

**INFLUENȚA CARACTERISTICILOR FIZICO – CHIMICE
ASUPRA EVALUĂRII CALITĂȚII VINULUI ROȘU**

by Irimia Mihaela

Cuprins

Contents

Introducere	4
2. Prezentarea bazei de date	6
3. Determinarea importanței factorilor de predictie	6
3.1. Algoritmului Regresia liniara multipla (MRLM)	7
3.2. Algoritmi MRLM, Random Forest si XGBoost	7
3.2.1. Random Forest si XGBoost	7
3.2.1.1. Impartirea setului de date	8
3.2.1.2. Identificarea si ajustarea hiperparametrilor, selectia, antrenarea si testarea modelului	8
3.2.1.3. Importanta variabilelor	9
3.2.2. Modelul MRLM	9
3.3. Structura setului de date final	10
3.3.1. Operațiile preliminare și de transformare a variabilelor	10
3.3.2. Analiza descriptiva si grafica a variabilelor numerice	11
3.3.3. Analiza grafica a variabilelor numerice	11
3.3.4. Analiza descriptiva si grafica a variabilelor nenumerice	12
3.3.5. Analiza descriptive si grafica a variabilelor numerice in functie de categoriile variabilei categoriala quality_new	12
3.3.6. Identificarea valorilor extreme (outlierilor) si tratarea acestora	13
3.3.6.1. Identificarea valorilor extreme (outlierilor)	13
3.3.6.2. Tratarea outlierilor	14
3.3.6. Matricea de corelatie (MC)	15
3.3.7. Discretizarea variabilor	16
4. Obtinerea arborilor de clasificare prin algoritmi ML	17
4.1. Modele de clasificare	17
4.2. Rezultate si discutii	19
4.2.1. Modele asambiliste	19
4.2.1.1. Rezultatele obtinute pentru modelul generate cu algoritmi RF si XGB	19
4.2.2. Modele bazate pe arbori de decizie CART	21
4.2.2.1. Rezultatele obtinute pentru modelul generate cu algoritmul CART (tidymodels)	21
4.2.2.2. Rezultatele obtinute pentru modelul cu 3 variabile generat cu algoritmul CART (tidymodels)	22
4.2.2.3. Rezultatele obtinute pentru modelul cu trei variabile generat cu algoritmul CART (clasic)	22
4.2.3. Modele bazate pe arbori de decizie CHAID	24
4.2.3.1. Rezultatele obtinute pentru modelul generate cu algoritmul CHAID (tidymodels)	24
4.2.3.2. Rezultatele obtinute pentru modelul cu 3 variabile generat cu algoritmul CHAID (tidymodels)	24
4.2.3.3. Rezultatele obtinute pentru modelul generat cu algoritmul CHAID (clasic)	25
4.2.4. Modele bazate pe arbori de decizie C5.0	27
4.2.4.1. Rezultatele obtinute pentru modelul generate cu algoritmul C5.0	27
4.2.4.2. Rezultatele obtinute pentru modelul cu 3 variabile generat cu algoritmul C5.0 (tidymodels)	28

4.2.4.3. Rezultatele obtinute pentru modelul generat cu algoritmul C5.0 (clasic)	28
4.2.5. Modele bazate pe arbori de decizie C5.0 Boost	29
 4.2.5.1. Rezultatele obtinute pentru modelul generate cu algoritmul C5.0 Boost	29
 4.2.5.2. Rezultatele obtinute pentru modelul cu 3 variabile generat cu algoritmul C5.0 Boost (tidymodels)	30

Introducere

Dezvoltarea tehnologiei informationale a facut posibila colectarea, stocarea si procesarea volumelor mari de date si, adesea, complexe. Toate aceste seturi de date contin informatii valoroase ce pot fi folosite pentru a imbunataati procesul de luare a deciziilor si optimizarea sanselor de succes. Tehnicile de DM vizeaza extragerea de cunostinte de nivel inalt din date brute. Exista foarte multi algoritmi DM fiecare cu avantajele si dezavantajele lui. Abordarea clasica pentru modelarea datelor continue este regresia liniara simpla sau multipla. De la aparitia lor si pana in prezent NN (retele neuronale) reprezinta cel mai popular algoritm ML folosit. Recent, a fost propus si algoritmul SVM (machine vector support) datorita flexibilitatii si capacitatii de invatare neliniara atingand performante predictive inalte .

Cea mai mare problema atunci cand aplicam algoritmi de ML este selectia variabilelor importante si a modelului. Eliminarea variabielor irelevante/redundante conduce la obtinerea unor modele mai simple si mai usor de interpretat care, de obicei, sunt mai performante. Modelele complexe ar putea intampina probleme precum overfitting-ul ceea ce conduce la obtinerea unui model cu o capacitate mica de generalizare, in timp ce un model simplu va prezenta capabilitati de invatare limitate. Majoritatea lagoritmilor ML au hiperparametri ce trebuie reglati/ajustati pentru a obtine o precizie predictiva mai buna.

Sistemele de sprijinire a deciziilor utilizate in industria vinurilor, de catre factorii de decizie, sunt focalizate in principal pe productia de vin. In ciuda potentialului pe care il prezinta tehnicile DM de previzionare a calitatii vinului rosu pe baza proprietatilor fizico – chimice utilizarea lor este destul de rara si se realizeaza in general pe seturi mici de date.

In ultimi deceniu, consumatorii au devenit interesati de produse care provin din anumite zone, precum si de caracteristicile si calitatea acestora. Vinul fiind unul dintre aceste produse. Fiind o bautura alcoolica obtinuta printr-un proces de fermentatie totala sau parciala a strugurilor proaspeti (must), vinul este unul dintre cele mai consumate produse la nivel mondial, iar calitatea lui este influentata in mare masura de materia prima din care provin majoritatea compusilor chimici importanti ai acestuia. Diversitatea soiurilor de struguri, dar si a zonelor geografice in care sunt produsi struguri, conditiilor climatice, respectiv tehnica de vinificatie folosita de producatori au un impact covarsitor asupra produsului final, rezultand astfel o gama variata de vinuri. Provenienta (originea) geografica a fost considerata un factor important, motiv pentru care, multe tari au formulat o serie de linii directoare (Protected Denomination of Origin – PDO, stabuita de UE) si care au rolul de a proteja zonele viticole, dar si a traditiilor de vinificatie. Etichetele DPO adauga valoare vinurilor, care trebuie sa indeplineasca anumite criterii cu privire la soiul de structuri si tehniciile de vinificatie utilizate, respectiv produsul final poate fi replicabil prezentand caracteristici senzoriale similare, specifice zonei din care provine.

Calitatea si proprietatile organoleptice ale unui vin pot fi influente de parametri oenologici, cum ar fi soiul de struguri si procesul de vinificare, aspecte ce pot fi evaluate prin teste chimice/biologice si teste senzoriale sau folosind diferite modele de clasificare in cazul in care sunt colectate date suficiente si relevante. Datorita faptului ca acestei parametri sunt legati de calitatea/pretul vinului, posibilitatea de a gasi o relatie intre parametrii fizico-chimici si calitatea vinului poate fi interesanta, mai ales dacă caracterizarea vinului poate oferi rezultate bune care reduc, de asemenea, costurile operationale in comparatie cu alte metode, cum ar fi folosirea de experti pentru evaluarea calitatii acestuia.

Obiectivul principal al acestei lucrari este de a construi si dezvolta un model ML bazat pe arbori de clasificare pentru a previziona sau clasifica calitatea vinului rosu pe baza proprietatilor fizico-chimice masurabile. Studiul combină datele obtinute din analiza chimică a probelor de vin rosu cu datele senzoriale pentru a construi și dezvolta modele bazate pe arbori de decizie/clasificare capabile să prevadă gustul și miroslul (scorul de apreciere stabilit de un expert/somelier) vinului pe baza concentrației compușilor chimici identificati in urma analizei chimice. Prin acest studiu, se propune stabilirea unui cadru și a pașilor necesari pentru analiza volumelor mari de date (big data) din industria vinicolă. De asemenea, majoritatea studiilor realizate pe seturile RWDS si WWDS sunt focalizate pe aplicarea algoritmilor ML si pe compararea rezultatelor obtinute, insa sunt putine lucrari care analizeaza hierparametri optimi obtinuti pentru modelele generate. Prin urmare, in acest studiu ne propunem sa previzionam calitatea vinului pe baza celor mai importanti predictori prin implementarea unor modele ML pe baza de arbori de clasificare, respectiv determinarea hiperparametrilor optimi sau a modelului optim pentru previzionarea calitatii vinului indicand si valorile utilizate pentru hiperparametri.

Problema de cercetare

Problema de cercetare a acestui studiu este: Cat de eficient este un model ML bazat pe arbori de clasificare pentru a previziona calitatea vinului rosu pe baza proprietatilor fizico - chimice ale acestuia?

Obiective de cercetare propuse

Principalele obiective de cercetare propuse sunt: realizarea cercetarii utilizand setul de date RWDS; implementarea algoritmilor ML: RF, XGBoost, CHAID, CART, QUEST, C5.0 in previzionarea clasei din care face parte o proba de vin rosu (in clasificarea calitatii vinului rosu); identificarea algoritmilor ML performanti si rolul pe care acesti algoritmi ii joaca in predictia calitatii vinului rosu.

Studiile anterioare au demonstrat că analiza proprietăților fizico-chimice ale vinului roșu și utilizarea algoritmilor de învățare automată pot oferi o abordare eficientă pentru predicția calității și gustului acestuia. Cu toate acestea, este necesară o cercetare mai amplă pentru a identifica cele mai bune practici și abordări în acest domeniu. Această lucrare își propune să contribuie la acest obiectiv prin compararea performanțelor diferitelor modele ML în previzionarea calității vinului roșu și identificarea factorilor importanți care influențează această caracteristica.

Lucrarea este structurata in cinci parti importante. In prima parte a cercetarii facem o scurta introducere in tema de cercetare si motivam importanta cercetarii realizate. De asemenea, prezintam problema de cercetare si obiectivele urmarite in cadrul studiului realizat si a lucrarilor relevante pe acest subiect incercand sa identificam elementele mai putin studiate sau unde am putea aduce o contributie sau am putea clarifica anumite aspecte. In a treia parte prezintam metodologia cercetarii abordata, respectiv prezintate informatii cu privire la setul de date RWDS, preprocesarea datelor si analiza exploratorie a datelor, respectiv analiza metodelor ML folosite pentru selectia caracteristicilor semnificative. In a cincea parte a lucrarii prezintam modul de implementare si evaluire a tehnicielor ML bazat pe arbori de decizie utilizati si comparam rezultatele obtinute cu scopul de a identifica algoritmul ML cel mai performant in previzionarea calitatii vinului rosu. Si in final sunt prezентate principalele concluzii care se desprind din acest studiu, directii de cercetare si eventual posibile aplicatii ale rezultatelor obtinute.

Literature review

In analiza RWDS cercetatorii folosesc diverse abordari pentru a analiza factorilor ce influenteaza calitatea unui vin rosu si a relatiilor dintre acesti factori de influenta. Aceste abordari includ metode statistice, metode DM etc. Metodele DM sunt utilizate pentru extragerea informatiilor utile din setul de date, permitand determinarea relatiilor de cauzalitate nedescoperite anterior sau relevante. Scopul DM-ului consta in detectarea, interpretarea si previzionarea patternuri calitative si cantitative in setul de date analizat, generand astfel noi informatii. Tehnicile DM sunt instrumente ce permit analiza legaturii dintre diferite varabile ale setului de date. Aceste tehnici pot fi utilizate pentru clasificare, clusterizare, previzionare, optimizare si realizarea unor rezumate ale datelor. In industria viticola, tehniciile DM sunt utilizate pentru a realiza recomandari cu privire la achizitia unui anumit vin sau la tip de vin pe care consumatorii il pot achizitiona, pe baza evaluarii realizate de un expert in domeniu, pretului si parerea exprimata de consumatori. Exista un numar mare de site-uri si aplicatii mobile ce fac recomandari cu privire la modul in care se alege vinul pe baza unor informatii (<https://www.wine-searcher.com/>, <https://www.go-wine.com/>, <https://www.vivino.com/> etc.).

In ciuda potentialului mare de previzionare a calitatii vinului pe baza rezultatelor obtinute din analize achimice, tehniciile de DM nu sunt utilizate pe scara larga pentru realizarea acestei sarcini.

Motivele pentru care algoritmi de inductie poate devini ineficient sunt adesea legate de: complexitatea setului de date (date lipsa sau incomplete, date zgomotoase (cu valori eronate sau aberante), seturi de date neechilibrate (care conduc la modele cu performante bune pentru clasa dominanta)), selectia algoritmului (alegerea algoritmilor ML nepotrivit, ajustarea ineficienta a hiperparametrilor, complexitatea modelului (aparitia efectului de overfitting/underfitting), evaluarea modelului folosind tehnici de crossvalidare etc.), modul in care sunt gestionate varabilele (selectia inadecvata a variabilelor, prezenta colinearitatii), etc. In plus, prezenta variabilele irelevante pot incurca algoritmii, determinandu-i sa traga concluzii gresite si sa ofere rezultate slabe. Prin urmare, selectia variabilelor are doua caracteristici esentiale: imbunatatirea performantelor algoritmilor si costuri de colectare reduse. Datorita acestor avantaje, selectia varabilelor a inceput, in ultimul timp, sa atraga atentia atat in domeniul ML, cat si in domeniul DM. De-a lungul timpului, au fost dezvoltate mai multe tehnici de selectie a datelor, insa unii dintre cei mai cunoscuti algoritmi de selectie a varabilelor semnificative sunt: MRLM, XGBoost si RF. Analiza semnificatiei varialelor, evalueaza contributia fiecarei varabile la capacitatea de predictie a modelului. Scorul de semnificatie al fiecarei varabile este calculat de unele modele (ex. DT si RF) in functie de cat de frecvent o varibila este utilizata pentru a diviza setul de date si de reducerea impuritatilor pe care il ofera. Prin urmare, pentru a imbunatati performantele unui model si interpretabilitatea acestuia, selectia varabilelor joaca un rol important.

In literatura de specialitate, au fost raportate diverse metode de predictie a calitatii vinului rosu si nu numai, cum ar fi: arbori de decizie si retele neuronale, regresia logistica, SVM, K-NN, XGBoost, Random Forest etc. Un alt studiu a observat ca majoritatea vinurilor din setul de date aveau o calitate medie (5 si 6), iar vinurile de calitate slabă erau mult mai numeroase decât cele de calitate bună. Conform acestui studiu, cel mai bun model pentru prezicerea calității vinului roșu a fost modelul Support Vector Machine (SVM) cu o acuratețe de 92.5%. Ajustarea parametrilor de tuning a fost sugerată pentru a îmbunătăți acuratețea modelului. Un alt studiu a comparat diverse modele de învățare automată pentru prezicerea calității vinului roșu, concluzionând că Gradient Boosting a avut cea mai bună performanță. Deși rețelele neuronale artificiale (ANN) nu au performat la fel de bine, studiul sugerează că mărirea setului de date ar putea îmbunătăți performanța acestora. Această cercetare evidențiază importanța selecției variabilelor și utilizarea unor modele diverse pentru a obține o predicție mai precisă a calității vinului.

Nu există o definitie generală și unanim acceptată în literatura de specialitate cu privire la calitatea vinului fiind o construcție cu mai multe fatete. De obicei, pentru a defini calitatea unui vin se folosesc teste senzoriale (care sunt bazate pe cunoștințele și experiența unui expert), însă pot fi folosite proprietățile fizico – chimice obținute prin analize de laborator. Experti sau somelieri își folosesc întreaga experiență și cunoaștere pentru a evalua calitatea unui vin și prin urmare scorul acordat este subiectiv ceea ce face dificila definirea calitatii cu precizie. Un studiu arată că 10% dintre somelieri sau experți au reusit să acorde același scor la reevaluarea calitatii aceluiași vin.

Tehnicile de învățare automată, cum ar fi algoritmii de învățare automată, reprezintă o metodă importantă pentru extragerea de cunoștințe din date brute. Fiecare algoritm ML are avantajele și dezavantajele sale, iar selectarea modelului și a variabilelor adecvate sunt cruciale pentru obținerea unor rezultate precise și interpretabile.

În cadrul lucrării de fata prezentăm clasificarea probelor de vin roșu pe baza proprietăților fizico – chimice rezultate dintr-o analiză chimică. Acest studiu poate fi valoros atât pentru producători din industria viticolă, cât și pentru consumatori pentru a îmbunătăți procesul de producție sau pentru a face o selecție, după caz, însă poate fi utilizat și de firmele care exportă/importă vinuri pentru a sprijini modul în care acestea evaluatează calitatea vinurilor importate/exportate și pentru a crește viteza și calitatea deciziilor luate, dar și de expertii din domeniul viticol pentru a-și îmbunătăți viteza și calitatea deciziilor lor și pentru a evalua vinul.

2. Prezentarea bazei de date

Setul de date conține informații cu privire la calitatea vinului rosu și este disponibil online în scopuri de cercetare <https://www.kaggle.com/datasets/ruthgn/wine-quality-data-set-red-white-wine>. Setul de date conține 1599 de observații și 12 variabile. Intrările includ teste efectuate asupra unor probe/mostre de vin care au evaluat o serie de proprietăți fizico – chimice (ex. valorile pH-ului) și rezultatul este bazat pe date senzoriale (mediana a cel puțin 3 evaluări facute de experți în vin). Fiecare expert a dată o nota cu privire la calitatea vinului cuprinsă între 3 (foarte rau) și 8 (foarte bun). Informațiile cuprinse în setul de date sunt informații cu privire la vinul rouș portughez „Vinho Verde”. Caracteristicile măsurate pentru probele/mostrelle de vin roșu analizate sunt: aciditatea fixă și volatila, aciditatea citrică, zaharul rezidual, cloruri, oxidul său liber, oxidul de sulf total, densitatea, pH, sulfatii și alcoolul prezentate pe larg în Anexa 1. Scopul setului de date este de a permite previzionarea calitatii vinului, evaluare pe care o face un expert (somelier sau degustator autorizat) pe baza proprietăților fizico - chimici pe care le prezinta o probă/mostra de vin. Presupunem că, datorita problemelor de confidențialitate și logistică, doar proprietățile fizico – chimice (variabilele independente) și evaluarea senzorială (variabila dependentă) date de un expert sunt prezentate în setul de date analizat deoarece nu avem informații cu privire la tipul de struguri din care este facut vinul, marca de vin, pretul de vânzare etc.

3. Determinarea importantei factorilor de predictie

Dimensiunea setului de variabile utilizat pentru previzionarea calitatii vinului a fost redus, fiind retinute doar caracteristicile relevante prin utilizarea unor metode de selecție a variabilelor, precum RF, XGBoost, MRLM. Utilizarea metodelor ML pe seturi de date cu dimensiuni foarte mari sunt predispușe la o serie de provocări. Datorită numărului mare de variabile, modelele obținute cu ajutorul algoritmilor ML au tendința de a fi supraadaptate, prezintând o performanță scăzută pe date necunoscute. Pentru reducerea dimensiunii, selecțarea variabilelor semnificative este o metodă utilizată pe scară largă și permite alegerea unui subset de caracteristici relevante din totalul acestora conducând la obținerea unor modele mai performante și mai ușor de interpretat. Pentru a reduce dimensiunea setului de date RWDS și pentru a obține un set de date optim pe baza caracteristicilor disponibile propum trei algoritmi: modelul de regresie liniar multiplu (MRLM), modelul Random Forest (RF) și modelul XGBoost (XGB).

3.1. Algoritmului Regresia liniara multipla (MRLM)

Modelele bazate pe regresia liniara simpla sau multipla evaluateaza legatura dintre variabila obiectiv si doua sau mai multe variabile independente. Este utilizata pentru a previziona valoarea varabilei analizate pe baza valorilor variabilelor predictive, permitand determinare contributiei relative a fiecarei varabile la variabila dependenta. In acest caz NU am impartit setul de date initial in doua subseturi ci am utilizat intreg setul de date. Dupa cum se poate observa modelul MRLM este semnificativ statistic. De asemenea, factorii cei mai importanți care influentează calitatea vinului sunt continutul in alcool, aciditatea volatila, continutul in dioxizi de sulf si sulfati, respective continutul in sare (Fig. 1.). Restul variabilelor neavand un impact semnificativ asupra calitatii vinului rosu.

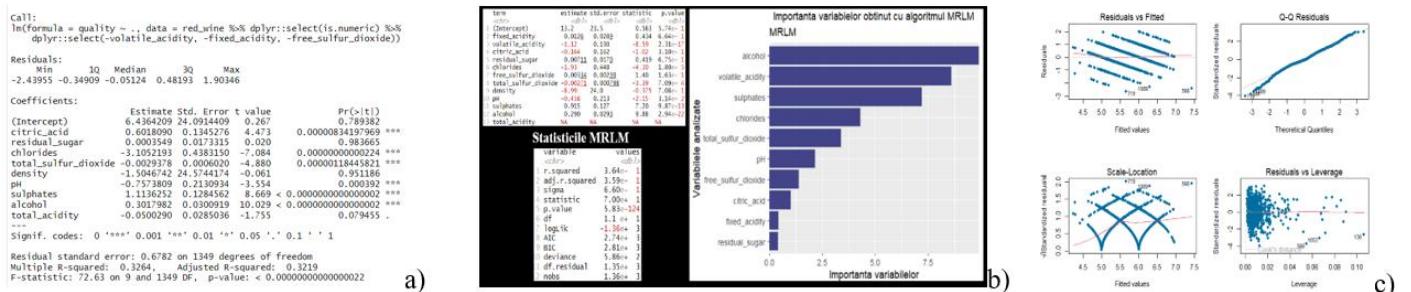


Fig. 2.1. Prezentam a) statisticile modelului de regresie inainte de eliminarea variabilelor, c) importanta variabilelor si b) graficele de diagnosticare a erorilor.

De asemenea, se poate observa ca daca construim modelul doar cu variabilele semnificative are loc o usoara imbunatatire a raportului de determinatie de la 32,64% la 36,19%, respectiv a raportului de determinatie ajustat (Fig. 1. a) si Fig. 2. a)). Observam ca ipotezele cu privire la variabila eroare nu sunt incalcate (Fig. 1. c) si Fig. 2. c)).

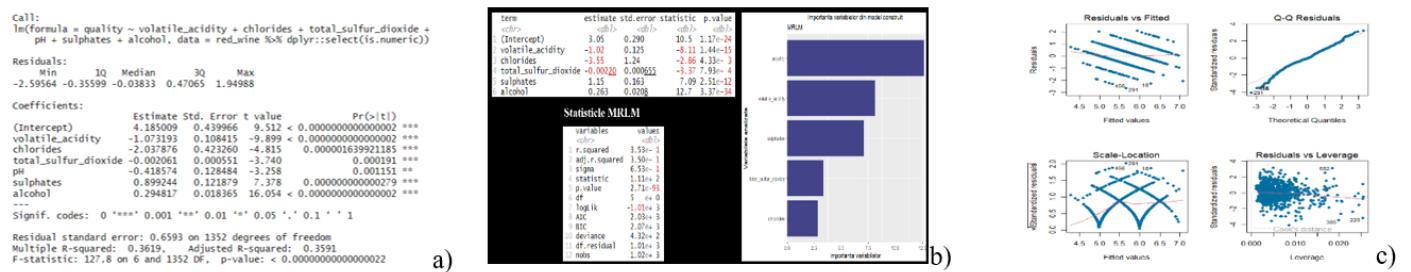


Fig. 2.2. Prezentam a) statisticile modelului de regresie dupa eliminarea variabilelor nesemnificative, b) importanta variabilelor si graficele de diagnosticare a erorilor.

3.2. Algoritmi MRLM, Random Forest si XGBoost

In ML lucrurile sunt un pic diferite, in sensul ca nu se foloseste intreg setul de date pentru construirea modelului, ci doar o parte din date, urmand ca setul de testare sa fie folosit pentru a evalua performantele modelului antrenat.

3.2.1. Random Forest si XGBoost

Algoritmi RF si XGBoost sunt doua metode recursive care permit clasificarea predictorilor in functie de importanta lor. Aceste metode, permit obtinerea predictorilor optimi pe baza caracteristicilor disponibile in setul initial urmand o serie de etape: impartirea setului de date (set de antrenare si set de testare) si seturile pentru crossvalidare (presupune impartirea setului de antrenare, creat in prima faza, la randul sau in seturi de antrenare si seturi de testare si are ca scop, pe de o parte, crearea mai multor versiuni de simulare a setului de antrenare, respectiv permite reglarea hiperparametri modelelor, iar pe de alta parte permite evitarea overfittingului); specificarea modelului de regresie ce ar putea explica cel mai bine datele analizate (regresia liniara multipla, RF, XGBoost); ingineria caracteristicilor etapa in care construim reteta aplicata setului de antrenare si al final asupra setului de testare si care include tratarea valorilor lipsa, tratarea varaielor cu varainta constanta, transformarea variabilelor in factor etc., crearea fluxului si executarea acestuia etapa in care asamblam reteta si modelul specificat si ii aplicam asupra setului de antrenare. Are loc construirea modelelor, modele obtinute sunt antrenate si in functie de performantele acestora este selectat cel mai bun model.

3.2.1.1. Impartirea setului de date

Nu exista un algoritm fix care sa poata previziona cu o acuratete ridicata. Pentru a putea genera si evalua modele optime, avem nevoie de un set de testare pentru a evalua acuratetea modelului generat, selectat si antrenat pe setul de antrenare. Este dificil sa gasim seturi de testare bune. K-Fold Cross Validation este o metoda statistica care ne permite sa obtinem modele mai performante. Procesul de generare a modelelor are loc pe mai multe seturi de date cu scopul de a obtine mai multi indicatori de performanta. Setul de antrenare este impartit intr-un numar de seturi/folds. Crearea unui numar mare de seturi de testare reduce riscul de a obtine un model optim din intamplare. Precizia finala pe care o obtinem va fi media celor k folds testate.

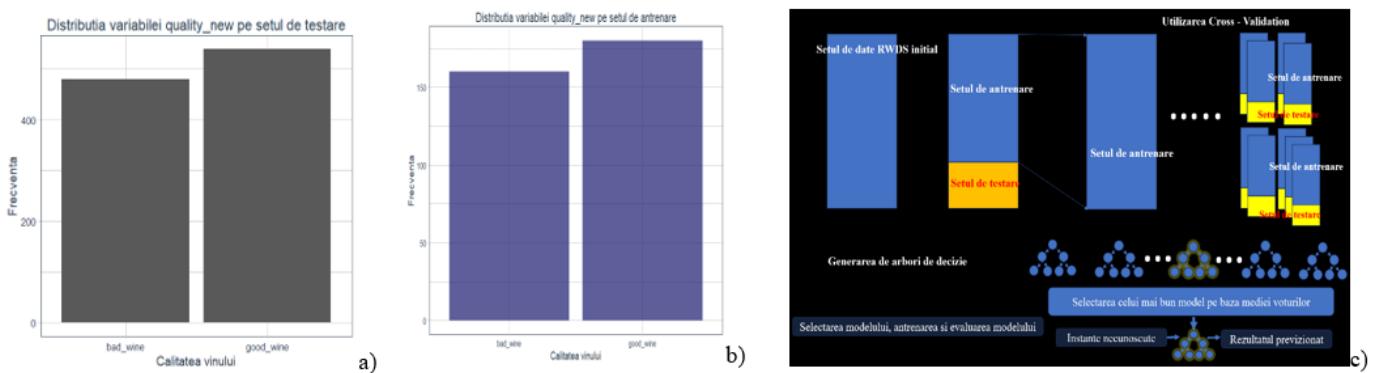


Fig. 3.1. Distributia variabilei quality_new pe cele doua seturi de date a) setul de antrenare si b) setul de testare.

O serie de experimente au fost realizate pentru a evalua performanta algoritmilor ML selectati. Evaluarea s-a realizat in doua moduri: metoda cross validation cu 10 folds si o repetitie si metoda setului de testare (Fig. 3.1.). Rezultatele obtinute pot fi granatate cu o probabilitate de 95% (respectiv ne asumam un risc de 5%). Pentru a evalua efectele algoritmilor utilizati, am folosit o serie de metri standard pentru evaluarea performantei, precum: curba ROC-AUC, sensibilitatea, specificitatea, precizia etc.

3.2.1.2. Identificarea si ajustarea hiperparametrilor, selectia, antrenarea si testarea modelului

Fiecare model de regresie generat are un set de hiperparametri care nu pot fi stabiliti de la inceput ci sunt acordati/reglati in functie de datele din setul de date. Astfel, in cazul modelului de regresie RF avem trei hiperparametri importanți dintre care: trees este stabilit de noi, restul hiperparametrilor min_n si mtry sunt ajustati in functie de datele din setul de date. Hiperparametrul trees este setat la 500 daca stabilesc o valoare mai mare performanta nu creste, in timp ce ntry (este numarul de predictori) si min_n (numarul minim de instante dintr-un nod necesar pentru a continua procesul de descompunere a arborelui) sunt acordati (tune()) in functie de datele din setul de date. Construirea arborilor de decizie este controlata prin acesti hiperparametri (Fig. 3.2.).

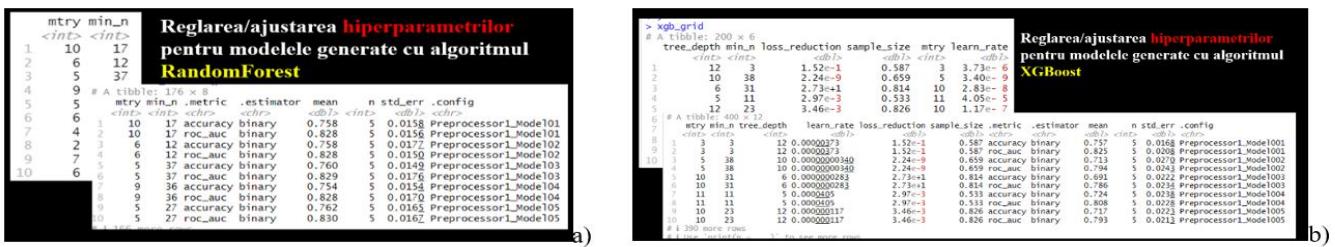


Fig. 3.2. Reglarea hiperparametrilor pentru modelele generate cu algoritmi a) RF si b) XGBoost.

Pachetele ranger si xgboost sunt tehnici asambliste pentru generarea de arbori de decizie si sunt considerati cei mai buni algoritmi de ML pentru clasificare si regresie. Generam mai apoi o lista de valori posibile a acestor hiperparametri care sunt stabiliți in functie de datele din setul de date (se ia in considerare numarul de predictori, valorile predictorilor etc.). Din cele 100/200 de combinatii de valori a acestor hiperparametri vom alege cea mai buna combinatie (Fig. 3.2.).

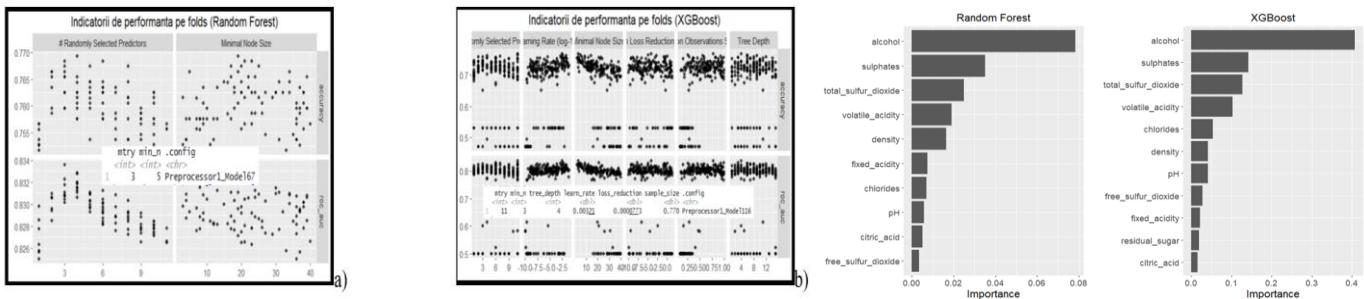


Fig. 3.3. Selectarea hiperparametrilor optimi pentru modelele generate cu algoritmi a) RF si b) XGBoost. Importanta variabilelor.

Astfel, pentru combinatia de valori ai hiperparametrilor corespunzatoare modelului de regresie implementat cu algoritmul ML RF avem doua metri: roc_auc si accuracy. Aceleasi metri le obtinem si pentru modelul de regresie implementat cu algoritmul ML XGBoost pentru fiecare combinatie de valori ai celor 6 hiperparametri. Pentru selectarea celui mai bun model se foloseste metrica roc_auc, accuracy care trebuie sa tinda catre 100%. In Fig. 3.3. a) prezentam evolutia celor doua metri in functie de valorile hiperparametrilor pentru modelul de regresi generate cu algoritmul ML RF. Observam ca valorile roc_auc pentru hiperparametrul ntry ating un maxim dupa care incep sa scada pe masura ce numarul de predictori creste, respectiv accuracy nu prezinta un pattern odata cu cresterea numarului minim de instante din nod. In Fig. 3.3. b) prezentam evolutia celor doua metri in functie de valorile hiperparametrilor pentru modelul de regresi implementat cu algoritmul ML XGB.

Pentru a genera un model optim vom folosi cea mai buna combinatie de valori pentru hiperparametri obtinuti. Construim modelul final doar alegand combinatia cea mai buna de valori pentru hiperparametri si generam graficele de importanta a variabilelor (Fig. 3.4.).

3.2.1.3. Importanta variabilelor

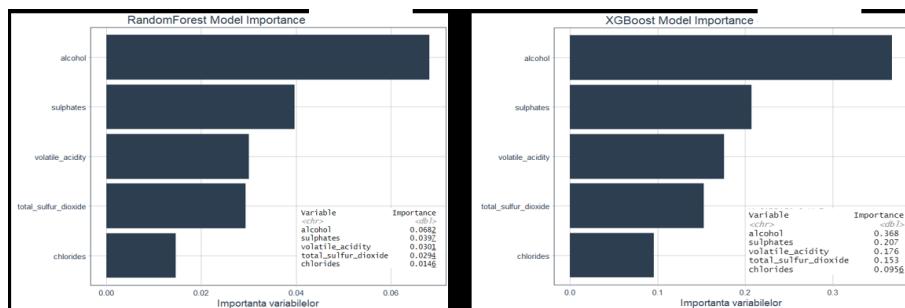


Fig. 3.4. Importanta variabilelor pentru modelele construite doar cu variabilele semnificative.

3.2.2. Modelul MRLM

Modelul MRLM

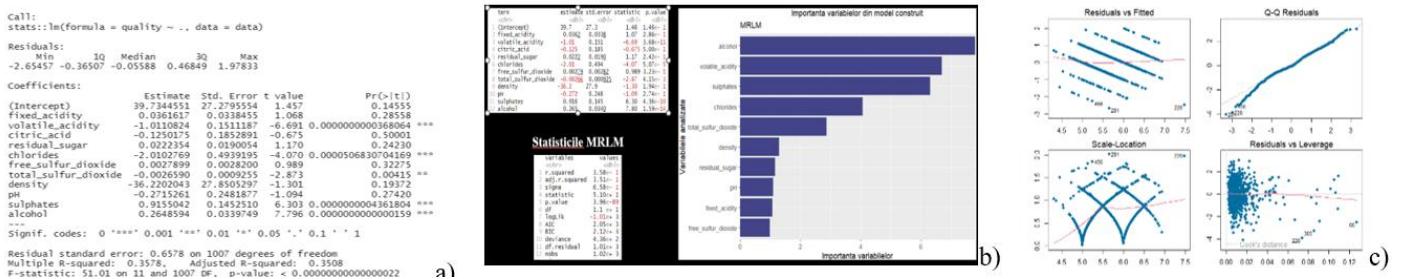


Fig. 3.5. Prezentam a) statisticile modelului de regresie, b) importanta variabilelor si graficele de diagnosticare a erorilor.

Din Anexa 2 observam ca avem probleme de colinearitate (intre variabilele fixed_acidity si density). Prin urmare, vom elimina variabila fixed_acidity deoarece este o variabila redundanta si generam din nou modelul fara a include variabila fixed_acidity. In urma eliminarii varabilei fixed_acidity problemele de colinearitate s-au rezolvat (Anexa 2). Verificarea ipotezelor cu privire la variabila reziduu si variabilele independente si observam ca acestea nu sunt incalcate. Observam ca in urma eliminarii colinearitatii performanta modelului este un pic mai scazuta, valoarea obtinuta pentru raportul de determinatie este de 0.357.

```
Call:
stats::lm(formula = quality ~ volatile_acidity + citric_acid +
  residual_sugar + chlorides + free_sulfur_dioxide +
  total_sulfur_dioxide + pH + sulphates + alcohol, data = data)

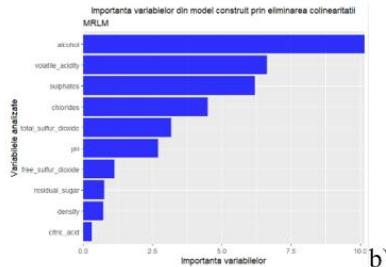
Residuals:
Min      1Q    Median      3Q     Max 
-2.60596 -0.36558 -0.04602  0.47188 1.98624 

Coefficients:
Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.726951 0.1672700 99.999 0.31816
volatile_acidity -1.001284 0.1599950 -6.446 0.000000000493 ***
chlorides 0.126995 0.0185430 6.770 0.74446
residual_sugar 0.0189543 0.0169530 0.770 0.44446
free_sulfur_dioxide -2.1462579 0.4777389 -4.497 0.0000000005087 ***
total_sulfur_dioxide 0.0011731 0.0027971 1.134 0.25691
pH 0.2324848 0.0011126 2.042 0.04589 ***
sulphates 0.890533 0.1432945 6.211 0.000000000670 0.0000000000022 ***
alcohol 0.2852783 0.0289922 10.153 < 0.0000000000000022 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6578 on 1008 degrees of freedom
Multiple R-squared: 0.3571
Adjusted R-squared: 0.3507
F-statistic: 55.99 on 10 and 1008 DF, p-value: < 0.0000000000000022
```

a)



b)

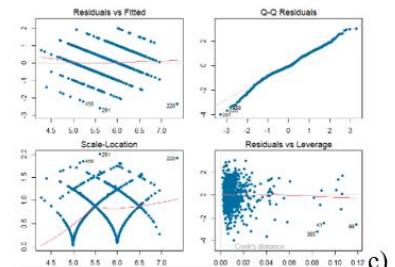


Fig. 3.6. Prezentam a) statisticile modelului de regresie dupa eliminarea variabilelor fixed_acidity, b) importanta variabilelor si graficele de diagnosticare a erorilor.

```
Call:
stats::lm(formula = quality ~ volatile_acidity + chlorides +
  total_sulfur_dioxide + pH + sulphates + alcohol, data = data)

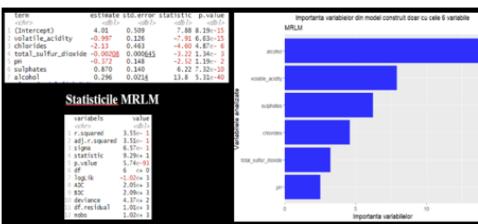
Residuals:
Min      1Q    Median      3Q     Max 
-2.58746 -0.35242 -0.05005  0.47132 1.96590 

Coefficients:
Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.0093780 0.5085621 7.884 0.000000000000819 ***
volatile_acidity -0.126335 0.1599950 -7.914 0.000000000493 ***
chlorides 0.126995 0.0185430 6.770 0.74446
total_sulfur_dioxide -0.020753 0.0006452 -3.217 0.00134 ***
pH 0.3720888 0.1432945 2.619 0.03194 **
sulphates 0.890533 0.0289922 13.836 < 0.0000000000000022 ***
alcohol 0.2963669 0.0214203 13.836 < 0.0000000000000022 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6575 on 1012 degrees of freedom
Multiple R-squared: 0.3553
Adjusted R-squared: 0.3514
F-statistic: 92.94 on 6 and 1012 DF, p-value: < 0.0000000000000022
```

a)



b)

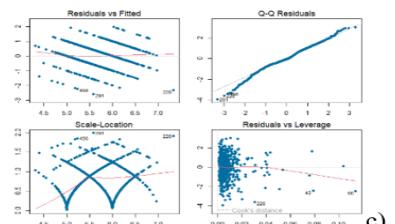


Fig. 3.7. Prezentam a) statisticile modelului de regresie construit cu variabilele semnificative, b) importanta variabilelor si graficele de diagnosticare a erorilor.

In final construim modelul de regresie care ia in considerare doar variabilele semnificative. Observam ca modelele obtinute sunt semnificative statistic, respectiv valorile obtinute pentru raportul de determinatie nu se imbunatatesc semnificativ, indiferent de model, acesta ramand aproape constant ~ 35%. De asemenea, se poate observa din Fig.... ca primele 5 variabile care influenteaza calitatea vinului sunt aceleasi indiferent de model si ordinea acestora relativ se pastreaza. Ipotezele cu privire la variabila reziduu si variabilele independente sunt respectate (Anexa 2).

3.3. Structura setului de date final

Setul de date final care va fi supus analizei este format din urmatoarele variabile (Fig. 3.8. a)): **Variabile categoriale:** quality si **Variabile numerice:** volatile_acidity, chlorides, total_sulfur_dioxide, pH, sulphates si alcohol. După cum se poate observa din Fig. 3.8. c) nici o variabilă din setul de date analizat nu prezintă valori lipsă.

3.3.1. Operațiile preliminare și de transformare a variabilelor

Construim variabila quality_new pe baza variabilei quality, care este o variabila categoriala discrete cu 7 categorii, pe baza variabilei quality, astfel: 1. *probele de vin rosu care au primit un scor mai mic de 5 le consideram vinuri de slaba calitate (bad_wine);* in timp ce 2. *probele de vin rosu care au primit un scor mai mare de 5 le consideram vinuri de buna calitate (good_wine);*

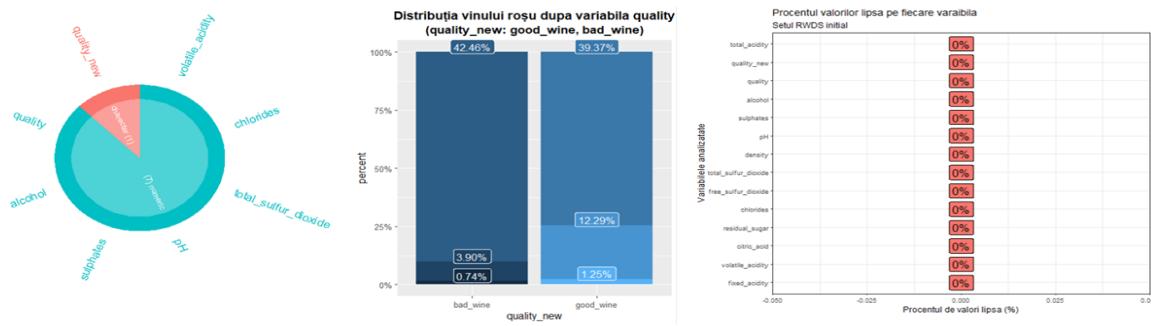


Fig. 3.8. Prezentam a) structura setului de date analizat si b) distributia vinului rosu in functie de varaiibila quality si c) valorile null.

În urma selecției variabilelor setul de date final are în componență: cele 6 variabile numerice si o variabila categorială: quality_new: "good_wine"/"bad_wine". Putem observa că în urma eliminării dupicatelelor setul de date contine un număr de 1359 de probe de vin rosu, este de tip data.frame si denumirea variabilor a ramas aceeași ca cea stabilita in setul de date initial (Fig. 3.8. a)).

3.3.2. Analiza descriptiva si grafica a variabilelor numerice

volatile_acidity	chlorides	total_sulfur_dioxide	pH	sulphates	alcohol	quality	quality_new
Min. :0.1200	Min. : 0.01200	Min. : 6.00	Min. :2.74	Min. :0.3300	Min. : 8.40	Min. :3.000	bad_wine :640
1st Qu.:0.3900	1st Qu.:0.07000	1st Qu.: 22.00	1st Qu.:3.21	1st Qu.:0.5500	1st Qu.: 9.50	1st Qu.:5.000	good_wine:719
Median :0.5200	Median :0.07900	Median : 38.00	Median :3.31	Median :0.6200	Median :10.20	Median :6.000	
Mean :0.5295	Mean :0.08812	Mean : 46.83	Mean :3.31	Mean :0.6587	Mean :10.43	Mean :5.623	
3rd Qu.:0.6400	3rd Qu.:0.09100	3rd Qu.: 63.00	3rd Qu.:3.40	3rd Qu.:0.7300	3rd Qu.:11.10	3rd Qu.:6.000	
Max. :1.5800	Max. :0.61100	Max. :289.00	Max. :4.01	Max. :2.0000	Max. :14.90	Max. :8.000	

variables	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
volatile_acidity	1	1.36e+03	0.529	0.183	0.52	0.518	0.185	0.12	1.58	1.46	0.728	1.23	0.00496
chlorides	2	1.36e+03	0.0881	0.0494	0.079	0.0802	0.0148	0.012	0.611	0.599	5.49	38.4	0.00134
total_sulfur_dioxide	3	1.36e+03	46.8	33.4	38	42.1	28.2	6	289	283	1.54	4.01	0.906
pH	4	1.36e+03	3.31	0.155	3.31	3.31	0.148	2.74	4.01	1.27	0.232	0.866	0.00421
sulphates	5	1.36e+03	0.659	0.171	0.62	0.637	0.119	0.33	2	1.67	2.4	11	0.00463
alcohol	6	1.36e+03	10.4	1.08	10.2	10.3	1.04	8.4	14.9	6.5	0.858	0.15	0.0294
quality	7	1.36e+03	5.62	0.824	6	5.58	1.48	3	8	5	0.192	0.33	0.0223

Fig. 3.9. Prezentarea principalelor indicatori ai statisticii descriptive.

In Fig. 3.9. sunt prezentati principali indicatori statistici pentru variabilele numerice, din care extragem urmatoarele informatii: in medie, o proba de vin rosu are o aciditate volatila de 0,53 g/dm³, un continut de sare de 0,0881 g/dm³, un nivel mediu de dioxid de sulf total de 46,83 mg/dm³, un pH acid de 3,31, un continut al sulfitilor de 0,6587 g/dm³ si un continut de alcool de 10,43%. Calitatea vinului rosu are un scor cuprins intre 3 si 8, minimul, respectiv maximul sunt egale cu aceste valori. Observam ca 75% dintre vinuri inregistreaza un continut de alcool de pana la 11,10%, in timp ce 25% dintre acestea inregistreaza o un continut de acol peste aceasta valoare. Jumătate din vinuri inregistreaza un pH de pana la 3,31%, in timp ce celalta jumătate peste 3.31% dolari. Continutul maxim de sare este de 0.611 g/dm³, respectiv o valoare minima de 0,012 g/dm³. Media continutui in sulfiti se abate de la nivelul mediu cu plus/minus 0.17 g/dm³, in timp ce 25% dintre vinurile rosii inregistreaza un continut al sulfitilor de pana la 0.55 g/dm³, in timp ce 75% dintre vinuri peste 0.55 g/dm³. Erorile de estimare (SE – standrad error indicator al statisticii inferentiale) a mediei la nivelul populatiei sunt mici (acuratetea mediei calcultae la nivelul esantionului comparativ cu media calculata la nivelul populatiei). Valorile obtinute pentru coeficienti de asimetrie arata ca distributia probelor de vin rosu dupa calitate, continutul in sare, continutul in alcool, pH, continutul in sulfiti, continutul total de dioxid de sulf si aciditatea volatila sunt usor asimetrice la dreapta sau pozitive apropiindu-se de o distributie normala (skew = 0). Observam, de asemenea, ca valorile obtinute pentru coeficienti de boltire pentru variabilele numerice analizate sunt pozitive, ceea ce inseamna ca distributia probelor de vin rosu dupa varaiibile analizate sunt distributii leptocurte (Fig. 3.10.).

3.3.3. Analiza grafica a variabilelor numerice

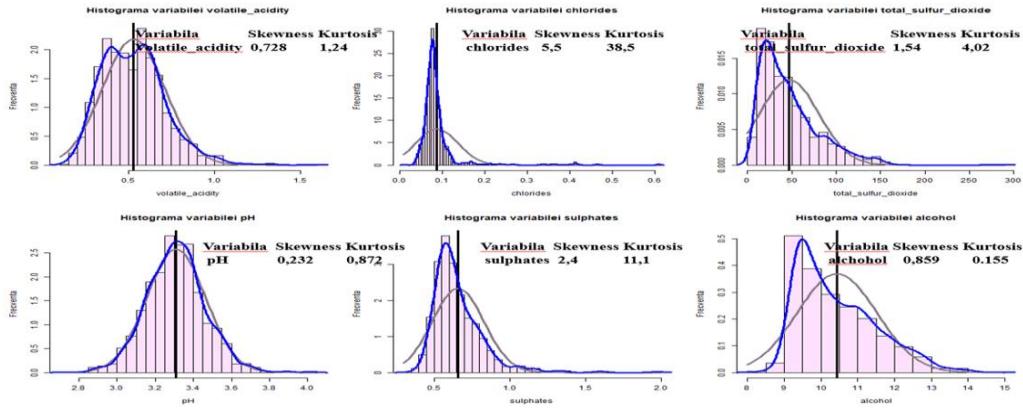


Fig. 3.10. Distributia probelor de vin rosu dupa varaiabilele analizate.

In Fig. 3.10. sunt prezentate distributiile probelor de vin rosu dupa variabilele analizate. Observam ca distributiile probelor de vin rosu dupa variabilele studiate prezinta o distributie asimetrica la dreapta (valoarea coeficientului skew este pozitiva) lucru sustinut si de valoarea obtinuta pentru coeficientul de asimetrie. Curvele frecventelor sunt alingite spre valori mari ale variabilelor, cuada curbelor se intinde spre dreapta. Din Fig. 3.10. a) se observam, de asemenea, ca distributia probelor de vin rosu dupa variabilele analizate prezinta o distributie leptocurrica lucru sustinut si de valorile obtinute pentru coefficienti de boltire.

3.3.4. Analiza descriptiva si grafica a variabilelor nenumericice

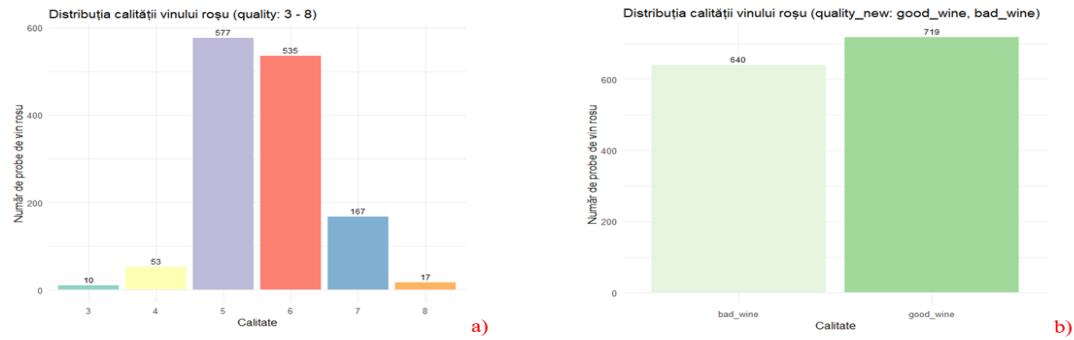


Fig. 3.11. Distributia probelor de vin rosu a) dupa scorurile acordate de experti si b) dupa variabila construita cu doua categorii (good_wine, bad_wine).

Observam ca cele mai multe probe de vin rosu sunt cele care au obtinut un scor de 5, adica un numar de 577 probe de vin rosu (Fig. 3.11. a)). Rezultatele prezentate in Fig. 3.11. a) arata ca cele mai multe probe de vin rosu fac parte din categoria good_wine, adica 719 probe de vin. Cu alte cuvinte 52,91% din totalul probelor de vin rosu fac parte din categoria vinurilor bune (good_wine), si doar 47,09% din totalul probelor de vin fac parte din categoria vinurilor mai putin bune (bad_wine) (Fig. 3.11. b)). Distributia probelor de vin rosu dupa scorul acordat de exeprti arata ca cele mai multe probe de vin sunt din categoria 5, 39,37% sunt din categoria 6, 12,29% fac parte din categoria 7, urmate de categoriile 4, 8, si 3 cu un procent de 3,9%, 1,25% si 0,74% (Fig. 3.11. a)).

3.3.5. Analiza descriptive si grafica a variabilelor numerice in functie de categoriile variabilei categoriala quality_new

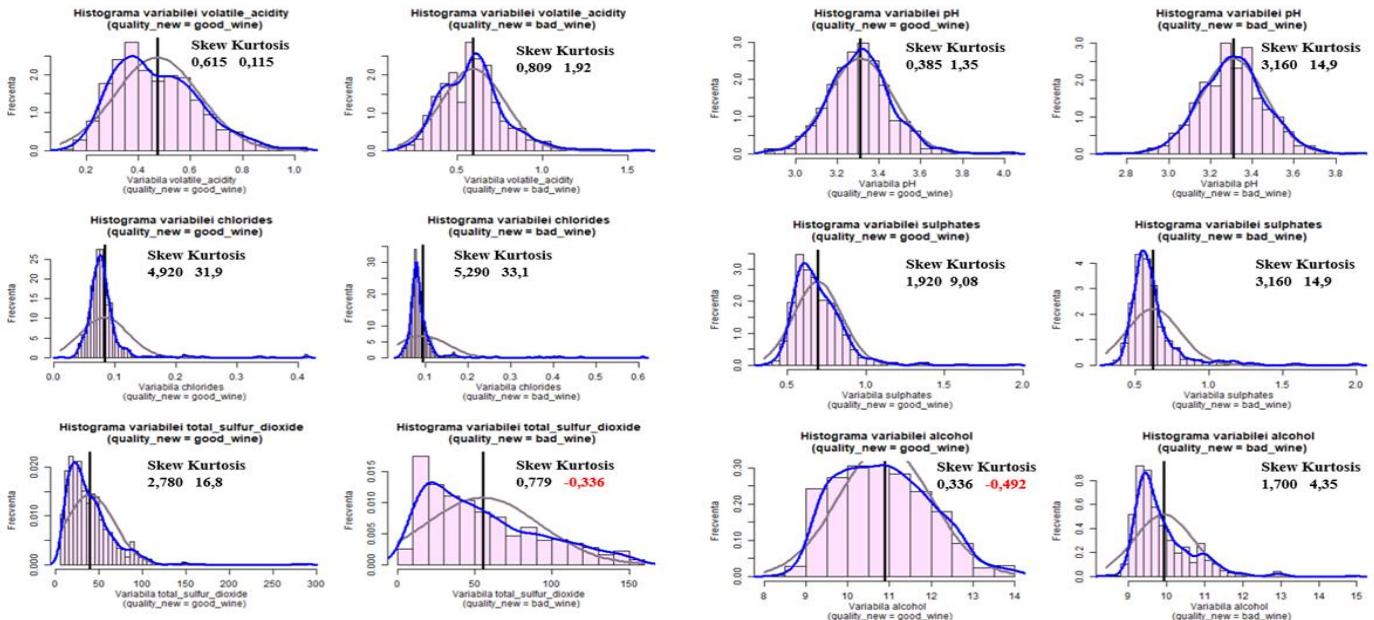


Fig. 3.12. Distributia probelor de vin rosu dupa variabilele numerice analizate in functie de categoriile variabilei tipul vinului.

In Fig. 3.12. sunt prezentate distributiile probelor de vin dupa variabilele analizate in functie de categorie: good_wine si bad_wine. Observam ca majoritatea distributiilor sunt asimetrice la dreapta, lucru sustinut si de rezultatele obtinute pentru valorile coeficientilor de asimetrie si boltire. Dupa cum se poate observa cele mai multe probe de vin au un continut in alcool de 9,5% si un continut in sare de 0,08 g/dm³. De asemenea, cele mai multe vinuri din categoria bad_wine au primit un scor de la experti de 5, in timp ce cele mai multe probe de vin din categoria good_wine au primit un scor de 6. In Fig. 3.12. sunt prezentate rezultatele obtinute pentru coeficienti de asimetrie si boltire pentru cele doua categorii de vinuri. Observam ca majoritatea distributiilor prezinta o asimetrie pozitiva si sunt leptocurtice, cu exceptia distributiei dupa variabila quality, total_sulfur_dioxide si variabila alcohol.

3.3.6. Identificarea valorilor extreme (outlierilor) si tratarea acestora

3.3.6.1. Identificarea valorilor extreme (outlierilor)

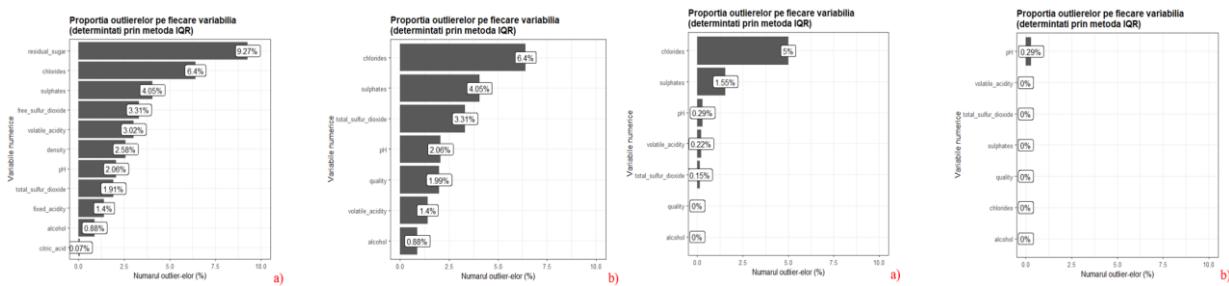
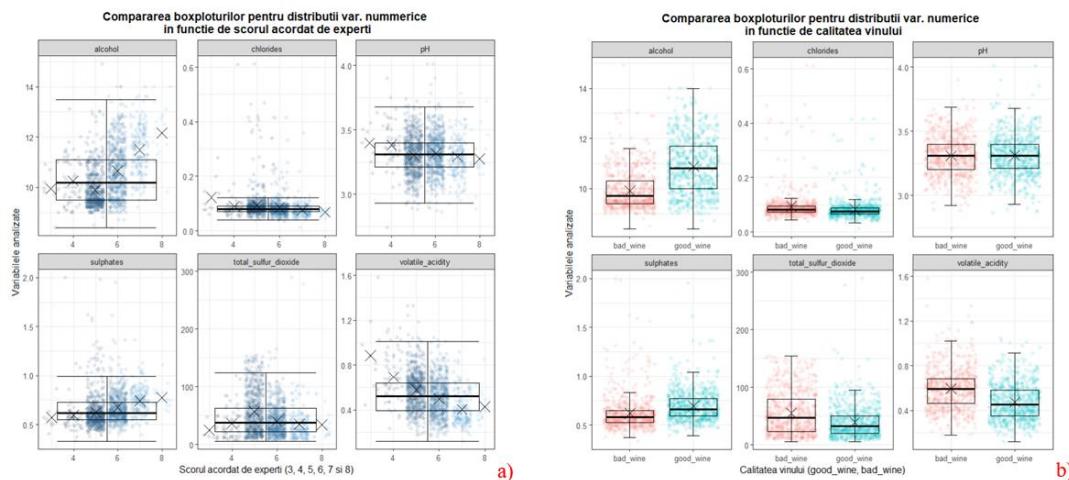


Fig. 3.13. Proprietatea outlierelor pe fiecare variabila analizata a) pe intreg setul de date si b) pe setul de date final ce nu inculde duplicatele, respectiv proprietatea outlierelor pe fiecare variabila analizata dupa tratarea autlierilor a) pe intreg setul de date si b) pe setul de date final.

Valorile extreme pot fi identificate fie pe cale vizuala/grafica (reprezentand grafic distributiile sau histogramele variabilelor sau prin trasarea diagramele de tip boxplot) sau pot fi determinate prin metoda IQR. Diagramele de tip boxplot au ca principal avantaj faptul ca pun in evidenta foarte rapid prezenta valorilor extreme. Aceste valori pot influenta probabilitatea de aparitie a erorilor, pot influenta valoarea abaterii standard, respectiv influenta lor este cu atat mai importanta cu cat dimensiunea esantionului este mai mica. In Fig. 3.13. sunt prezentate proprietatea outlierelor pe fiecare variabila analizata in cadrul lucrarii. Observam ca majoritatea variabilelor analizate prezinta outlieri. De asemenea, in Fig. 3.13. sunt prezentate graficele in care prezentam proprietatea outlierelor pe fiecare variabila analizata dupa tratarea autlierilor.

In Anexa 3 sunt prezentate boxploturile corespunzatoare variabilelor analizate in cadrul lucrarii. Dupa cum se poate observa se face o analiza comparativa intre boxploturile corespunzatoare variabilei analizate si boxplotul obtinut pentru o distributie normala de medie si varianta egala cu cea a variabilei analizate.



In Fig. 3.14. a) sunt prezentate diagramele boxplot in functie de variabila categoriala quality. Observam odata cu restarea scorului acordat de experti creste si continutul in alcool, continutul in sulfiti. In ceea ce priveste continutul in sare, acesta scade odata cu scorul acordat de experti. De asemenea, observam o usoara scadere a pH, respectiv o scadere accentuata a aciditatii volatile odata cu cresterea scorului acordat de experti. In ceea ce priveste continutul total de dioxid de sulf acesta creste usor pentru probele de vin din categoria bad_wine, dupa care incepe sa scada usor pentru probele de vin din categoria good_wine odata cu cresterea scorului (variabila total_sulfur_dioxide inregistreaza un maxim la un scor de 5 dupa care incepe usor sa scada). In Fig. 3.14. b) sunt prezentate diagramele boxplot in functie de variabila categoriala quality_new (calitatea vinului: good_wine si bad_wine). Observam ca vinurile din categoria good_wine prezinta un continut in acol si sulfati mai mare, respectiv inregistreaza valori mai mici pentru pH, continutul in SiO₂, continutul in sare si aciditatea volatila spre deosebire de vinurile din categoria bed_wine.

3.3.6.2. Tratarea outlierilor

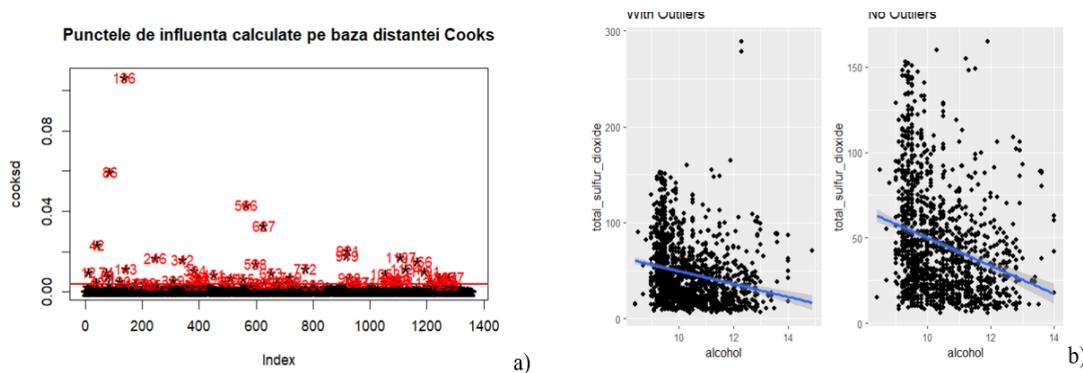


Fig. 3.15. a) Punctele de extreime/influenta si b) Compararea liniei de regresie dintre variabilele analizate a) inainte si dupa eliminarea punctelor de influenta si a outlierilor si b) punctele de influenta.

In Fig. 3.15. a) sunt reprezentate punctele de influenta prezente in setul de date analizat. Se poate observa ca o serie de observatii au valori ce depasesc distanta Cook's. Vom considera aceste observatii ca fiind puncte de influenta cu un impact negativ asupra modelului de regresie, respectiv asupra analizei si modelelor pe care vrem sa le construim. Prin urmare, avem un numar de 123 de puncte de influenta (sau outlieri) la care nu vom reninta si pe care le vom transforma (distribui in cadrul distributiei) astfel incat distributia variabilei sa isi pastreze forma initiala, respectiv pentru a nu influenta rezultatul analizei. Astfel, setul de date final va avea acelasi numar de observatii (1359) si 8 variabile.

In Fig. 3.15. b) prezentam o comparatie a modulului in care linia de regresie este influentata de aceste puncte de influenta si/sau outlieri. Astfel, comparam linia de regresie obtinuta in prezenta punctelor de influenta cu linia de regresie obtinuta in absenta acestora. In cazul nostru modificarile nu sunt foarte semnificative, insa sunt importante. Putem observa ca in unele cazuri linia de regresie obtinuta dupa eliminarea punctelor de influenta fitaaza mai bine datele analizate.

Asa cum s-a mentionat, ipoteza de heteroscedasticitate nu este respectata si, de asemenea, observam exista unor outlieri si/sau puncte de influenta. Daca ne uitam la graficul Residuals vs. Leverage observam ca o parte din outlieri au o valoare semnificativa mai mare decat de 2.5 ori valoarea medie (Fig. 3.5. – 3.7. b)). Dupa eliminarea punctelor de influenta, se poate observa din Fig. 3.16. a), ca valoarea raportului de determinatie si a raportului de determinatie ajustat s-au imbunatatit semnificativ. Prin urmare, modelul de regresie s-a imbunatatit substantial (Fig. 3.16. a)). Comparativ cu graficele de diagnostic obtinute la primul model de regresie, cele obtinute pentru modelul de regresie in care am transformat punctele de influenta s-au imbunatatit. De asemenea, graficul Residuals vs. Leverage nu mai prezinta puncte de influenta care sa aiba o valoare semnificativa conducand la obtinerea unui model de regresie mai bun. In Fig. 3.16. prezentam rezultatul obtinut pentru modelul de regresie construit folosind ca variabila dependenta caracteristica quality si ca variabile independente toti predictori.

```
Call:
lm(formula = quality ~ ., data = red_wine_without_outliers %>%
  select_if(is.numeric))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.61073 -0.33477 -0.03994  0.41174  1.92095 

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)    
(Intercept) 3.9968801  0.4248284  9.408 < 0.000000000000002 ***
volatile_acidity -1.0205473  0.1044883 -9.767 < 0.000000000000002 ***
chlorides    -2.1283799  0.4316273 -4.931 0.000000929848914992 ***
total_sulfur_dioxide -0.0022934  0.0005156 -4.448 0.000009443736391913 ***
pH          -0.4127978  0.1240601 -3.327   0.000903 ***
sulphates    0.9753895  0.1171739  8.324 0.00000000000000224 ***
alcohol      0.3068700  0.0174861 17.549 < 0.000000000000000224 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5868 on 1229 degrees of freedom
Multiple R-squared:  0.4235, Adjusted R-squared:  0.4207 
F-statistic: 150.5 on 6 and 1229 DF, p-value: < 0.0000000000000022
```

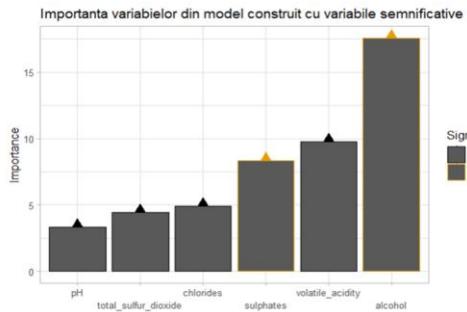


Fig. 3.16. Selectarea variabilelor cu un impact semnificativ asupra variabilei quality.

Valoarea obtinuta pentru raportul de determinatie este de 36,32% (respectiv 42,35% pentru modelul pentru care am tratat valorile extreme), in timp ce valoarea obtinuta pentru raportul de determinatie ajustat este de 35,99%. Prin urmare, 35,99% din variația calitatii vinului rosu este influentata de variația simultana a variabilelor independente incluse in model.

In Fig. 3.17. sunt prezentate statisticile obtinute pentru modelului de regresie care include toate variabilele independente si modelul de regresie care include doar variabilele semnificative identificate in primul model. Observam o imbunatatire a raportului de determinatie pentru modelul care include doar variabilele semnificative si pentru care s-a tratat punctele de extrem. Valoarea R² ajustat este mai mica de 50% ceea ce indica faptul ca exista unul sau mai multi factori pe care nu i-am inclus in model, iar cei inclusi explica doar o parte din variația varibilei dependente (sau predictor nu sunt buni pentru a realiza o predictie buna a calitatii vinului). In plus, valorile p-values sunt mai mici ca 0.05 ceea ce indica faptul ca R² ajustat este semnificativ diferit de zero.

3.3.6. Matricea de corelatie (MC)

Analiza MC ne ofera o serie de informatii cu privire la tipul, sensul si intensitatea legaturi dintre variabilele analizate. Observam ca majoritatea variabilelor sunt slab corelate intre ele. Variabilele volatile_acidity, sulphates, alcohol, citric_acid si total_sulfur_dioxide prezinta o corelatie semnificativa cu variaibila quality (Fig. 3.17.). In studiul nostru, am folosit si indicatorul VIF (Variance Inflation Factor) ca tehnica de selectie a variabilelor si eliminarea caracteristicilor ce prezinta o valoarea VIF ridicata (VIF > 10). Multicolinearitatea poate introduce redundanta sau instabilitate in modelele predictive. Aceasta ne permite sa observam interdependentele dintre variabile si sa identificam legaturile puternice dintre variabilele predictor si variaibila quality. Corelatia pozitiva indica faptul ca ambele variabile au acelasi sens de variație, sugerand un impact pozitiv asupra calitatii vinului. In timp ce, o corelatie negativa indica faptul ca sensul de variație a celor doua variabile este diferit, sugerand un impact negativ asupra calitatii vinului.

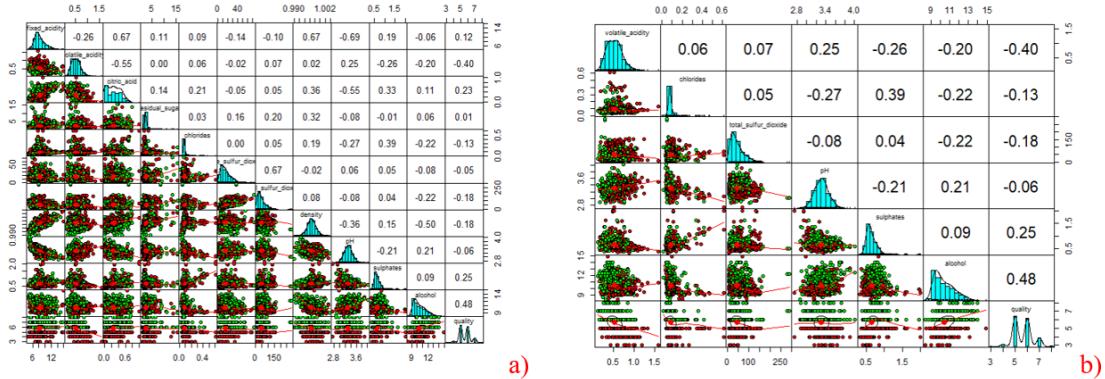


Fig. 3.17. Matricea de corelatie obtinuta pentru a) toate variabilele, respectiv b) cele semnificative.

Comparativ s-a realizat MC pentru setul de date initial si pentru setul de date ce include doar variabilele semnificative selectate. MC a evidențiat faptul ca aciditatea citrica, sulfati si alcoolul sunt corelate pozitiv cu calitatea vinului, ceea ce sugereaza faptul ca un nivel mai ridicat al acestor caracteristici sunt asociate cu un vin de calitate superioara. Observam ca aciditatea volatila, pH-ul si SiO₂ sunt corelate negativ cu calitatea vinului, ceea ce sugereaza faptul ca un continut ridicat al aciditatii volatile, un pH scazut si un nivel ridicat in SiO₂ sunt asociate cu un vin de slaba calitate. In Fig. 3.17. sunt prezentate MC pentru cele doua seturi de date, care pune in evidenta legatura dintre variabile, respectiv arata variabilele care au un impact pozitiv sau negativ asupra calitatii vinului. Spre deosebire de vinurile de masa/slabe, observam ca vinurile de calitate prezinta o valoare medie mai mare pentru continutul in alcool, sulfati si acid citric, respectiv inregistreaza valori scazute pentru concentratia de dioxidul de sulf total.

3.3.7. Discretizarea variabielor

Tabel 3.1. Structura setului RWDS corespunzator variabilelor utilizate in algoritmi ML.

Numele variabilei si simbolul	Valori	Code	Structura/ valori unice		Structura dupa discretizare		Tipul variabilei		
			f _i	%	f _i	%	MS	IDV/DV	
quality (Y)	3	bad_wine	10	0.7%	640	47.1%	OV ↓ NV	DV	
	4		53	3.9%					
	5		577	42.5%					
	6	good_wine	535	39.4%	719	52.9%	OV ↓ NV		
	7		167	12.3%					
	8		17	1.3%					
volatile_acidity (X ₁)	Numeric continua	mai_mare_0.55 mai_mic_0.55	141	10.4%	595 764	43.8% 56.2%	RV ↓ NV	IDV	
chlorides (X ₂)	Numeric continua	mai_mare_0.069 0.08-0.069 mai_mic_0.08	95	7.0%	618 409 332	45.5% 30.1% 24.4%	RV ↓ OV	IDV	
		mai_mare_82 54-82 mai_mic_54			206 211 942	15.2% 15.5% 69.3%	RV ↓ OV		
total_sulfur_dioxide (X ₃)	Numeric continua	mai_mare_0.65 mai_mic_0.65	80	5.9%	537 822	39.5% 60.5%	RV ↓ NV	IDV	
		mai_mare_10 mai_mic_10			722 637	53.1% 46.9%	RV ↓ NV		
alcohol (X ₅)	Numeric continua								

Nota: acronimele si simbolurile utilizate in tabel au urmatoarea semnificatie: MS = scala de masura; NV = Variabila nominala; OV – Variabila ordinala; RV – variabila raport; RV→OV = transformarea variabilei numerice (raport) in variabila categoriala (ordinala). f_i – frecventa; IDV – variabila independenta; DV – variabila dependenta.

Pentru a putea construi modele bazate pe arbori cu ajutorul algoritmilor ML s-a realizat o discretizare a variabilelor. Variabila quality este definita ca variabila dependenta si este o variabila nominala cu doua clase/categorii astfel incat modelele create sa poata fi bazate pe criteriul de divizare χ^2 .

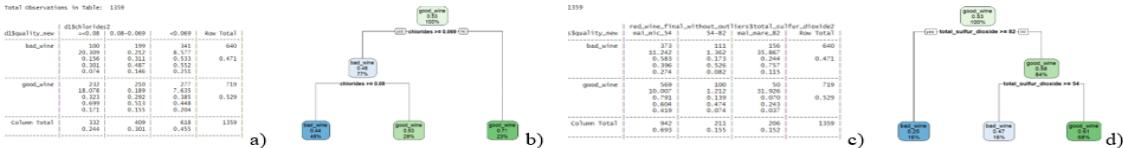


Fig. Tabelul de contingenta, respectiv a AD obtinute pentru variabila a) – b) chlorides si c) – d) total_sulfur_dioxide.



Fig. Tabelul de contingenta, respectiv a AD obtinute pentru variabila a) – b) alchool si c) – d) volatile_acidity.

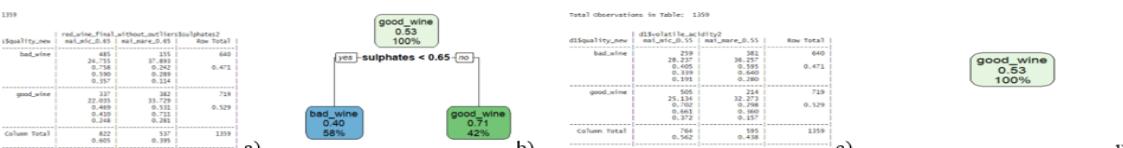


Fig. Tabelul de contingenta, respectiv a AD obtinute pentru variabila a) – b) sulphates si c) – d) pH.

Fig.3.18. Discretizarea variabilelor cu ajutorul arborilor de decizie.

Variabilele prezentate in Tabelul. 3.1. si marcate cu simbolul X_i sunt variabile independente. Acestea sunt definite ca variabile nominale sau ordinale. Majoritatea variabilelor au fost initial variabile numerice/variabile raport, insa pentru a le putea incorpora in algoritmul CHAID, a fost necesara o transformare a acestora in variabile categoriale. Discretizaarea variabilelor s-a facut cu ajutorul arborilor (Fig. 3.18.) de decizie care a sugerat si intervalul de valori dupa care sa se faca discretizarea.

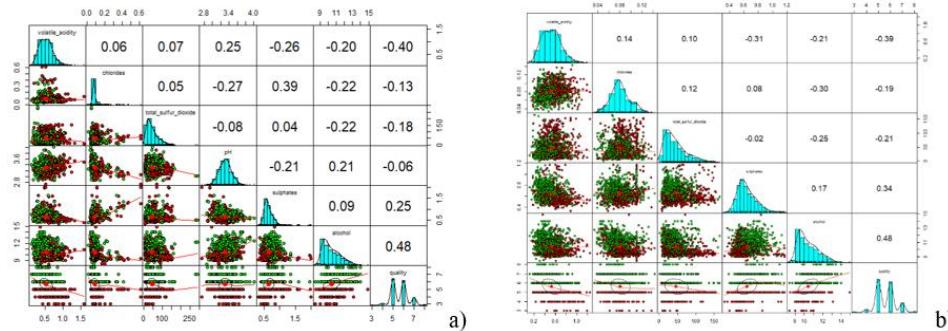


Fig. 3.19. Matricea de corelatie obtinuta pentru setul de date care a) contine puncte extreme, respectiv b) nu contine puncte extreme.

In Fig. 3.19. prezintam matricea de corelatie realizata luand in considerare variabilele numerice raw, respectiv a variabilelor semnificative dupa tratarea punctelor extreme. Se observa o usoara imbunatatire a coeficientilor de corelatie. Astfel, variabila quality este corelata negativ cu dioxidul de sulf, sareala si aciditatea volatila si corelata pozitiv cu variabilele alcool si sulfati. De asemenea, observam ca forma distributie nu se modifica.

4. Obtinerea arborilor de clasificare prin algoritmi ML

Am utilizat clasificatori ML bazati pe arbori de clasificare pentru previonarea clasei pe baza proprietatilor probei de vin rosu, precum: CART, CHAID, C5.0, QUEST, Random Forest, XGBoost.

4.1. Modele de clasificare

Metodele asambliste precum RF si XGB pot imbunatati acuratetea predictiilor facute de algoritmul CART prin combinarea mai multor arbori de decizie:

XGBoost: construiese modele CART in mod secential, in care fiecare model nou generat este focalizat pe corectarea erorilor din colectia de modele existente, permitand optimizarea si imbunatatirea graduala a acuratetii modelelor. XGBoost utilizeaza o serie de tehnici pentru a imbunata performanta, inclusiv pruning-ul arborilor de decizie, regularizarea si optimizarea gradientului stocastic.

Random Forest: este o colectie de arbori de decizie CART construiti in paralel in care fiecare arbore a fost antrenat cu ajutorul setului de antrenare si are un anumit set de caracteristici, respectiv un scor de probabilitate asociat. La final se calculeaza probabilitatea medie luand in considerare toti arborii generati. Acest lucru permite evitarea overfittingului si imbunatatestea stabilitatea arborilor. Construieste mai multi arbori de decizie in paralel si combina rezultatele pentru a obtine o predictie finala. Principalul atu consta in constructia arborilor in mod aleator, prin selectarea aleatorie a esantioanelor si a caracteristicilor pentru a evita overfittingul si pentru a imbunata generalizarea. Este robust in fata datelor lipsa si a valorilor aberante, si este potrivit pentru o varietate de probleme de invatare supervizata, inclusiv clasificare si regresie.

C5.0: este o versiune imbunatatita a algoritmului C4.5 (care la randul sau este o versiune imbunatatita a algoritmului ID3) mostenind toate caracteristicile si avantajele stramosilor sai. Acest algoritm foloseste raportul Gain care este o normalizarea a indicatorului information Gain folosind valorile splitinfo. Raportul Gain este folosit ca masura a impuritatilor si divizarea se face prin mai multe cai de divizare. Ca regula de oprire, numarul minim de instante pe nodul copil este setat la 2. Este utilizat pentru clasificare si regresie si utilizeaza o varietate de tehnici pentru a imbunata performanta si interpretarea modelului. C5.0 utilizeaza un set de reguli pentru a face predictii si poate gestiona datele lipsa si variabilele continue fara prelucrare speciala. Unul dintre avantajele sale este capacitatea de a gestiona seturi de date mari si complexe.

CART: construieste un arbore unde fiecare nod reprezinta o intrebare, iar ramurile nodurilor reprezinta raspunsurile posibile la intrebare. Scopul CART este de a diviza setul de date in subgrupuri omogene din punct de vedere al variabilei de raspuns. Diviziunea optimă a datelor este determinată prin minimizarea unei funcții de cost, cum ar fi impuritatea Gini sau entropia.

QUEST: este un algoritm care utilizeaza inferenta statistică pentru a lua decizii de divizare pentru construirea arborilor. Acest algoritm a fost conceput pentru a oferi o alternativă rapidă și eficientă la alte metode de generare a arborilor. Principalul sau atu constă în utilizarea testelor statistice pentru a determina cele mai bune divizări ale datelor, reducând astfel timpul necesar pentru construirea arborilor.

CHAID: este un alt algoritm folosit în special pentru analiza datelor de tip întrebare-raspuns. Spre deosebire de CART și QUEST care sunt bazati pe metode non-parametrice, CHAID este un algoritm bazat pe teste parametrice (utilizează teste statistice pentru a identifica interacțiuni semnificative între variabilele explicative și variabila de răspuns). Algoritmul construieste un arbore prin selectarea diviziunii care maximizează statisticile testului χ^2 pentru a evidenția asociațiile semnificative între variabile.

Tabelul 4.1. Caracteristicile algoritmilor ML utilizati.

Algoritm	QUEST	CART	CHAID	C5.0	XGBoost	RF
Tipul variabilei	continuu sau categoriale	continuu sau categoriale	categoriale	continuu sau categoriale	continuu sau categoriale	continuu sau categoriale
Numar de ramuri	mai mult de 2	2	mai mult de 2	2	mai mult de 2	mai mult de 2
Variabile de ramificatie	1 sau mai multe O singura variabila la fiecare pas	1 sau mai multe	2	1 sau mai multe O singura variabila la fiecare pas	1 sau mai multe	1 sau mai multe
Regula de divizare	Testul F/Chi-square	Indexul Gini	Testul Chi-square	Testul F/Chi-square	Greutatea totala a gradientului	Testul F/Chi-square
Pruning	Cross – validation/ testarea pe setul de testare	Cross validation/ – testarea pe setul de testare	– Cross validation/ – testarea pe setul de testare	– Cross validation/ – testarea pe setul de testare	Utilizeaza reguli de oprire	– Cross validation/ – testarea pe setul de testare

Regula de divizare greutatea gradientului presupune ca, la fiecare pas al constructiei arborelui, sa se calculeaza un indicator care ia in considerare cat de mult influenteaza fiecare variabila predictia modelului, iar caracteristica cu cea mai mare greutate totala a gradientului este aleasa pentru a fi divizata la fiecare pas. Algoritmul XGBoost nu foloseste tehnici de taiere si utilizeaza reguli de oprire. In timpul construirii arborelui, nu se aplica tehnici de taiere, in schimb, algoritmul foloseste tehnici de oprire pentru a determina cand sa opreasca construirea arborelui. Regulile de oprire pot include numarul minim de instante din nodul parinte sau copil. Acest lucru ajuta la evitarea overfittingului si la obtinerea unui model mai generalizat. Aplicarea tehnicii de corss – validare are ca obiectiv evaluarea performantei

modelului (intr-un mod robust si obiectiv), evitarea overfittingului si optimizarea hiperparametrilor (prin indentificarea combinatiei optime care maximizeaza performanta modelui).

4.2. Rezultate si discutii

4.2.1. Modele asambiliste

4.2.1.1. Rezultatele obtinute pentru modelul generate cu algoritmi RF si XGB

In Anexa 4 prezentam evolutia celor doua metrii (roc_auc si accuracy) in functie de valorile hiperparametrilor pentru modelele cu toate variabilele, respectiv cu 3 variabile generate cu RF si XGB. Din multitudinea de valori generate se selecteaza combinatia optima (cei mai buni hiperparametri) de hiperparametri si se adauga la fluxul de lucru pentru a genera si antrena modelele utilizand setul de antrenare. In final este selectat cel mai bun model. In Anexa 4 sunt prezentate valorile obtinute pentru hiperparametri modelelor optime selectate.

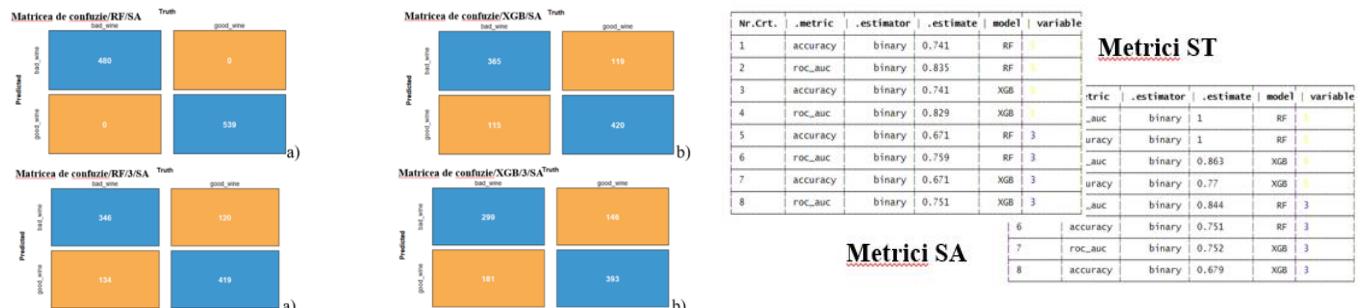


Fig. 4.1. Matrice de confuzie obtinute pe setul de antrenare metriile asociate.

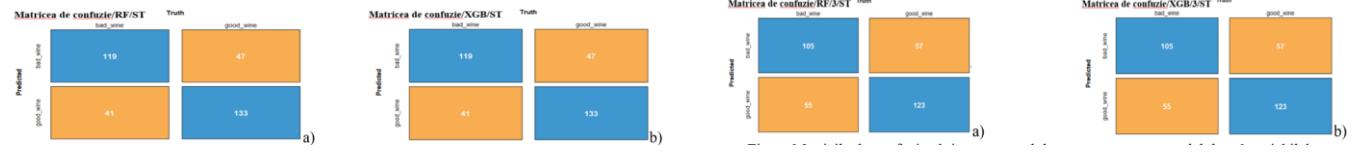


Fig. 4.2. Matricile de confuzie obtinute pe setul de testare pentru modelul cu 3 variabile.

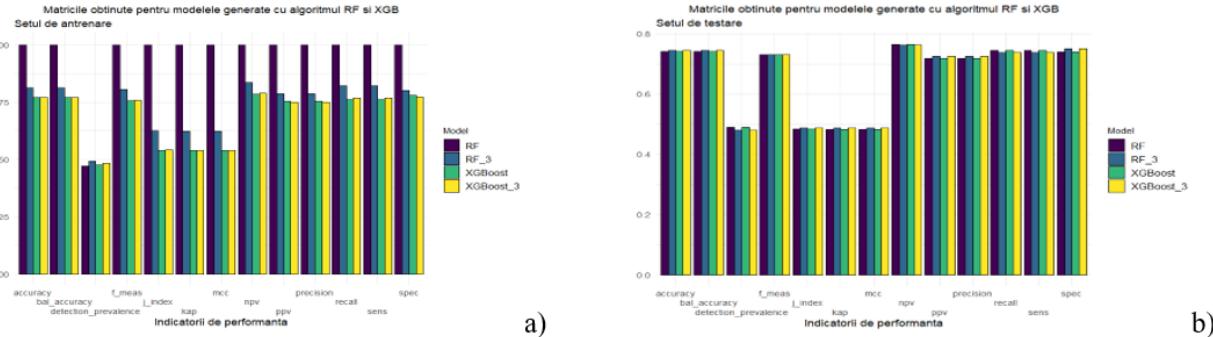


Fig. 4.3. Indicatorii de performanta obtinuti pentru modelele generate (cu toate variabilele si cu 3 variabile) cu algoritmi RF si XGB pe setul a) de antrenare si b) testare.

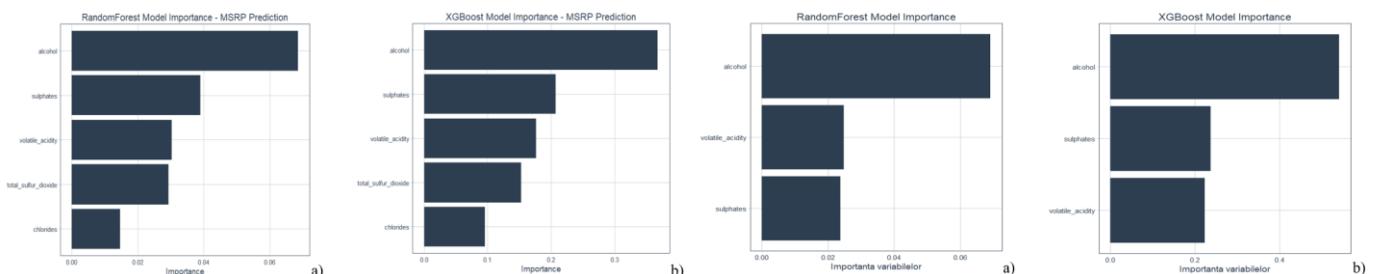


Fig. 4.4. Importanta variabilelor in cadrul modelelor cu toate/3 variabile generate cu algoritmul a) RF si b) XGBoost.

Comparand performanta modelelor pe setul de antrenare si pe cel de testare observam ca atat modelele cu 3 variabile, cat si cele cu toate variabilele prezinta o performanta mai buna pe setul de antrenare spre deosebire de performanta modelului obtinuta pe setul de testare, ceea ce este de asteptat deoarece modelele sunt antrenate pe setul de antrenare si cunoaste instantele din setul de date. Observam de asemenea ca performantele modelelor obtinute pe setul de antrenare si cel de testare sunt similara sau variaza foarte putin ceea ce indica faptul ca modelele nu prezinta overfitting sau underfitting. In ceea ce urmeaza vom pune accentul pe performanta modelului obtinuta pe setul de testare, intrucat dorim sa obtinem un model care sa prezinte o performanta optima si care sa prezinte o putere mare de generalizare pe date noi.

Nr.Crt.	.metric	.estimate	model	date
1	accuracy	0.741	RF_5	test
2	sens	0.744	RF_5	test
3	spec	0.739	RF_5	test
4	ppv	0.717	RF_5	test
5	npv	0.764	RF_5	test
6	precision	0.717	RF_5	test
7	accuracy	0.741	XGB_5	test
8	sens	0.744	XGB_5	test
9	spec	0.739	XGB_5	test
10	ppv	0.717	XGB_5	test
11	npv	0.764	XGB_5	test

Nr.Crt.	.metric	.estimate	model	date
1	accuracy	1	RF_5	train
2	sens	1	RF_5	train
3	spec	1	RF_5	train
4	ppv	1	RF_5	train
5	npv	1	RF_5	train
6	precision	1	RF_5	train
7	accuracy	0.77	XGB_5	train
8	sens	0.76	XGB_5	train
9	spec	0.779	XGB_5	train
10	ppv	0.754	XGB_5	train
11	npv	0.785	XGB_5	train

Fig. 4.5. Indicatorii de performanta obtinuti pentru modelul cu toate variabilele a) pe setul de testare si b) pe setul de antrenare.

In cazul modelului generat cu algoritmul XGB pe setul de antrenare valorile obtinute pentru metrii arata că modelul prezinta o performanță moderată în clasificarea vinurilor. Sensibilitatea și specificitatea sunt relativ ridicate, indicând că modelul este bun atât în detectarea vinurilor de calitate bună, cât și a celor de calitate slabă, însă acuratețea nu este la nivelul ridicat, ceea ce sugerază că modelul poate fi îmbunătățit în clasificarea corectă a instantelor. În ansamblu, acuratețea modelului este relativ bună, iar între sensibilitate și specificitate există un echilibru. Modelele obtinute pot fi îmbunătățite, mai ales în ceea ce privește precizia și sensibilitatea. Coeficiențul Kappa (0.539) indică un acord moderat între predictiile modelului și realitatea observată. 75% (PPV) din predictiile modelului pentru vinurile clasificate ca fiind vinuri de buna calitate sunt corecte, în timp ce 78,8% (NPV) dintre predictiile modelului pentru vinurile clasificate ca fiind de calitate slabă sunt corecte. Coeficientul de corelație Matthews (mcc) care măsoară calitatea de generalizare a modelului este de 0.539, indicând un acord moderat între predictia modelului și realitatea observată. Între predictiile modelului și realitatea observată există un grad moderat de similaritate (0.540, indicele Jaccard măsoară gradul de similaritate între predictia modelului și realitatea observată). 47,5% din observațiile din setul de date sunt etichetate ca fiind pozitive.

Nr.Crt.	.metric	.estimate	model	date
1	accuracy	0.744	RF_3	test
2	sens	0.738	RF_3	test
3	spec	0.75	RF_3	test
4	ppv	0.724	RF_3	test
5	npv	0.763	RF_3	test
6	precision	0.724	RF_3	test
7	accuracy	0.744	XGB_3	test
8	sens	0.738	XGB_3	test
9	spec	0.75	XGB_3	test
10	ppv	0.724	XGB_3	test
11	npv	0.763	XGB_3	test

Nr.Crt.	.metric	.estimate	model	date
1	accuracy	0.812	RF_3	train
2	sens	0.823	RF_3	train
3	spec	0.801	RF_3	train
4	ppv	0.787	RF_3	train
5	npv	0.836	RF_3	train
6	precision	0.787	RF_3	train
7	accuracy	0.77	XGB_3	train
8	sens	0.769	XGB_3	train
9	spec	0.772	XGB_3	train
10	ppv	0.75	XGB_3	train
11	npv	0.789	XGB_3	train

Fig. 4.6. Indicatorii de performanta obtinuti pentru modelul cu 3 variabilele a) pe setul de testare si b) pe setul de antrenare.

Performantele modelului cu 3 variabile generate cu XGB și RF pe setul de testare sunt foarte asemănătoare, prezintând rezultate similare pentru toți indicatorii de evaluare, iar alegerea uneia dintre cele două modele se va face, fie tinând cont de: complexitatea modelului, timpul de antrenare sau interpretabilitatea rezultatelor. Astfel, în cazul în care interpretabilitatea modelului și timpul de antrenare sunt importante, iar setul de date nu este foarte mare, RF poate fi o alegere mai bună. În cazul în care, performanța este importantă și setul de date este mare, XGBoost ar putea fi o opțiune bună.

În ceea ce privește modelul RF observăm că rezultatele obtinute pe setul de antrenare indică o performanță perfectă (ceea ce suferă ca modelul să-a potrivit foarte bine cu datele de antrenare, clasificând corect toate instantele), în timp ce pe setul de testare

performanta acestuia scade, prezintand o acuratete de 74% ceea ce indica faptul ca modelul nu are o putere mare de generalizare pe seturi de date noi, insa si faptul ca prezinta overfitting.

In ceea ce priveste, valorile obtinute pentru precizie si factorul F acestea sunt comparabile (RF_3 având o precizie de 72.4% și un factor F de 73.1%, iar XGB_3 având o precizie de 71.7% și factor F de 73%) ceea ce denota capacitatea modelelor de a previziona vinurile de buna calitate si cele de slaba calitate. Observam ca valorile obtinute pentru sensitivitate si specificitate sunt aproximativ egale pentru cele doua modele (RF_3 având o sensibilitate de 73.8% și o specificitate de 75%, în timp ce XGB_3 are o sensibilitate de 74.4% și o specificitate de 73.9%) ceea ce denota ca modelele au o capacitate echilibrata de identificare a probelor de vin de calitate buna, respectiv slaba. De asemenea, capacitatea modelului RF_3 de clasificare a probelor de vin este usor mai buna, lucru sustinut de valorile obtinute pentru acuratete (acuratetea modelului XGB_3 (74,4%) este usor mai mare fata de cea a modelului RF_3 (74,1%)).

4.2.2. Modele bazate pe arbori de decizie CART

4.2.2.1. Rezultatele obtinute pentru modelul generate cu algoritmul CART (tidymodels)

In Anexa 4 prezentam evolutia metricilor (roc_auc si accuracy) in functie de valorile hiperparametrilor pentru modelele cu toate variabilele, respectiv cu 3 variabile generate cu algoritmul CART. Din multitudinea de valori generate se selecteaza combinatia optima (cei mai buni hiperparametri) de hiperparametri si se adauga la fluxul de lucru pentru a genera si antrena modelele utilizand setul de antrenare. In final este selectat cel mai bun model.

```

parsnip model object
n= 1019
node), split, n, loss, yval, (yprob)
  * denotes terminal node

1) root 1019 480 good_wine (0.4710500 0.5289500)
  2) alcohol< 9.975 429 123 bad_wine (0.7132867 0.2867133)
    4) total_sulfur_dioxide>=99.5 65 0 bad_wine (1.0000000 0.0000000) *
    5) total_sulfur_dioxide< 99.5 364 123 bad_wine (0.6620879 0.3379121)
      10) volatile_acidity>=0.6075 147 28 bad_wine (0.8095238 0.1904762) *
      11) volatile_acidity< 0.6075 217 95 bad_wine (0.5622120 0.4377880)
        22) sulphates< 0.645 133 47 bad_wine (0.6466165 0.3533835) *
        23) sulphates>=0.645 84 36 good_wine (0.4285714 0.5714286) *
  3) alcohol>=9.975 590 174 good_wine (0.2949153 0.7050847)
    6) alcohol< 11.15 343 141 good_wine (0.4110787 0.5889213)
      12) sulphates< 0.625 154 64 bad_wine (0.5844156 0.4155844) *
      13) sulphates>=0.625 189 51 good_wine (0.2698413 0.7301587)
        26) total_sulfur_dioxide>=84 10 2 bad_wine (0.8000000 0.2000000) *
        27) total_sulfur_dioxide< 84 179 43 good_wine (0.2402235 0.7597765) *
    7) alcohol>=11.15 247 33 good_wine (0.1336032 0.8663968)
      14) volatile_acidity>=0.77 15 6 bad_wine (0.6000000 0.4000000)
        28) alcohol< 11.65 11 2 bad_wine (0.8181818 0.1818182) *
        29) alcohol>=11.65 4 0 good_wine (0.0000000 1.0000000) *
      15) volatile_acidity< 0.77 232 24 good_wine (0.1034483 0.8965517) *

```

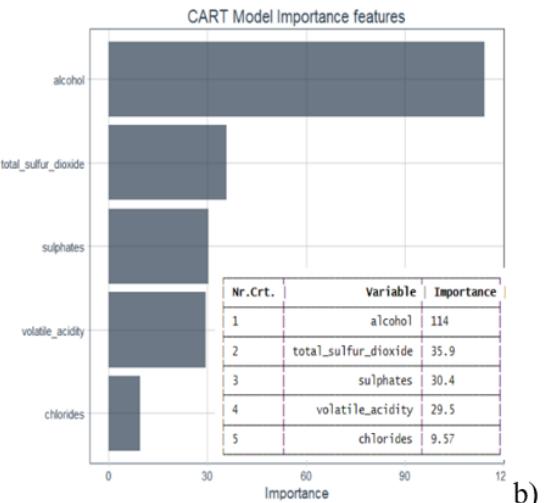


Fig. 4.7. Arboarele de decizie pentru modelul generat cu algoritmul a) CART si b) importanta variabilelor.

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.759	antrenare_CART
2	sens	0.785	antrenare_CART
3	spec	0.735	antrenare_CART
4	mcc	0.519	antrenare_CART
5	precision	0.725	antrenare_CART
6	recall	0.785	antrenare_CART
7	f_meas	0.754	antrenare_CART

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.735	testare_CART
2	sensitivity	0.8	testare_CART
3	specificity	0.678	testare_CART
4	recall	0.8	testare_CART
5	precision	0.688	testare_CART
6	f_meas	0.74	testare_CART
7	mcc	0.479	testare_CART
8	roc_auc	0.791	testare_CART

Fig. 4.8. Matricile obtinute pentru modelul CART optim cu toate variabilele pe setul de a) antrenare si b) testare.

Modelul generat cu algoritmul CART prezintă o performanță moderată în clasificarea vinurilor lucru observat din metricile obținute pe setul de testare. Valoarea acurateței indică faptul că în jur de trei sferturi din predicțiile modelului sunt corecte. Valoarea obținuta pentru sensibilitate este relativ mare (80%) ceea ce sugerează că modelul CART este eficient în detectarea vinurilor de calitate bună, identificând-le corect în majoritatea cazurilor. Cu toate acestea, specificitatea (67.78%) este mai mică decât sensibilitatea, indicând o tendință mai slabă de a identifica vinurile de calitate slabă. Aceasta ar putea însemna că modelul are o tendință de a supraestima vinurile de calitate bună, clasificând mai multe probe de vin ca fiind de calitate bună decât în realitate. Totuși, metricile preciziei și masurii F sugerează că modelul CART este echilibrat în capacitatea sa de a identifica corect atât vinurile de calitate bună, cât și cele de calitate slabă. Coeficientul de corelație Matthews (MCC) de aproximativ 47.91% indică o corelație moderată între predicțiile

modelului și realitatea observată. Tinând cont de valorile obținute pentru modelul generat cu algoritmul CART putem considera că acesta poate fi utilizat pentru clasificarea vinurilor. Mai mult, dacă ne uităm la valorile obținute pentru specificitate modelul poate fi îmbunătățit astfel încât să previzioneze corect ambele tipuri de probe devin.

4.2.2.2. Rezultatele obținute pentru modelul cu 3 variabile generat cu algoritmul CART (tidymodels)

```

parsnip model object
n= 1019

(node), split, n, loss, yval, (yprob)
  * denotes terminal node

1) root 1019 480 good_wine (0.4710500 0.5289500)
  2) alcohol< 9.975 429 123 bad_wine (0.7132867 0.2867133)
    4) sulphates< 0.575 179 29 bad_wine (0.8379888 0.1620112) *
    5) sulphates>=0.575 250 94 bad_wine (0.6240000 0.3760000)
      10) volatile_acidity>=0.405 196 61 bad_wine (0.6887755 0.3112245) *
      11) volatile_acidity< 0.405 54 21 good_wine (0.3888889 0.6111111) *
  3) alcohol>=9.975 590 174 good_wine (0.2949153 0.7050847)
    6) alcohol< 11.15 343 141 good_wine (0.4110787 0.5889213)
      12) sulphates< 0.625 154 64 bad_wine (0.5844156 0.4155844) *
      13) sulphates>=0.625 189 51 good_wine (0.2698413 0.7301587)
        26) volatile_acidity>=0.6375 30 14 bad_wine (0.5333333 0.4666667) *
        27) volatile_acidity< 0.6375 159 35 good_wine (0.2201258 0.7798742) *
    7) alcohol>=11.15 247 33 good_wine (0.1336032 0.8663968)
      14) volatile_acidity>=0.77 15 6 bad_wine (0.6000000 0.4000000)
        28) alcohol< 11.65 11 2 bad_wine (0.8181818 0.1818182) *
        29) alcohol>=11.65 4 0 good_wine (0.0000000 1.0000000) *
      15) volatile_acidity< 0.77 232 24 good_wine (0.1034838 0.8965517) *

```

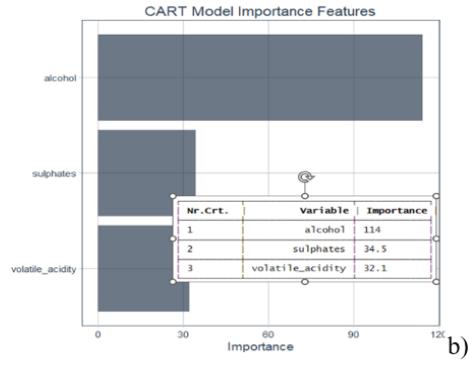


Fig. 4.9. Arboarele de decizie pentru modelul generat cu algoritmul a) CART și b) importanța variabilelor.

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.755	antrenare_CART3
2	sens	0.833	antrenare_CART3
3	spec	0.685	antrenare_CART3
4	mcc	0.521	antrenare_CART3
5	precision	0.702	antrenare_CART3
6	recall	0.833	antrenare_CART3
7	f_meas	0.762	antrenare_CART3

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.726	testare_CART3
2	sensitivity	0.812	testare_CART3
3	specificity	0.65	testare_CART3
4	recall	0.812	testare_CART3
5	precision	0.674	testare_CART3
6	f_meas	0.737	testare_CART3
7	mcc	0.466	testare_CART3
8	roc_auc	0.789	testare_CART3

Fig. 4.10. Matricile obținute pentru modelul CART cu 3 variabile pe setul de a) antrenare și b) testare.

4.2.2.3. Rezultatele obținute pentru modelul cu trei variabile generat cu algoritmul CART (clasic)

Pentru a construi și a afișa arboarele de decizie utilizând algoritmul CART, s-au folosit librăriile "rpart" și „rpart.plot”. Arboare de decizie au fost generati atât pe setul de antrenament, cât și pe setul de testare. Vizualizarea arborului de decizie rezultat ne-a permis să interpretăm și să înțelegem mai bine relațiile dintre variabilele independente și calitatea vinului.

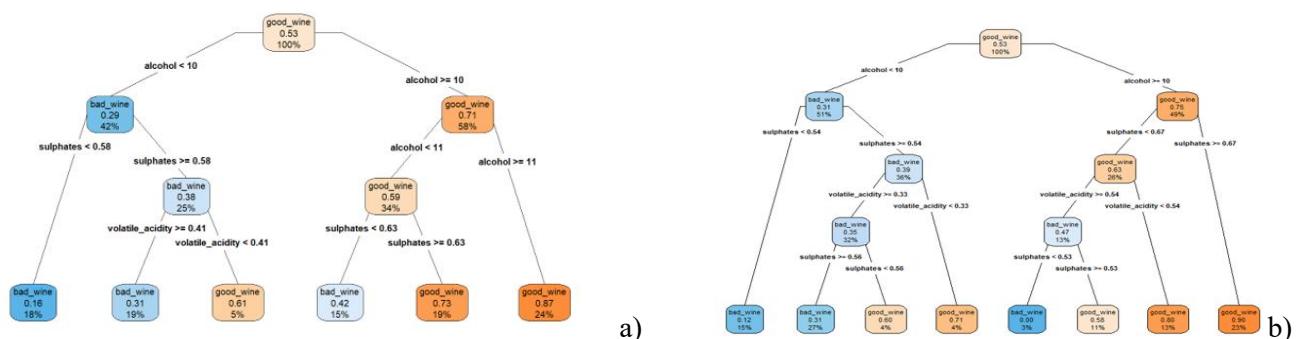


Fig. 4.11. Arboarele generat cu algoritmul CART obținut pe a) setul de antrenament b) pe setul de testare

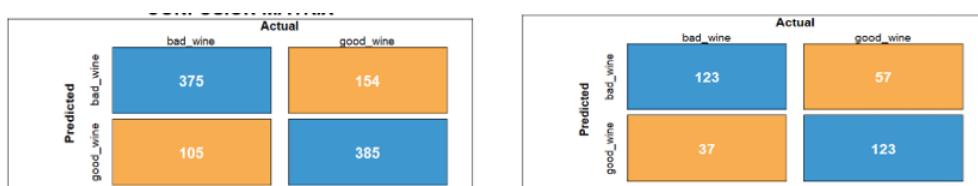


Fig. 4.12. Matricile de confuzie obtinute pe setul de testare/antrenare pentru modelul cu 3 variabilele.

După cum se poate observa în Fig.1, arborele inițial CART este dificil de citit, datorită complexității sale. Prin urmare, s-a încercat simplificarea acestuia cu ajutorul parametrului de complexitate (CP).

În Figura 1, putem observa că arborele inițial CART este dificil de interpretat din cauza complexității sale. Pentru a îmbunătăți înțelegerea și utilizarea acestuia, am aplicat o procedură de fasonare folosind **parametrul de complexitate (CP)**. Pentru a valida acest proces și a alege cea mai potrivită valoare a parametrului CP, am utilizat funcțiile **printcp()** și **plotcp()** din pachetul „rpart” în R. Parametrul CP este o măsură a complexității arborelui de decizie, iar găsirea unei valori optime ne permite să obținem un arbore mai generalizabil și mai puțin supradadaptat. În general, alegem CP-ul care **minimizează eroarea de validare încrucișată**, astfel încât să putem decide cât de mult dorim să fasonăm arborele. Funcția **printcp()** furnizează un tabel de valori optime bazate pe CP, în timp ce funcția **plotcp()** generează un grafic care ne arată modul în care se modifică eroarea relativă și parametrul de complexitate (CP) odată cu creșterea numărului de împărțiri ale arborelui.

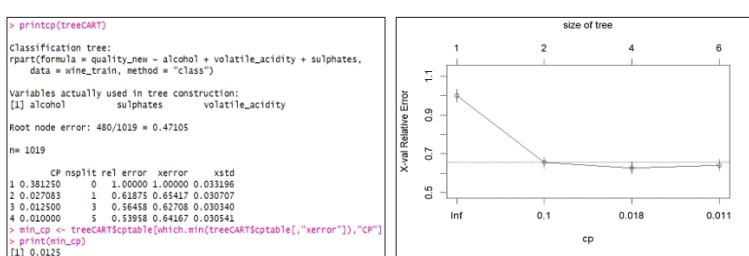


Fig.4.13. Valorile obtinute pentru cp, selectia valorii optime si arborele de decizie fasonat.

In Fig. 3 sunt prezentate valorile parametrului CP. Observam ca valoarea optimă a parametrului CP este 0.0125 și care are cea mai mică eroare asociată. La construirea arborelui final folosim această valoare pentru a fasona arborele CART inițial pe setul de antrenament, rezultatul fiind prezentat în Fig....

Pentru evaluarea performantei arborelui de decizie fasonat pe setul de testare, pe baza matricei de confuzie, s-au calculat metricile. Astfel, arborele fasonat, a prezis corect 133 din vinurile de calitate slabă și 112 din vinurile de calitate bună. Însă, 27 de vinuri au fost identificate ca fiind de calitate bună, ceea ce nu este corect, iar 68 de vinuri au fost identificate ca fiind de calitate slabă („bad_wine”), ceea ce este, de asemenea fals. Acuratețea arborelui de decizie după fasonare pe setul de testare este de aproximativ 72.67%, putin mai mare fata de acuratetea arborelui obținut inițial fără fasonare care avea 72.6%. De asemenea, arborele fasonat este mult mai simplu și usor de interpretat.

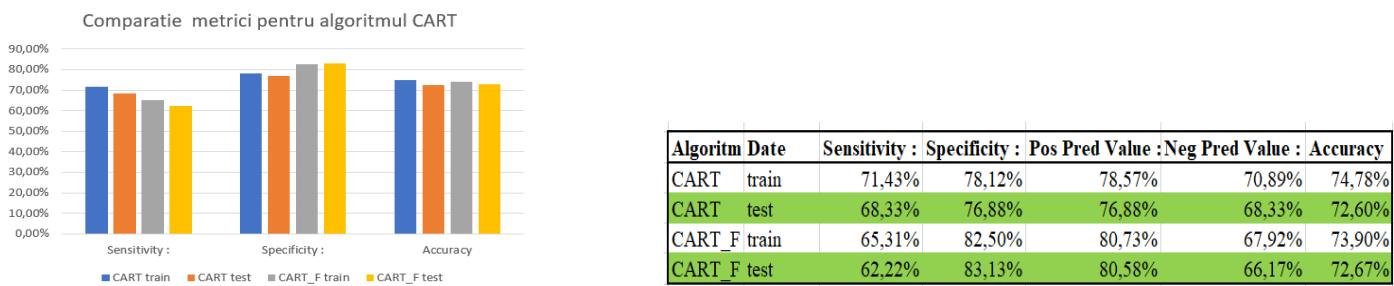


Fig.4.14. Compararea metricilor obtinute pentru modelul CART.

Acuratețea modelului pe setul de testare, este de aproximativ 72.67%. Sensibilitatea modelului (rata de detectare a vinurilor de calitate bună), este de aproximativ 62.22%, în timp ce specificitatea (rata de a detecta vinurile de calitate slabă), este de aproximativ 83.13%. Având în vedere că valorile obținute pentru acuratețe și specificitate sunt relativ ridicate, dar sensibilitatea este scăzută, performanța modelului este moderată. Valorile ridicate obținute pentru acuratețe și specificitate sugerează că modelul este eficient în identificarea vinurilor de calitate slabă, însă o valoare scăzută a sensibilității indică faptul că modelul poate fi îmbunătățit în ceea ce privește identificarea vinurilor de calitate bună. Astfel, performanța modelului în identificarea vinurilor de calitate bună poate fi îmbunătățită prin ajustări și optimizări suplimentare.

4.2.3. Modele bazate pe arbori de decizie CHAID

4.2.3.1. Rezultatele obtinute pentru modelul generat cu algoritmul CHAID (tidymodels)

```

parsnip model object

Model formula:
quality_new ~ volatile_acidity + chlorides + total_sulfur_dioxide +
sulphates + alcohol

Fitted party:
[1] root
| [2] alcohol <= 9.95
| | [3] volatile_acidity <= 0.605
| | | [4] total_sulfur_dioxide <= 98: bad_wine (n = 217, err = 43.8%)
| | | [5] total_sulfur_dioxide > 98: bad_wine (n = 42, err = 0.0%)
| | | [6] volatile_acidity > 0.605: bad_wine (n = 170, err = 16.5%)
| [7] alcohol > 9.95
| | [8] volatile_acidity <= 0.585
| | | [9] alcohol <= 11.1
| | | | [10] sulphates <= 0.71: good_wine (n = 134, err = 44.8%)
| | | | [11] sulphates > 0.71: good_wine (n = 88, err = 13.6%)
| | | | [12] alcohol > 11.1: good_wine (n = 192, err = 8.3%)
| | | [13] volatile_acidity > 0.585
| | | | [14] alcohol <= 11.4
| | | | | [15] volatile_acidity <= 0.84: bad_wine (n = 112, err = 50.0%)
| | | | | [16] volatile_acidity > 0.84: bad_wine (n = 26, err = 15.4%)
| | | | [17] alcohol > 11.4: good_wine (n = 38, err = 21.1%)

Number of inner nodes: 8
Number of terminal nodes: 9

```

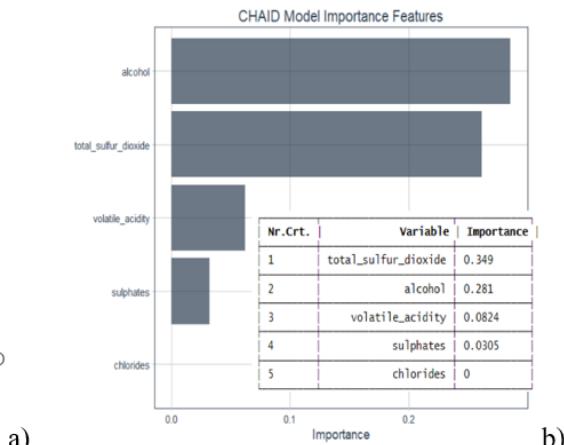


Fig. 4.15. Arborele de decizie pentru modelul generat cu algoritmul a) CHAID si b) importanta variabilelor.

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.726	antrenare_CHAID
2	sens	0.8	antrenare_CHAID
3	spec	0.66	antrenare_CHAID
4	mcc	0.463	antrenare_CHAID
5	precision	0.677	antrenare_CHAID
6	recall	0.8	antrenare_CHAID
7	f_meas	0.734	antrenare_CHAID

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.712	testare_CHAID
2	sensitivity	0.781	testare_CHAID
3	specificity	0.65	testare_CHAID
4	recall	0.781	testare_CHAID
5	precision	0.665	testare_CHAID
6	f_meas	0.718	testare_CHAID
7	mcc	0.433	testare_CHAID
8	roc_auc	0.772	testare_CHAID

Fig. 4.16. Matricile obtinute pentru modelul CHAID optim cu toate variabilele pe setul de a) antrenare si b) testare.

Performanța modelului generat cu algoritmul CHAID este moderată în clasificarea vinurilor, tinând cont de valorile obtinute pentru metrii pe setul de testare. Clasificatorul prezintă o rată de predicție a instantelor corecte de 71.18%. Valoare obtinută pentru sensibilitate este relativ mare (78.12%) ceea ce sugerează că modelul CHAID este eficient în detectarea vinurilor de calitate bună, clasificându-le corect în majoritatea cazurilor. Observăm că, valoarea obtinută pentru specificitate (65%) este mică, indicând o tendință mai slabă de a identifica vinurile de calitate slabă. Aceasta poate însemna că modelul are o tendință de a supraestima vinurile de calitate bună, clasificând mai multe vinuri ca fiind de calitate bună decât în realitate. Cu toate acestea, valorile obtinute pentru precizia și factorul F sugerează că modelul CHAID este echilibrat în identificarea corecta a vinurilor de ambele clase. Coeficientul de corelație Matthews (MCC) indică o corelație moderată între predicțiile modelului și realitate (43.29%).

Importanța variabilelor din cadrul modelului CHAID, arată că SiO₂ este cel mai relevant factor în clasificarea vinurilor, urmat de alcool, aciditate volatilă, sulfati și sare, fiind considerați factori importanți în evaluarea și clasificarea calității vinului. Un echilibru între acești factori poate contribui la producerea unui vin echilibrat de bună calitate. De asemenea, în funcție de tipul de vin, de regiunea de producție și de preferințele consumatorilor poate varia și legatura dintre caracteristicile fizico – chimice și calitatea vinului. Observăm că ordinea primele trei variabile se păstrează și pentru modelul cu 3 variabile (Fig. 4.17.).

4.2.3.2. Rezultatele obtinute pentru modelul cu 3 variabile generat cu algoritmul CHAID (tidymodels)

```

parsnip model object

Model formula:
quality_new ~ alcohol + sulphates + volatile_acidity

Fitted party:
[1] root
| [2] alcohol <= 9.95
| | [3] volatile_acidity <= 0.605
| | | [4] sulphates < 0.64: bad_wine (n = 170, err = 27.6%)
| | | [5] sulphates > 0.64: good_wine (n = 89, err = 46.1%)
| | [6] volatile_acidity > 0.605: bad_wine (n = 170, err = 16.5%)
[7] alcohol > 9.95
| [8] volatile_acidity <= 0.585
| | [9] alcohol <= 11.1
| | | [10] sulphates <= 0.71: good_wine (n = 134, err = 44.8%)
| | | [11] sulphates > 0.71: good_wine (n = 88, err = 13.6%)
| | [12] alcohol > 11.1: good_wine (n = 192, err = 8.3%)
| [13] volatile_acidity > 0.585
| | [14] alcohol < 11.1
| | | [15] volatile_acidity <= 0.84: bad_wine (n = 112, err = 50.0%)
| | | [16] volatile_acidity > 0.84: bad_wine (n = 26, err = 15.4%)
| | [17] alcohol > 11.1: good_wine (n = 38, err = 21.1%)

Number of inner nodes: 8
Number of terminal nodes: 9

```

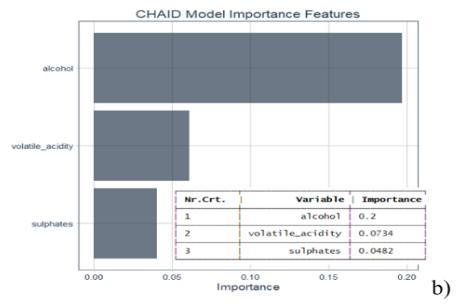


Fig. 4.17. Arborele de decizie pentru modelul generat cu algoritmul a) CHAID si b) importanta variabilelor.

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.733	antrenare_CHAID3
2	sens	0.715	antrenare_CHAID3
3	spec	0.75	antrenare_CHAID3
4	mcc	0.464	antrenare_CHAID3
5	precision	0.718	antrenare_CHAID3
6	recall	0.715	antrenare_CHAID3
7	f_meas	0.716	antrenare_CHAID3

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.712	testare_CHAID3
2	sensitivity	0.694	testare_CHAID3
3	specificity	0.728	testare_CHAID3
4	recall	0.694	testare_CHAID3
5	precision	0.694	testare_CHAID3
6	f_meas	0.694	testare_CHAID3
7	mcc	0.422	testare_CHAID3
8	roc_auc	0.777	testare_CHAID3

Matricea de confuzie

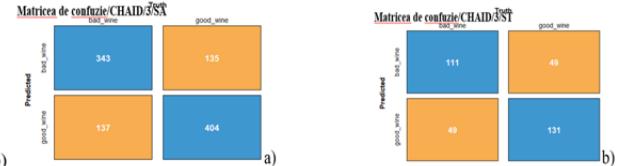


Fig. 4.18. Matricile obtinute pentru modelul CHAID cu toate variabilele pe setul de a) antrenare si b) testare.

4.2.3.3. Rezultatele obtinute pentru modelul generat cu algoritmul CHAID (clasic)

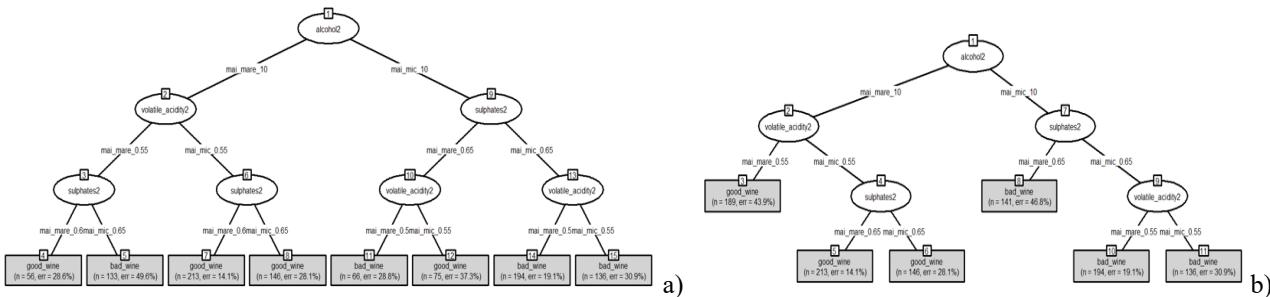


Fig. 4.19. Arborele generat cu algoritmul CHAID obtinut a) fara reguli de oprire si b) fasonat (variabile discretizate)

Arborele generat cu algoritmul CHAID pe set de antrenament a fost realizat fara criterii de oprire (sunt folosite cele prestatibile). Dupa cum se poate observa acesta are o complexitate destul de crescută, ceea ce face dificila interpretarea sa.

```

Model formula:
quality_new ~ alcohol2 + volatile_acidity2 + sulphates2

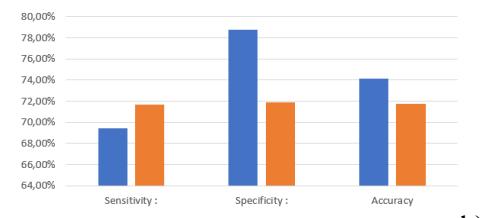
Fitted party:
[1] root
| [2] alcohol2 in mai_mare_10
| | [3] volatile_acidity2 in mai_mare_0.55: good_wine (n = 189, err = 43.9%)
| | [4] volatile_acidity2 in mai_mic_0.55
| | | [5] sulphates2 in mai_mare_0.65: good_wine (n = 213, err = 14.1%)
| | | [6] sulphates2 in mai_mic_0.65: good_wine (n = 146, err = 28.1%)
| | [7] alcohol2 in mai_mic_10
| | | [8] sulphates2 in mai_mare_0.65: bad_wine (n = 141, err = 46.8%)
| | | [9] sulphates2 in mai_mic_0.65
| | | | [10] volatile_acidity2 in mai_mare_0.55: bad_wine (n = 194, err = 19.1%)
| | | | [11] volatile_acidity2 in mai_mic_0.55: bad_wine (n = 136, err = 30.9%)

```

Number of inner nodes: 5
Number of terminal nodes: 6

a)

Comparatie metrii pentru algoritmul CHAID



b)

Fig. 4.20. Prezentam a) setul de reguli pentru arborele fasonat si b) metricele obtinute pentru cele doua modele (variabile discretizate).

Pentru generarea arborelui fasonat am stabilit urmatoarele valori pentru hiperparametri: minsplit = 200, minbucket = 100, maxheight = 3. Dupa cum se poate observa acuratetea arborelui fasonat (71,77%) este usor mai mica comparativ cu cea a arborelui initial (74,10%), respectiv complexitatea acestuia a scazut. La fiecare iteratie divizarea setului de date

se face pe baza unei caracteristici și a unui prag de decizie, minimizând o anumită metrică, cum ar fi impuritatea nodului sau eroarea de clasificare. Pentru fiecare clasa previzionată modelul precizează și eroarea de clasificare. Astfel, observăm că erorile de predictie pe regula 1 și 4 au crescut, însă nu pierdem foarte mult din informații.

Acuratețea modelului reprezentată de proporția de clasificări corecte în raport cu numărul total de clasificări, este de aproximativ 71.77%. Sensibilitatea modelului (rata de detectare a vinurilor de calitate bună), este de aproximativ 71.67%, în timp ce specificitatea (rata de a detecta vinurile de calitate slabă), este de aproximativ 71.88%. Având în vedere că valorile obținute pentru acuratețe și sensibilitate sunt relativ ridicate, dar apropiate ca valoare de valoarea specificitatii, performanța modelului în ceea ce privește clasificarea vinurilor este moderată. Cu toate acestea, există o tendință de a nu face diferența clară între vinurile de calitate bună și cele de calitate proastă, ceea ce poate însemna că modelul nu este la fel de precis în identificarea vinurilor de calitate proastă. Acest lucru poate duce la erori de clasificare, în special în cazul vinurilor de calitate proastă, deoarece modelul poate să le clasifice greșit ca fiind de calitate bună sau invers. Prin urmare, performanța modelului în identificarea vinurilor de calitate proastă poate fi mai slabă decât în identificarea celor de calitate bună. Astfel, performanța modelului poate fi îmbunătățită prin ajustarea și optimizarea acestuia pentru a obține o specificitate mai bună.

ceea ce înseamnă că modelul are tendința de a supraestima vinurile de calitate bună.

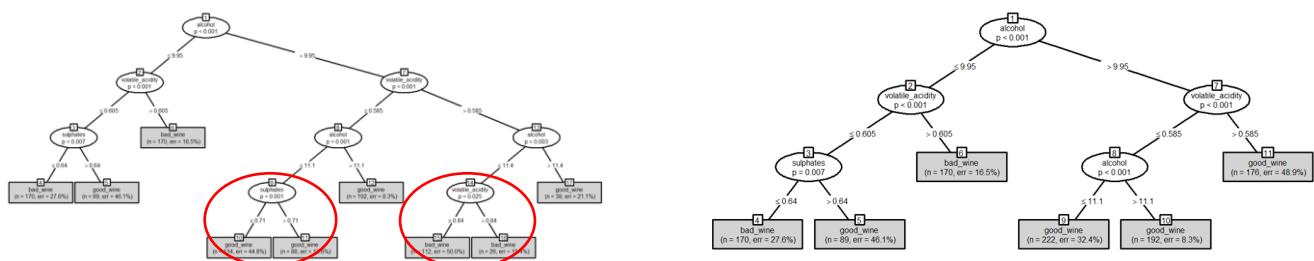


Fig. 4.21. Arborele generat cu algoritmul CHAID obținut a) fără reguli de oprire și b) fasonat (variabile numerice)

In Fig. 9 este prezentat arborele generat cu algoritmul CHAID, pe set de antrenare, realizat fără criterii de oprire. Acesta are o complexitate destul de crescută, comparativ cu arborele fasonat. Dupa cum se poate observa arborele complex prezintă un set de 9 reguli, în timp ce arborele fasonat prezintă un set de 6 reguli.

Astfel, arborele complex a fost fasonat aplicând urmatoarele criterii de oprire: o adâncime maximă de 3 și un număr minim de 200 de observații necesare pentru a diviza un nod. Rezultatul obținut este un arbore cu 5 noduri și 6 frunze, aşa cum este prezentat și în Fig.11. Dupa cum se poate observa nodul 13 a fost tăiat și nodul 9 a fost transformat într-un nod frunză.

```

Model formula: quality_new ~ volatile_acidity + sulphates + alcohol
Fitted party:
[1] root
 [2] alcohol <= 9.95
   [3] volatile_acidity <= 0.605
     [4] sulphates <= 0.64: bad_wine (n = 170, err = 27.6%)
     [5] sulphates > 0.64: good_wine (n = 89, err = 46.1%)
   [6] volatile_acidity > 0.605: bad_wine (n = 170, err = 16.5%)
 [7] alcohol > 9.95
   [8] volatile_acidity <= 0.585
     [9] alcohol <= 11.1
       [10] sulphates <= 0.71: good_wine (n = 134, err = 44.8%)
       [11] sulphates > 0.71: good_wine (n = 88, err = 13.6%)
     [12] alcohol > 11.1: good_wine (n = 192, err = 8.3%)
   [13] volatile_acidity > 0.585
     [14] alcohol <= 11.4
       [15] volatile_acidity <= 0.84: bad_wine (n = 112, err = 50.0%)
       [16] volatile_acidity > 0.84: bad_wine (n = 26, err = 15.4%)
     [17] alcohol > 11.4: good_wine (n = 38, err = 21.1%)
Number of inner nodes: 8
Number of terminal nodes: 9

Model formula: quality_new ~ volatile_acidity + sulphates + alcohol
Fitted party:
[1] root
 [2] alcohol <= 9.95
   [3] volatile_acidity <= 0.605
     [4] sulphates <= 0.64: bad_wine (n = 170, err = 27.6%)
     [5] sulphates > 0.64: good_wine (n = 89, err = 46.1%)
   [6] volatile_acidity > 0.605: bad_wine (n = 170, err = 16.5%)
 [7] alcohol > 9.95
   [8] volatile_acidity <= 0.585
     [9] alcohol <= 11.1: good_wine (n = 222, err = 32.4%)
     [10] alcohol > 11.1: good_wine (n = 192, err = 8.3%)
   [11] volatile_acidity > 0.585: good_wine (n = 176, err = 48.9%)
Number of inner nodes: 5
Number of terminal nodes: 6

```

Fig. 4.22. Setul de reguli generat obținut pentru a) arborele complex și b) arborele fasonat (variabile numerice)

Pentru evaluarea performanței arborelui de decizie CHAID, am analizat matricea de confuzie pe setul de testare. Pe baza acesteia putem spune că arborele din fig. 11, a clasificat corect 90 din vinuri ca fiind de calitate slabă („bad_wine”), și 151 de vinuri ca fiind de calitate bună („good_wine”). Totuși, 70 de vinuri au fost clasificate incorect ca fiind de calitate slabă, iar 29 de vinuri au fost clasificate incorect ca fiind de calitate bună.

Tabelul Metrici obținute pentru modelul CHAID.

Algoritm	Date	Sensitivity : Specificity	Pos Pred Value	Neg Pred Val	Prevalen	Accuracy
CHAID	test	69,44%	78,75%	78,62%	69,61%	52,94% 74,10%
CHAID_F	test	71,67%	71,88%	74,14%	69,28%	52,94% 71,77%

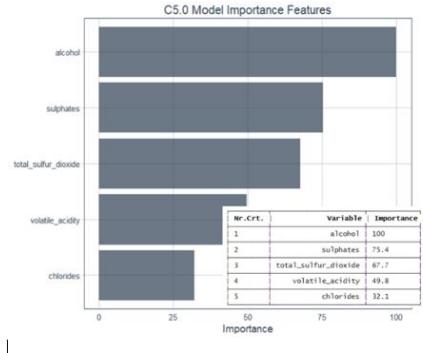
Acuratețea modelului complex, respectiv sensibilitatea și specificitatea sunt de aproximativ 69,44%, 78,75%, respectiv 74,10%. Prin urmare, modelul complex are o performanță moderată în ceea ce privește clasificarea vinurilor. Valori realtiv ridicate pentru acuratețe și sensibilitate indică o bună capacitate a modelului de a clasifica probele de vin din categoria good_wine, cat și din categoria bad_wine, însă specificitatea mai mică sugerează că modelul poate avea dificultăți în detectarea vinurilor de calitate slabă. Astfel, modelul ar putea fi optimizat pentru a îmbunătăți capacitatea acestuia de a distinge corect vinurile de calitate slabă. În ceea ce privește, modelul optimizat, performanța acestuia ramane moderată în ceea ce privește clasificarea vinurilor. Sensibilitatea și specificitatea sunt relativ echilibrate, insă valoarea predictivă pozitivă este destul de bună. Cu toate acestea, valoarea predictivă negativă ar putea fi îmbunătățită pentru a crește capacitatea modelului de a face predicții corecte în cazul vinurilor de calitate slabă.

În ceea ce privește importanța variabilelor în cadrul modelului fasonat, rezultatele arată că concentrația de alcool este cel mai relevant factor (100%) în clasificarea vinurilor, urmat de sulfuri (82,04%) și aciditatea volatilă (49,66%) și sunt considerate factori importanți în evaluarea și clasificarea calității vinului. Un echilibru adecvat între acești factori poate contribui la producerea unui vin echilibrat și placut. Relația dintre aceste caracteristici și calitatea vinului poate varia în funcție de tipul de vin, de regiunea de producție și de preferințele consumatorilor.

4.2.4. Modele bazate pe arbori de decizie C5.0

4.2.4.1. Rezultatele obținute pentru modelul generat cu algoritmul C5.0

```
Read 1019 cases (6 attributes) from undefined.data
Decision tree:
alcohol <= 10.2:
:...total_sulfur_dioxide > 98: bad_wine (68)
:...total_sulfur_dioxide <= 98:
:...sulphates <= 0.55:
:...chlorides > 0.097: bad_wine (18)
:...chlorides <= 0.097:
:...alcohol <= 9.7:
:...total_sulfur_dioxide <= 88: bad_wine (82/10)
:...total_sulfur_dioxide > 88: good_wine (6/1)
:...alcohol > 9.7:
:...volatile_acidity <= 0.45: good_wine (8/1)
:...volatile_acidity > 0.45: bad_wine (33/8)
:...sulphates > 0.365:
:...volatile_acidity <= 0.365:
:...chlorides <= 0.105: good_wine (39/6)
:...chlorides > 0.105: bad_wine (2)
:...volatile_acidity <= 0.365:
:...alcohol <= 9.4: bad_wine (101/31)
:...alcohol > 9.4:
:...volatile_acidity <= 0.53: good_wine (72/28)
:...volatile_acidity > 0.53: bad_wine (106/40)
alcohol > 10.2:
:...alcohol > 11.4: good_wine (183/17)
alcohol <= 11.4:
:...sulphates <= 0.64:
:...volatile_acidity <= 0.86: bad_wine (14/2)
:...volatile_acidity > 0.86:
:...sulphates <= 0.47: bad_wine (9/1)
:...sulphates > 0.47: good_wine (123/56)
sulphates > 0.64:
:...total_sulfur_dioxide > 71:
:...total_sulfur_dioxide > 96: bad_wine (5)
:...total_sulfur_dioxide <= 96:
:...total_sulfur_dioxide > 78: good_wine (5)
:...total_sulfur_dioxide <= 78:
:...sulphates <= 0.76: bad_wine (4)
:...sulphates > 0.76: good_wine (2)
total_sulfur_dioxide <= 71:
:...chlorides <= 0.089: good_wine (113/11)
:...chlorides > 0.089:
:...sulphates > 0.75: good_wine (8)
:...sulphates <= 0.75:
:...total_sulfur_dioxide > 42: good_wine (4)
:...total_sulfur_dioxide <= 42:
:...alcohol <= 11.2: bad_wine (12/2)
:...alcohol > 11.2: good_wine (2)
```



a)

b)

Fig. 4.23. Arboarele de decizie pentru modelul generat cu algoritmul a) C5.0 și b) importanța variabilelor.

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.79	antrenare_c5.0
2	sens	0.75	antrenare_c5.0
3	spec	0.826	antrenare_c5.0
4	mcc	0.578	antrenare_c5.0
5	precision	0.793	antrenare_c5.0
6	recall	0.75	antrenare_c5.0
7	f_meas	0.771	antrenare_c5.0

a)

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.735	testare_c5.0
2	sensitivity	0.688	testare_c5.0
3	specificity	0.778	testare_c5.0
4	recall	0.688	testare_c5.0
5	precision	0.733	testare_c5.0
6	f_meas	0.71	testare_c5.0
7	mcc	0.468	testare_c5.0
8	roc_auc	0.791	testare_c5.0

b)

Fig. 4.24. Matricile obținute pentru modelul C5.0 cu toate variabilele pe setul de a) antrenare și b) testare.

Acuratețea modelului generat cu algoritmul C5.0, respectiv sensibilitatea și specificitatea, pe setul de testare, sunt 73,5%, 68,8%, respectiv 77,80%. Modelul C5.0 prezintă o performanță moderată și are o capacitate relativ echilibrată

de a clasifica vinurile. Astfel, clasificatorul prezinta o performanță moderată în ceea ce priveste predicatia calitatii vinului. Valoarea realtiv mica pentru sensibilitate indică o capacitate redusa a modelului de a clasifica probele de vin din categoria good_wine comparativ cu modelul CHAID, însă specificitatea mai mare comparativ cu cea a modelului CHAID sugerează că modelul C5.0 prezinta o capacitate buna de detectare a vinurilor de calitate slabă. Valorile preciziei si factorului F sunt moderate, indicând faptul că modelul C5.00 poate identifica corect vinurile de ambele clase.

In ceea ce priveste importanta variabilelor in cadrul modelului C5.0, rezultatele arată că concentrația de alcool este cel mai relevant factor (100%) în clasificarea vinurilor, urmat de sulfuri (75,04%), SiO₂ (67,7%), aciditatea volatilă (49,8%) si sare (32,1%) acestea fiind considerate factori importanți în evaluarea și clasificarea calității vinului. Un echilibru între acești factori poate contribui la producerea unui vin echilibrat și placut. Relația dintre aceste caracteristici si calitatea vinului poate varia în funcție de tipul de vin, de regiunea de producție și de preferințele consumatorilor.

4.2.4.2. Rezultatele obtinute pentru modelul cu 3 variabile generat cu algoritmul C5.0 (tidymodels)

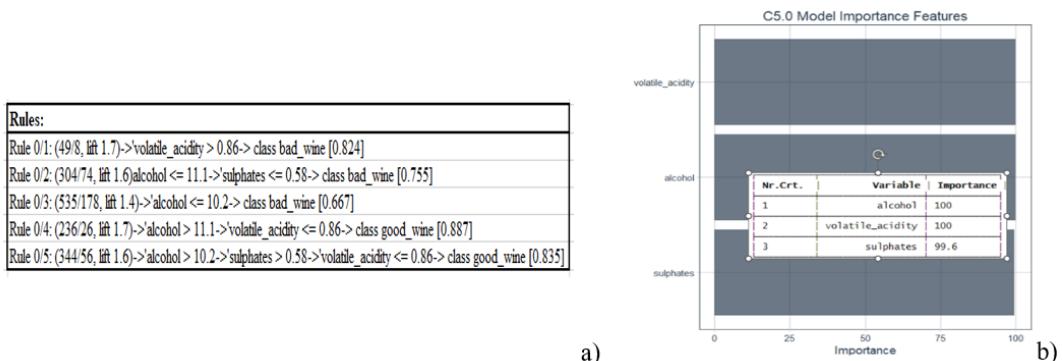


Fig. 4.25. Arborele de decizie pentru modelul generat cu algoritmul a) C5.0 si b) importanta variabilelor.

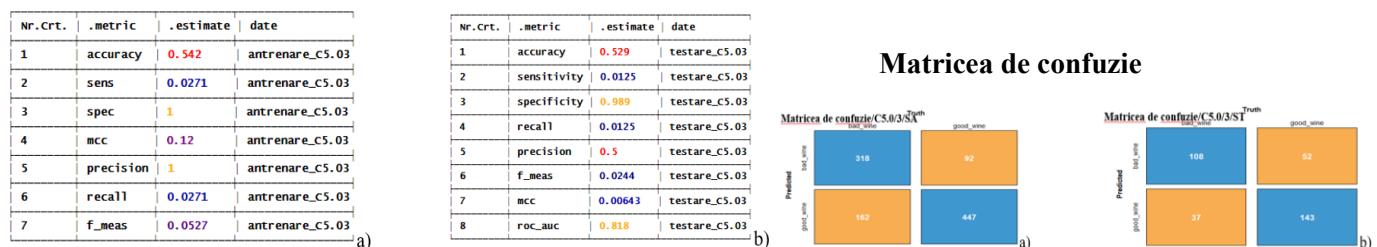


Fig. 4.26. Matricile obtinute pentru modelul C5.0 cu 3 variabile pe setul de a) antrenare si b) testare.

4.2.4.3. Rezultatele obtinute pentru modelul generat cu algoritmul C5.0 (clasic)

Arborele initial generat cu algoritmul C5.0 a fost realizat fara reguli de oprire (Fig. ...). Acesta are marimea de 9 si o eroare totala de 24.8% . Pentru construirea arborelui algoritmul a folosit in procent de 100% variabila “alcohol”, 80.47% a folosit “ sulphates”, si 64.57% volatile_acidity.

Acuratețea modelului, respectiv sensibilitatea si sepecificitatea sunt de aproximativ 69,44%, 78,75%, respectiv 74,10%. Prin urmare, modelul complex are o performanță moderată în ceea ce priveste clasificarea vinurilor. Valori realtiv ridicate pentru acuratețe și sensibilitate indică o bună capacitate a modelului de a clasifica vinurile în general, însă specificitatea mai mică sugerează că modelul poate avea dificultăți în detectarea vinurilor de calitate slabă. Astfel, modelul ar putea fi optimizat pentru a îmbunătăți capacitatea acestuia de a distinge corect vinurile de calitate slabă.

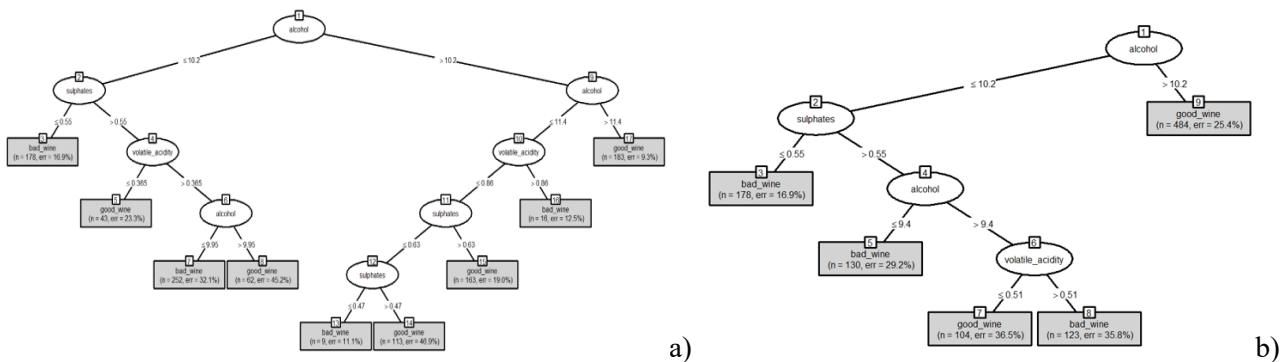


Fig.4.27. Arborele generat cu algoritmul C5.0 obtinut a) fara reguli de oprire si b) fasonat (variabile numerice).

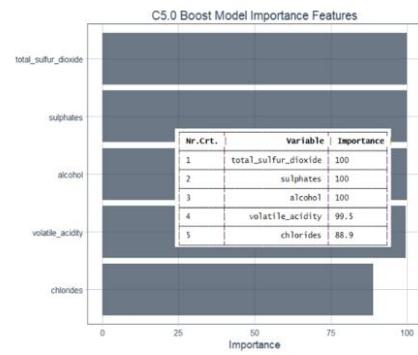
Pe baza acestor valori, putem spune că modelul complex are o performanță moderată în ceea ce privește clasificarea vinurilor. Deși acuratețea și sensibilitatea sunt relativ ridicate, specificitatea ar putea fi îmbunătățită pentru a crește capacitatea modelului de a identifica corect vinurile de calitate slabă. În ceea ce privește modelul fasonat, acesta are o performanță moderată în ceea ce privește clasificarea vinurilor. Sensibilitatea și specificitatea sunt relativ echilibrate, iar acuratețea este destul de bună. Cu toate acestea, valoarea predictivă pozitivă și negativă ar putea fi îmbunătățite pentru a crește capacitatea modelului de a face predicții corecte, în cazul vinurilor de calitate bună și slabă.

Algoritm	Date	Sensitivity : Specificity	Pos Pred Value : Neg Pred Val	Prevalen	Accuracy
C5.0	test	78,89% : 67,50%	73,20% : 73,97%	52,94%	73,19%
C5.0_F	test	71,46% : 76,62%	73,13% : 75,09%	47,11%	74,19%

4.2.5. Modele bazate pe arbori de decizie C5.0 Boost

4.2.5.1. Rezultatele obtinute pentru modelul generat cu algoritmul C5.0 Boost

```
Decision tree:
alcohol > 9.8:
...alcohol > 11.5: good_wine (41.2)
: alcohol <= 11.5:
: ...volatile_acidity > 0.67: bad_wine (102.7/37.9)
: ...volatile_acidity <= 0.67:
:   ...total_sulfur_dioxide > 86: bad_wine (27.8/8.7)
:   ...total_sulfur_dioxide <= 86:
:     ...sulphates > 0.63: good_wine (143.3/20.6)
:     ...sulphates <= 0.63:
:       ...sulphates <= 0.47: bad_wine (12.5/1.4)
:       ...sulphates > 0.47:
:         ...total_sulfur_dioxide <= 26:
:           ...total_sulfur_dioxide > 19: good_wine (65.8/25.3)
:           ...total_sulfur_dioxide > 19: bad_wine (21.9/2)
:         ...total_sulfur_dioxide > 26:
:           ...alcohol > 11: good_wine (11.3)
:           ...alcohol <= 11:
:             ...alcohol <= 10.9: good_wine (70.9/18.9)
:             ...alcohol > 10.9: bad_wine (8.8)
alcohol <= 9.8:
...total_sulfur_dioxide > 91: bad_wine (37.3)
total_sulfur_dioxide <= 91:
: ...sulphates > 0.67:
:   ...sulphates > 1.06: bad_wine (6.7)
:   ...sulphates <= 1.06:
:     ...total_sulfur_dioxide <= 19: bad_wine (16.8/4.1)
:     ...total_sulfur_dioxide > 19:
:       ...total_sulfur_dioxide <= 53: good_wine (64.4/14.6)
:       ...total_sulfur_dioxide > 53: bad_wine (26.5/9.8)
sulphates <= 0.67:
...sulphates <= 0.52: bad_wine (36.6/2.3)
sulphates > 0.52:
...volatile_acidity > 0.585: bad_wine (67.7/8.6)
...volatile_acidity <= 0.585:
...chlorides > 0.1: bad_wine (8.9)
chlorides <= 0.1:
...chlorides > 0.088: good_wine (13.1/7.6)
chlorides <= 0.088:
...volatile_acidity < 0.365: good_wine (13.5/4.1)
...volatile_acidity > 0.365:
...total_sulfur_dioxide <= 16: bad_wine (10.5)
...total_sulfur_dioxide > 16:
...  ...total_sulfur_dioxide <= 26: good_wine (19.8/5.6)
...  ...total_sulfur_dioxide > 26: bad_wine (77.1/17.6) a)
```



b)

Fig. 4.28. Arborele de decizie pentru modelul generat cu algoritmul a) C5.0Boost si b) importanta variabilelor.

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.816	antrenare_C5.0_Boost
2	sens	0.798	antrenare_C5.0_Boost
3	spec	0.833	antrenare_C5.0_Boost
4	mcc	0.632	antrenare_C5.0_Boost
5	precision	0.81	antrenare_C5.0_Boost
6	recall	0.798	antrenare_C5.0_Boost
7	f_meas	0.804	antrenare_C5.0_Boost

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.729	testare_C5.0_Boost
2	sensitivity	0.725	testare_C5.0_Boost
3	specificity	0.733	testare_C5.0_Boost
4	recall	0.725	testare_C5.0_Boost
5	precision	0.707	testare_C5.0_Boost
6	f_meas	0.716	testare_C5.0_Boost
7	mcc	0.458	testare_C5.0_Boost
8	roc_auc	0.786	testare_C5.0_Boost

Matricea de confuzie

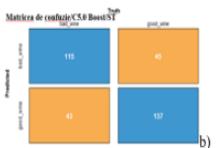
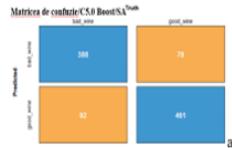


Fig. 4.29. Matricile obtinute pentru modelul C5.0 Boost cu toate variabilele pe a) antrenare si b) testare.

4.2.5.2. Rezultatele obtinute pentru modelul cu 3 variabile generat cu algoritmul C5.0 Boost (tidymodels)

```
Read 1019 cases (4 attributes) from undefined.data

Decision tree:

alcohol <= 10.2:
: ...sulphates <= 0.55: bad_wine (178/30)
:   sulphates > 0.55:
:     : ...volatile_acidity <= 0.365: good_wine (43/10)
:       volatile_acidity > 0.365: bad_wine (314/115)
alcohol > 10.2:
: ...alcohol > 11.4: good_wine (183/17)
alcohol <= 11.4:
: ...sulphates > 0.64: good_wine (155/30)
  sulphates <= 0.64:
    : ...sulphates <= 0.57: bad_wine (73/28)
    sulphates > 0.57: good_wine (73/31)
```

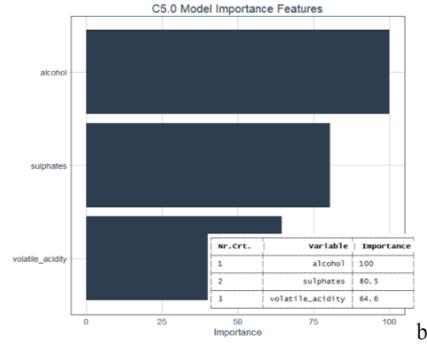


Fig. 4.30. Arborele de decizie pentru modelul generat cu algoritmul a) C5.0 Boost si b) importanta variabilelor.

In Fig. ... este prezentat setul de reguli generat cu algoritmul C5.0 Boost, respectiv importanta variabilelor in cadrul modelului. Observam ca concentrația de alcool este cel mai relevant factor (100%) în clasificarea vinurilor, urmat de sulfuri (80,5%) și aciditatea volatălă (64,6%) acestea fiind considerate factori importanți în evaluarea și clasificarea calității vinului.

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.752	antrenare_C5.03
2	sens	0.71	antrenare_C5.03
3	spec	0.788	antrenare_C5.03
4	mcc	0.501	antrenare_C5.03
5	precision	0.749	antrenare_C5.03
6	recall	0.71	antrenare_C5.03
7	f_meas	0.729	antrenare_C5.03

Nr.Crt.	.metric	.estimate	date
1	accuracy	0.735	testare_C5.03
2	sensitivity	0.675	testare_C5.03
3	specificity	0.789	testare_C5.03
4	recall	0.675	testare_C5.03
5	precision	0.74	testare_C5.03
6	f_meas	0.706	testare_C5.03
7	mcc	0.468	testare_C5.03
8	roc_auc	0.788	testare_C5.03

Matricea de confuzie

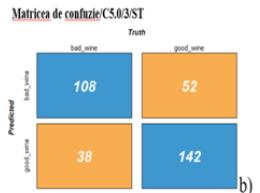
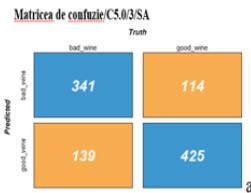


Fig. 4.31. Matricile obtinute pentru modelul C5.0 Boost cu toate variabilele pe a) antrenare si b) testare.

Limitele studiului si directii viitoare de cercetare

Studiul realizat poate fi îmbunătățit utilizarea unor seturi mai mari de date. Deoarece in cadrul acestei lucrari am utilizat doar setul de date pentru vinul roșu, cercetările viitoare pot include și date despre vinul alb pentru a compara rezultatele și a oferi sugestii mai cuprinzătoare. În plus, setul de date folosit în acest articol conține doar 1599 de observații. Pentru a obține o predicție mai exactă, ar fi mai bine să se mărească numărul de observații din setul de date. De asemenea, setul de date este din anul 2009, iar rezultatele pot să nu fie relevante pentru piața vinurilor din 2022 și ar

trebuie actualizate pentru a reflecta situația actuală. Mai mult, setul de date conține doar 11 predictori (dintre care un număr mare dintre acestia oferă informații similare, respectiv un număr mic de predictori au un impact semnificativ asupra variabilei obiectiv), însă există multe alte variabile care ar putea influența calitatea vinului roșu. În cercetările viitoare vom refață analiza și vom standardiza variabilele.

De asemenea, vom aplica standardizarea variabilelor și vom analiza algoritmi ML selectați. Prin standardizare variabilelor, ne vom asigura că variabilele independente sunt aduse la același numitor comun, conducând astfel la obținerea unor modele îmbunătățite, mai performante, și la o înțepătire mai bună a rezultatelor. În studiu nostru, variabila *quality* are două clase balansate.

Concluzii

În lucrarea de față ne-am propus să facem o analiză comparativă a performanțelor modelelor obținute cu ajutorul mai multor modele ML bazati pe arbori de clasificare, modele utilizate pentru a previziona/estima calitatea vinului roșu pe baza proprietăților fizico – chimice măsurate în etapa de certificare a vinului roșu. Variabila obiectiv este furnizată printr-o analiză senzorială care evaluatează modul în care răspunsul (calitatea vinului) este afectat atunci când valorile variabilelor independente se modifică. Cu scopul de a obține parametri optimi dintr-un volum mic de date cu efort minim, am analizat 6 modele diferite ML pe baza de arbori de clasificare: CHAID, CART, QUEST, C5.0, XGBoost, RandomForest.

Pentru a evalua precizia clasificărilor realizate de un model obținut cu ajutorul algoritmului ML utilizat și analizați în cadrul studiului (RF, XGB, CHAID, C5.0, CART, C5.0 Boost), am atrenat, calibrat și ajustat modelele bazate pe arbori de decizie/clasificare obținute cu clasificatori: RF, XGB, CHAID, C5.0, CART, C5.0 Boost. De asemenea, selectarea variabilelor importante, gestionarea problemei colinearității și limitarea numărului de predictori fără a compromite acuratețea modelului s-a realizat în două moduri: prima metodă presupune utilizarea MRLM și a doua metodă presupune folosirea a trei algoritmi ML, și anume: MRLM, RF și XGBoost.

Spre deosebire de XGBoost care permite, cu o ajustare atentă a hiperparametrilor, obținerea unor modele performante, algoritmul RF sunt predispuși la overfitting. În cadrul lucrării s-a luat în considerare utilizarea unor algoritmi ML pentru selectarea variabilelor pe baza importanței acestora. Desi, fiecare algoritm ML prezintă anumite avantaje și poate fi aplicat pe diferite tipuri de date, este esențială selecția variabilelor și a modelelor ML adecvate, intrucât un model prea simplu poate să nu transmită eficient ideea de bază, iar un model prea complex poate supraesantiona datele.

Anexa 1

Tabelul.1.1. Descrierea variabilelor prezente in setul de date red wine¹.

Variabila	Descrierea variabilei	Tipul variabilei
Fixed acidity (aciditate fixa)	Majoritatea zidzi cintrici. Exprimat in g / dm ³ .	Numeric, continuu
Volatile acidity (aciditatea volatila)	Reprezinta cantitatea de acid acetic din vin, care daca prezinta valori ridicate vinul are un gust neplacut de otet. Aciditatea volatila corespunde acizilor organici cu lanturi scurte ce pot fi extrasi din porba de vin prin procese de distilare: acid formic, acid acetic, acid propionic și acid butiric. Exprimat in g / dm ³ .	Numeric, continuu
Citric acidity (acidul citric)	Daca se gaseste in cantitati mici, acidul citric poate adauga prospetime si savoare vinului. Aciul citric este un acid organic slab fara culoare si este continut in mod natural in fructele citrice. Exprimat in g / dm ³ .	Numeric, continuu
Residual sugar (zaharul rezidual)	Cantitatea de zahar care ramane in vind dupa ce procesul de fermentatie a luat sfarsit si sunt extrem de putine vinuri care contin mai putin de 1g/l. Cantitatea de zahar rezidual afecteaza gustul dulce al vinului. Exprimat in g / dm ³ .	Numeric, continuu
Chlorides (clorurile)	Reprezinta cantitatea de sare din vin. Continutul de cloruri prezent in vinul rosu se datoreaza ionilor extrasi si prezenti in pielea bobului de strugure. Exprimat in g / dm ³ .	Numeric, continuu
Free sulfure dioxide (bioxidul de sulf liber)	SO ₂ liber se afla in echilibru intre SO ₂ molecular (sub forma de gaz dizolvat) si ioni bisulfit. Aceste sulfiti liberi reactioneaza si confera vinului proprietati germicide si antioxidantice. Sulfite legati sunt cei care au reactionat (atat reversibil cat si ireversibil) cu alte molecule prezente in vin (molecule, pigmenți, aldehide sau zaharuri). Suma sulfitilor liberi si legati reprezinta concentratia totala de sulfite. Exprimat in mg / dm ³ .	Numeric, continuu
Total sulfur dioxide (dioxidul de sulf total)	Reprezinta cantitatea de SO ₂ liber si legat. Daca concentratia de SO ₂ liber este mica acesta nu poate fi detectat. Exprimat in mg / dm ³ .	Numeric, continuu
Density (densitatea)	Densitatea vinului este apropiata de cea a apei si este influentata de procentul de alcool si zahar. Cu ajutorul hidrometrului se masoara densitatea mustului, a vinului fermentat si a vinului in raport cu apa pura, raport denumit greutate specifica. Exprimat in g / cm ³ .	Numeric, continuu
pH	Descrie cat de acid sau bazic este un vin pe o scala cuprinsa intre 0 (foarte acid) si 14 (foarte bazic). Majoritatea vinurilor au aciditatea cuprinsa intre 3 – 4. Vinurile care au valori ale pH-ului mai mari de 3.65 prezinta o serie de provocari in timpul procesului de fermentatie si invecire, in timp ce vinurile cu aciditate ridicata de alterare. Prin urmare, in procesul de fermentatie se foloseste dioxidul de sulf sub forma de metabisulfit de potasiu pentru a mentine calitatea si stabilitatea vinului in timpul procesului de fermentatie si invecire.	Numeric, continuu
Sulphates (sulfati)	Un aditiv care contribuie la cresterea concentratiei de SO ₂ si care actioneaza ca un antibactericid. Sulfite din vin se gasesc in mod natural in concentratii mici si sunt	Numeric, continuu

¹ <https://www.kaggle.com/code/halimedogan/red-wine-quality-prediction>, accesat la data de 12.04.2024.

	unul dintre miile de subproduse schimice create in timpul procesului de fermentatie. Producatorul adauga sulfiti pentru a pastra si proteja vinul de bacterii. Exprimat in g / dm^3.	
Alcohol (alcoolul)	Reprezinta procentul de alcool din vin. Exprimat in procente (%).	Numeric, continuu
Quality (calitatea)	Variabila de iesire (obtinuta pe baza datelor senzoriale si are un scor cuprins intre 3 si 8)	Numeric discret

Tabelul 1.1. Descrierea variabilelor prezente in setul de date red wine².

pH	alcohol	density	quality	chlorides	sulphates	citric_acid	fixed_acidity	free_sulfur_dioxide	residual_sugar	total_sulfur_dioxide	volatile_acidity
Min. :2.74	Min. : 8.40	Min. :0.9901	Min. :3.000	Min. :0.01200	Min. :0.3300	Min. :0.0000	Min. : 4.600	Min. : 1.00	Min. : 0.900	Min. : 6.00	Min. :0.1200
1st Qu.:3.21	1st Qu.: 9.50	1st Qu.:0.9956	1st Qu.:5.000	1st Qu.:0.07000	1st Qu.:0.5500	1st Qu.:0.0900	1st Qu.: 7.100	1st Qu.: 7.00	1st Qu.: 1.900	1st Qu.: 22.00	1st Qu.:0.3900
Median :3.31	Median :10.20	Median :0.9967	Median :6.000	Median :0.07900	Median :0.6200	Median :0.2600	Median : 7.900	Median :14.00	Median : 2.200	Median : 38.00	Median :0.5200
Mean :3.31	Mean :10.43	Mean :0.9967	Mean :5.623	Mean :0.08812	Mean :0.6587	Mean :0.2723	Mean : 8.311	Mean :15.89	Mean : 2.523	Mean : 46.83	Mean :0.5295
3rd Qu.:3.40	3rd Qu.:11.10	3rd Qu.:0.9978	3rd Qu.:6.000	3rd Qu.:0.09100	3rd Qu.:0.7300	3rd Qu.:0.4300	3rd Qu.: 9.200	3rd Qu.:21.00	3rd Qu.: 2.600	3rd Qu.: 63.00	3rd Qu.:0.6400
Max. :4.01	Max. :14.90	Max. :1.0037	Max. :8.000	Max. :0.61100	Max. :2.0000	Max. :1.0000	Max. :15.900	Max. :72.00	Max. :15.500	Max. :289.00	Max. :1.5800

Tabelul 1.2. Descrierea variabilelor semnificative din setul de date.

```
> stat.desc(red_wine_final)
      volatile acidity chlorides total_sulfur_dioxide pH sulphates alcohol quality quality_new
nbr.val    1359.000000000 1359.000000000 1359.000000000 1359.000000000 1359.000000000 1359.000000000 1359.000000000 NA
nbr.null   0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 NA
nbr.na     0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 NA
min       0.120000000 0.012000000 6.0000000 2.740000000 0.330000000 8.400000000 3.000000000 NA
max       1.580000000 0.611000000 289.0000000 4.010000000 2.000000000 14.900000000 8.000000000 NA
range     1.460000000 0.599000000 283.0000000 1.270000000 1.670000000 6.500000000 5.000000000 NA
sum       719.560000000 119.760000000 63636.5000000 4498.000000000 895.180000000 14177.51666667 7642.000000000 NA
median    0.520000000 0.079000000 38.0000000 3.310000000 0.620000000 10.200000000 6.000000000 NA
mean      0.529477557 0.088123620 46.8259750 3.309786608 0.658704930 10.43231543 5.62325239 NA
SE.mean   0.004964959 0.001339411 0.9062605 0.004205559 0.004629558 0.02935241 0.02234061 NA
CI.mean.0.95 0.009739822 0.002627539 1.7778224 0.008250097 0.009081862 0.05758099 0.04382585 NA
var       0.033500463 0.002438075 1116.1576532 0.024036258 0.029127188 1.17086564 0.67828072 NA
std.dev   0.183031318 0.049376862 33.4089457 0.155036311 0.170666891 1.08206545 0.82357800 NA
coef.var  0.345682863 0.560313594 0.7134704 0.046841785 0.259094600 0.10372246 0.14645937 NA

```

² <https://www.kaggle.com/code/halimedogan/red-wine-quality-prediction>, accesat la data de 12.04.2024.

Anexa 2

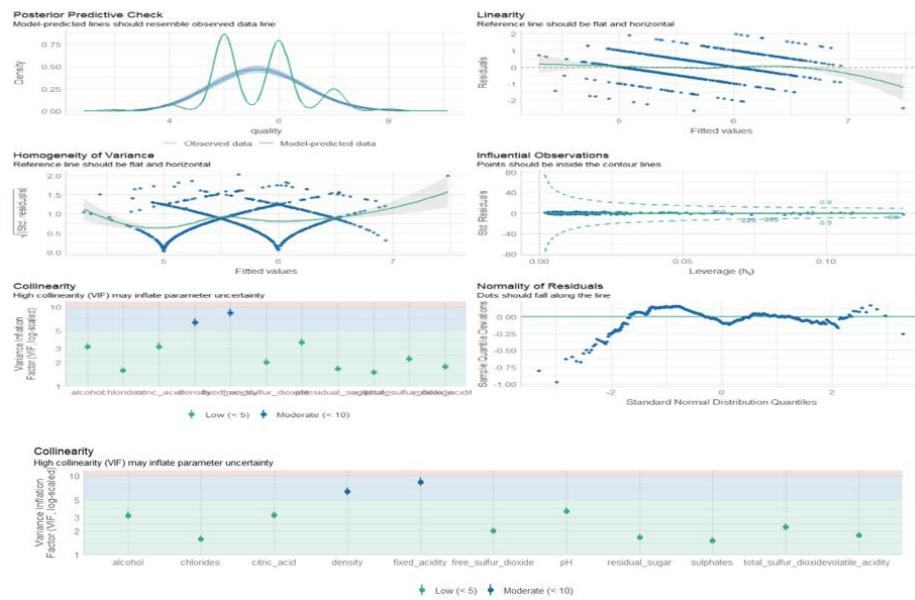


Fig. Verificarea ipotezelor cu privire la variabila reziduu si variabilele independente pentru modelului de regresie MRLM construit cu toate variabilele.

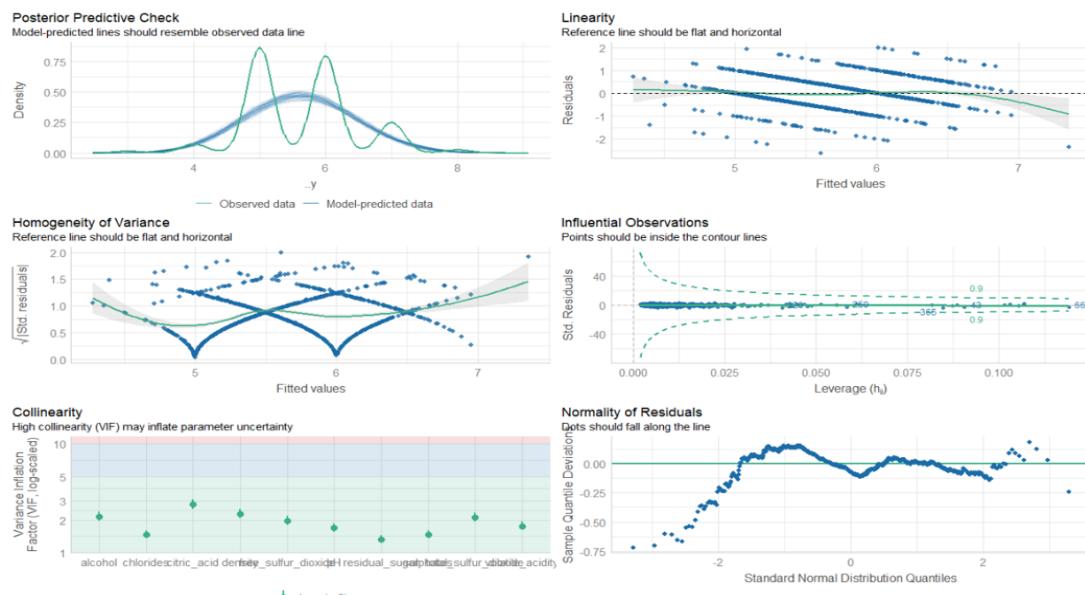


Fig. Verificarea ipotezelor cu privire la variabila reziduu si variabilele independente pentru modelului de regresie MRLM construit fara a include variabila fixed_acidity.

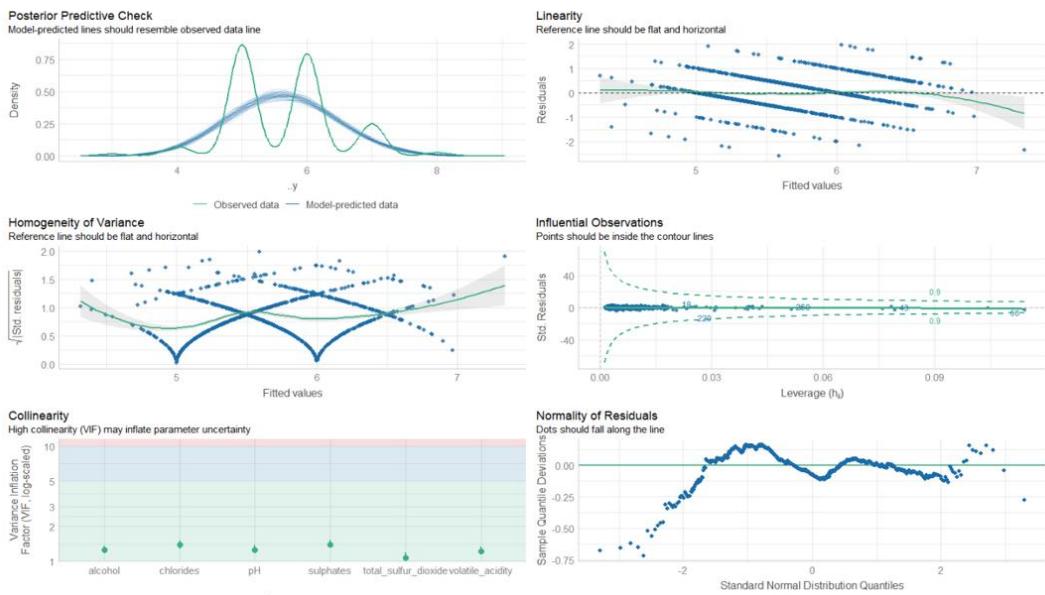


Fig. Verificarea ipotezelor cu privire la variabila reziduu si variabilele independente pentru modelului de regresie MRLM construit fara a include doar variabilele semnificative.

Anexa 3

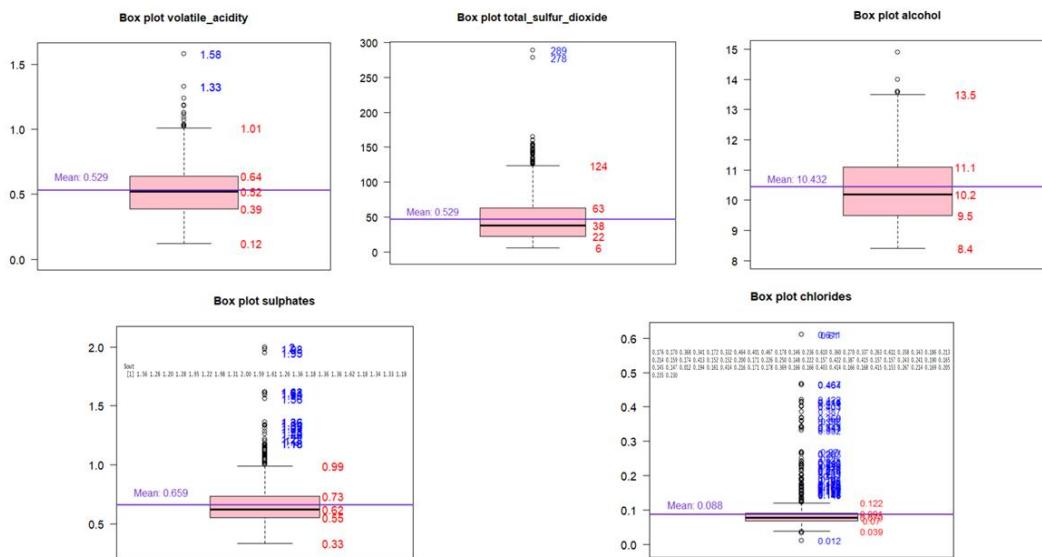


Fig. Diagramale boxplot corespunzatoare varibilelor analizate si indicarea outlierilor.

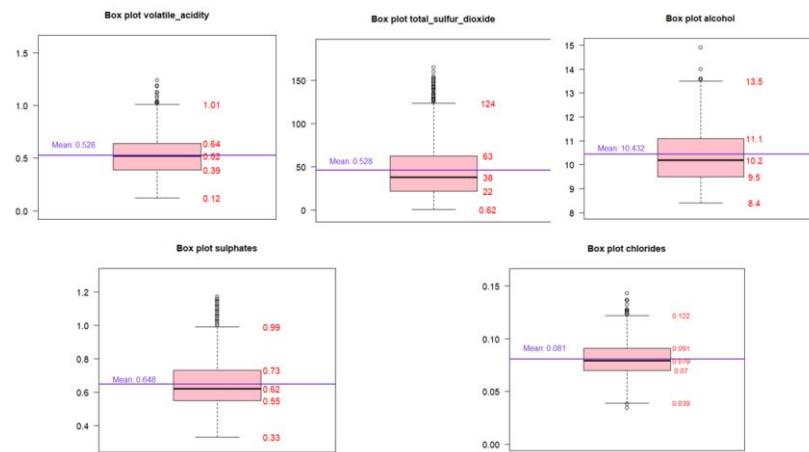


Fig. Diagramale boxplot corespunzatoare varibilelor analizate dupa eliminarea outlierilor.

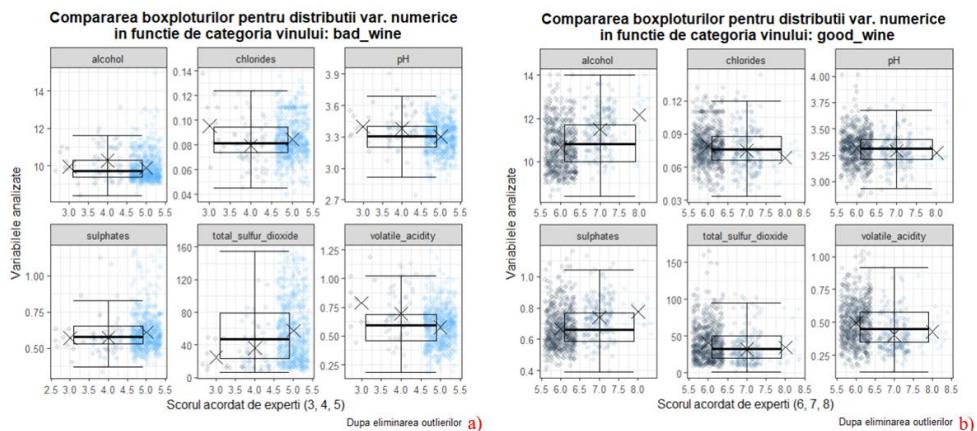


Fig. Boxploturile varabilelor analizate in functie de a) scorul acordat de experti si b) calitatea vinului

Anexa 4

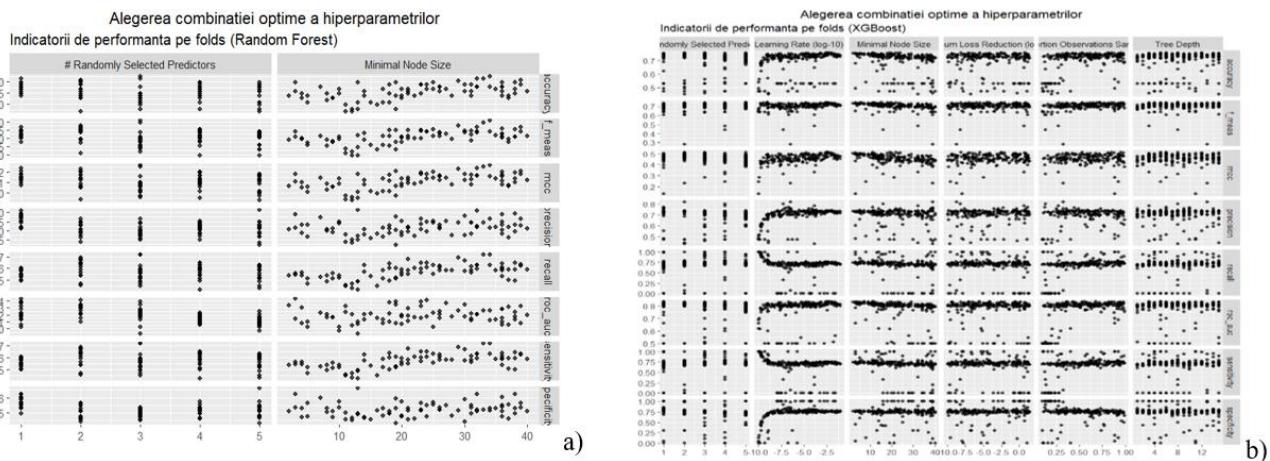


Fig. Identificarea hiperparametrilor pentru modelul cu toate variabilele generat cu algoritmul a) RF si b) XGBoost.

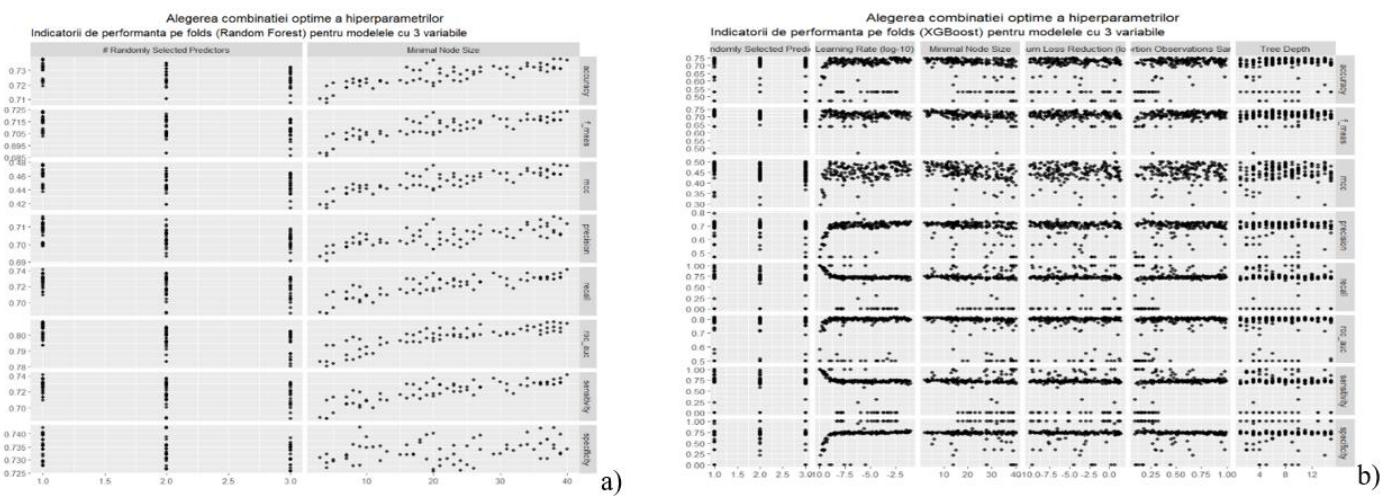


Fig. Identificarea hiperparametrilor pentru modelul cu 3 (alcool, sulfiti si aciditate volatila) variabile generat cu algoritmul a) RF si b) XGBoost.

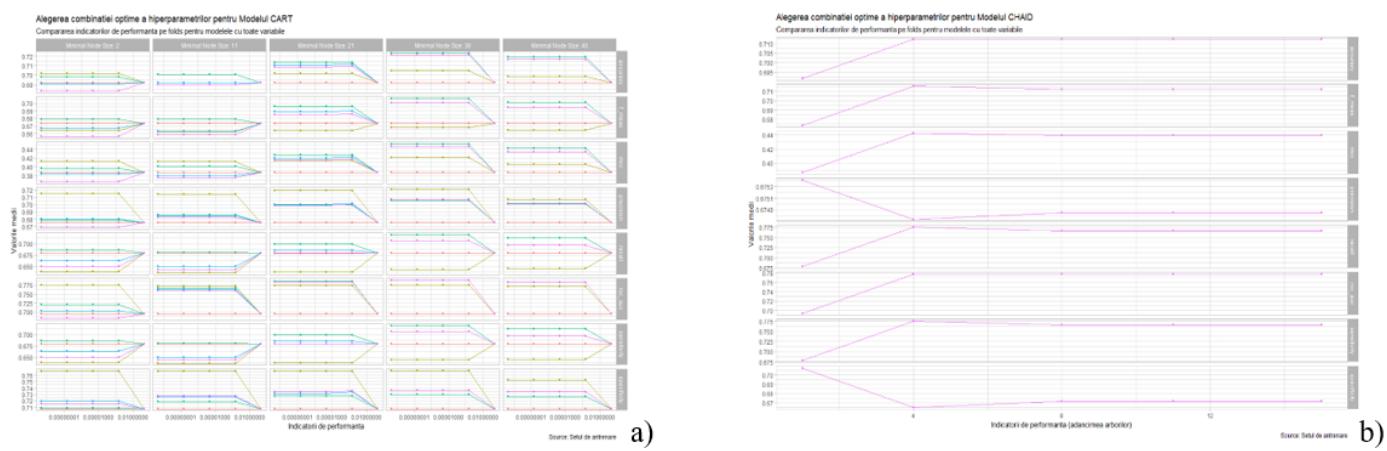


Fig. Identificarea hiperparametrilor pentru modelul cu toate variabilele generat cu algoritmul a) CART si b) CHAID.

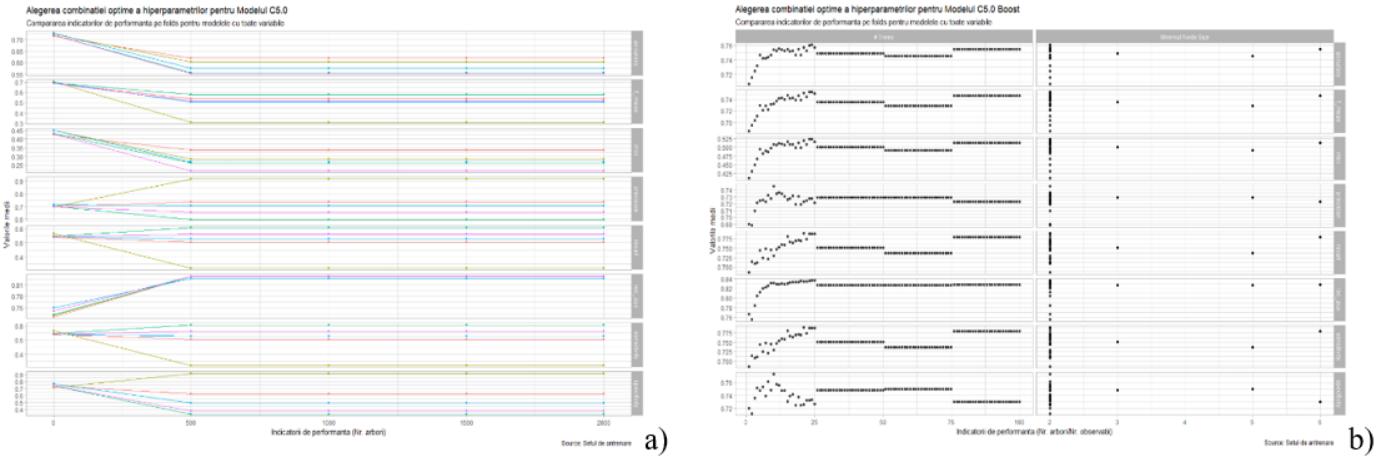


Fig. Identificarea hiperparametrilor pentru modelul cu toate variabilele generat cu algoritmul a) C5.0 si b) C5.0 Boost.

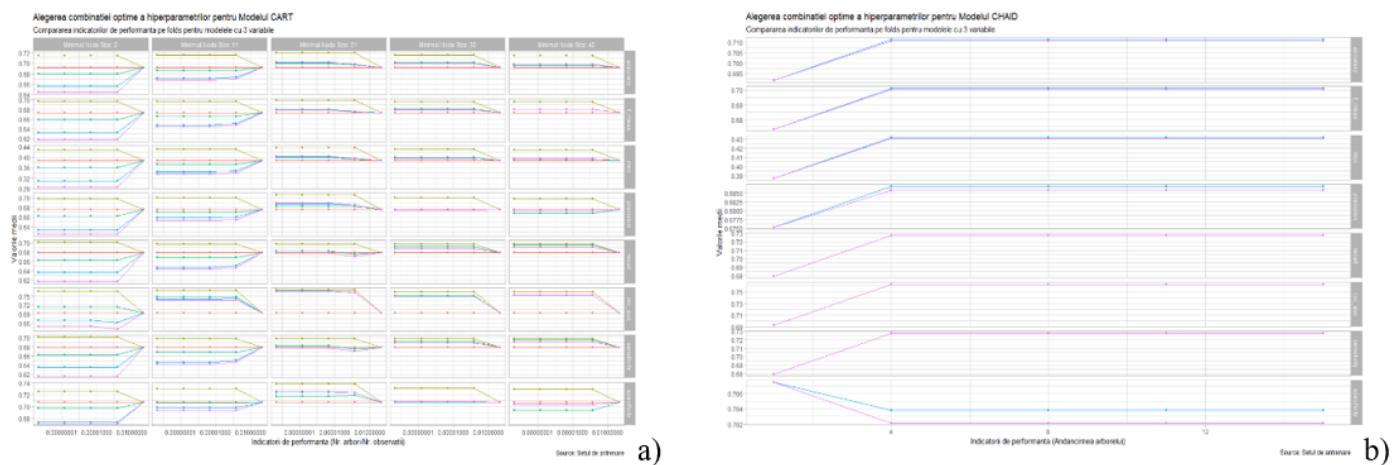


Fig. Identificarea hiperparametrilor pentru modelul cu 3 (alcool, sulfiti si aciditate volatila) variabilele generat cu algoritmul a) CART si b) CHAID.

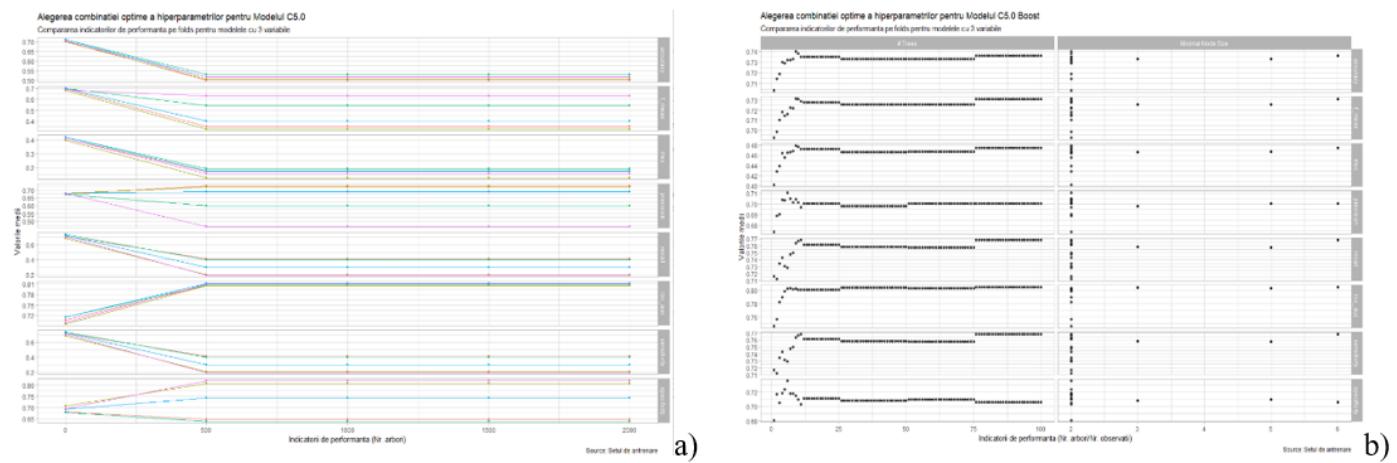


Fig. Identificarea hiperparametrilor pentru modelul cu 3 (alcool, sulfiti si aciditate volatila) variabilele generat cu algoritmul a) C5.0 si b) C5.0 Boost.

Bibliografie

D.B. Radosavljevic, S. Ilic, S.N. Pitulic, A data mining approach to wine quality prediction, 2019, International Scientific Conference UNITECHAt: Gabrovo, Bulgaria.

P. Cortez, A. Cerdeira, F. D. Almeida, T. Matos, J. L. Reis, Modeling wine preferences by data mining from physicochemical properties, 2009, Decision Support Systems, 47(4), 547-553, https://www.researchgate.net/publication/221612614_Using_Data_Mining_for_Wine_Quality_Assessment.

https://en.wikipedia.org/wiki/Protected_designation_of_origin, accesat la data de 03.05.2024.

D. Pawar, A. Mahajan, S. Bhoithe, Wine Quality Prediction using Machine Learning, 2019, International Journal of Computer Applications Technology and Research, 8, 09, pp. 385-388.

R. Chandra, K. Chaudhary, A. Kumar, The Combination and Comparison of Neural Networks With Decision Trees for Wine Classification, https://www.academia.edu/905652/The_Combination_and_Comparison_of_Neural_Networks_With_Decision_Trees_for_Wine_Classification, accesat la data 03.05.2024.

G. Astray, J. C. Mejuto, V. Martínez-Martínez, I. Nevares, M. Alamo-Sanza, J. Simal-Gandara, Prediction Models to Control Aging Time in Red Wine, 2019, Molecules, 24, 5, 826, <https://doi.org/10.3390/molecules24050826>.

N. Sharma, Quality Prediction of Red Wine based on Different Feature Sets Using Machine Learning Techniques, IJSR, ISSN: 2319-7064, 2019, <https://www.ijsr.net/archive/v9i7/SR20718002904.pdf>, accesat la data de 03.04.2024.

N. Khilari, P. Hadawale, H. Shaikh, S. Kolase, Analysis of Machine Learning Algorithm to predict WineQuality, International Journal of Scientific Research in Science, Engineering and Technology, 9, 2, pp. 231 – 236.

T. M. Geethanjali, M. Y. Sowjanya, S. N. Rohith, B. E. Shubashree, A. Charan, Prediction of wine quality using machine learning, 2021 JETIR, 8, 11, pp. 175 – 182.

D. Radosavljevic, S. Ilic, S. Pitulic, A data mining approach to wine quality prediction, 2019, International scientific conference 15 – 16,

M. Basalekou, P. Tataridis, K. Georgakis, C. Tsintonis, Measuring Wine Quality and Typicity, 2023, 9, 2, 41, <https://doi.org/10.3390/beverages9020041>.

R. T. Hodgson, An Examination of Judge Reliability at a major U.S. Wine Competition, Journal of Wine Economics, Volume 3, Issue 2, Fall 2008, Pages 105–113.

B. Zhan, Forecasting red wine quality: A comparative examination of machine learning approaches, 2024, Applied and Computational Engineering, 32, 58-65, https://www.researchgate.net/publication/377818483_Forecasting_red_wine_quality_A_comparative_examination_of_machine_learning_approaches?fbclid=IwZXh0bgNhZW0CMTAAAR2XfGLDMh0gM4-Sycu49ZOUz7jrvUCjMu0Ui9PApdY7TzLb1Y32XJ1Y8PE_aem_AazpUchX9FlYqoqreWtsFvmVidqdZoeVLP7eeWsnPot8WYEyrWpl08Chptq04pOEEllep_n8liMKNM7JohAqvZPc.

P. Cortez, A. Cerdeira, F. D. Almeida, T. Matos, J. L. Reis, Modeling wine preferences by data mining from physicochemical properties, 2009, Decision Support Systems, 47(4), 547-553, https://www.researchgate.net/publication/221612614_Using_Data_Mining_for_Wine_Quality_Assessment.

The effectiveness of PCA and various hyperparameter settings in SVM and KNN for wine quality estimation, 2024, Applied and Computational Engineering, 31, pp. 86-95, https://www.researchgate.net/publication/377826298_The_effectiveness_of_PCA_and_various_hyperparameter_settings_in_SVM_and_KNN_for_wine_quality_estimation?fbclid=IwZXh0bgNhZW0CMTAAAR1SdK4s0gW1j1Hj8x-u5yi55fMbBWtvPWDChji-O1xRXBp2-wVjf2tqllw_aem_AZ9NiSPA9U-C9naUBIQvpLs5CJ5WxoSnEk49RDPAqstxVc9CnVloCesHHFjvo8ywMw4fcNbs_uxwr6hJoOEj5Faa.

<https://archive.ics.uci.edu/datasets>, accesat la data de 10.04.2024.

<https://www.kaggle.com/datasets/piyushgoyal443/red-wine-dataset>, accesat la data de 20.04.2024.

<https://www.vinhoverde.pt/en/grape-varieties>, accesat la data 20.04.2024.

B.M. Greenwell, B. C. Boehmke, Variable Importance Plots—An Introduction to the *vip* Package, 2020, The R Journal, <https://koalaverse.github.io/vip/articles/vip.html>, <https://github.com/koalaverse/vip/blob/master/rjournal/RJwrapper.pdf>, accesat la data de 15.04.2024.

Y. Gupta, Selection of important features and predicting wine quality using machine learning techniques, Procedia Computer Science 125, 2018, 305-312, https://www.sciencedirect.com/science/article/pii/S1877050917328053?ref=pdf_download&fr=RR-2&rr=87357b5d7bfc0550, accesat la data de 13.04.2024.