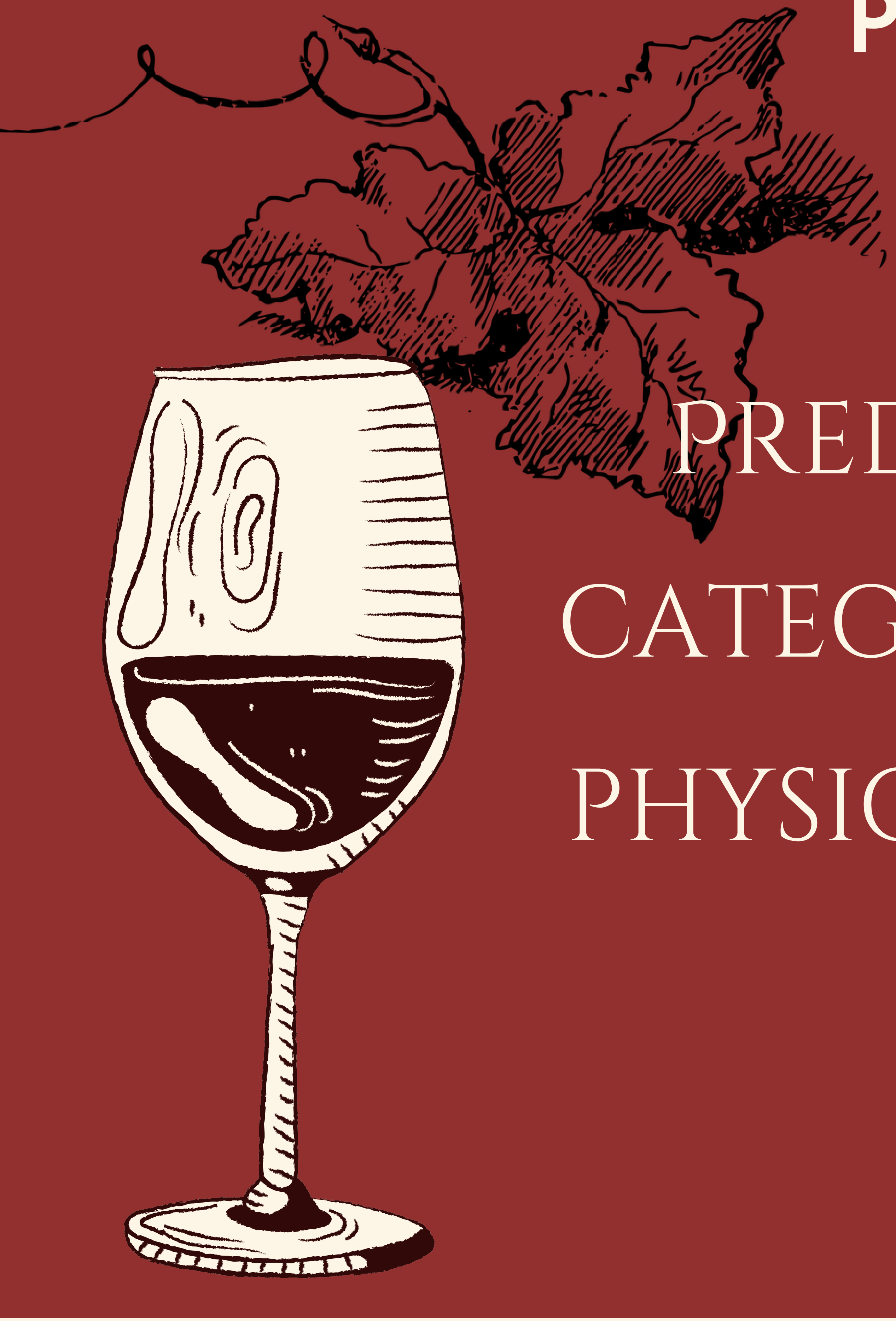
Primatijada 2024



PREDICTING THE PRIMARY CATEGORY OF WINE BASED ON PHYSICOCHEMICAL QUALITIES

Emilija Trajkovska

I

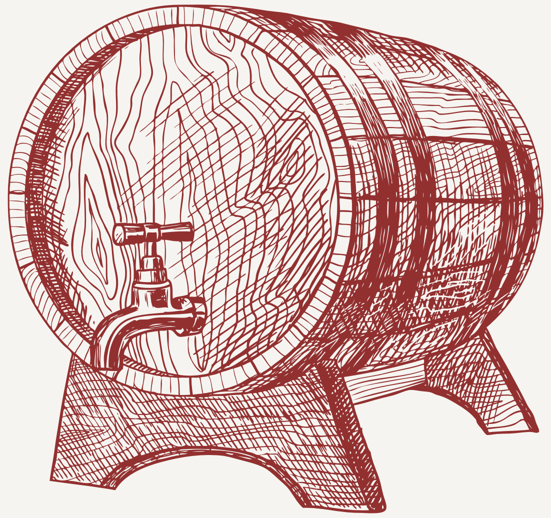
INTRODUCTION

2

MATERIALS AND METHODS

DATA SETS

Variants of Portuguese Vinho Verde wine



Unicorn winery

FEATURES

category of wine

Color of wine.

quality

Score between 0 (very bad) and 10 (very excellent) by wine experts.

fixed acidity

The acids that naturally occur in the grapes used to ferment the wine and carry over into the wine.

volatile acidity

Acids that evaporate at low temperatures.

citric acid

Used as an acid supplement which boosts the acidity of the wine.

residual sugar

The amount of sugar remaining after fermentation stops.

chlorides

The amount of salt in the wine.

free sulfur dioxide

The free form of SO2 which exists as a dissolved gas.

total sulfur dioxide

The amount of free and bound forms of S02.

density

The density of wine juice depending on the percent alcohol and sugar content.

pH

How acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic).

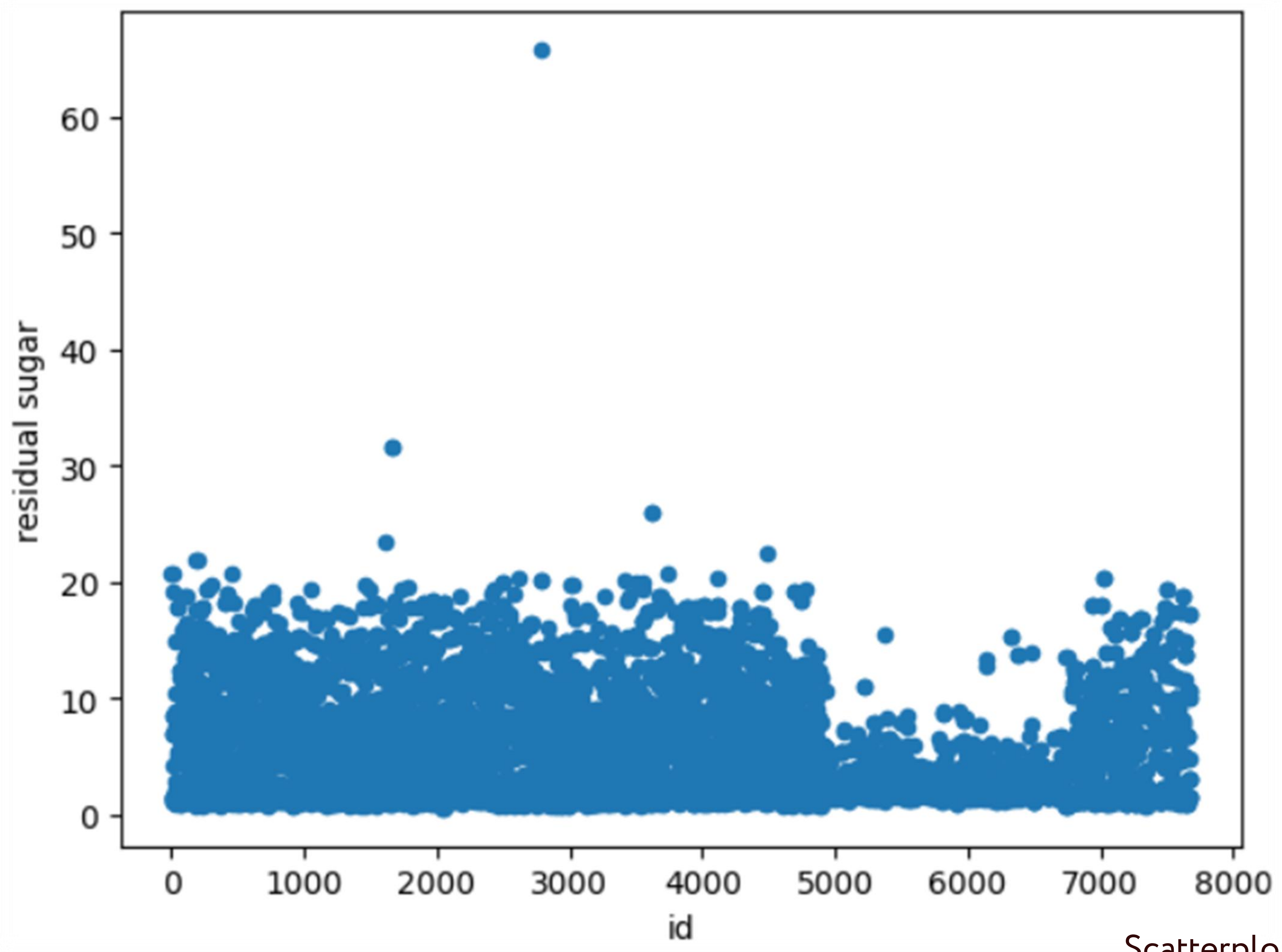
sulphates

Amount of potassium sulphate as a wine additive which can contribute to sulfur dioxide gas (S02) levels.

alcohol

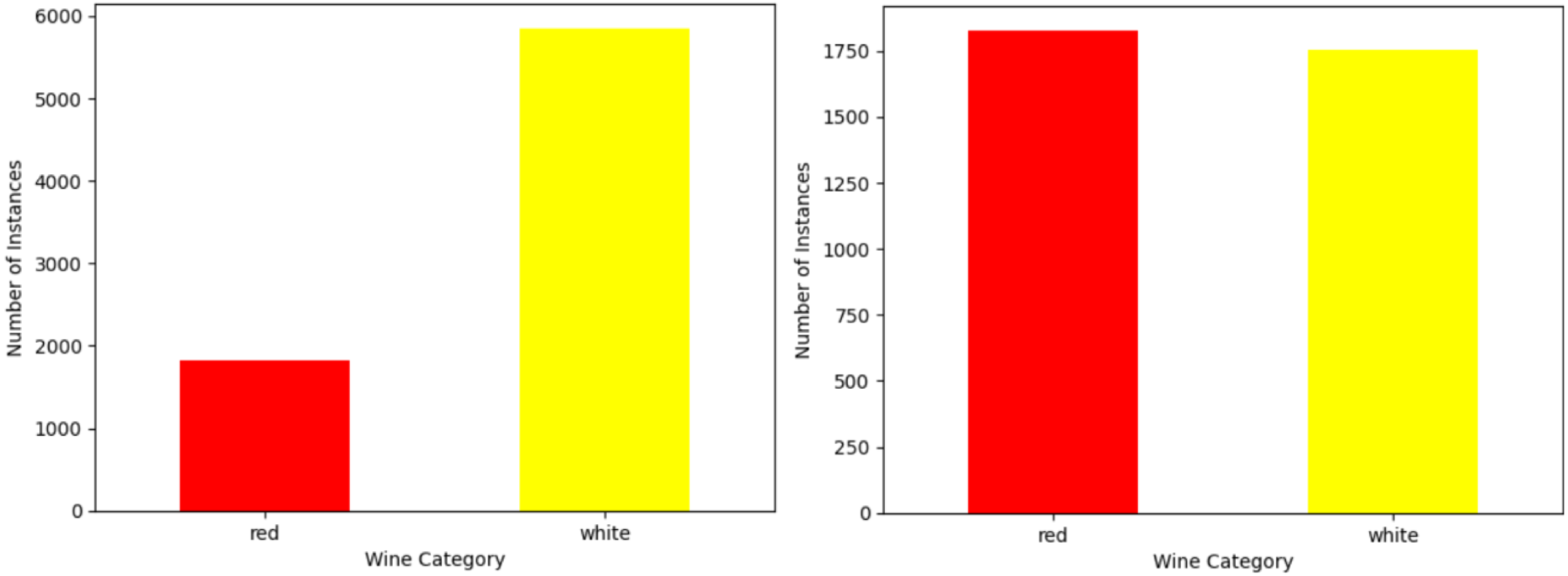
The percent alcohol content of the wine.

DATA PREPARATION



Scatterplot of the feature residual sugar

DATA PREPARATION



Histograms of before and after balancing the data set.

DATA MINING APPROACH

K-fold cross-validation

randomly dividing the set of observations into k groups, or folds, of approximately equal size

Percentage split

the dataset is randomly divided into two parts: training and testing dataset

DATA MINING TECHNIQUES

- Decision Tree Classifier
- Random Forest Classifier
- Logistic Regression
- Histogram-based Gradient Boosting
- K-Nearest Neighborhood
- K-Means Clustering

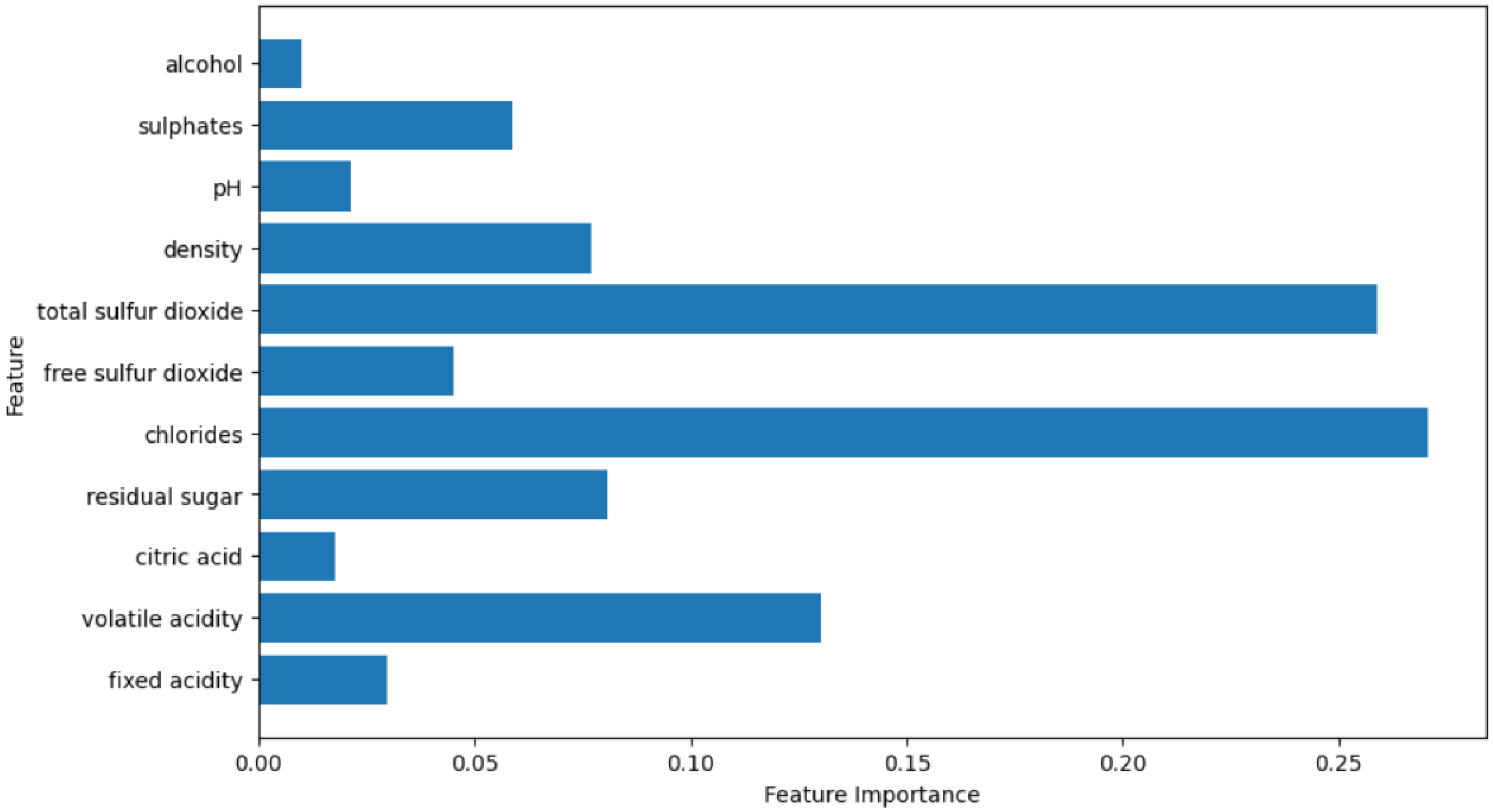
3

EXPERIMENTAL RESULTS

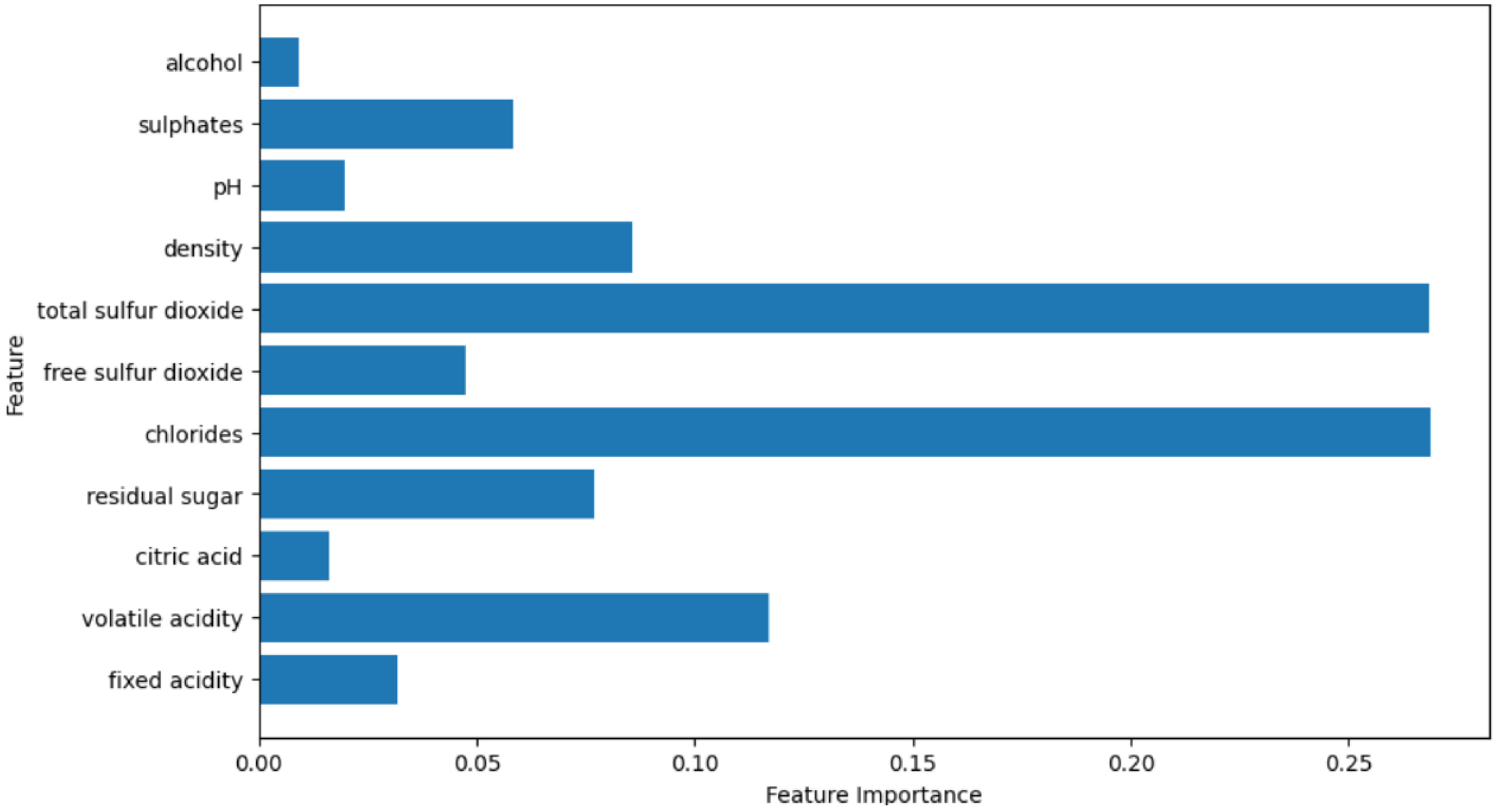
PERFORMANCE RESULTS OF THE EMPLOYED MODELS

Model	CV accuracy (%)	TT accuracy (%)
LogisticRegression	97.3	97.2
DecisionTreeClassifier	97.8	97.3
RandomForestClassifier	99.3	99.3
HistGradientBoostingClassifier	99.3	99.2
KNearestNeighborsClassifier	92.7	98.1

FEATURE IMPORTANCES

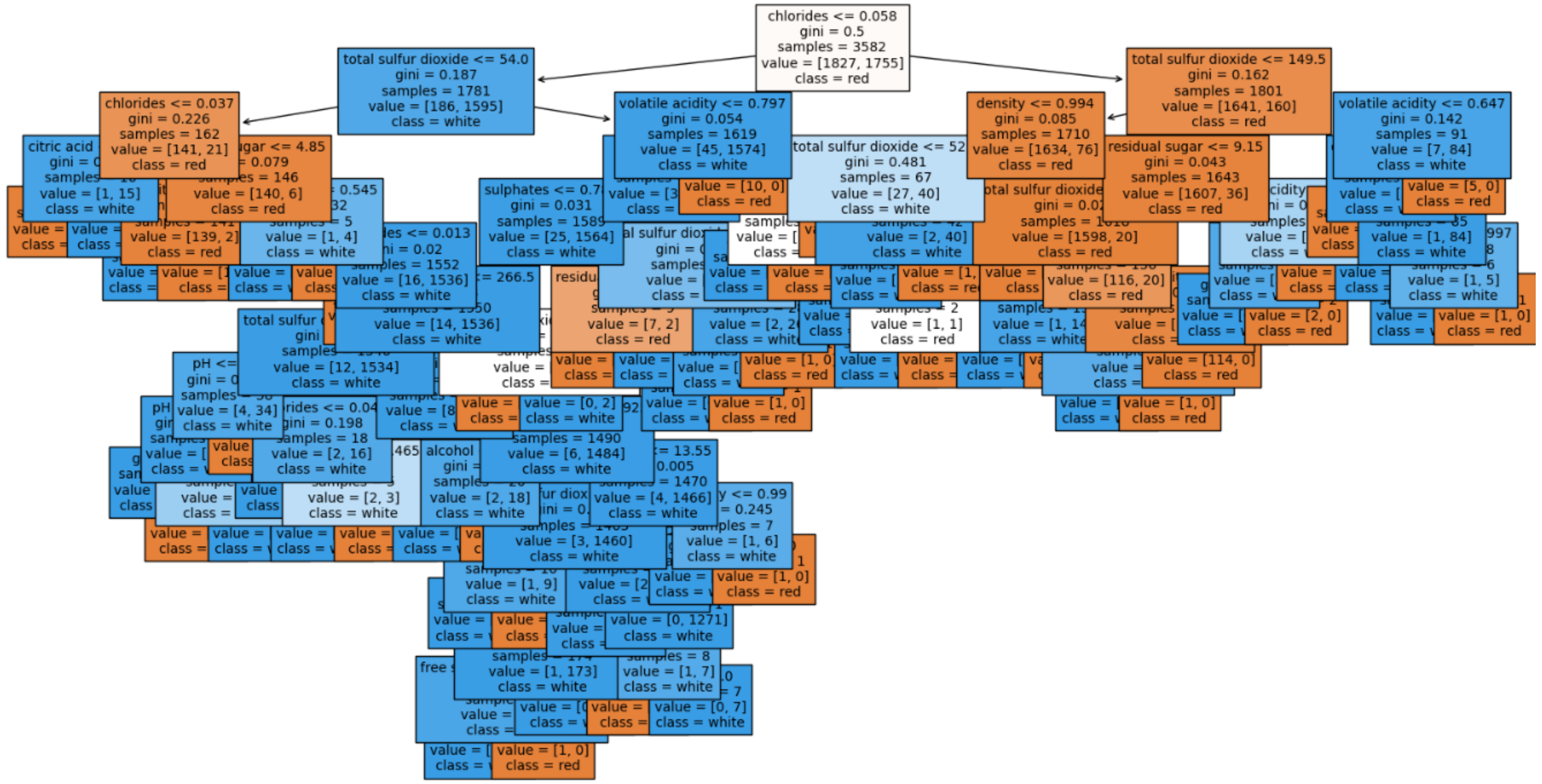


Cross-validation Random Forest Classifier feature importances

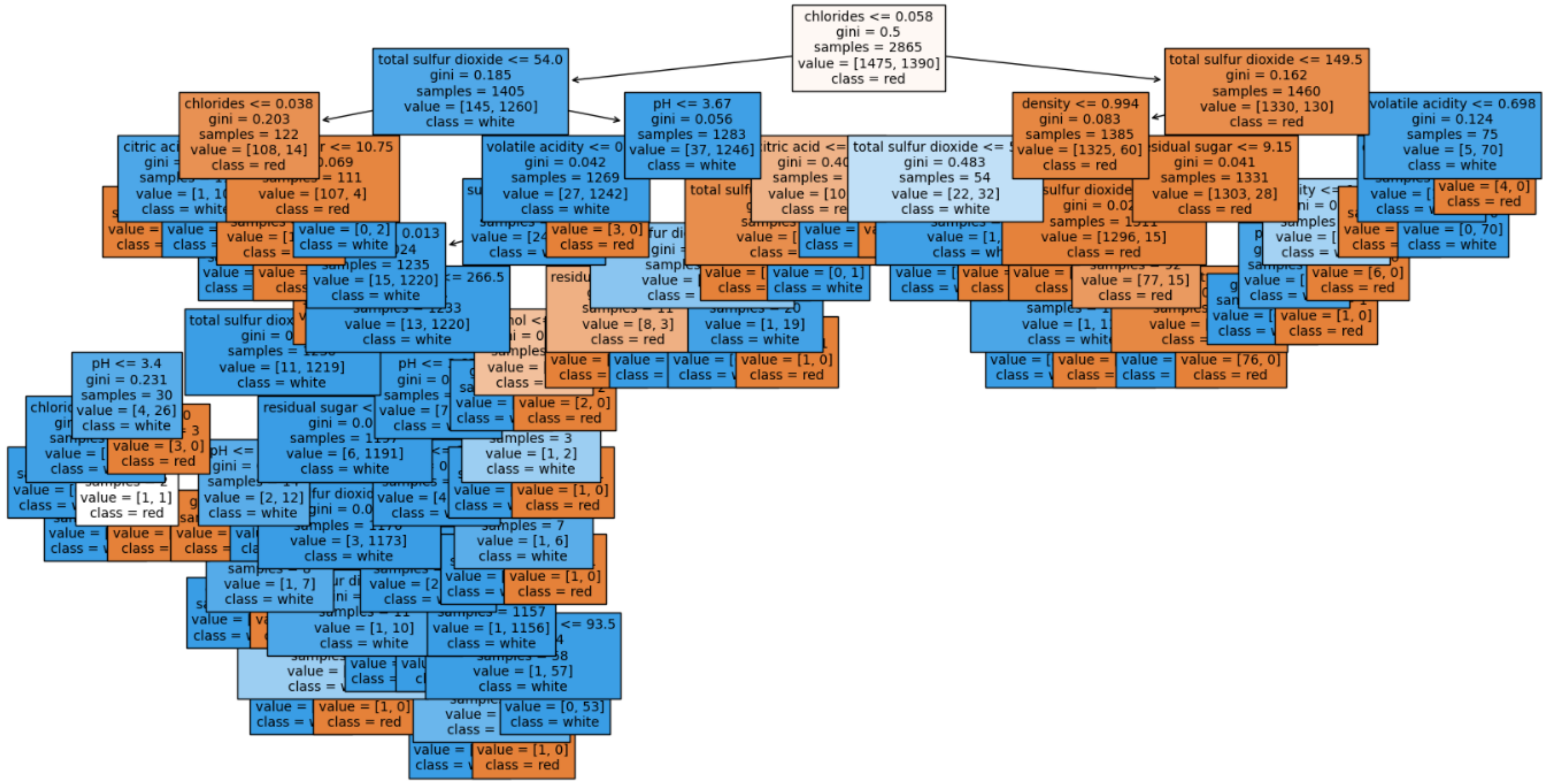


Train-test Random Forest Classifier feature importances

DECISION TREES

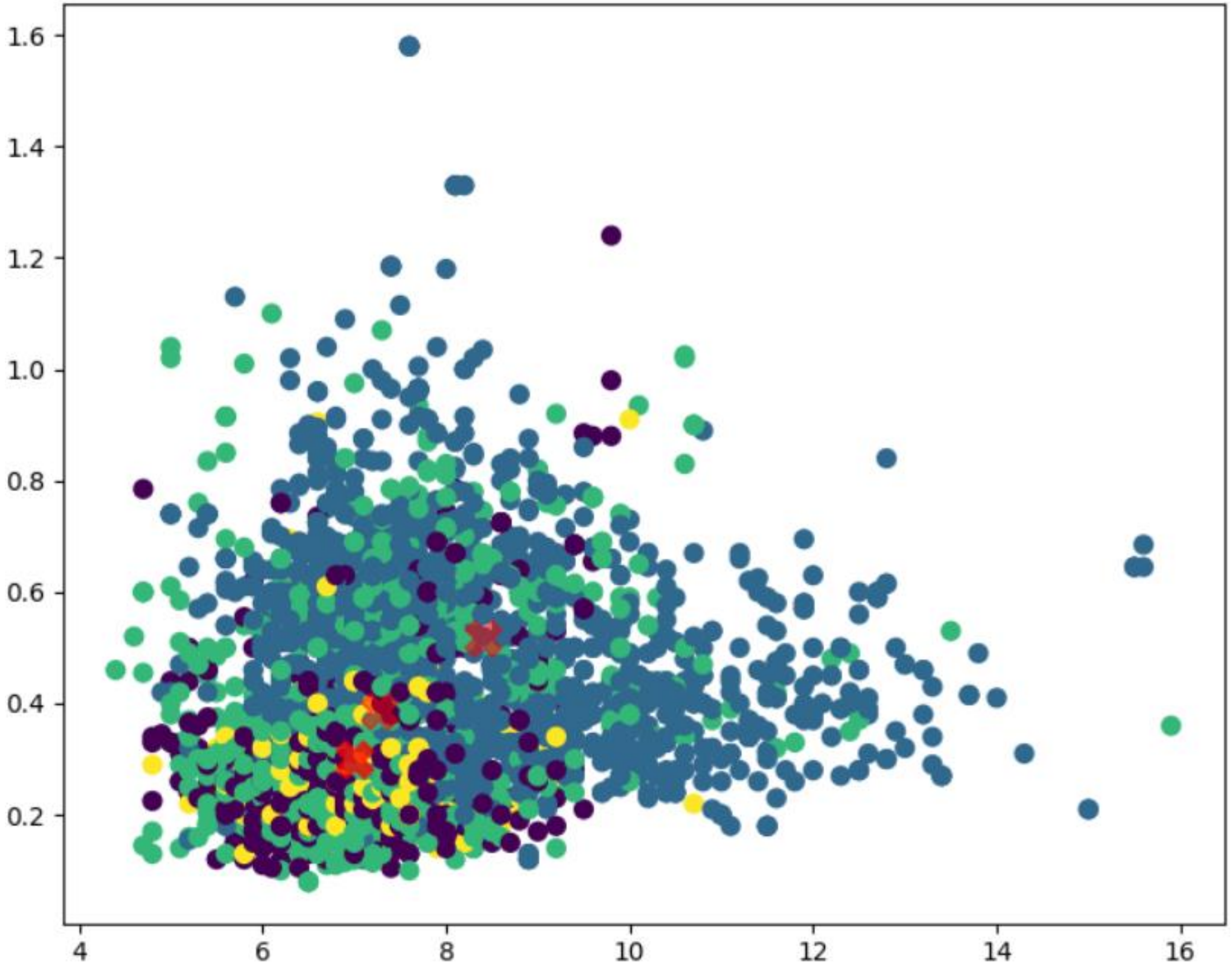


Cross-validation Decision Tree.

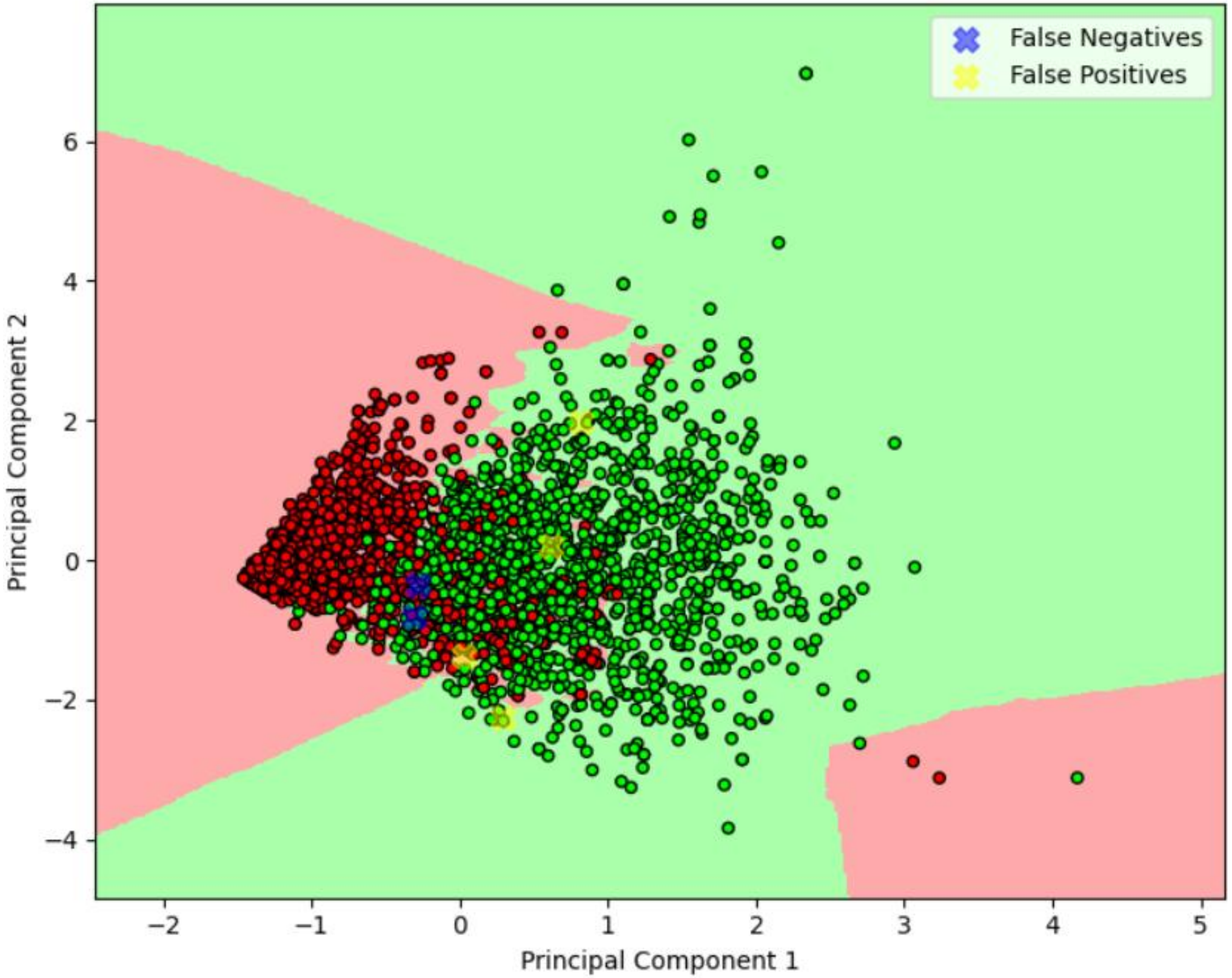


Train-test Decision Tree.

K-MEANS CLUSTERING



K-NEAREST NEIGHBORHOOD



4

CONCLUSIONS

THANK YOU

ANY QUESTIONS?