Proiect Fundamente de Big Data

Clasificarea satisfacției clienților pentru o companie aeriană

Proiect realizat de:
David Larisa Patricia, larisa.david@stud.ubbcluj.ro
Gâta Alexandra Denisa, alexandra.gata@stud.ubbcluj.ro

Cuprins

1. Introducere

Companiile sunt întotdeauna interesate de a crește satisfacția clienților lor. Există o mulțime de factori ce pot influența percepția consumatorilor asupra produselor sau serviciilor oferite. De aceea, în urma studiilor de marketing privind părerea clienților, este important să putem analiza ce factori sunt relevanți în creearea opiniei finale.

În acest proiect, vom analiza cazul unei companii aeriene ce a efectuatuat un studiu privind satisfacția clienților săi. Serviiciile companiilor aeriene nu includ doar transportul propriu-zis, ci și o gama largă de alte servicii prestate înainte și în timpul zborului, cu scopul de a crește satisfacția consumatorilor și de a se departaja de concurență. Astfel, este esențial pentru acestea să își poată analiza importanța serviciilor prestate.

Setul de date ce a fost folosit este rezultat în urma unui chestionar privind experiența clienților în cadrul ultimului lor zbor cu compania analizată. Datele preluate au fost referitoare la diverse servicii și facilități oferite. Acestea vor fi detaliate în capitolul „Setul de date”.

Setul de întrebări de cercetare pentru care ne propunem să realizăm analiza este:

1. Există vreo legătură între datele referitoare la client (Gender, Customer Type, Age, Flight distance) și gradul de satisfacție?
2. În cazul în care există, cât de puternică este legătura?
3. Este posibilă realizarea unei estimări dacă clientul va fi sau nu satisfăcut, având în vedere părerea acestuia referitoare la serviciile oferite de compania aeriană?

Analiza în funcție de datele referitoare la client (Gender, Customer Type, Age, Flight distance) este relevantă pentru a evita subiectivitatea dată de anumite segmente de consumatori. De exemplu, persoanele care călătoresc distanțe lungi ar putea fi predispuse să aibă o părere negativă indiferent de restul serviciilor prestate. Acest exemplu este doar o posibilă presupunere, ce trebuie verificată pentru a putea creea o concluzie cu acuratețe.

În cazul datelor referitoare la părerea consumatorului față de serviciile oferite de compania aeriană, în urma analizei, dorim să verificăm dacă va putea fi realizată o predicție dacă clientul este sau nu satisfăcut. Această predicție este folositoare în detectarea clienților predispuși la a fi nesatisfăcuți, astfel încât compania să poată interveni prin diverse mijloace pentru a diminua gradul de insatisfacție creat. Acestea ar putea fi oferirea de cupoane de reducere, servicii suplimentare sau alte beneficii pentru a îmbunătăți experiența acestuia.

Prevenirea insatisfacției consumatoriilor este importantă pentru a creea o imagine pozitivă a companiei și a preveni recenziile negative. Astfel, acordând atenție acestei analize, se poate îmbunătății experiența clienților, iar ca rezultat, compania va avea o creștere în performanță.

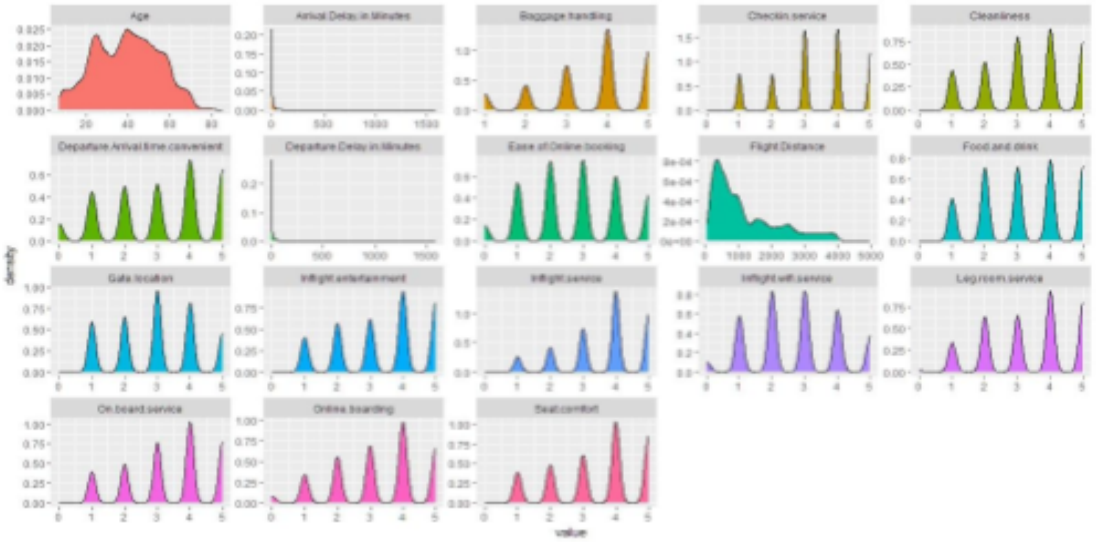
2. Setul de date

Setul de date utilizat în realizarea acestui proiect a fost preluat de pe site-ul https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction. Acesta reprezintă un set de date colectate în urma unui chestionar de opinie a clienților ce au folosit serviciile companiei aeriene.

Atributele din setul de date sunt: **Gender:** genul clientului (Female, Male), **Customer Type:** tipul clientului (Loyal customer, disloyal customer), **Age:** vârsta actuală a clientului, **Type of Travel:** scopul zborului (Personal Travel, Business Travel), **Class:** clasa la care a zburat (Business, Eco, Eco Plus), **Flight distance:** distanța călătoriei, **Inflight wifi service:** nivelul satisfacției serviciilor wi-fi din durata zborului (0:Nu se aplică; 1-5), **Departure/Arrival time convenient:** nivelul de satisfacție față de ora de decolare/ sosire, **Ease of Online booking:** nivelul de satisfacție ușurița rezervării online, **Gate location:** nivelul de satisfacție față de amplasarea porții de îmbarcare, **Food and drink:** nivelul de satisfacție față de mâncare și băutură, **Online boarding:** nivelul de satisfacție rezervării online, **Seat comfort:** nivelul de satisfacție a confortului scaunelor, **Inflight entertainment:** nivelul de satisfacție față de divertismentul la bord, **On-board service:** nivelul de satisfacție serviciile de la bord, **Leg room service:** nivelul de satisfacție spațiul pentru picioare, **Baggage handling:** nivelul de satisfacție față de serviciile de manipulare a bagajelor, **Check-in service:** nivelul de satisfacție a serviciilor de check-in, **Inflight service:** nivelul de satisfacție servicii din durata zborului, **Cleanliness:** nivelul de satisfacție față de curățenie, **Departure Delay in Minutes:** minute de întârziere la decolare, **Arrival Delay in Minutes:** minute de întârziere la aterizare. Atributul țintă este **satisfaction** ce arată clasa din care face parte clientul: „satisfied” sau „neutral or dissatisfied”.

Primul pas este eliminarea datelor nule. În plus, atributele ce aveau categorii stocate în date de tip string (Gender, Customer Type, Type of travel, Class și satisfaction) au fost transformate în factori.

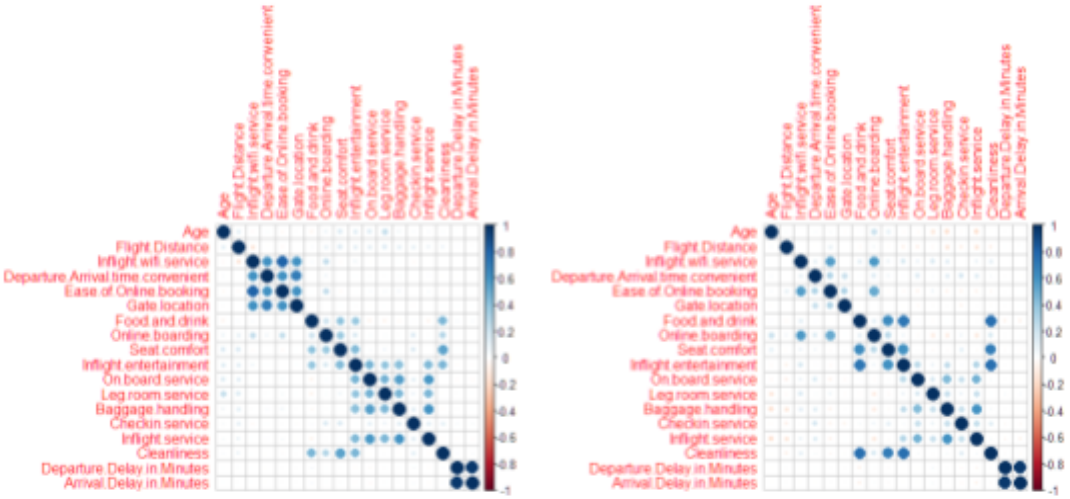
Din cauza faptului că atributele erau numeroase și exista un risc de corelare mare între ele, ceea ce ar duce la scăderea acurateței metodelor, am realizat un proces de curățare. Primul pas a fost vizualizarea datelor numerice.



Se poate observa ca majoritatea sunt date extrase din chestionare cu variante de la 0 la 5 pentru a clasifica nivelul de satisfacție față de diferite servicii prestate. Aceste date vor trebui transformate în factori pentru a putea fi interpretate corect în realizarea metodelor de predicție. Atributele ce vor fi transformate în factori sunt: Inflight wifi service, Departure/Arrival time

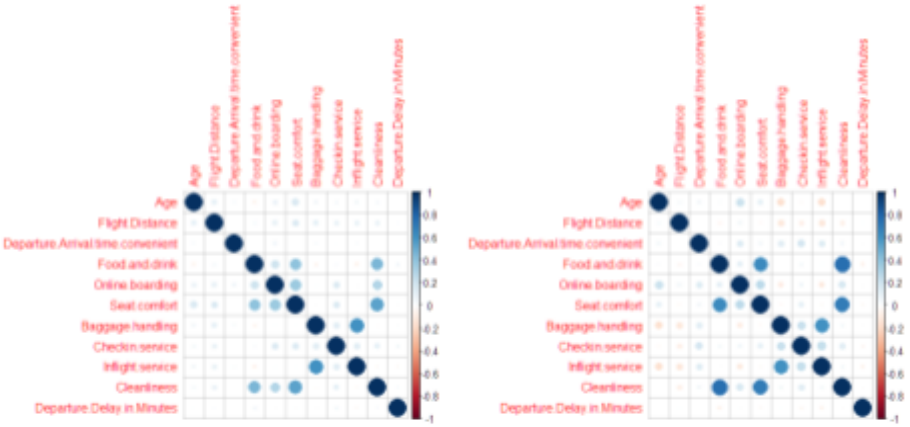
convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Check-in service, Inflight service și Cleanliness.

De asemenea, am verificat corelarea între atributele numerice. Prima reprezentare este pentru clienții din clasa „satisfied”, iar a doua din clasa „neutral or dissatisfied”.



Se pot observa mai multe corelări dintre atribute. De asemenea, se pot observa și anumite atribute care conțin valori lega de aceleași servicii. De exemplu, Leg room service este inclus în atributul Seat comfort.

Luând în cosiderare observațiile anterioare, am eliminat o serie de atribute: Arrival Delay in Minutes, Inflight wifi service, Ease of Online booking, Gate location, Inflight entertainment, On-board service și Leg room service. Am realizat din nou reprezentarea corelării dintre atribute. Prima reprezentare este pentru clienții din clasa „satisfied”, iar a doua din clasa „neutral or dissatisfied”.



Se poate observa că nu au fost eliminate complet corelăriile dintre atribute. Am decis acest lucru deoarece chiar dacă par a fi corelate, atributele Cleanliness, Food and drink și Seat comfort se referă fiecare la câte un serviciu diferit. Astfel, am decis să păstrăm aceste valori. Rezultatele în urma curățării efectuate vor fi folosite în continuare pentru fiecare metodă.

3. Rezultate și discuții

3.1. Prezentare generală

Metodele alese sunt Naive Bayes, Regresia logistică și Arbori de decizie. Aceste metode au fost alese pentru a putea realiza o predicție de clasificare și a putea compara rezultatele obținute în urma fiecărei metode. Comparația va fi efectuată pentru a putea alege metoda cu acuratețea cea mai bună, astfel asigurând calitatea rezultatelor.

3.2. Naive Bayes

Prima metodă utilizată a fost Naive Bayes. Implementarea acesteia a fost realizată pe setul de date al companiei aeriene ce a fost curățat după cum este descris în capitolul „Setul de date”.

Procesul de predicție începe prin împărțirea datelor în date de antrenament (70%) și date de test (30%). Datele de test vor fi utilizate o singură dată, pentru a verifica situația de overfitting. Distribuția pe cele doua clase este:

Date de training	Date de test
neutral or dissatisfied satisfied 41087 31427	neutral or dissatisfied satisfied 17610 13470

Se observă că există mai multe date în clasa „neutral or dissatisfied”.

Pentru a obține o acuratețe mai mare folosind acest set de date, am utilizat metoda de validare cross-validation cu k=5. Modelul rezultat va avea o acuratețe de 84,64158% dacă nu este folosit kernel și 84,93670% dacă este folosit kernel. Matricea de confuzie rezultată (folosind kernel) este următoarea:

```
                          Reference
Prediction                neutral or dissatisfied satisfied
  neutral or dissatisfied                    49.7       8.1
  satisfied                                   7.0      35.2

 Accuracy (average) : 0.8494
```

Acuratețea este bună, fiind de 84,94%. Se observă o preferință pentru prezicerea clasei „neutral or dissatisfied”, înregistrându-se 49,7% true negatives față de 35,2% true positives. Valorile pentru false positives (7,0%) și false negatives (8,1%) sunt mici. Astfel, modelul creat este bun, dar poate fi îmbunătățit.

În continuare am creat un nou model pe metoda Naive Bayes, căutând o combinație optimă de parametri. Parametrii ce vor fi ajustați în căutare sunt: dacă va fi sau nu folosit kernel, factorul Laplace va fi setat la 0,5, iar intervalul de ajustare a kernelului este de la 0 la 5, cu pasul de 1. Modelul cel mai optim rezultat a fost cel cu kelnel cu ajustare 4 și factor Laplace egal cu 0,5. Acesta a rezultat în acurateţe de 85,25940%, iar matricea de confuzie este următoarea:

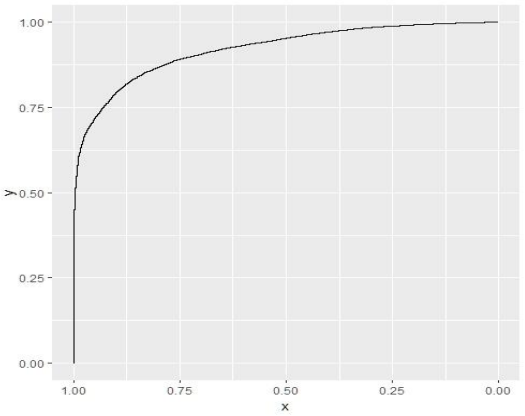
```
                          Reference
Prediction                neutral or dissatisfied satisfied
  neutral or dissatisfied                    50.6       8.7
  satisfied                                   6.1      34.7
```

Se observă o creștere pentru prezicerea clasei „neutral or dissatisfied”, înregistrându-se 50,6% true negatives, dar o scădere în true positives la 34,7%. Valorile pentru false positives au scăzut la 6,1% și false negatives au crescut la 8,7%. Deoarece este mai importantă detectarea clienților ce nu sunt mulțumiți, evoluția valorilor este una favorabilă.

În continuare am creat predicția pe setul de testare cu modelul ajustat. Rezulatele în urma comparării cu valorile reale sunt următoarele:

```
                          Reference
Prediction                neutral or dissatisfied satisfied
  neutral or dissatisfied                   15743      2711
  satisfied                                  1867     10759

               Accuracy : 0.8527
                 95% CI : (0.8487, 0.8566)
    No Information Rate : 0.5666
    P-Value [Acc > NIR] : < 2.2e-16

            Sensitivity : 0.8940
            Specificity : 0.7987

       'Positive' Class : neutral or dissatisfied
```

Acuratețea finală este de 85,27%, iar intervalul de încredere este (0,8487, 0,8566). P-value este mic (< 2.2e-16), rezultând o acuratețe mai mare decât NIR (No Information Rate) care se înregistra la 56,66%. Sensitivitatea (0,8940) este mai mare ca specificitatea (0,7987), aratând faptul ca modelul poate identifica mai bine clienții din categoria „neutral or dissatisfied” față de cei „satisfied”.



În final am generat curba ROC, pentru a putea analiza reprezentarea grafică a ratei obținutînută pentru true positives (sensitivitate), raportată la rata de false positives (specificitate). Aria curbei ROC este mare, aratând eficiența metodei Naive Bayes ajustată.

3.3. Regresie Logistică

A doua metodă utilizată a fost Regresia logistică. Implementarea acesteia a fost realizată pe setul de date al companiei aeriene ce a fost curățat după cum este descris în capitolul „Setul de date".

În realizarea primului model de regresie logistică, au fost incluse toate atributele. Am descis acestă metodă în favoarea analizei individuale a fiecărui atribut, deoarece astfel atributele sunt numeroase și am putut analiza care sunt relațiile între ele, evitând confounding-ul. Modelul creat a avut următorul rezultat:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.929e+01	2.225e+02	-0.087	0.93091
Seat.comfort1	-1.248e+00	2.052e+02	-0.006	0.99515
Seat.comfort2	-1.527e+00	2.052e+02	-0.007	0.99406
Seat.comfort3	-2.441e+00	2.052e+02	-0.012	0.99051
Seat.comfort4	-1.992e+00	2.052e+02	-0.010	0.99226
Seat.comfort5	-1.464e+00	2.052e+02	-0.007	0.99431
Inflight.service1	1.179e+01	1.035e+02	0.114	0.90934
Inflight.service2	1.179e+01	1.035e+02	0.114	0.90928
Inflight.service3	1.164e+01	1.035e+02	0.112	0.91049
Inflight.service4	1.250e+01	1.035e+02	0.121	0.90390
Inflight.service5	1.303e+01	1.035e+02	0.126	0.89981
Cleanliness1	1.297e+01	5.768e+01	0.225	0.82204
Cleanliness2	1.310e+01	5.768e+01	0.227	0.82027
Cleanliness3	1.359e+01	5.768e+01	0.236	0.81367
Cleanliness4	1.350e+01	5.768e+01	0.234	0.81501
Cleanliness5	1.378e+01	5.768e+01	0.239	0.81118

În urma acestor rezultate, după analiza p-value a fiecărui atribut se vor elimina atributele: Seat comfort, Inflight service și Cleanliness. Eliminarea acestora a fost deoarece p-value avea o valoare mare, rezultând în o influență nesemnificativă asupra modelului final a atributelor. În continuare, a fost realizat încă un model de regresie, dar care nu avea decât atributele rămase. Însă au fost înregistrate valori mari a p-value pentru Checkin services și Flight Distance:

	Estimate	Std. Error	z value	Pr(>|z|)
Checkin.service1	8.523e+00	7.246e+01	0.118	0.906368
Checkin.service2	8.647e+00	7.246e+01	0.119	0.905013
Checkin.service3	9.109e+00	7.246e+01	0.126	0.899970
Checkin.service4	9.061e+00	7.246e+01	0.125	0.900486
Checkin.service5	9.757e+00	7.246e+01	0.135	0.892895
Flight.Distance	1.351e-05	1.422e-05	0.950	0.342328

Astfel, pentru următorul model a fost eliminate și aceste atribute. În plus, în realizarea următorului model datele au fost împărțite în date de antrenament (70%) și date de testare (30%). Rezultatele pe setul de antrenament sunt:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.8130069	0.3425544	8.212	< 2e-16	***
GenderMale	0.0909008	0.0248291	3.661	0.000251	***
Customer.TypeLoyal Customer	2.3815849	0.0387665	61.434	< 2e-16	***
Age	-0.0051836	0.0008991	-5.765	8.15e-09	***
Type.of.TravelPersonal Travel	-3.4938383	0.0424709	-82.264	< 2e-16	***
ClassEco	-0.7231214	0.0297616	-24.297	< 2e-16	***
ClassEco Plus	-0.8993076	0.0505866	-17.778	< 2e-16	***
Departure.Arrival.time.convenient1	-0.5463402	0.0632707	-8.635	< 2e-16	***
Departure.Arrival.time.convenient2	-0.5499732	0.0624894	-8.801	< 2e-16	***

```
Departure.Arrival.time.convenient3 -0.4899504  0.0617416  -7.935 2.10e-15 ***
Departure.Arrival.time.convenient4 -0.5278244  0.0581688  -9.074  < 2e-16 ***
Departure.Arrival.time.convenient5 -0.5352999  0.0599510  -8.929  < 2e-16 ***
Food.and.drink1                    -2.4076420  0.3331428  -7.227 4.94e-13 ***
Food.and.drink2                    -2.0030888  0.3325232  -6.024 1.70e-09 ***
Food.and.drink3                    -1.9886686  0.3324025  -5.983 2.19e-09 ***
Food.and.drink4                    -1.6948816  0.3324529  -5.098 3.43e-07 ***
Food.and.drink5                    -1.6763848  0.3328052  -5.037 4.73e-07 ***
Online.boarding1                   -3.4091621  0.0765379 -44.542  < 2e-16 ***
Online.boarding2                   -3.7148040  0.0730487 -50.854  < 2e-16 ***
Online.boarding3                   -3.7534944  0.0710751 -52.810  < 2e-16 ***
Online.boarding4                   -1.4853596  0.0667396 -22.256  < 2e-16 ***
Online.boarding5                    0.5346019  0.0703932   7.595 3.09e-14 ***
Baggage.handling2                  -0.1352036  0.0566528  -2.387 0.017008 *
Baggage.handling3                  -0.1959209  0.0528245  -3.709 0.000208 ***
Baggage.handling4                   1.1646320  0.0495057  23.525  < 2e-16 ***
Baggage.handling5                   2.0696802  0.0530220  39.034  < 2e-16 ***
Departure.Delay.in.Minutes         -0.0044881  0.0003328 -13.484  < 2e-16 ***
```

Se poate observa ca valorile p-value sunt mici, rezultând că atributele rămase afectează într-un mod semnificativ predicția finală. Atributele ce influențează decizia spre clasa „satisfied” sunt Gender (Male), Customer type (Loyal Customer), Online boarding (5) și Baggage handling (4 și 5), iar celalte influențează decizia spre clasa „neutral or dissatisfied”.

Setând ca valorile rezultate în urma predicției să fie împărțite în „neutral or dissatisfied” dacă sunt sub 0,5 și „satisfied” dacă sunt peste 0,5, matricea obținută este:

```
       neutral or dissatisfied satisfied
FALSE                    15865      2098
TRUE                      1745     11372
```

Datele arată o predicție mai bună a clasei „neutral or dissatisfied” față de „satisfied”. De asemenea, numărul de false positives și false negatives sunt mici.

Următorul model a fost realizat folosind cross-validation cu toate atributele. Acuratețea rezultată este de 88,97105%. Matricea de confuzie arată în continuare o estimare mai bună a clasei „neutral or dissatisfied” și un număr mic de valori pentru false positives și false negatives.

```
                         Reference
Prediction               neutral or dissatisfied satisfied
  neutral or dissatisfied                    51.9       6.3
  satisfied                                   4.7      37.0
```

În urma predicției pe setul de testare, valorile obținute sunt următoarele:

```
              Accuracy : 0.8879
                95% CI : (0.8844, 0.8914)
   No Information Rate : 0.5666
   P-Value [Acc > NIR] : < 2.2e-16

                 Kappa : 0.7705

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9198
            Specificity : 0.8463

      'Positive' Class : neutral or dissatisfied
```

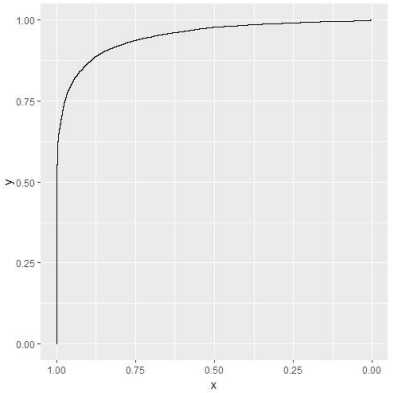
Acuratețea finală este de 88,79%, iar intervalul de încredere este (0,8844, 0,8914). P-value este mic (< 2.2e-16), rezultând o acuratețe mai mare decât NIR (No Information Rate) care se înregistra la 56,66%. Sensitivitatea (0,9198) este mai mare ca specificitatea (0,8463),

aratând faptul ca modelul poate identifica mai bine clienții din categoria „neutral or dissatisfied” față de cei „satisfied”.

În continuare am realizat un model în care am exclus atributele ce în urma analizei anterioare s-au presupus că nu influențează predicția. Însă rezultatele obținute au fost mai slabe:

```
                          Reference
Prediction                neutral or dissatisfied satisfied
  neutral or dissatisfied                   15865      2098
  satisfied                                  1745     11372

               Accuracy : 0.8764
                 95% CI : (0.8726, 0.88)
    No Information Rate : 0.5666
    P-Value [Acc > NIR] : < 2.2e-16

            Sensitivity : 0.9009
            Specificity : 0.8442

       'Positive' Class : neutral or dissatisfied
```

În concluzie, modelul ales va fi acela folosind cross-validation cu toate atributele, iar rezultatele acestuia au fost detaliate anterior. Curba ROC a fost construită pe acest model, fiind reprezentată grafic:



3.4. Arbore

Ultima metoda utilizată a fost a Arborilor de decizie. Implementarea acesteia pe setul de date al companiei aeriene ce a fost curățat după cum este descris în capitolul „Setul de date”. În plus, pentru aceasta metodă am ales să transformăm variabilele numerice de la Age și Flight.Distance în intervale, pentru a fi mai ușor atunci când arborele împarte pe noduri, de asemenea făcându-l mai ușor de interpretat.

Procesul de predicție începe prin împărțirea datelor în date de antrenament (70%) și date de test (30%). Datele de test vor fi utilizate o singură dată, pentru a verifica situația de overfitting. Distribuția pe cele două clase este:

Date de training	Date de test
neutral or dissatisfied satisfied 41087 31427	neutral or dissatisfied satisfied 17610 13470

Am continuat cu crearea primului arbore m1. De aici, reiese că cele mai importante variabile sunt Online boarding, Seat comfort, Type of travel, Class, Cleanliness, Age group și Food and drink.

```
Online.boarding    Seat.comfort  Type.of.Travel           Class
             33              15              13              13
    Cleanliness       Age_Group  Food.and.drink
             10               9               7
```

Pe baza acestor variabile, s-a creat arborele, urmând să realizăm matricea de confuzie.

```
                          Reference
      Prediction          neutral or dissatisfied        satisfied
      neutral or dissatisfied                15845             3061
      satisfied                               1765            10409
```

De aici observăm că am avut 15845 de instanțe de true negatives și 10409 de true positives. Instanțele de false negatives (3061) și false positives (1765) reprezintă o valoare mică comparativ cu cele prezise corect. În continuare, am analizat statisticile rezultate în urma predicției:

```
               Accuracy : 0.8447
                 95% CI : (0.8406, 0.8487)
    No Information Rate : 0.5666
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6802

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8998
            Specificity : 0.7728
         Pos Pred Value : 0.8381
         Neg Pred Value : 0.8550
             Prevalence : 0.5666
         Detection Rate : 0.5098
   Detection Prevalence : 0.6083
      Balanced Accuracy : 0.8363

       'Positive' Class : neutral or dissatisfied
```

Avem valori relativ bune, atât la acuratețe , cât și la sensitivitate și specificitate, dar ar putea fi îmbunătățite. De aceea, am decis să creăm și un arbore netăiat, ca să analizăm dacă diferența rezultatelor va fi una majoră. Astfel, avem arborele m2, la care am setat valoarea complexity parameter cp=0, pentru ca arborele nostru să aibă numărul maxim de noduri și splituri, având un arbore complex fără pruning. Rezultatele obținute de acesta pe setul de antrenament este unul cu valori bune, dar nu aduce o îmbunătățire majoră, de aceea am decis să nu folosim acest arbore, din cauza dimensiunilor foarte mari pe care le are, iar rezultatele nu au fost suficient de favorabile. Am decis însă, să facem un nou arbore m2_prunned , unde am setat complexity parameter cp = 0.02. În urma analizării matricii de confuzie, am constatat că acesta are valori identice cu cele obținute la m1. De aceea, varianta finală pentru care am folosit ca și parametru de optimizare eroarea , este arborele m1.

Deoarece ne-am dorit să obținem valori și mai bune, am decis să construim arbori și pe baza altor parametri de optimizare și anume entropy și gini.

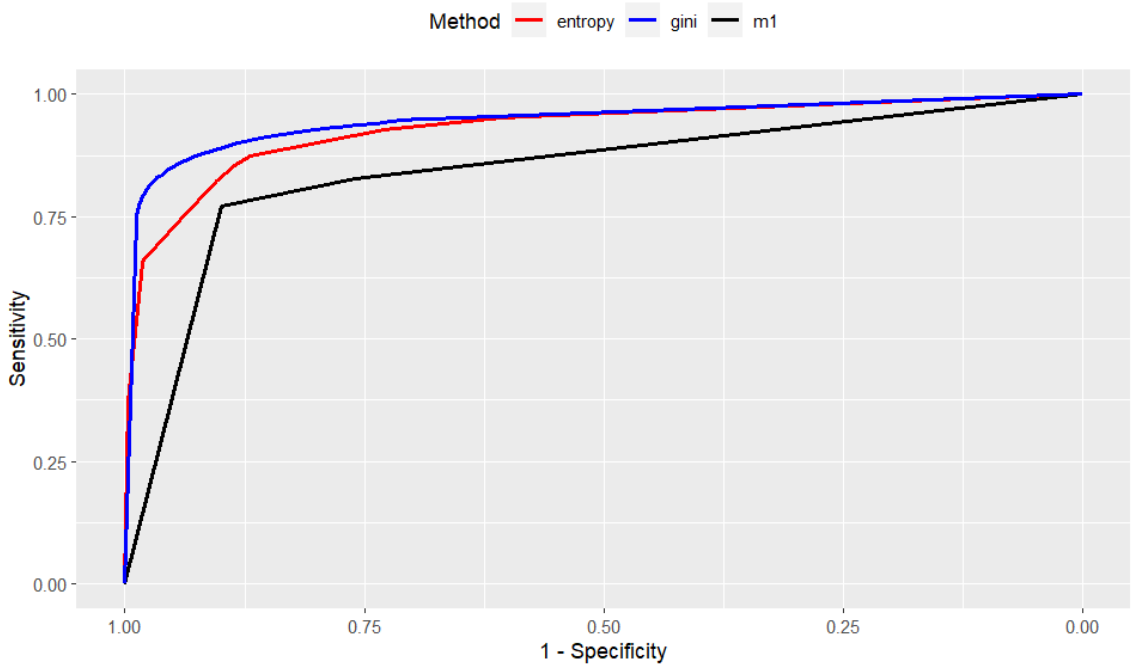
Rezultatele matricii de confuzie pentru arborele cu parametrul entropy:

```
Confusion Matrix and Statistics

                          Reference
Prediction                 neutral or dissatisfied satisfied
  neutral or dissatisfied                   15593      1945
  satisfied                                  2017     11525

               Accuracy : 0.8725
                 95% CI : (0.8688, 0.8762)
    No Information Rate : 0.5666
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7406

 Mcnemar's Test P-Value : 0.2593

            Sensitivity : 0.8855
            Specificity : 0.8556
         Pos Pred Value : 0.8891
         Neg Pred Value : 0.8511
             Prevalence : 0.5666
         Detection Rate : 0.5017
   Detection Prevalence : 0.5643
      Balanced Accuracy : 0.8705

       'Positive' Class : neutral or dissatisfied
```

Rezultatele matricii de confuzie pentru arborele cu parametrul gini:

```
Confusion Matrix and Statistics

                          Reference
Prediction                 neutral or dissatisfied satisfied
  neutral or dissatisfied                   16657      1922
  satisfied                                   953     11548

               Accuracy : 0.9075
                 95% CI : (0.9042, 0.9107)
    No Information Rate : 0.5666
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.81

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9459
            Specificity : 0.8573
         Pos Pred Value : 0.8965
         Neg Pred Value : 0.9238
             Prevalence : 0.5666
         Detection Rate : 0.5359
   Detection Prevalence : 0.5978
      Balanced Accuracy : 0.9016

       'Positive' Class : neutral or dissatisfied
```

În urma acestor rezultate, am observat că arborele cu cea mai mare acuratețe, un p-value mic și totodată cele mai bune valori pentru specificitate si senzitivitate, este arborele care folosește ca și parametru de optimizare gini. Am ales să facem și o reprezentare a curbelor ROC, pentru a se vedea cât mai clar, care dintre variante este cea mai optimă.



3.5. Metoda optimă

Pentru stabilirea metodei optime pentru problema de clasificare prezentată, am luat în considerare rezultatele obținute în urma analizei setului de date prin metodele prezentate anterior.

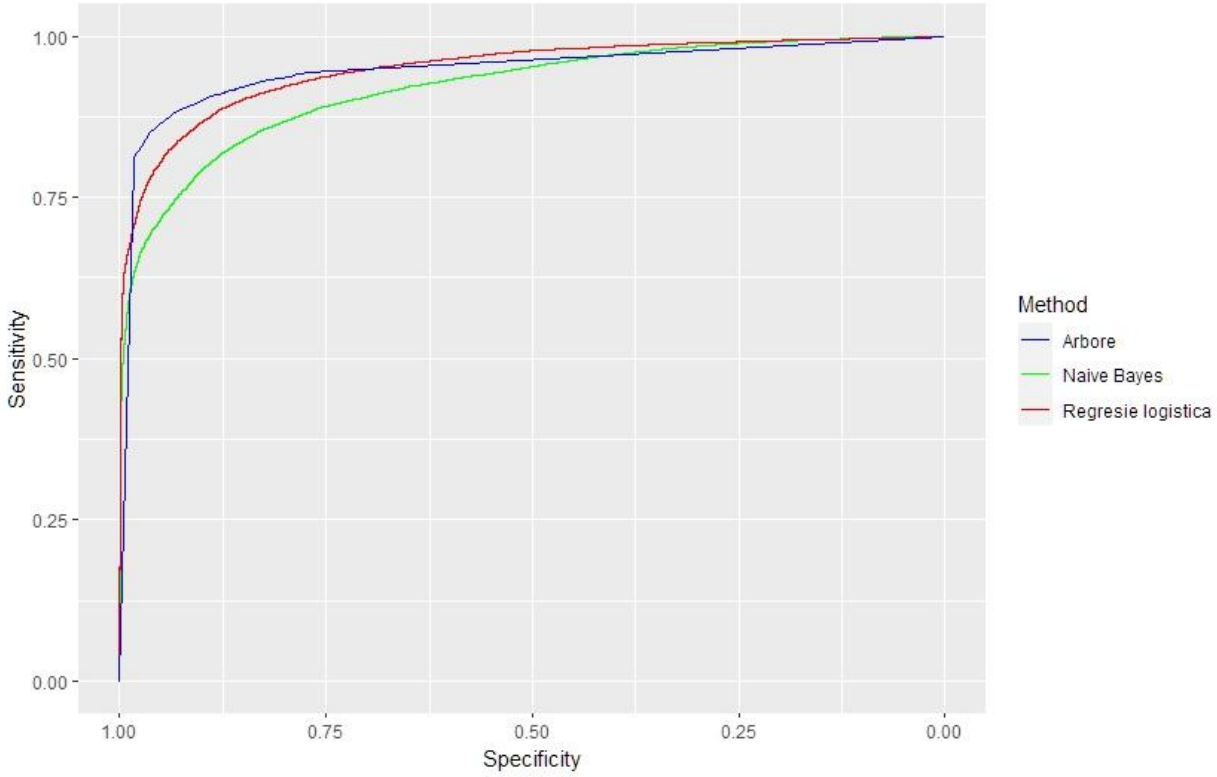
Am realizat o comparație între metodele Naive Bayes, Regresie logistică și Arborele cu parametru gini. În această comparație am utilizat valorile acuratețe, senzitivitate și specificitate. Astfel, am putut analiza ce metoda este mai eficientă în estimarea clasei corecte. În plus, am putut observa cat de corecte sunt estimările pentru fiecare clasă.

Toate metodele au înregistrat o senzitivitate mai mare decât specificitatea, astfel acestea oferă o estimare mai bună a clasei „neutral or dissatisfied”. Acest rezultat este favorabil pentru setul de întrebări de cercetare ales.

Se observă că cele mai bune valori sunt înregistrate pentru metoda Arborelui cu parametru gini, iar cele mai slabe sunt cele ale metodei Naive Bayes.

	Naive-Bayes	**Regresie logistică**	**Arbore (cu parametru gini)**
Acuratețe	0,8527	0,8764	0,9075
Senzitivitate	0,8940	0,9009	0,9459
Specificitate	0,7987	0,8442	0,8573

În continuare, am comparat curbele ROC obținute prin fiecare metodă. În urma reprezentării grafice de mai jos se poate observa din nou că rezultatele cele mai bune sunt obținute prin folosirea metodei Arborelui cu parametru gini.



În concluzie, metoda aleasă ca cea mai optimă în estimarea clasei pe setul de date analizat este metoda Arborelui cu parametru gini.

3.6. Limitări ale analizei

Primele limitări sunt date de setul de date analizat. Numărul de înregistări este limitat, iar de aceea a trebuit să fie utilizată metoda cross-validation. Datorită acestui aspect, se poate ca unele cazuri să nu fi fost abordate. De exemplu, în cadrul Naive Bayes se poate să nu fi existat anumite perechi, remedierea fiind oferită de factorul Laplace.

O altă limitare adusă de acest set este reprezentată de atributele ce au fost înregistrate. O parte dintre ele erau asemănătoare, iar în cadrul unui chestionar puteau aduce confuzie clienților, rezultând în atribute cu corelare mare.

Aceste limitări pot duce la o scădere a rezultatelor obținute prin metodele prezentate anterior.

4. Concluzia

În urma analizei prezentate anterior, se va folosi metoda Arborelui cu parametru gini pentru a putea răspunde la setul de întrebări de cercetare.

Pentru primele întrebări „Există vreo legătură între datele referitoare la client (Gender, Customer Type, Age, Flight distance) și gradul de satisfacție?” și „În cazul în care există, cât de puternică este legătura?”, a fost analizat gradul de importanță al atributelor. Cele mai

importante atribute au fost în ordine: Online boarding, Seat comfort, Type of travel, Class, Cleanliness, Age group și Food and drink. Astfel, majoritatea atributelor importante sunt referitoare la serviciile prestate. Excepția este atributul Age group care arată o influență a categoriei de vârsta asupa satisfacției. Însă se poate afirma ca aspectele cele mai importante în determinarea satisfacției clienților sunt cele referitoare la servicii, iar datele personale au o influență mai mică.

Ultima întrebare „Este posibilă realizarea unei estimări dacă clientul va fi sau nu satisfăcut, având în vedere părerea acestuia referitoare la serviciile oferite de compania aeriană?" are un răspuns pozitiv. Se va putea estima satisfacția clientului în funcție de opiniile referitoare la serviciile prestate de compania aeriană. Estimarea va fi cu o acuratețe bună și se vor putea detecta în mod special clienții nemulțumiți spre a se putea încerca o remediere a situației.

În concluzie, analiza a dat rezultate pozitive în raport cu întrebările de cercetare propuse. Astfel, metoda aleasă poate oferi o predicție de încredere pentru gestionarea satisfacției clienților în cadrul companiei aeriene.