HOMEMADE SOMMELIER

Università degli Studi di Milano - Bicocca
Machine Learning 2022 / 2023

Papa Emanuele - 844888
Furini Andrea - 845118

INTRODUZIONE

Negli ultimi anni, il mercato del vino sta vedendo una fascia sempre più ampia di consumatori e, per sostenerne la continua crescita, l'industria del vino sta impegnando nuove tecnologie nei processi di vinificazione e di vendita.

In passato, sono state effettuate molte ricerche sulla qualità del vino basate principalmente su studi empirici.
La qualità del vino non è facile da definire e ci sono molti fattori che ne influenzano quella percepita. Questi fattori includono caratteristiche intrinseche (visive, gustative, olfattive), ambientali (clima, regione, sito) e ingredienti fisico-chimici derivanti dalle pratiche viticole (acido, pH, solfati e solfuri).

I focus dell'innovazione del mercato riguardano la certificazione del vino e la valutazione della sua qualità. La certificazione previene l'adulterazione illegale dei vini, per salvaguardare la salute dell'uomo, e garantisce la qualità per il mercato del vino. La valutazione della qualità è parte del processo di certificazione e può essere utilizzata per migliorare la vinificazione individuando i fattori più influenti e per stabilire i prezzi.

La certificazione del vino è generalmente valutata mediante test fisico-chimici e test sensoriali.
I test di laboratorio fisico-chimici utilizzati solitamente per caratterizzare il vino includono la determinazione dei valori di densità, alcol o pH, mentre i test sensoriali si basano su individui esperti.

Essendo il gusto uno dei sensi umani meno sviluppati, la classificazione del vino risulta un compito arduo. Inoltre, la relazione tra l'analisi fisico-chimica e quella sensoriale rimane complessa e ancora non del tutto compresa.

OBIETTIVI RICERCA

Lo scopo di questo progetto sarà quello di analizzare e determinare, tramite un approccio di apprendimento automatico, la qualità di un determinato vino basandosi sulla sua composizione.

Inoltre, ai fini dello studio, verrà svolta un'ulteriore ricerca per determinare la tipologia del vino dai suoi elementi fisico-chimici.

Entrambi gli obiettivi prevederanno una classificazione binaria dei valori da prevedere.

ANALISI ESPLORATIVA DATASET

I dati presi in considerazione rappresentano le informazioni relative alle composizioni delle varianti di bianco e di rosso del vino portoghese *Vinho Verde* raccolte tra il Maggio 2004 e il Febbraio 2007.

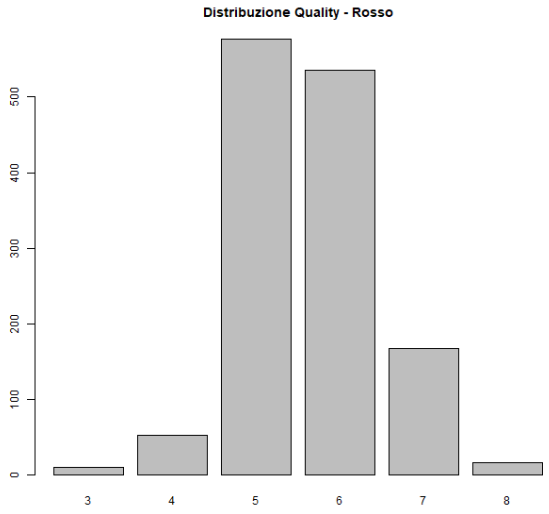
In particolare, ci si concentra sugli elementi fisico-chimici utilizzati e al grado di qualità espresso in un valore compreso tra 0 (bassissima qualità) e 10 (altissima qualità) che rappresenta indirettamente il nostro attributo di target.

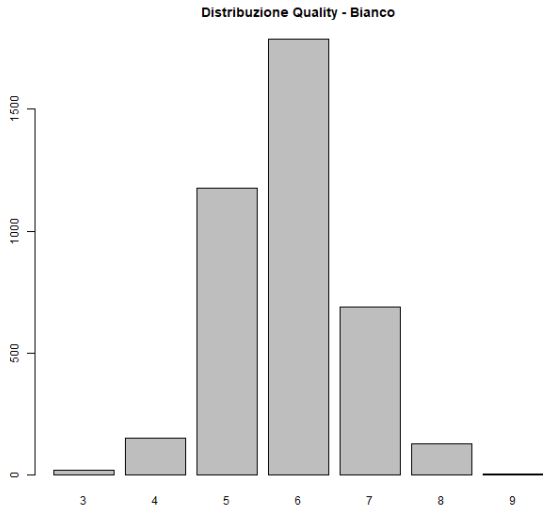
I dati studiati sono stati raccolti da due dataset differenti rappresentanti una variante di vino ciascuno e uniti insieme in uno solo. Le specifiche versioni dei dataset utilizzati sono disponibili al seguente link.

Nel corso di questo capitolo verranno analizzati il dataset, con le varie modifiche apportate, e i risultati delle analisi esplorative su di questo.

STRUTTURA DEL DATASET

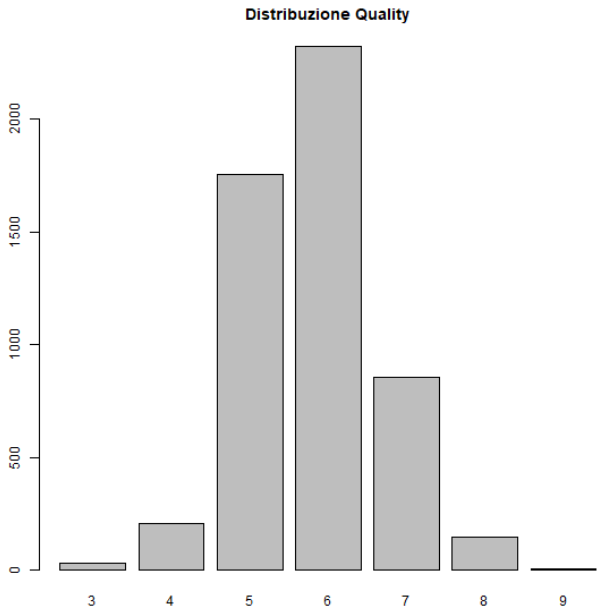
Come anticipato nella parte precedente, il set di dati su cui è stata effettuata la ricerca è composto da due dataset distinti: *WineQT-Red* (1599 istanze) e *WineQT-White* (4898 istanze).





Ponendo una maggiore attenzione alla sistemazione dei dati in funzione dell'attributo rappresentante la qualità del vino, si può notare come la maggioranza dei valori si disponga attorno all'elemento centrale della scala di valutazione, con una piccolissima quantità di valori estremi.

Successivamente, unendo i due dataset, rimuovendo le istanze duplicate e quelle contenenti valori nulli, si ottiene un unico set di dati formato da 5320 istanze.



Siccome tra i due dataset figli vi è una sostanziale differenza di numero di record, nel dataset composto si può vedere una situazione simile a quella riportata per la variante di vino bianco. Si nota dunque una distribuzione di valori concentrati nella metà della scala di valutazione della qualità con una bassa presenza di valori esterni.

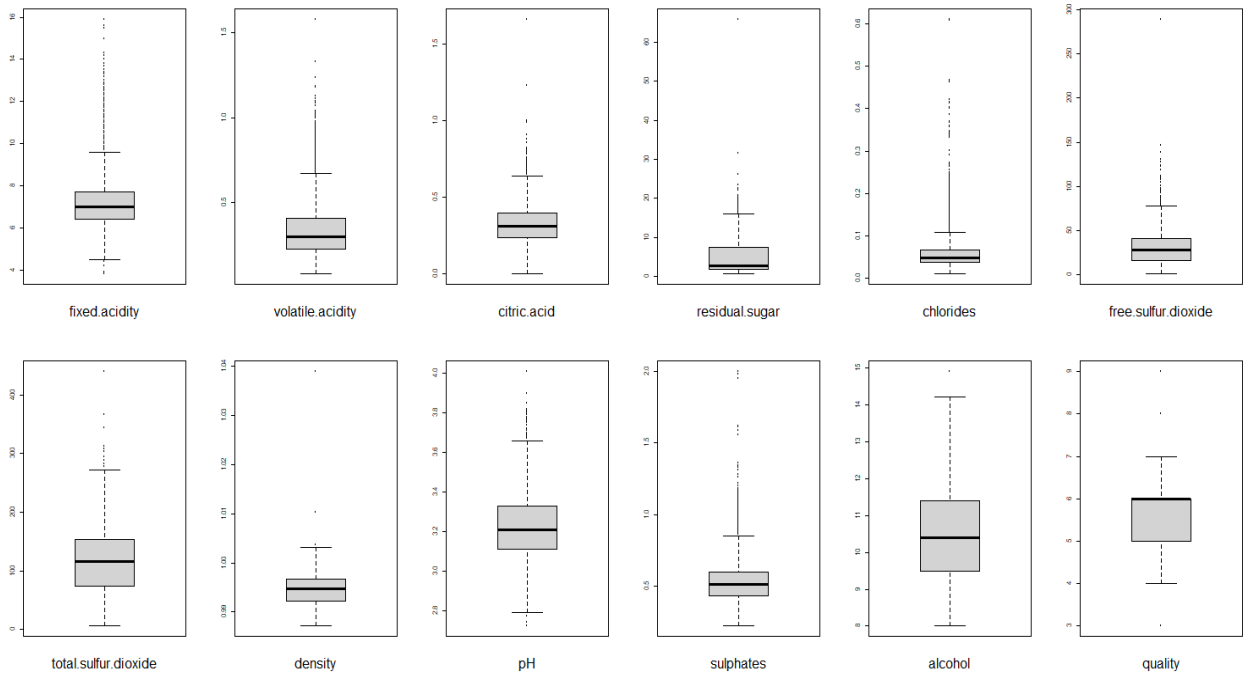
Poiché la ricerca si baserà su uno studio di classificazione binaria, l'informazione più rilevante che ricaviamo da questo grafico riguarda la simmetria della sistemazione dei dati. Difatti, i dati tendono a disporsi nelle varie classi con una distribuzione simile a quella di una normale.

ATTRIBUTI

Il dataset preso in considerazione per lo studio è composto da 12 attributi tutti rappresentati in valori numerici di tipo reale, eccetto per il grado di qualità del vino che viene espresso in forma di numero intero.

Nello specifico, gli attributi contenuti all'interno del dataset sono:

- Fixed Acidity
- Volatile Acidity
- Citric Acid
- Residual Sugar
- Chlorides
- Free Sulfur Dioxide
- Total Sulfur Dioxide
- Density
- PH
- Sulphates
- Alcohol
- Quality



Come si evince dai boxplot di ciascun attributo, si può notare come la maggior parte dei valori siano equamente raggruppati intorno alla loro mediana. Inoltre, per alcuni attributi come **Fixed Acidity**, **Volatile Acidity**, **Chlorides** e **Sulphates** si possono vedere come i dati superiori al terzo quartile e i valori outliers superiori tendono a distaccarsi molto tra di loro. Nonostante questo comportamento anomalo, non sono state effettuate modifiche agli attributi sotto questo aspetto in quanto, nel complessivo, i valori ottenuti non causano troppo rumore nelle analisi e dunque nello studio che ne seguirà.

MODIFICA FEATURES

Per poter continuare la nostra ricerca, risulta doveroso fare degli aggiustamenti agli attributi di target del dataset.

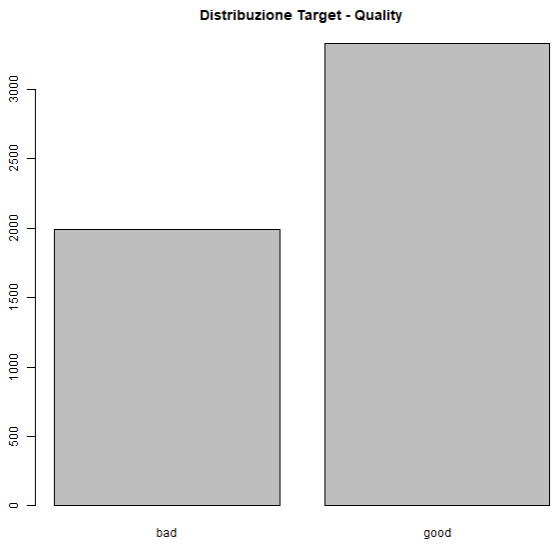
Innanzitutto, dato che lo studio prevederà un problema di classificazione binaria, è stato deciso di modificare la colonna **Quality** assegnando i valori riportati dalla feature in due classi distinte.
I dati vengono suddivisi in "bad" (cattivo) e in "good" (buono) seguendo la logica della votazione scolastica: un vino viene considerato "cattivo" se la sua qualità assume un valore minore o uguale a 5; viceversa viene considerato "buono" se il valore è uguale a 6 o superiore.
A seguito di questa rielaborazione, la colonna **Quality** è stata dunque sostituita con la colonna custom **Target.Quality**.

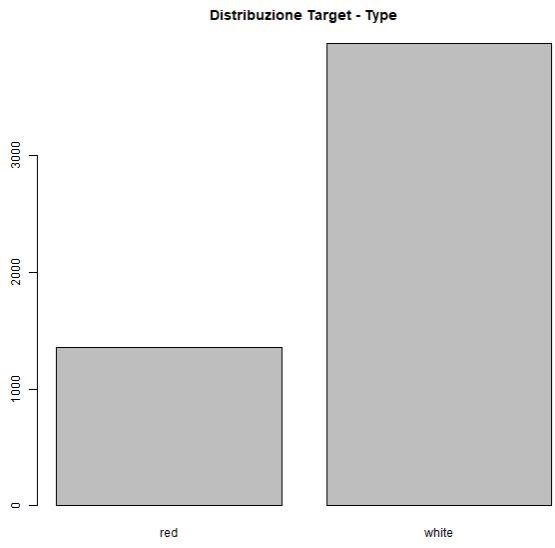
```
target.quality <- c(wine$quality)
target.quality[target.quality > 5] <- 'good'
target.quality[target.quality < 6] <- 'bad'
wine['target.quality'] <- target.quality
wine$target.quality <- as.factor(wine$target.quality)
```

Un'altra modifica effettuata sulle features del dataset è stata quella di aggiungere una nuova colonna custom **Target.Type** riguardante la variante di vino.
Questa operazione è stata svolta assegnando, ad ogni istanza, il dataset di provenienza, prima di svolgere l'unione tra i due insiemi di dati iniziali.

```
wine.red['target.type'] <- 'red'
wine.red$target.type <- as.factor(wine.red$target.type)
wine.white['target.type'] <- 'white'
wine.white$target.type <- as.factor(wine.white$target.type)
wine <- merge(wine.red, wine.white, all = TRUE)
```

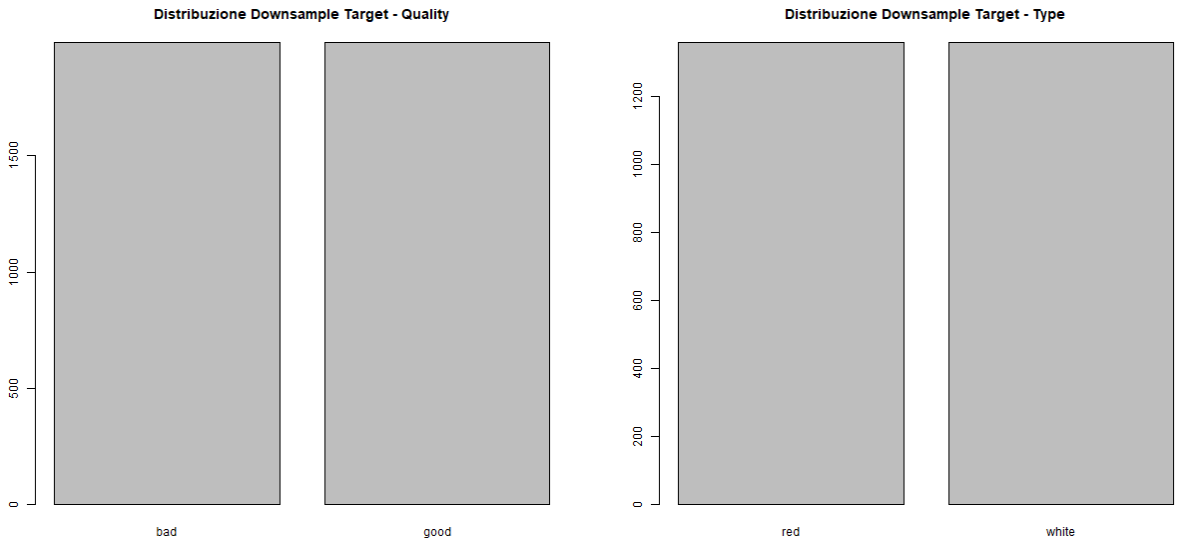
Concluse le modifiche alle colonne di target, si nota uno scenario a grandi linee aspettato.
Difatti, per quanto riguarda l'attributo **Target.Quality**, dato che la suddivisione delle qualità dei vini seguiva l'andamento di una curva Gaussiana, non vi è una sostanziale differenza tra le distribuzioni delle due classi ottenute.
Invece, poiché tra i due dataset minori selezionati per lo studio vi è una grande disparità di grandezza, per la colonna **Target.Type**, possiamo notare come vi siano più del triplo delle istanze che riportano il valore "bianco" rispetto al valore "rosso".





DOWNSAMPLING DEI DATASET

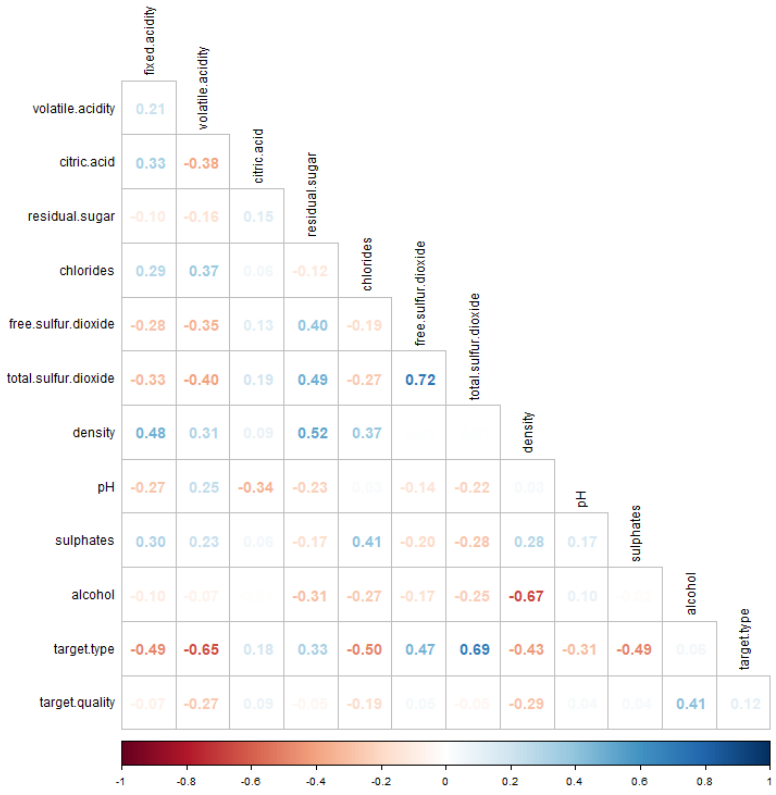
Per poter ovviare al problema riguardante lo sbilanciamento dei valori di target, è stata svolta un'operazione di ridimensionamento del numero di istanze in maniera casuale, in riferimento alle colonne **Target.Quality** e **Target.Type**.



Siccome per la ricerca sono stati posti due obiettivi distinti tra loro e non esiste un modo semplice per ottenere un unico dataset che abbia la stessa distribuzione su entrambi i target, per svolgere l'operazione di downsampling sono stati creati due dataset su cui verranno effettuati gli studi. Si avranno infatti i dataset ridotti **wine.quality** e **wine.type** sui quali verranno rispettivamente svolte le ricerche per prevedere la qualità e la variante del vino.

MATRICE DI CORRELAZIONE

La matrice di correlazione è una tabella quadrata che mostra i coefficienti di correlazione tra le varie coppie di features. Si tratta di un prospetto che permette di valutare, nell'insieme, il grado di interdipendenza di una serie di grandezze.



Dalla matrice di correlazione ottenuta dal dataset principale, si può notare come sia presente un numero ristretto di casi con forti correlazioni (positive e negative) tra i vari attributi. Nello specifico, alcune features, come **Residual.Sugar** e **Citric.Acid,** non presentano relazioni particolari con le altre caratteristiche.

Per quanto riguarda le colonne di target, l'attributo **Target.Type** riporta molte relazioni alternate tra loro; mentre **Target.Quality** sembra essere influenzata solamente dalla feature **Alcohol**.

In conclusione, dati i coefficienti ottenuti in funzione degli attributi di target, si può stimare come sarà l'andamento delle previsioni delle varie classi. Ci si aspetterà, infatti, che i modelli selezionati faranno molta più fatica a classificare i valori della feature **Target.Quality** rispetto a quelli di **Target.Type**.

PRINCIPAL COMPONENT ANALYSIS

Considerando l'alto numero di features del dataset, è necessario l'utilizzo di uno strumento in grado di applicare una tecnica di riduzione delle dimensionalità, in modo tale da ridurre il costo computazionale dell'addestramento dei modelli classificatori.

La Principal Component Analysis (PCA) si basa sul concetto che, all'interno dell'insieme di dati, vi siano uno o più attributi che forniscano informazioni già apprese da altri.

La PCA si pone quindi l'obiettivo di individuare le features più significative, ossia in grado di descrivere al meglio il dataset in questione, senza condizionare particolarmente il risultato finale dell'analisi e delle predizioni. Questo procedimento avviene identificando i pattern significativi nei dati ed esprimendoli per evidenziare le loro somiglianze e differenze.

Per la ricerca, l'analisi delle PCA effettuata verrà svolta sul dataset principale, al quale sono stati rimossi gli attributi di target **Target.Quality** e **Target.Type**.

CALCOLO PCA

Per la selezione delle feature principali, bisogna prima di tutto analizzare gli autovalori ottenuti dal calcolo della PCA rispettando determinati criteri:

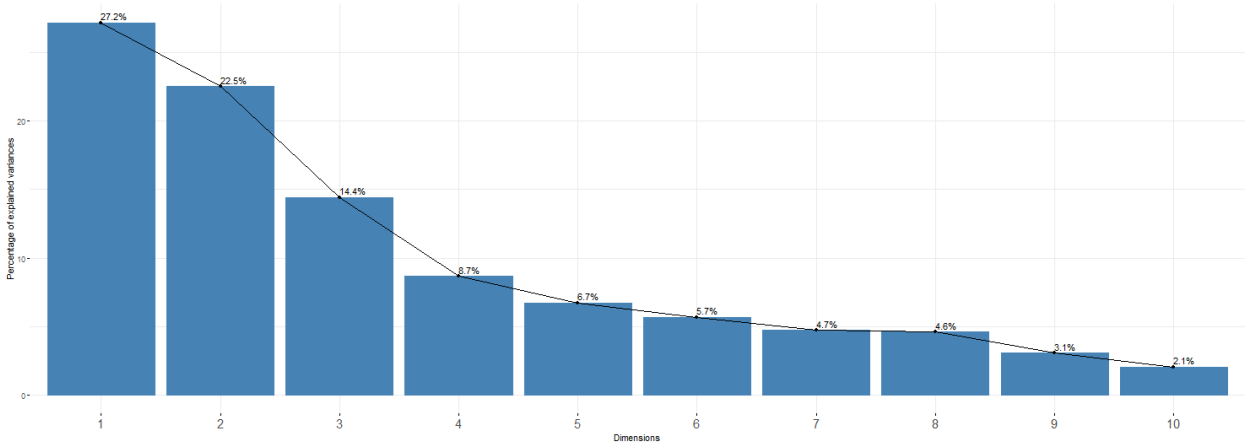
1. **Autovalore maggiore di 1**: scegliere i componenti con autovalore superiore a 1.
2. **Varianza totale spiegata superiore del 70%**: limitare il numero delle componenti in modo tale che rappresentino una certa frazione della varianza totale.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.98837552	27.1670502	27.16705
Dim.2	2.47526679	22.5024253	49.66948
Dim.3	1.58728240	14.4298400	64.09932
Dim.4	0.95340157	8.6672870	72.76660
Dim.5	0.74223529	6.7475936	79.51420
Dim.6	0.62703796	5.7003452	65.21454
Dim.7	0.52103998	4.7367271	89.95127
Dim.8	0.50759857	4.6145325	94.56580
Dim.9	0.33672096	3.0610997	97.62690
Dim.10	0.22595817	2.0541652	99.68107
Dim.11	0.03508278	0.3189344	100.00000

Seguendo i criteri riportati precedentemente, si può notare come solamente 3 features riportino un autovalore superiore a 1. Mentre, 4 attributi risultano avere almeno il 70% della varianza spiegata.

Qualora volessimo accertarci dei risultati ottenuti, un metodo alternativo per determinare il numero di componenti principali è osservare i valori ottenuti dallo scree plot.
Uno scree plot mostra sempre gli autovalori in una curva discendente, ordinando gli autovalori dal più grande al più piccolo. Secondo lo scree test, i valori rappresentati alla sinistra del "gomito" del grafico in cui gli autovalori sembrano stabilizzarsi dovrebbero essere ritenuti i più significativi.

Il risultato ottenuto dall'analisi dello scree plot ci permette di rafforzare le informazioni ricavate dall'analisi precedente. Difatti, dal grafico si evince che la scelta delle features ricadrebbe sulle prime quattro rappresentate.

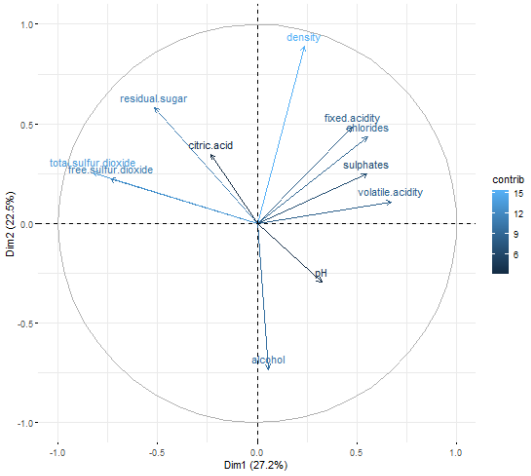


Incrociando dunque i dati estrapolati dalle analisi, si è optato per selezionare le prime quattro dimensioni.

SCELTA ATTRIBUTI

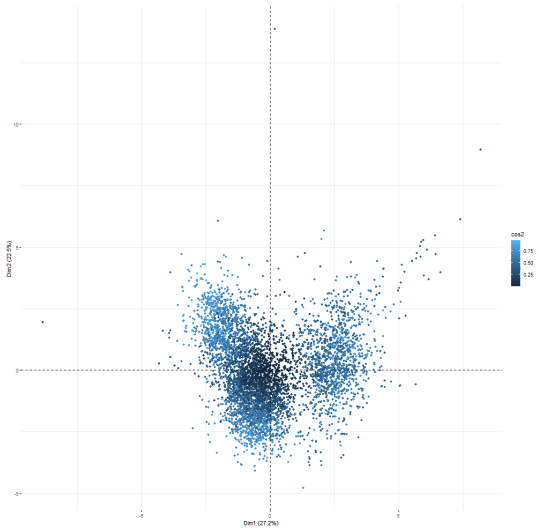
La selezione degli attributi viene effettuata analizzando i dati ottenuti dal grafico delle variabili.

All'interno del grafico, l'angolo riportato tra i vettori mostra la correlazione tra le variabili: un piccolo angolo denota che le variabili sono correlate positivamente; un grande angolo indica, invece, che sono correlate negativamente. Inoltre, la lunghezza del vettore rivela quanto la feature risulti essere significativa.



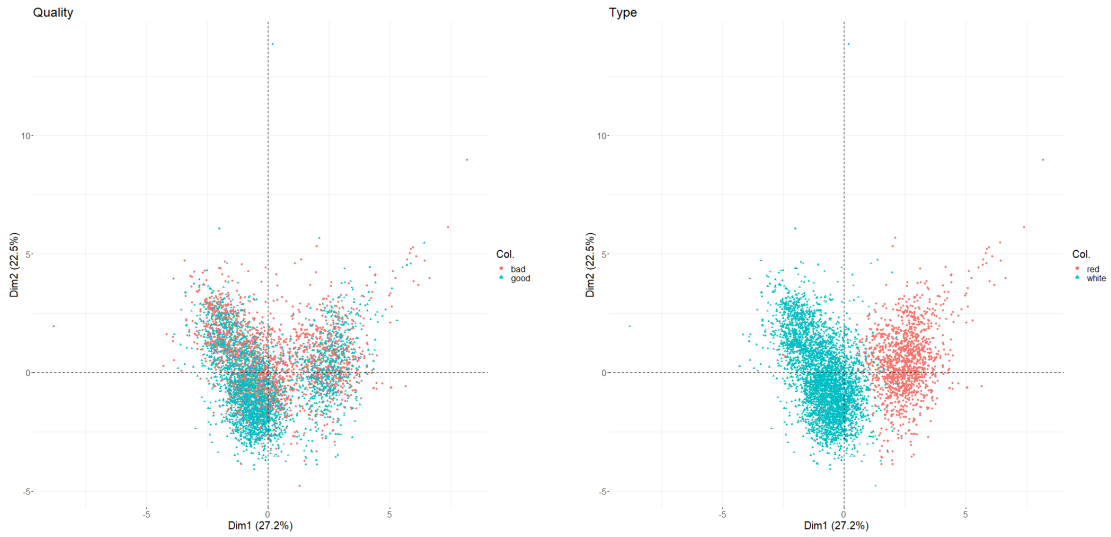
Insieme a quest'ultimo grafico, è molto spesso ragionevole associare lo studio degli individui del dataset. In questo modo è possibile vedere come le istanze presenti nel set di dati vengano descritte meglio in funzione dei vari attributi.

I valori riportati dentro lo scatter plot ci permettono di comprendere come vi sia una grossa ridondanza dei dati non perfettamente rappresentati dalle componenti. Nonostante ciò, i dati tendono comunque a disporsi nell'interezza dello spazio.

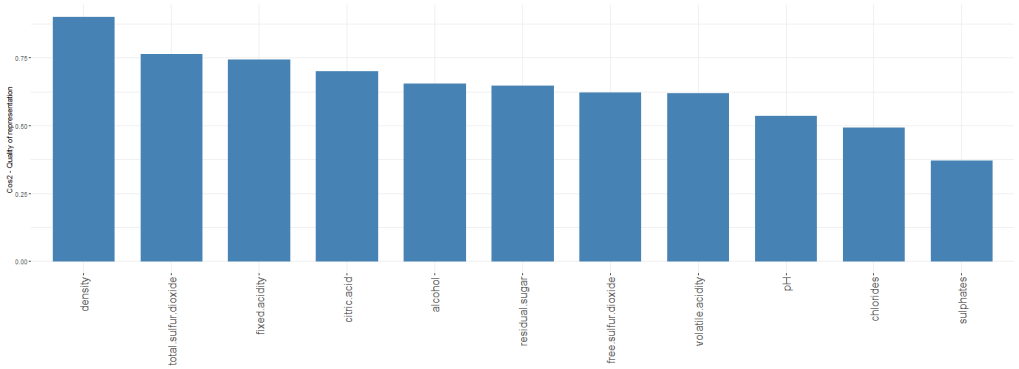


Lo scatter plot risulta utile, inoltre, per comprendere meglio i *clusters* riguardanti le distribuzioni dei vari individui dato un determinato attributo.

Dai grafici ottenuti in funzione delle features di target, si possono notare due scenari fortemente diversi tra loro. Difatti, per l'attributo **Target.Quality** viene mostrata una disposizione dei valori alquanto caotica, mentre, per **Target.Type** si riescono a distinguere chiaramente i due clusters.



Un ulteriore metodo efficiente per determinare quali attributi selezionare è tramite l'utilizzo del grafico del Cos2. Nel grafico delle qualità, maggiore è il valore del Cos2 e più la variabile risulta essere rilevante.



Incrociando le informazioni ottenute dal grafico delle variabili, dal grafico del Cos2 e dalle analisi esplorative precedenti, per la ricerca sono stati selezionati gli attributi **Density**, **Total.Sulfur.Dioxide**, **Fixed.Acidity** e **Alcohol**.

Queste features verranno utilizzate per la fase di addestramento dei modelli per entrambi gli obiettivi stabiliti. In questo modo si avrà un confronto più valido tra i due studi effettuati.

MODELLI

Nel mondo del Machine Learning non esiste un modello universale per ciascun problema di classificazione e di modellazione dei dati.
La scelta su quale utilizzare è determinata inizialmente dalla natura del problema e successivamente su preferenza personale.

I modelli selezionati per lo studio sono quelli strutturati su Naive Bayes e su Neural Network della libreria "caret". In questo modo, il metro di giudizio per la comparazione dei risultati ottenuti avrà un ulteriore fattore comune tra i modelli utilizzati.

L'analisi prenderà in considerazione il subset di attributi, selezionati tramite PCA, al quale sono stati aggiunti gli attributi di target.

NAIVE BAYES

Naive Bayes è un algoritmo di classificazione che utilizza i dati degli eventi passati per provare a prevedere gli eventi futuri.
Questo approccio, come suggerisce il nome, è basato sul teorema di Bayes nella sua versione "naive", ovvero, sull'assunzione dell'indipendenza delle variabili. In altre parole, un classificatore Naive Bayes assume che la presenza di una specifica feature non sia correlata con la presenza di altre features.

Dopo l'analisi dei dati ottenuti dalla PCA, si è deciso di utilizzare questo modello poiché le variabili rappresentate graficamente riportano una buona indipendenza tra di loro, e risultano quindi adatte all'assunzione che questo tipo di modello si porta con sé.

NEURAL NETWORKS

Una rete neurale è un sistema computazionale che genera previsioni, composto da unità elementari chiamate neuroni. Nello specifico, una rete neurale è strutturata con un gruppo interconnesso di nodi, che coinvolge uno strato di input, uno di nodi nascosti e lo strato di output.
Ciascun nodo si connette ad un altro tramite un peso e una soglia ad esso associata. Se l'output di qualsiasi singolo nodo è al di sopra del valore di soglia specificato, tale nodo viene attivato, inviando i dati al livello successivo della rete. Una volta raggiunto lo strato di output, la rete neurale è in grado di poter dare la propria predizione sui valori analizzati.

Si è optato di utilizzare questo modello per provare a cogliere le particolarità e le raffinatezze della qualità dei vini e comprendere il comportamento della struttura con le tipologie di vino.

TRAINING E TESTING

Indipendentemente dal tipo di set di dati utilizzato, per addestrare qualsiasi modello di Machine Learning, è necessario suddividere casualmente il dataset nei set di *training* e di *test*.

Durante la fase di addestramento, al modello vengono fornite le osservazioni contenute all'interno del set di training in modo che questo possa stimare i possibili *outcome*. Mentre, il set di test viene utilizzato, nella fase successiva, per vedere il comportamento del modello su dati non ancora analizzati.

NORMALIZZAZIONE DEI VALORI

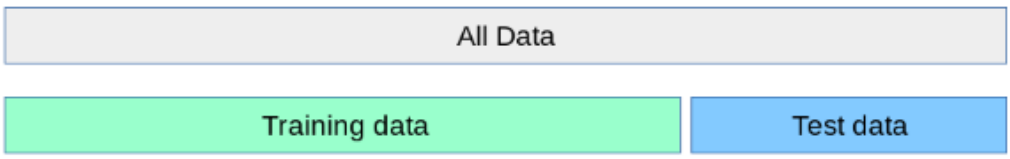
Prima di poter iniziare la fase di training e testing dei modelli, è stato svolto, su tutti i valori numerici del dataset, un processo di standardizzazione dei dati.
In questo modo, ciascuna feature presenterà un valore medio di 0 e una deviazione standard di 1.

Questo passaggio risulta fondamentale per evitare anomalie durante l'addestramento dei modelli utilizzati.

SPLITTING

Durante lo splitting delle osservazioni, i dati dovrebbero essere suddivisi in modo tale che vi sia un'elevata quantità di valori per l'addestramento del modello.

Per lo studio si è optato per ripartire il dataset con un rapporto 70%-30%.

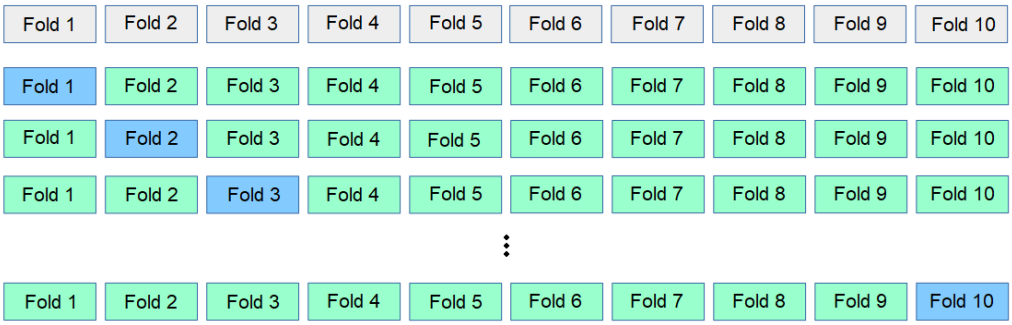


```
split.data = function(data, p = 0.7, s = 1) {
  set.seed(s)
  index = sample(1:dim(data)[1])
  train = data[index[1:floor(dim(data)[1] * p)],]
  test = data[index[((ceiling(dim(data)[1] * p)) + 1):dim(data)[1]],]
  return(list(train = train, test = test))
}
```

CROSS-VALIDATION

La procedura della *k-fold cross validation* consiste nella suddivisione del dataset totale in k-parti di uguale numerosità dove, ad ogni iterazione, la k-esima parte rappresenta il set di test e quella restante costituisce l'insieme di addestramento.

In questo modo, il modello viene addestrato per ognuna delle k-parti, evitando quindi i problemi di overfitting e di campionamento asimmetrico del campione osservato, tipici della suddivisione tramite splitting.



Inoltre, per garantire una migliore ridistribuzione dei valori e un bacino di modelli più numeroso su cui fare media, la procedura della cross-validation verrà ripetuta 3 volte.

Alla fine della fase di training, dunque, si avranno un totale di 30 modelli generati dai 3 gruppi di k=10 fold.

RISULTATI

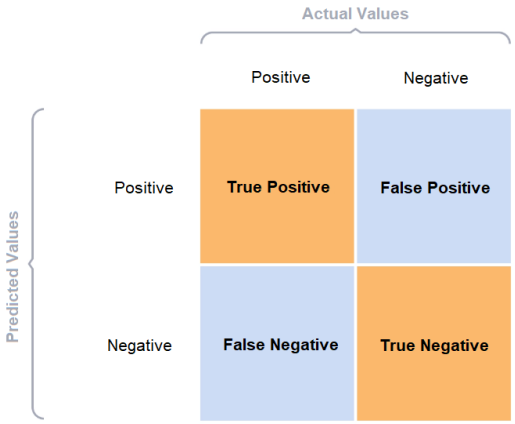
I dataset utilizzati per la ricerca, e dunque per l'analisi dei risultati, saranno quelli ottenuti dalla riduzione, in funzione dell'attributo target, e formati dalle 4 features selezionate, tramite PCA, unite agli attributi di target.
Difatti, data la differenza di numerosità tra le classi dei target, scegliere di eseguire le analisi su tutto il dataset avrebbe portato a risultati inconsistenti.

Dunque, una volta eseguita la fase di training e di testing, è possibile confrontare tra loro i modelli implementati, per stabilire quale sia il migliore dei due, e quindi, determinare quale meglio prevede la classificazione e la tipologia del vino.

METRICHE

Per valutare le differenti prestazioni dei modelli ottenute a seguito della fase di training e testing, verranno utilizzate una serie di metriche ricavabili tramite la matrice di confusione.

La matrice di confusione è la rappresentazione dell'accuratezza del classificatore. In particolare ogni colonna della matrice rappresenta i valori attuali delle osservazioni, mentre ogni riga rappresenta i valori predetti dal classificatore.



Possono verificarsi quattro casi:

- *True positive* (TP)
  Se la classe prevista è positiva ed è uguale alla classe effettiva, si tratta di un caso di true positive (vero positivo).
- *True negative* (TN)
  Se la classe prevista è negativa ed è uguale alla classe effettiva, si tratta di un caso di true negative (vero negativo).
- *False positive* (FP)
  Se la classe prevista è positiva ma è diversa dalla classe effettiva, si tratta di un caso di false positive (falso positivo).
- *False negative* (FN)
  Se la classe prevista è negativa ma è diversa dalla classe effettiva, si tratta di un caso di false negative (falso negativo).

Per quanto riguarda le matrici di confusione riferite ai dataset sviluppati, i valori ritenuti positivi per gli attributi di target saranno "good" e "white" rispettivamente per il dataset **Wine.Quality** e **Wine.Type**.

Dalla matrice di confusione è possibile ottenere:

- **Accuracy**
  Capacità di un modello di trovare tutti i casi veritieri, ovvero la percentuale di istanze classificate correttamente.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**
  Capacità di un modello di classificazione di identificare correttamente i valori positivi rispetto a tutte le osservazioni positive.

$$\frac{TP}{TP + FP}$$

- **Recall**
  Capacità di un modello di trovare tutti i casi pertinenti all'interno di una serie di dati, ovvero la capacità di identificare i valori positivi.

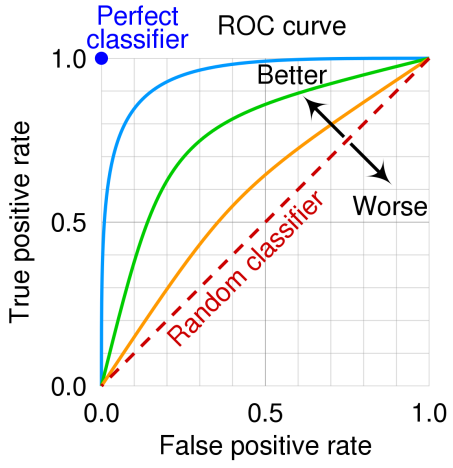
$$\frac{TP}{TP + FN}$$

- **F1-Score**
  Media armonica di precisione e recall, prese entrambe le metriche.

$$\frac{TP}{2TP + FP + FN}$$

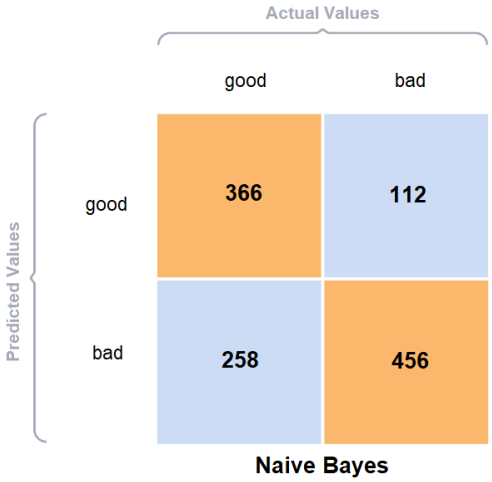
Oltre a queste 4 metriche, come metro di paragone, verranno utilizzate la curva ROC e la AUC.

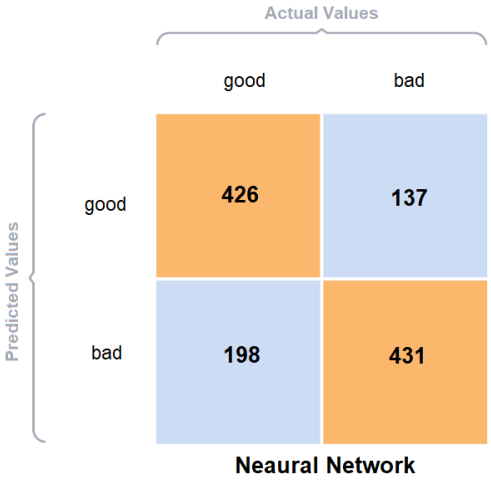
La Receiver Operating Characteristic (ROC) è un grafico che illustra le prestazioni di un sistema di classificazione binaria e traccia il tasso vero positivo contro il tasso di falsi positivi per diversi punti di taglio. Attraverso l'analisi di questa curva, e in particolare della sua Area Under the Curve (AUC), riusciamo ad avere una misura di bontà del modello.



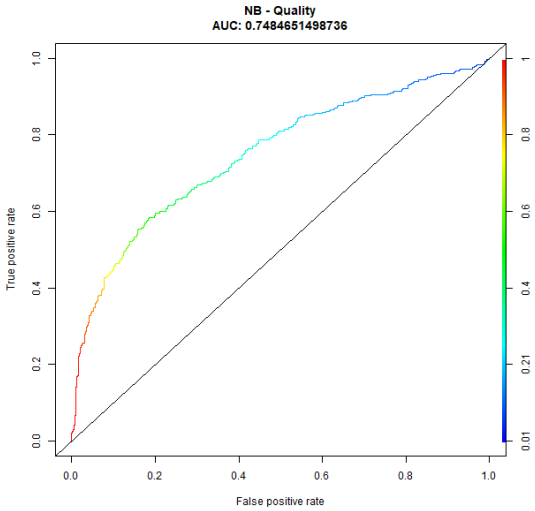
RISULTATI - SPLITTING

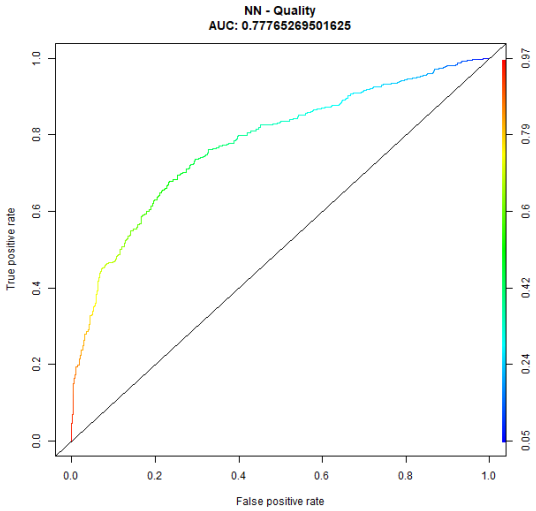
- Quality



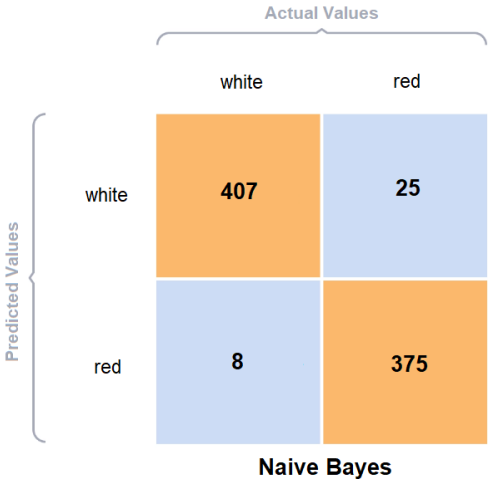


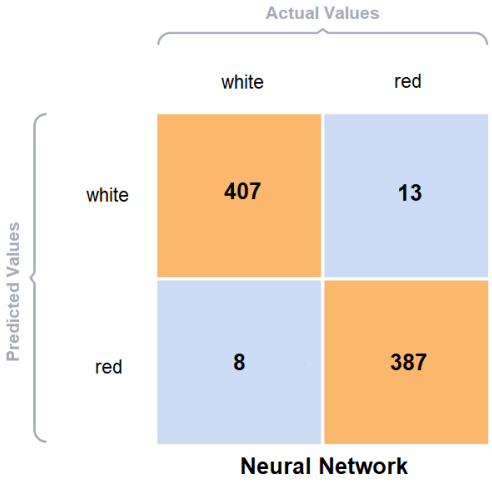
	Accuracy	Precision	Recall	F1-Score
NB	0.6896	0.7657	0.5865	0.6642
NN	0.7190	0.7567	0.6827	0.7178



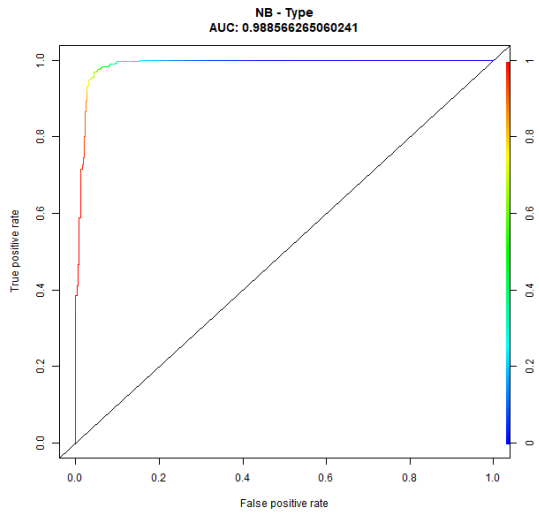


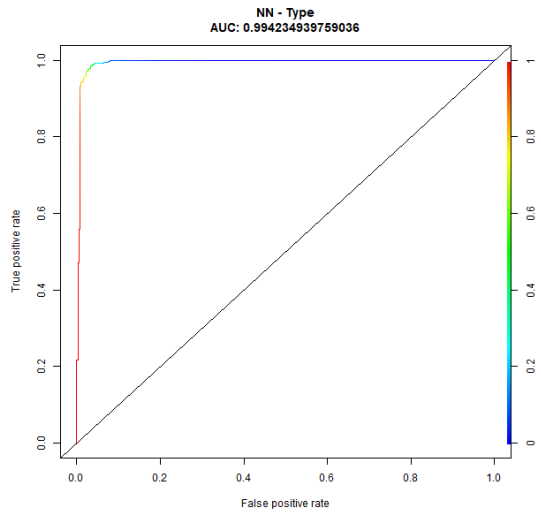
- Type





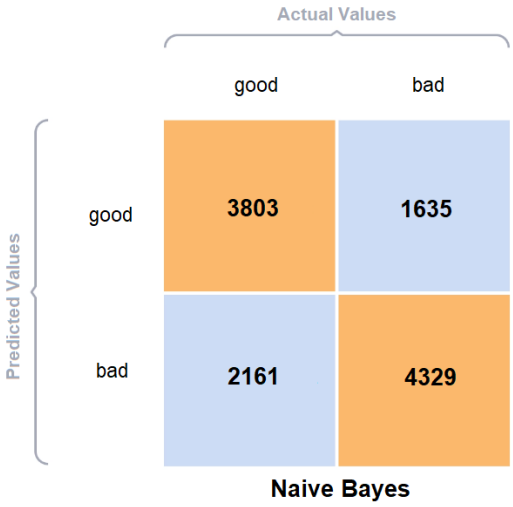
	Accuracy	Precision	Recall	F1-Score
NB	0.9595	0.9421	0.9807	0.9610
NN	0.9742	0.9690	0.9807	0.9749

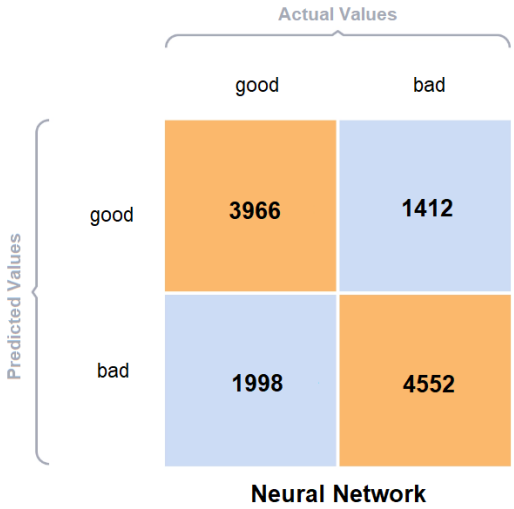




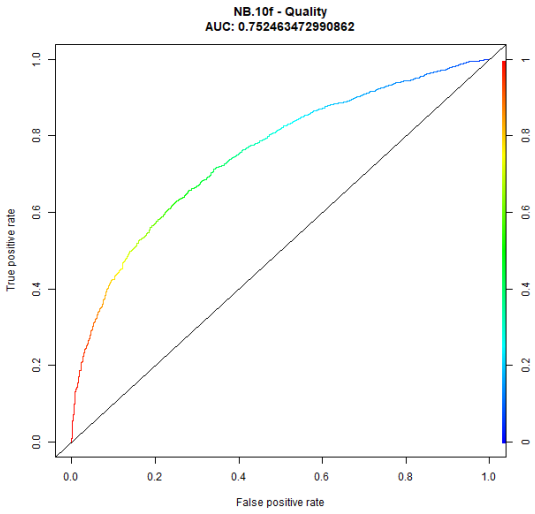
RISULTATI - CROSS VALIDATION

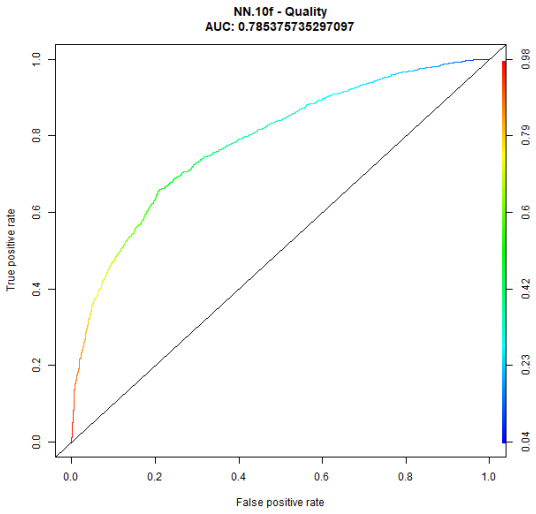
- Quality



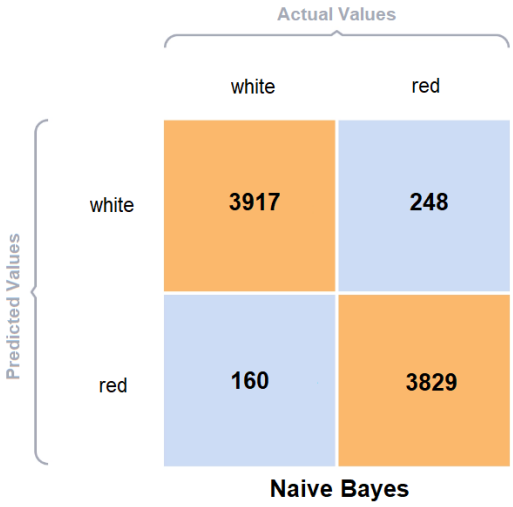


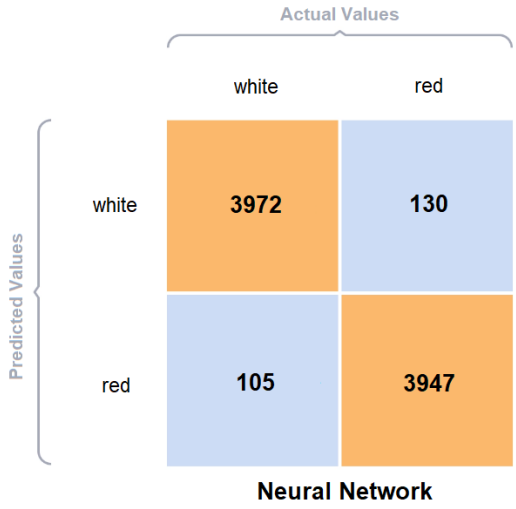
	Accuracy	Precision	Recall	F1-Score
NB	0.6818	0.6993	0.6377	0.6671
NN	0.7141	0.7374	0.6650	0.6993



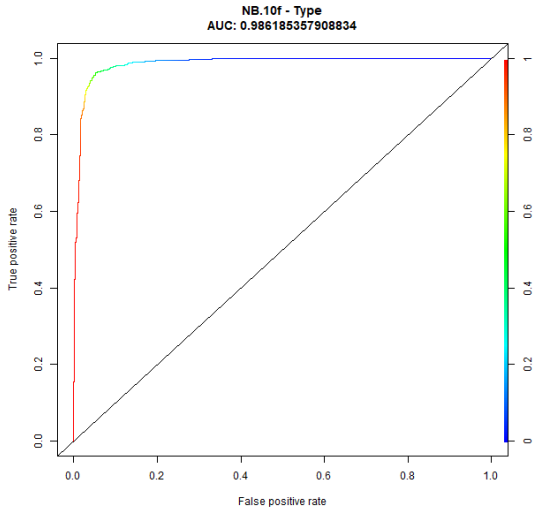


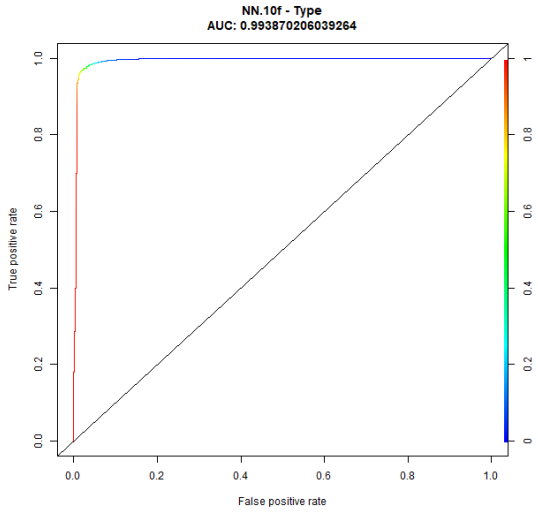
- Type





	Accuracy	Precision	Recall	F1-Score
NB	0.9500	0.9405	0.9608	0.9505
NN	0.9712	0.9683	0.9742	0.9713





CONCLUSIONI

Analizzando i risultati ottenuti dalla fase di testing, possiamo osservare che entrambi i modelli svolgono in maniera estremamente efficace il compito di prevedere la variante di vino. Non si può, però, dire lo stesso per quanto riguarda le previsioni della qualità.

Tuttavia, i valori ottenuti per le due ricerche rispecchiano le aspettative che ci eravamo posti sin dall'inizio. Difatti, la tipologia di vino risulta molto più deducibile dati i suoi elementi fisico-chimici rispetto alla sua qualità, la quale è principalmente condizionata da un parere umano e soggettivo.

Nel complessivo, per tutti gli studi effettuati, si può notare come entrambi i modelli riportino valori simili tra loro. In particolare, il modello riguardante la rete neurale restituisce esiti leggermente migliori in cambio, però, di un maggior tempo computazionale rispetto al modello bayesiano.

Inoltre, una particolare attenzione va posta sui risultati della cross-validation che non si allontanano troppo dai valori ottenuti dallo splitting.