**Depistarea factorilor care au influențat asupra mortalității pe Titanic**

¹Universitatea Tehnică a Republicii Moldova, Facultatea Calculatoare Informatică și Microelectronică, Departamentul Informatică și Ingineria Sistemelor, Informatica Aplicată

**Abstract**

Interacțiunile sociale joacă un rol important în comportamentul uman și în statutul social. Titanicul a fost exemplul perfect, deoarece pasagerii aparțineau unor grupuri cu venituri mari, medii și mici. Este interesant de văzut cum factorii sociali au influențat cine va supraviețui. Datele au fost colectate de pe site-ul "Kaggle.com", iar algoritmii de învățare automată au fost aplicați după efectuarea unei analize exploratorii și vizuale. Ipoteza conform căreia femeile și copiii au fost salvați (care a devenit celebră după filmul Titanic (1975) al lui Steven Spielberg) a fost testată cu ajutorul algoritmului random forest, precum și ipoteza conform căreia densitatea familiei a jucat un rol major în supraviețuire. Rezultatele au arătat că titlul și sexul au fost cei mai importanți factori care influențează dacă pasagerul va supraviețui.

Incidentul Titanic i-a determinat pe oamenii de știință și anchetatori să înțeleagă ce ar fi putut determina supraviețuirea câtorva călători și moartea celorlalți. Mulți algoritmi de învățare automată au contribuit la prezicerea ratei de supraviețuire a pasagerilor. Pe lângă aceasta, a fost folosit un set de date de 891 de rânduri care include atributele și anume Vârstă, ID pasager, Sex, Nume, Imbarcat, Tarif etc. În această lucrare, supraviețuirea pasagerilor este stabilită folosind diferite tehnici de învățare automată, și anume arborele de decizie, regresia logistică, SVM liniar și multe alte tehnici. Obiectivul principal al acestei lucrări este de a analiza rata de supraviețuire a călătorului pe baza preciziei.

**Introducere**

Dezvoltarea tehnologiei a adus beneficii stilului nostru de viață. Unul dintre avantajele adăugate de tehnologie este că o categorie largă de date poate fi solicitată cu ușurință. Cu toate acestea, nu este atât de ușor fezabil să acumulați statisticile adecvate. Înregistrările brute și simple care sunt accesate fără efort din sursele de internet nu au sens și trebuie procesate pentru a servi unui sistem de recuperare a datelor. În acest sens, metodele de extragere a caracteristicilor și algoritmii de învățare automată a sistemului îndeplinesc o funcție esențială în această metodă.

Scopul studiului lor este de a obține rezultate cât mai sigure și cât mai eficient posibil din înregistrările brute lipsă prin utilizarea algoritmilor de învățare automată și a tehnicilor de extragere a caracteristicilor. Prin urmare, se folosește unul dintre cele mai faimoase seturi de date din data mining, Titanic. Acest set de date reprezintă diverse funcții ale pasagerilor de pe Titanic, care include cine a supraviețuit și cine nu. Este de departe realizat că unele funcții lipsă și necorelate au scăzut performanța predicției. Pentru o evaluare mai detaliată a înregistrărilor, a fost investigat efectul diferitelor caracteristici. În consecință, unele funcții noi sunt introduse în setul de date și unele capabilități existente sunt eliminate din setul de date.

S-au implementat statistici descriptive folosind funcția summary() pentru a obține un rezultat al setului de date care include minimul, mediana, media etc. S-a utilizat biblioteca ggplot2 pentru a crea diagrame cu bare, pentru a vizualiza relația dintre variabile și rata de supraviețuire.

A avut loc și transformarea unelor variabile din setul de date. De exemplu: extragerea titlului din coloana Name și crearea unei noi coloane Title. La fel s-a creat o coloană FamilySize adăugând coloanele SibSp și Parch și o caracteristică Title din Name. Sunt date numerice precum PassengerId, Age, SibSp, Parch, Fare. Sunt date categoriale precum Survived, Pclass, Name, Ticket, Cabin, Embarked.

Analiza computațională bazată pe date devine din ce în ce mai importantă în domeniul biomedical cercetare, deoarece cantitatea de date generată continuă să crească. Cu toate acestea, lipsa de practici de partajare a rezultatelor cercetării, cum ar fi date, cod sursă și metode, efecte

transparența și reproductibilitatea studiilor, care sunt esențiale pentru dezvoltarea ştiinţă.

**Date și statistici sumare**

Depistarea factorilor care au influențat asupra mortalității pe Titanic este o temă de cercetare importantă. Am utilizat setul de date Titanic, care oferă informații despre pasagerii de pe Titanic. Fiecare intrare din setul de date reprezintă un pasager și conține informații precum clasa biletului, sexul, vârsta, numărul de frați/surori sau soți/soții la bord, numărul de părinți/copii la bord, tariful biletului, cabina și portul de îmbarcare.

Variabila noastră de răspuns, supraviețuirea, este marcată cu 1 pentru supraviețuire și cu 0 pentru deces. Din eșantion, aproximativ 32% dintre pasageri au supraviețuit. Mai multe variabile pot fi asociate cu supraviețuirea, inclusiv clasa biletului, sexul, vârsta și tariful biletului. Pentru a înțelege ce factori sunt legați de supraviețuire, este crucial să analizăm aceste date. Potrivit datelor, rata de supraviețuire a scăzut odată cu scăderea clasei biletului. De asemenea, majoritatea copiilor care au murit erau în clasa a treia. Supraviețuirea nu a fost doar bazată pe sex sau vârstă, ci și pe clasă. Pentru variabilele cu valori lipsă, cum ar fi vârsta, am înlocuit aceste valori lipsă cu valorile lor mediane corespunzătoare. Datele au fost analizate folosind programul statistic R versiunea 4.0.5.

Această analiză ne ajută să înțelegem mai bine factorii care au influențat supraviețuirea pe Titanic și să tragem concluzii importante despre impactul statutului socio-economic asupra supraviețuirii în situații de urgență.

**Metode**

Regresia liniară este o tehnică statistică utilizată pentru a modela relația dintre o variabilă dependentă (sau de răspuns) și una sau mai multe variabile independente (sau predictoare). Aceasta este o abordare liniară, ceea ce înseamnă că relațiile sunt modelate folosind funcții liniare ale căror parametri necunoscuți sunt estimați din date.

Există două tipuri principale de regresie liniară precum regresia liniară simplă stabilește relația dintre două variabile folosind o linie dreaptă. Regresia liniară multiplă stabilește relația dintre o variabilă dependentă și mai multe variabile independente.

Regresia liniară este utilizată în multe domenii, inclusiv în economie, biologie și învățare automată, pentru a numi doar câteva. Este adesea utilizată pentru a face predicții sau pentru a determina relațiile dintre variabile. De exemplu, în contextul setului de date Titanic, am utilizat regresia liniară pentru a prezice șansele de supraviețuire ale unui pasager pe baza caracteristicilor sale, cum ar fi vârsta, sexul și clasa biletului.

Putem estima eroarea de testare folosind metodele de validare a setului. Am calculat eroarea de validare încrucișată și eroarea setului de validare pentru fiecare model luat în considerare, apoi am ales modelul cu cea mai mică eroare de testare ca fiind specificația noastră preferată.

Această metodă poate fi aplicată și în alte situații de selecție a modelelor, mai ales atunci când numărul de covariante din model este dificil de estimat sau când varianța erorii σ^2 este complexă. Abordarea setului de validare: Pentru a evalua rata de eroare de testare asociată cu o

anumită metodă pe un set de eșantioane, am folosit abordarea setului de validare. Aceasta

constă în împărțirea aliatoare a eșantioanelor disponibile într-un set de instruire și un set de

validare. Modelul este antrenat folosind setul de instruire, iar apoi este utilizat pentru a prezice

răspunsurile din setul de validare. Rata de eroare a setului de validare este evaluată în mod

obișnuit utilizând eroarea medie pătratică (MSE).

Table 1 prezintă variabilele, minimul, primul cvartil, mediana, media, al treilea cvartil, maximul și numărul de valori lipsă pentru fiecare grup (“Boys” și “Miss”).

**Table 1.** Descriere statistică

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variabila | Minim | 1st | Mediana | Medie | 3rd | Maxim | Valori lipsă |
| Boys | 0.330 | 2.00 | 4.00 | 5.483 | 9.00 | 14.50 | 8 |
| Miss | 0.17 | 15.00 | 22.00 | 21.77 | 30.00 | 63.00 | 50 |

**Rezultate**

Figura 1 reprezintă un raport de clasificare pentru un model de predicție. Iată ce înseamnă fiecare termen:

**Precision** reprezintă curatețea pozitivă predictivă, adică proporția de identificări pozitive corecte (de exemplu, cât de des modelul a prezis corect că un pasager va supraviețui).

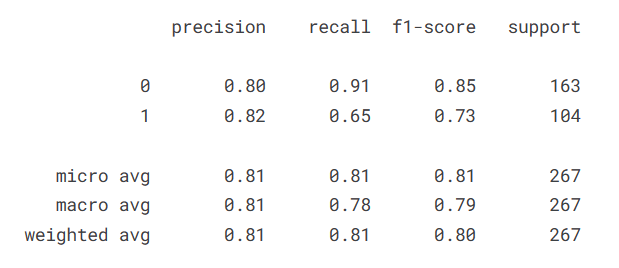
**Recall** reprezintă sensibilitatea sau rata de adevărat pozitiv, adică proporția de pozitive reale pe care modelul le identifică corect (de exemplu, cât de multe dintre supraviețuitorii reali au fost prezise corect de model).

**F1-score** este media armonică a preciziei și recall-ului. O valoare mai mare indică o performanță mai bună a modelului.

**Support** reprezintă numărul de eșantioane de răspuns adevărat care se află în acea clasă.

În acest caz, avem două clase: 0 (nu a supraviețuit) și 1 (a supraviețuit). Deci, pentru clasa 0 (nu a supraviețuit), modelul are o precizie de 0.80, un recall de 0.91 și un F1-score de 0.85, cu un suport de 163 de eșantioane. Pentru clasa 1 (a supraviețuit), modelul are o precizie de 0.82, un recall de 0.65 și un F1-score de 0.73, cu un suport de 104 de eșantioane.

Media ponderată (weighted avg) ia în considerare suportul (numărul de eșantioane pentru fiecare clasă) atunci când calculează media metricilor. Media micro (micro avg) calculează media globală a metricilor, indiferent de distribuția claselor. Media macro (macro avg) calculează media neponderată a metricilor pentru fiecare clasă.

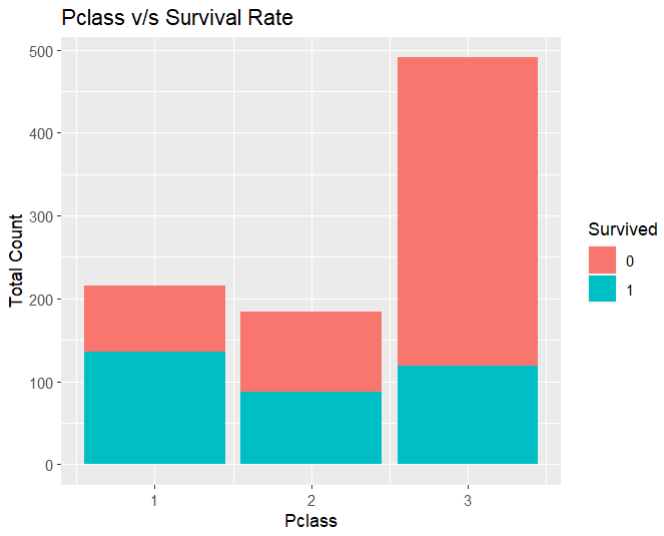


**Figura 1.** Raport de clasificare pentu un model de predicție

Figura pe care o vedeți mai jos este un grafic de bare care arată rata de supraviețuire în funcție de clasa de călătorie (Pclass) pe Titanic. Există trei clase, 1, 2 și 3. Bara roșie indică numărul de persoane care nu au supraviețuit, iar bara albastru-verzuie indică numărul celor care au supraviețuit.

Din acest grafic, putem observa că rata de supraviețuire scade odată cu creșterea numărului clasei. Clasa 1 are cea mai mare rata de supraviețuire, cu mai mult de jumătate dintre pasageri supraviețuind. Clasa 3 are cea mai mică rata de supraviețuire, cu o mică parte a pasagerilor supraviețuind în comparație cu cei care care nu au supraviețuit.

Figura 2 sugerează că clasa de călătorie a avut un impact semnificativ asupra șanselor de supraviețuire ale pasagerilor pe Titanic. Acest lucru poate indica faptul că pasagerii din clasele superioare au avut un acces mai bun la bărcile de salvare sau au primit prioritate în caz de urgență. Rezultatele arată că rata de supraviețuire în clasa 1 este mai mare decât rata de supraviețuire în clasa 2 și 3.

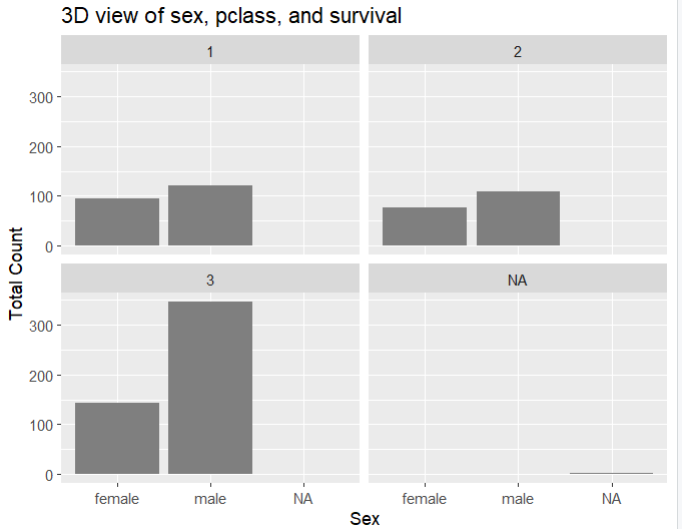


**Figura 2.** Analiza exploratorie

Graficele de mai jos (Figura 3) reprezintă datele privind sexul, clasa de călătorie (Pclass) și rata de supraviețuire pe Titanic. Există patru grafice de bare separate în imagine, fiecare corespunzând unei clase diferite: 1, 2, 3 și NA. Fiecare bară este împărțită în două culori: roșu indică numărul de persoane care nu au supraviețuit, iar albastru-verzui indică numărul celor care au supraviețuit.

Din acest grafic, putem observa că rata de supraviețuire variază în funcție de sex și clasa de călătorie. De exemplu, în clasa 1, rata de supraviețuire a femeilor este mult mai mare decât cea a bărbaților. În schimb, în clasa 3, rata de supraviețuire este mult mai mică pentru ambele sexe, dar în special pentru bărbați.

Această figură sugerează că atât sexul, cât și clasa de călătorie au avut un impact semnificativ asupra șanselor de supraviețuire ale pasagerilor pe Titanic. Acest lucru poate indica faptul că femeile și pasagerii din clasele superioare au avut un acces mai bun la bărcile de salvare sau au primit prioritate în caz de urgență. Această schemă oferă o perspectivă vizuală asupra modului în care diferiți factori, cum ar fi sexul și clasa de călătorie, au influențat rata de supraviețuire pe Titanic.



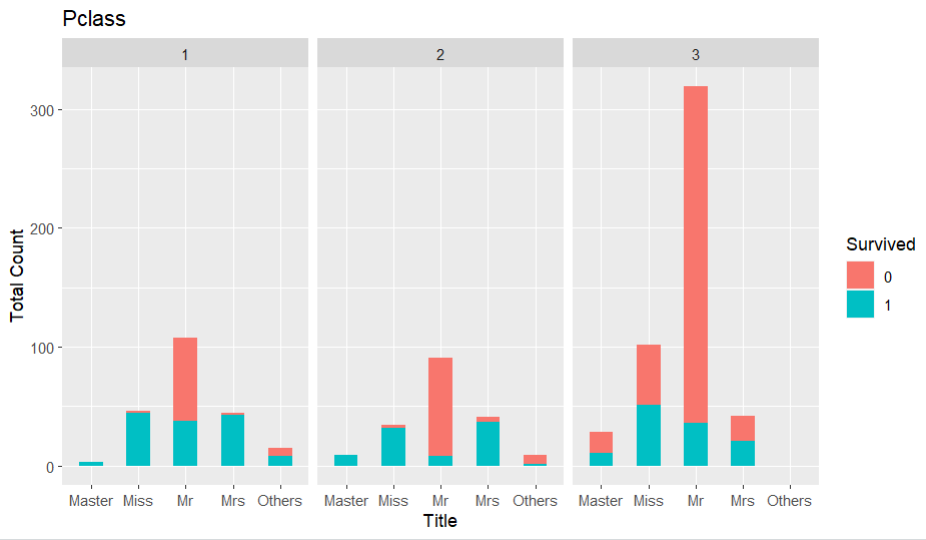
**Figura 3.** Reprezentarea relației în trei direcții dintre sex, pclass și supraviețuire

În Figura 4 este reprezentată o diagramă de bare grupate care reprezintă numărul total de persoane, categorizate după titlurile lor (Master, Miss, Mr, Mrs, Others) și clasa (1, 2, 3), care au supraviețuit sau nu. Starea de supraviețuire este indicată prin culoare: roșu pentru cei care nu au supraviețuit (0) și albastru pentru cei care au supraviețuit (1).

Din acest grafic, putem observa că rata de supraviețuire variază în funcție de titlu și clasa de călătorie. De exemplu, în clasa 1, rata de supraviețuire a persoanelor cu titlul "Master" și "Miss" este mai mare decât cea a persoanelor cu titlul "Mr" și "Mrs". În schimb, în clasa 3, rata de supraviețuire este mult mai mică pentru toate titlurile, dar în special pentru "Mr".

Această figură sugerează că atât titlul, cât și clasa de călătorie au avut un impact semnificativ asupra șanselor de supraviețuire ale pasagerilor pe Titanic. Acest lucru poate indica faptul că copiii, femeile și pasagerii din clasele superioare au avut un acces mai bun la bărcile de salvare sau au primit prioritate în caz de urgență.

Această schemă oferă o perspectivă vizuală asupra modului în care diferiți factori, cum ar fi titlul și clasa de călătorie, au influențat rata de supraviețuire pe Titanic.



**Figura 4.** Diagrama de bare grupate cu numărul total de persoane

Boxplot-ul, cunoscut și sub numele de diagramă cu mustăți, este o reprezentare grafică care rezumă distribuția unui set de date. Iată ce reprezintă fiecare componentă a boxplot-ului:

**Cutia** reprezintă intervalul interquartil (IQR), care conține 50% din date. Linia de jos a cutiei indică primul quartil (Q1), linia de sus indică al treilea quartil (Q3), iar linia din mijlocul cutiei reprezintă mediana (Q2).

**Mustățile r**eprezintă datele din afara IQR. Mustățile se întind de la Q1 și Q3 până la valoarea minimă și maximă a datelor care nu sunt considerate outlieri.

**Outlierii** reprezentați prin puncte individuale deasupra sau dedesubtul mustăților, acestea sunt valori care cad în afara a 1.5 \* IQR deasupra Q3 sau dedesubtul Q1.

În acest boxplot specific:

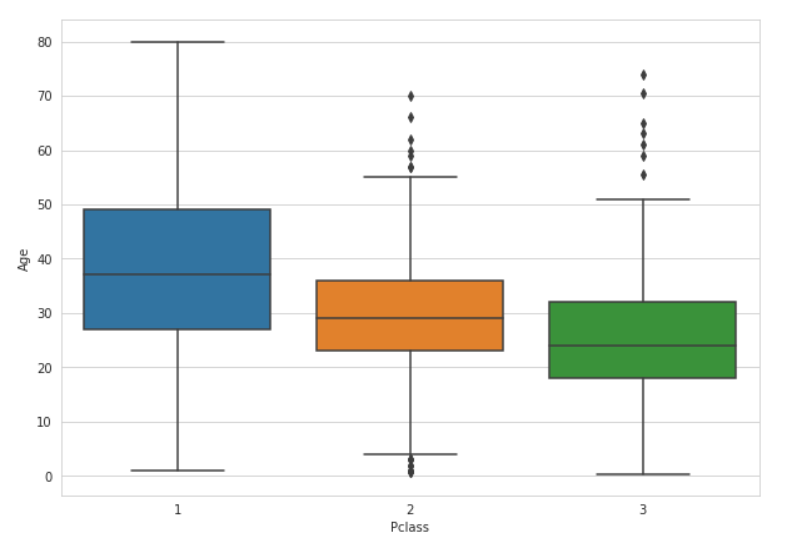
Axa x reprezintă diferite clase etichetate ca 1, 2 și 3.

Axa y reprezintă vârsta.

Fiecare clasă are o cutie colorată care reprezintă IQR, conținând 50% din date.

Linia din interiorul fiecărei cutii indică mediana vârstei pentru fiecare clasă.

Outlierii sunt reprezentați prin puncte negre deasupra și dedesubtul mustăților.



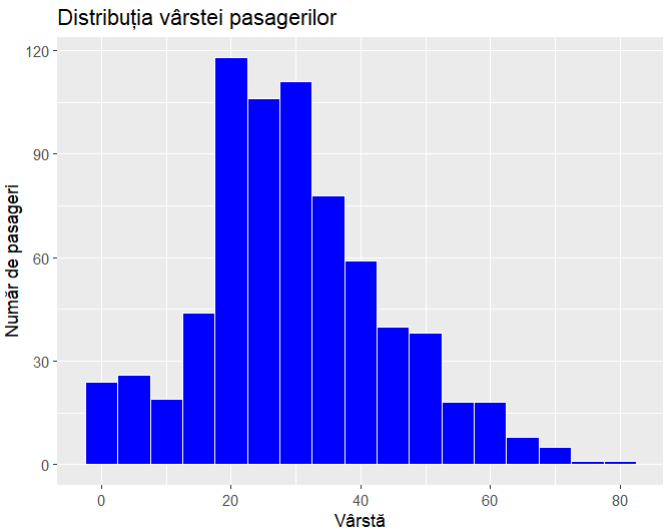
**Figura 5.** Distribuția datelor reprezentată prin boxplot

Figura 6 reprezintă distribuția vârstei pasagerilor pe Titanic. Este un grafic de bare cu vârsta pe axa x și numărul de pasageri pe axa y. Majoritatea pasagerilor sunt între 20 și 30 de ani.

Din acest grafic, putem observa că majoritatea pasagerilor de pe Titanic erau tineri, cu vârste cuprinse între 20 și 30 de ani. Există, de asemenea, un număr semnificativ de pasageri cu vârste cuprinse între 30 și 40 de ani. Numărul de pasageri scade odată cu creșterea vârstei, cu excepția unui mic picaj pentru copiii cu vârsta sub 10 ani.

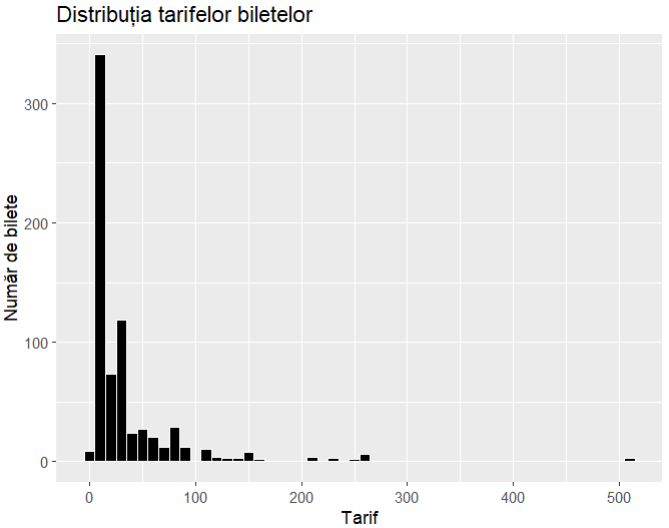
Această distribuție poate fi utilă pentru a înțelege demografia pasagerilor de pe Titanic și poate oferi informații valoroase atunci când se analizează factorii care au influențat rata de supraviețuire. De exemplu, vârsta ar putea fi un factor important de luat în considerare, deoarece copiii și tinerii ar fi putut avea mai multe șanse să supraviețuiască datorită priorității acordate la salvare.

Distribuția este utilă în multe domenii. În statistici, ne ajută să înțelegem și să prezicem comportamentul variabilelor aleatoare. În economie, ne permite să optimizăm procesele de livrare a produselor către clienți. În lingvistică, ne ajută să înțelegem structura și utilizarea limbajului. Deci, în funcție de context, distribuția poate avea diferite aplicații și beneficii.



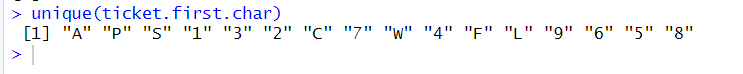
**Figura 6.** Distribuția vârstei pasagerilor

În Figura 7 este o diagramă de bare care reprezintă distribuția tarifelor biletelor. Barele negre verticale indică numărul de bilete vândute la diferite tarife. Majoritatea biletelor au un tarif sub 100, cu un vârf semnificativ în jurul valorii de 0-50. Există și un număr mic de bilete care costă între 100 și 500, dar acestea sunt mult mai puține comparativ cu biletele ieftine. Graficul sugerează că majoritatea pasagerilor au plătit tarife mici pentru biletele lor, în timp ce un număr mic de pasageri au plătit sume semnificative. Pe axa Y este etichetat "Număr de bilete", iar pe axa X este etichetat "Tarif". Majoritatea barelor sunt concentrate la începutul axei X, indicând că majoritatea biletelor au avut un preț mai mic. Există câteva bare izolate spre capătul drept al axei X, arătau că existau și bilete cu prețuri considerabil mai mari. Titlul graficului "Distribuția tarifelor biletelor" este afișat în partea de sus a imaginii.



**Figura 7.** Distribuția tarifelor biletelor

Comanda unique(ticket.first.char) este utilizată pentru a extrage valori unice din variabila “ticket.first.char”. Rezultatul acestei comenzi este afișat sub formă de vector și include litere și cifre: “A”, “P”, “S”, cifrele de la 1 la 9 și “C”. Acest lucru sugerează că variabila “ticket.first.char” conține aceste valori unice. Acest cod poate fi folosit în analiza datelor pentru a înțelege diversitatea datelor din variabila “ticket.first.char”.



**Figura 8.** Simboluri unice cu primul caracter al ticketului

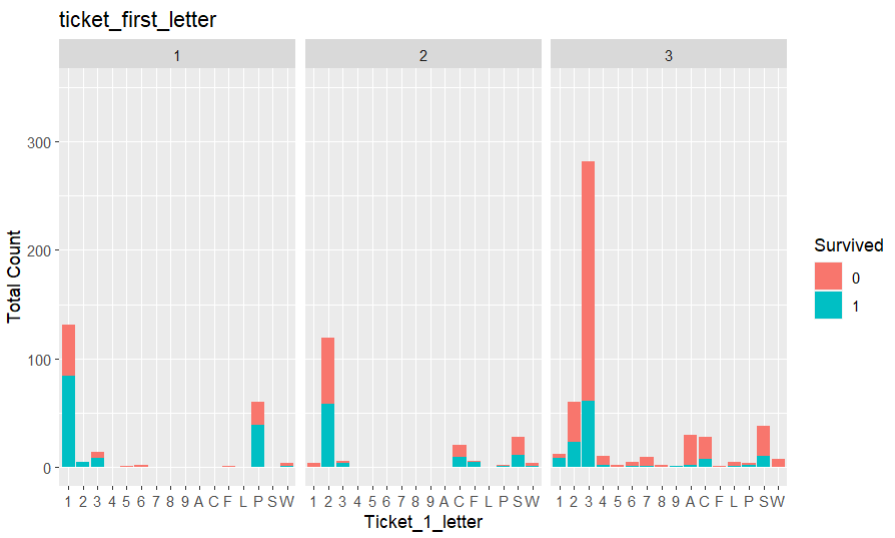
În Figura 9 este o diagramă de bare care reprezintă numărul total de persoane care au supraviețuit (indicat în albastru) și nu au supraviețuit (indicat în roșu), în funcție de prima literă a numărului lor de bilet. Există trei grafice separate etichetate ca 1, 2 și 3, fiecare arătând litere diferite și numărări corespunzătoare.

În secțiunea unu, biletul care începe cu '3' are cel mai mare număr de non-supraviețuitori.

În secțiunea doi, biletul care începe cu 'C' are un număr notabil de non-supraviețuitori comparativ cu supraviețuitorii.

În secțiunea trei, biletul care începe cu ‘A’ are un număr covârșitor de non-supraviețuitori comparativ cu supraviețuitorii.

Aceste date sugerează că prima literă a numărului de bilet poate avea o anumită corelație cu rata de supraviețuire, deși ar fi necesară o analiză suplimentară pentru a determina semnificația statistică a acestor rezultate.



**Figura 9.** Distribuția supraviețuirii

Random Forest este un algoritm de clasificare care a fost dezvoltat de Breiman și Cutler. Este un algoritm care se încadrează în categoria învățării supravegheate, care este flexibil, ușor de utilizat în învățarea automată. Este un algoritm utilizat pe scară largă datorită simplității sale, alături de utilizarea atât cu regresie, cât și cu clasificare. Pădurea pe care o construiește acest algoritm este un ansamblu de Decision Tress care este antrenat în cea mai mare parte folosind o metodă de pungi. Conceptul din spatele utilizării unei metode în pungi este că combinarea mai multor algoritmi se îmbunătățește este că învățarea crește, rezultând o creștere generală a performanței. Pădurea aleatorie adaugă mai multă aleatorie modelului în timp ce se află în stadiul de creștere.

Spre deosebire de alte modele, acest model caută cele mai bune caracteristică în subsetul aleatoriu de date, altele decât caracteristicile importante. Marea preocupare în utilizarea acestui algoritm este modul în care tinde să încetinească și să fie ineficient atunci când există un număr mare de trees.



**Figura 10.** Model ce prezice supraviețuirea

Acest model de Random Forest este creat pentru a prezice supraviețuirea pe Titanic, luând în considerare clasa pasagerului (`Pclass`), titlul (`title`) și numărul de frați/surori sau soți/soții prezenți pe navă (`SibSp`). Iată ce face fiecare linie de cod:

`set.seed(1234)`: Aceasta setează semința generatorului de numere aleatorii la 1234, asigurând că rezultatele sunt reproductibile.

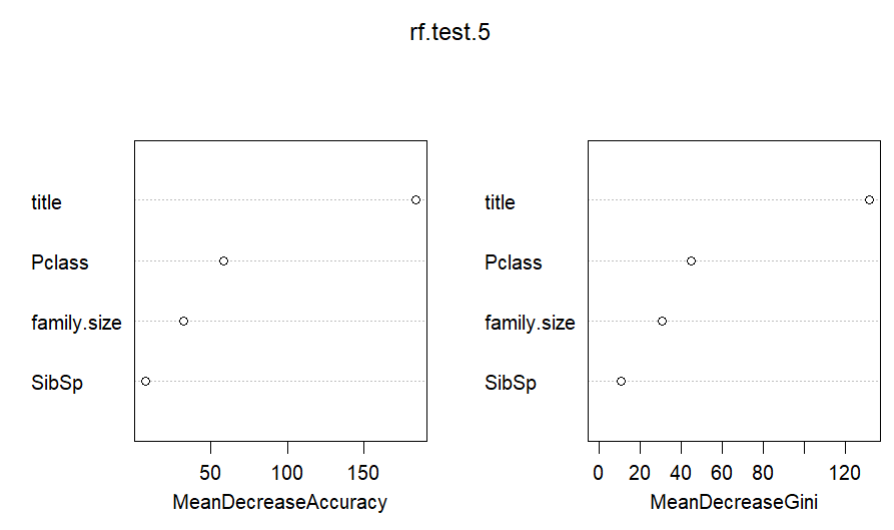
`rf.2=data.combined[1:891,c("Pclass","title","SibSp")]`: Aceasta creează un subset de date din `data.combined`, selectând doar primele 891 de rânduri (care corespund setului de date de antrenament) și coloanele `Pclass`, `title` și `SibSp`.

`rf.test.2=randomForest(x=rf.2,y=rf.res,importance=TRUE,ntree = 1000)`: Aceasta construiește modelul de pădure aleatoare, folosind `rf.2` ca set de date de intrare și `rf.res` ca răspunsuri. Parametrul `importance=TRUE` indică faptul că importanța variabilelor trebuie calculată, iar `ntree = 1000` specifică numărul de arbori din pădurea aleatoare.

`rf.test.2`: Aceasta afișează rezultatele modelului de pădure aleatoare.

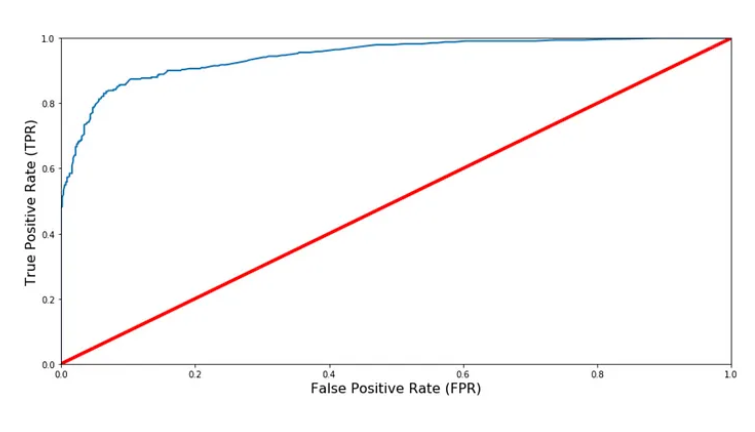
`varImpPlot(rf.test.2)`: Aceasta generează un grafic al importanței variabilelor în modelul de pădure aleatoare.

În concluzie, acest program este folosit pentru a analiza care dintre variabilele `Pclass`, `title` și `SibSp` sunt cele mai importante pentru a prezice supraviețuirea pe Titanic. Acesta este un exemplu de analiză predictivă folosind datele Titanic.



**Figura 11.**Model Random Forest

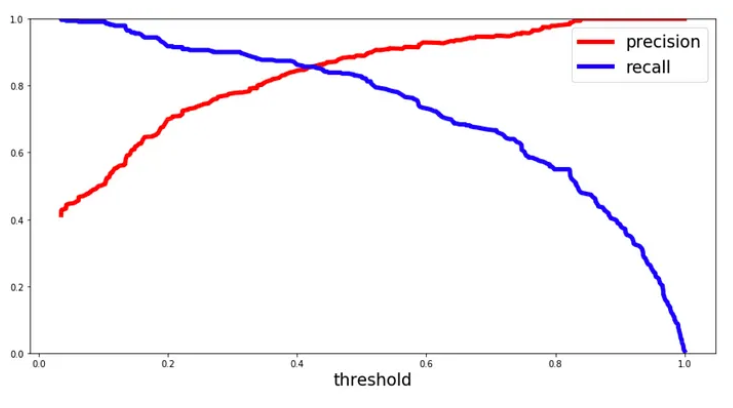
O altă modalitate de a evalua și compara clasificatorul dvs. binar este oferită de curba ROC AUC. Această curbă grafică rata pozitivă adevărată (numită și reamintire) față de rata fals pozitivă (raportul cazurilor negative clasificate incorect), în loc să grafice precizia față de reamintire.



**Figura 12.** Modelul Random Forest

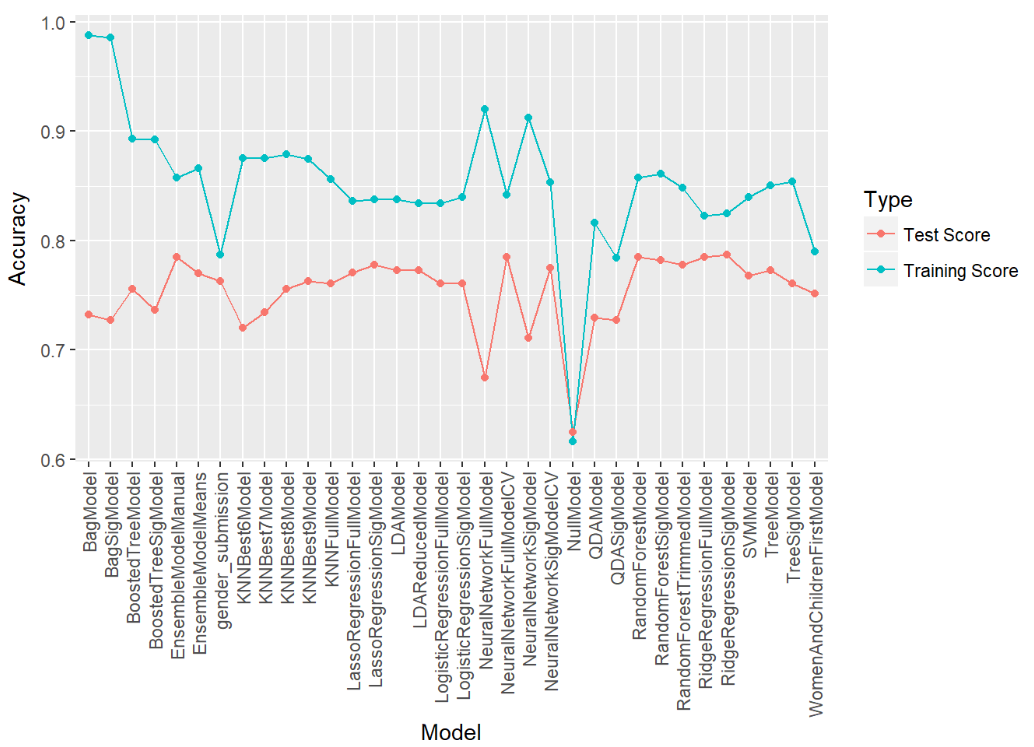
Linia roșie din mijloc reprezintă un clasificator pur aleatoriu (de exemplu, o monedă) și, prin urmare, clasificatorul dvs. ar trebui să fie cât mai departe de acesta. Modelul nostru Random Forest pare să facă o treabă bună. Desigur, avem și aici un compromis, deoarece clasificatorul produce mai multe false pozitive, cu atât rata pozitivă adevărată este mai mare.

Pentru fiecare persoană pe care algoritmul Random Forest trebuie să o clasifice, calculează o probabilitate pe baza unei funcții și clasifică persoana ca supraviețuită (când scorul este mai mare decât pragul decât) sau ca nesupraviețuit (când scorul este mai mic decât pragul) . De aceea, pragul joacă un rol important.



**Figura 13.** Precision și Recall

Mai sus puteți vedea clar că rechemarea scade rapid la o precizie de aproximativ 85%. Din acest motiv, dacă dorim să selectăm compromisul precizie/rechemare înainte de aceasta - poate la aproximativ 75 %. Acum se poate alege un prag, care oferă cel mai bun compromis de precizie/rechemare pentru problema aceasta.



**Figura 14.** Modelul

Să presupunem că toate aceste rezultate de antrenament au fost mai mult sau mai puțin egale - unele la o acuratețe de 0,84 setului de antrenament validat încrucișat, altele la o precizie de 0,83 setului de antrenament validat încrucișat și unul la 0,85 acuratețe a setului de antrenament validat încrucișat. Care ar fi atunci criteriul de alegere a unui model? La prima vedere, ar putea fi evident să alegeți cel care are cea mai mare acuratețe a setului de antrenament validat încrucișat de 0,85. Dar dacă modelul respectiv este foarte complex, cu multă parametrizare și există un model mult mai simplu care a marcat 0,84. Atunci ar fi mai bine să alegeți modelul mai simplu, dar cu scoruri mai mici? Este diferența dintre 0,84 și 0,85 semnificativă statistic sau diferența se datorează doar unei șanse aleatorii?

Dacă ar fi să aleg cel mai „cel mai bun” model dintre acestea, probabil că ar fi modelul de regresie a crestei. Penalizarea de regularizare la pătrat ajută la reducerea coeficienților unora dintre predictori mai bine decât modelul similar de regresie lazo.

**Discuții**

În cazul în care rata de eroare globală scade sub 20%, modelul este mai bine pregătit să prezică decesul (linia roșie) decât supraviețuirea (linia verde).

Învățarea automată este un proces care ne ajută să ne apropiem de o nouă etapă în informatică, iar aceasta este o abstractizare. În această lucrare, au fost utilizate două abordări de învățare automată pentru a găsi factorii determinanți care au jucat un rol semnificativ în prezicerea supraviețuirii pasagerilor. Deoarece variabilele prezente în setul de date au fost legate de clasificările sociale ale pasagerilor, domeniul de aplicare al studiului este mai degrabă sociologic decât tehnic. Cu toate acestea, algoritmul ne oferă dovezi solide că titlul, sexul și tariful au fost primele trei variabile care au decis soarta pasagerilor.

Regulamentele British Board of Trade stipulau că Titanicul putea naviga fără a avea suficiente locuri pentru fiecare pasager; astfel, lipsa bărcilor de salvare de pe Titanic a fost principalul motiv pentru care au existat doar câțiva supraviețuitori. Nava avea la bord douăzeci de bărci, care au oferit refugiu pentru 52% din numărul total de pasageri în timpul călătoriei inaugurale a navei și pentru 30% din numărul de pasageri atunci când era plin.

Cum este modelul? Perspective

**Concluzii**

În lumea de astăzi bazată pe cunoștințe, obținerea de rezultate din date brute și neprocesate prin utilizarea tehnicilor de învățare automată și de extracție a caracteristicilor a devenit foarte importantă. În această lucrare, există trei modele propuse și anume, Random Forest, Gradient Boosting Machine și Support Vector Machine pentru prezicerea supraviețuirii pasagerilor la bord în călătoria inaugurală a Titanicului. În primul rând, a fost efectuat un studiu asupra diferitelor caracteristici ale setului de date pentru a stabili dacă acestea pot fi utilizate pentru procesul de corelare sau dacă sunt irelevante pentru modelul predictiv în curs de construire. În etapa de pre-procesare au fost excluse câteva caracteristici, cum ar fi Nume, Ticket, Cabină etc., iar valorile lipsă din celelalte caracteristici au fost completate. În cele din urmă, au fost utilizați algoritmii de învățare automată și algoritmii de clasificare pentru a prezice supraviețuirea membrilor echipajului și a pasagerilor la bord în timpul călătoriei inaugurale a Titanicului.

**Referințe**

1. Analyzing Titanic disaster using machine learning algorithms-Computing, Communication and Automation (ICCCA), 2017 International Conference on 21 December 2017, IEEE.
2. Eric Lam, Chongxuan Tang, "Titanic Machine Learning From Disaster", LamTang-Titanic Machine Learning From Disaster, 2012.
3. S. Cicoria, J. Sherlock, M. Muniswamaiah, L. Clarke, "Classification of Titanic Passenger Data and Chances of Surviving the Disaster", Proceedings of Student-Faculty Research Day CSIS, pp. 1-6, May 2014.
4. Corinna Cortes, Vlasdimir Vapnik, “Support-vector networks”, Machine Learning, Volume 20, Issue 3,pp 273-297.
5. L Breman- “random forests”, Machine Learning, 2001 Ng. CS229 Notes, Standford University, 2012.
6. SJ Russsel P Norvig-“Artificial intelligence: A modern approach”-2016.
7. Lonnie Stevans, David L. Gleicher, ”Who Survived the Titanic? A logistic regression analysis”-Article in International Journal of Maritime History, December 2004.
8. MICHAEL AARON WHITLEY, Using statistical learning to predict survival of passengers on the RMS Titanic by Michael Aaron Whitley, 2015.
9. Kunal Vyas, Zeshi Zheng, Lin Li, Titanic- Machine Learning From Disaster- 2015.
10. EECS 349 Titanic- Machine Learning From Disaster, Xiaodong Yang, Northwestern University.
11. Prediction of Survivors in Titanic Dataset: A Comparitive Study using Machine Learning Algorithms, Tryambak Chatterlee, IJERMT-2017.
12. An Introduction to Logistic Regression Analysis and Reporting by Chao-Yig Joanne Peng, Kuk Lida Lee & Gary M. Ingersoll, April 2010.
13. Zhenyan Liu, Yifei Zeng, Yida Yan, Pengfei Zhang and Yong Wang, Machine Learning for Analyzing Malware, Journal of Cyber Security and Mobility, Vol: 6 Issue: 3, July 2017.
14. Andy Liaw and Metthew Wiener, Classification and Regression by Random Forest, vol. 2/3, December 2002.
15. Galit Shmueli and Otto R. Koppius MIS Quarterly, Predictive Analytics in Information System Research, , Vol. 35, No. 3(September 2011), pp. 553-572.
16. john D. Kelleher, Brain Mac Namee, Aoife D’Arcy Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms .
17. Dr. Neeraj Bhargava, Girja Sharma, Decision Tree Analysis on J48 Algorithm for Data Mining. Volume 3, Issue 6, June 2013.
18. Data Mining: Practical Machine Learning Tools and Techniques, by Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal.
19. A Comparison of Goodness of Fit Tests for the Logistic Regression Model, D.W. Hosmer, T. Hosmer, S. Le Cessie and S. Lemeshow