Wine Quality Dataset - Ordinal Analysis

Thiago Vazquez
Master of Science Program in Data Science
University Of Delaware
Newark, US
thiagomv@udel.edu

Ayush Jagani
Master of Science Program in Data Science
University Of Delaware
Newark, US
jayushj@udel.edu

Manisha Kandel
Master of Civil, Construction and Environmental Engineering
University Of Delaware
Newark, US
mkandel@udel.edu

Pratyush
Master of Science Program in Robotics
University Of Delaware
Newark, US
probot@udel.edu

***Abstract*—Our project explores the application of ordinal classification techniques to predict wine quality using physicochemical attributes from the UCI Wine Quality dataset. Specifically, we implemented and compared Ordinal Logistic Regression and Ordinal Support Vector Machine (SVM) against their non-ordinal counterparts. We applied SMOTE to oversample under-represented classes and ensure more balanced model training, given the inherent class imbalance and ordinal nature of the target variable. We then performed feature importance analysis using a Random Forest classifier, which confirmed all input features were relevant, allowing us to retain the full feature set. Our findings demonstrate that Ordinal SVM outperformed both Standard Multiclass SVM and Ordinal Logistic Regression across all evaluation metrics, indicating the effectiveness of ordinal-aware models for quality ranking tasks. Our results basically highlight the importance of preserving the ordered structure of labels in classification and suggest promising directions for future research in ordinal learning.**

I. Introduction

A. Overview of The Problem:

We have two datasets related to red and white vinho verde wine samples, from the north of Portugal. Our goal is to model wine quality based on physicochemical tests. We will try to determine if Ordinal Classifiers like Ordinal SVM and Logistic regression perform better than their ordinary counterparts and also how Ordinal method's performance compare against each other (i.e. Ordinal SVM vs Ordinal Logistic regression).

B. Dataset Description:

Our Dataset is a twofold (i.e. red and white Wine) Dataset with multivariate characteristics, containing 11 predictive real features and around 6500 (red and white combined) instances. As mentioned in the dataset information [3], the classes are ordered and the dataset is not balanced (e.g. there are many more normal wines than excellent or poor ones). Also, it is unknown if all input variables are relevant, so we will perform feature selection using random forest.

C. Results:

The results show that after performing feature selection and smote for some minority classes, Ordinal SVM gives the best performance, followed by Standard SVM and at last Ordinal logistic regression giving us the worst performance out the of 3.

II. Related work

An article published on ScienceDirect titled "Modeling Wine Preferences by Data Mining from Physicochemical Properties" aimed to predict wine quality using machine learning based on physicochemical attributes such as acidity, alcohol content, and pH [1]. The authors applied traditional classification and regression methods, including decision trees, k-nearest neighbors, support vector machines, and neural networks. In their approach, wine quality ratings (ranging from 0 to 10) were treated either as a regression problem with continuous outputs or as a multi-class classification problem with discrete categories. However, a key limitation of their work was the failure to account for the ordinal nature of the quality scores, where higher ratings are incrementally closer to adjacent values (e.g., a rating of 7 is more similar to 6 than to 3).

Another study titled "An Ordinal Classification Approach for Predicting Student Academic Performance" focused on forecasting student outcomes using machine learning models designed to handle ordinal data [2]. The researchers categorized academic performance into ordered levels such as Fail, Pass, Good, and Excellent, and explored the limitations of treating these as nominal classes. They implemented ordinal logistic regression and support vector machines with ordinal constraints, mainly to better capture the inherent order in the data. These methods were compared to traditional classifiers like Naive Bayes and Decision Trees. The results showed that ordinal models not only improved predictive accuracy but also provided more interpretable results, highlighting the value of preserving order in target variables. This reinforces the importance of using ordinal-specific models for tasks like

wine quality prediction, where scores are ranked rather than purely categorical.

III. Methods

We employed Ordinal Logistic Regression and Ordinal Support Vector Machines (SVM), both of which are well-suited for problems involving ordered categorical responses. Unlike conventional classification algorithms that treat quality scores as nominal classes, ordinal methods explicitly model the natural ordering present in the labels, which is particularly important for our dataset, where (as mentioned earlier) quality ratings follow a ranked scale (e.g., a score of 7 is inherently closer to 6 than to 3). In simple words, the difference between both methods is that Ordinal Logistic Regression models the cumulative probabilities of the response variable, while Ordinal SVM optimizes hyperplanes that respect the order constraints among categories. Below are the objective functions that summarize the two ordinal classifiers we have used.

Ordinal Logistic regression can be defined as:

$$P(y \leq k \mid \mathbf{x}) = \frac{1}{1 + \exp\left(-(\theta_k - \mathbf{x}^\top \boldsymbol{\beta})\right)} = \sigma(\theta_k - \mathbf{x}^\top \boldsymbol{\beta}) \tag{1}$$

Components :

- y: Ordinal response variable (e.g., a Likert scale: low, medium, high)
- k: Category threshold (e.g., the probability that $y \leq k$)
- x: Vector of predictors
- β : Coefficient vector (same across all categories — proportional odds assumption)
- θ_k : threshold parameter (also called cutpoints) specific to each category k.
- $\sigma(z)$: Logistic sigmoid function

Where in (1) "P" represents the cumulative probability function (i.e.CDF) which is equivalent to the sigmoid function in ordinary Logistic Regression

Ordinal SVM is defined as:

$$\min_{\mathbf{w},\{b_k\},\{\xi_i^-,\xi_i^+\}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i^- + \xi_i^+) \tag{2}$$

Components :

- $\mathbf{n}$: Total no. of instances
- $\mathbf{w} \in \mathbb{R}^d$: shared weight vector for all thresholds (like linear SVM)
- b_k: thresholds (or cut-points) between ordinal classes
- ξ_i^-, ξ_i^+ : slack variables to allow for soft margin violations
- C : regularization parameter controlling the trade-off between margin size and training error

(2) represents the Objective function of Ordinal SVM (In Ordinal SVM there are "k-1" thresholds for K classes unlike ordinary SVM which has 1 hyperplane threshold separating 2 classes).

IV. Experiments

A. Data Pre-processing

First we grabbed the Wine Dataset from [3] and combined both red and white wine samples. We converted all physicochemical features to numeric types, and normalized column names to lowercase for consistency. Our target variable quality was an ordinal label typically ranging from 3 to 9.

B. Exploratory Data Analysis

We then performed an exploratory analysis to assess data was suitable for modeling. We generated a count plot, which confirmed a heavy class imbalance, with most samples falling into the mid-range quality scores. We proceeded by performing feature importance by using Random Forest, however all features were revealed important so none of them were removed from the dataset. Additionally, Variance Inflation Factor (VIF) scores were calculated to evaluate multicollinearity, but no features were removed at this stage either.

C. Class Imbalance Handling

We observed that wine quality scores of 3 and 9 were extremely rare in the dataset, representing a negligible proportion of total instances, so we removed them to reduce noise and ensure model stability, especially for ordinal classifiers, which depend on reliable class distributions. We then proceeded by applying the Synthetic Minority Oversampling Technique (SMOTE) to classes 4, 7, and 8 in order to mitigate the effects of class imbalance. SMOTE generates artificial samples for minority classes by interpolating between existing examples. We chose this approach to preserve feature distributions while ensuring that ordinal classes were better represented during training. It is important to mention that we applied SMOTE only to the training set to prevent data leakage.

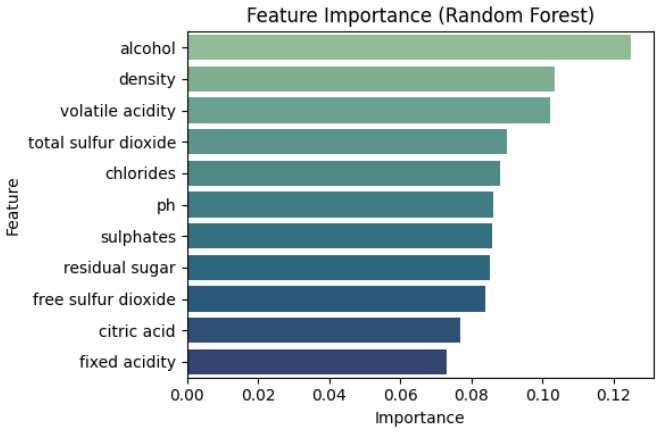

Fig. 1. Feature Importance plotted using Random Forest for Explanatory Data Analysis (Section B)

TABLE I
EXPERIMENT SETUP ACROSS MODELS

Experiment Parameters	Ordinal Logistic Regression	SVM (RBF)	Ordinal SVM (RBF)
Target Classes	4, 5, 6, 7, 8	4, 5, 6, 7, 8	4, 5, 6, 7, 8
Train-Test Split	Stratified 80-20	Stratified 80-20	Stratified 80-20
SMOTE	On Training Data	On Training Data	On Training Data
SMOTE Sampling Strategy	{4:400, 7:1600, 8:1000}	{4:400, 7:1600, 8:1000}	{4:400, 7:1600, 8:1000}
SMOTE `k_neighbors`	5	5	5
C (Regularization)	—	100	100
Gamma	—	0.1	0.1
Decision Strategy	Multiclass	One-vs-Rest	Custom One-vs-One
Decision Threshold	Default	Default	0.55
Performance Metrics	Accuracy, Precision, Recall, F1-score	Accuracy, Precision, Recall, F1-score	Accuracy, Precision, Recall, F1-score

TABLE II
PERFORMANCE COMPARISON OF MODELS ACROSS METRICS

Metrics \ Models		Ordinal LR	SVM RBF	Ordinal SVM RBF
Accuracy		0.47	0.60	**0.62**
Precision	Macro Average	0.43	0.48	**0.52**
	Weighted Average	0.49	0.60	**0.62**
Recall	Macro Average	0.33	**0.49**	0.48
	Weighted Average	0.47	0.60	**0.62**
F1-score	Macro Average	0.32	0.48	**0.49**
	Weighted Average	0.46	0.60	**0.61**

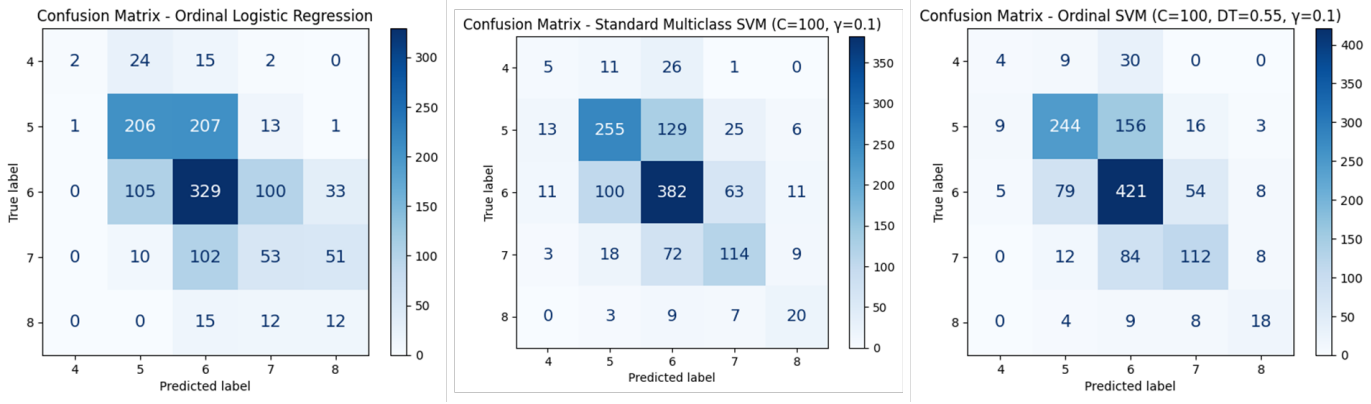

Fig. 2. Confusion matrices comparing the performance of Ordinal Logistic Regression, Standard Multiclass SVM, and Ordinal SVM for wine quality classification.

D. Modeling Approach

Given the categorical nature of the target variable, alongside applying Standard Multi-class SVM on the dataset, we employed the two ordinal classification methods previously mentioned:

- **Ordinal Logistic Regression (OLR)**
- **Ordinal Support Vector Machine (Ordinal SVM)**

Surprisingly, when we first tested the three models using the standard default range of $C = 1$ to 10 (i.e., the regularization parameter) for both SVMs (Ordinal and Standard), we found that the Standard Multiclass SVM outperformed the Ordinal SVM. However, after fine-tuning the parameters, more specifically setting $C = 100$ and $\gamma = 0.1$, the Ordinal SVM finally outperformed the Standard Multiclass SVM, as shown in Fig. 2. There could be several reasons for this. One possibility is that, in the presence of strong multicollinearity (as identified through VIF analysis during exploratory data analysis), the Standard SVM, being more flexible, performs better. Additionally, the regularization parameter C and the gamma value are critical factors, and fine-tuning them can significantly impact the performance of both models. This explains why, at $C = 100$, the Ordinal SVM ultimately demonstrated strong and superior performance.

V. Discussion and Conclusion

A. Conclusions

Based on our evaluation metrics in **Table 2**, the Ordinal SVM achieved the highest performance across all key indicators, including accuracy (0.62), weighted F1-score (0.61), and weighted precision/recall, outperforming both the Standard Multiclass SVM and Ordinal Logistic Regression. Notably, the standard multiclass SVM also demonstrated competitive performance, surpassing the ordinal logistic model by a significant margin.

The confusion matrices offer further insights into model behavior. While all models tended to confuse neighboring quality classes, which is a natural challenge given the ordinal nature of the labels, Ordinal SVM better preserved the ordinal structure, especially by reducing misclassifications to distant labels. This reflects its design intention of penalizing ordinal violations more heavily.

B. Ideas of Future Work

There are several directions for future research that could extend and improve the findings in our study. One potential path is to explore Support Vector Regression (SVR) techniques, treating wine quality as a continuous variable and discretizing the output afterward. This approach could better capture the ordinal relationships than strict classification models. Additionally, while SMOTE was applied during training to address class imbalance, future work could investigate applying SMOTE-like techniques to the test data (with caution) to evaluate their effect on generalization and consistency. Another promising strategy is consolidating adjacent quality classes—such as merging rarely occurring labels like 3 and 4 with neighboring categories, potentially reducing noise and improving classification accuracy. These extensions, combined with further model comparisons and hyperparameter tuning, may lead to more robust and interpretable ordinal prediction frameworks.

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, Nov. 2009. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0167923609001377

[2] M. Mohammadi, F. A. Harun, and D. G. Rajpathak, "Ordinal classification in artificial intelligence: Techniques and applications," *Applied Intelligence*, vol. 54, pp. 6314–6340, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10489-023-04810-2

[3] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Wine Quality [Dataset]," UCI Machine Learning Repository, 2009. [Online]. Available: https://doi.org/10.24432/C5653T