JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Great Learning

Wine Data

Practical Implementation of ML Algorithms

Agenda

Quick Introduction to Dataset

Create code to look at the data

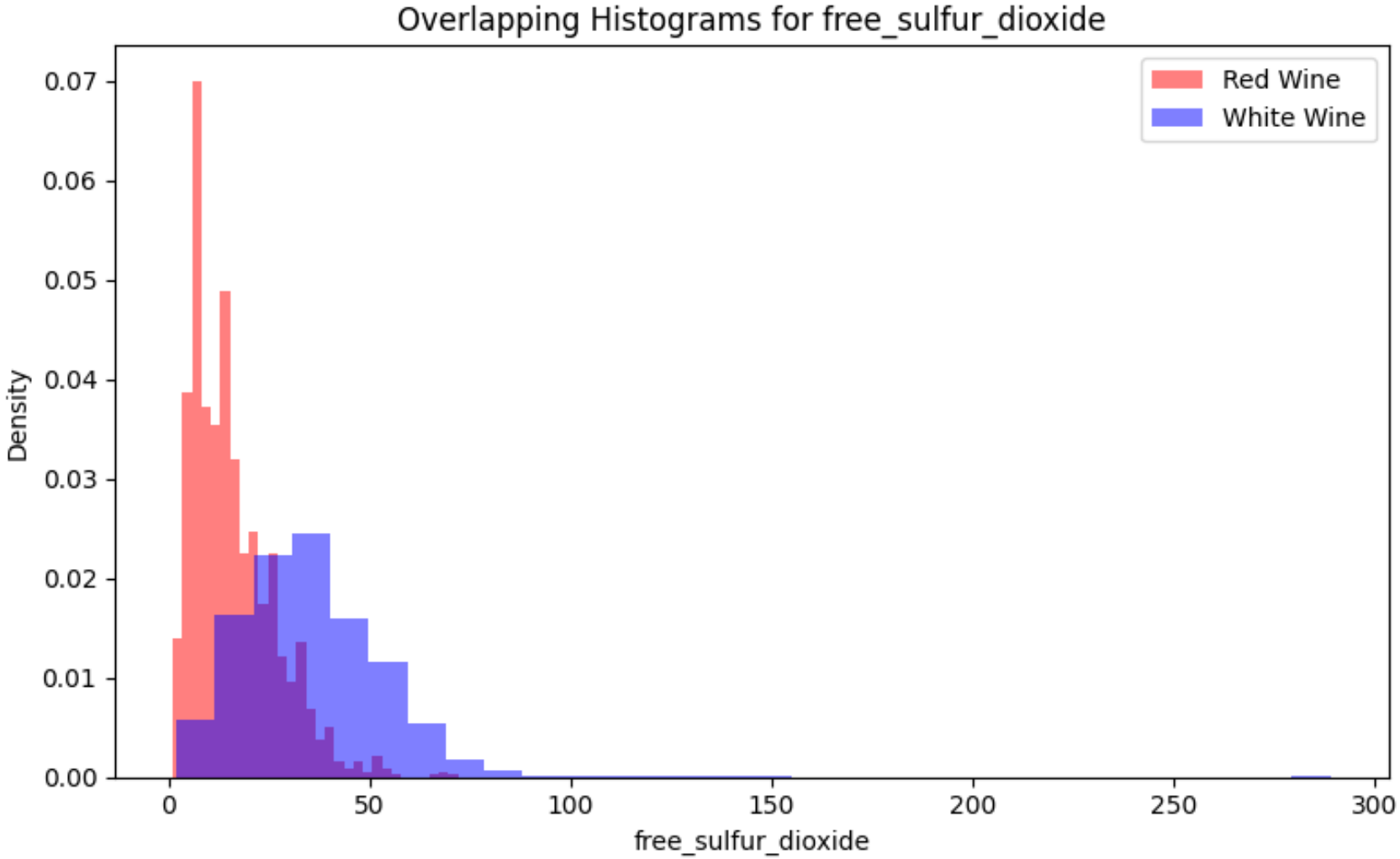
Create code to prepare it for machine learning algorithms

Wine Dataset

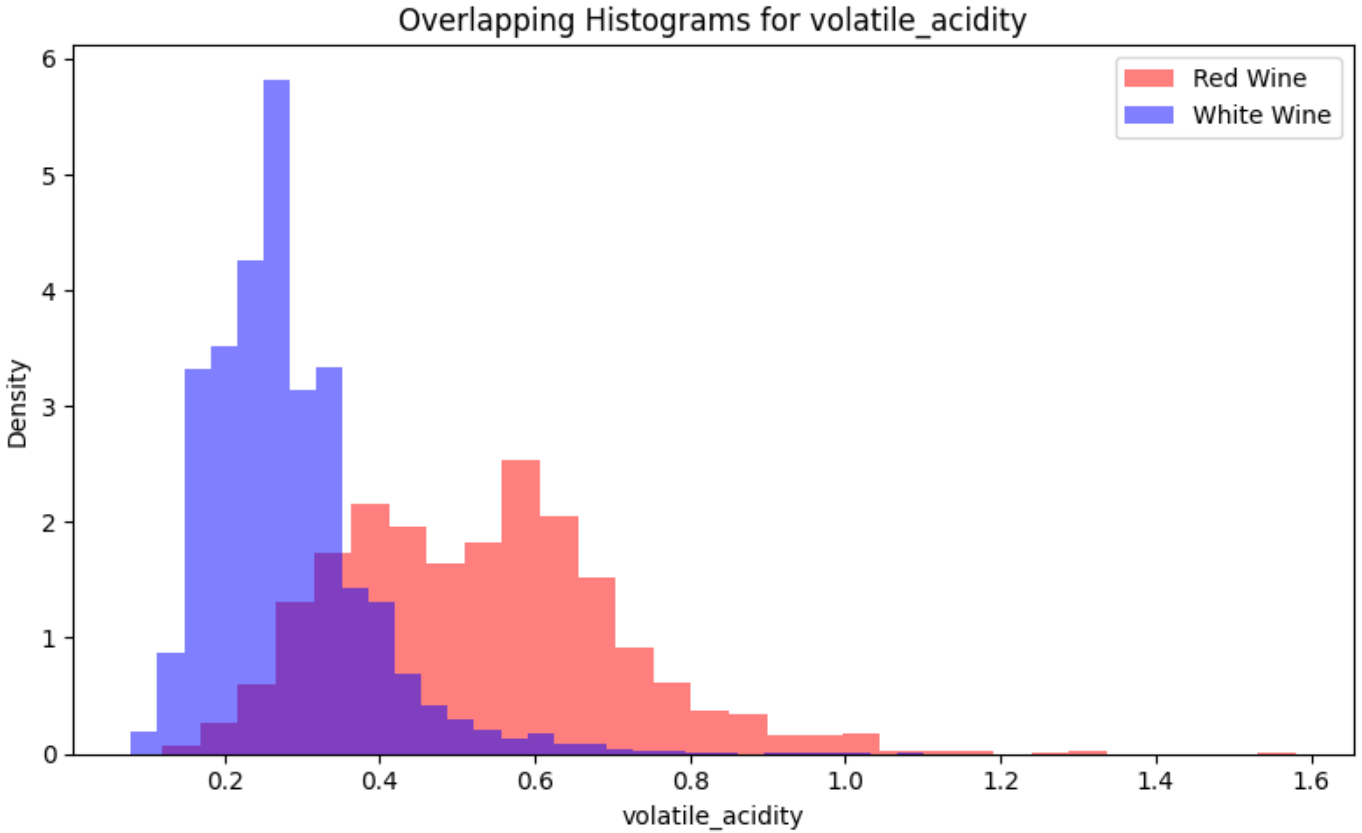
- Classic and very simple dataset with information about red and white variants of the Portuguese *Vinho Verde* wine
  - Vinho Verde implies very "young" wines
  - One file for the white wines and one for red wines. Each files contains:
    - **12** metrics from objective tests (e.g. acidity levels, PH values, ABV, etc…)
    - **1** metric from with a subjective score based on wine experts — median of at least 3 evaluations made by the experts. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent)
    - It contains information from 1599 red wine samples and 4898 white wine sample, which means it is an unbalanced dataset
    - Due to privacy and logistic issues, there is no data about grape types, wine brand, and wine selling price.

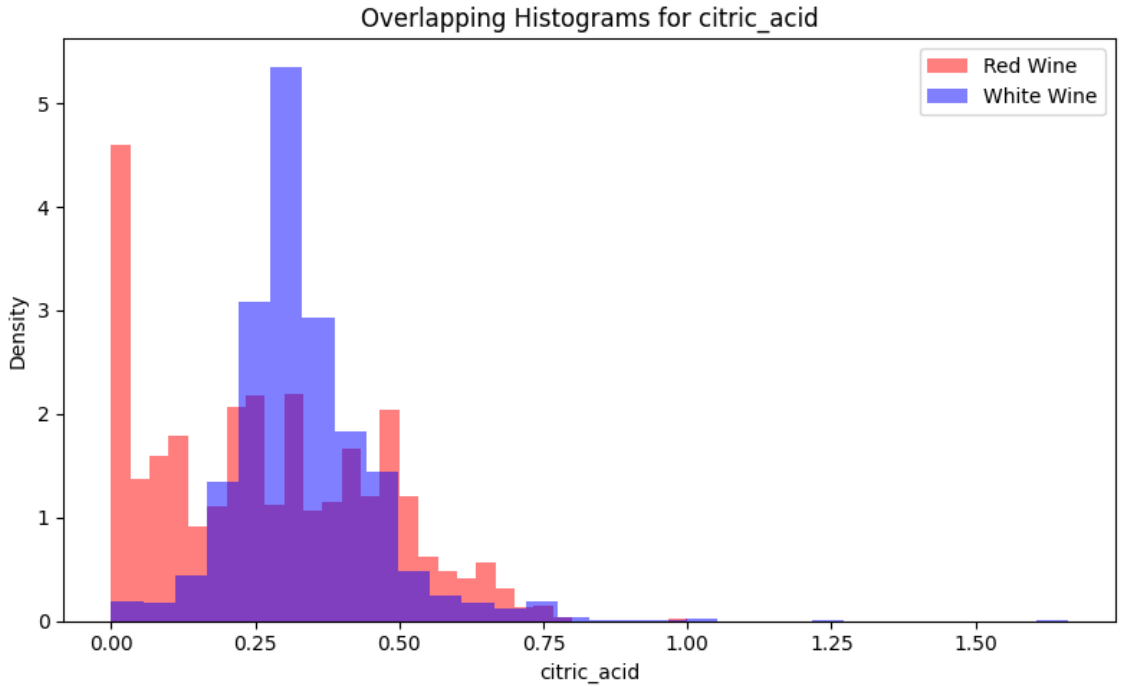
https://www.kaggle.com/datasets/ruthgn/wine-quality-data-set-red-white-wine/data

Metric Histogram per Wine Color

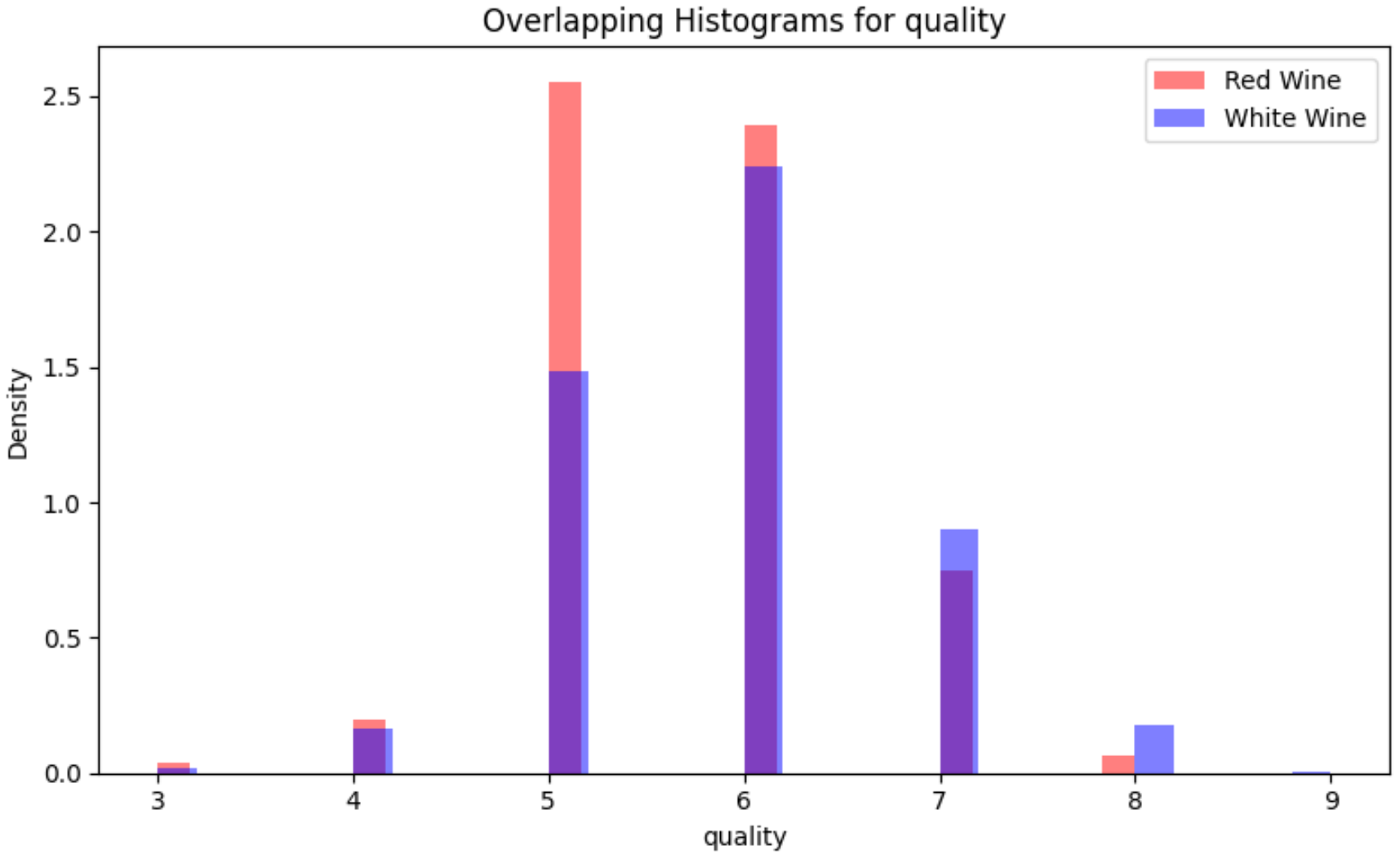


Metric Histogram per Wine Color



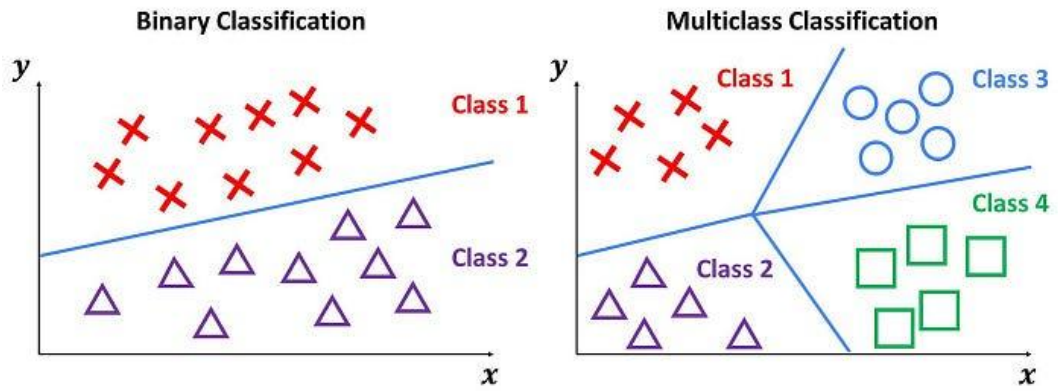


Metric Histogram per Wine Color



Classification Code

- Prepare data to create ML classifiers to:
  - Predict the wine type given all the wine metrics (except quality) – binary classifier
  - Predict the wine quality given all the wine metrics – multiclass classifier
- ML Classifiers to use in our code:
  - SVM
  - Naïve Bayes
  - Decision Tree
  - Random Forest
  - Neural Network
  - Logistic regression



Agenda

Quick Introduction to Dataset

Create code to look at the data

Create code to prepare it for machine learning algorithms

Summary

- Neural networks (NN) are machine learning algorithms inspired by the central nervous system
  - Various layers of "neurons" connected by adaptive weights
  - First computer implementation in 1957
  - Can solve simple problems