**PROGETTO DI STATISTICA NUMERICA**

“*Classificazione delle qualità dei vini*”

Nicolas

Cola

**Pre-Processing**

In questa fase ho controllate che non ci fossero NaN con la relativa funzione, e che non ci fossero valori fuori soglia utilizzando dei Boxplot, da cui si può notare come in molti features siano presenti. Ciò sicuramente influenzerà nella corretta classificazione delle qualità. Tuttavia, non avendo informazioni sulla causa di questi valori fuori soglia (errori di misura, errori tipografici o tendenze non previste), ho ritenuto opportuno non fare nulla a riguardo.

Immagine che contiene testo, diagramma, Parallelo, schermata

Descrizione generata automaticamente

**Exploratory Data Analysis (EDA)**

In questa fase ho utilizzato dei Catplot per vedere le tipologie di qualità presenti nel dataset che mi ha mostrato come fossero presenti solo le qualità 3, 4, 5, 6, 7, 8 di cui le qualità medie (5 e 6) in maggior quantità. Sulla base di ciò ho ridenominato le qualità in:

* non buono le qualità 3 e 4 assegnando come valore 0
* medio buono le qualità 5 e 6 assegnando come valore 1
* buono le qualità 7 e 8 assegnando come valore 2

In questo modo il dataset è più sbilanciato sulle qualità medie, il che mi sembrava una rappresentazione più corretta.

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente Immagine che contiene testo, schermata, diagramma, Rettangolo

Descrizione generata automaticamente

Ho utilizzato dei grafici a barre per vedere come al variare di una variabile variasse la qualità, in modo da cercare di capire da cosa potesse essere determinata. Ottenendo che:

* All'aumentare dell’acidità volatile la qualità diminuisce,
* All'aumentare dell’acido citrico la qualità aumenta
* All'aumentare dei cloruri la qualità diminuisce
* All'aumentare dei solfati la qualità aumenta
* All'aumentare dell'alcol la qualità aumenta

Immagine che contiene testo, schermata, Rettangolo, diagramma

Descrizione generata automaticamente Immagine che contiene schermata, testo, diagramma, Rettangolo

Descrizione generata automaticamente Immagine che contiene testo, schermata, diagramma, Rettangolo

Descrizione generata automaticamente Immagine che contiene testo, schermata, Rettangolo, diagramma

Descrizione generata automaticamente Immagine che contiene schermata, testo, Rettangolo, Policromia

Descrizione generata automaticamente

Successivamente ho utilizzato la matrice di correlazione, rappresentata da una Heatmap, per visualizzare il grado di correlazione tra le variabili, trovando che alcune di esse sono fortemente correlate negativamente (-0.7) o positivamente (0.7).

Immagine che contiene testo, schermata, diagramma, Piano

Descrizione generata automaticamente

**SPLITTING**

Ho suddiviso il dataset in Training-Set di 1119 righe, e Validation-Set e Test-Set di 200 righe entrambi

**REGRESSIONE**

Ho scelto due coppie di variabili fortemente correlate:

1. X = fixed-acidity e Y = citric acid con correlazione = 0.7 ottenendo:

* Coefficiente angolare = 0.0752, indica che per ogni unità di aumento di fixed-acidity, citric-acid aumenta in media di 0.0752 unità.
* Intercetta = -0.3543,
* Coefficiente R^2 = 0.45, indica che il 45% della variabilità nel citric acid è spiegata dalla fixed-acidity. Questo suggerisce una relazione moderata ma non particolarmente forte.
* Mean Squared Error = 0.0208
* I residui non seguono una distribuzione normale (p-value:0.0000000000), la non normalità dei residui può suggerire che il modello non è adeguato, che ci potrebbero essere variabili non incluse nel modello che influenzano la variabile dipendente oppure potrebbe essere dovuta dalla presenza di valori fuori soglia nelle variabili.

Immagine che contiene testo, schermata, mappa, linea

Descrizione generata automaticamente Immagine che contiene diagramma, Diagramma, schermata, linea

Descrizione generata automaticamente Immagine che contiene testo, linea, Diagramma, diagramma

Descrizione generata automaticamente

1. X= fixed-acidity e Y = pH con correlazione = -0.7 ottenendo:

* Coefficiente angolare: -0.0606, indica che per ogni unità di aumento di fixed-acidity, pH diminuisce in media 0.0606 unità.
* Intercetta: 3.8150
* Coefficiente R^2: 0.47, indica che il 47% della variabilità nel pH è spiegata dalla fixed-acidity. Questo suggerisce una relazione moderata ma non particolarmente forte.
* Mean Squared Error: 0.0127
* I residui non seguono una distribuzione normale (p-value:0.0000000001)

Immagine che contiene testo, schermata, linea, Diagramma

Descrizione generata automaticamente Immagine che contiene diagramma, schermata, Diagramma, linea

Descrizione generata automaticamente Immagine che contiene testo, linea, Diagramma, diagramma

Descrizione generata automaticamente

In questa fase ho utilizzato scatterplot per visualizzare la retta di regressione e istogramma e qq-plot per la normalità dei residui

**HYPERPARAMETER TUNING**

Dopo aver addestrato il modello sul Training-Set, in questa fase ho ricercato la combinazione ottimale di parametri verificando come al variare del parametro kernel e al variare dei relativi cost, gamma e degree variasse l’accuratezza del modello. Da questa analisi ho notato che nel caso si usasse un kernel di tipo **radiale** si otteneva un’accuratezza migliore. In particolare, con **gamma>3** e **C qualsiasi** si ottiene l’accuratezza massima di 0.835. Pertanto, questi saranno i parametri scelti per predire il Test-Set e valutarne la performance.

**VALUTAZIONE DELLA PERFORMANCE**

Per valutare le performance del modello una volta scelti gli iperparametri, ho utilizzato misure di errore come:

* Errore di misclassificazione = 31
* tasso di misclassificazione = 0.155
* l’accuratezza = 0.845

Inoltre, in modo da definire i punti di forza e debolezza del modello, ho utilizzato la matrice di confusione dalla quale ho potuto capire come il modello, quando la qualità è:

 Vero **0:**

* Predice come 0: 5 volte (corretto)
* Predice come 1: 25 volte (errore)
* Predice come 2: 0 volte (corretto)

 Vero **1:**

* Predice come 0: 0 volte (corretto)
* Predice come 1: 162 volte (corretto)
* Predice come 2: 0 volte (corretto)

 Vero **2:**

* Predice come 0: 0 volte (corretto)
* Predice come 1: 6 volte (errore)
* Predice come 2: 0 volte (errore)

Ciò evidenzia come il modello sia forte nel predire la qualità di tipo 1, dovuto al fatto che ne sono presenti in quantità maggiori, quindi addestrato meglio su di esse, ma debole nel predire le qualità di tipo 0 e 2.

Immagine che contiene testo, schermata, diagramma, Rettangolo

Descrizione generata automaticamente

**STUDIO STATISTICO SUI RISULTATI DELLA VALUTAZIONE**

In questa fase, ripetendo le fasi di addestramento e testing per 50 volte, ho ottenuto dei SRS (50) per le metriche di errore di misclassificazione, tasso di misclassificazione e accuratezza:

1. **Errore di misclassificazione**:

* Media: 27.4200
* Deviazione Standard: 4.5315
* Min: 19.0000
* Primo quartile: 24.2500
* Mediana (Secondo Quartile): 27.0000
* Terzo quartile: 29.7500
* Max: 38.0000
* Range Interquartile: 5.50
* Intervallo di confidenza della media al 95%: (26.163961680649862, 28.67603831935014)

Immagine che contiene diagramma, linea, Diagramma, testo

Descrizione generata automaticamente Immagine che contiene testo, schermata, linea, diagramma

Descrizione generata automaticamente

1. **Tasso di misclassificazione**:

* Media: 0.1371
* Deviazione Standard: 0.0227
* Min: 0.0950
* Primo quartile: 0.1212
* Mediana (Secondo Quartile): 0.1350
* Terzo quartile: 0.1487
* Max: 0.1900
* Range Interquartile: 0.03
* Intervallo di confidenza della media al 95%: (0.1308198084032493, 0.1433801915967507)

Immagine che contiene diagramma, linea, Diagramma, testo

Descrizione generata automaticamente Immagine che contiene testo, schermata, linea, diagramma

Descrizione generata automaticamente

1. **Accuratezza**:

* Media: 0.8629
* Deviazione Standard 0.0227
* Min: 0.8100
* Primo quartile: 0.8512
* Mediana (Secondo Quartile): 0.8650
* Terzo quartile: 0.8788
* Max: 0.9050
* Range Interquartile: 0.03
* Intervallo di confidenza della media al 95%: (0.8566198084032494, 0.8691801915967509)

Immagine che contiene diagramma, linea, Diagramma, testo

Descrizione generata automaticamente Immagine che contiene testo, schermata, linea, diagramma

Descrizione generata automaticamente