# *AMS 561 - Project Presentation*
# A STUDY ON RED WINE QUALITY

By: Thi-Thanh-Lam Vu & Sangjun Hong & Xiangji Li

May 2025

Stony Brook University

# Introduction

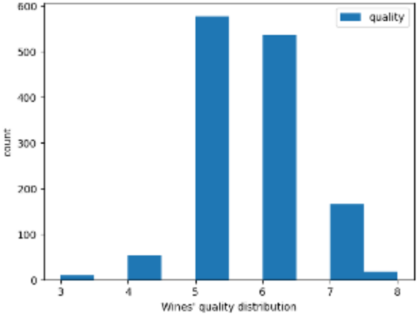
- Wine quality is assessed using both physicochemical measurements and sensory evaluations.
- Physicochemical tests analyze properties such as density, alcohol content, and pH.
- Sensory evaluation is subjective, and the connection between taste and chemistry remains unclear.
- Project's objectives:
  - Analyze how physicochemical properties influence wine quality
  - Develop an accurate classification model for wine quality
  - Identify clusters of wines based on their chemical profiles

# Techniques and Tools

- Data pre-processing: Remove outliners and duplicates.
- Wines' quality classification
  - Machine Learning Models: K-Nearest Neighbors (KNN), Gradient Boosting Machines (GBM), Decision Trees, Support Vector Machines (SVM), Random Forests, and Neural Networks.
  - Techniques to handle imbalanced dataset
    - SMOTE: a sampling method to generate new synthetic samples.
    - Principal component analysis (PCA): a dimensionality reduction method
    - Weighted Cross-Entropy Loss and Focal Loss: add weights to each class
    - One-vs-All Training: adapt binary classifiers to handle multi-class classification problems.
- Wines clustering:
  - K-means and Elbow method.
  - Louvain clustering: a popular algorithm for community detection in networks.

# Data Exploration and Analysis

- 11 associate physicochemical properties were measured, such as pH level, alcohol percentage, wine density, chlorides,..
- Wines are graded from 1 to 10.




- The dataset is highly imbalanced.

- There are some trends in the relationship between wine quality and measured variables.

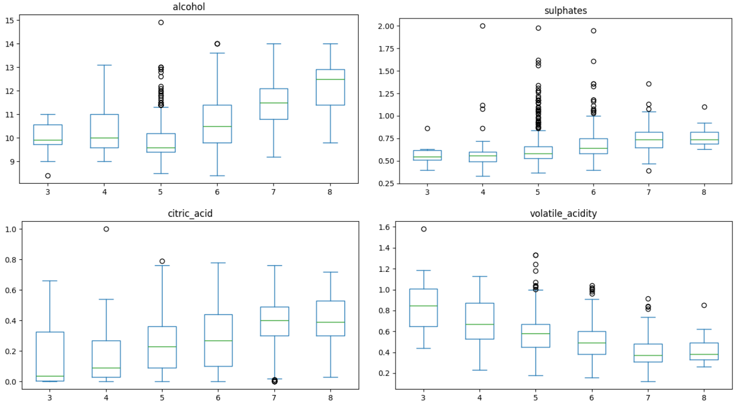
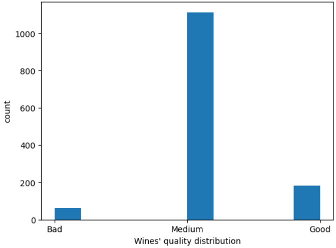



Figure: Boxplot the distribution of some variables based on wines' quality

# Wines' quality classification

- We divide wines into 3 groups: Bad, Medium, and Good.




- PCA does not improve the classification performance.
- SMOTE generally improves the performance of classical machine learning models.

| | KNN | GBM | Decision Trees | SVM | Random Forests |
|---|---|---|---|---|---|
| Original | 0.47 | 0.53 | 0.51 | 0.50 | 0.49 |
| SMOTE | 0.48 | 0.54 | 0.49 | 0.58 | 0.61 |

Figure: Macro F1-score of classical machine learning models

# Wines' quality classification

- SMOTE improves the classification performance on minority classes.

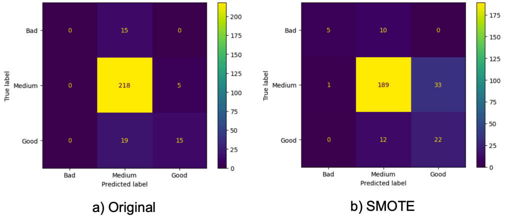



Figure: Confusion matrix of Random Forests classifier

- The neural network and its variants outperform the classical classifier when training with SMOTE data.

| Baseline | Weighted CE Loss | Focal Loss | One-vs-All |
|---|---|---|---|
| 0.61 | 0.64 | 0.65 | 0.68 |

Figure: Macro F1-scores of Neural Network variants

- Adding weights to loss function does improve the performance of Neural Networks.
- One-vs-All training achieves the most balanced results, significantly improving accuracy across all classes.

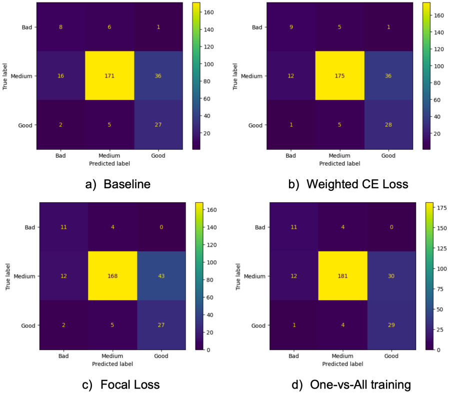
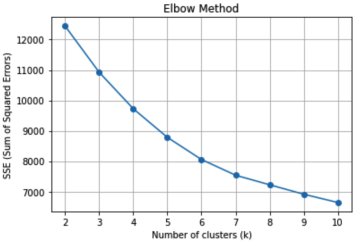


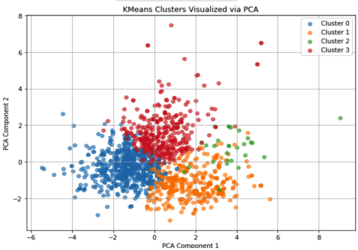

Figure: Confusion matrices of variants of NNs

# Wines clustering

- K-mean fails to capture non-linear or high-dimensional structures.




(a) Error graph for elbow method




(b) Clustering based on K-mean

# Wines clustering

- Louvain clustering generated 4 clusters. Louvain gives a modularity (0.2228)

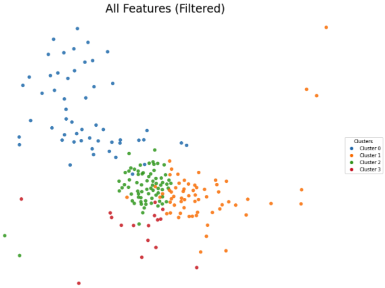



Figure: Clustering based on Louvain

# End!

THANK YOU!