Predicția calității vinului

1. Contextul problemei (Introducere)

Lucrarea de față are ca și scop determinarea factorilor din compoziția unui vin care influențează calitatea acestuia și astfel să ajutăm brandurile de vin să își perfecționeze procesul de producție pentru a-și putea crește vânzările.

Dimensiunea pieței vinului din 2020 este estimată la 325 de miliarde de dolari în întreaga lume și se estimează că va ajunge la 445 de miliarde de dolari până în 2027. Companiile investesc pentru a îmbunătăți producția și vânzarea de vin. Evaluarea calității este crucială și depinde în mare măsură de gustul subiectiv al experților în vinuri.

Vinificarea este procesul prin care se produce vinul și include toate etapele prin care trece mustul până se transformă în vin. Astfel procesul de producție este foarte important în determinarea calității vinului și presupune mai mulți pași în care trebuie să se țină cont de o serie de condiții pentru a obține calitatea dorită. De exemplu la recoltarea strugurilor se urmărește nivelul de zahăr din struguri cât și maturitatea fenolică (dată, de exemplu, de culoarea și gustul pieliței sau semințelor strugurelui), aciditatea, pH-ul boabelor etc. iar în ceea ce privește procesul de fermentație contează foarte mult temperatura, timpul și cantitatea de drojdie și alte ingrediente adăugate în timpul acestui proces. De asemenea, compoziția și gustul produsului finit pot fi influențate și dacă se îndepărtează sau nu ciorchinele înainte de zdrobire sau dacă se lasă sau nu strugurii la fermentat cu tot cu pieliță.

După finalizarea tuturor acestor procese cei care produc vin în scopuri comerciale supun vinul și unor teste de laborator, printre care se numără: test de aciditate totală, test de aciditate volatilă, dioxid de sulf liber, dioxid de sulf total, teste pentru stabilirea concentrației alcoolice, test de pH, zahăr total, zahăr reducător etc. Aceste atribute fizico-chimice ale vinului care afectează gustul pot fi măsurate în mod obiectiv și pot fi modificate de vinificatori.

Fiindcă gusturile diferă de la om la om este destul de dificil să determini printr-un număr limitat de degustări ce combinație de parametrii finali va fi cea mai preferată de public pentru a alege rețeta perfectă de vin. Cu ajutorul unui set de date vast format din note date produselor de către clienții diferitelor branduri de vin ne propunem să studiem care sunt factorii finali care afectează cel mai mult calitatea vinului iar în funcție de aceștia firmele își vor putea ajusta procesul de producție astfel încât să obțină combinația perfectă de parametrii.

Prin intermediul acestui studiu dorim să obținem cea mai bună ecuație a unui vin de calitate superioară și să răspundem la întrebarea care o au toți producătorii de vin de pe piață: „**Ce face un vin să fie bun?**”.

Alte întrebări care ne vor ajuta să ajungem la răspunsurile dorite sunt: „Cât de puternică este relația dintre rezultatele finale ale testelor de laborator și calitatea vinului?”, „Ce parametrii finali influențează cel mai mult calitatea vinului?” și „Cu ce acuratețe putem prezice calitatea unui vin știindu-i rezultatele testelor finale de laborator?”.

2. Descrierea setului de date

Setul de date pe care s-a realizat acest studiu a fost procurat de pe siteul *kaggle.com* și se numește „Red Wine Quality”. În acest set de date se găsesc 1599 de observații, 11 variabile numerice independente continue, reprezentate de atribute fizico-chimice, dintre care unele pot fi corelate și o variabilă numerică discretă, care reprezintă date senzoriale pe o scară de la 1 la 10 (notele date de către diverși degustători vinurilor).

Acest set de date este potrivit pentru a răspunde întrebărilor de cercetare deoarece acești parametrii finali (nivel de zahar, concentrație de alcool, sulf, etc.) influențează gustul produsului final și în consecință ar trebui să aibă o influență semnificativă asupra calității vinului și a notelor date de către clienți. De asemenea sunt ușor de obținut fiind măsurați la sfârșitul oricărei producții de vin și din ei se poate deduce ecuația pentru obținerea unui vin de calitate care să ajute producătorii să-și ajusteze procesele, timpii și ingredientele în scopul de a obține anumiți parametrii finali preferați de clienți.

Descrierea variabilelor:

• **Aciditate fixă:** dată de acizi nevolatili care nu se evaporă ușor;

• **Aciditate volatilă:** acid acetic din vin care duce la un gust neplăcut de oțet;

• **Acid citric:** acționează ca un conservant pentru creșterea acidității. Când este în cantități mici, adaugă prospețime și aromă vinurilor;

• **Zahar rezidual:** este cantitatea de zahăr rămasă după oprirea fermentației. Cheia este să ai un echilibru perfect între dulceață și acru;

• **Cloruri:** cantitatea de sare din vin;

• **Dioxid de sulf liber:** previne creșterea microbiană și oxidarea vinului;

• **Dioxid de sulf total:** este cantitatea de forme libere + legate de dioxid de sulf.

• **Densitate:** vinurile mai dulci au o densitate mai mare;

• **pH:** descrie nivelul de aciditate pe o scară de 0-14. Majoritatea vinurilor sunt întotdeauna între 3-4 la scara pH-ului;

• **Alcoolul:** disponibil în cantități mici în vinuri;

• **Sulfat:** un aditiv pentru vin care contribuie la nivelurile de dioxid de sulf și acționează ca antimicrobian și antioxidant;

• **Calitate:** care este variabila de ieșire, o notă de la 1 la 10.

În continuare vom analiza setul de date pentru a determina modul în care sunt distribuite datele în funcție de parametrii care îi avem la dispoziție și influența pe care o au atributele față de nota finală dată de către clienți.

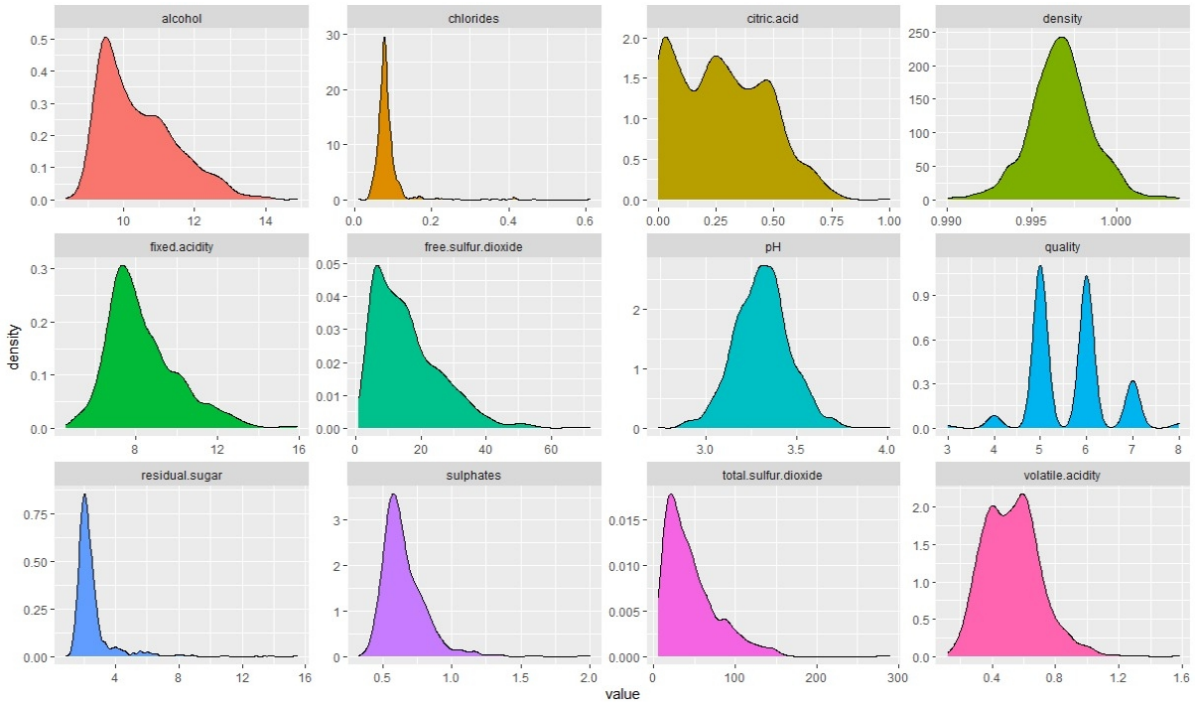


Figura 1- Distribuția datelor

În Figura 1 se mai pot observa intervalele în care sunt concentrate majoritatea valorilor pentru fiecare proprietate. Distribuția alcoolului, a acidității fixe, a zahărului residual, a clorurilor, a dioxidului de sulf liber, a dioxidului de sulf total și a sulfatului au o inclinație pozitivă (cu un număr mare de valori scăzute și cu un număr din ce în ce mai mic al valorilor ridicate). Ph-ul și densitatea au distribuții normale aproape perfecte (valorile medii au cea mai frecventă apariție). Distribuțiile acidului citric și a acidității volatile sunt neregulate.

Se mai poate concluziona că toate datele cu excepția calității sunt de tip continuu, calitatea fiind de tip discret luând doar valori întregi de la 3 la 8. De asemenea se mai poate observa că datele legate de calitate nu sunt distribuite uniform existând mai multe valori pentru vinuri normale (note de 5,6) decât excelente (7, 8) sau de calitate scăzută (3,4). Vom crea o coloană suplimentară numită rating, aceasta va lua următoarele valori discrete: „bad” pentru vinurile cu calitate mai mica decât 5, „average” pentru vinurile cu calitate între 5 și 7 și „good” pentru vinurile cu calitate între 7 și 10.

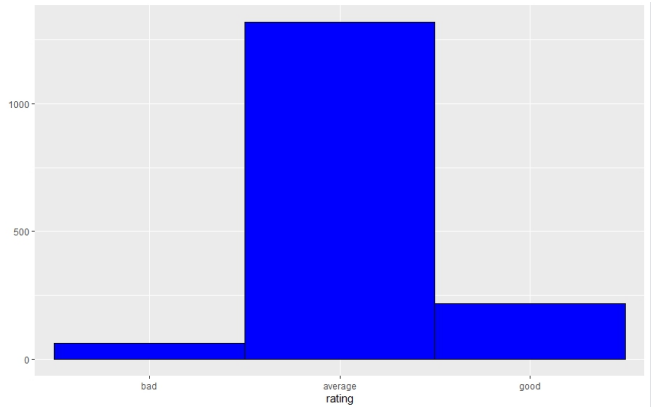


Figura 2 – Distribuția variabilei rating

În figura 2 se poate observa mai bine distribuția calității vinului pe categorii. Deoarece vinurile de bună calitate și de proastă calitate sunt aproape ca cele anormale aici, ar putea fi dificil să obțineți un model precis al calității vinului.

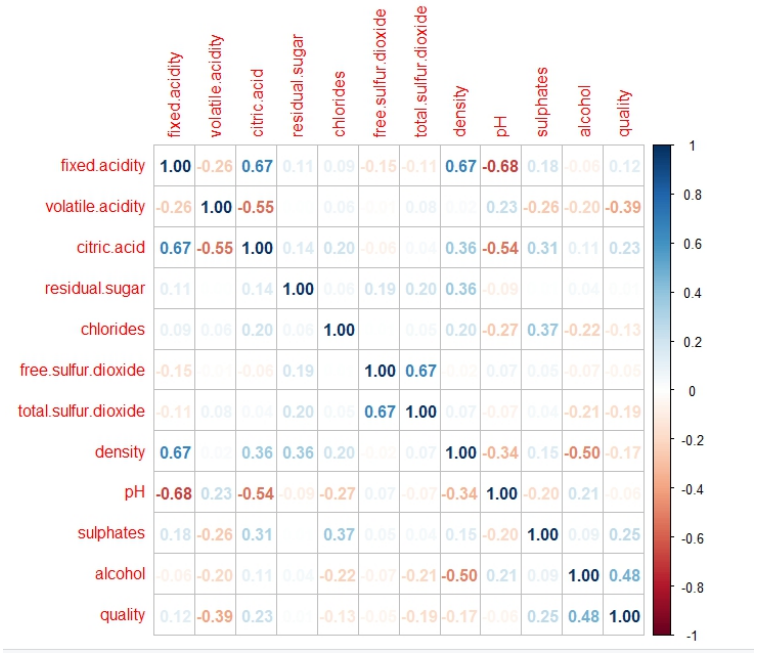
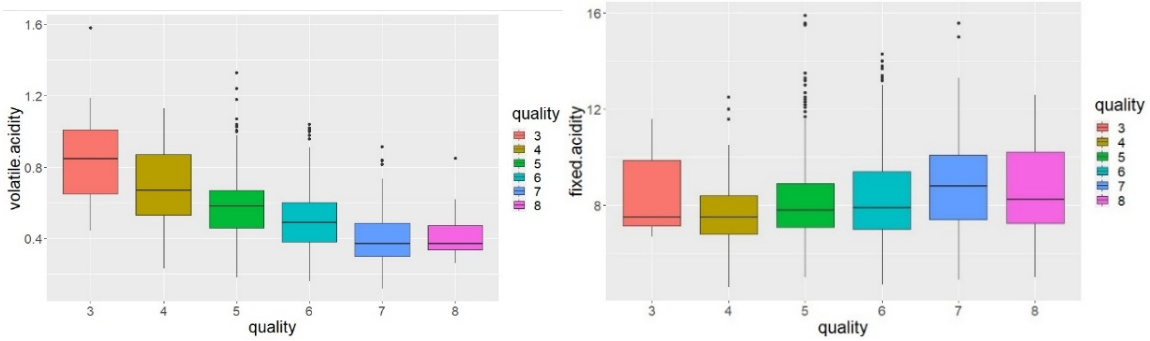


Figura 3 – Corelația între date

În graficul din Figura 3 se observă corelația între atributele înregistrărilor din tabel, din care se poate deduce că calitatea este cel mai mult influențată de valorile atributelor alcool, sulfați, acid citric și aciditate volatilă. De asemenea aciditatea fixă are o corelație foarte puternică cu densitatea, acidul citric și o corelație negativă puternică cu ph-ul ceea ce este logic deoarece ph-ul scade când aciditatea crește (de aceea se pot observa corelații negative puternice între ph și orice atribut care indică valori ale acidității). Alcoolul are o corelație negativă cu densitatea lucru care este evident din faptul că densitatea apei este mai mare decât densitatea alcoolului.



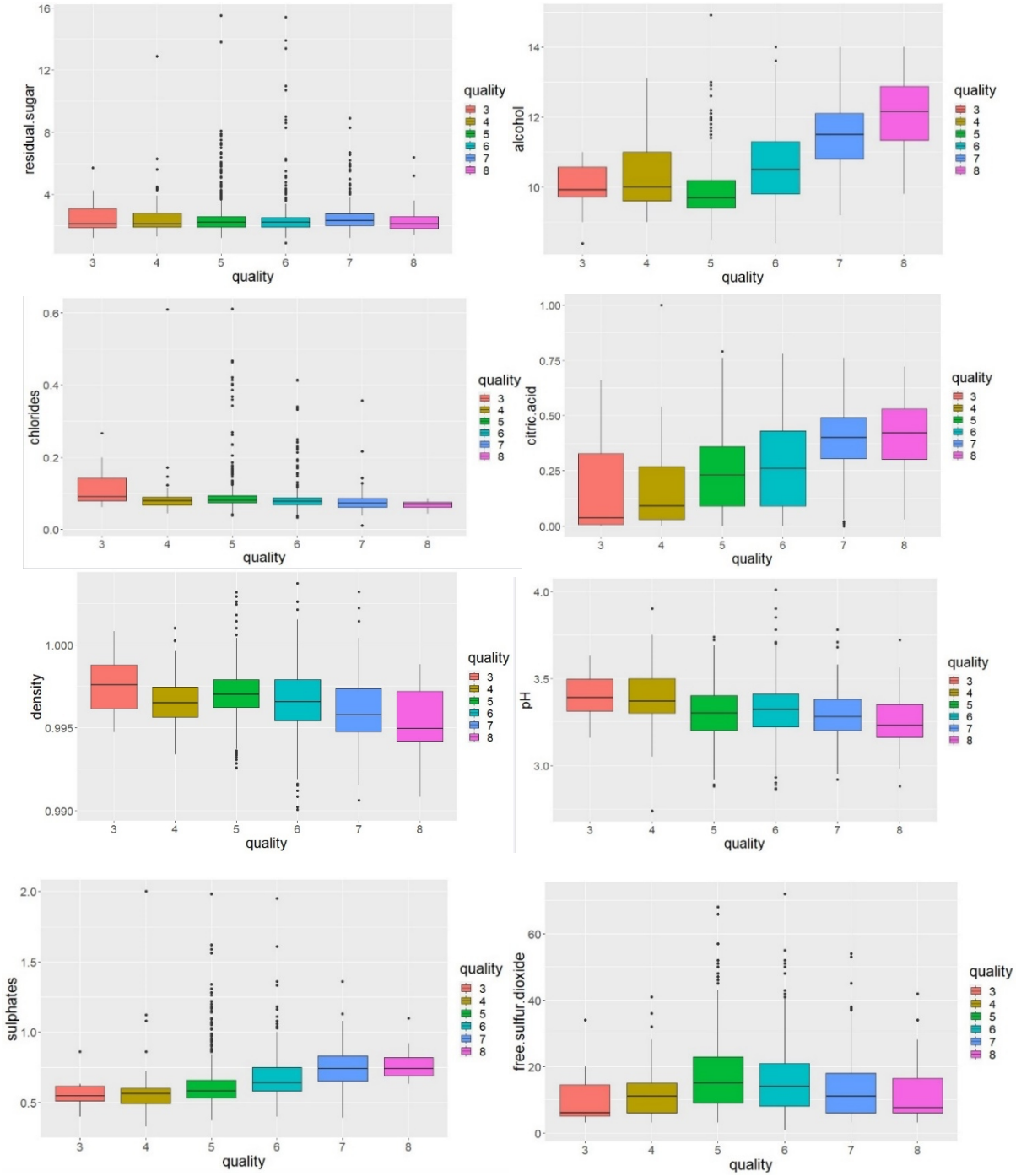


Figura 4 – Candle Plots - variabila calitate și ceilalți factori

Prin analiza graficelor din Figura 4 putem să deducem, la fel ca în cazul graficului de corelație că proprietățile alcool, aciditate volatilă, acid citric și sulfați au o corelație mai mare cu calitatea vinului. Cu cât calitatea vinului crește se poate observa că și media variabilei alcool crește, același lucru este valabil pentru acid citric și sulfat însă acidul volatil pare să aibă un impact negativ asupra calității vinului. Pe măsură ce nivelul acidului volatil crește, calitatea vinului se degradează. Aciditatea fixă, zahărul rezidual, clorurile, dioxidul de sulf par să nu aibă un impact mare asupra calității. Se mai poate observa o corelație mică între calitate și densitate și ph. Cu cât aceste 2 variabile sunt mai mici cu atât pare să crească

calitatea. Corelația între densitate și calitate poate fi explicată prin corelația existentă între calitate și alcool (procentul de alcool influențând densitatea vinului după cum am văzut în Figura 3).

3. Analiza datelor, rezultate și discuții

În continuare vom încerca să creăm niște modele care să ne ajute să înțelegem mai bine legătura dintre calitatea unui vin și atributele chimice ale acestuia și care să ne ajute să prezicem calitatea unui vin având datele necesare despre aceste atribute.

Pentru crearea modelelor vom folosi:

1. Regresia liniară: aceasta ne va ajuta să detectăm care sunt atributele semnificative statistic și să testăm interacțiunile dintre atribute. Datorită distribuției neuniforme a variabilei calitate datorată faptului că această variabilă este de tip discret, preconizăm că regresia liniară nu ne va ajuta să obținem un model care să prezică cu acuratețe calitatea. De aceea am considerat că trebuie să folosim și regresia pentru a putea obține un model cu acuratețe mai ridicată.

2. Regresie logistică: vom defini un rating de 6 sau mai mare ca bun (1) și sub 6 ca rău (0)

3. Arbori de decizie și Random forest pentru o exemplificare mai potrivită a rezultatelor obținute pentru un cititor nespecializat.

3.1. Regresia liniară

Pentru început vom construi un model de regresie liniară cu toți parametrii pe care îi avem la dispoziție pentru a putea observa relevanța fiecărui parametru în predicția calității.

	Coeficient	Std. error	t-statistic	p-value
intercepta	1.997e+01	2.119e+01	0.942	0.3463
aciditate fixă	2.499e-02	2.595e-02	0.963	0.3357
aciditate volatilă	-1.084e+00	1.211e-01	-8.948	< 2e-16
aciditate citrică	-1.826e-01	1.472e-01	-1.240	0.2150
zahăr rezidual	1.633e-02	1.500e-02	1.089	0.2765
cloruri	-1.874e+00	4.193e-01	-4.470	8.37e-06
dioxid de sulf liber	4.361e-03	2.171e-03	2.009	0.0447
dioxid de sulf total	-3.265e-03	7.287e-04	-4.480	8.00e-06

densitatea	-1.788e+01	2.163e+01	-0.827	0.4086
pH	-4.137e-01	1.916e-01	-2.159	0.0310
sulfați	9.163e-01	1.143e-01	8.014	2.13e-15
alcool	2.762e-01	2.648e-02	10.429	< 2e-16

Measure	Value	p-value
RSE	0.648	
R^2	0.3561	
F statistic	81.35	< 2.2e-16

Tabel 1 Regresie liniară cu toate atributele și matricea de confuzie

Coeficienții din primul tabel sunt folosiți pentru construirea ecuației calității astfel:

$$quality \approx \beta0 + \beta1 \times X1 + \beta2 \times X2 + \beta3 \times X3 + ... + \beta n \times Xn,$$

β0- este interceptaa și reprezintă nivelul calității atunci când toate variabilele care influențează sunt egale cu 0. X1,X2,X3 etc. sunt valorile variabilelor pentru diferite înregistrări/cazuri. Acestea sunt înmulțite cu coeficienții lor și adunate pentru a obține calitatea prezisă de model.

Std. error ne spune (în medie) cu cât estimarea coeficientului respectiv diferă de valoarea reală a acestuia. Se folosesc pentru determinarea intervalelor de încredere pentru parametru.

T-statistic: $\beta i / SE(\beta i)$ ne spune numărul de deviații standard cu care estimarea parametrului βi se depărtează de valoarea reală.

P-value indică probabilitatea ca să observăm o asociere între predictor și variabila dependentă datorită șansei. O valoare mica a lui p-value ne permite să tragem concluzia că există o asociere între predictor și variabila dependentă. În consecință putem deduce că există 7 atribute care au o corelație semnificativă statistic cu calitatea vinului. Dintre acestea, aciditatea volatilă, clorurile, dioxidul de sulf total și pH-ul sunt corelate negativ cu calitatea vinului, în timp ce dioxidul de sulf liber, sulfații și alcoolul sunt corelați pozitiv.

R^2 este de 35,6%, ceea ce înseamnă că 35,6% din variația ratingului vinului este explicată de model. Aceasta este o valoare mică ceea ce ne sugerează că modelul nu explică variația calității cu o acuratețe mare.

RSE reprezintă mărimea medie cu care variabila dependentă quality va devia de la linia de regresie. O valoare mică indică o potrivire mai mare a modelului.

Valori mari ale lui F statistic indică că există o relație între variabila prezisă și predictori.

	Coeficient	Std. error	t-statistic	p-value
intercepta	2.4300987	0.4029168	6.031	2.02e-09
aciditatea volatilă	-1.0127527	0.1008429	-10.043	< 2e-16
cloruri	-2.0178138	0.3975417	-5.076	4.31e-07
dioxid de sulf liber	0.0050774	0.0021255	2.389	0.017
dioxid de sulf total	-0.0034822	0.0006868	-5.070	4.43e-07
pH	-0.4826614	0.1175581	-4.106	4.23e-05
sulfați	0.8826651	0.1099084	8.031	1.86e-15
alcool	0.2893028	0.0167958	17.225	< 2e-16

Table 3 Regresie liniară cu atributele semnificative

În continuare vom crea un model doar cu variabilele semnificative statistic.

Measure	Value	p-value
RSE	0.6477	
R^2	0.3595	
F statistic	127.6	< 2.2e-16

Tabel 4 Rezultatele regresiei liniare cu atributele semnificative

În cazul acestui model RSE și R^2 s-au îmbunătățit însă nesemnificativ. Valoarea lui F statistic a crescut ceea ce indică că în cadrul acestui model predictorii au o relație mai strânsă cu valoarea prezisă.

Înainte de a trage concluzii vom încerca să combinăm variabilele între ele pentru a observa cum interacționează acestea și pentru a încerca să obținem un model mai bun.

	Coeficient	Std. error	t-statistic	p-value
intercepta	3.793e+00	4.530e+00	0.837	0.402463
aciditatea volatilă	-3.208e+00	2.312e+00	-1.388	0.165464
cloruri	6.637e-01	1.362e+01	0.049	0.961151
dioxid de sulf liber	-3.343e-02	5.182e-02	-0.645	0.518963
dioxid de sulf total	1.974e-02	1.689e-02	1.169	0.242697
pH	-5.061e-01	1.333e+00	-0.380	0.704165
sulfați	-3.920e+00	2.781e+00	-1.410	0.158818
alcool	3.560e-01	3.379e-01	1.053	0.292365
aciditatea volatilă: cloruri	1.086e+00	2.510e+00	0.433	0.665211
aciditatea volatilă: dioxid de sulf liber	-1.731e-02	1.399e-02	-1.237	0.216236
aciditatea volatilă: dioxid de sulf total	1.562e-02	4.619e-03	3.382	**0.000737**

aciditatea volatilă: pH	2.671e-01	6.622e-01	0.403	0.686685
aciditatea volatilă: sulfați	4.224e-01	6.581e-01	0.642	0.521100
aciditatea volatilă: alcool	6.144e-02	1.107e-01	0.555	0.578947
cloruri: dioxid de sulf liber	2.401e-02	5.833e-02	0.412	0.680676
cloruri: dioxid de sulf total	-2.356e-02	2.537e-02	-0.929	0.353240
cloruri: pH	2.473e+00	4.130e+00	0.599	0.549464
cloruri: sulfați	-3.096e-01	1.809e+00	0-.171	0.864116
cloruri: alcool	-9.703e-01	5.744e-01 .	-1.689	0.091369
dioxid de sulf liber: dioxid de sulf total	-1.510e-05	4.544e-05	-0.332	0.739654
dioxid de sulf liber: pH	1.284e-02	1.508e-02	0.851	0.394890
dioxid de sulf liber: sulfați	-1.941e-02	1.397e-02	-1.389	0.165039
dioxid de sulf liber: alcool	1.690e-03	2.412e-03	0.700	0.483798
dioxid de sulf total: pH	-7.655e-03	4.695e-03	-1.631	0.103179
dioxid de sulf total: sulfați	-8.747e-03	3.912e-03	-2.236	**0.025496**
dioxid de sulf total: alcool	1.048e-04	7.319e-04	0.143	0.886201
pH: sulfați	8.473e-01	8.006e-01	1.058	0.290034
pH:alcool	-7.044e-02	9.769e-02	-0.721	0.470995
sulfați: alcool	2.747e-01	1.222e-01	2.248	**0.024697**

Tabel 5 Regresie liniară cu interacțiunile dintre atribute

Din acest model putem deduce că există o interacțiuni între aciditate volatilă și dioxidul total de sulf, între dioxid total de sulf și sulfați și între sulfați și alcool.

În continuare vom construi un ultim model compus din parametrii semnificativi și interacțiunile semnificative dintre aceștia, după care vom realiza un tabel pentru a compara calitatea modelului format doar din parametrii semnificativi cu cel care include și interacțiunile.

	Coeficient	Std. error	t-statistic	p-value
intercepta	4.086651	0.829468	4.927	9.23e-07
aciditatea volatilă	-1.128660	0.162560	-6.943	5.58e-12
cloruri	-1.812381	0.411686	-4.402	1.14e-05
dioxid de sulf liber	-0.003508	0.004755	-0.738	0.460801
dioxid de sulf total	0.005670	0.001686	3.364	0.000787
pH	-0.490760	0.115803	-4.238	2.39e-05
sulfați	-1.445383	1.144476	-1.263	0.206803
alcool	0.074937	0.072617	1.032	0.302250
aciditatea volatilă: dioxid de sulf liber	0.015806	0.008292	1.906	0.056823
dioxid de sulf total: sulfați	-0.014133	0.002389	-5.915	4.06e-09

sulfați: alcool	0.319117	0.106774	2.989	0.002845

Tabel 6 Regresie liniară cu elementele semnificative și dependențele semnificative

	Value	p-value
RSE	0.6357	
R^2	0.3804	
F statistic	99.13	< 2.2e-16

Tabel 7 Rezultatele regresie liniară cu elementele semnificative și dependențele semnificative

Acest model are valoarea lui R^2 de 38% deci s-a îmbunătățit față de modelul precedent. Prezența unei interacțiuni semnificative indică faptul că efectul unui atribut asupra calității vinului este diferit la diferite valori ale celuilalt atribut.

Model	RSE	R^2	F statistic
Model cu parametrii semnificativi	0.6477	0.3595	127.6
Model cu parametrii semnificativi și interacțiuni	0.6357	0.3804	99.13

Tabel 8 Tabel comparativ între modelele cu parametri semnificativi și parametri nesemnificativi

Modelul fără interacțiune are valoarea R^2 mai mică față de cea a modelului cu interacțiune, adică nu poate explica la fel de bine variația calității. De asemenea RSE-ul este mai mic la modelul cu interacțiune ceea ce înseamnă că prezice calitatea cu o mai mare acuratețe. În final vom alege modelul cu interacțiune ca model final pentru a trage o serie de concluzi.

Concluzii:

-aciditatea volatilă, clorurile, sulfații, alcoolul, dioxidul de sulf total și ph-ul influențează în cea mai mare măsură calitatea vinului;

-un nivel mai scăzut de aciditate volatilă, cloruri, sulfați și pH duce la o mai bună calitate a vinului;

-conținutul ridicat de alcool pare să producă vinuri mai bune;

- Există interacțiuni semnificative între: alcool și sulfați - corelație pozitivă cu calitatea vinului;

dioxid de sulf total și sulfați - corelație negativă cu ratingul vinului. Deci pentru o calitate cât mai bună același nivel de sulfați ar trebui să se potrivească cu un nivel mai ridicat de alcool și un nivel mai scăzut de dioxid de sulf total;

- R-pătratul ajustat este de doar 38,0%, ceea ce implică un nivel limitat de potrivire a modelului.

Regresia logistică

Presupunerea de la începutul acestei lucrări conform căreia regresia liniară nu va oferi un model cu o precizie ridicată s-a adeverit în discuția de mai sus, astfel vom încerca să găsim un model cu o acuratețe mai ridicată cu ajutorul regresiei logistice.

După cum am descoperit în studierea variabilelor setului de date calitatea vinului este o variabilă discretă cu valori între 3 și 8, motiv pentru care aleg să factorizez acest element și să împart valorile în vin cu o calitate slabă (pentru 3,4 și 5) și vin cu o calitate ridicată pentru ratingurile de 6,7 și 8. După o astfel de împărțire se formează o împărțire echilibrată cu 744 de observații pentru calitatea scăzută și 855 de observații pentru o calitate ridicată.

În primul rând vom crea o regresie ce conține toate atributele puse la dispoziție pentru a studia relevanța lor în regresie logistică. Astfel de obțin rezultatele:

	Coeficient	Std. error	t-statistic	p-value
intercepta	43.714113	96.875379	0.451	0.6518
aciditate fixă	0.135187	0.119837	1.128	0.2593
aciditate volatilă	-2.683749	0.567630	-4.728	2.27e-06
aciditate citrică	-1.221435	0.661039	-1.848	0.0646
zahăr rezidual	0.047584	0.063060	0.755	0.4505
cloruri	-2.387465	1.723719	-1.385	0.1660
dioxid de sulf liber	0.028169	0.009786	2.878	0.0040
dioxid de sulf total	-0.017624	0.003332	-5.289	1.23e-07
densitatea	-51.230665	98.875759	-0.518	0.6044
pH	-0.812178	0.852951	-0.952	0.3410
sulfați	2.455264	0.522958	4.695	2.67e-06
alcool	0.934836	0.126298	7.402	1.34e-13

Tabel 9 Regresie logistică cu atributele setului de date

Se poate observa că atributele care au relevanță statistică au rămas aceleași ca și în cazul regresie liniare și anume:aciditatea volatilă, aciditatea citrică, dioxid de sulf liber, dioxid de sulf total, sulfați și alcool. Având in vedere rezultatele obținute in modelul de mai sus și faptul că există atribute care nu au o relevanță mare pentru modelul creat, se creează un model în care să punem proprietățile relevante, eliminând proprietățile care au valori ridicate pentru p-value.

Conform matricei de confuzie cu un prag de 0.4 de mai jos acest model prezice din cele din cele 223 de observații care încadrează vinul ca fiind de o calitate proastă modelul identifică 143 de observații corect, având o acuratețe de 64,12%, iar în cazul vinurilor cu o calitate mai ridicată modelul identifică 210 valori corect dintr-un total de 256, având o acuratețe de 82,03%. Acuratețea totală a modelului este de 73,36%.

O altă informație importantă este aceea că regresia logistică prezintă o acuratețe semnificativ mai ridicată față de rezultatele regresiei liniare observații prezentate în tabelul 2 (Măsurătorile regresiei liniare cu atributele setului de date).

0.4		Valori reale	
		bad	good
Valori prezise	FALSE	143	46
	TRUE	80	210

Tabelul 10 Matricea de confuzie a regresie logistică cu atributele setului de date

Pentru un studiu mai amănunțit am selectat doar atributele care au o semnificație mai mare și am obținut următoarele rezultate:

	Coeficient	Std. error	t-statistic	p-value
intercepta	-5.348992	1.722945	-3.105	0.001906
aciditate volatilă	-2.135124	0.452047	-4.723	2.32e-06
cloruri	-3.462229	1.587890	-2.180	0.029228
dioxid de sulf liber	0.032512	0.009496	3.424	0.000618
dioxid de sulf total	-0.019458	0.003147	-6.183	6.31e-10
pH	-1.219403	0.512457	-2.380	0.017335
sulfați	2.317394	0.497739	4.656	3.23e-06
alcool	0.951241	0.087092	10.922	< 2e-16

0.4		Valori reale	
		bad	good
Valori prezise	FALSE	103	26
	TRUE	120	230

Tabel11 Regresia atributelor relevante și matricea de confuzie

În acest caz modelul reușește să identifice corect 230 de observații care au o calitate a vinului nu foarte bună dintr-un total de 256 (acuratele de 89,84%) însă reușește să explice destul de puține cazuri în care calitatea vinului este un bună având o acurateță de 46,18%. Acuratețea acestui model este de de 69,51% reușind să identifice corect 333 de observații dintr-un total de 479.

Având în vedere aceste rezultate am combinat toate atributele din setul de date și dependențele cele mai puternice pentru vedea acuratețea unui astfel de model. Astfel se obțin următoarele rezultate:

	Coeficient	Std. error	t-statistic	p-value
intercepta	38.453044	99.462135	0.387	0.699045
aciditate fixă	0.093341	0.120883	0.772	0.440020
aciditate volatilă	-2.423944	0.576366	-4.206	2.6e-05
aciditate citrică	-1.413020	0.669280	-2.111	0.034751

zahăr rezidual	0.042756	0.064358	0.664	0.506465
cloruri	-3.020422	1.897328	-1.592	0.111399
dioxid de sulf liber	0.027069	0.009773	2.770	0.005608
dioxid de sulf total	0.013075	0.008322	1.571	0.116167
densitatea	-42.547844	100.661305	0.423	0.672526
pH	-1.436943	0.865572	-1.660	0.096893
sulfați	0.654148	6.787429	0.096	0.923222
alcool	0.623888	0.447929	1.393	0.163672
dioxid de sulf total: sulfați	-0.048126	0.012421	-3.874	0.000107
sulfați: alcool	0.516030	0.663892	0.777	0.436993
aciditatea volatilă: dioxid de sulf liber	0.018915	0.040274	0.470	0.638596

0.4		Valori reale	
		bad	good
Valori prezise	FALSE	145	50
	TRUE	78	206

Tabelul 12 Rezultatele și matricea de confuzie a regresia logistică cu atributele setului de date și interacțiunile semnificative

Pentru acest model se poate observa in matricea de confuzie rezultate mai bune, astfel pentru un vin cu calitate scăzută modelul reușește să identifice 65,02% din valori, un rezultat mai bun față de modelul anterior on care acuratețea era de 64,12%. Pentru observațiile care încadrează vinul la o calitate bună se poate observa și aici o îmbunătățire reușind să identifice din cele 256 de valori 206, față de modelul anterior unde identifica 210 (acurateț̦e 80,46%). În consecință și acuratețea totală a modelului are o valoare mai ridicată: 73,27%.

Cu toate că punem în model și variabilele care au o influentă ridicată în calitatea vinului se poate observa că modelul devine nu unul mai bun, ci din potrivă chiar poate identifica mai puține date corect decât în cazul modelului fără aceste variabile, cu toate că marja de eroare nu este una foarte mare ci de 2 observații în cazul vinurilor cu o calitate scăzută și de 4 pentru cazul în care calitatea nu este una foarte bună.

Mai departe am creat un model pentru atributele relevante ale setului de date și pentru interacțiunile cele mai semnificative unde s-au obținut următoarele rezultate:

	Coeficient	Std. error	t-statistic	p-value
intercepta	-3.893057	4.367357	-0.891	0.372715
aciditate volatilă	-2.067335	0.795868	-2.598	0.009388

cloruri	-4.105455	1.769264	-2.320	0.020318
dioxid de sulf liber	0.021694	0.023718	0.915	0.360359
dioxid de sulf total	0.010045	0.008278	1.213	0.224994
pH	-1.427069	0.520175	-2.743	0.006080
sulfați	1.129573	6.593358	0.171	0.863972
alcool	0.675579	0.409281	1.651	0.098811
aciditatea volatilă: dioxid de sulf liber	0.018915	0.040274	0.470	0.638596
dioxid de sulf liber total: sulfați	-0.045864	0.012528	-3.661	0.000251
sulfați: alcool	0.436917	0.641584	0.681	0.495872

0.4		Valori reale	
		bad	good
Valori prezise	FALSE	116	29
	TRUE	107	227

Tabel 13 Regresie logistică cu atributele relevante și dependențele cele mai puternice

Acest model identifică 116 observații corect din totalul de 223 care clasifică vinul ca fiind unul de calitate scăzută și 227 de înregistrări corecte pentru un vin de calitate superioară dintr-un total de 256, acuratețea acestui model fiind de 71,6%.

Modelul care a explicat corect cele mai multe înregistrări a fost cel care a fost creat pe baza cu toate atributele setului de date și cu dependențele cele mai semnificative, conform tabelului de mai jos:

	Acuratețe calitate scăzută	Acuratețe calitate ridicată	Acuratețe model
Toate atributele	64,12%	82,03%	73,36%
Atribute relevante	89,84%	46,18%	69,51%
Toate atributele și dependențele relevante	65,02%	80,46%	73,27%
Atribute relevante și dependențe relevante	52,01%	88,67%	71,6%

Tabel 14 Comparația modelelor regresie logistică

După cum am preconizat la începutul acestei lucrări cu ajutorul regresiei logistice s-a putut crea un model care are o acuratețe mai ridicată decât am putea obține cu ajutorul unei regresii liniare unde acuratețea de 38,0%. De altfel se poate observa că toate cele 4 modele create au o precizie considerabil mai mare decât precizia regresiei liniare.

Pentru o ipoteză relevantă pentru a alege un vin cu o calitate superioară, având în vedere rezultatele obținute la regresia liniară și la cea logistică este recomandat să se folosească un model creat cu ajutorul regresiei logistice. Pentru a preconiza un vin cu o calitate mai mare cel mai potrivit model este cel în care sunt incluse toate atributele care au un p-value mic, el reușind să identifice cele mai multe înregistrări corect. Pentru a preconiza un vin cu o calitate slabă cel mai bun rezultat a fost dat de modelul în care au fost incluse atributele relevante și dependențele cele mai strânse, având exactitate de 88,67%.

Arbori de decizie și Random forest

Arborii de decizie sunt printre cele mai intuitive modele de clasificare pe care oamenii pot să le interpreteze.

Construirea arborilor de decizie constă în împărțirea setului de date în subseturi mai mici cărora li se asociază clasa cu cele mai multe apariții din subset. Împărțirea se face prin intermediul variabilelor predictor care sunt selectate astfel încât să se minimizeze rata de eroare E, indexul Gini sau entropia.

$$E = 1 - \max p_{mk}$$

unde p_{mk} reprezintă proporția instanțelor din subsetul m care au clasa țintă k.

Indexul Gini: măsoară varianța totală pe toate cele K clase.

Entropia: va lua valori mici dacă un nod este pur.

Vom construi 3 arbori, fiecare va folosi unul dintre aceste criterii de selecție a testului din nodurile arborelui.

1. **Arborele obținut folosind ca și criteriu de selecție rata de eroare E.**

Inițial am împărțit setul de date în două subseturi, unul pentru antrenarea modelului și unul pentru testarea lui.

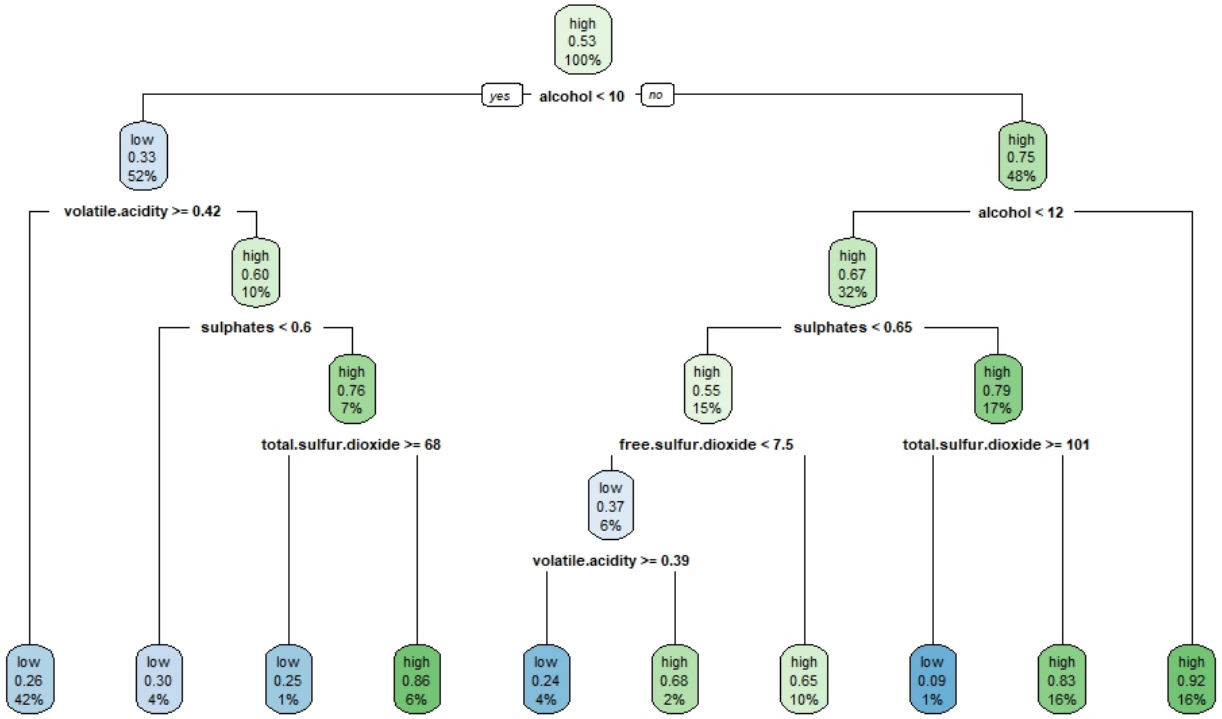


Figura 5 – Arbore de decizie folosind rata de eroare E

În figura 5 primul nod al arborelui (nodul rădăcină) semnifică că 100% din date au în medie o probabilitate de 53% să fie de calitate superioară. După ce datele sunt împărțite în două seturi în funcție de testul ales de algoritm, în cazul nostru *alcool<10*, nodul din stânga indică că 52% din datele din setul inițial au fost repartizate în acest subset, în care există o probabilitate de doar 33% ca vinul să fie de o calitate superioară iar nodul din dreapta ne indică că 48% din datele din setu inițial au fost repartizate în acest subset în care există o probabilitate de 75% ca vinul să fie de calitate superioară. Deci se poate observa că predictorul cel mai semnificativ este procentul de alcool. Valori mari ale acestui predictor tind să crească calitatea vinului. Predictorii aleși de algoritm în nodurile superioare ale arborelui tind să aibă o influență mai puternică asupra valorii prezise. Aciditatea volatilă, sulfații, alcoolul, dioxidul de sulf total joacă cel mai important rol în determinarea calității vinului, concluzie la care am ajuns și după analiza modelului obținut prin regresie liniară.

```
          Reference
Prediction low high
      low  172   81
      high  52  176

               Accuracy : 0.7235
                 95% CI : (0.6812, 0.763)
    No Information Rate : 0.5343
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.449

 Mcnemar's Test P-Value : 0.01519

            Sensitivity : 0.7679
            Specificity : 0.6848
         Pos Pred Value : 0.6798
         Neg Pred Value : 0.7719
             Prevalence : 0.4657
         Detection Rate : 0.3576
   Detection Prevalence : 0.5260
      Balanced Accuracy : 0.7263

       'Positive' Class : low
```

Figura 6 Rezultate pentru arborele obținut folosind ca și criteriu de selecție rata de eroare E.

Am testat acest model prin intermediul setului de date de test și am obținut următoarele rezultate:

Acuratețea: proporția de instanțe prezise corect:

$$172+176/172+176+52+81=0.72$$

Senzitivitatea: proporția vinurilor de calitate inferioară care sunt identificate corect:

$$172/172+52=0.76$$

Specificitate: proporția vinurilor de calitate superioară care sunt identificate corect:

$$76/176+81=0.68$$

Din valorile obținute putem deduce că modelul obținut poate prezice calitatea unui vin cu o acuratețe de 72%, poate prezice care sunt vinurile de o calitate inferioară cu o acuratețe de 76% iar vinurile de calitate superioară cu o acuratețea de 68%.

2. În continuare vom folosi ca și criteriu de selectare a predictorilor **entropia.**

```
1) root 1118 1544.000 high ( 0.46512 0.53488 )
  2) alcohol < 10.25 580  736.500 low ( 0.66897 0.33103 )
    4) sulphates < 0.575 250  235.700 low ( 0.82000 0.18000 ) *
    5) sulphates > 0.575 330  453.500 low ( 0.55455 0.44545 )
     10) total.sulfur.dioxide < 59 218  298.600 high ( 0.43578 0.56422 )
       20) volatile.acidity < 0.555 127  149.500 high ( 0.27559 0.72441 ) *
       21) volatile.acidity > 0.555 91  116.700 low ( 0.65934 0.34066 ) *
     11) total.sulfur.dioxide > 59 112  116.400 low ( 0.78571 0.21429 ) *
  3) alcohol > 10.25 538  599.500 high ( 0.24535 0.75465 )
    6) alcohol < 11.55 362  457.100 high ( 0.32597 0.67403 )
     12) sulphates < 0.645 172  237.000 high ( 0.45349 0.54651 ) *
     13) sulphates > 0.645 190  195.600 high ( 0.21053 0.78947 )
       26) total.sulfur.dioxide < 100.5 179  161.800 high ( 0.16760 0.83240 ) *
       27) total.sulfur.dioxide > 100.5 11    6.702 low ( 0.90909 0.09091 ) *
    7) alcohol > 11.55 176   97.740 high ( 0.07955 0.92045 ) *
```

Figura 6 Arborele obținut cu ajutorul predicatorilor entropia

Arborele obținut folosind această metodă a plasat procentul de alcool în rădăcină la fel ca și arborele anterior ceea ce scoate în evidență importanța acestuia. Spre deosebire de arborele anterior, sulful și dioxidul de sulf total sunt plasați pe nivele mai înalte decât aciditatea volatilă însă predictorii semnificativi rămân aceeași.

```
          Reference
Prediction low high
      low  142   62
      high  82  195

               Accuracy : 0.7006
                 95% CI : (0.6575, 0.7412)
    No Information Rate : 0.5343
    P-Value [Acc > NIR] : 7.146e-14

                  Kappa : 0.3949

 Mcnemar's Test P-Value : 0.1133

            Sensitivity : 0.6339
            Specificity : 0.7588
         Pos Pred Value : 0.6961
         Neg Pred Value : 0.7040
             Prevalence : 0.4657
         Detection Rate : 0.2952
   Detection Prevalence : 0.4241
      Balanced Accuracy : 0.6963

       'Positive' Class : low
```

Figura 7 Rezultatele arborelui obținut cu metoda entropia

Acest model are o acuratețe mai mică (70%) față de modelul anterior (72%) și senzitivitate mai mică (63% în comparație cu 76%) ceea ce înseamnă că nu poate prezice la fel de bine care sunt vinurile de calitate inferioară. Cu toate acestea specificitatea este mai mare la acest model ceea ce înseamnă că poate prezice mai bine vinurile de calitate înaltă. Specificitatea ar putea fi considerată

Modelul va fi ales în funcție de ce urmărește fiecare comerciant. De exemplu un model cu o specificitate mai bună, cum este acesta, poate fi bun în cazul brandurilor care pun foarte mult accent pe calitate și imaginea pe care o au pe piață motiv pentru care ar prefera să riște ca unele vinuri bune să fie prezise ca fiind de calitate inferioară decât să fie un vin de calitate inferioară identificat ca fiind bun iar cesta să fie lansat pe piață și să dăuneze imaginii brandului.

3. Pentru modelul următor am folosit **indexul Gini** pentru a face selecția predictorilor.

Și în cazul acestui arbore predictorii semnificativi aleși sunt: procentul de alcool, sulfații, aciditatea volatilă, dioxidul de sulf total, clorurile și ph-ul. Acidul citric, densitatea și zahărul rezidual au fost și ele luate în considerare pe ramurile mai inferioare ale arborelui.

Indicii obținuți în urma testelor realizate se pot observa mai jos.

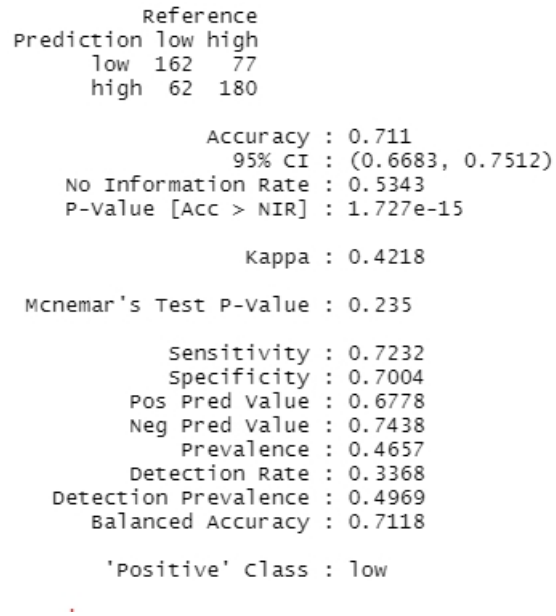


Figura 8 Rezultatele arborelui cu indexul Gini

Pentru a compara mai bine modelele obținute am realizat tabelul de mai jos.

Model	Acuratețe	Senzitivitate	Specificitate
eroare E	72%	76%	68%
entropie	70%	63%	75%
index Gini	71%	72%	70%

Modelul cu cea mai buna acuratețe este cel care folosește eroarea E, cel cu cea mai bună specificitate este cel care folosește entropia și cel cu cea mai bună senzitivitate este cel care folosește eroarea E. Modelul obținut prin indexul Gini pare însă cel mai echilibrat cu valori ridicate pentru toți indecșii analizați. După cum am specificat, modelul ales depinde de nevoile fiecărui beneficiar

Putem trage ca și concluzie că prin intermediul acestor modele care se folosesc de parametrii fizico-chimici ai vinului se poate prezice calitatea cu o acuratețe bună iar factorii care influențează cel mai mult calitatea vinului sunt procentul de alcool, sulfații, aciditatea volatilă și dioxidul de sulf total.

Random forests

În final vom încerca să obținem un model și mai bun folosind metoda Random Forests. Am ales această metodă deoarece în cazul procedurii bagging, modelele rezultate de la diferite eșantioane de bootstrap tind să fie similare, capacitatea de a reduce varianța este limitată iar random forests injectează un element aleatoriu suplimentar. Se obțin mai multe eșantioane bootstrap din setul de date principal pe care se învață câte un arbore CART iar la fiecare proces de împărțire a setului, căutarea atributului este realizată doar pe o submulțime (extrasa aleatoriu) de m din cei p predictori (în general, m = p/3). Pentru a se obține valoarea de predicție, se face o medie a predicțiilor individuale obținute pentru fiecare arbore obținut.

Indicii obținuți se pot observa în figura ce urmează.

```
          Reference
Prediction low high
      low  175   53
      high  49  204

               Accuracy : 0.7879
                 95% CI : (0.7487, 0.8236)
    No Information Rate : 0.5343
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5744

 Mcnemar's Test P-Value : 0.7664

            Sensitivity : 0.7812
            Specificity : 0.7938
         Pos Pred Value : 0.7675
         Neg Pred Value : 0.8063
             Prevalence : 0.4657
         Detection Rate : 0.3638
   Detection Prevalence : 0.4740
      Balanced Accuracy : 0.7875

       'Positive' Class : low
```

Figura 9 Rezultate pentru metoda Random Forest

Acest model are o acuratețe (78%), senzitivitate (78%) și specificitate (79%) mai mare față de modelele la care predicția se face pe un singur arbore. Acest model este cel mai bun pentru a putea realiza predicții asupra calității unui fiind dacă se știu parametrii fizico-chimici ai acestuia. De asemenea mai putem deduce că acești parametrii au un impact semnificativ asupra calității vinului.

4. Concluzia

În urma rezultatelor metodelor de mai sus se poate extrage ideea că alcoolul este factorul care influențează cel mai mult calitatea vinului. Acest lucru se poate extrage în mod vizual din arborele de decizie folosind rata de eroare E, figura 5 din capitolul anterior. Regresiile demonstrează statistic faptul că o dată cu creșterea nivelului de alcool crește și calitatea vinului, cu o probabilitate de 53% să fie de o calitate de superioară, informație extrasă din arborele referit anterior.

Figura corelațiilor dintre date din capitolul anterior (fig. 3) arată că atributele care sunt strâns corelate între ele sunt: aciditatea citrică și sulfați, aciditatea volatilă și dioxidul de carbon, dioxidul de sulf și sulfați și nu în ultimul rând sulfați și alcool. Informație demonstrată statistic și în modelul create în cadrul regresiei liniare care conține dependențele dintre atribute.

Atributul care influențează cel mai mult calitatea vinului este alcoolul lucru demonstrat atât cu ajutorul regresiei liniare cât și vizual în arbore de decizie folosind rata de eroare E unde primul criteriu pe care se face împărțirea este concentrația alcoolică. Astfel

dacă crește concentrația de alcool crește și posibilitatea vinului de a avea o calitate mai bune, lucru descoperit în cadrul regresiilor unde coeficientul acestui atribut este mereu pozitiv. Alte atribute semnificative sunt: sulfați care sunt direct proporționali cu rating-ul, și aciditatea citrică și aciditate volatilă care sunt invers proporționali, calitatea vinului crește dacă acestea scad.

Cea mai mare acuratețe cu care se poate prezice calitatea unui vin pe acest set de date este de 89,84%, model în care sunt luate toate rezultatele cu o relevanță relativ mare.