Wine Quality Prediction Analysis

Data Science Capstone Project

present by Yan Zhang

Nov 27th2023

What's the problem?

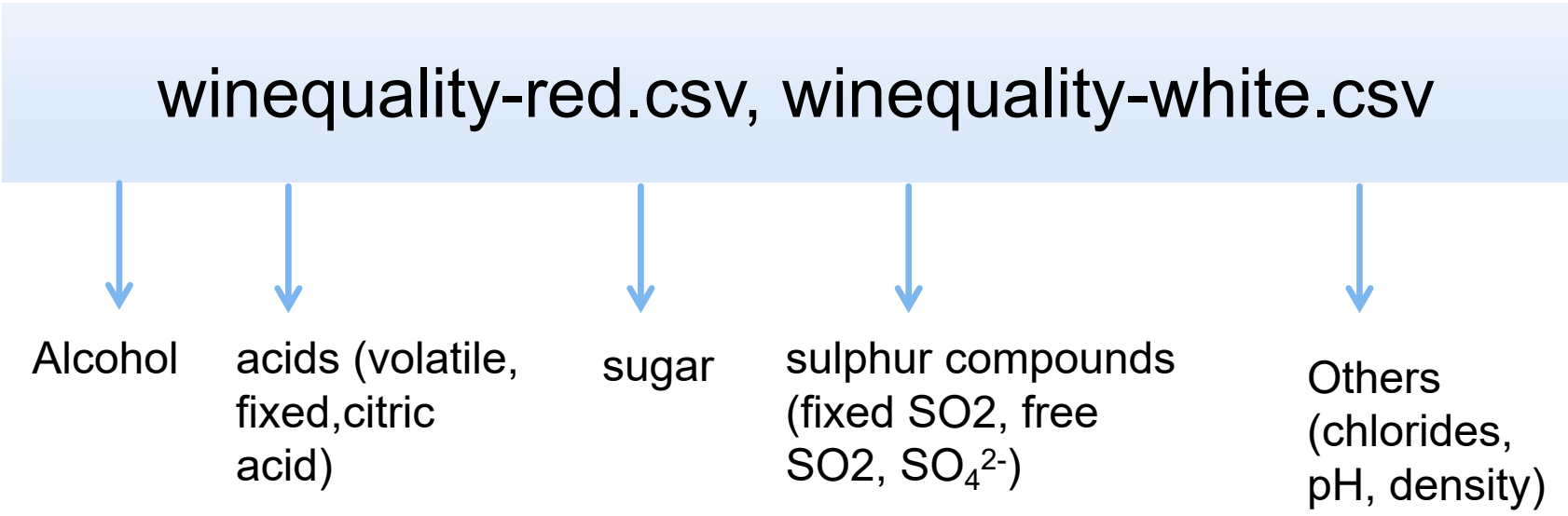
- Wine product Vinho verde quality rate: 0 (very bad) to 10 (very good).
- much more normal wines than excellent or poor ones.

- what features affect wine quality?
- Can we predict the likehood of good wine quality?

Who cares?

- Wine companies in Minho (northwest) region of Portugal.
- Wine companies produce red/white wine
- Vinho Verde marketing

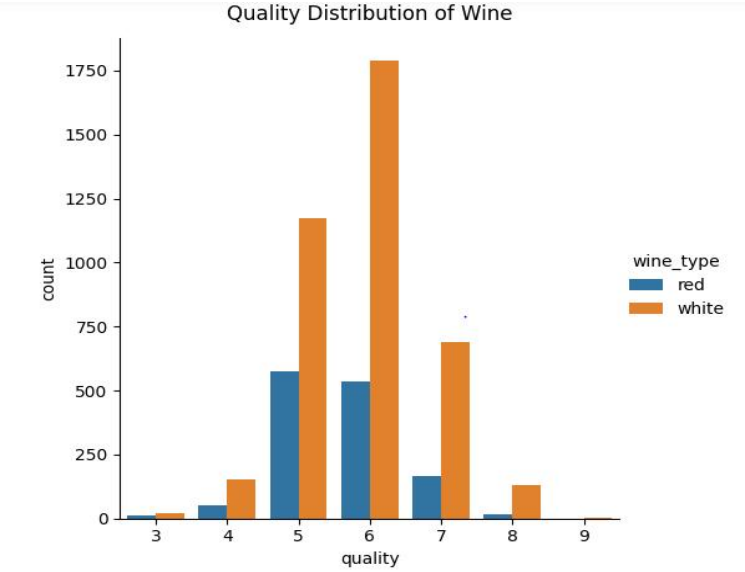
What might affect wine quality?

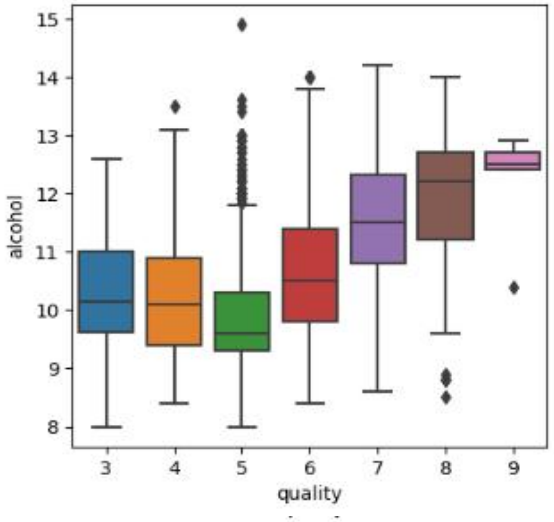


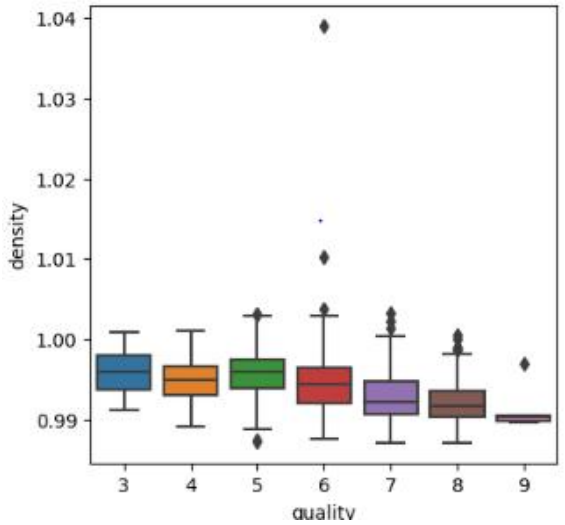
data source:https://archive.ics.uci.edu/dataset/186/wine+quality
with 11 physicochemical features and quality (score 0 to 10)

Data Wranling

- missing values?duplication?
- data type/range
- feature vs quality
- features vs features

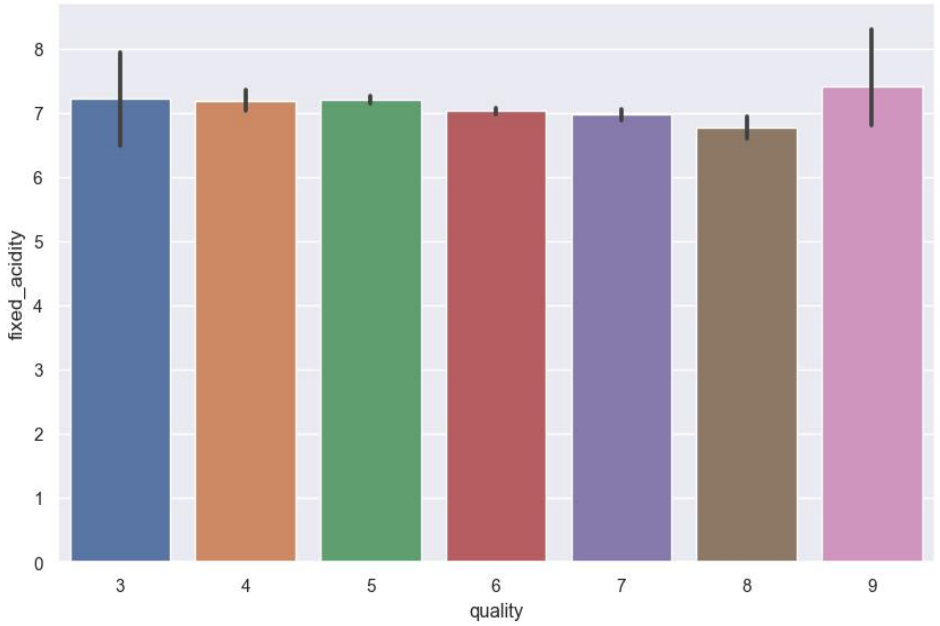


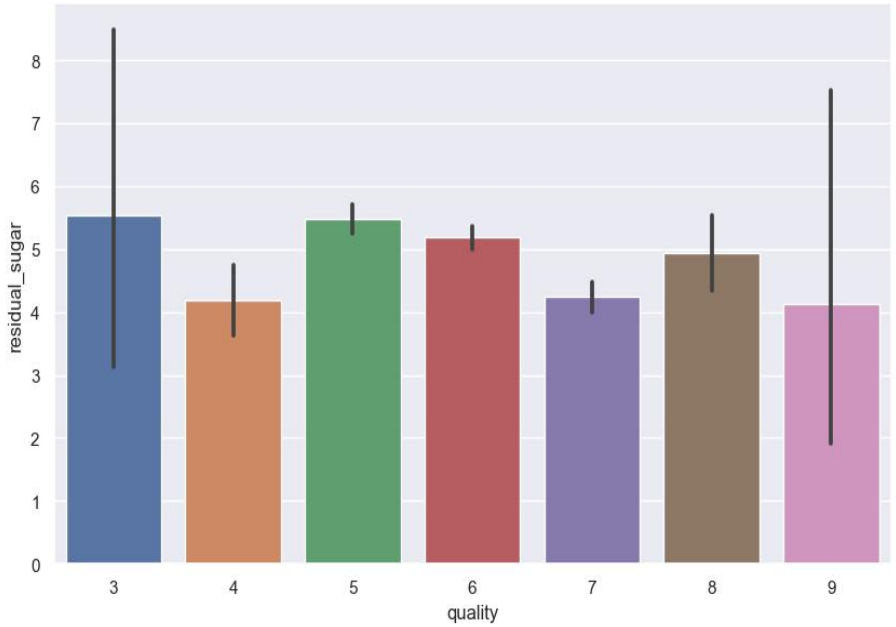




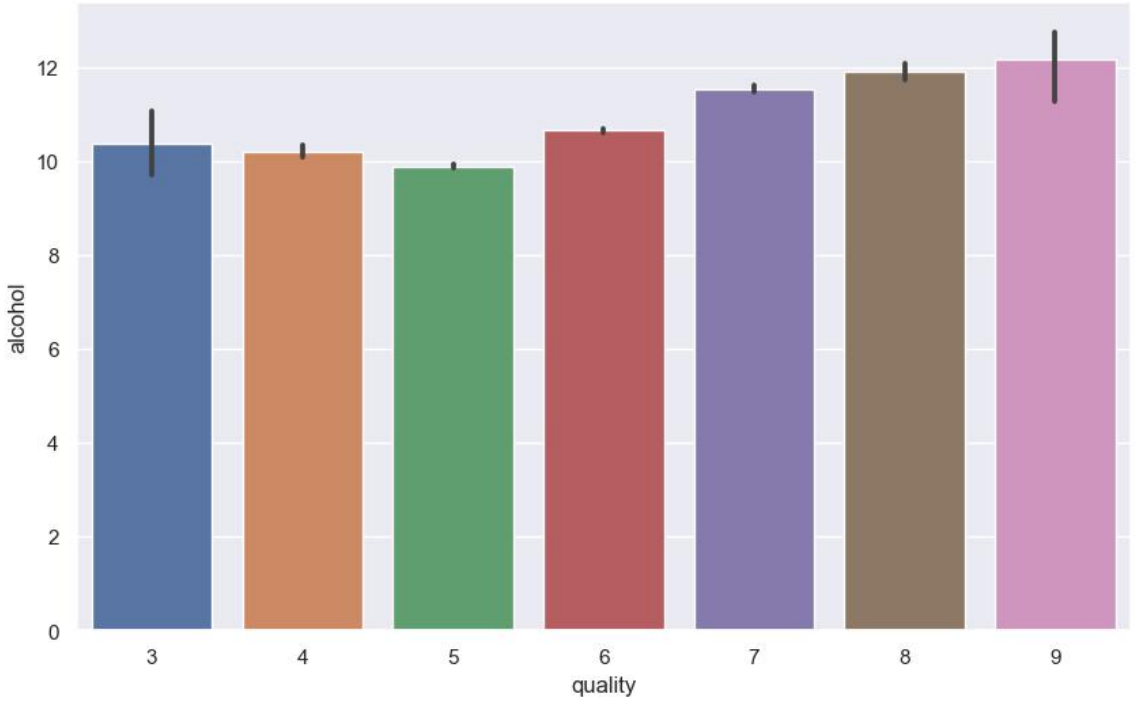
18% good wine quality (>=7),
76% wine quality of 5 & 6.

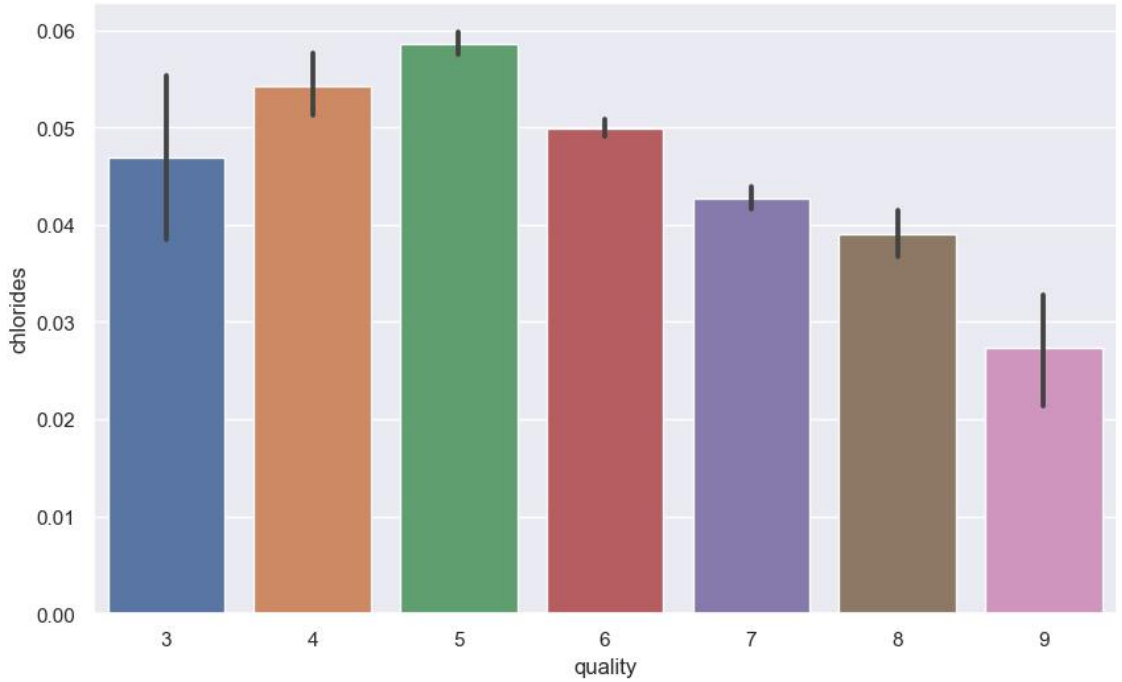
EDA





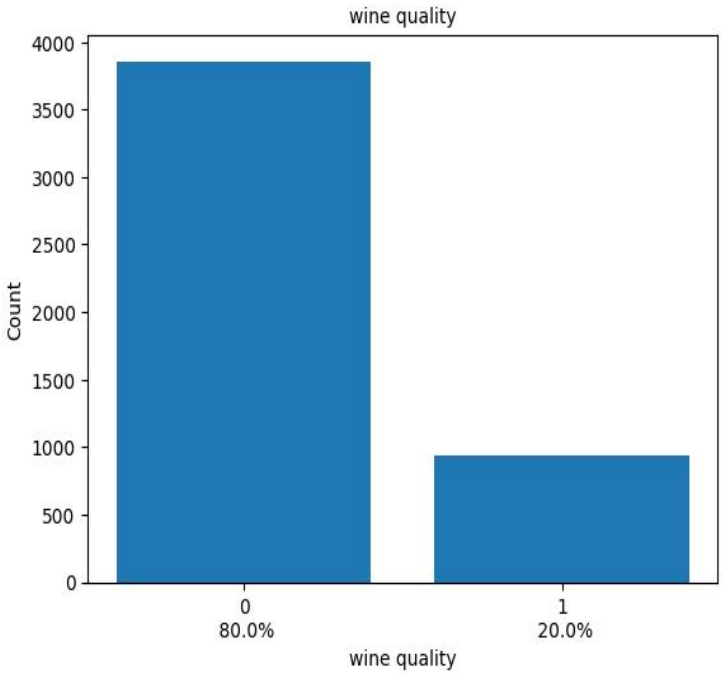
Remove outlier

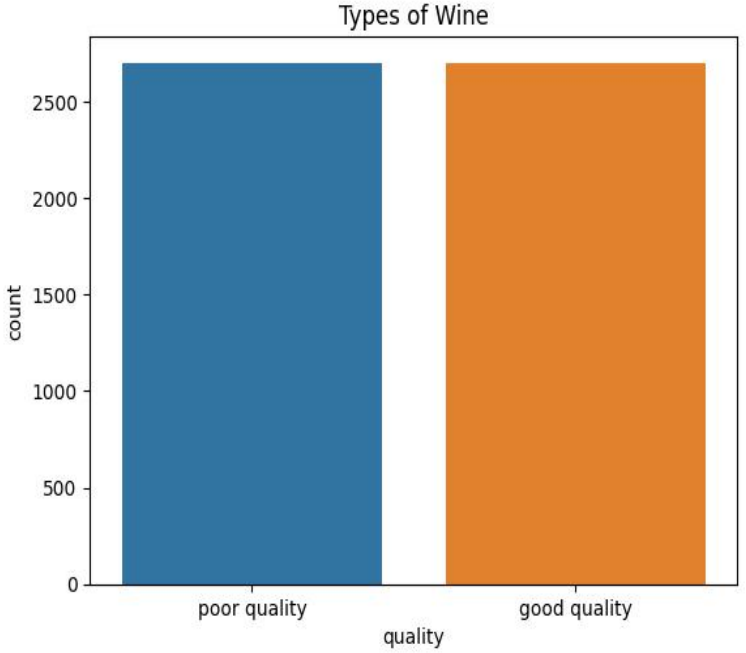

alcohol
0
2
4
6
8
10
12
3
4
5
6
7
8
9
quality


chlorides
0.00
0.01
0.02
0.03
0.04
0.05
0.06
3
4
5
6
7
8
9
quality

data preprocessing

- catagoreical featrue : get_dummy
- bucket “quality” into class 0,1
- scaling
- train/test split (0.7/0.3)
- SMOTE





Modeling

- Dummy Classifier (baseline)
- Logistic regression
- Random forest
- XGBoost

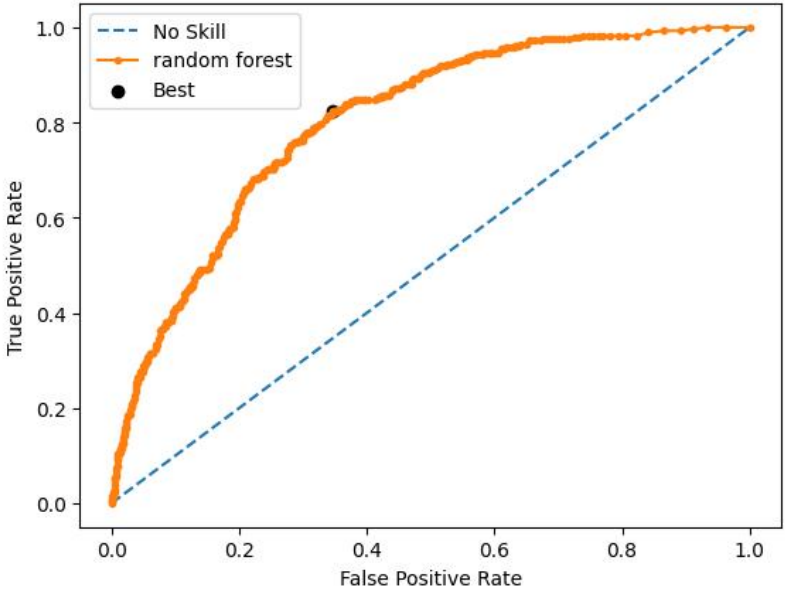
- modeling metrics: predict class 1 (positive): recall
- The stability of each modeling:Cross validation scores

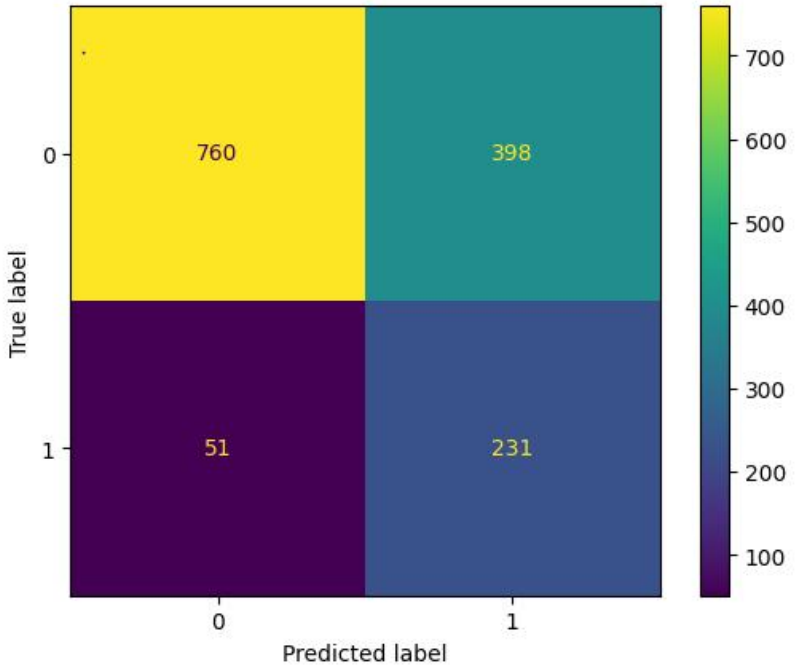
Model Metrics Results

- Pipeline(SelectKBest, Classifier)
- DummyClassifier: 0.31(recall)

	cv	score	roc_auc	precision	recall	accuracy
LogisticRegression		0.78	0.75	0.39	0.83	0.71
Randomforest		0.90	0.74	0.37	0.82	0.69
XGBoost		0.84	0.74	0.40	0.77	0.73

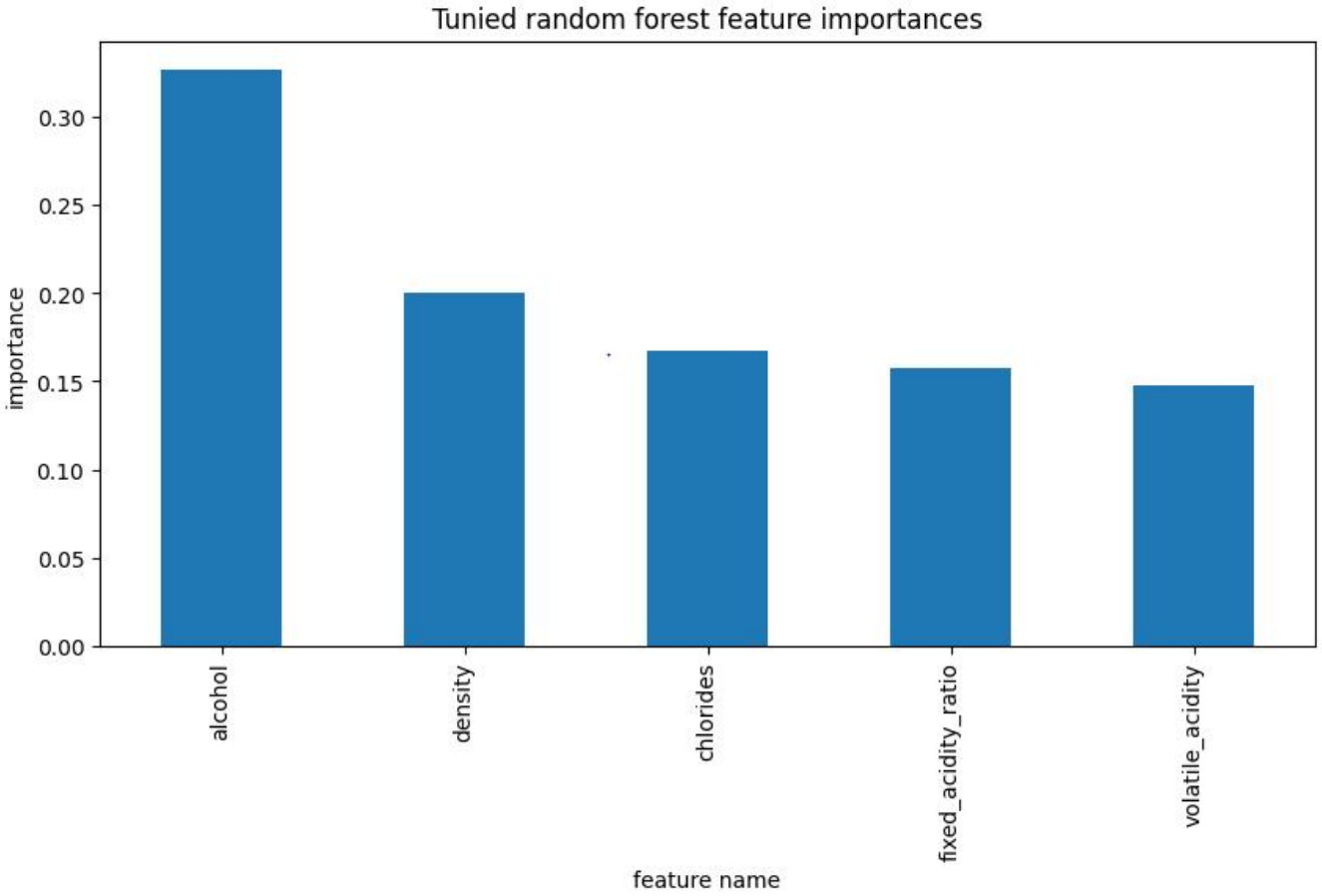
Model Metrics Results





	precision	recall	f1-score	support
0	0.94	0.66	0.77	1158
1	0.37	0.82	0.51	282
accuracy			0.69	1440
macro avg	0.65	0.74	0.64	1440
weighted avg	0.83	0.69	0.72	1440

Featuring Importance



Summary

- Out of 3 supervised classification models, thre random forest model is the best one. The recall score is 0.82.
- Out of 15 features, used 5 features for the best model using SelectKBest
- With more delicate feature engineering, the model can be improved in the future.

Future Work

- Will test the modeling if new data will be provided.
- Can the model reach higher recall? may go back to do more feature enginnering (address the skewed distribution), modify modeling parameters (k values in SelectKbest, learning rate)
- Will try other modeling such as SVM, KNN.