FAKULTETA ZA INFORMACIJSKE ŠTUDIJE

V NOVEM MESTU

P R O J E K T N A N A L O G A

VISOKOŠOLSKEGA STROKOVNEGA ŠTUDIJSKEGA PROGRAMA

PRVE STOPNJE

JANEZ BUČAR

FAKULTETA ZA INFORMACIJSKE ŠTUDIJE

V NOVEM MESTU

PROJEKTNA NALOGA

KLASIFIKACIJA KAKOVOSTI VIN Z UPORABO BAYESOVIH MREŽ

Mentor (oz. Mentorica): Dr. Goran Klepac

Novo mesto, marec 2025

Janez Bučar

Kazalo vsebine

Kazalo Slik

1. UVOD

Za izdelavo Bayesovega napovednega modela smo uporabili programsko okolje GeNIe 5.0, dostopno na bayesfusion.com.

Podatke smo predhodno ustrezno diskretizirali (nizek / visok) in jih nato uvozili v GeNIe. Postopek čiščenja in priprave podatkov je v ločeni Notebook datoteki.

2. UVOZ PODATKOV

Na spodnji sliki je prikazan del vhodnih podatkov, uporabljenih za učenje strukture in parametrov Bayesove mreže.

	alcohol_bin	density_bin	volatile acidity_bin	chlorides_bin	citric acid_bin	qualit
▶	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	nizek	visok	visok	visok
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	nizek
	nizek	nizek	visok	visok	nizek	visok
	nizek	visok	visok	visok	nizek	visok
	visok	visok	visok	visok	visok	nizek
	nizek	visok	visok	visok	nizek	nizek
	visok	visok	visok	visok	visok	nizek
	nizek	nizek	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	nizek
	visok	visok	nizek	visok	visok	visok
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	visok	visok
	nizek	visok	nizek	visok	visok	visok
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	visok
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	nizek
	nizek	visok	visok	visok	nizek	nizek

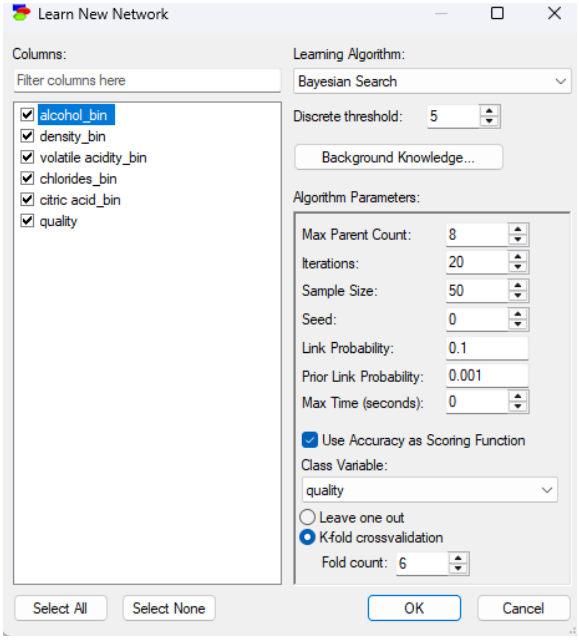
Slika 1: Uvoženi podatki

Po uvozu podatkov smo nadaljevali z učenjem Bayesovega modela. Uporabili smo funkcionalnost Learn Network, kjer smo izbrali algoritem Bayesian Search. Ta algoritem nam omogoča avtomatsko iskanje najboljše mrežne strukture glede na podatke.

V oknu za učenje smo nastavili:

- Discrete threshold = 5, da GeNIe obravnava binned spremenljivke kot diskretne
- Max Parent Count = 8 in 20 iteracij, kar omogoča dovolj kompleksnosti za strukturo
- Označili smo Use Accuracy as Scoring Function, da se mreža optimizira glede na klasifikacijsko natančnost
- Nastavili smo spremenljivko "**Quality**" kot ciljno spremenljivko.

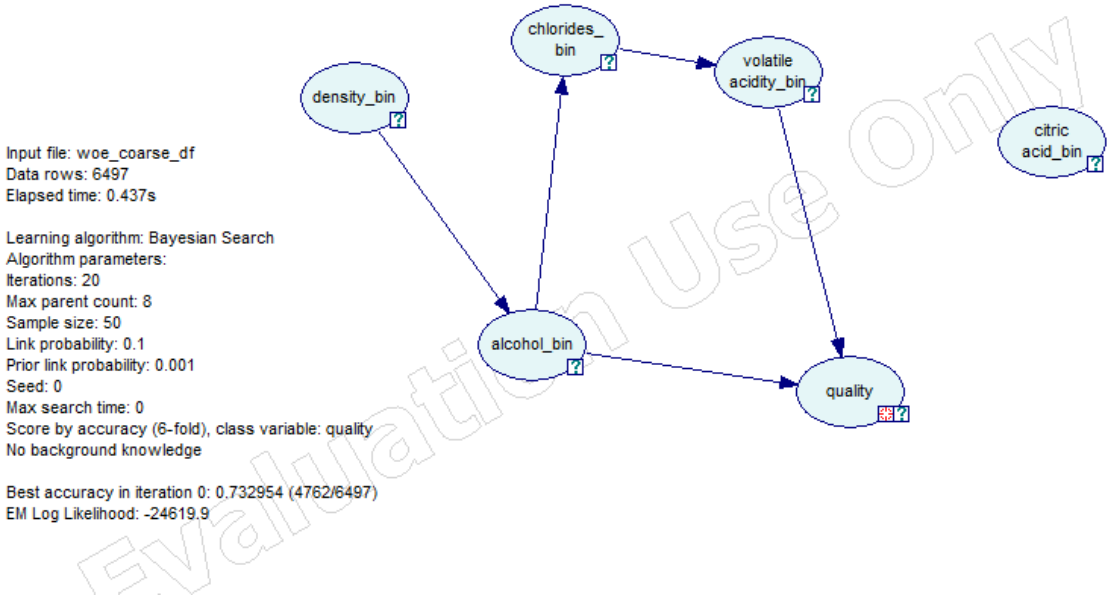
Za preverjanje posplošitvene sposobnosti modela smo uporabili metodo K-fold crossvalidation (*6-fold*), kar pomeni, da se podatki delijo na 6 delov – 5 jih služi za učenje, 1 za testiranje, in ta postopek se ponovi za vsak del.



Slika 2: Nastavitve učenja modela

3. MODEL

Na spodnji sliki je prikazan končni model Bayesove mreže, ki je bil ustvarjen na podlagi učenih podatkov. Opazimo več neposrednih in posrednih odvisnosti med spremenljivkami.



Slika 3: Model

Spremenljivka density_bin vpliva na alcohol_bin, ki ima pomembno vlogo pri napovedi ciljne spremenljivke quality. Poleg tega alkohol vpliva tudi na chlorides_bin in volatile_acidity_bin, pri čemer slednja prav tako neposredno vpliva na kakovost.

Zanimivo je, da citric_acid_bin nima povezave z nobeno drugo spremenljivko, kar pomeni, da v tem modelu ni pomembna za napoved kakovosti vina.

Ključna ugotovitev iz modela je to, da ima alcohol_bin največji vpliv na kakovost. Višja vrednost alkohola močno poveča verjetnost, da bo vino ocenjena kot visoke kakovosti – kar je skladno z našim pričakovanjem.

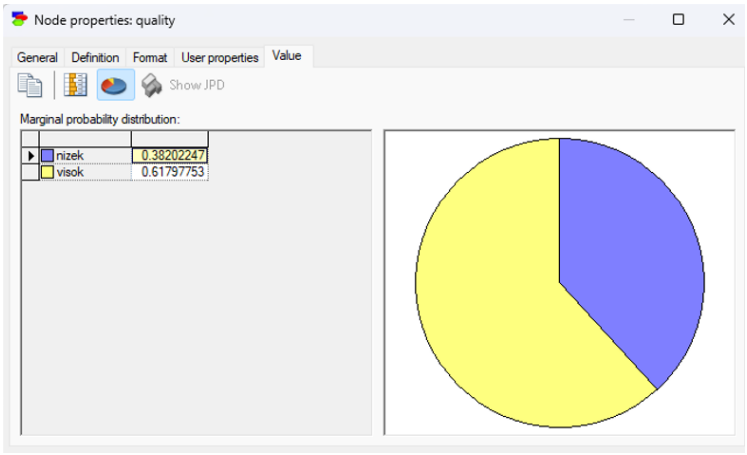
4. RESCORE PODATKOV

V tem koraku smo preverili napoved modela za specifičen primer, kjer smo **vse vhodne spremenljivke** nastavili na vrednost **nizek** (Set Evidence).

Na sliki 4 je prikazana marginalna porazdelitev ciljne spremenljivke quality za ta scenarij.

Opazimo, da je kljub nizkim vrednostim vseh vhodov verjetnost za visoko kakovost vina še vedno kar 61,8 %, medtem ko je verjetnost za nizko kakovost 38,2 %.

To kaže, da ima model sposobnost razločevanja tudi pri manj ugodnih vhodnih pogojih in da verjetnost ni porazdeljena enakomerno, temveč temelji na naučenih povezavah iz podatkov.

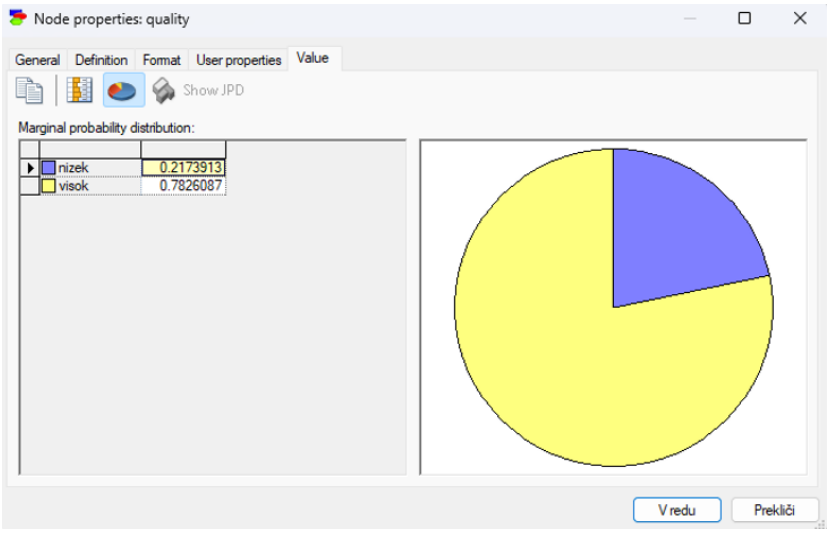


Slika 4: Rezultat - Vse spremenljivke "Nizek"

V nadaljevanju smo izvedli še napoved za scenarij, kjer so **vse vhodne spremenljivke nastavljene na vrednost visok.** Model nam v tem primeru vrne verjetnost za visoko kakovost vina kar 78,3 %, medtem ko je verjetnost za nizko kakovost le še 21,7 % (slika 5)..

To potrjuje, da model uspešno upošteva pozitivne vplive vhodnih spremenljivk in pravilno poveča verjetnost visoke kakovosti, ko so pogoji ugodni.

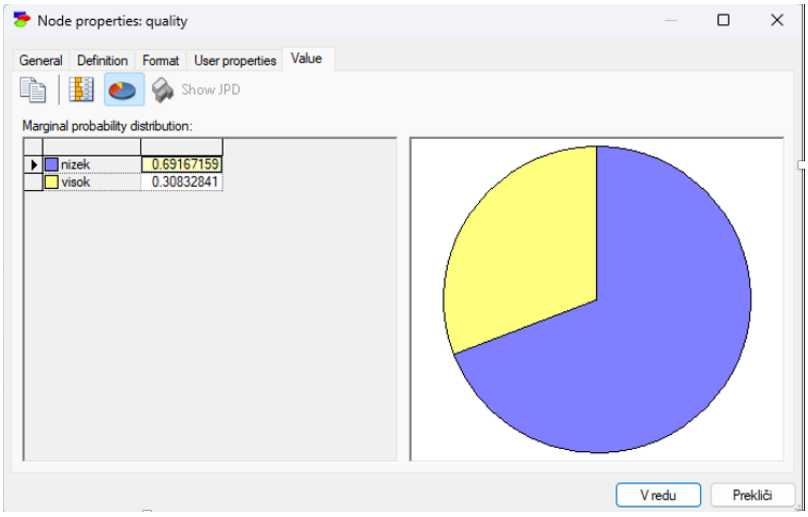
Tak rezultat je v skladu s pričakovanji – višje vrednosti alkohola, gostote in nižje kisline so namreč značilne za kakovostnejša vina.



Slika 5: Rezultat - Vse spremenljivke "Visoka"

V naslednjem koraku smo preverili scenarij, kjer **so vse vhodne spremenljivke nastavljene na visok, razen alcohol_bin, ki je nizek**. Rezultat napovedi je pokazal, da je v tem primeru verjetnost za nizko kakovost vina kar 69,8 %, kar pomeni, da model alkohol dojema kot ključno spremenljivko za končni izid (slika 6).

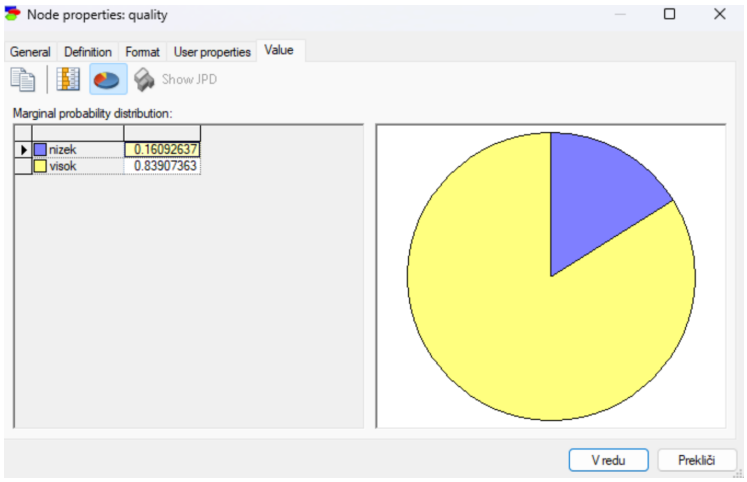
To je v skladu z opazovanjem iz strukture modela in tornado diagrama: nizka vsebnost alkohola ima največji negativen vpliv na kakovost vina – tudi če so ostali pogoji optimalni.



Slika 6: Rezultat - Vse spremenljivke "Visoka", razen alkohol "Nizka"

Za zaključek smo preverili še obratno situacijo – **vse spremenljivke so nastavljene na nizek, razen alcohol_bin, ki je nastavljen na visok**. V tem primeru model napove kar 83,9 % verjetnost za visoko kakovost vina (slika 7).

To dodatno potrjuje, da ima alkohol daleč največji pozitiven vpliv med vsemi vhodnimi spremenljivkami. Kljub neugodnim pogojem drugih spremenljivk model sklepa, da visoka vsebnost alkohola močno zviša verjetnost za kakovostno vino.

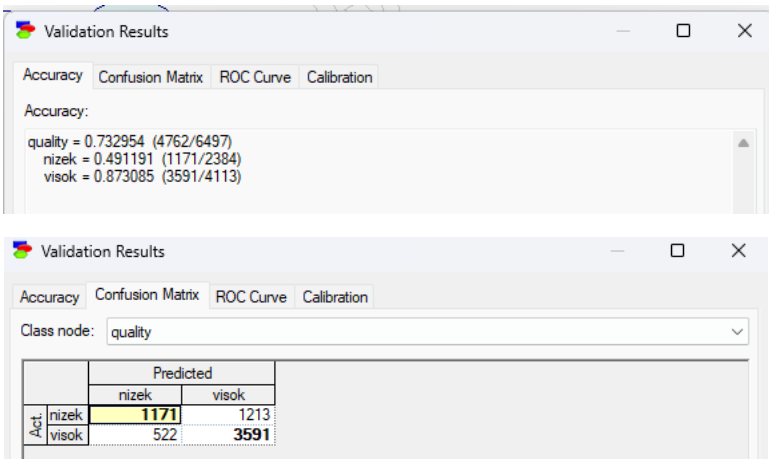


Slika 7: Vse spremenljivke "Nizka", razen alkohol "Visoka"

5. ROC KRIVULJA

Za oceno napovedne sposobnosti modela smo izvedli K-fold navzkrižno validacijo s 5 deli (5-fold crossvalidation), kjer smo za razredno spremenljivko izbrali quality.

Model je dosegel skupno natančnost (accuracy) 73,3 %, pri čemer je bila natančnost za razred visok kar 87 %, medtem ko je bila za razred nizek nekoliko slabša (49 %). To nakazuje, da model bolje prepoznava primere visoke kakovosti, kar potrjuje tudi matrika zmede.

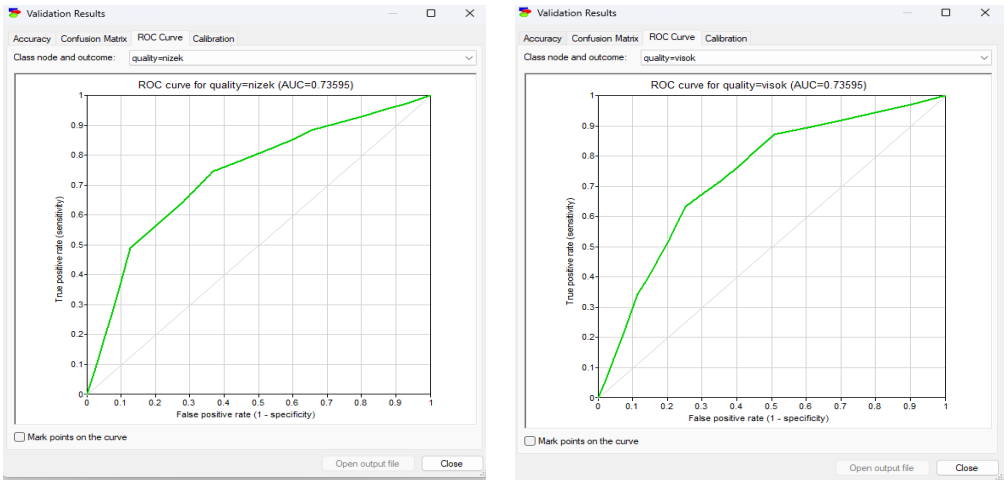


Slika 8: Natančnost in Matrika zmede

Za dodatno oceno ločilne sposobnosti smo analizirali ROC krivuljo za oba razreda. Model je dosegel:

- **AUC = 0.736** za razred nizek
- **AUC = 0.736** za razred visok

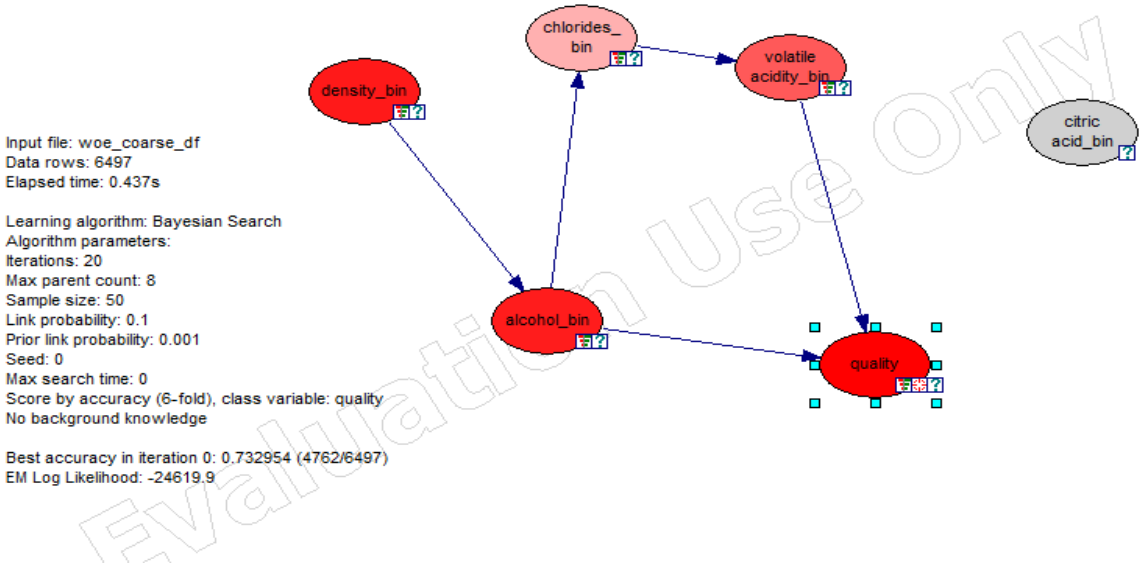
Obe krivulji kažeta jasno sposobnost modela, da razlikuje med razredoma bolje kot naključje (AUC > 0.7), kar pomeni, da je model dobro napovedno naravnan.



Slika 9: ROC krivulja

6. SENSITIVNA ANALIZA

Za analizo vpliva vhodnih spremenljivk na napoved cilja quality smo izvedli sensitivity analysis. Ta analiza omogoča vpogled v to, katera vhodna vozlišča imajo največji vpliv na verjetnost izida ciljne spremenljivke.



Slika 10: Sensitivna analiza

Na zgornji sliki so barvno označene spremenljivke glede na njihov vpliv:

- Temno rdeče barve označujejo močan vpliv (npr. alcohol_bin, volatile_acidity_bin, density_bin)
- Siva barva (citric_acid_bin) označuje, da spremenljivka ni povezana s ciljem in nima vpliva

Iz slike lahko razberemo, da ima največji vpliv spremenljivka alcohol_bin, ki posredno in neposredno vpliva na kakovost, kar smo že zaznali pri analizi modela in napovedih. Poleg alkohola pomembno vplivata še volatile_acidity_bin in density_bin, saj vplivajo na vmesne spremenljivke in posledično na kakovost.

Slika 11: Tornado Diagram - Quality – Nizek

Za podrobnejši vpogled smo uporabili tudi Tornado Diagram, ki prikazuje, kako občutljiv je izid **quality** = nizek glede na spremembe posameznih vhodnih vrednosti.

Na grafu so prikazane kombinacije spremenljivk, ki imajo največji vpliv na napovedano verjetnost:

- Prva vrstica prikazuje, da kombinacija **alcohol_bin = nizek** in **volatile_acidity_bin = visok** močno poveča verjetnost za nizko kakovost (dolga rdeča vrstica).
- Nasprotno kombinacija **alcohol_bin = visok** in **volatile_acidity_bin = nizek** zmanjša verjetnost za nizek izid (dolga zelena vrstica).
- Tudi posamezne vrednosti **density_bin** in **chlorides_bin** imajo opazen vpliv, vendar bistveno manjši kot alkohol.

Tornado diagram torej potrdi ključni vpliv alkohola in razkrije tudi interakcije med spremenljivkami, ki imajo največji skupni vpliv na napoved cilja.

Slika 12: Tornado Diagram - Quality - Visok

Tornado diagram za razred **quality = visok kaže,** kako se napovedana verjetnost spremeni ob spremembah vhodnih spremenljivk. Največji pozitiven vpliv ima kombinacija alcohol_bin = visok in volatile_acidity_bin = nizek, ki zviša verjetnost za visoko kakovost na skoraj 85 %. Nasprotno kombinacija alcohol_bin = nizek in volatile_acidity_bin = visok zmanjša verjetnost na okoli 61 %.

To dodatno potrjuje, da je alkohol najmočnejši napovednik kakovosti in da ima pomembno vlogo tudi volatile_acidity_bin. Kombinacija teh dveh močno vpliva na rezultat modela, medtem ko imajo drugi atributi, kot sta density_bin in chlorides_bin, manjši vpliv.

7. ZAKLJUČEK

Analizo smo začeli z uvozom podatkov in njihovo pripravo v okolju Python, kjer smo s pomočjo NumPy in Pandas izvedli čiščenje podatkov. Odstranili smo manjkajoče vrednosti, preverili porazdelitve, izračunali WoE, IV in vse vhodne spremenljivke diskretizirali v razrede nizek in visok (binarna pretvorba), s čimer smo pripravili podatke na format, združljiv s programom GeNIe.

V okolju GeNIe smo nato zgradili Bayesov model na podlagi strukture in parametrov, naučenih iz podatkov. Za preprečevanje prenaučenosti smo uporabili K-fold crossvalidation (5-fold). Model je dosegel skupno natančnost 73,3 % in AUC = 0.736, kar pomeni, da ima dobro ločilno sposobnost pri napovedovanju kakovosti vina

S pomočjo rescore analize smo testirali vpliv različnih kombinacij vhodnih spremenljivk. Rezultati so pokazali, da je alkohol najpomembnejši dejavnik, saj lahko sam močno poveča ali zmanjša verjetnost kakovostnega vina – tudi če so vse druge spremenljivke neugodne. Nazadnje smo izvedli še sensitivity analysis, ki je z barvnim prikazom in tornado diagrami potrdila, da ima največji vpliv na cilj prav alkohol (sledi mu volatilna kislina), medtem ko citronska kislina nima vpliva.

Celoten postopek je pokazal, da je Bayesova mreža učinkovito in razložljivo orodje za napovedovanje kvalitativnih ciljev na podlagi več vhodnih dejavnikov, še posebej tam, kjer so pomembne odvisnosti med spremenljivkami.

VIRI

1. *GeNIe Modeler – BayesFusion*. (b. d.). Pridobljeno 2. april 2025, s https://www.bayesfusion.com/genie/

2. Paulo Cortez, A. C. (2009). *Wine Quality* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T