkező utasok esetében a túlélési változó értéke különbözik. A megtalált duplikátumokat a 21. ábra szemlélteti.

|  |  |  |
| --- | --- | --- |
| Utas azonosító | Változók alapján képzett azonosító | Túlélte |
| 659 | 21230001310028 | 0 |
| 735 | 21230001310028 | 0 |
| 567 | 31190000810020 | 0 |
| 878 | 31190000810020 | 0 |
| 284 | 31190000810026 | 1 |
| 373 | 31190000810026 | 0 |
| 520 | 31320000810019 | 0 |
| 580 | 31320000810019 | 1 |
| 415 | 31440000810025 | 1 |
| 604 | 31440000810025 | 0 |

**21. ábra:** **Bemeneti változók alapján képzett utas azonosítók és a célváltozó szemléltetése**

**Forrás:** saját szerkesztés

Ahogyan az kitűnik a 21. ábrából, tíz utas esetén páronként előfordul a duplikált input-konstelláció, tehát összesen 5 alkalommal. Ebből 3 esetben a célváltozó különböző értékeket vesz fel, mely arra enged következtetni, hogy a rendelkezésre álló adatok alapján nem lehetséges a tökéletes becslés a tanulás során, tehát a vizsgált inputokhoz képest úm. a véletlennek is szerepe van abban, hogy adott utas túlélte-e a katasztrófát vagy sem. Ezek az ellentmondások feloldása csak pl. az eredeti utazonosítók (pl. jegyazonosító, kabinazonosító) számmisztikai (pl. prímszám, oszthatósági szabályok, mint státuszváltozók) értelmezése mellett kísérelhető meg.

Nemek szerinti becslés

Vitathatatlanul az összes változó közül különböző statisztikai eljárásokat elemezve a Sex (Nem) változó volt az, amely a legnagyobb hatással volt a célváltozó értéke. Az összes döntési fa esetében a legelső olyan változó volt, mely szabályképzővé vált. A férfi utasok esetén a becslések nagyrésze egyértelműen a Survived (Túlélte) változó „elhunyt” értékét vette fel a legtöbb esetben, míg a nők esetén különböző modellek, különböző eredményeket hoztak, mely feltételezi, hogy nők esetén a célváltozó értékének becslése kiszámíthatatlanabb, mely jelenséget érdemes a továbbiakban kutatások keresztüzébe helyezni. A női természet komplex működését alátámasztandó másik kutatás, a Tinder randialkalmazás használatát elemzi, melynek eredményeképp az a megállapítás nyert teret, hogy míg a férfiak esetében nagy mértékben lehetséges gépi tanuló módszerek segítségével előrejelezni, hogy mely hölgy nyeri el tetszését, nők esetében ez sokkal jobban kiszámíthatatlanabb a jóval változatos ízlés miatt. (dailymail.co.uk, 2016)

1. Irodalomjegyzék

Dailymail.co.uk (2016): Why Tinder is making women MISERABLE: Men swipe right for an ego boost with no intention of speaking to matches. Letöltve: <http://www.dailymail.co.uk/sciencetech/article-3710505/What-Tinder-strategy-Men-likely-casually-swipe-right-women-match-attract-them.html> Letöltés ideje: 2018. 01. 15.

Hámori Gábor (2001): A CHAID alapú döntési fák jellemzői. Statisztikai szemle, 79. évf., (8): 703-710. p.

Kaggle Dataset: <https://www.kaggle.com/c/titanic/data> Letöltés ideje: 2017.11.01

Pitlik László (jun), Pitlik László (2014): Occam borotvája finomhangolva – MIAÚ No.185 <http://miau.gau.hu/miau/185/occams_razor_finetuned.doc>

rd.hu (é. n.): A Titanic teljes története. Letöltve: [http://www.rd.hu/A\_Titanic\_  
teljes\_t%C3%B6rt%C3%A9nete](http://www.rd.hu/A_Titanic_teljes_t%C3%B6rt%C3%A9nete) Letöltés ideje: 2018. 01. 08.

1. Mellékletek
2. számú melléklet: Python kód RandomForest

import pandas as pd

from sklearn.ensemble import RandomForestClassifier

import numpy as np

import matplotlib.pyplot as plt

df\_titanic = pd.read\_csv('D:\Egyetem\Statisztikai programcsomagok\TitanicTrns4.csv')

df\_titanic.columns = ['PassengerId','Survived','Pclass','Sex','Age', 'SibSp','Parch','Fare','Embarked\_Southhampton','Embarked\_Cherbourg','Embarked\_Queenstown', 'LenName']

X, y = df\_titanic.iloc[:, 2:].values, df\_titanic.iloc[:, 1].values

feat\_labels = df\_titanic.columns[2:]

forest = RandomForestClassifier(n\_estimators=10000, random\_state=0, n\_jobs=-1)

forest.fit(X, y)

importances = forest.feature\_importances\_

indices = np.argsort(importances)[::-1]

for f in range(X.shape[1]):

print("%2d) %-\*s %f" % (f + 1, 30, feat\_labels[indices[f]], importances[indices[f]]))

pass

plt.bar(range(X.shape[1]), importances[indices], color='lightblue', align='center')

plt.xticks(range(X.shape[1]), feat\_labels[indices], rotation=90)

plt.xlim([-1, X.shape[1]])

plt.tight\_layout()

plt.show()

1. számú melléklet: Feldolgozott fájlok



Szerzői adatok:

Barta Gergő

PhD hallgató

Gazdálkodás és Szervezéstudományok Doktori Iskola, Szent István Egyetem

SZIE, MY-X kutatócsoport, 2100 Gödöllő, Páter K. u. 1., [barta.gergo@phd.uni-szie.hu](mailto:barta.gergo@phd.uni-szie.hu)

Pitlik László

Egyetemi docens

Társadalomtudományi és Tanárképző Intézet, Szent István Egyetem

SZIE, MY-X kutatócsoport, 2100 Gödöllő, Páter K. u. 1., [pitlik@miau.gau.hu](mailto:pitlik@miau.gau.hu)

1. [↑](#footnote-ref-1)
2. <https://www.kaggle.com/c/titanic/data> [↑](#footnote-ref-2)
3. A névhossz és a milyen előkelő osztályon utazás közötti kapcsolat (egy fajta konzisztencia-ellenőrzésként) azt mutatta az elvárásoknak megfelelően, hogy a hosszabb nevű utasok előkelőbb osztályon utaztak. Az átlagos névhossz 27.7 karakter, ebből az első osztályon utazók átlagos névhossza 30.2, a második osztályon utazók átlagos névhossza 28.4, míg a harmadik osztályon utazók átlagos névhossza 26 karakter. [↑](#footnote-ref-3)
4. A társadalomtudományi kutatásokban általános elfogadott 0.05 alatti szignifikancia szint alatti érték miatt. [↑](#footnote-ref-4)
5. A Cabin (Kabin sorszáma) változó 687 esetben nem tartalmazott adatot, így értelemszerű a változó elhagyása az információtartalmát eltekintve is. [↑](#footnote-ref-5)
6. http://miau.gau.hu/miau/185/occams\_razor\_finetuned.doc [↑](#footnote-ref-6)
7. https://www.kaggle.com/c/titanic/leaderboard [↑](#footnote-ref-7)