

# ML Assignment

## 2020 - 2021

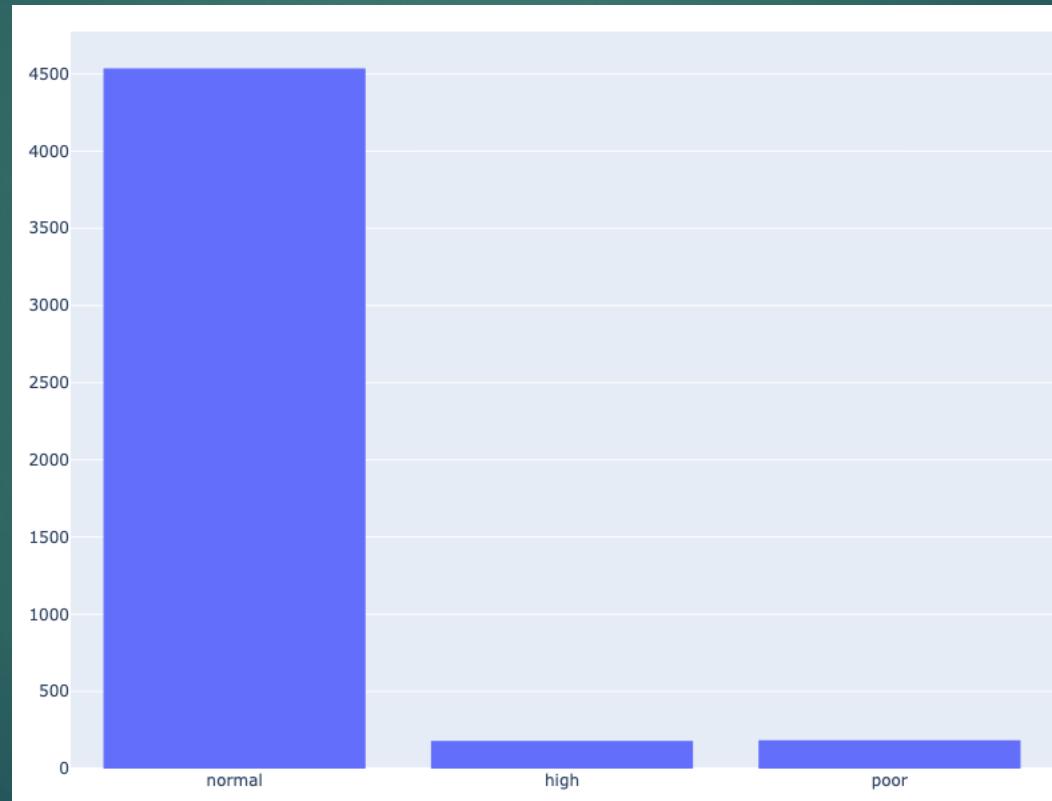
WINE QUALITY PREDICTION  
STERGIOS GIANNIOS – MTN 2003

# Problem Definition

- ▶ The Dataset (from UCI):
  - ▶ 4898 examples of white wines
  - ▶ 11 physicochemical variables e.g. (acids, pH, residual sugar)
  - ▶ Quality range -> [3,9]
  - ▶ Modeled as a classification task
    - ▶ [3,4] -> poor quality
    - ▶ [5,7] -> normal quality
    - ▶ [8,9] -> high quality

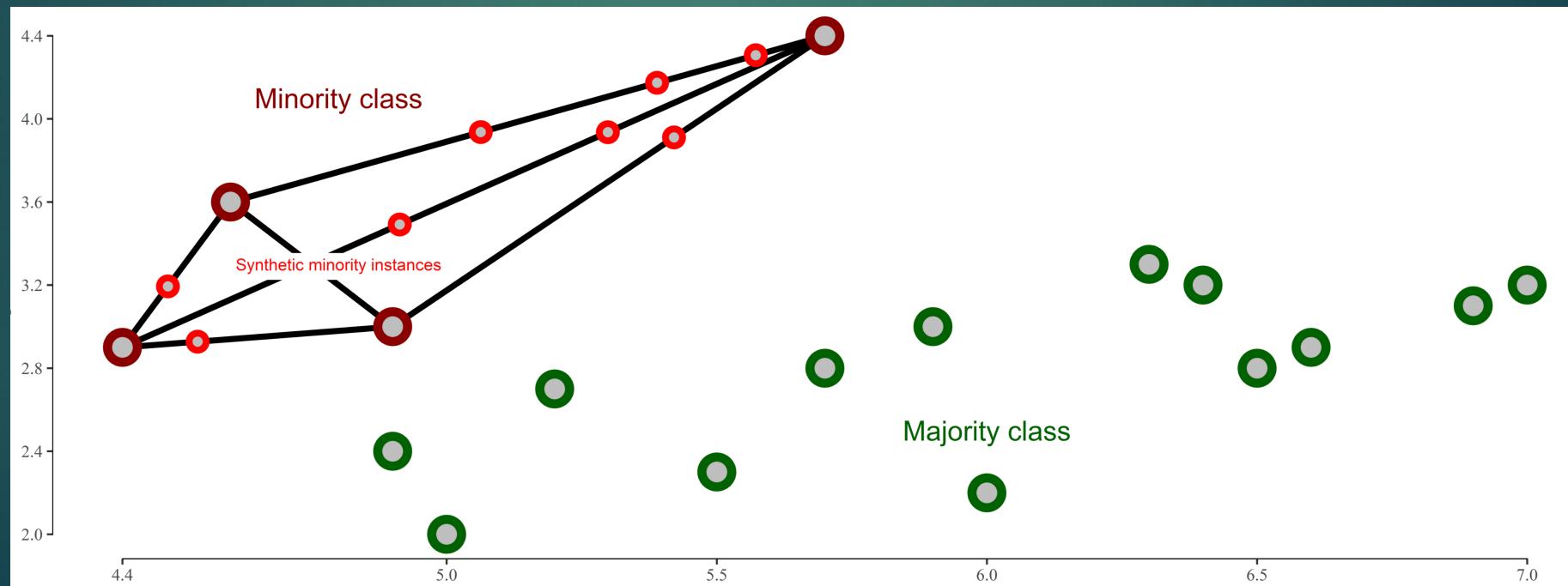
# Class Imbalance

93% of the examples belong to the majority class



# SMOTE

- ▶ The Synthetic Minority Oversampling Technique (SMOTE) is an oversampling method to address the problem of imbalanced distribution of data.
- ▶ SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.



# Classifiers

- ▶ Logistic Regression
- ▶ Gaussian Naïve Bayes
- ▶ K Neighbors
- ▶ SVM
- ▶ Decision Tree
- ▶ Random Forest
  
- ▶ Grid Search CV for hyperparameter tuning.

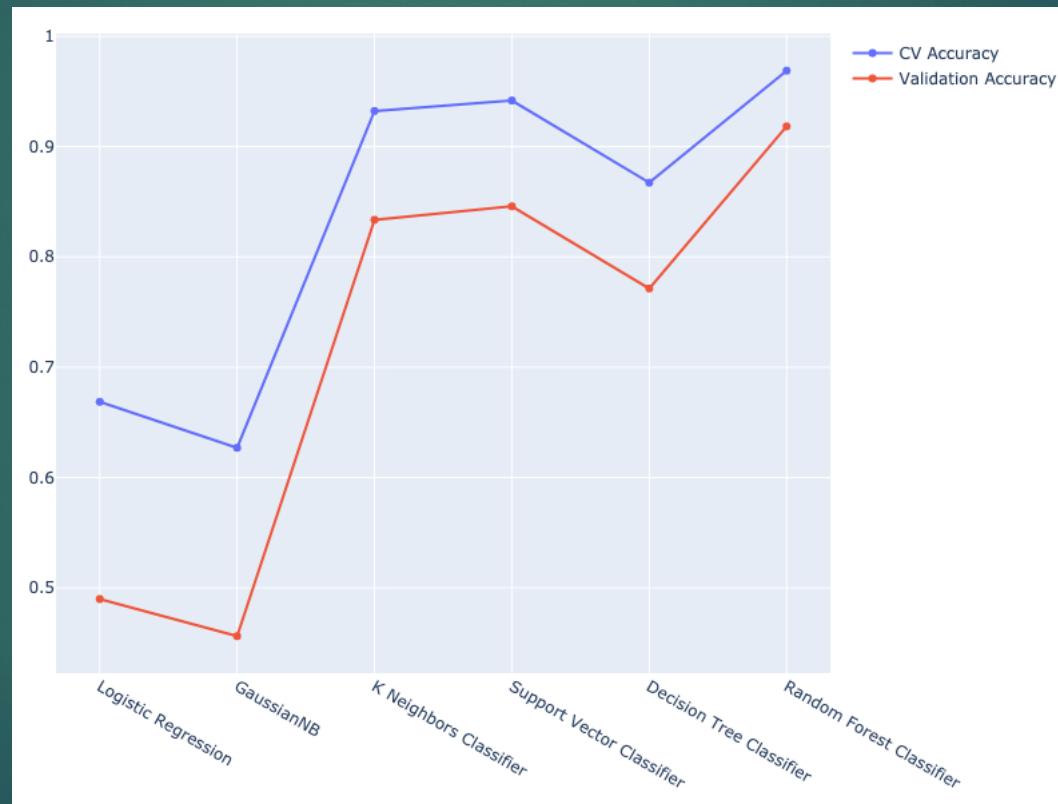
# Model Evaluation

- ▶ CV, Validation Accuracy -> improve training.
- ▶ Test Accuracy and ROC Curves -> generalization capabilities.

# Model Comparison (1/2)

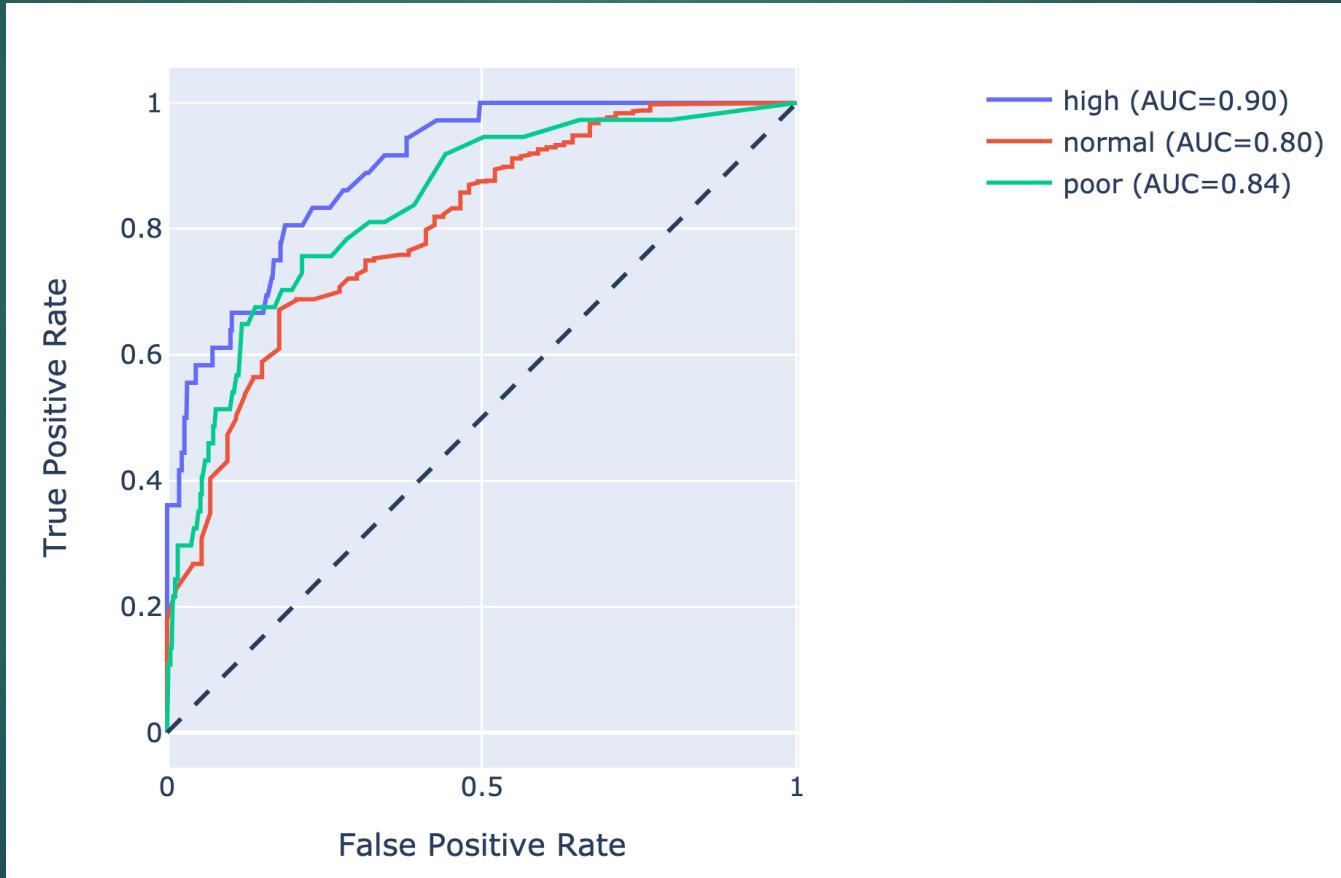
	Classifier	Tuned Hyperparameters	Mean CV Accuracy	Validation Accuracy	Test Accuracy
0	Logistic Regression	{'C': 0.1, 'penalty': 'l1'}	0.668653	0.489796	0.478571
1	GaussianNB	{}	0.626943	0.456122	0.457143
2	K Neighbors Classifier	{'algorithm': 'ball_tree', 'n_neighbors': 3, '...}	0.932383	0.833673	0.845918
3	Support Vector Classifier	{'C': 10, 'kernel': 'rbf'}	0.941839	0.845918	0.857143
4	Decision Tree Classifier	{'criterion': 'entropy', 'max_depth': 10}	0.867358	0.771429	0.734694
5	Random Forest Classifier	{'max_depth': 21, 'n_estimators': 90}	0.968912	0.918367	0.907143

# Model Comparison (2/2)



# Best Model

Random Forest ROC Curve



Questions?