Title: Comparative Analysis of Classification Models for Wine Quality Dataset from Kaggle

Team members: Fardad Homafar(90319914), Gervasio Martinez(82363524), Hyejeong Choi(88566492)

Introduction

According to recent research, the interest in applying machine learning algorithms to wine quality is still worldwide.[1, 2] In addition, the certification of wine quality is essential to the wine industry.[3, 4] However, as machine learning models have continuously been developed, finding the appropriate model for a particular dataset, such as wine quality, is still an issue for machine learning researchers. A myriad of previous research has modeled versatile methods comparing classification and regression models.[5, 6] Nevertheless, there is limited research that focuses only on Boosting methods. In this paper, we will concentrate on boosting techniques to train the wine quality variables and compare them to conventional models in order to analyze both the suitability of these models and their benefits and shortcomings.

The complexity of the wine quality dataset arises because of its few and imbalanced samples. Traditional machine learning models are biased towards the majority class, leading to poor performance on the minority class. On the other hand, boosting is a machine learning ensemble technique that combines the predictions of multiple weak learners to create a strong learner. The primary goal of boosting is to sequentially improve the accuracy of the model by giving more emphasis to the misclassified instances in the training set. With their ability to focus on and correct misclassification, boosting algorithms can address imbalances and enhance the model's performance, especially in the minority class.

This project aims to compare multiple boosting methods (Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), and Light Gradient-Boosting Machine (LightGBM)) for multi-class classification applied to the wine quality dataset from Kaggle. Then, in order to analyze the advantages and disadvantages of boosting methods, we will compare their results with conventional methods, such as Support Vector Machines (SVM) and Random Forests (RF) [7, 8]

Methodology

In this project, we will be using the following three boosting algorithms:

- **XGBoost,** an efficient gradient boosting for structured data with decision tree ensembles, is known for preventing overfitting through regularization. Widely used in competitive machine learning, it supports custom loss functions and handles missing data effectively.[9]

- **AdaBoost,** an adaptive boosting algorithm, iteratively improves weak learners, typically simple decision trees. Focusing on challenging instances, it creates a robust classifier by assigning greater weight to misclassified examples in each iteration, commonly applied in classification tasks.[10]

- **LightGBM,** developed by Microsoft, is a high-speed gradient-boosting framework with a unique tree-building strategy. It reduces memory usage and accelerates training using a histogram-based

approach for feature binning. It excels in distributed training on large datasets and is versatile for both classification and regression.[11]

Observation: We may consider other boosting methods if they are relevant to the application (for example, Category Boosting (CatBoost), Stochastic Gradient Boosting (SGD), and Gradient Boosting Machine (GBM)).

Dataset

Our analysis will utilize the Wine Quality Dataset, which is available on Kaggle. This dataset is related to red variants of the Portuguese "Vinho Verde" wine. It describes the amount of various chemicals in wine and their effect on its quality, which has 11 features and 1143 examples. These features are summarized into 11 continuous numerical features such as fixed acidity, volatile acidity, and citric acid, whereas the classification target (quality) discreetly ranges from 3 to 9.

Expected Contribution

In assessing the three models mentioned earlier, we will conduct a detailed analysis focusing on aspects such as accuracy, computational efficiency, and how well they handle the specific features of the Wine Quality Dataset. Additionally, we will compare these models with conventional methods to highlight their strengths and weaknesses. The goal of this comparative study is to offer valuable insights into the effectiveness and applicability of boosting techniques to datasets with similar characteristics.

References

[1] D. F. Török, "Machine Learning for Predicting Wine Quality and its Key Determinants Based on Physicochemical Properties," *Sage Science Review of Applied Machine Learning*, vol. 6, no. 11, pp 1-21, Nov. 2023.

[2] P. Bhardwaj, P. Tiwari, K. Olejar Jr, W. Parr, and D. Kulasiri, "A machine learning application in wine quality prediction," *Machine Learning with Applications*, vol.8, p. 100261, Jun. 2022.

[3] E. Parga-Dans, P. Alonso González, and R. Otero-Enríquez, "The role of expert judgments in wine quality assessment: the mismatch between chemical, sensorial and extrinsic cues," *British Food Journal*, vol. 124, no. 12, pp. 4286-4303, Nov. 2022.

[4] M. Fiore, M. Giacomarra, M. Crescimanno, and A. Galati, "Quality certifications' impact on wine industry assets performance," *Bulgarian Journal of Agricultural Science*, vol. 26, no. 2, pp. 257-267, Mar. 2020.

[5] K. R. Dahal, J. N. Dahal, H. Banjade, and S. Gaire, "Prediction of wine quality using machine learning algorithms," *Open Journal of Statistics*, vol. 11, no. 2, pp. 278-289, Mar. 2021.

[6] S. Mani, R. A. Krishnankutty, S. Swaminathan, and P. Theerthagiri, "An investigation of wine quality testing using machine learning techniques," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 747-754 June. 2023.

[7] *Support vector machine*, Wikipedia, The Free Encyclopedia, Nov. 2023. [Online]. Available: https://en.wikipedia.org/wiki/Support_vector_machine

[8] *Random Forest*, Wikipedia, The Free Encyclopedia, Nov. 2023, [Online], Available: https://en.wikipedia.org/wiki/Random_forest

[9] C. Tianqi, and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785-594, doi: 10.1145/2939672.2939785.

[10] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, Sep. 1999.

[11] G. Ke, M. Qi, F. Thomas, W. Taifeng, C. Wei, M. Weidong, Y. Qiwei, and L. Tie-Yan, "Lightgbm: A highly efficient gradient boosting decision tree," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3149-3157, doi: 10.5555/3294996.3295074.