WINE QUALITY PREDICTION USING MACHINE LEARNING

A Comprehensive Study of Chemical Properties and Clustering Techniques

INFO6105 – DATA SCIENCE ENGINEERING METHODS AND TOOLS
BHUVANESHWARI KALVAKUNTLA
NORTHEASTERN UNIVERSITY

PROBLEM STATEMENT & MOTIVATION

Traditional wine quality assessment depends on human tasters, which is subjective, inconsistent, and not scalable.
There is a need for a data-driven, objective method to classify wine quality using measurable chemical properties.

The idea of predicting something sensory like wine taste using only chemical attributes was fascinating.
It gave me the chance to explore ensemble learning models, clustering, and visualization techniques in a meaningful way.
The dataset was well-structured and challenging enough to build a comprehensive ML pipeline

METHODS & TECHNIQUES USED

Preprocessing. - Cleaned and structured a dataset of 1,599 red wine samples created binary target:
Exploratory Data Analysis (EDA)

- Correlation heatmap: Found strong relationships with alcohol and sulphates
- Boxplot: Visualized alcohol content by quality score
- Category distribution: Showed imbalance toward Low-quality wines

Machine Learning Models Used

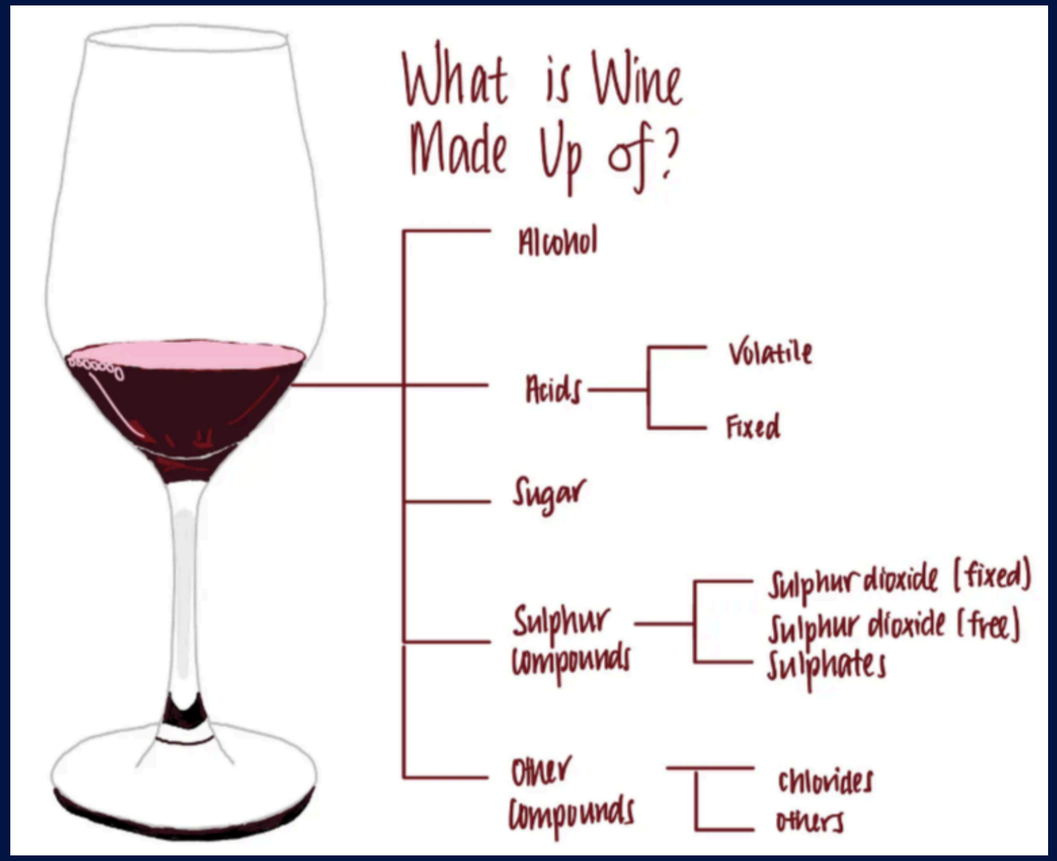
- Gradient Boosting (GBM)
- XGBoost
- Random Forest

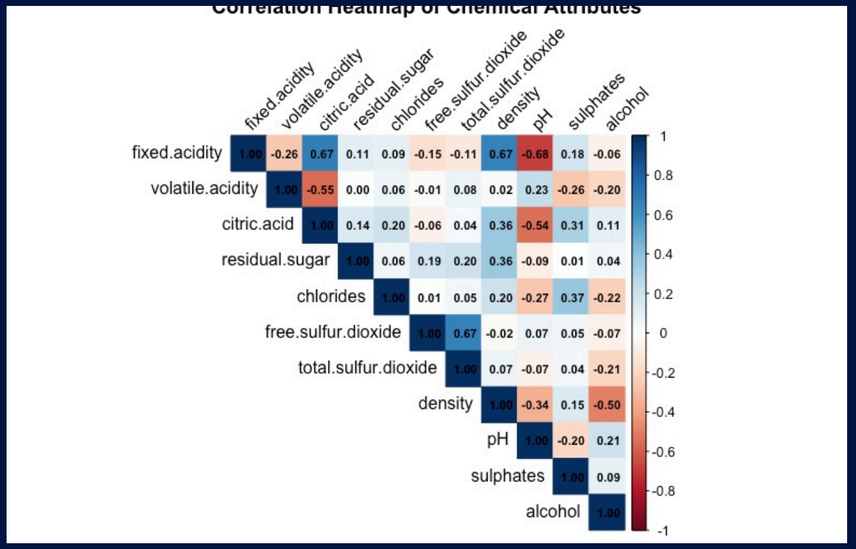
Evaluation Metrics:

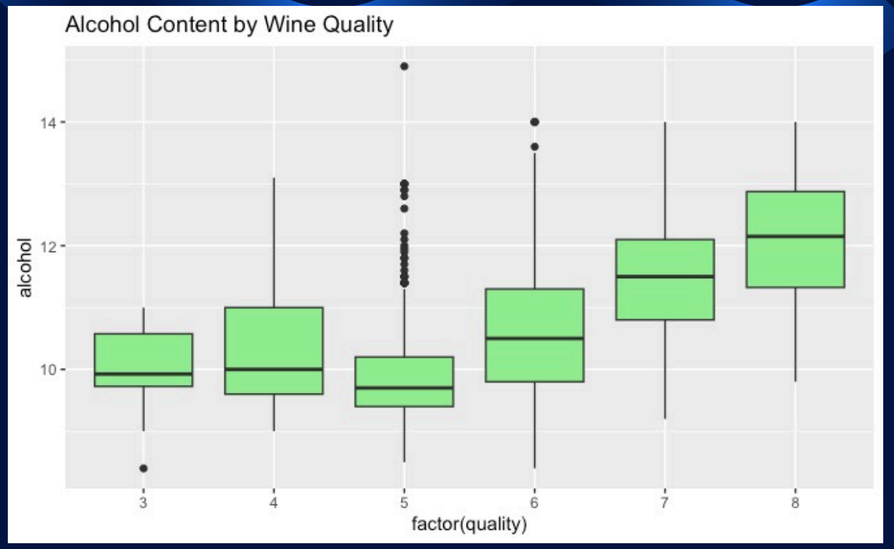
- Accuracy
- Precision
- Recall
- F1-score

Unsupervised Learning

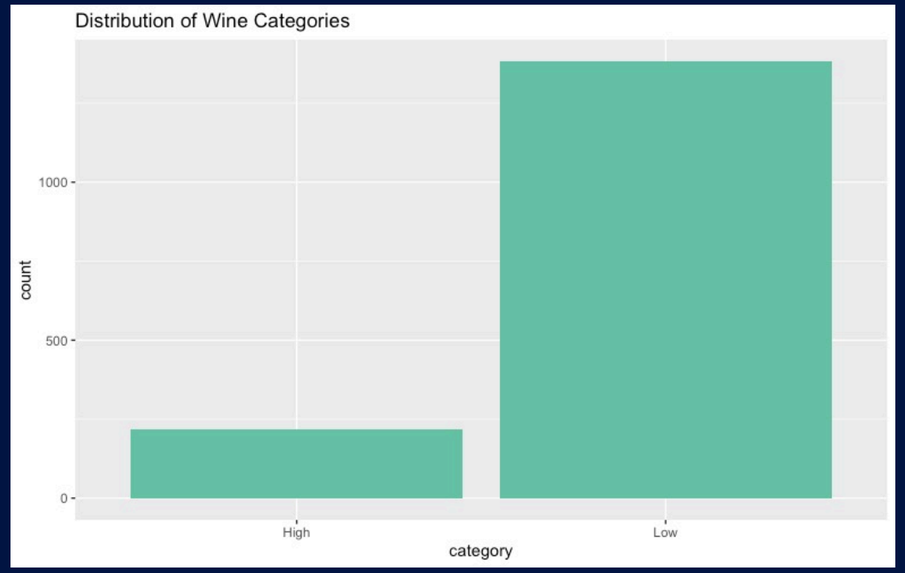
- KMeans Clustering to find natural wine groupings
- PCA for visualizing clusters in 2D

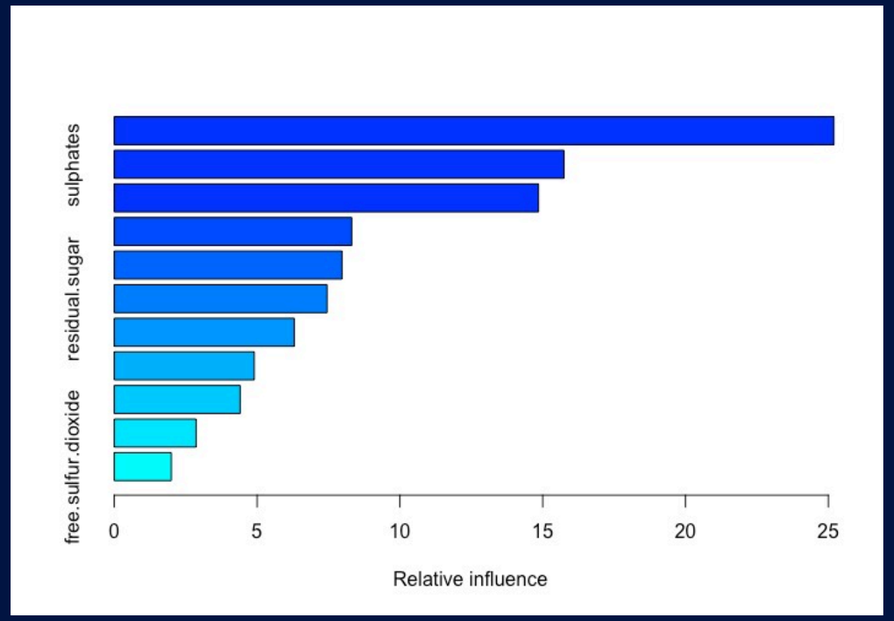


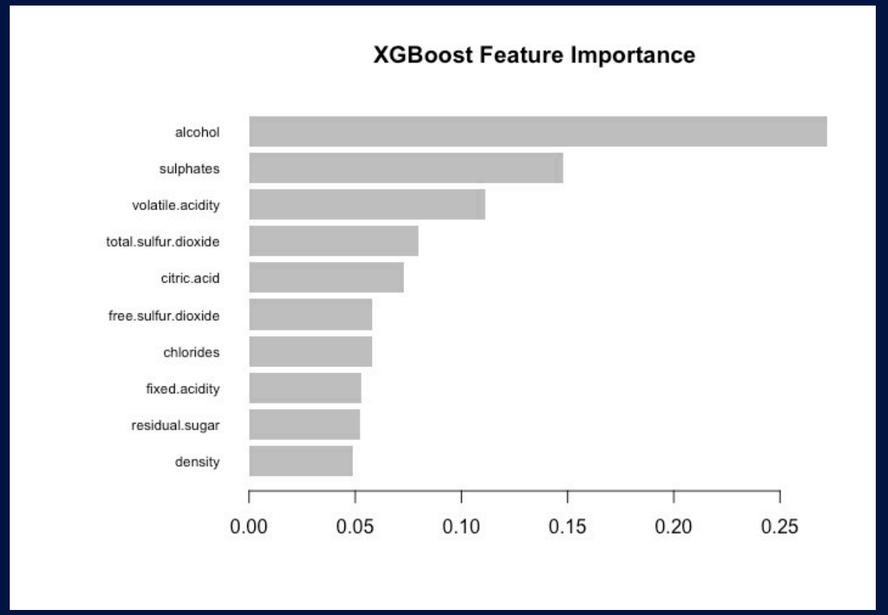




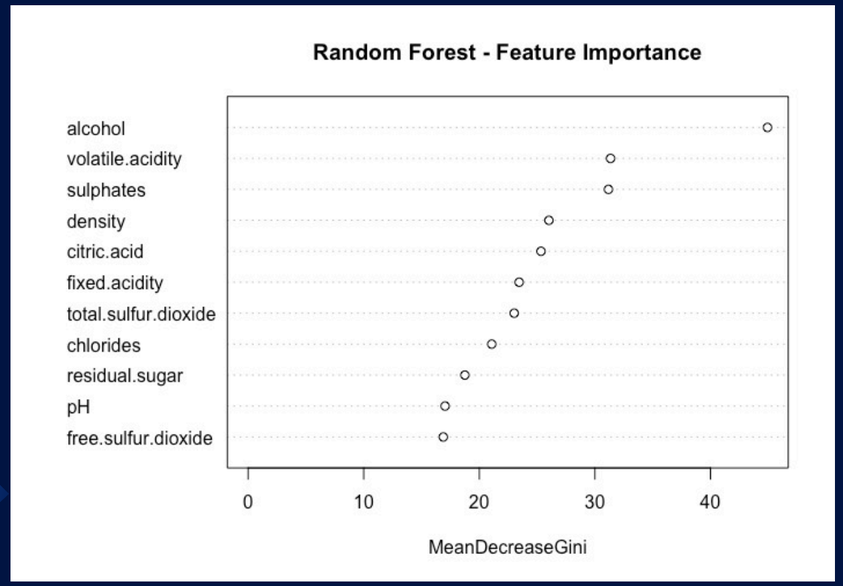
ALCOHOL AND SULPHATES ARE THE STRONGEST POSITIVE PREDICTORS OF QUALITY, WHILE VOLATILE ACIDITY IS NEGATIVELY ASSOCIATED.

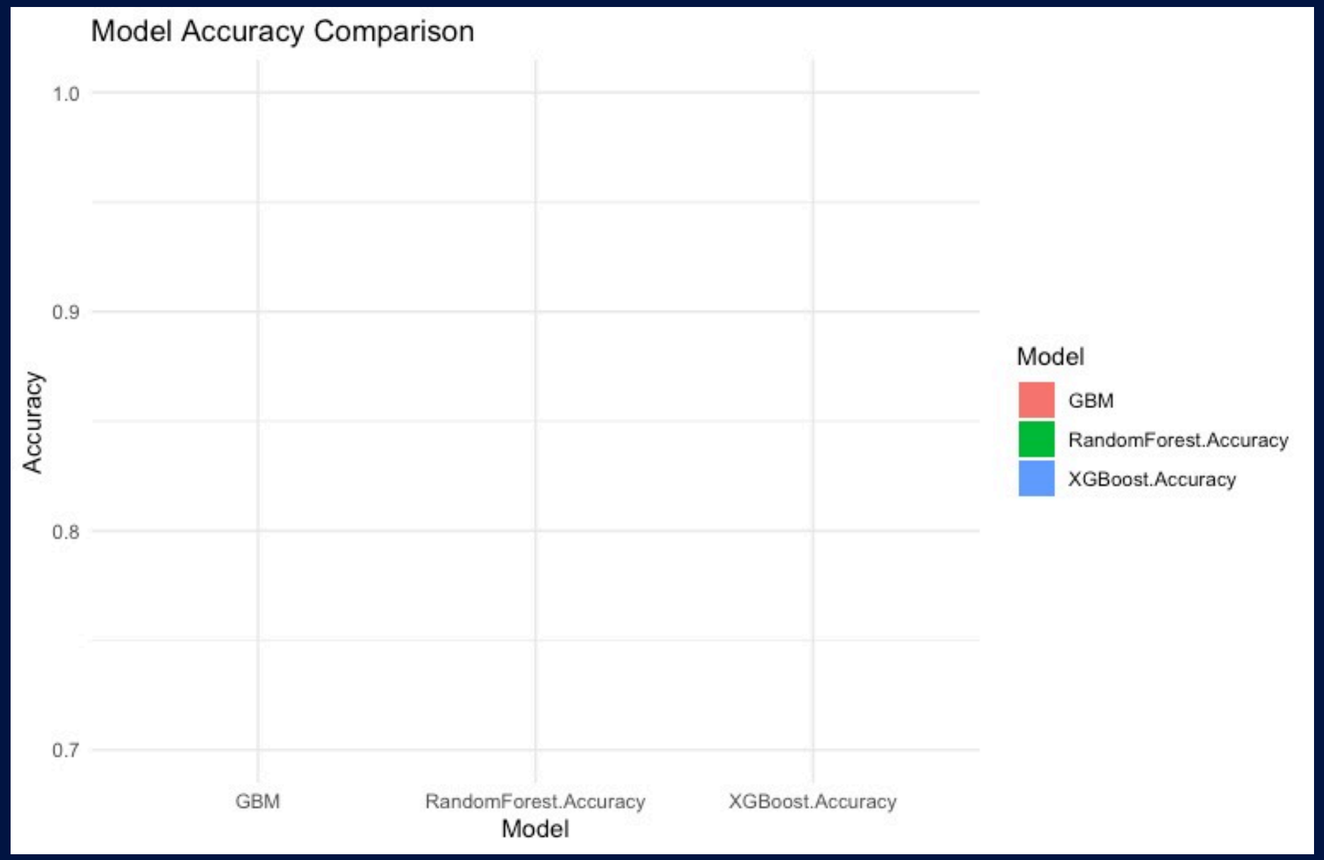




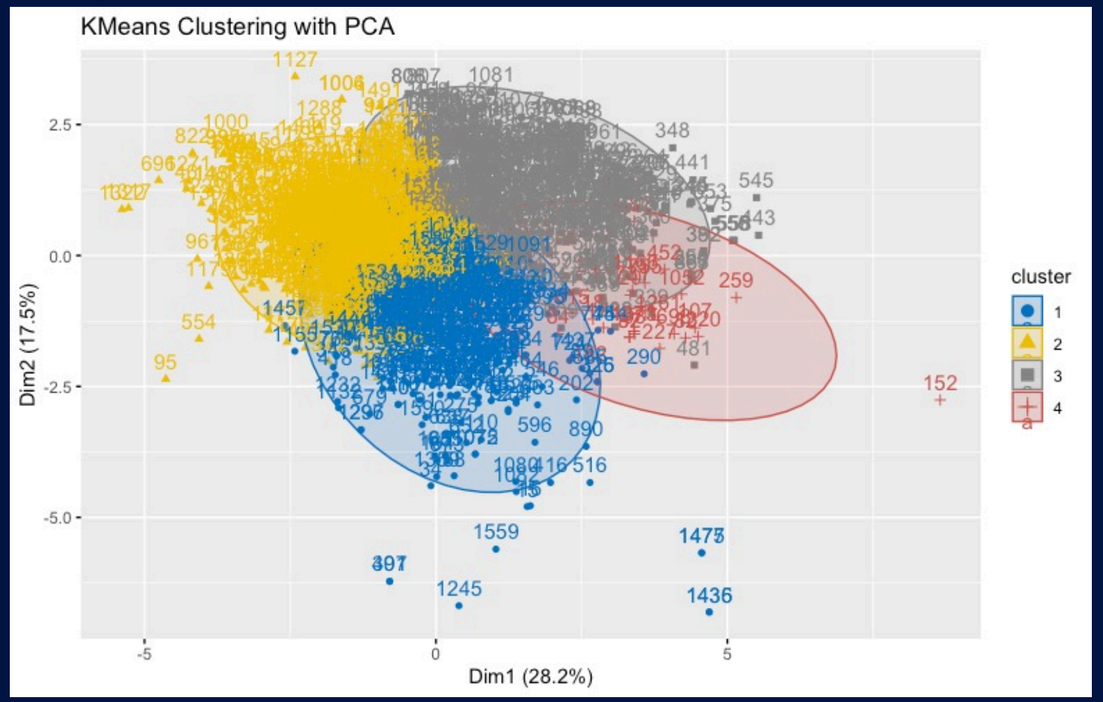


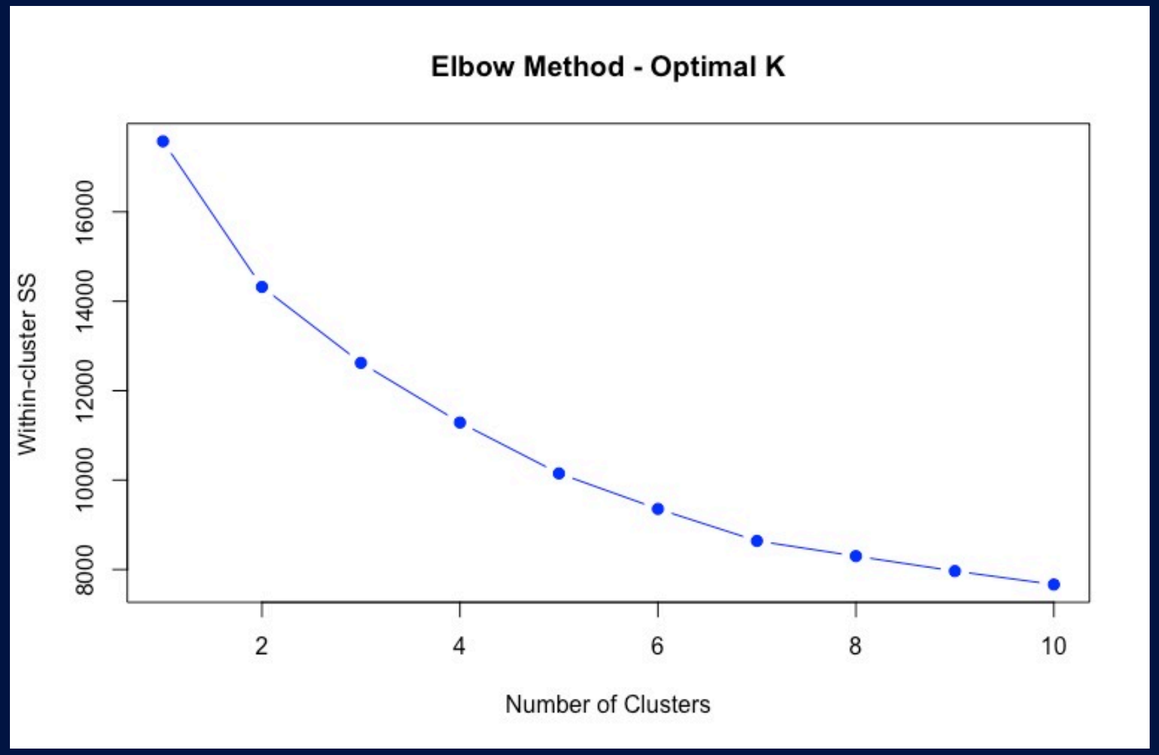
GRADIENT BOOSTING, XGBOOST, AND RANDOM FOREST — TO CLASSIFY WINE QUALITY. DATA WAS SCALED, THE TARGET VARIABLE WAS BINARIZED, AND MODELS WERE EVALUATED USING STANDARD CLASSIFICATION METRICS. RANDOM FOREST PERFORMED BEST OVERALL.





RANDOM FOREST ACHIEVED THE BEST RESULTS WITH AN ACCURACY OF 91.7%, FOLLOWED BY XGBOOST. GBM WAS SLIGHTLY LOWER AT 87%. ACROSS ALL MODELS, ALCOHOL AND SULPHATES WERE KEY PREDICTORS, WHILE VOLATILE ACIDITY NEGATIVELY IMPACTED WINE QUALITY





THE ELBOW METHOD INDICATED THAT 4 CLUSTERS GAVE THE BEST SEPARATION. PCA REDUCED THE DATA TO 2D SO I COULD VISUALIZE THESE CLUSTERS. WINES IN THE SAME GROUP SHOWED SIMILAR CHEMICAL PROPERTIES, LIKE ALCOHOL AND ACIDITY LEVELS.

CONCLUSION

THIS PROJECT SUCCESSFULLY DEMONSTRATED THE USE OF MACHINE LEARNING TO CLASSIFY RED WINE QUALITY USING CHEMICAL ATTRIBUTES.

AMONG THE MODELS TESTED, RANDOM FOREST ACHIEVED THE HIGHEST ACCURACY (91.7%), FOLLOWED BY XGBOOST AND GBM.

KEY PREDICTORS OF WINE QUALITY INCLUDED ALCOHOL, SULPHATES, AND VOLATILE ACIDITY.

CLUSTERING AND PCA FURTHER REVEALED NATURAL GROUPINGS AMONG WINES, COMPLEMENTING THE CLASSIFICATION TASK.

FUTURE SCOPE

- INCORPORATE SENSORY FEATURES LIKE TASTE, AROMA, AND COLOR
- TEST WITH LARGER AND MORE DIVERSE WINE DATASETS
- EXPLORE DEEP LEARNING MODELS OR AUTOML FOR ENHANCED PERFORMANCE AND TUNING