WINE QUALITY CLASSIFIER

DOCUMENTAZIONE CASO DI STUDIO
INGEGNERIA DELLA CONOSCENZA 2024-2025

Progetto realizzato da:
-Alessandro Olivieri, matricola: 703269, a.olivieri14@studenti.uniba.it

GitHub: https://github.com/a1esm0ke/classiwiner.git

SOMMARIO

1) INTRODUZIONE

L'industria vinicola, rappresenta un campo fertile per l'applicazione di tecniche analitiche avanzate che possono contribuire a garantire e migliorare la qualità del vino. L'evoluzione tecnologica degli ultimi anni ha reso possibile l'impiego di modelli di machine learning per affrontare questa sfida, offrendo nuove prospettive per l'analisi predittiva in questo settore tradizionalmente legato a metodi di valutazione più soggettivi e artigianali.
In questo contesto, il presente studio esplora l'efficacia di vari modelli di classificazione, tutti radicati nell'apprendimento supervisionato, che promettono di rivoluzionare il modo in cui la qualità del vino viene valutata e classificata. Utilizzando approcci computazionali avanzati, miriamo a delineare quale tra questi modelli si adatta meglio alla previsione delle caratteristiche qualitative del vino, basandosi su parametri misurabili in modo oggettivo.
I modelli analizzati comprendono:

- K-Nearest Neighbors (KNN): Un approccio basato sulla prossimità ai campioni più simili, che stabilisce la classificazione basandosi sulle caratteristiche degli elementi vicini.
- Naïve Bayes: Un modello che applica principi di probabilità, spesso usato per la sua semplicità e l'efficacia anche in presenza di basi dati di grandi dimensioni.
- Random Forest e AdaBoost: Due potenti modelli di ensemble che combinano più algoritmi di apprendimento per ottenere risultati più accurati e robusti rispetto ai singoli classificatori.
- Support Vector Machine (SVM): Nota per la sua capacità di operare efficacemente in spazi multidimensionali e per la sua versatilità in varie applicazioni di classificazione.

Oltre a valutare la performance di questi modelli, si approfondisce la comprensione dei meccanismi attraverso cui questi algoritmi processano e interpretano i dati. Da un lato, fornisce una base solida per ulteriori ricerche in questo ambito; dall'altro, offrire spunti applicativi concreti che possano essere utilizzati dall'industria vinicola per affinare i processi di produzione e di controllo qualità.
Con una metodologia rigorosa e un approccio critico, miriamo a delineare non solo quale modello è il più efficace, ma anche perché alcuni modelli riescono meglio di altri a catturare le caratteristiche che determinano la qualità di un vino.

2)Linguaggio e Ambienti Utilizzati

Il progetto di classificazione della qualità del vino è stato interamente sviluppato utilizzando Python 3.8, una scelta motivata dalla sua estesa adozione nella comunità scientifica e di analisi dei dati. Python offre una sintassi intuitiva e flessibile, supportata da un ampio ecosistema di librerie per il data processing e il machine learning, rendendolo ideale per prototipare e implementare complessi algoritmi di analisi.
Per garantire un ambiente di sviluppo pulito e controllato, è stato configurato un ambiente virtuale Python. Questo approccio isola le librerie utilizzate nel progetto dalle installazioni globali, prevenendo conflitti e facilitando la condivisione del progetto con setup minimi. L'ambiente virtuale può essere facilmente replicato su altri sistemi operativi o configurazioni hardware, garantendo così la riproducibilità degli esperimenti.

3)Librerie Utilizzate

Le seguenti librerie Python hanno svolto un ruolo cruciale nello sviluppo e nell'analisi del progetto:

- NumPy: Fondamentale per la manipolazione ad alta performance di grandi array di dati, con supporto per complesse operazioni matematiche e algebriche.
- Pandas: Essenziale per il caricamento, la manipolazione e l'analisi dei dati tramite strumenti che semplificano la gestione di DataFrame complessi e la pre-elaborazione dei dati.
- Matplotlib e Seaborn: Queste librerie forniscono vasti strumenti di visualizzazione, essenziali per l'analisi esplorativa dei dati e la presentazione dei risultati attraverso grafici intuitivi e professionali.
- Scikit-learn: Il pilastro per la modellazione statistica e il machine learning, utilizzato per implementare e valutare modelli di classificazione, gestire la selezione delle features, la normalizzazione dei dati e la divisione del dataset.

4)Manuale d'Uso

Il programma è progettato per facilitare l'analisi e la classificazione della qualità del vino attraverso diverse fasi:
Preparazione del Dataset:

- Assicurati che il dataset sia pre-elaborato conformemente agli standard descritti, con features normalizzate e target categorizzati. Verifica che il dataset non contenga valori mancanti e che sia pronto per l'analisi.

Esecuzione del Programma:

- Avvio: Esegui il file main.py usando il comando python main.py in un terminale che opera nell'ambiente virtuale configurato.
- Processo: Durante l'esecuzione, il programma caricherà il dataset, eseguirà la PCA per ridurre la dimensionalità, e addestrerà i modelli di classificazione come Random Forest e SVM.
- Visualizzazione dei Risultati: Il software visualizzerà grafici e metriche di valutazione in tempo reale, permettendo di monitorare la performance dei modelli e di valutare l'importanza delle diverse features.

Analisi dei Risultati:

- Consulta i grafici di performance e le matrici di confusione generati per comprendere l'efficacia dei vari modelli e identificare il più performante per il set di dati di vino in esame.

Iterazione e Validazione:

- Il programma è progettato per essere iterativo; puoi modificare i parametri dei modelli o i metodi di preprocessing e ripetere l'esecuzione per testare diverse configurazioni o applicare il sistema ad altri dataset simili.

5)Base di Conoscenza

La base di conoscenza del sistema di classificazione della qualità del vino è progettata per identificare automaticamente le relazioni tra le caratteristiche chimiche e sensoriali dei vini e la loro classificazione in termini di qualità. Utilizziamo questa base per guidare il processo decisionale del modello e per fornire spiegazioni comprensibili sulle previsioni.
Struttura della Knowledge Base
La base di conoscenza è costruita attorno a un sistema di regole che definiscono come specifiche

caratteristiche del vino influenzano l'assegnazione alla categoria di qualità "GOOD" o "BAD". Le regole sono supportate da fatti, che rappresentano le osservazioni derivanti dall'analisi dei dati.
Regole di Classificazione Le regole stabiliscono criteri basati su soglie scientificamente validate o statisticamente derivabili dalle caratteristiche del vino, come segue:

- Good Quality ⟵ (alcohol > 12%) ∧ (sulphates > 0.65) ∧ (citric acid > 0.4)
- Bad Quality ⟵ (volatile acidity > 0.6) ∧ (pH > 3.5)

Queste regole aiutano a categorizzare i vini e sono essenziali per il training degli algoritmi di apprendimento supervisionato, fornendo una guida su quali feature guardare durante la classificazione.
I fatti sono suddivisi in:

- Fatti Dichiarati: Caratteristiche osservate nei dati, come l'acidità volatile, il contenuto di alcol, i livelli di zuccheri, ecc.
- Fatti Derivati: Classificazioni ottenute applicando le regole di classificazione, che assegnano ogni campione di vino a una categoria di qualità.

Categorie Identificate La base di conoscenza permette di categorizzare i vini in "GOOD" e "BAD", basandosi sulle regole definite. Questa categorizzazione è cruciale per la fase di apprendimento del modello e per la successiva valutazione delle performance.
Supporto Decisionale In caso di incertezza nella classificazione, il sistema può suggerire di esaminare ulteriori dati o di riconsiderare le soglie usate nelle regole. Questo approccio iterativo migliora continuamente la base di conoscenza, adattandola ai nuovi dati e alle nuove scoperte nel campo della vinificazione.
Utilizzare questa struttura per la base di conoscenza non solo rende il modello di classificazione trasparente ma anche fornisce una spiegazione logica alle decisioni prese, facilitando l'interpretazione dei risultati da parte degli utenti e migliorando la fiducia nel sistema di classificazione automatizzato.

6)DATASET

Il dataset impiegato in questo studio è tratto dalla repository dell'UCI (University of California, Irvine) e si concentra esclusivamente sulla variante rossa del “Vinho Verde”, una denominazione di origine protetta portoghese. Questo dataset è composto da 4898 campioni di vini, con i nomi omessi per privacy. Le informazioni raccolte sono utilizzate per analizzare le correlazioni tra le composizioni chimiche dei vini e la loro qualità percepita, valutata su una scala da 1 a 10. Di seguito, sono elencate le caratteristiche (features) analizzate:

1. Fixed Acidity: L'acidità fissa si riferisce agli acidi che non evaporano facilmente. L'equilibrio degli acidi nel vino è fondamentale per la sua freschezza e può influenzare notevolmente il gusto.
2. Volatile Acidity: Misura la quantità di acido acetico nel vino, essenziale per valutare la tendenza del vino a sviluppare un aroma di aceto se presente in quantità eccessive.
3. Citric Acid: Utilizzato in piccole quantità, l'acido citrico può migliorare la freschezza e il sapore del vino.
4. Residual Sugar: Rappresenta lo zucchero residuo post-fermentazione. La quantità di zucchero residuo classifica i vini in categorie da secco a dolce.
5. Chlorides: Indica la quantità di sale nel vino, che può influenzare il sapore generale.
6. Free Sulfur Dioxide: La presenza di anidride solforosa libera aiuta a limitare l'ossidazione e preserva il vino da batteri indesiderati.

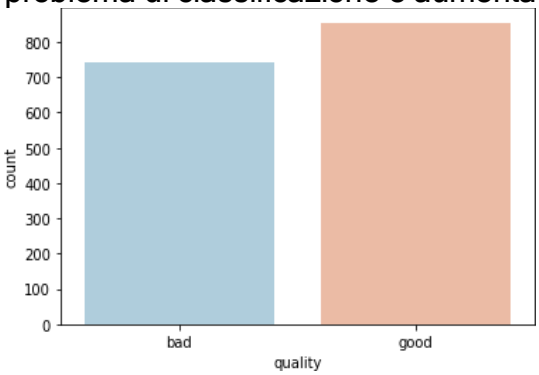
7. Total Sulfur Dioxide: La somma di anidride solforosa libera e legata. A livelli alti, può diventare evidente all'olfatto e al gusto.
8. Density: La densità del vino è influenzata dall'alcol e dal contenuto zuccherino. Questa misura è importante per la determinazione del grado alcolico.
9. pH: Il pH del vino è un indicatore del suo livello di acidità, che può influenzare colore, fermentazione e microbiologia.
10. Sulphates: I solfati sono additivi che possono contribuire ai livelli di anidride solforosa, influenzando la resistenza del vino contro microorganismi e ossidazione.
11. Alcohol: La percentuale di alcol determina non solo la potenza del vino ma anche influenze sulla percezione di vari aromi e sapori.
12. Quality: La qualità del vino, valutata da esperti e espressa su una scala da 1 a 10, serve come variabile di risposta nel modello.

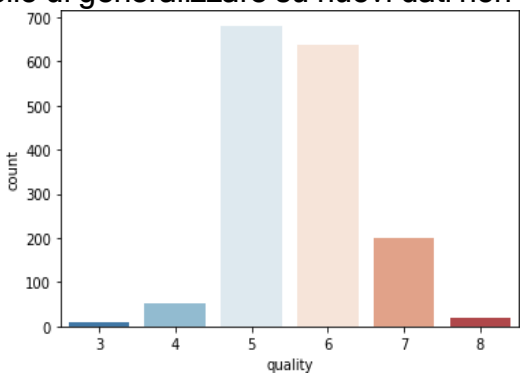
Rilevanza del Dataset

L'uso di questo dataset fornisce un'opportunità unica per esplorare come variazioni sottili in componenti chimici possano alterare la percezione della qualità del vino. L'analisi approfondita di questi dati aiuta a identificare quali caratteristiche hanno il maggiore impatto sulla qualità percepita, offrendo intuizioni preziose ai produttori di vino su come migliorare i loro prodotti. Inoltre, i risultati di questo studio possono essere utilizzati per educare i consumatori sulle proprietà chimiche che influenzano il loro apprezzamento del vino.

7)PREPROCESSING

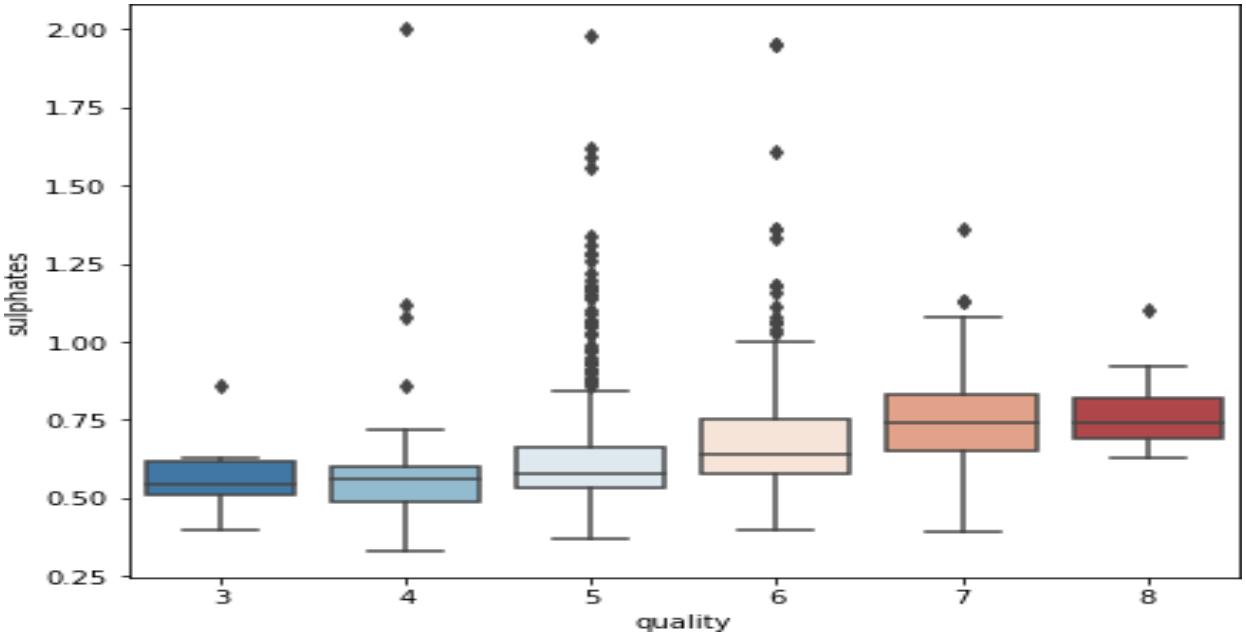
Il dataset utilizzato per questo studio non presenta valori mancanti (NaN), il che facilita le fasi iniziali di pulizia dei dati. Tuttavia, una sfida significativa nel dataset è rappresentata dallo sbilanciamento delle classi della feature target, come evidenziato dall'analisi esplorativa dei dati.
Sbilanciamento dei dati: Come illustrato nel grafico sottostante, la distribuzione delle valutazioni di qualità mostra una prevalenza di voti concentrati intorno ai valori 5 e 6. Questo sbilanciamento può portare a modelli di machine learning che sono inclini a favorire le classi maggioritarie, riducendo di conseguenza l'accuratezza della predizione per le classi minoritarie.
Raggruppamento delle classi: Per mitigare l'effetto dello sbilanciamento e rendere il modello più robusto, ho optato per un approccio di binarizzazione delle etichette di qualità. Specificamente, i vini con una valutazione inferiore a 5.5 sono stati categorizzati come "BAD", mentre quelli con una valutazione di 5.5 o superiore come "GOOD". Questa trasformazione riduce la complessità del problema di classificazione e aumenta la capacità del modello di generalizzare su nuovi dati non visti

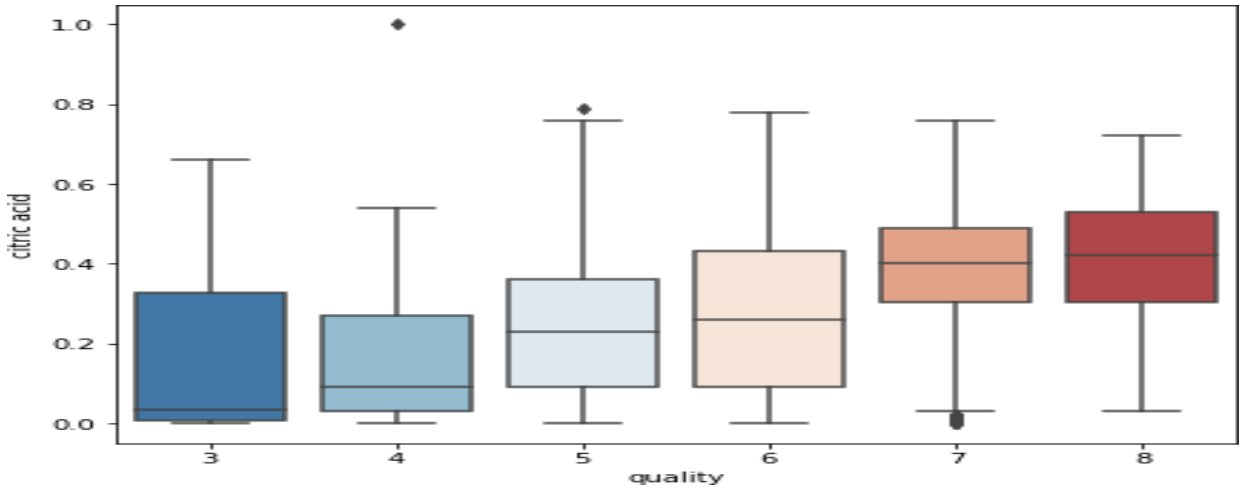


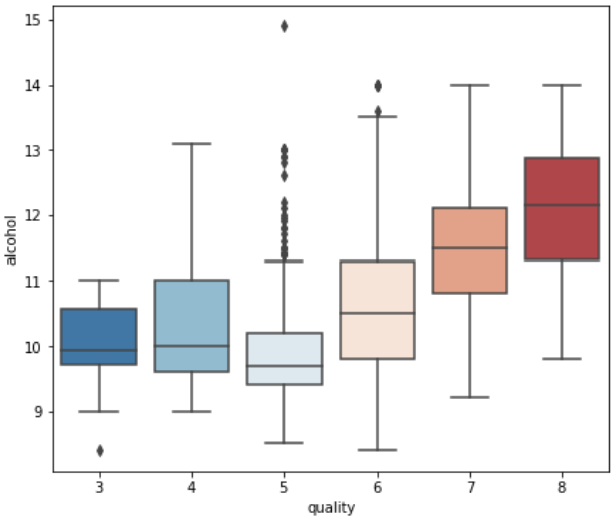


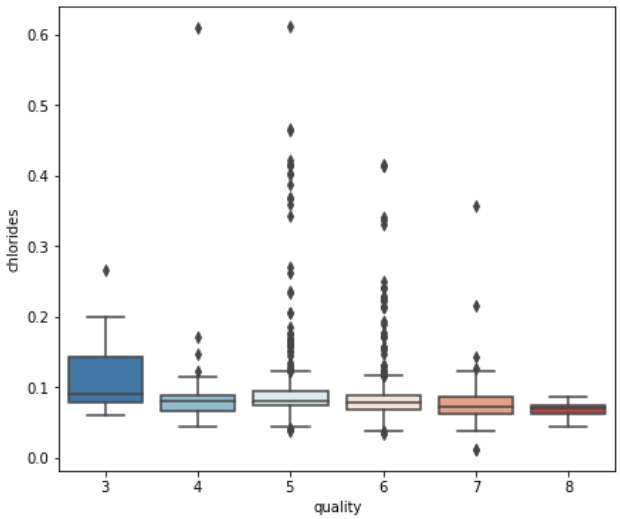
8)CORRELAZIONE INPUT/OUTPUT

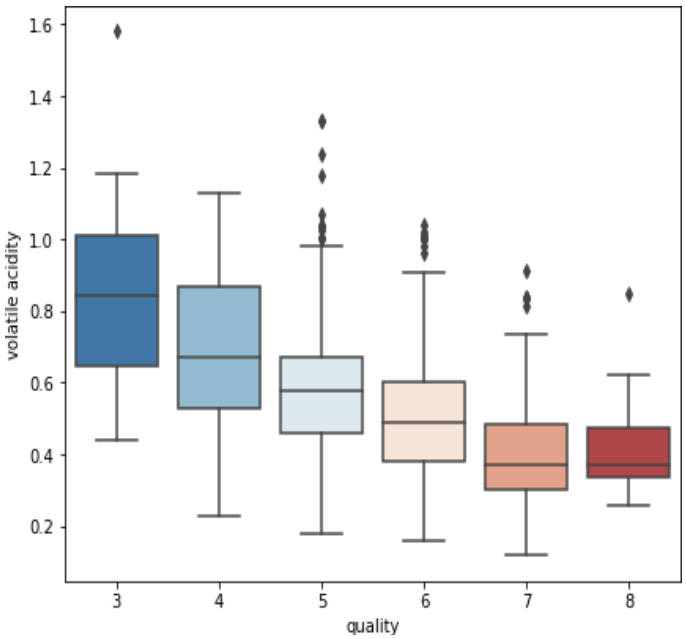
Analisi della Correlazione: Prima di procedere alla classificazione, ho esaminato attentamente la correlazione tra le features di input e la feature obiettivo, ossia la qualità del vino. Questo passaggio preliminare è cruciale per identificare quali componenti chimico-fisiche influenzano maggiormente la percezione della qualità del vino. Utilizzando le librerie seaborn e matplotlib, ho generato una serie di boxplot che illustrano visivamente le relazioni tra la qualità e ciascuna delle features. Questi grafici sono strumenti potenti per identificare tendenze e anomalie nei dati.
Osservazioni dai Boxplot: Dall'analisi dei boxplot emerge che:
Un aumento dell'acido citrico è generalmente associato a un miglioramento della qualità del vino, suggerendo che l'acidità contribuisce positivamente alla freschezza e al gusto complessivo.
Analogamente, incrementi nei solfati e nell'alcool sono correlati a valutazioni più elevate, il che riflette l'importanza di questi componenti nella stabilizzazione e nel miglioramento del profilo aromatico del vino.
D'altra parte, un aumento dei cloridi e degli acidi volatili tende a essere associato a una diminuzione della qualità, indicando che un eccesso di questi elementi può alterare negativamente il sapore e l'aroma del vino.
Applicazione della PCA: Data la complessità del dataset e l'intercorrelazione tra molte features, ho applicato l'Analisi delle Componenti Principali (PCA) per semplificare il modello riducendo la dimensionalità dei dati. Questo metodo di riduzione lineare trasforma le features originarie in un nuovo set di variabili, i componenti principali, che sono ortogonali tra loro e catturano la massima varianza possibile. I dati sono stati standardizzati prima dell'applicazione della PCA per assicurare che ogni feature contribuisse equamente all'analisi.
Visualizzazione dell'Effetto della PCA: Il grafico della varianza spiegata cumulativa mostra chiaramente che le prime otto componenti principali rappresentano la maggior parte dell'informazione utile. Questo risultato indica che possiamo concentrarci su queste componenti per i successivi modelli di classificazione, migliorando così l'efficienza computazionale e riducendo il rischio di overfitting.

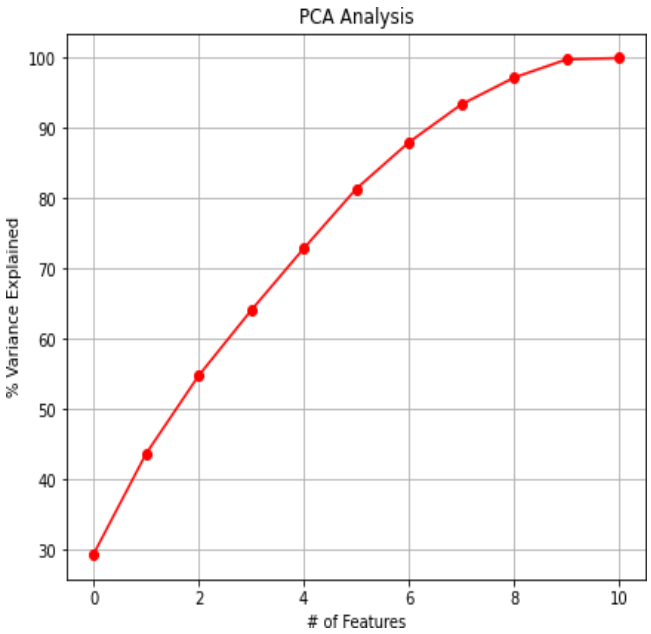



citric acid
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
quality


alcohol
15
14
13
12
11
10
9
3
4
5
6
7
8
quality


chlorides
0.6
0.5
0.4
0.3
0.2
0.1
0.0
3
4
5
6
7
8
quality


volatile acidity
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
3
4
5
6
7
8
quality


PCA Analysis
% Variance Explained
100
90
80
70
60
50
40
30
0
2
4
6
8
10
# of Features

9)CLASSIFICAZIONE

Utilizzando la libreria sklearn sono stati costruiti diversi modelli di classificazione.

Support Vector Machine:
il Support Vector Machine ha l'obiettivo di identificare l'iperpiano che meglio divide i vettori di supporto in classi. Per farlo esegue i seguenti step:
Cerca un iperpiano linearmente separabile o un limite di decisione che separa i valori di una classe dall'altro. Se ne esiste più di uno, cerca quello che ha margine più alto con i vettori di supporto, per migliorare l'accuratezza del modello.
Se tale iperpiano non esiste, SVM utilizza una mappatura non lineare per trasformare i dati di allenamento in una dimensione superiore (se siamo a due dimensioni, valuterà i dati in 3 dimensioni). In questo modo, i dati di due classi possono sempre essere separati da un iperpiano, che sarà scelto per la suddivisione dei dati.
I nuovi esempi sono mappati nell'iperpiano e la predizione della categoria alla quale appartengono viene fatta individuando il lato dell'iperpiano nel quale ricade.
L'algoritmo SVM ottiene la massima efficacia nei problemi di classificazione binari.

Random Forest:
È un modello ottenuto dall'aggregazione tramite bagging di alberi di decisione. Esso è un meta-stimatore che si adatta ad una serie di alberi decisionali addestrati su vari sotto-campioni del dataset e utilizza la media di ogni singolo output di ogni albero per migliorare l'accuratezza predittiva e il controllo del sovradattamento. Il Random Forest deve essere dotato di due matrici: una matrice X sparsa che contiene i campioni di addestramento e una matrice Y di dimensioni che contiene i valori target.

GaussianNB:
l'algoritmo supporta dati numerici continui e presuppone che i valori di ciascuna caratteristica siano normalmente distribuiti (ossia ricadono da qualche parte su una curva a campana). In altre parole, Naive Bayes può essere esteso ad attributi a valori reali, più comunemente assumendo una distribuzione gaussiana o normale. Secondo questa assunzione è sufficiente trovare la media e la deviazione standard di ciascuna probabilità per ogni attributo e per ogni singola classe.
Sostituendo tali valori nella funzione di densità di probabilità gaussiana (detta anche Gaussian Probability Density Function) si ricava una probabilità che permette di ricavare le varie probabilità di classe. Il valore di probabilità di classe più alto così ottenuto rappresenta la classe da associare alla nuova istanza che si vuole categorizzare.

K-Nearest Neighbors:
Viene chiamato algoritmo lazy learner perché non apprende immediatamente dal set di addestramento, ma memorizza il set di dati e al momento della classificazione esegue un'azione sul set di dati. Infatti, calcola la somiglianza tra un nuovo esempio e gli esempi disponibili nel dataset assegnandone l'etichetta più simile alle categorie disponibili. In altre parole, memorizza tutti i dati disponibili e classifica un nuovo esempio in base alla somiglianza.

AdaBoost:
AdaBoost è un modello di ensemble boosting che utilizza alberi decisionale. L'output del meta-classificatore (alberi decisionali) è dato dalla somma pesata delle predizioni dei singoli modelli. Ogni qual volta un modello viene addestrato, ci sarà una fase di ripesaggio delle istanze.
L'algoritmo di boosting tenderà a dare un peso maggiore alle istanze misclassificate, nella speranza che il successivo modello sia più esperto su quest'ultime. Sostanzialmente, ad ogni iterata calcola il tasso di errore ponderato dell'albero decisionale, ovvero il numero di predizioni sbagliate sul totale delle predizioni, che dipende dai pesi associati agli esempi nel dataset;
successivamente in base all'errore calcola il learning rate dell'albero decisionale.

maggiore è il tasso di errore di un albero, minore sarà il potere decisionale che l'albero avrà durante la predizione successiva minore è il tasso di errore di un albero, maggiore sarà il potere decisionale assegnato all'albero durante la predizione successiva.

10)Analisi dei Grafici di Regressione

Nel progetto sulla classificazione della qualità del vino, è incluso l'analisi dei grafici di regressione per esplorare le relazioni tra le caratteristiche chimiche e sensoriali del vino e la loro qualità percepita. I grafici di regressione, come quello mostrato nell'immagine, rappresentano una visualizzazione potente per comprendere come variabili specifiche, in questo caso il contenuto alcolico, influenzino la qualità del vino.
Importanza dei Grafici di Regressione
Questi grafici sono particolarmente utili per:

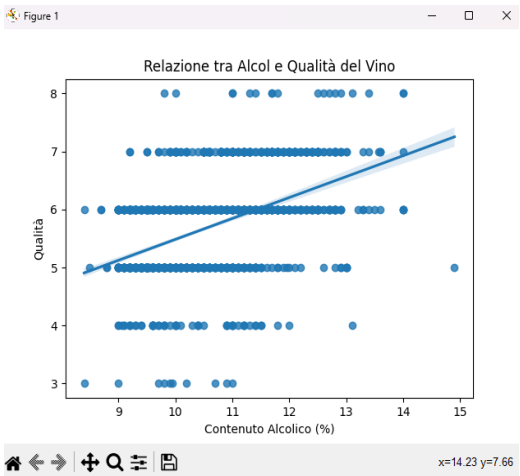
- Valutare la Relazione Lineare: Il grafico mostra una chiara tendenza ascendente che indica che, generalmente, un maggiore contenuto alcolico è associato a una qualità del vino percepita come superiore.
- Identificare Outlier: Attraverso il grafico, possiamo identificare facilmente eventuali dati anomali che non seguono la tendenza generale. Questo è utile per ulteriori indagini o per pulire i dati.
- Verificare Assunzioni di Modellazione: Nella regressione lineare, è fondamentale che la relazione tra le variabili sia lineare. Questi grafici aiutano a confermare o a mettere in dubbio tale assunzione.

Interpretazione del Grafico Specifico

- Dispersione dei Punti: Notiamo che i dati sono raggruppati principalmente intorno a valori di qualità da 5 a 7, con una distribuzione più densa attorno al contenuto alcolico del 9% al 12%.
- Linea di Regressione: La linea blu nel grafico rappresenta la migliore linea di adattamento che mostra la tendenza generale. L'ombreggiatura blu chiaro attorno alla linea indica l'intervallo di confidenza, che ci dà un'idea della variabilità attesa nella previsione della qualità del vino basata sul contenuto alcolico.

Implicazioni e Utilizzo nel Progetto

- Orientamento alla Classificazione: Anche se il focus è sulla classificazione, comprendere le relazioni lineari tra le variabili può aiutare a migliorare la selezione delle features e l'interpretazione dei modelli di classificazione.
- Supporto nella Selezione delle Features: Le insights ottenute dall'analisi dei grafici di regressione possono guidare decisioni informate sulla selezione delle variabili da includere nei modelli predittivi, migliorando così l'efficacia e l'efficienza del modello finale.



Questo è un grafico di regressione che mostra la relazione tra il contenuto alcolico dei vini e la loro qualità. Questo tipo di grafico è molto utile per visualizzare tendenze o correlazioni tra due variabili quantitative:

- **Asse X (orizzontale):** Rappresenta il contenuto alcolico del vino, espresso in percentuale.
- **Asse Y (verticale):** Rappresenta la qualità del vino, che presumibilmente è stata valutata su una scala numerica (probabilmente da 1 a 10).

La **linea di tendenza** (linea blu) mostra la relazione generale tra il contenuto alcolico e la qualità del vino. La zona ombreggiata attorno alla linea indica l'intervallo di confidenza per la stima della regressione, che aiuta a comprendere dove si prevede che i dati si adattino con una certa probabilità, assumendo che la relazione lineare sia corretta.
Interpretazione:

- **Pendenza positiva:** Indica che, in generale, all'aumentare del contenuto alcolico, aumenta anche la qualità percepita del vino. In altre parole, i vini con un maggiore grado alcolico tendono ad avere valutazioni di qualità più alte.

11)IPERPARAMETRI

Con GridSearchCV, si sono generati in maniera esaustiva i possibili candidati (iperparametri) attraverso una griglia di valori specificata opportunamente dal parametro "param_grid", caratterizzato da un range di valori per ogni singolo parametro specificato dall'utente. In maniera del tutto automatica, vengono valutate tutte le possibili combinazioni di assegnazioni degli iperparametri e viene mantenuta la combinazione migliore. Al termine di tale processo, verranno
mostrati quelli che sono gli iperparametri migliori per un determinato modello di classificazione. Gli iperparametri utilizzati sono:

C: 0.1, 1, 10, 100, 1000; gamma: 1, 0.1, 0.01, 0.001, 0.0001; kernel: rbf, sigmoid; per SVM

n_estimators: 100, 250, 500; max_features: auto, log2; criterion :gini, entropy per RF

n_neighbors:1,4,5,6,7,8; leaf_size:1,3,5,10; per KNN

I migliori sono stati:

C: 10, gamma: 0.1, kernel: rbf

criterion: gini, max_features: auto, n_estimators: 500

leaf_size: 1, n_neighbors: 1

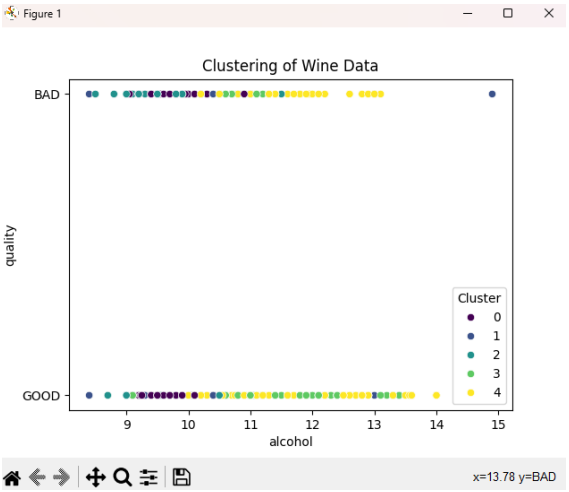
12)Clustering nel Classificatore di Vini

Il clustering è un metodo di apprendimento non supervisionato usato per scoprire strutture e pattern nascosti nei dati. Nel mio studio sul vino, ho utilizzato l'algoritmo KMeans per esaminare le relazioni tra le varie caratteristiche chimiche e sensoriali dei vini, il che aiuta a capire come i vini si raggruppano in base alle loro proprietà e come queste influenzano la percezione della qualità.
Scopo e Utilizzo del Clustering
L'obiettivo del clustering nel progetto è di identificare gruppi naturali all'interno dei dati che potrebbero corrispondere a diverse qualità o tipi di vino. Questo approccio fornisce una nuova prospettiva sulle caratteristiche che definiscono un vino di alta qualità rispetto a uno di qualità inferiore e offre una validazione indipendente delle etichette di classificazione utilizzate nei nostri modelli di apprendimento supervisionato.

Implementazione del Clustering con KMeans:
Preparazione dei Dati: normalizzati i dati per garantire un contributo equo di ogni caratteristica nel processo di clustering, evitando distorsioni dovute alle differenze di scala tra le variabili.
Applicazione di KMeans: ho utilizzato l'analisi della silhouette e il metodo del gomito per determinare il numero ottimale di cluster. Ciò ha permesso di definire una quantità di gruppi che meglio rappresenta la variabilità dei dati senza incorrere in sovraadattamento.

Analisi dei Cluster: sono state esaminate le caratteristiche medie e la distribuzione delle qualità del vino all'interno di ogni cluster per identificarne le proprietà comuni.

Visualizzazione:
Utilizzo scatter plot con dati colorati per cluster per visualizzare le divisioni tra i gruppi e comprendere le loro proprietà distintive. Questo aiuta a visualizzare come proprietà come l'alcol e l'acidità influenzano la formazione dei cluster.
Risultati e Implicazioni:
L'analisi ha mostrato che il contenuto alcolico è una delle proprietà chimiche che maggiormente influenzano la formazione dei gruppi, con vini ad alto contenuto alcolico che tendono a raggrupparsi insieme. Questo suggerisce una forte correlazione tra alcol e percezioni di qualità superiore.
Reporting:
Documentiamo meticolosamente il processo di clustering e i risultati ottenuti, includendo descrizioni dettagliate delle tecniche utilizzate, grafici esplicativi, e discussioni sulle implicazioni dei risultati per le strategie di produzione e marketing del vino.
Questo approccio non solo migliora la comprensione dei dati, ma apre anche nuove possibilità per strategie di analisi e sviluppo di prodotti basate su un'approfondita comprensione delle qualità che caratterizzano i vari tipi di vino.



Il grafico rappresenta il risultato di un'analisi di clustering sui dati relativi alla qualità del vino, utilizzando come variabili il contenuto alcolico e la qualità del vino categorizzata come "BAD" o "GOOD". Qui sono stati impiegati colori diversi per distinguere i vari cluster identificati tramite l'algoritmo KMeans. Ogni colore rappresenta un cluster diverso, indicando gruppi di vini che condividono caratteristiche simili in termini di contenuto alcolico e percezione di qualità.

Interpretazione del grafico:

- Asse X (Contenuto Alcolico %): Mostra il contenuto alcolico dei vini. L'asse va da circa 9% a 15%, indicando una varietà di vini da quelli meno forti a quelli più forti in termini di gradazione alcolica.
- Asse Y (Qualità): Categorizza i vini come "BAD" o "GOOD", mostrando una chiara separazione verticale tra queste due classificazioni. Questo suggerisce che il clustering ha potuto distinguere efficacemente tra vini considerati di qualità inferiore e superiore basandosi sui dati disponibili.
- Colori/Cluster: I diversi colori rappresentano i cluster formatisi durante l'analisi di clustering. Ogni cluster raggruppa vini che sono simili per contenuto alcolico e valutazione di qualità. Ad esempio, i vini nel cluster giallo tendono ad avere un alto contenuto alcolico e sono classificati come "BAD", mentre quelli nel cluster blu scuro sono prevalentemente classificati come "GOOD" e hanno un contenuto alcolico medio-alto.
- **Analisi del Consumatore:** Questo tipo di visualizzazione può aiutare i produttori di vino a comprendere meglio le preferenze dei consumatori, associando il contenuto alcolico alla percezione di qualità.
- **Segmentazione del Mercato:** Identificare cluster specifici può aiutare nella creazione di strategie di marketing mirate, per esempio focalizzandosi su consumatori che preferiscono vini forti ma di qualità percepite come inferiore, o viceversa.

- **Sviluppo del Prodotto:** Le informazioni derivanti dal clustering possono guidare lo sviluppo di nuovi vini o l'adattamento delle ricette esistenti per soddisfare o creare nicchie di mercato specifiche.

Questo grafico non solo offre un'istantanea visiva della distribuzione dei dati di qualità del vino rispetto al contenuto alcolico ma apre anche a discussioni su come queste variabili influenzino la percezione del consumatore e come possono essere utilizzate per migliorare la produzione e la commercializzazione del vino.

13)Sviluppi Futuri

Il progetto attuale fornisce una robusta base di classificazione e analisi dei vini, tuttavia, esistono numerose direzioni attraverso cui il sistema potrebbe essere migliorato e ampliato per rispondere a nuove esigenze e incorporare tecnologie emergenti.

Integrazione di Algoritmi di Deep Learning:
L'introduzione di algoritmi di deep learning potrebbe notevolmente incrementare l'accuratezza della classificazione dei vini. L'uso di modelli come le reti neurali convoluzionali (CNN) o le reti neurali ricorrenti (RNN) potrebbe permettere la rilevazione di pattern complessi nelle caratteristiche chimico-sensoriali dei vini, che i classificatori tradizionali potrebbero non riuscire a catturare. L'implementazione di questi modelli richiederebbe l'utilizzo di librerie avanzate come TensorFlow o PyTorch e potrebbe necessitare la creazione di un dataset più ampio e diversificato.

Miglioramenti nell'Algoritmo di Clustering:
Attualmente, il sistema impiega l'algoritmo K-Means per il clustering. In futuro, si potrebbero esplorare algoritmi di clustering più sofisticati come DBSCAN o modelli basati su miscele gaussiane (Gaussian Mixture Models, GMM), che possono gestire meglio la diversità delle forme dei cluster e la presenza di rumore. Questo miglioramento potrebbe rafforzare la capacità del sistema di identificare sottogruppi di vini con caratteristiche simili, migliorando la personalizzazione delle strategie di marketing e produzione.

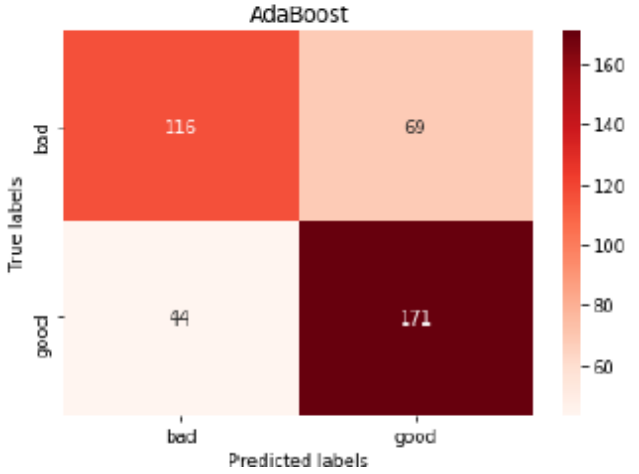
Interfaccia Utente e Funzionalità Interattive:
Sviluppare un'interfaccia utente grafica (GUI) renderebbe il sistema più accessibile agli utenti non tecnici, facilitando operazioni come il caricamento dei dati, la visualizzazione delle analisi e la ricezione di raccomandazioni. Integrare funzionalità interattive, come la possibilità per gli utenti di etichettare manualmente i vini o di fornire feedback sulle classificazioni, potrebbe notevolmente migliorare l'apprendimento del sistema e la sua precisione attraverso un feedback continuo o l'apprendimento attivo.

Adattamento alla Variabilità dei Vini:
Mentre si espande la base di dati con nuovi vini da diverse regioni o annate, potrebbe essere necessario aggiustare i modelli per tenere conto delle variazioni intrinseche. L'introduzione di algoritmi che possono adattarsi dinamicamente ai cambiamenti nel dataset garantirebbe che il sistema rimanga efficace e accurato nel tempo.

Questi sviluppi non solo potrebbero potenziare le capacità analitiche del sistema ma anche aprire nuove strade per la comprensione approfondita e la gestione ottimale della produzione e distribuzione dei vini.

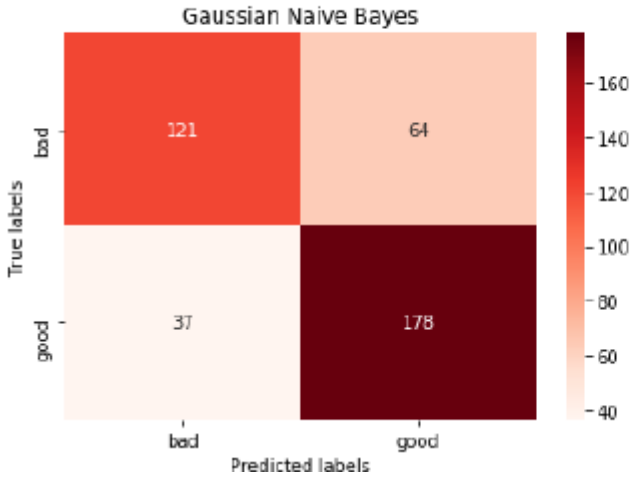
14)RISULTATI

Adaboost:



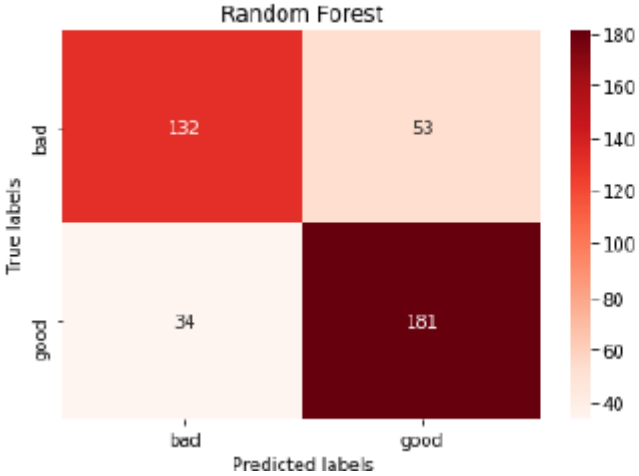
	precision	recall	f1-score
bad	0.72	0.63	0.67
good	0.71	0.80	0.75
accuracy	0.72		
Macro avg	0.72	0.71	0.71
Weighted avg	0.72	0.72	0.72

GaussianNB:



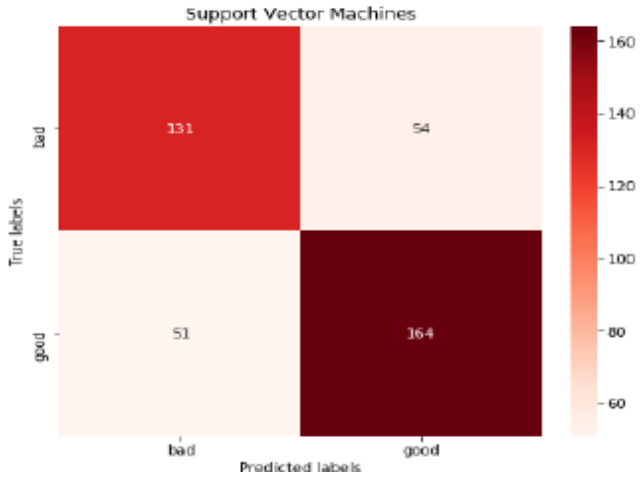
	precision	recall	f1-score
bad	0.77	0.65	0.71
good	0.74	0.83	0.78
accuracy	0.75		
Macro avg	0.75	0.74	0.74
Weighted avg	0.75	0.75	0.75

Random Forest:



	precision	recall	f1-score
bad	0.80	**0.71**	0.75
good	0.77	**0.84**	0.81
accuracy	**0.78**		
Macro avg	0.78	0.78	0.78
Weighted avg	0.78	0.78	0.78

Support Vector machines:



	precision	recall	f1-score
bad	0.72	0.71	0.71
good	0.75	0.76	0.76
accuracy	0.74		
Macro avg	0.74	0.74	0.74
Weighted avg	0.74	0.74	0.74

Algoritmo	**Accuracy**
SVM	0.74
Random forest	0.78
K-NN	0.73
Adaboost	0.72
Naïve Bayes	0.75