

Haladó Adatelemzés Labor Dokumentáció

Pap Arion - YJWQMP
Bata Dániel - B95Q0I
Nyiri Balázs - REMAIO

Absztrakt

A feladatunk a Titanic tragédia adatait feldolgozó, széles körben használt Kaggle-dataset elemzésén keresztül mutatja be a túlélés előrejelzésének lehetőségeit különböző modellek segítségével. A vizsgálat során változók, például nem, életkor, utasosztály, jegyár és családi kapcsolatok kerültek elemzésre, az adat-előkészítés és tisztítás (pl. hiányzó adatok pótlása, adatok feldolgozása) kiemelt szerepet kapott. A legbefolyásosabb változónak a nem, az utasosztály és a jegyár bizonyultak. A feladat során három modellt (Random Forest, Gradient Boosting Machine, Extreme Gradient Boosting) hasonlítottunk össze, hiperparaméter optimalizálással finomítva. A Random Forest és a Gradient Boosting modellek bizonyultak a legpontosabbnak, míg az Extreme Gradient Boosting modell lemaradt teljesítményben.

Bevezetés

A Titanic-dataset a Kaggle "Titanic: Machine Learning from Disaster" versenyéből származik, és az RMS Titanic első útján utazó 891 utas tragédiáját dokumentálja. Ez az esemény több szempontból is rendkívül fontos: egyrészt rámutatott a korabeli tengeri közlekedés biztonsági hiányosságaira és az utasok társadalmi elhelyezkedésének élet-halál kérdéseiben játszott szerepére; másrészt a túlélési arányok és az egyéni jellemzők (például nem, életkor, fedélzeti osztály, családi kapcsolatok) közötti összefüggések feltárása élő, valós történeti és társadalomtudományi kontextust biztosít. Az vizsgálatunk elsődleges célja, hogy a rendelkezésre álló demográfiai és szocioökonómiai jellemzők segítségével több "túlélési" (binary classification) modellt építsünk és összehasonlítsuk ezeket.

Minden rekord egy utas profilját tartalmazza (ezekről később). Az adatokat tanító és teszt halmazokra bontották, hogy a tanító halmazon ismert túlélési eredmények alapján építhessünk osztályozó modelleket, majd alkalmazhassuk azokat a teszt halmazon. Kezdők körében népszerű mérföldkőként szolgál ennek az adathalmaznak a használata, hiszen jól demonstrálja az adat-előkészítés (például hiányzó Age vagy Cabin értékek kezelése), a adat transzformáció (például a nevekből kiemelt titulusok), valamint a gépi tanulási algoritmusok összehasonlításának nehézségeit.

Irodalomkutatás

A Titanic-dataset-tel több cikk is foglalkozott modellek kiértékelésével és adatok feldolgozásának technikáival kapcsolatban. Mi a feladatvégzésünk előtt négy cikkel foglalkoztunk a legtöbbet. Két cikk is foglalkozott modellek kiértékeléssel a probléma megoldására. Mindkettő a Logistic Regression, Decision Tree és Random Forest modelleket értékelték ki.

Az első cikk [1] eredményei:

- Logistic Regression (LR):
 - Keresztvalidációs pontszám: 79%
 - Precision: 88.2%
 - Recall: 82.9%
 - AUROC: 0.81

- Decision Tree (DT):
 - Mélység: 45 és 75
 - Keresztvalidációs pontszám: 77%
 - Precision: 86.8%
 - Recall: 84.6%
- Random Forest (RF):
 - Mélység: 45 és 75
 - Pontszám: 79.8%
 - Precision: 91.9%
 - Recall: 87.2%

Második cikk [2] eredményei:

- Decision Tree: 99.29% pontosság (tréning adatokon)
- Logistic Regression: 81.11% pontosság (teszt adatokon)
- Random Forest: 97.53% (tréning adatokon), 80.41% (teszt adatokon)

Másik kettő cikk főleg az adattisztítással és feldolgozással és adatmegfigyeléssel foglalkozik [3], [4]. A nevekből kinyertek titulusokat, aminek segítségével egészítették ki a hiányos Age értékeket az azonos titulussal rendelkező utasok átlag életkorával. A Cabin változó hiányzó értékeit általában egy Unknown érték bevezetésével oldották meg. Bevezettek újabb változókat is a modellek tanítására, például jegyár költsége személyre leosztva, ugyanis vannak, akiknek a jegyára az egész családjának az összesített jegyárában van megadva.

Következtetés

A négy cikkből származó megközelítések közös pontokat mutatnak a Titanic túlélés előrejelzésének modellezésében. A Random Forest modellek következetesen jobban teljesítettek, míg a logisztikus regresszió a pénzügyi modellezésben bevált egyszerűsége miatt továbbra is alapmodell marad. Az új jellemzők bevezetése — különösen a címek, családméret és osztály kombinációk — jelentősen hozzájárult a modellek pontosságához.

Adatok bemutatása

Adat neve	Adat jellemzése
PassengerId	Adatsor id-je
Survived	Bináris érték. 1:Túlélte, 0: Nem élte túl
PClass	Utas osztálya (első, másod, harmad)
Name	Utas neve
Sex	Utas neme
Age	Utas kora
SibSp	Az Utas testvéreinek és vagy férjének/feleségének száma a fedélzeten.
Parch	Az Utas gyerekeinek vagy szüleinek száma a fedélzeten.
Ticket	A jegy azonosítója (string)
Fare	A jegy ára
Cabin	Az Utas kabinjának azonosítója (string)
Embarked	Kikötő ahol az utas felszállt

Adatelemzés és feldolgozás

Az adatok minőségét az első lépésekben a `df.info()` és a `df.isnull().sum()` függvények segítségével vizsgáltuk, ami azt mutatta, hogy az egyes attribútumok közül az **Age** mezőben 263, a **Cabin** mezőben 1014, az **Embarked**-ban 2, a **Fare**-ben pedig 1 hiányzó érték található (a **Survived**-ban csak a teszthalmazon hiányzik 418 esetben, ennyi a teszthalmaz mérete). Különösen a **Cabin** oszlop tartalmaz sok NaN értéket – több mint 77 % hiányzik –, ezért a további elemzések egyszerűsítése érdekében, felmérve az adat alapvető hasznosságát is, arra a döntésre jutottunk, hogy eltávolítjuk ezt az oszlopot. A **PassengerId** és a **Ticket** oszlopokat azért dobtuk el, mert önmagukban nem nyújtanak előrejelző értéket, viszont a zajt és a felhasznált erőforrásokat növelik. A hiányzó **Embarked** értékeket a jegyár-mediánokhoz viszonyított abszolút különbség alapján töltöttük ki, a **Fare** egyetlen hiányzó értékét pedig az adott kikötő szerinti csoport mediánjával pótoltuk. Az **Age** hiányzó értékeinek kezeléséhez a **Name** mezőből kivont titulus kategóriák (**Title**) szerint csoportosítottuk az utasokat („Mr”, „Mrs”, „Miss” stb.), ezek medián korát kiszámoltuk, és ezekkel a mediánokkal helyettesítettük az ismeretlen korokat.

Ugyanakkor új, modellezés szempontjából releváns jellemzőket vettünk fel: az előbb említett **Title** oszlop beemelésével finomabb kor- és státuszfelosztást teszünk lehetővé,

továbbá a **Sex** attribútumot bináris **Sex_male** változóra alakítottuk az **Embarked** értékeit pedig One-Hot-Encodeoltuk, hogy az algoritmusok számára könnyebben kezelhetővé váljanak.

Ezzel a bővített változó könyvtárral a modell felkészítése során alaposabban tudjuk megragadni a túlélés esélyét befolyásoló tényezőket.

Adatvizualizáció

Korrelációs Mátrix

	Pclass	Age	Fare	SibSp	Parch	Sex_male	Embarked_C	Embarked_Q	Embarked_S	Survived
Pclass	1.000000	-0.356248	-0.549500	0.083081	0.018443	0.131900	-0.251139	0.221009	0.081720	-0.338481
Age	-0.356248	1.000000	0.100406	-0.261951	-0.184460	0.092715	0.050160	-0.061560	-0.005388	-0.071470
Fare	-0.549500	0.100406	1.000000	0.159651	0.216225	-0.182333	0.273614	-0.117216	-0.166603	0.257307
SibSp	0.083081	-0.261951	0.159651	1.000000	0.414838	-0.114631	-0.061970	-0.026354	0.070941	-0.035322
Parch	0.018443	-0.184460	0.216225	0.414838	1.000000	-0.245489	-0.013725	-0.081228	0.063036	0.081629
Sex_male	0.131900	0.092715	-0.182333	-0.114631	-0.245489	1.000000	-0.090223	-0.074115	0.125722	-0.543351
Embarked_C	-0.251139	0.050160	0.273614	-0.061970	-0.013725	-0.090223	1.000000	-0.149345	-0.784064	0.174718
Embarked_Q	0.221009	-0.061560	-0.117216	-0.026354	-0.081228	-0.074115	-0.149345	1.000000	-0.496624	0.003650
Embarked_S	0.081720	-0.005388	-0.166603	0.070941	0.063036	0.125722	-0.784064	-0.496624	1.000000	-0.155660
Survived	-0.338481	-0.071470	0.257307	-0.035322	0.081629	-0.543351	0.174718	0.003650	-0.155660	1.000000

A korrelációs mátrix alapján több fontos összefüggés rajzolódik ki az egyes jellemzők és a túlélés (**Survived**) között, valamint a jellemzők egymás közötti viszonyában:

1. Túlélésre gyakorolt hatás

- A legerősebb negatív kapcsolatot a **Sex_male** mutatja a túléléssel ($\approx -0,54$), azaz a férfiak túlélési esélye szignifikánsan alacsonyabb volt.
- A **Pclass** és a túlélés korrelációja is viszonylag magas negatív értékkel bír ($\approx -0,34$): az alacsonyabb (3-as) osztályból származó utasok nagyobb eséllyel nem éltek túl a katasztrófát.
- A **Fare** pozitív korrelációt mutat a túléléssel ($\approx +0,26$), ami arra utal, hogy a magasabb jegyár – gyakran jobb elhelyezkedés a hajón – növelte a túlélés esélyét.
- A **Embarked_C** enyhe pozitív hatása ($\approx +0,17$) azt sugallja, hogy a Cherbourg-ból beszállók valamivel jobb túlélési arányt értek el, míg az **Embarked_S** (Southampton) enyhén negatív ($\approx -0,16$). Ebből igazából a kikötő környékén lévő élelszínvonalra lehet következtetni, ha számításba vesszük azt is, hogy gazdagabb, jobb jeggyel rendelkező utasok nagyobb eséllyel éltek túl a történeteket. Az **Embarked_Q** korrelációja elhanyagolható.

2. Egyéb jellemzők

- Az életkor (**Age**) korrelációja a túléléssel elenyésző ($\approx -0,07$), ami arra utal, hogy önmagában lineárisan nincs erős összefüggés (de nem zárja ki az életkor és más változók közötti nemlineáris hatásokat).
- A családi kapcsolatok (**SibSp**, **Parch**) gyenge, gyakorlatilag semleges hatást mutatnak ($\approx -0,035$, illetve $+0,081$), bár közülük a **Parch** kismértékben pozitívabb.

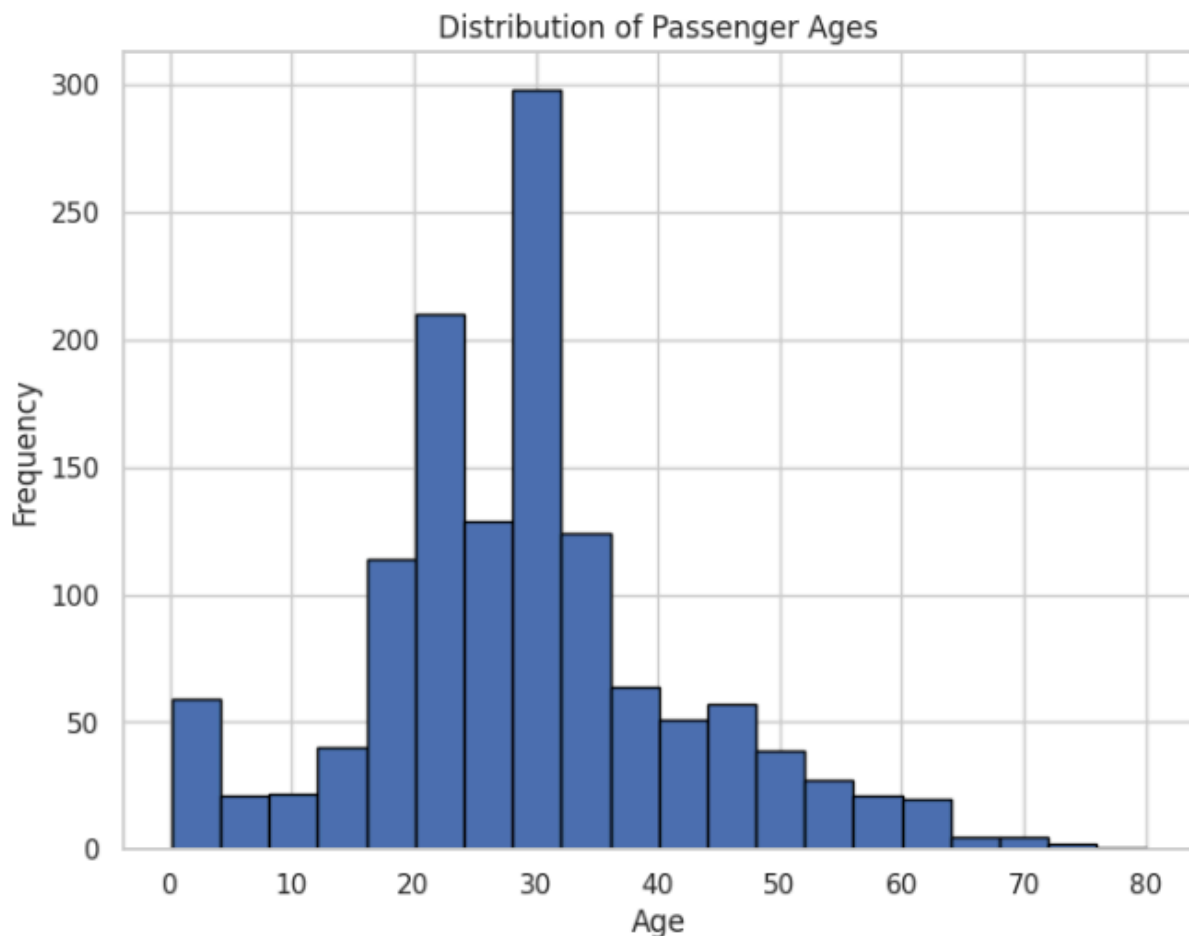
3. Jellemzők közötti összefüggések

- A **Pclass** és a **Fare** erősen negatív korrelációban vannak ($\approx -0,55$), ami azt tükrözi, hogy az első osztályú jegyek jellemzően magasabb áron keltek el.
- A testvérek/hozzá tartozók száma (**SibSp**) és a szülő/gyermek kapcsolat (**Parch**) viszonylag erősen együtt mozog ($\approx +0,41$), hiszen gyakran nagyobb családok utaztak együtt.

Összességében a mátrix jól alátámasztja, hogy a túlélés szempontjából a legfontosabb prediktorok a nem (**Sex_male**), a jegyosztály (**Pclass**) és a jegyár (**Fare**), míg az életkor és a családi kapcsolatok lineáris hatása gyengébbnek bizonyult. Ugyanakkor a jellemzők közti erős korrelációk (multikollinearitás) figyelembevételre – például **Pclass** vs. **Fare** – fontos lehet a modellek stabilitásának megőrzéséhez.

Érdeemes kiemelni, hogy a Pearson-féle korreláció csak a lineáris kapcsolatokat ragadja meg, ezért nem érzékeli jól a nemlineáris hatásokat (például a gyermekek magasabb túlélési arányát), illetve nagy mértékű outlierok (tipikusan a nagyon magas jegyárak) torzíthatják az eredményt. A „Fare” erősen ferde eloszlása indokoltá teheti a log-transzformáció alkalmazását, hogy finomabban modellezhessük az árbeli különbségek hatását. Ugyancsak befolyásolja a korrelációs értékeket az imputációs stratégia: az „Age” esetén titulus-alapú mediánokkal pótoltuk hiányzó adatokat, ami homogenizálja az életkor eloszlását, és elrejtheti a valós életkori mintázatokat. A one-hot-kódolt „Embarked” változók közti erős negatív korrelációk (**C** vs. **S**) pedig egyszerűen a one-hot kódolás következményei, nem valódi kölcsönhatást jeleznek. Érdeemes lehet továbbá a **FamilySize = SibSp + Parch + 1** összefoglaló változó bevezetése, mert így a családi csoport hatása erősebben látható, mint amikor a **SibSp** és **Parch** külön-külön szerepel. Végül pedig ne feledjük: a korreláció nem feltétlenül jelent ok-okozati viszonyt; például a magasabb „Fare” és a túlélés közti összefüggés mögött a jobb anyagi státusz is állhat, így a modellezés során érdemes lehet további, kauzális vagy többváltozós elemzéseket is beiktatni.

Utasok osztály és kor szerinti felbontása

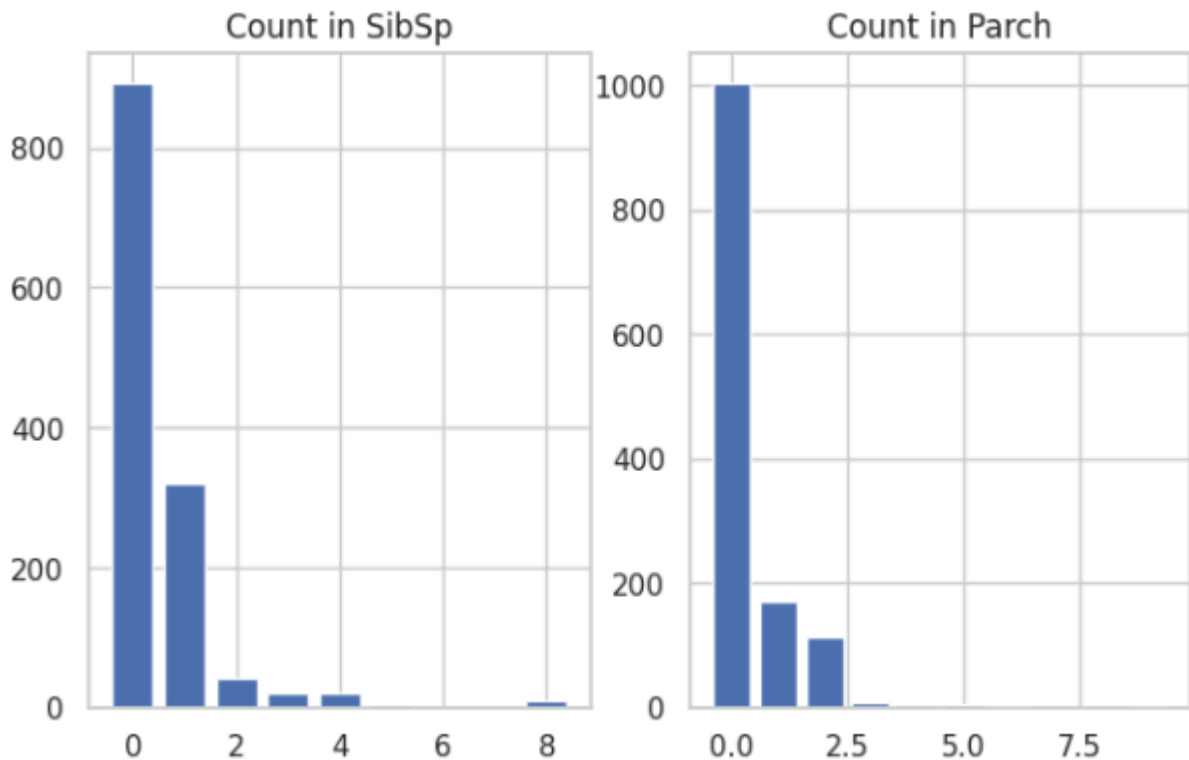


A koreloszlás-hisztogram egy jól definiált csúcsot mutat a 25–30 éves korosztálynál, ami jelzi, hogy a hajó utasainak többsége fiatal felnőtt volt. A két kiugró érték valószínűsítőleg a mediánnal feltöltött előzőleg ismeretlen életkor értékek okozzák. Az eloszlás enyhén jobbra ferdült, azaz bár a középpérték körül torlódnak az adatok, a 40–80 évesen a hosszú, elnyúlt farokban jelennek meg. A 0–5 éves sávban látható kisebb kiugrásnál kérdéses, hogy tényleg ennyire sok kisgyerek volt, vagy pedig régen pár ismeretlen értéket 0-val töltöttek volna fel.

Következtetések és javaslatok:

- Az életkor nemlineáris hatásait (például a gyermekek túlélési előnyét vagy az idősek kiemelt kockázatát) érdemes korcsoportokra („Child”, „Young Adult”, „Adult”, „Senior”) bontva modellezni.
- A kovariancia-csökkentés érdekében alternatív imputációs stratégiákat (pl. prediktív imputáció) vagy a kor bináris átalakítását (pl.: $\text{Age} \leq 16?$) is megfontolhatjuk.
- A szélsőséges életkori értékek hatását robosztusabb modellezési technikák (pl. rang-transzformáció, winsorizálás) alkalmazásával mérsékelhetjük, hogy az illesztés ne torzuljon a ritka, extrém esetek miatt.

Családi kapcsolatok



A két oszlopdiagram a családi kíséretre vonatkozó jellemzők eloszlását mutatja:

- **SibSp (Sibling/Spouse)**

Itt jól látszik, hogy az utasok közel 900-an egyedül utaztak (SibSp = 0), további ~300-an egy testvérrel vagy házastárssal, majd gyorsan zuhan a darabszám: 2–4 kísérő esetén már csak pártucatnyi utas, és egy extrém érték (8) is megjelenik. Ez erősen jobbra ferdén eloszlott, nagy többségben egyedülálló vagy csak egypár fős utazókat mutat.

- **Parch (Parent/Child)**

Hasonlóan a SibSp-hez, itt is az egyedül utazók dominálnak: 1000 feletti számmal Parch = 0, majd ~170 utasnak volt egy szülője vagy gyermeke, ~100-nak kettő, és alig néhányan utaztak háromnál több édesanya/apa- vagy gyermek társaságban.

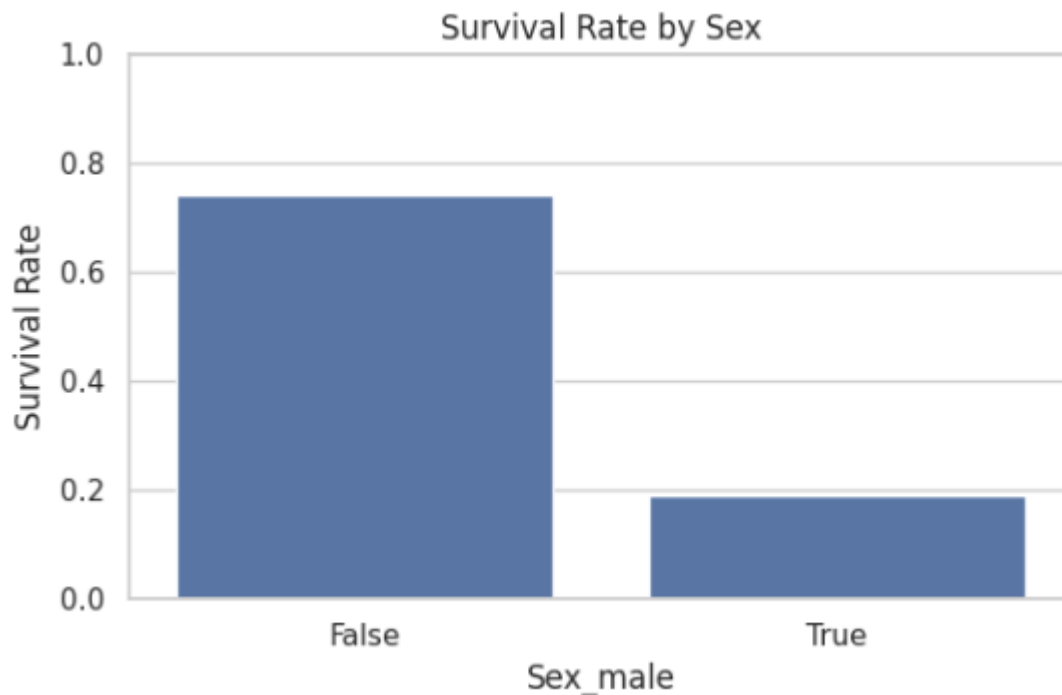
Következtetés & javaslatok

- Mindkét esetben rendkívül egyoldalú, jobbra ferdén eloszlást látunk, ami azt jelzi, hogy a családi kapcsolatok ritkán terjednek ki nagy csoportokra.
- A nagyon kis elemszámú, nagyobb SibSp vagy Parch értékeknél érdemes lehet aggregált kategóriákat (pl. „nagy család” – $\text{SibSp} + \text{Parch} \geq 3$) képezni, hogy a ritka esetek ne dobják fel aránytalanul a zajt a modellben.
- Javasolt továbbá az előre összesített **FamilySize = SibSp + Parch + 1** változó használata, amely egyszerre ragadja meg az utas körüli teljes családi csoport

nagyságát és lineárisan semlegesíti a külön-külön nagyon alacsony gyakoriságú értékeket.

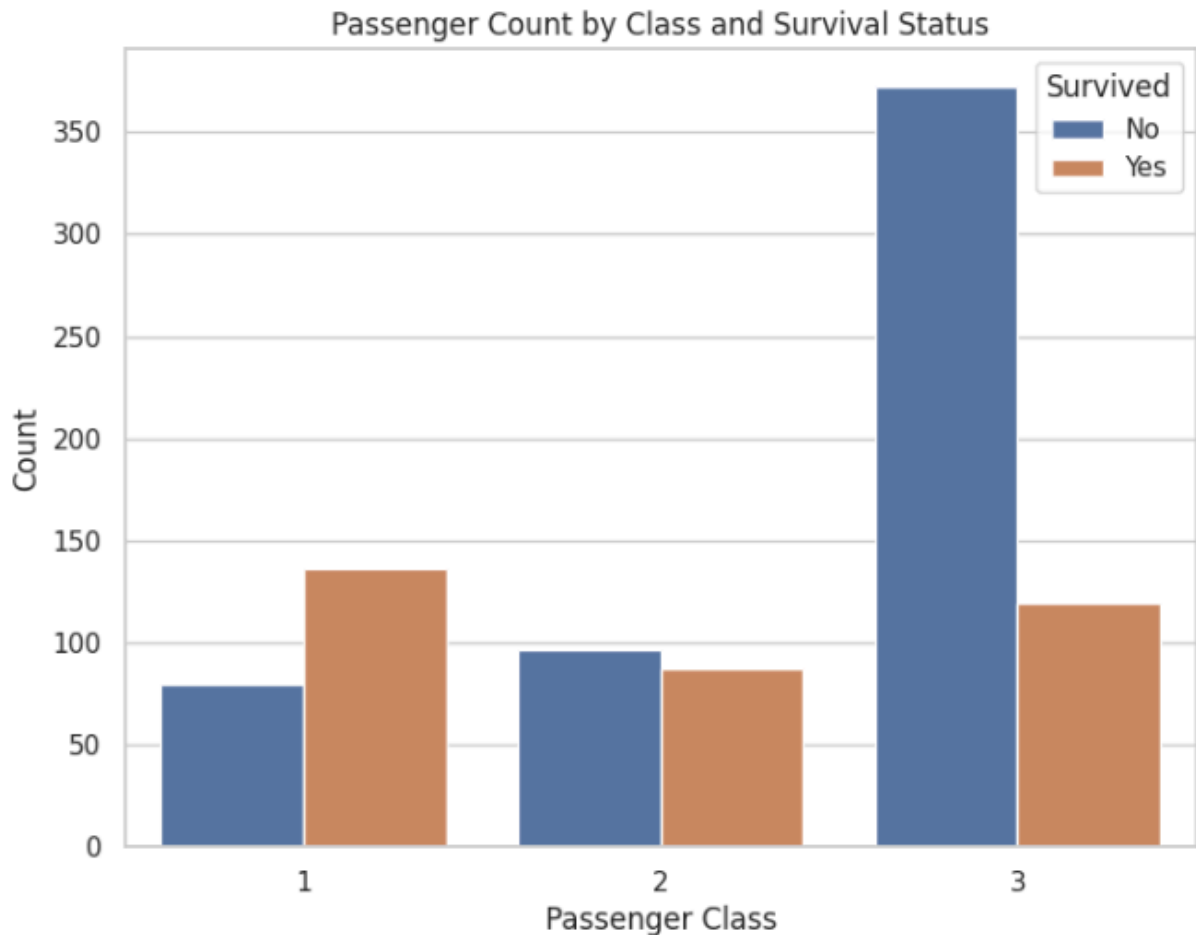
- A családi méret hatása a túlélésre nem feltétlenül lineáris, ezért érdemes lehet életciklus-kategóriákban (egyedül, kis család, nagy család) vizsgálni a túlélési arányokat.

Nemek és túlélés kapcsolata



A diagram világosan mutatja a nemek közötti markáns túlélési különbséget: a női utasok (`Sex_male = False`) körülbelül 74 %-a élte túl a katasztrófát, míg a férfiaké (`Sex_male = True`) mindössze mintegy 18 %. Ez a „nők és gyerekek előre” elv drámai megnyilvánulása – a nők létszámuknál jóval nagyobb arányban kerültek mentőcsónakba. Modellépítés szempontjából éppen ezért a `Sex_male` változó a legerősebb egyedi prediktor a túlélés előrejelzéséhez.

Utassosztály és túlélés kapcsolata



Az oszlopdiagram jól szemlélteti, hogy a hajó utasainak társadalomban betöltött szerepe drámai hatással volt a túlélésre:

- **I. osztály:** körülbelül 135 utas élte túl, míg mintegy 80-an veszítették életüket (túlélési arány ~63 %).
- **II. osztály:** itt már kiegyenlítettebb a kép, nagyjából 87 túlélő és 96 áldozat (túlélési arány ~48 %).
- **III. osztály:** a legnagyobb számú csoportban óriási volt a lemaradás: kb. 370 halott szemben mintegy 120 túlélővel (túlélési arány ~25 %).

Ebből egyértelműen látszik, hogy minél magasabb osztályon utazott valaki, annál nagyobb eséllyel menekült meg – azaz a **Pclass** erős prediktora a túlélésnek. A modellépítés során érdemes lehet tovább vizsgálni a **Pclass** és más változók (például **Sex_male** vagy **Age**) közötti kölcsönhatásokat, hiszen például az első osztályú nők túlélési aránya még inkább kiemelkedő lehet a teljes populációhoz képest.

Modellek

A feladat elvégzéséhez 3 modellt választottunk hogy megfigyeljük milyen hatékonysággal képesek a túlélés kikövetkeztetésére az adathalmaz alapján. A hiperparaméter optimalizálás elvégzéséhez mindhárom modell esetén alkalmaztunk GridSearch algoritmust, a hatékonyabban tanított modellek kialakításának érdekében. Az algoritmus futtatását csak pár alkalommal végeztük el, mivel minden futtatása hosszabb időbe telt (akár 20-30 perc), az eredmények megszerzése után újrafuttatni az algoritmust amúgy is felesleges lenne. Kiértékelést a modelleken elvégeztük a Gridsearch algoritmus használata előtt, illetve utána is, hogy látható legyen a különbség amit a hiperparaméter optimalizálással el tudtunk érni a modellek esetében.

Általánosságban elmondható, hogy a feladat megoldásához a modelleknek nincs szüksége nagyobb `n_estimator` számra mint 100, illetve a döntési fák maximális mélysége, `max_depth` értéke sem haladja meg a 10-et, sőt általában 5 mélységet talált megfelelőnek.

Random Forest

A Random Forest egy ensemble gépi tanulási módszer, amely több döntési fa kombinációján alapul a predikciós pontosság növelése és a túlillesztés csökkentése érdekében. A Random Forest könnyen kezel hiányzó értékeket, rangsorolni tudja a bemeneti változók fontosságát, és általánosan jó általánosítási képességgel rendelkezik, ezért széles körben alkalmazzák mind osztályozási, mind regressziós feladatokban.

Gradient Boosting Machine

A Gradient Boosting is egy ensemble módszer, mely gyenge tanulókból—gyakran sekély döntési fákból —állít össze erős prediktív modellt. Ahelyett, hogy független fákat építenénk (mint a Random Forestnél), itt minden új fa a korábbi fák által elkövetett hibák – a kiválasztott veszteségfüggvény gradiense alapján számított – korrekciójára specializálódik. Hátránya viszont, hogy érzékenyebb a zajra és a túlillesztésre, valamint a tanítási idő általában hosszabb, különösen nagy adathalmazokon.

Extreme Gradient Boosting Machine

Az Extreme Gradient Boosting (XGBoost) a klasszikus Gradient Boosting továbbfejlesztett, nagy teljesítményre optimalizált változata. Emellett sparsity-aware technikákat alkalmaz a hiányzó értékek kezelésére, és Tree Pruning lépésekkel vágja le az irreleváns ágakat, hogy jobban általánosítsanak a modellek.

Kiértékelés:

A modellek kiértékelésére több metrikát is alkalmaztunk, például:

- Accuracy
- Precision
- Recall
- ROC görbe és AUROC

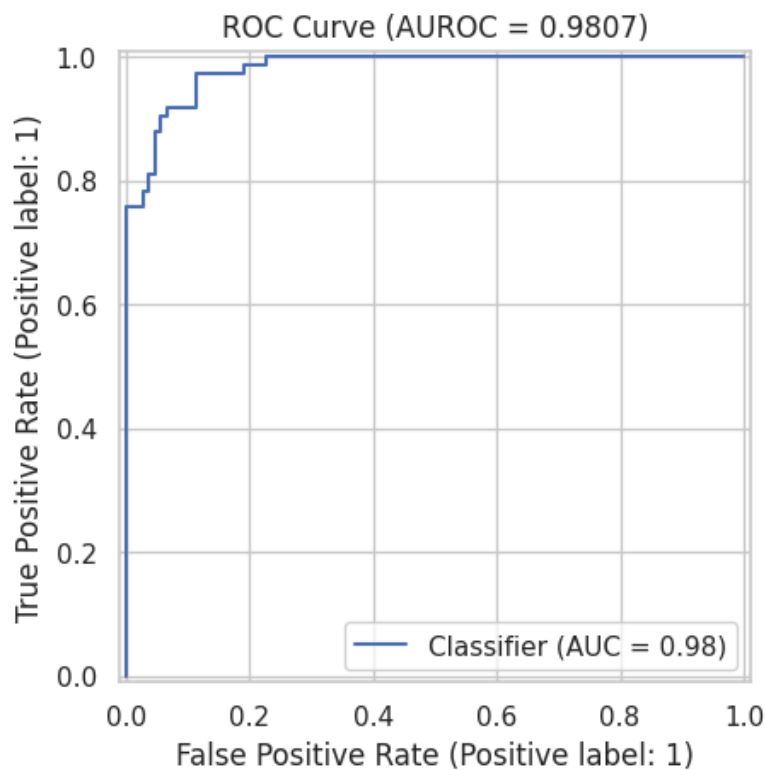
Random Forest

Classification Report:				
	precision	recall	f1-score	support
0.0	0.84	0.88	0.86	105
1.0	0.81	0.76	0.78	74
accuracy			0.83	179
macro avg	0.82	0.82	0.82	179
weighted avg	0.83	0.83	0.83	179

A Random Forest modell paraméter optimalizáció előtt 0.8268 pontosságot ért el a túlélők kikövetkeztetésében, miközben a tévesztések száma aránylag nagyobb volt azok esetében akik nem éltek túl, mint azok esetében akik túléltek, ami látható a Recall értékek közötti különbségben.

Classification Report:				
	precision	recall	f1-score	support
0.0	0.9252	0.9429	0.9340	105
1.0	0.9167	0.8919	0.9041	74
accuracy			0.9218	179
macro avg	0.9210	0.9174	0.9190	179
weighted avg	0.9217	0.9218	0.9216	179

A hiperparaméter optimalizálás után az Accuracy értéke 0.9218 lett, viszont a modell a tévesztések arányában továbbra is rosszabbul becsülte meg ha valaki nem élt túl. Ez látható itt is a Recall értékek közötti különbségben, ami arra utalhat, hogy a Random Forest modell jobban rátanult a túlélők kikövetkeztetésére, mivel több minta áll rendelkezésére mint a túlélők közül. Azt is jelentheti, hogy a modell előbb tippeli azt az adott utasra, hogy nem élte túl, ha bizonytalan a döntésben.



Látható a ROC görbe ábráján, hogy a a bal felső sarokhoz elég közel halad a görbe, ami mutatja, hogy a modell jól képes teljesíteni. A görbe illetve a magas AUROC érték mutatja, hogy a modell jó arányban képes megkülönböztetni azokat az utasokat akik túléltek az utat illetve azokat az utasokat akik nem.

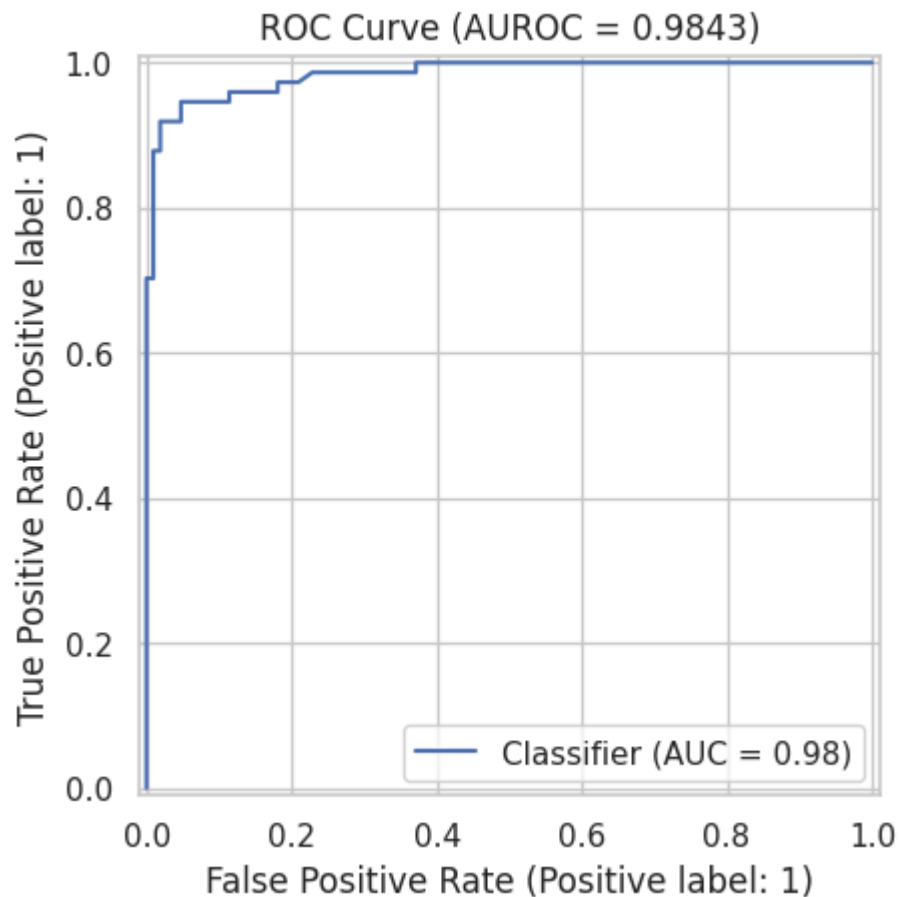
Gradient Boosting Machine

Classification Report:				
	precision	recall	f1-score	support
0.0	0.82	0.89	0.85	105
1.0	0.82	0.73	0.77	74
accuracy			0.82	179
macro avg	0.82	0.81	0.81	179
weighted avg	0.82	0.82	0.82	179

A Gradient Boosting Machine modell hiperparaméter optimalizáció előtt 0.8212 pontosságot ért el, miközben a tévesztések száma aránylag sokkal nagyobb volt azok esetében akik nem éltek túl, mint azok esetében akik túléltek.

Classification Report:				
	precision	recall	f1-score	support
0.0	0.9524	0.9524	0.9524	105
1.0	0.9324	0.9324	0.9324	74
accuracy			0.9441	179
macro avg	0.9424	0.9424	0.9424	179
weighted avg	0.9441	0.9441	0.9441	179

A hiperparaméter optimalizálás után az Accuracy értéke 0.9441 lett, illetve a modell egyenletesen hibázott a célváltozó mindkettő értékének kikövetkeztetésében. Látható a kiértékelési metrikákban, hogy a modell jól teljesített, illetve a Recall értékek közel vannak egymáshoz, ami arra utal, hogy a modell jelentős mértékben javult a túlélők és nem túlélők kikövetkeztetésének egyenletes tanulásában.



Látható a ROC görbe ábráján, hogy a a bal felső sarokhoz elég közel halad a görbe, ami mutatja, hogy a modell jól képes teljesíteni. A görbe illetve a magas AUROC érték mutatja, hogy a modell nagyon jó arányban képes megkülönböztetni azokat az utasokat akik túléltek az utat illetve azokat az utasokat akik nem.

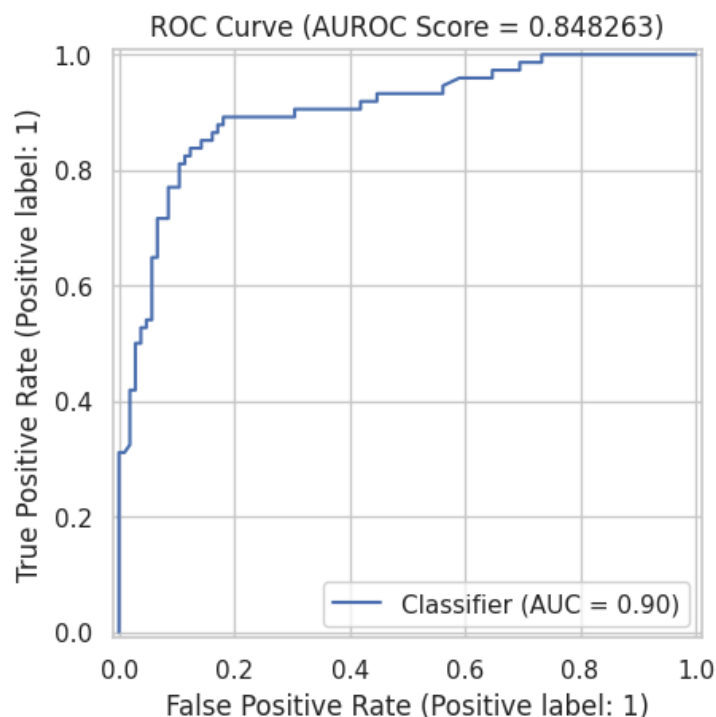
Extreme Gradient Boosting Machine

Classification Report:				
	precision	recall	f1-score	support
0.0	0.83	0.91	0.87	105
1.0	0.86	0.73	0.79	74
accuracy			0.84	179
macro avg	0.84	0.82	0.83	179
weighted avg	0.84	0.84	0.84	179

Az Extreme Gradient Boosting Machine modell paraméter optimalizáció előtt 0.8380 pontosságot ért el a túlélők kikövetkeztetésében. A Recall értékek a kettő célváltozó érték esetében jelentős mértékben különbözik, mivel a modell sokkal több esetben következtette ki hibásan egy utasról, hogy nem élte túl az utat mint túlélte, mint annak fordítottját. Ez megfigyelhető volt eddig mind a három modellre.

Classification Report:				
	precision	recall	f1-score	support
0.0	0.87	0.89	0.88	105
1.0	0.83	0.81	0.82	74
accuracy			0.85	179
macro avg	0.85	0.85	0.85	179
weighted avg	0.85	0.85	0.85	179

A hiperparaméter optimalizálás után a pontossága 0.8547 lett, ami jelentősen alacsonyabb mint a két másik modell esetében volt. Látható, hogy bár javult a Recall értékek között a különbség, de még továbbra is nagyobb mint a többi modell esetében volt. Itt a GridSearch algoritmus használata nem javított olyan sokat a teljesítményen.



Az eredmények a ROC görbén is jól láthatóak, bár az AUROC érték még mindig elfogadható, de jelentősen lemaradt a másik két modell értékétől. A görbe láthatóan nem tudja olyan mértékben megközelíteni az ideális ROC görbét, ami azt jelenti, hogy nem volt képes a modell olyan mértékben megkülönböztetni azokat az utasokat akik túléltek és azokat akik nem a kapott adatok alapján.

Összehasonlítás

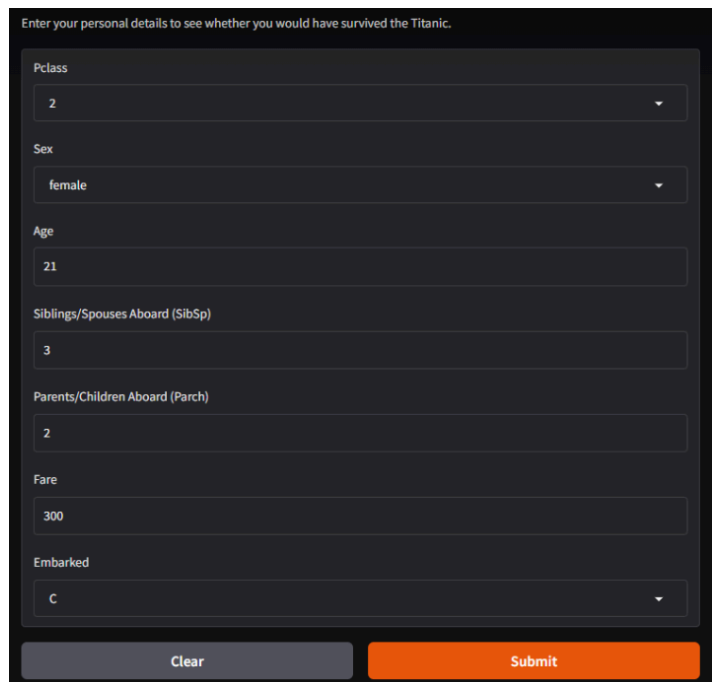
Az első modell a Random Forest, aminek a pontossága végül 0.9218. A ROC görbe közel halad el a bal felső sarokhoz, ami jó teljesítményt jelent. A második modell a Gradient Boosting, aminek pontossága 0.9441. A ROC görbe közelebb halad a bal felső sarokhoz, mint az RF esetében. A Extreme Gradient Boosting modell pontossága 0.8547, ami lényegesen elmaradt a RF és GBM eredményétől. A ROC görbéről is hasonló adatok olvashatók le, nem közelíti meg olyan mértékben az ideális modellt, mint az előző kettő.

Ezek a megoldások mellett is belátható, hogy mindegyik modell meghaladta a pontosságát a cikkekben bemutatott modelleknek, mégha a XGB modell le is maradt a többitől. A legjobb pontosságot a GBM modell érte el, sőt ez a modell tudta a legegyszerűsebben megtanulni a túlélő és nem túlélő utasok kikövetkeztetését az adatok alapján.

Applikáció

Egy Gradio alapú webalkalmazást hoztunk létre, amely a felhasználó által megadott bemenetek alapján a túlélés esélyét becsüli meg egy gépi tanulási modell segítségével. Mivel a GBM-el értük el a legkiemelkedőbb eredményt, ezt a modellt használtuk az alkalmazásban is. A felhasználó megadhatja az alábbi adatokat:

- Pclass (hajóosztály: 1, 2, 3),
- Nem (férfi vagy nő),
- Életkor,
- Testvérek/házastársak száma a hajón (SibSp),
- Szülők/gyermekek száma a hajón (Parch),
- Jegy ára (Fare),
- Beszállási kikötő (C, Q, S).

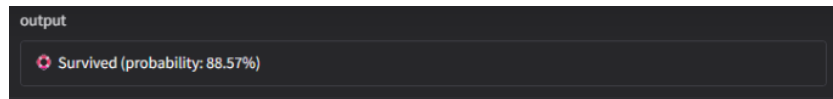


The screenshot shows a web application titled "Enter your personal details to see whether you would have survived the Titanic." The form contains several input fields: "Pclass" (a dropdown menu with "2" selected), "Sex" (a dropdown menu with "female" selected), "Age" (a text input field with "21"), "Siblings/Spouses Aboard (SibSp)" (a text input field with "3"), "Parents/Children Aboard (Parch)" (a text input field with "2"), "Fare" (a text input field with "300"), and "Embarked" (a dropdown menu with "C" selected). At the bottom of the form, there are two buttons: "Clear" and "Submit".

Ezek alapján a modell kiszámítja:

- Hogy az illető túlélte-e vagy sem a Titanic katasztrófáját,
- A túlélés valószínűségét százalékosan.

A végeredmény egy szöveges üzenet, amely megjelenik a Gradio felületén.



Irodalomjegyzék

- [1] <https://wepub.org/index.php/TCSISR/article/view/2428/2645>
- [2] A. Dasgupta, V. P. Mishra, S. Jha, et al. Predicting the Likelihood of Survival of Titanic's Passengers by Machine Learning, 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE). IEEE, 2021pp: 52-57
- [3] <https://triangleinequality.wordpress.com/2013/09/08/basic-feature-engineering-with-the-titanic-data/>
- [4] Analyzing Titanic Disaster using Machine Learning Algorithms Dr. Prabha Shreeraj Nair Dean Research, Tulsiramji Gayakwade Patil College of Engineering and Technology, Nagpu