

# CLASIFICACIÓN

Ryoji Takahashi

November 30, 2024

## Contents

### 1 Description of the code

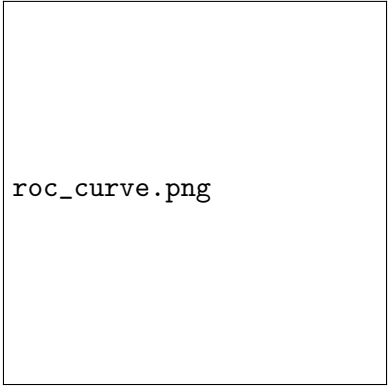
The project of the classification objections are following.

1. Choose an open classification dataset other than Iris (reference source)
2. In a Jupyter notebook, perform exploratory analysis (EDA), including data cleaning, transformations, aggregations and visualizations as appropriate.
3. Select, train and test the model(s) considered appropriate.
4. Justify the chosen model based on performance metrics.
5. Draw conclusions from the exercise carried out.
6. Prepare a deliverable with all the files necessary to reproduce the analysis and put the trained model into production (DevOps integration).

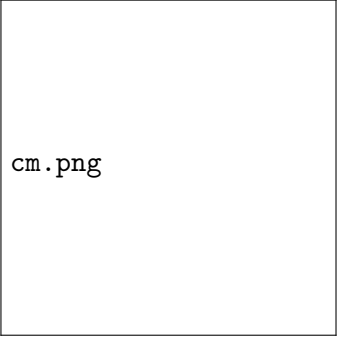
The **wine quality** dataset was chosen. As in the descriptions in the code, quality was ranked from 3 to 8, (8 is highest). I could use multi-classification methods, however, to simplify, I set up a binary classification good quality (1) is higher the 7 and rests were not good (0). As in the code, good quality has 217. not good quality has 1382. This is imbalanced datasets.

To deal with imbalanced datasets, there are several options. The popular one is SMOTE (Synthetic Minority Over-sampling Technique), threshold moving techniques, and ensemble techniques. Among ensemble methods, Random Forest (ensemble of multiple decision trees) and XGBoost (bagging with L1, L2 regularizations) are most popular ones. Therefore, I deployed these methods to build. Apart of scaling (transforming  $\mu = 0$ ,  $\sigma = 1$ ), I have directly applied two methods for the classification problem.

In Random forest, I applied grid search to improve the classification then default settings, however, it did not improve them. I have left line of codes for grid search in the notebook. In XGBoost, optuna was used for hyperparameter tuning. As the final results, XGBboost has better results.



roc\_curve.png



cm.png

...

## 2 Conclusions and Future Work

bias-variance of triad off, feature engineering are also important.