UNIVERSITATEA "BABEȘ-BOLYAI" CLUJ-NAPOCA

FACULTATEA DE ȘTIINȚE ECONOMICE ȘI GESTIUNEA AFACERILOR

Specializarea: Informatică Economică

Proiect Big Data

Tema aleasă:

Clasificarea victimelor de pe Titanic

Comșa Ionela-Florentina (email: ionela.comsa@stud.ubbcluj.ro)

Kovacs Adela-Maria (email: adela.kovacs@stud.ubbcluj.ro)

I. Introducere:

Analiza care se propune a se realiza în cadrul acestui proiect, va studia diverși factori care au impactat într-o anumită măsură moartea persoanelor de pe Titanic, implicit supraviețuirea lor.

Titanic a fost un transatlantic britanic de dimensiuni foarte mari care în cadrul primei călătorii efectuate s-a lovit într-un mod tragic de un iceberg în data de 14-15 aprilie 1912. Numeroase persoane au murit în urma acestui eveniment, însă altele au avut norocul de a supraviețui.

Deși la o primă vedere pare că acei oameni au avut din pură întâmplare șansa de a supraviețui, în urma unor analize efectuate s-a ajuns la concluzia că moartea acestora a putut fi influențată de diverși factori ce țin de fiecare individ în parte. Aspecte importante de luat în considerare pe tot parcursul proiectului, pe lângă factorii contribuabili existenți au mai fost și alte informații care ar putea duce la o rată mai mare de mortalitate în cadrul bărbaților de exemplu: majoritatea femeilor și copiilor au fost prioritizați și trimiși cu bărci de salvare de pe transatlantic.

Câțiva dintre factorii despre care s-au menționat anterior, precum și posibila însemnătate a lor în acest studiu, sunt:

- vârsta, cei înaintați aveau o probabilitatea mai mică să trăiască în situații extreme precum aceasta.
- dacă a avut frați și părinți îmbarcați deoarece o astfel de persoană nu s-a prioritizat doar pe ea însăși, ci a avut grija și de aceștia.
- portul de îmbarcare, pentru care există 3 variante posibile: Queenstown, Cherbourg și Southampton. Deși această informație pare să nu aibă o anumită relevanță cu probabilitatea de supraviețuire, s-a rezultat faptul că portul poate exprima distribuția socio-economică a persoanelor, unele fiind mai înstărite, iar altele nu, fapt ce poate influența tichetul cumpărat. Privind din acest punct de vedere, oamenii cu o situație financiară mai slabă cel mai probabil au optat pentru clase ce au fost dispuse mai jos în Titanic, fapt ce ar fi putut cauza o moarte mai sigură, nivelul 3 al transatlanticului fiind primul pătruns de apă.
- clasa aleasă a reprezentat un alt factor. Așa cum s-a menționat anterior, distribuția persoanelor pe diverse nivele a avut o influență asupra șansei de supraviețuire.

Întrebările la care s-a propus a se răspunde prin intermediul diferitelor modele de clasificare create sunt următoarele:

1. Care sunt factorii contribuabili cu cea mai mare semnificație în analiza supraviețuirii persoanelor îmbarcate pe Titanic?
2. În ce măsură putem să prezicem supraviețuirea persoanelor analizând factorii dați?

Posibilele audiențe interesate de acest studiu efectuat ar putea fi în primul rând publicul în sens larg, aceasta fiind o poveste tragică și captivantă de o lungă perioadă de timp. De asemenea, această analiză poate ajuta la evaluarea riscurilor asociate cu situații extreme, subiect ce poate reprezenta un interes pentru companiile generatoare de asigurări. Acestea își pot ajusta politicile după considerente proprii, rezultate după înțelegerea factorilor producători de riscuri.

II. Setul de date:

Pentru a se putea realiza analiza tematicii alese, un următor pas a fost căutarea unui set de date care să furnizeze informații despre indivizii îmbarcați în momentul scufundării transatlanticului. Setul de date găsit care s-a potrivit cel mai bine a fost preluat de pe Kaggle și se poate găsi la adresa următoare: Set_de_date_Titanic.

Acest set de date este relevant pentru studiul propus, întrucât conține factorii influențatori despre care am menționat și la capitolul precedent. Respectivii factori sunt reprezentați ca și câmpuri diferite în setul de date astfel:

Câmp	Definiție	Însemnătate
Survived	0 = No 1 = Yes	Sugerează dacă persoana a murit sau nu.
Pclass	Biletul aferent clasei: 1 = nivel înalt 2 = nivel de mijloc 3 = nivel jos	Clasele s-au repartizat pe nivele diferite, iar persoanele care s-au aflat la nivelul de jos au avut șansă mai mică de supraviețuire, fiind primul nivel afectat de inundare.
Sex	Feminin/Masculin	Femeile au fost prioritizate și au fost trimise într-un număr mai mare cu bărcile de salvare, mecanismul de întâietate funcționând pe principiul ”femeile și copiii primii”.
Age	Vârsta în ani	Copiii au avut și ei prioritate față de vârstnici.
SibSp	Numărul de frați aflați la bord	Persoanele care au avut frați îmbarcați au fost preocupate de salvarea acestora.
Parch	Numărul de părinți/copii aflați la bord, pe care îi are un individ: 0 = copii însoțiți de dădace 1,2,3,4 = număr de părinți/copii îmbarcați ai individului X.	Preocuparea acestor persoane este aceeași ca și în rândul persoanelor care au avut frați îmbarcați.
Ticket	Numărul aferent biletului	Poate juca rolul de identificator unic corespunzător fiecărui individ îmbarcat.
Fare	Taxa de îmbarcare	Poate sugera situația financiară a persoanei, fapt ce îi crește/scade probabilitatea de supraviețuire în funcție de clasa achitată prin respectiva taxă. Preț mai mare => clasă superioară. Preț mai mic => clasă inferioară.

Cabin	Numărul cabinei de forma: LiteraNumăr (ex: A12)	Fiecare literă reprezintă o anumită punte, iar numărul indică unde mai exact pe această punte se află cabina. Distribuția cabinelor s-a făcut de la nivelul cel mai înalt în jos, începând cu literele în ordine alfabetică. Persoanele care se aflau în cabinele notate cu primele litere din alfabet, se aflau la nivele mai înalte, unde de asemenea se aflau și bărcile de salvare.
Embarked	Portul de unde s-au îmbarcat: C = Cherbour Q = Queenstown S = Southampton	Fiind analizat împreună cu alți factori poate sugera caracteristici socio-economice.

Tabel 1. Câmpuri și însemnătatea lor

În Tabelul 1 se prezintă atributele de dinainte de a fi prelucrate. Deoarece s-a constatat că unele câmpuri influențau în mod negativ realizarea de modele de clasificare, s-au făcut următoarele modificări:

- întrucât existau câteva date lipsă în cadrul coloanei Age, s-a calculat o medie rotunjită a valorilor existente, iar valorile nule au fost înlocuite cu rezultatul respectiv.
- datele din cadrul câmpului Cabin erau insuficiente. Ca prima opțiune, s-a încercat să se completeze informațiile cu date extrase din resurse online. Cu toate acestea, după ce s-a analizat mai amănunțit problema s-a găsit informația că datele nu există în mod implicit, deoarece multe dintre acestea s-au pierdut la momentul scufundării Titanicului. Prin urmare, s-a ales să se excludă acest atribut, având oricum numărul clasei ca și câmp, câmp ce sugerează în principiu același lucru ca și cabina (a se consulta Tabel 1 pentru însemnătatea atributelor).
- Tichetul, fiind pe post doar de identificator unic, s-a ales să fie eliminat.
- Ca variabila țintă, anume Survived, să fie mai sugestivă, i s-au substituit valorile astfel: 1 =Yes, 0 =No.

Toate aceste modificări s-au efectuat cu ajutorul limbajului R. Formatul setului de date final pe seama căruia s-au făcut modelele de clasificare se poate observa la Figura 1.

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
No	3	male	22.00	1	0	7.2500	S
Yes	1	female	38.00	1	0	71.2833	C
Yes	3	female	26.00	0	0	7.9250	S
Yes	1	female	35.00	1	0	53.1000	S
No	3	male	35.00	0	0	8.0500	S
No	3	male	30.00	0	0	8.4583	Q

Figura 1. Format set de date final

III. Rezultate și discuții:

În cadrul acestui capitol se propune detalierea diverselor metode de clasificare folosite în scopul obținerii de răspunsuri la întrebările stabilite.

Clasificarea este o problemă de predicție unde variabila dependentă este calitativă, fapt ce implică atribuirea fiecărei instanțe într-o clasă. În contextul analizei alese, variabila țintă se numește Survived, iar clasele în care se propune a se asigna respectivele instanțe sunt: Yes (a supraviețuit) și No (nu a supraviețuit).

Înainte de a se trece mai departe la utilizarea metodelor, s-a ales să se facă o analiză asupra datelor numerice pentru a putea observa care dintre acestea ar trebui transformate în variabile de tip factor.

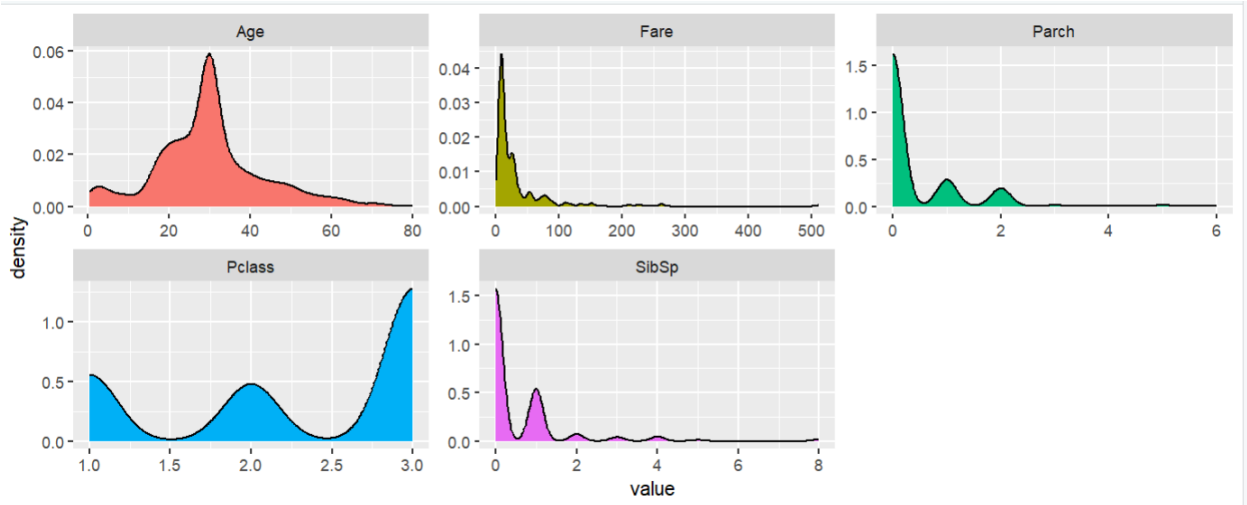


Figura 2. Analiza datelor numerice

În urma vizualizării acestor grafice, s-a constatat faptul că Pclass are o distribuție mai anormală. Acest lucru sugerează faptul că atributul este de tip feature, însă cu toate acestea, el este reprezentat sub formă numerică. Prin urmare, o transformare a respectivei date în factor este decizia potrivită. Alte variabile, care nu apar în Figura 2, dar care au fost transformate în factori sunt: variabila țintă Survived, Embarked și Sex.

În continuare, s-au realizat alte 2 grafice pentru a vizualiza matrici destinate evidențierii corelației dintre atributele numerice. Primul grafic reprezintă corelația dintre variabilele care au fost identificate ca aparținând de clasa Yes (a supraviețuit), iar cel de al doilea grafic arată același lucru doar că pentru clasa No (nu a supraviețuit).

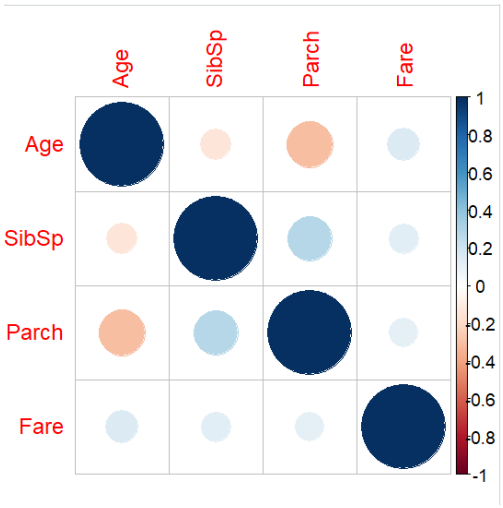


Figura 3. Corelație -Yes

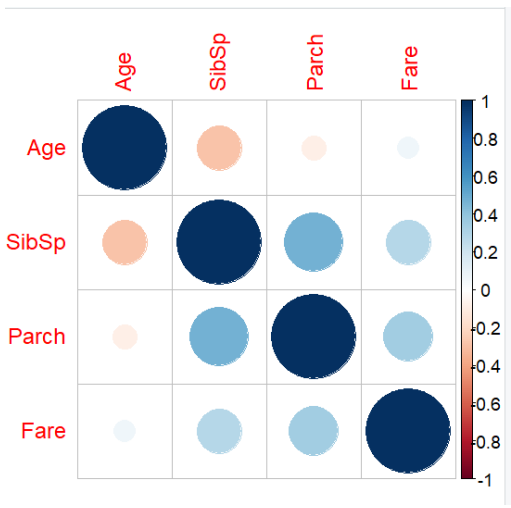


Figura 4. Corelație - No

Diagonala din aceste grafice va fi întodeauna 1, ceea ce este și firesc. În cadrul graficului de la Figura 3 observăm o corelație pozitivă mai slabă între Parch și SibSp fată de corelația acelorași factori din Figura 4. Acest lucru ar putea sugera faptul că în rândul persoanelor care au supraviețuit nu s-a pus un accent mare pe existența altor membri de familie aflați la bord, pe când în cazul celor care au murit, acest aspect a fost unul important. Această observație a fost menționată cu scop demonstrativ, motiv pentru care nu s-au detaliat și alte observații desprinse din grafice.

În continuare, pentru a se putea trece la utilizarea modelelor de clasificare, folosind tehnica hold-out, setul de date a fost împărțit în 2 subseturi: set de antrenare 70% și set de validare 30%.

Ca și obiectiv principal, se dorește ca modelele folosite să aibă putere de generalizare și să se obțină rezultate bune și pe setul de testare. În caz contrar, va apărea problema overfitting-ului.

1. Naive Bayes:

Înainte de a detalia modele create se vor explica câteva metrice luate în considerare pentru evaluarea acestora:

- Acuratețea este reprezentată de un procent ce sugerează cât la sută din predicțiile făcute sunt corecte.
- Specificitatea face referire la clasa minoritară. Odată cu studierea modelelor se va observa ca respectiva clasă este reprezentată de categoria persoanelor care au supraviețuit evenimetului. De asemenea, specificitatea ne indică în ce măsură din predicțiile pentru clasa Yes, sunt de fapt reale și corecte.
- Senzitivitatea face referire la clasa majoritară, anume clasa No sau clasa personalor care nu au supraviețuit. Aceasta ne indică câți din clasa majoritară au fost prezișI în mod corect.
- P-value, în cazul în care este reprezentat de un număr mic, sugerează faptul că între predictor și variabila dependentă există o corelație.

Naive Bayes constă în faptul că încearcă să genereze o probabilitate de apartenență la o clasă, folosind teorema lui Bayes. Acesta funcționează după următoarul principiu: se generează 2

probabilități, una pentru clasa Yes și una pentru clasa No. Se compară între ele, iar valoarea mai mare va sugera clasa de apartenență. Un aspect de ținut cont, este că această metodă este ”naivă”, întrucât consideră toate variabilele independente statistic.

Pentru construirea modelului s-a procedat astfel: toate variabilele înafară de Survived au fost setate ca variabile independente, iar Survived a fost setat ca și variabila dependentă. Folosind metoda 10-folds Cross Validation (cv), subsetul a fost împărțit în 10 părți de aproximativ aceeași dimensiune, cu scopul de a se antrena modelul pe mai multe părți. Mai departe, s-a utilizat metoda Naive Bayes (nb) și s-a generat matricea de confuzie în urma căruia a rezultat o acuratețe de 78,58%.

```
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction   No  Yes
       No  56.2 15.8
       Yes  5.6 22.4

 Accuracy (average) : 0.7858
```

Figura 5. Acuratețe fără kernel

```
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction   No  Yes
       No  56.0 16.3
       Yes  5.8 21.9

 Accuracy (average) : 0.7794
```

Figura 6. Acuratețe cu kernel

Pentru a verifică dacă există posibilitatea îmbunătățirii acurateții, s-a creat un searchGrid în care s-a specificat opțiunea usekernel. După generarea matricii de confuzie acuratețea rezultată a fost 77,94%, deci a fost influențată într-un mod negativ. Prin urmare, s-a ales să nu se folosească kernel.

```
               Accuracy : 0.7799
                 95% CI : (0.7254, 0.828)
    No Information Rate : 0.6157
    P-Value [Acc > NIR] : 6.972e-09

                  Kappa : 0.5107

 Mcnemar's Test P-Value : 0.000712

            Sensitivity : 0.9030
            Specificity : 0.5825
         Pos Pred Value : 0.7760
         Neg Pred Value : 0.7895
             Prevalence : 0.6157
         Detection Rate : 0.5560
   Detection Prevalence : 0.7164
      Balanced Accuracy : 0.7428

       'Positive' Class : No
```

Figura 7. Acuratețe finală NB

În cele din urmă, s-a verificat acuratețea și pe setul de test și a rezultat un procentaj de 77,99%, cu o senzitivitate de 90,30% și o specificitate de 58,25%.

2. Arbori de decizie:

Arborii de decizie sunt algoritmi de clasificare și de regresie. În cadrul acestora fiecare nod reprezintă o decizie, iar răspunsurile care pot exista sunt ramurile arborelui. Această metodă deține un avantaj major, în sensul în care arborii sunt ușor de interpretat și conferă o claritate mare asupra distribuției variabilelor și a relevanței caracteristicilor. Totuși, în cazul în care se vorbește de un set de date cu informații incomplete sau modificări, chiar și minore ale acestora, arborii sunt sensibili și au neajunsuri în gestionarea respectivelor situații.

Pentru realizarea arborilor de decizie s-au folosit tree și rpart după cum urmează: rpart pentru arborii simpli din secțiunile 2.1 și 2.2 și tree pentru 2.3 și 2.4.

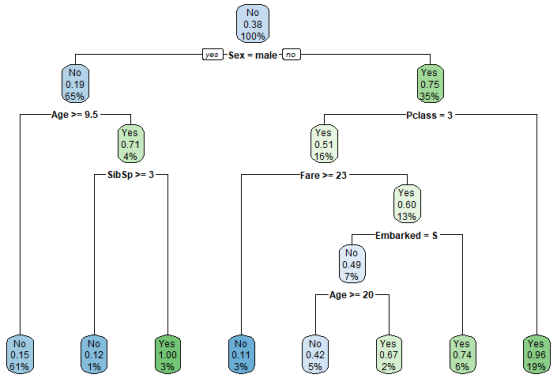
Librăria rpart presupune folosirea erorii ca metrică ce se optimizează. Aceasta are la bază metoda CART, prin intermediul căreia se dorește minimizarea devianței și a pruning-ului.

În cazul utilizării librăriei tree pentru arborele cu entropie și cel cu indexul gini, se abordează o rezolvare diferită în sensul în care de această dată se dorește maximizarea informației în fiecare moment.

2.1. Arbore simplu 1 (arbore1)

În cele ce urmează cu ajutorul librăriei rpart, care folosește eroarea ca și metrică care se optimizează, s-au creat doi arbori simpli care prezintă detaliat atributele care au influențat și în măsură supraviețuirea anumitor pasageri de pe Titanic.

După cum se poate observa și în cele două imagini de mai jos, arborele de decizie are în vârf 621 de valori dintre care majoritar întâlnim „No”, la scufundarea titanicului numărul decedaților a fost mai mare decât al supraviețuitorilor. Factorii care au determinat șansele acestora să supraviețuiască au fost în primul rând sexul persoanei, un pasager de sex feminin a avut șanse mai mari să trăiască catrastofei decât un bărbat. În cazul femeilor factorul primar care a determinat șansele lor de supraviețuire a fost totuși clasa și nu vârsta, cu cât discutăm de o clasă cu un număr mai mare, posibilitatea de supraviețuire scade. Bărbații și băieții cu vârste peste 9,5 ani au decedat în procentajul cel mai mare.



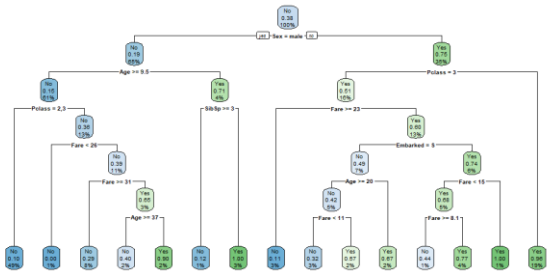

```
1) root 621 237 No (0.61835749 0.38164251)
  2) Sex=male 405  75 No (0.81481481 0.18518519)
    4) Age>=9.5 381  58 No (0.84776903 0.15223097) *
    5) Age< 9.5 24   7 Yes (0.29166667 0.70833333)
     10) SibSp>=2.5 8   1 No (0.87500000 0.12500000) *
     11) SibSp< 2.5 16   0 Yes (0.00000000 1.00000000) *
  3) Sex=female 216  54 Yes (0.25000000 0.75000000)
    6) Pclass=3 101  49 Yes (0.48514851 0.51485149)
     12) Fare>=23.35 18   2 No (0.88888889 0.11111111) *
     13) Fare< 23.35 83  33 Yes (0.39759036 0.60240964)
       26) Embarked=S 45  22 No (0.51111111 0.48888889)
         52) Age>=19.5 33  14 No (0.57575758 0.42424242) *
         53) Age< 19.5 12   4 Yes (0.33333333 0.66666667) *
       27) Embarked=C,Q 38  10 Yes (0.26315789 0.73684211) *
    7) Pclass=1,2 115   5 Yes (0.04347826 0.95652174) *
```

Figura 8. Arbore simplu 1

În ceea ce privește calitatea modelului, acesta poate fi considerat unul bun, deoarece are o acuratețe ridicată (80.22%) și un P-value foarte mic 3.778e-11. Totuși, datorită faptului că setul de date pe care se face analiza dispune un număr relativ mic de atribute, am considerat important să încercăm și un model cu cp = 0 (arbore2).

2.2. Arbore simplu 2 (arbore2)

Specificare opțiunii de control (cp) să fie 0 presupune, după cum se poate observa și în diagrama de mai jos, creare unui arbore mai complex, netăiat, deoarece implică eliminarea oricărui tip de penalizare pentru creare de ramuri. Totuși, această abordare deși, în cazul setului de date analizat, a creat un model mai bun care se potrivește mai bine pe setul de antrenament, există posibilitate mai mare de overfitting.



```
1) root 621 237 No (0.61835749 0.38164251)
  2) Sex=male 405  75 No (0.81481481 0.18518519)
    4) Age>=9.5 381  58 No (0.84776903 0.15223097)
      8) Pclass=2,3 303  30 No (0.90099010 0.09900990) *
      9) Pclass=1 78  28 No (0.64102564 0.35897436)
       18) Fare< 26.14375 7   0 No (1.00000000 0.00000000) *
       19) Fare>=26.14375 71  28 No (0.60563380 0.39436620)
         38) Fare>=30.5979 51  15 No (0.70588235 0.29411765) *
         39) Fare< 30.5979 20   7 Yes (0.35000000 0.65000000)
           78) Age>=36.5 10   4 No (0.60000000 0.40000000) *
           79) Age< 36.5 10   1 Yes (0.10000000 0.90000000) *
    5) Age< 9.5 24   7 Yes (0.29166667 0.70833333)
     10) SibSp>=2.5 8   1 No (0.87500000 0.12500000) *
     11) SibSp< 2.5 16   0 Yes (0.00000000 1.00000000) *
  3) Sex=female 216  54 Yes (0.25000000 0.75000000)
    6) Pclass=3 101  49 Yes (0.48514851 0.51485149)
     12) Fare>=23.35 18   2 No (0.88888889 0.11111111) *
     13) Fare< 23.35 83  33 Yes (0.39759036 0.60240964)
       26) Embarked=S 45  22 No (0.51111111 0.48888889)
         52) Age>=19.5 33  14 No (0.57575758 0.42424242)
          104) Fare< 10.825 19   6 No (0.68421053 0.31578947) *
          105) Fare>=10.825 14   6 Yes (0.42857143 0.57142857) *
         53) Age< 19.5 12   4 Yes (0.33333333 0.66666667) *
       27) Embarked=C,Q 38  10 Yes (0.26315789 0.73684211)
         54) Fare< 15.3729 31  10 Yes (0.32258065 0.67741935)
          108) Fare>=8.08335 9   4 No (0.55555556 0.44444444) *
          109) Fare< 8.08335 22   5 Yes (0.22727273 0.77272727) *
         55) Fare>=15.3729 7   0 Yes (0.00000000 1.00000000) *
    7) Pclass=1,2 115   5 Yes (0.04347826 0.95652174) *
```

Figura 9. Arbore simplu 2

Pentru arborele 2 s-a aplicat pruning, proces prin care se simplifică arborele prin eliminarea unor ramuri/subramuri care sunt considerate mai puțin importante, cu scopul de a îmbunătăți modelul și de a reduce overfitting-ul. Deși această tehnică ar putea aduce îmbunătățiri, în cazul modelului de față, datorită numărului redus de atribute, nu se observă modificări în ceea ce privește calitatea modelului.

2.3. Arbore cu entropie

Entropia măsoară cât de impură sau incertă este distribuția claselor într-un set de date. Practic cu cât se discută de clase mai amestecate și de o incertitudine mai mare în distribuția claselor cu atât crește și valorea entropiei. Aceasta va avea valori mici pentru nodurile pure. Se consideră despărțire optimă cazurile cu o entropie minimă.

Din analiza metricelor modelului se poate observa faptul că are o specificitate de 67,76%, o valorea destul de mică, deoarece doar o parte din pasagerii care au supraviețuit au fost găsiți.

Totuși, vorbim de un P- Value mai mic decât cel al modelelor de mai sus ceea de determină o posibilitate mai mare a respingerii ipotezei nule, făcând predicția mai semnificativă statistic.

```
               Accuracy : 0.806
                 95% CI : (0.7535, 0.8516)
    No Information Rate : 0.6157
    P-Value [Acc > NIR] : 1.464e-11

                  Kappa : 0.5793

 Mcnemar's Test P-Value : 0.07142

            Sensitivity : 0.8848
            Specificity : 0.6796
         Pos Pred Value : 0.8156
         Neg Pred Value : 0.7865
             Prevalence : 0.6157
         Detection Rate : 0.5448
   Detection Prevalence : 0.6679
      Balanced Accuracy : 0.7822

       'Positive' Class : No
```

Figura 10. Rezultate entropie

2.4. Index Gini

Indexul Gini este folosit pentru a evalua cât de calitativ au fost despărțite datele. În cazul unei despărțiri perfecte indexul va lua valorea 0, iar cazul în care nu se va reuși separarea în clase corecte, indexul va avea valorea de maxim 1. În cadrul arborelui de decizie se utilizează acest index pentru a măsura impuritățile în fiecare nod al arborelului și a determina în funcție de această măsurătoare care este cea mai bună metodă de a face split. În cazul nodurilor pure, unde elementele aparțin aceleiași clase se ia decizia să nu se mai facă split. Se consideră despărțire optimă cazurile cu un index Gini minim.

Folosind acest model am reușit să obținem o acuratețe de 73,88%, o valoare destul de mică comparativ cu modele realizate precedent, dar totuși arborele cu Gini inregistrează cea mai bună valoare pentru P-value, 1,429e-05. În ceea ce privește specificitatea și senzitivitatea înregistrăm valori relativ medii comparativ cu rezultatele tuturor modelelor create.

```
               Accuracy : 0.7388
                 95% CI : (0.6819, 0.7904)
    No Information Rate : 0.6157
    P-Value [Acc > NIR] : 1.429e-05

                  Kappa : 0.4357

 Mcnemar's Test P-Value : 0.1886

            Sensitivity : 0.8242
            Specificity : 0.6019
         Pos Pred Value : 0.7684
         Neg Pred Value : 0.6813
             Prevalence : 0.6157
         Detection Rate : 0.5075
   Detection Prevalence : 0.6604
      Balanced Accuracy : 0.7131

       'Positive' Class : No
```

Figura 11. Rezultate Gini

1. Arbori de decizie avansați:

Bagging este o procedură folosită în cadrul arborilor de decizie avansați astfel: procedura va extrage din setul de antrenament M baguri de dimensiunea m. Pe fiecare bag în parte se va construi un model, în cazul nostru acel model va fi un arbore de decizie, iar după aceea se va genera o predicție. Așadar vor rezulta M predicții între care se va realiza o medie, rezultând predicția finală. Un alt aspect important de reținut la bagging e faptul că validarea se va realiza pe instanțele din cadrul setului de antrenament, care au rămas înafara bag-ului respectiv.

Într-o primă fază s-a generat un model ce folosește implicit 25 bags, iar eroarea de predicție se poate observa în Figura 12:

```
Bagging classification trees with 25 bootstrap replications

Call: bagging.data.frame(formula = Survived ~ ., data = titanic_train,
    coob = TRUE)

Out-of-bag estimate of misclassification error:  0.1916
```

Figura 12. Eroare de predicție 1

În scopul reducerii erorii respective, s-a realizat un grafic care reprezintă variația erorii de predicție în funcție de numărul de bag-uri folosite și s-a observat că o valoare minimă a erorii s-ar obține la folosirea a 33 de bag-uri, motiv pentru care s-a considerat inutilă ideea de a trece de acest prag în ceea ce înseamnă crearea de noi bag-uri.

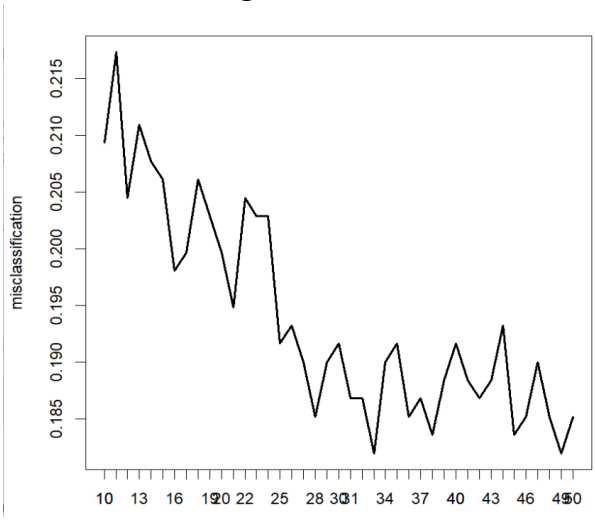


Figura 13. Variația erorii de predicție

În continuare, în funcția bagging() s-a menționat nbag-ul potrivit, descoperit în Figura 13 și s-a rezultat o eroare de predicție mai mică, anume: 0.1755.

```
Bagging classification trees with 33 bootstrap replications

Call: bagging.data.frame(formula = Survived ~ ., data = titanic_train,
    coob = TRUE, nbag = 33)

Out-of-bag estimate of misclassification error:  0.1755
```

Figura 14. Eroare de predicție 2

Rezultatele finale se pot observa după cum urmează:

```
               Accuracy : 0.7836
                 95% CI : (0.7294, 0.8314)
    No Information Rate : 0.6157
    P-Value [Acc > NIR] : 3.086e-09

                  Kappa : 0.5444

 Mcnemar's Test P-Value : 0.8955

            Sensitivity : 0.8182
            Specificity : 0.7282
         Pos Pred Value : 0.8282
         Neg Pred Value : 0.7143
             Prevalence : 0.6157
         Detection Rate : 0.5037
   Detection Prevalence : 0.6082
      Balanced Accuracy : 0.7732

       'Positive' Class : No
```

Figura 15. Rezultate Bagging

2. Compararea modelelor:

Pentru setul de date ales am considerat relevant să analizăm acuratețea și specificitatea modelelor, deoarece discutăm de un set cu date foarte disproporționale, cu toate acestea în cadrul tabelului de mai jos am specificat cu titlu informativ și P-value și senzitivitatea.

Model	**Acuratețe**	**P-Value**	**Specificitate**	**Senzitivitate**
Arbore simplu 1	80,22%	3,778e-11	61,17%	92,12%
Arbore simplu 2	82,84%	2,992e-14	67,96%	92,12%
Arbore Prunned	82,84%	2,992e-14	67,96%	92,12%
Arbore Entropie	80,6%	1,464e-11	67,96%	88,48%
Arbore Gini	73,88%	1,429e-05	60,19%	82,42%

Tabel 2. Compararea modelelor

Analizând tabelul de mai sus putem trage următoarele concluzii:

1. Cel mai bun model în ceea ce privește acuratețea sunt cei doi arbori (arborele 2 simplu cu CP = 0 și varianta sa prunned), aceștia atingând o valoarea a acurateții de 82,84% și o specificitate de 67,96%. Totuși se poate observa că există modele cu o specificitate mai

bună 76,70 - Bagged M1. Ținând cont de analiza pe care dorim să o facem, identificarea persoanelor care au supraviețuit – clasa minoritară în cazul nostru, considerăm relevant să alegem un model cu specificitate mai mare.

2. Modelul identificiat ca fiind cel mai dezavantajos este Naive Bayes. Cum am menționat și în capitolele precendente acesta consideră toate atributele ca fiind independente statistic, motiv pentru care metricele de analiză sunt mai slabe după cum urmează: acuratețe de 77,99% și specificitate de 58,25%. Senzititivitatea are, totuși, o valorea ridicată, însă ținând cont că ne interesează clasa minoritară, celelalte metrici ne interesează mai mult.

În figura de mai jos se poate observa curba ROC:

- Roșu – Arborele 1 simplu;
- Portocaliu – Arborele 2 (CP = 0);
- Verde – Arborele 2 prunned;
- Albastru – Arborele cu entropie;
- Mov – Arborele cu indexul Gini;
- Negru - Arborele avansat cu bagging M1;
- Galben – Arborele avansat cu bagging M33;
- Gri – Naive Bayes.

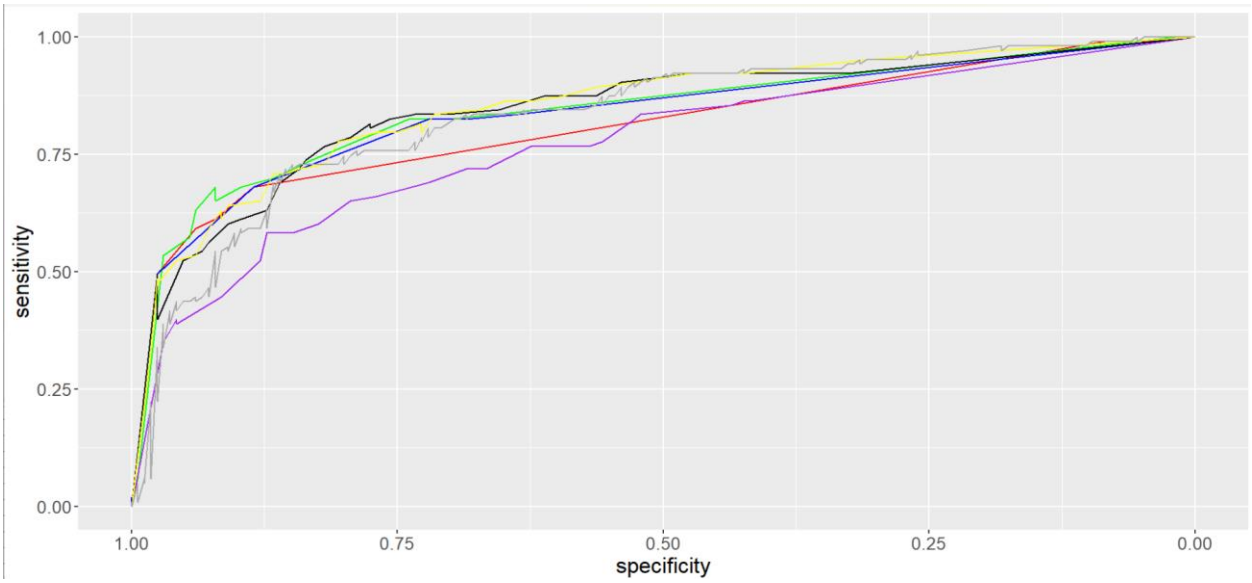


Figura 16. Curba ROC

IV. Concluzii:

Așadar analizând modele prezentate s-a reușit să se răspundă întrebărilor de cercetare precizate în introducere, astfel:

- Factorii care au cel mai mare impact asupra clasificării sunt: sexul, vârsta și clasa socială – indicată de Pclass (biletul aferent clasei). Sexul persoanei a repezentat un prim factor determinat în șansele de supraviețuire, în sensul în care procentajul femeilor care au supraviețuit a fost mai mare decât al bărbaților.
- În ceea ce privește măsura în care putem prezice supraviețuirea persoanelor analizând factorii dați am constatat că se poate obține o performanță de prezicere de 82,84%.

V. Limitări ale studiului:

În ceea ce privește limitările studiului câteva dintre acestea includ: lipsa parțială a unor date, precum cabina sau vârsta unor pasageri, lipsa informațiilor despre alți factori care ar fi putut determina supraviețuirea sau moartea unor pasageri, precum și posibile conflicte care ar fi putut apărea pe transatlantic în momentul în care s-a conştientizat iminența catastrofei.

Ca şi direcții viitoare de cercetare luăm în considerare segmentarea pasagerilor în grupuri de vârstă pentru a vedea mai clar care categorii au avut şanse mai mari de supraviețuire, cercetarea posibililor conflicte care au apărut şi determinarea poziției exacte pe navă a pasagerilor.