Proiect Analiza și procesarea dateleor prin tehnici de Învățare Automată

Wine Quality

# Introducere

Acest proiect are ca scop analiza unui set de date despre vinuri și construirea unor modele de predicție și clustering. Proiectul include utilizarea unei regresii liniare pentru a prezice calitatea vinului, precum și utilizarea algoritmului de clustering K-means pentru a grupa vinurile în funcție de caracteristicile lor.

## Setul de date

Setul de date utilizat în acest proiect conține informații despre caracteristicile chimice ale diferitelor vinuri, precum și scorurile de calitate atribuite acestora. Variabilele sunt:

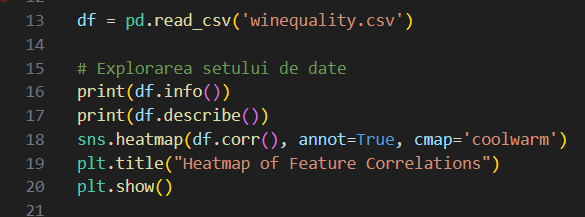
* **Input Variables** (bazate pe teste fizico-chimice):
  + Fixed Acidity
  + Volatile Acidity
  + Citric Acid
  + Residual Sugar
  + Chlorides
  + Free Sulfur Dioxide
  + Total Sulfur Dioxide
  + Density
  + pH
  + Sulphates
  + Alcohol
* **Output Variable** (bazată pe date senzoriale):
  + **Quality** (scor între 0 și 10)

# Pași de elaborare a proiectului

## 1. Analiza Exploratorie a Datelor

Primul pas al proiectului a fost analiza exploratorie a datelor pentru a înțelege mai bine structura și relațiile dintre variabile. S-au utilizat funcții precum info() și describe() pentru a obține informații generale și statistici descriptive.

În plus, s-a folosit o diagramă de corelație (heatmap) pentru a vizualiza relațiile dintre variabile. Acest lucru ne-a permis să identificăm ce variabile sunt cele mai puternic corelate cu scorul de calitate al vinului.



Acest cod a fost utilizat pentru a explora structura setului de date și pentru a vizualiza corelațiile dintre variabilele chimice ale vinului.

A screenshot of a game

Description automatically generated

## Curatarea si prepocesarea datelor

Pentru a ne asigura că setul de date este pregătit pentru modelare, s-au verificat valorile lipsă și s-au eliminat rândurile incomplete. În acest pas s-au folosit funcțiile *isnull().sum()* pentru identificarea valorilor lipsă și *dropna()* pentru eliminarea acestora.

A screen shot of a computer code

Description automatically generated

## Împărțirea datelor în seturi de antrenament și testare

Setul de date a fost împărțit în variabile independente (X) și variabila dependentă (y), unde y reprezintă scorul de calitate. Apoi, setul a fost împărțit în date de antrenament și testare utilizând train\_test\_split(). Acest pas este esențial pentru a asigura evaluarea corectă a modelului pe date noi.

A screen shot of a computer

Description automatically generated

Acest cod împarte setul de date în seturi de antrenament și testare, necesare pentru evaluarea performanței modelului.

## Construirea modelului de regresie liniară

Pentru a prezice calitatea vinului, s-a construit un model de regresie liniară folosind biblioteca *sklearn.* Modelul a fost antrenat pe setul de date de antrenament pentru a învăța relația dintre caracteristicile vinului și scorul său de calitate.

A computer screen with colorful text

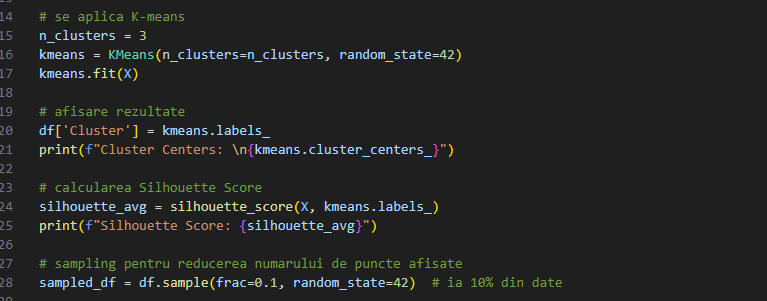
Description automatically generated

Acest cod construiește și antrenează un model de regresie liniară, apoi calculează eroarea medie pătratică (MSE) și scorul R2 pentru a evalua performanța modelului.

## Aplicarea Algoritmului K-means pentru clustering

Algoritmul de clustering K-means a fost aplicat pentru a grupa vinurile în funcție de caracteristicile lor. Acest lucru permite identificarea unor grupuri similare de vinuri care au caracteristici comune. Numărul de grupuri (n\_clusters) a fost setat la 3, iar modelul K-means a fost antrenat pe întregul set de date.

Acest cod aplică algoritmul K-means pentru a grupa vinurile și utilizează Silhouette Score pentru a evalua calitatea grupărilor.



Silhouette Score este o metrica utilizata pentru a evalua calitatea gruparilor generate de K-means. Valorile variaza intre **-1** si **1**. Se calculeaza distanta medie in interiorul clusterului, notata cu **a**, distanta medie pana la toate punctele din cel mai apropiat cluster vecin, notata cu **b**. S reprezinta media valorilor pentru toate punctele din setul de date.

## Vizualizarea și Interpretarea Grupărilor

S-a creat o diagramă de dispersie (scatterplot) folosind seaborn pentru a vizualiza grupările rezultate. În această vizualizare, grupurile sunt diferențiate prin culori, pentru a observa cum vinurile sunt distribuite în funcție de anumite caracteristici, cum ar fi concentrația de alcool și aciditatea volatilă.

A screen shot of a computer code

Description automatically generated

Blocul de cod de mai sus vizualizează grupările rezultate în urma algoritmului K-means, pentru a înțelege mai bine distribuția vinurilor.

## Salvarea rezultatelor

Prin intermediul liniei de cod:

**df.to\_csv('/mnt/data/winequality\_with\_clusters.csv', index=False)**

se salveaza rezultatele obtinute.

A screen shot of a computer screen

Description automatically generated

Graficul reprezinta rezultatele clusteringului realizat cu algoritmul K-means pe un set de date esantionate. Fiecare punct reprezinta un exemplu din setul de date ( o observatie unica despre vin).

Culoarea punctului indica clusterul din care face parte, determinat de algoritmul K-means. De exemplu, cu mov este reprezentat clusterul 0, cu galben clusterul 1 iar cu turcoaz clusterul 2.

Punctele rosii reprezinta pozitiile centroidelor fiecarui cluster. Un centroid este punctul mediu al tuturor punctelor dintr-un cluster si serveste ca “ reprezentant” al acelui grup.

Pentru intelegerea cu usurinta a distributiei, a fost folosita medoa **.sample()** pentru a afisa doar 10% din date.



**MSE** masoara eroarea medie patratica intre valorile reale y\_test si cele prezise y\_pred. Este o metrica care indica cat de bine se potrivesc predictiile modelului cu valorile reale. O valoare mai mica a MSE indica faptul ca modelul face predictii mai precise.

**R² Score (R Squared)** indica proportia variabilitatii din datele reale care este explicate de modelul de regresie. Este o valoare cuprinsa intre 0 si 1, 1 reprezinta faptul ca modelul explica perfect variabilele datelor iar 0 reprezinta opusul. Valoarea reprezentata mai sus, dupa rularea programului, sugereaza faptul ca modelul explica aproximativ 26.5% din variabilitatea valorilor reale.

# Concluzii

Acest proiect a demonstrat modul în care analiza datelor și tehnicile de învățare automată pot fi utilizate pentru a prezice și grupa vinurile în funcție de caracteristicile lor. Modelul de regresie liniară a permis predicția calității vinului, iar clusteringul K-means a oferit o perspectivă asupra grupărilor similare de vinuri.

Pentru îmbunătățirea modelului de predicție, se pot explora alte algoritme, precum regresia logistică sau arborii de decizie. De asemenea, numărul de grupuri în clustering poate fi optimizat utilizând metoda „elbow” pentru a determina valoarea optimă a n\_clusters.

# Idei pentru implementari si optimizari viitoare

## Optimizarea Clusteringului K-means

* **Metoda „Elbow”** pentru a determina numărul optim de grupuri (n\_clusters) pentru algoritmul K-means. Astfel, se pot identifica numărul de grupuri care minimizează distanța internă fără a crește inutil complexitatea.
* **Algoritmi Alternativi de Clustering**: Agglomerative Clustering sau DBSCAN, pentru a vedea dacă aceștia oferă o separare mai bună a grupurilor de vinuri.

## Integrarea unei Interfețe Utilizator

* **Aplicație Web**: constructia unei aplicatii web simplă utilizând Flask sau Streamlit pentru a permite utilizatorilor să încarce datele despre vinuri și să obțină predicții privind calitatea acestora.
* **Dashboard pentru Vizualizare**: folosirea unei librării precum Dash sau Tableau pentru a crea un dashboard interactiv care să prezinte vizualizările datelor și rezultatele clusteringului.