Al-Aqsa University
Faculty of Engineering
Intelligence Systems & Computer Engineering Dept.

جامعة الأقصى
كلية الهندسة
قسم هندسة النظم الذكية و الحاسوب

Pattern Recognition (ENGS3302)

Task 2: Supervised Learning Classifiers – Wine Quality Detection Report

Presented to: Eng. Ibraheem k. Shehada

Student name: **Samar M. Balousha**
Student number: **2320222601**

Link to Github

February 8, 2026

Semester: First Semester 2025/2026

Report: Wine Quality Classification Using SVM and XGBoost

a. Introduction

The objective of this project is to apply Pattern Recognition (PR) techniques to classify wine quality based on its physicochemical properties. In the wine industry, quality assessment is traditionally subjective; however, machine learning offers an objective approach by analyzing chemical components such as alcohol content, acidity, and pH levels.

This work evaluates two primary classification paradigms: **Support Vector Machines (SVM)**, exploring both Hard and Soft Margin configurations, and **XGBoost**, an advanced ensemble learning method. By comparing these models, we aim to determine the most effective boundary-seeking algorithm for chemical dataset classification.

b. Dataset Preparation

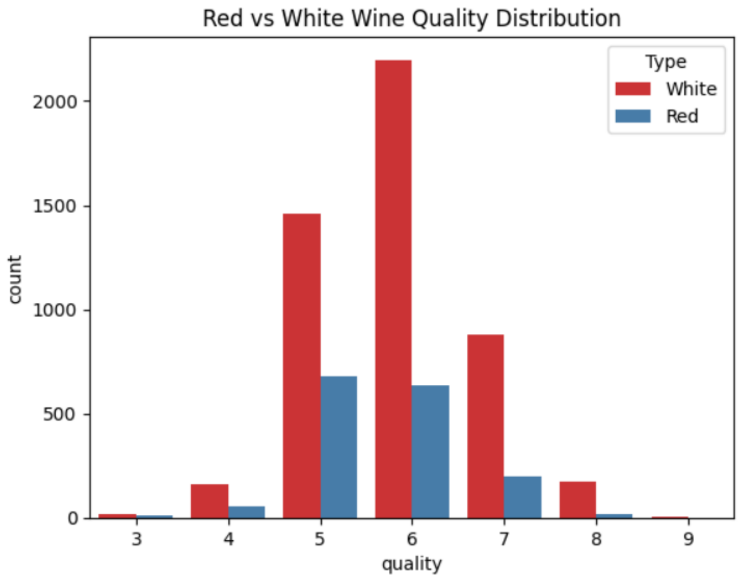
The study utilizes the UCI Wine Quality dataset, consisting of two subsets: red wine and white wine.

1. Data Integration and Cleaning

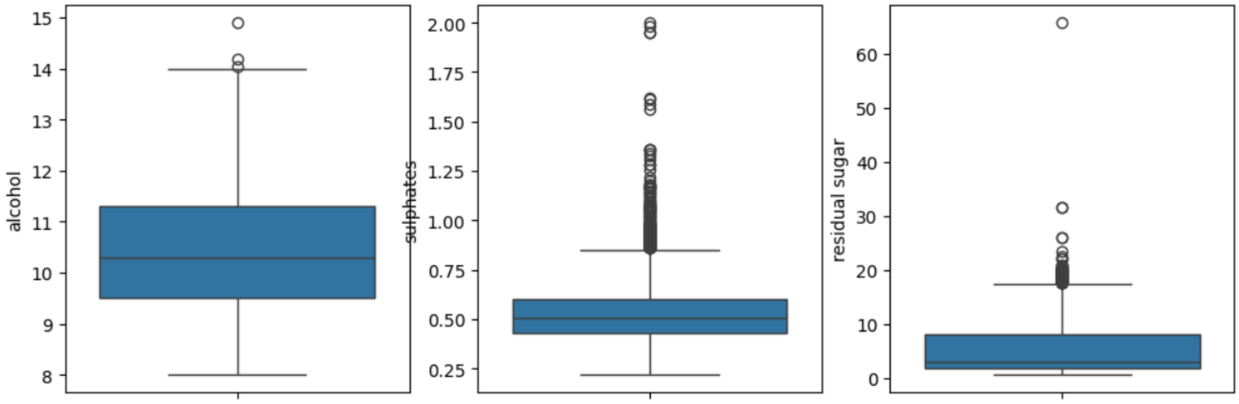
- **Merging:** To create a robust model, the red and white wine datasets were combined into a single dataframe.
- **Feature Engineering:** A "Type" feature (Red=1, White=0) was added to preserve the categorical distinction between the two wine varieties.
- **Filtering:** Samples with quality scores of 3 and 9 were removed. These classes were statistically rare, representing noise that could prevent the model from learning a generalized pattern.

2. Exploratory Data Analysis (EDA)

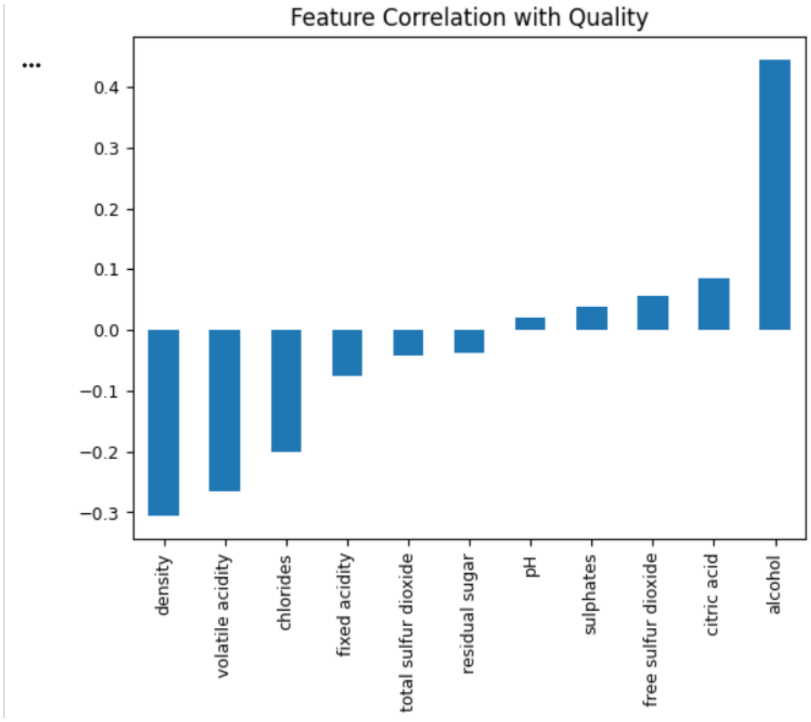
- **Class Distribution:** EDA revealed that the dataset is imbalanced, with most wines rated as 5 or 6.



- **Outlier Analysis:** Using Box Plots, outliers were identified in features like "residual sugar" and "chlorides." These were retained to maintain the natural variance of the chemical data.



- **Correlation:** A correlation heatmap showed that **Alcohol** has the strongest positive correlation with wine quality.



3. Data Preprocessing Steps

- **Target Transformation:** The target variable was converted to binary classification: **'Good' = 1 (Quality ≥ 6)** and **'Not Good' = 0 (Quality < 6)**.

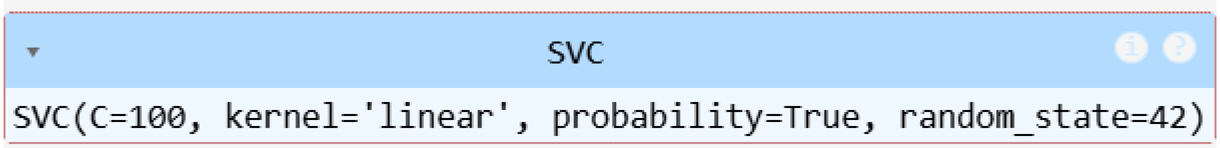
- **Standardization:** Since SVM is a distance-based classifier, all features were scaled using StandardScaler to ensure a mean of 0 and a variance of 1.

- **Splitting:** The data was partitioned into **80% training** and **20% testing** sets.

---

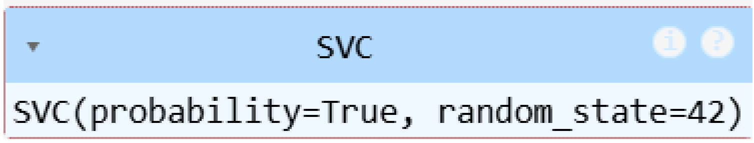
c. Classifiers' Description

1. Hard Margin SVM
The Hard Margin SVM seeks a hyperplane that perfectly separates the classes without allowing any misclassifications. In this project, this was simulated using a very high penalty parameter (C = 100). This model assumes that the data is linearly separable and does not tolerate noise.



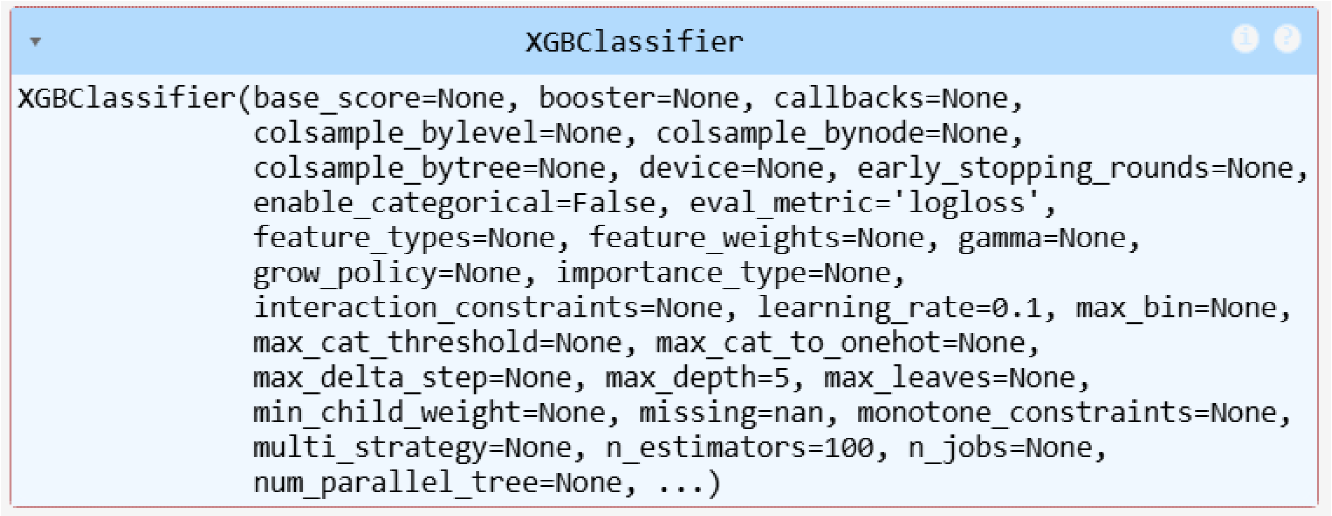
```
SVC(C=100, kernel='linear', probability=True, random_state=42)
```

2. Soft Margin SVM
Recognizing that real-world chemical data often overlaps, the Soft Margin SVM introduces "slack variables." This allows for some misclassifications to achieve a wider margin and better generalization. A lower C value (C = 1.0) with an RBF kernel was utilized to handle non-linear boundaries.



```
SVC(probability=True, random_state=42)
```

3. XGBoost Classifier
XGBoost (Extreme Gradient Boosting) is an ensemble method that builds decision trees sequentially. Each new tree corrects the errors of the previous ones. Unlike SVM, it is non-parametric and highly effective at capturing complex interactions between features without requiring a specific kernel.



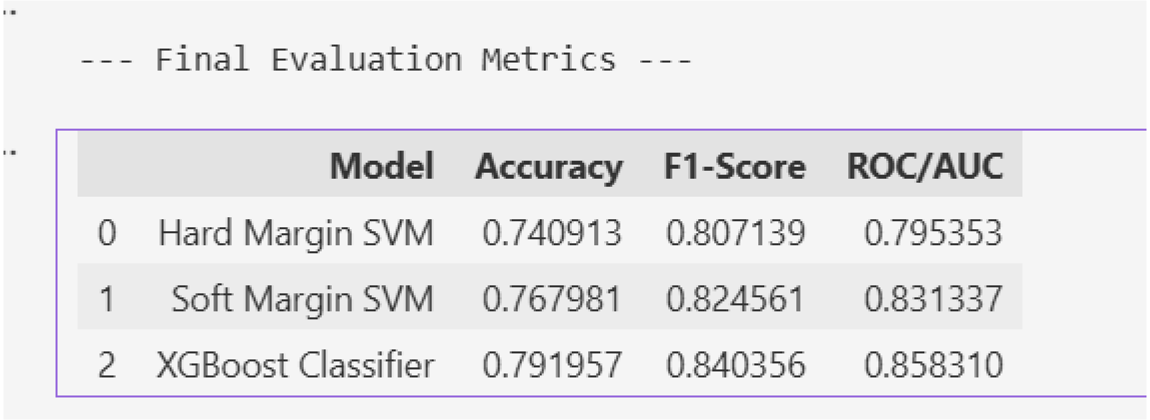
```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric='logloss',
              feature_types=None, feature_weights=None, gamma=None,
              grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=0.1, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=5, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=100, n_jobs=None,
              num_parallel_tree=None, ...)
```

---

d. Results and Discussion

1. Performance Metrics Summary

The table below summarizes the evaluation of the three classifiers. **XGBoost** outperformed both SVM variants across all metrics.
The models were evaluated based on Accuracy, F1-Score (to account for class imbalance), and the Area Under the ROC Curve (AUC).

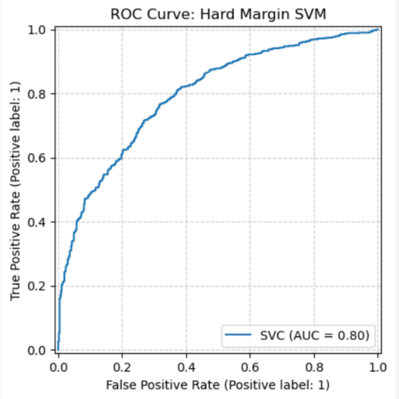


```
--- Final Evaluation Metrics ---
```

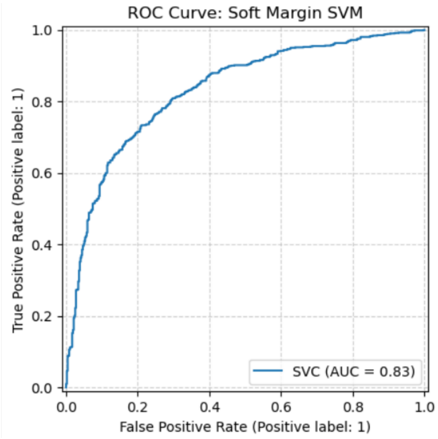
	Model	Accuracy	F1-Score	ROC/AUC
0	Hard Margin SVM	0.740913	0.807139	0.795353
1	Soft Margin SVM	0.767981	0.824561	0.831337
2	XGBoost Classifier	0.791957	0.840356	0.858310

2. Visual Comparison (ROC Curves)

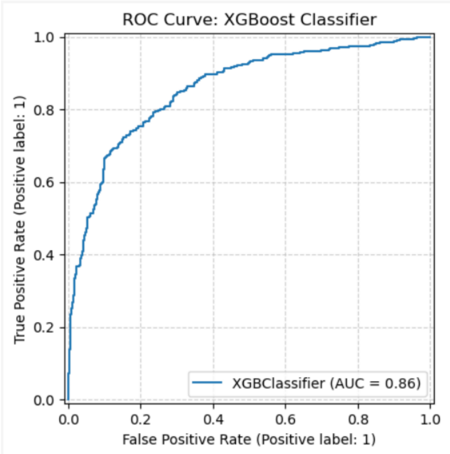
- **Curve - Hard Margin SVM**



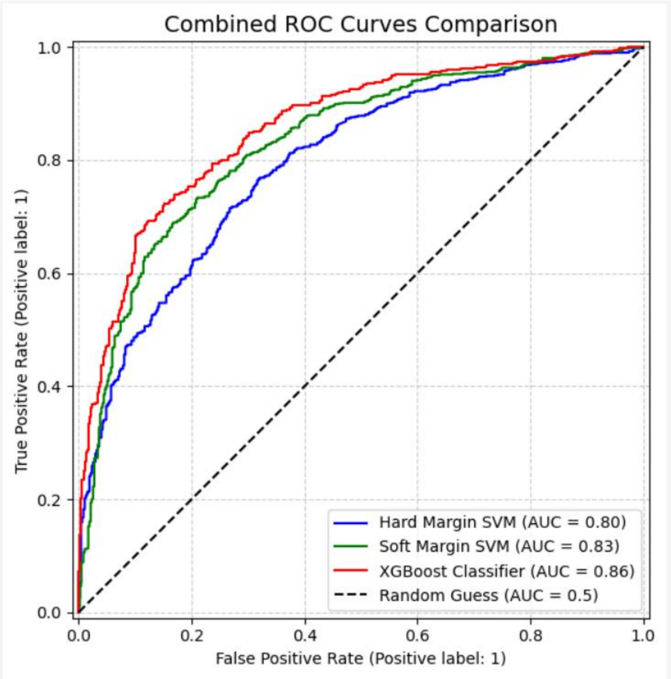
- **ROC Curve - Soft Margin SVM**



- ### ROC Curve – XGBoost



3. Discussion



- The **Hard Margin SVM** showed poor performance on the wine quality dataset. This is mainly because the model assumes perfectly separable data, while the dataset contains noise and overlapping classes. As a result, training was slow, and the model failed to generalize well.

- The **Soft Margin SVM** performed better by allowing margin violations, which helped it handle noisy samples more effectively. This flexibility improved both training stability and prediction accuracy compared to the Hard Margin approach.

- **XGBoost** achieved the best results among all models. Its ability to capture non-linear relationships and focus on hard-to-classify samples made it well suited for the wine quality dataset. The boosting mechanism helped reduce errors and improve overall robustness.

- Overall, **XGBoost outperformed both SVM models**, while Soft Margin SVM provided a reasonable compromise between complexity and performance. These results emphasize the importance of selecting models that align with real-world data characteristics.

e. Conclusion

The study concludes that **XGBoost** is the most robust model for wine quality classification. In terms of Pattern Recognition, the experiment confirms that **Soft Margin** approaches are vital for real-world datasets where class boundaries are inherently noisy and overlapping.