How to get drunk in style!

supervised machine learning algorithms to predict wine quality

Timo

Ironhack Bootcamp
Februray 2025

Project Overview

How to get drunk in style:

- dataset of wine features chosen
- quality of the wine rated according to these features
- decisive features selected and normalized
- supervised machine learning algorithms implemented to predict wine quality.

objective: allow users to select best wines for maximum pleasure

impact: everyone shall can get drunk in style!

Wine Quality Prediction

data set from kaggle by M Yasser H

usability score: 10.0

1143 entries

https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

11 key wine characteristics

Quality score (based on sensory data): 0 - 10

no name of the wines given! :-(

Comment: Portugese wines from Vinho Verde

https://en.wikipedia.org/wiki/Vinho_Verde

11 key wine characteristics: 4 categories

density: mass per unit volume, higher when sugar & alcohol higher = sweeter.
alcohol (% ethanol): higher = better balance & richness in flavor.
residual sugar: sugar left unfermented, affects sweetness.

pH: Low pH (acidic, 0-6) improves stability & freshness,
high pH (alkaline, 8-14) leads to flat & dull flavors.
fixed acidity: non-volatile acids (e.g. tartaric, malic acids)
-> moderate levels contribute to crisp & fresh taste,
too much can make the wine overly sour.
volatile acidity: mainly acetic acid -> causes vinegar taste
citric acid: enhances freshness, contributes to fruity taste.

free sulfur dioxide: free SO_2 = prevents oxidation & microbial growth,
preserves freshness, but excess gives unpleasant smell.
total sulfur dioxide: free + bound SO_2, preserve wine,
very high levels cause a pungent odor.
sulphates: Sulfur compounds act as preservatives,
enhance antimicrobial properties,
but excessive amounts can negatively affect taste.

chlorides: salt content, high levels give salty taste = undesirable

Data Cleaning & Preparation

Preprocessing of data for modelling:

- all column names conclusive & lower case
- all values numerical
- no null values
- no duplicated values

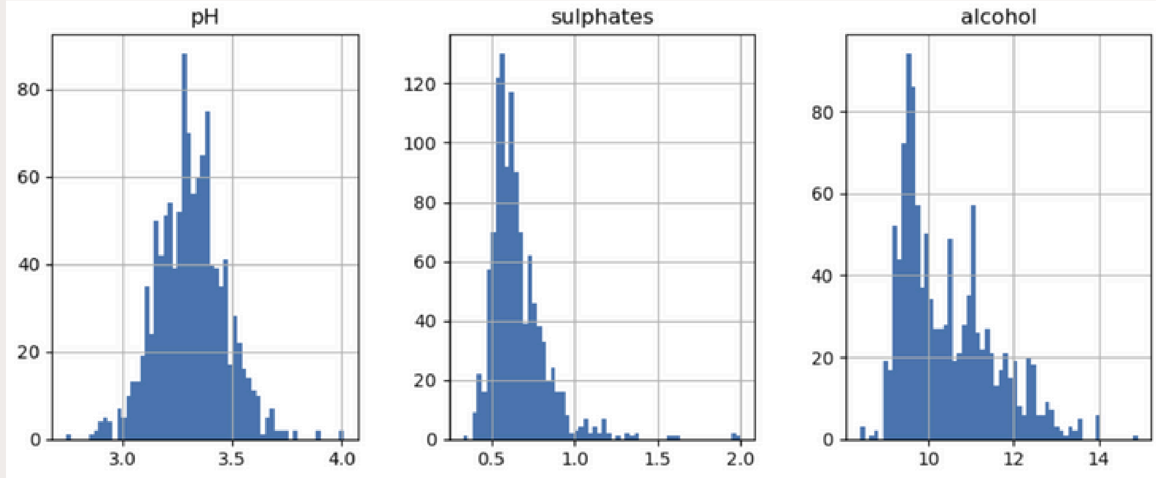
--> No extensive cleaning needed

--> More free time to get drunk early on Monday

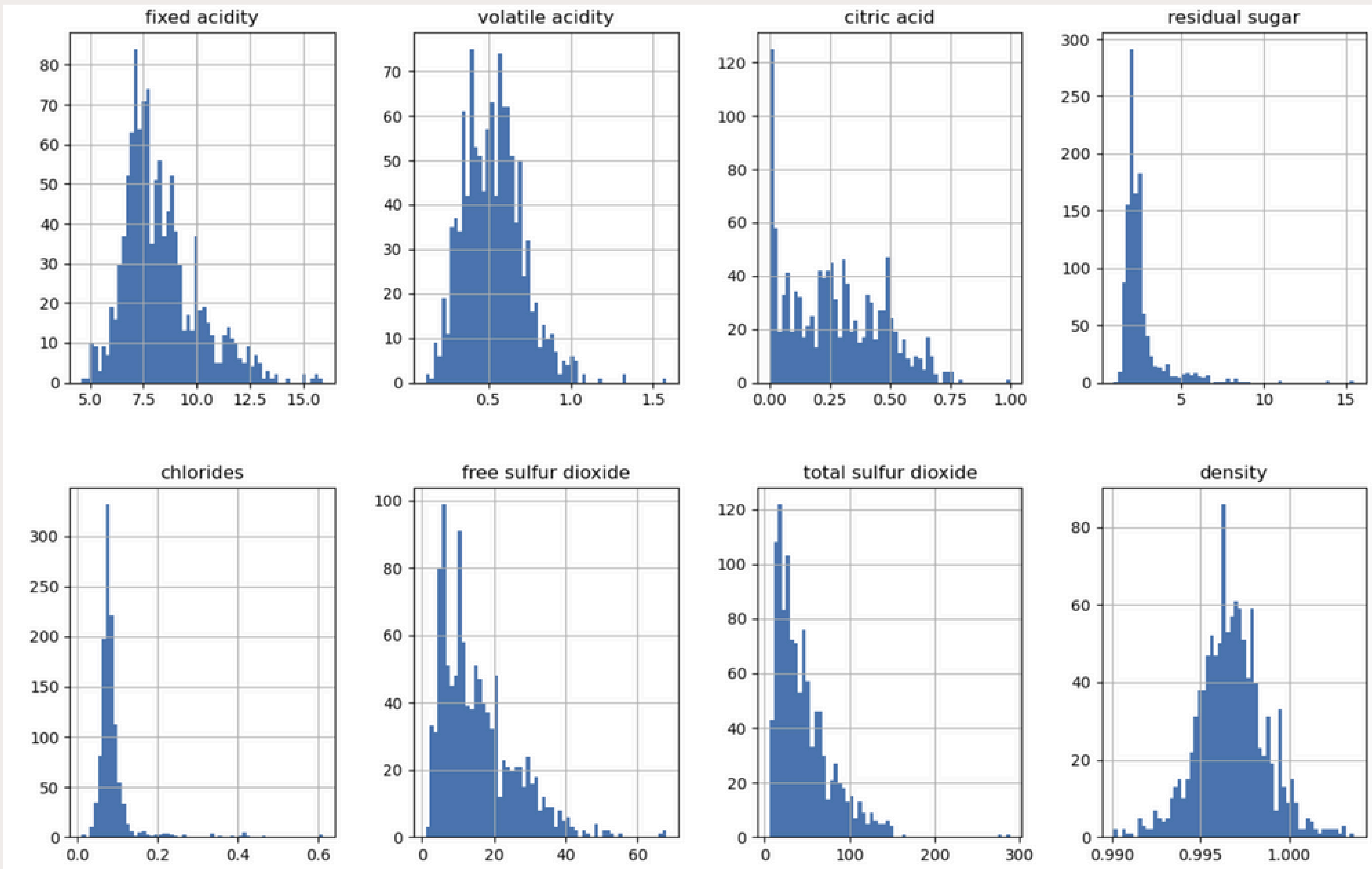
Exploratory data analysis

11 features in data

- histograms

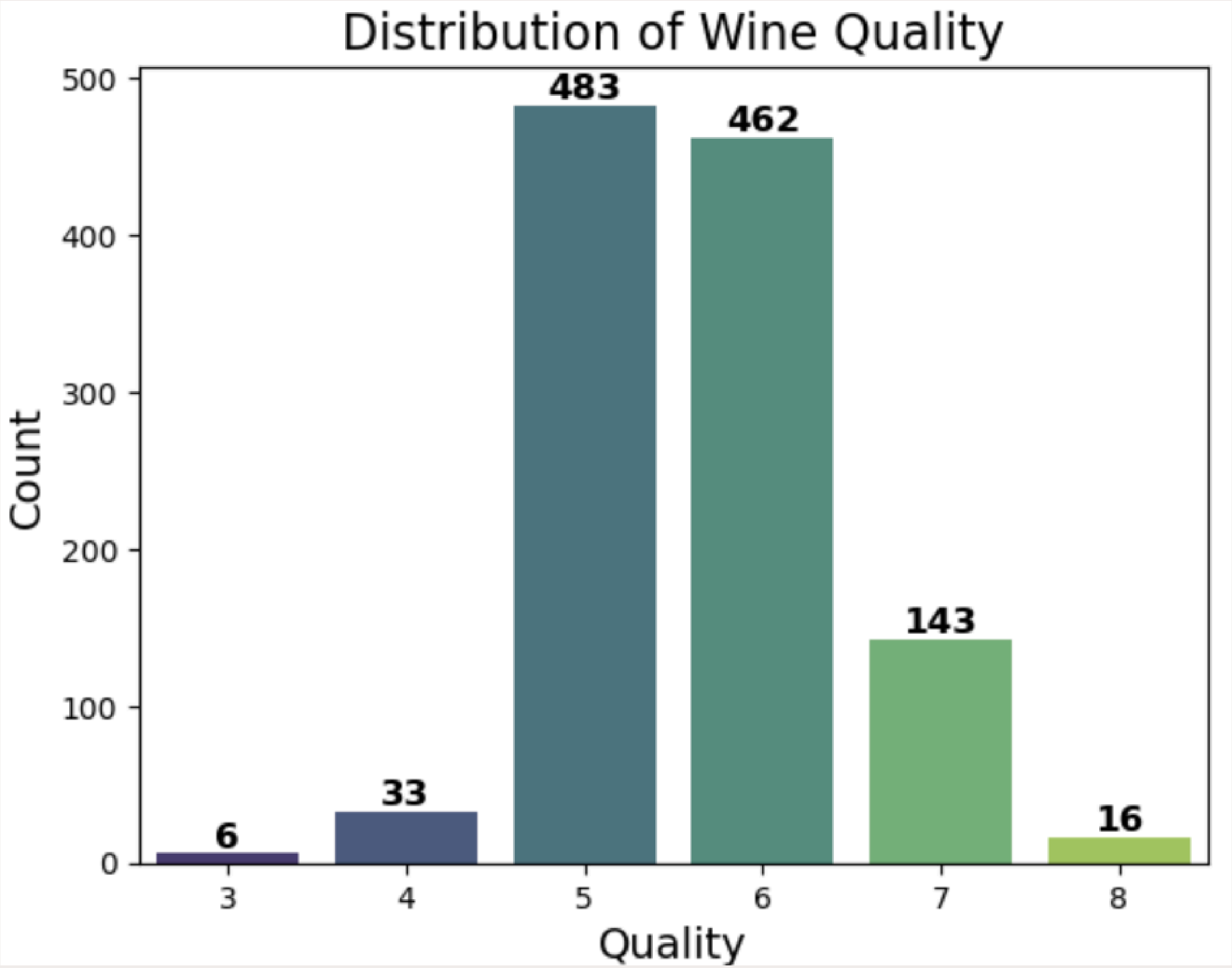


target is wine quality



Exploratory data analysis
- wine quality -

rating based on sensory data: 0 - 10

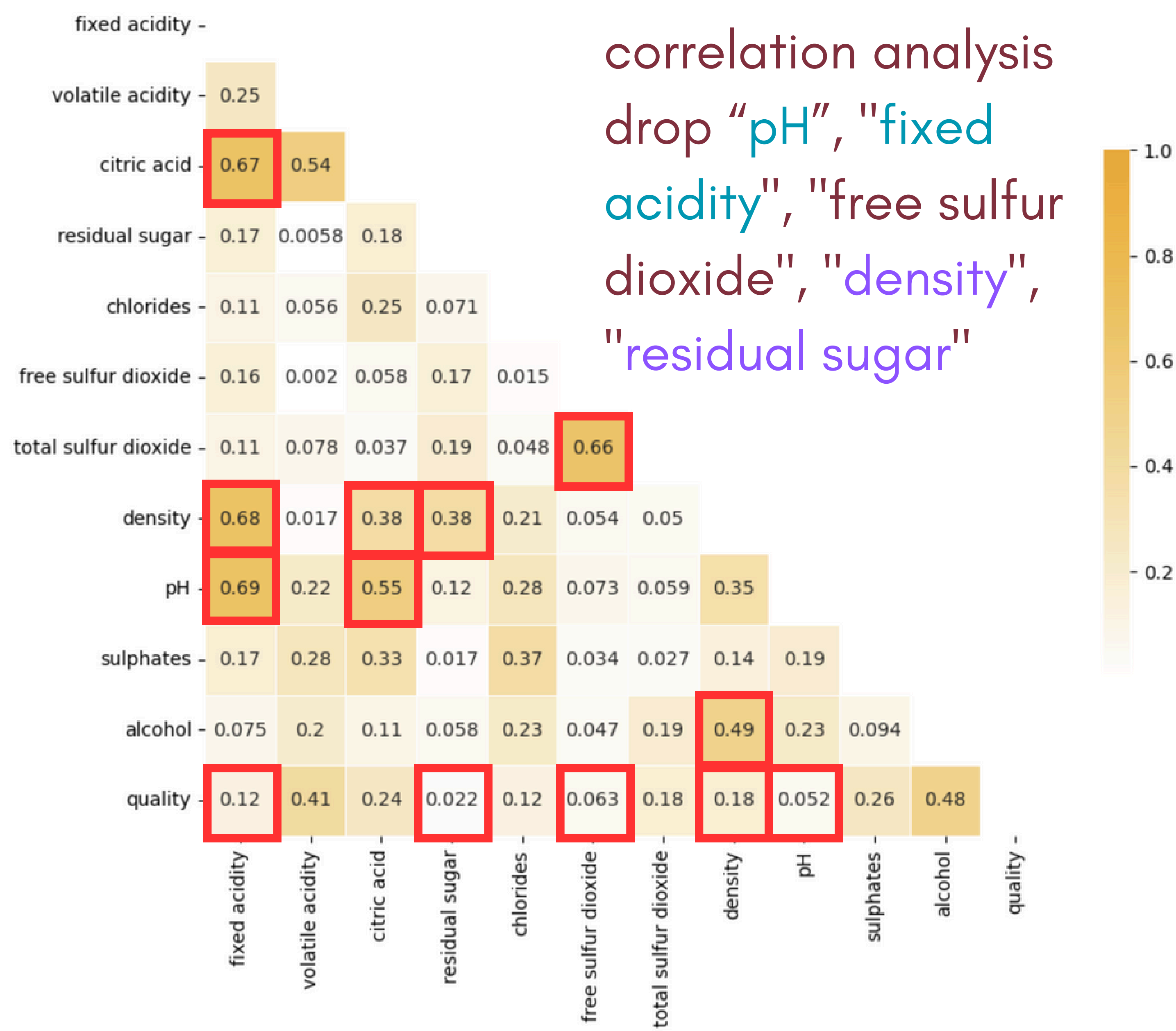


Feature Engineering & Selection

data set split: 80% training & 20% test

correlation analysis

drop “pH”, "fixed acidity", "free sulfur dioxide", "density", "residual sugar"



Model Building

3+4 different **classification models** to predict wine quality:

- **K Nearest Neighbour:** 11 non-norm features, 11 min-max-norm features, or 6 selected min-max-norm features
- **Logistic Regression:** 6 selected min-max-norm features
- **Decision Tree:** 6 selected min-max-norm features
- four different ensemble modelling techniques

ML Model Name	Features	Accuracy
KNN #1	11x non-normalized	45.4
KNN #2	11x MinMax-normalized	62.0
KNN #3	6x MinMax-normalized	61.6
Logistic Regression	6x MinMax-normalized	66.8
Decision Tree	6x MinMax-normalized	31.8
LogReg + Bagging	6x MinMax-normalized	65.9
Random Forest	6x MinMax-normalized	64.2
Gradient Boosting	6x MinMax-normalized	53.3
LogReg + Adapt. Boosting	6x MinMax-normalized	65.1

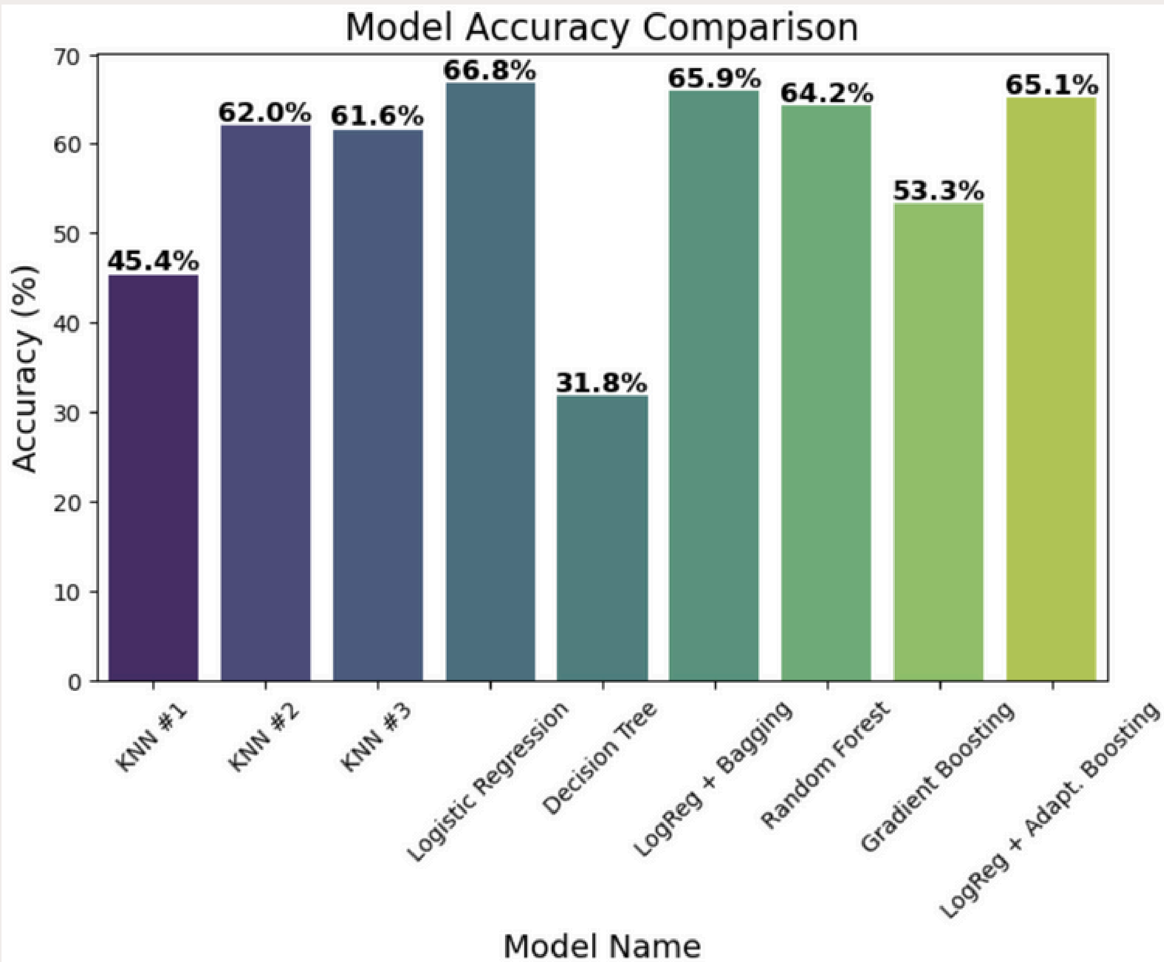
Model Evaluation

Model performance: **accuracy** of wine quality prediction:

LogReg: 66.8% > Ensemble: 53.3-65.9% >

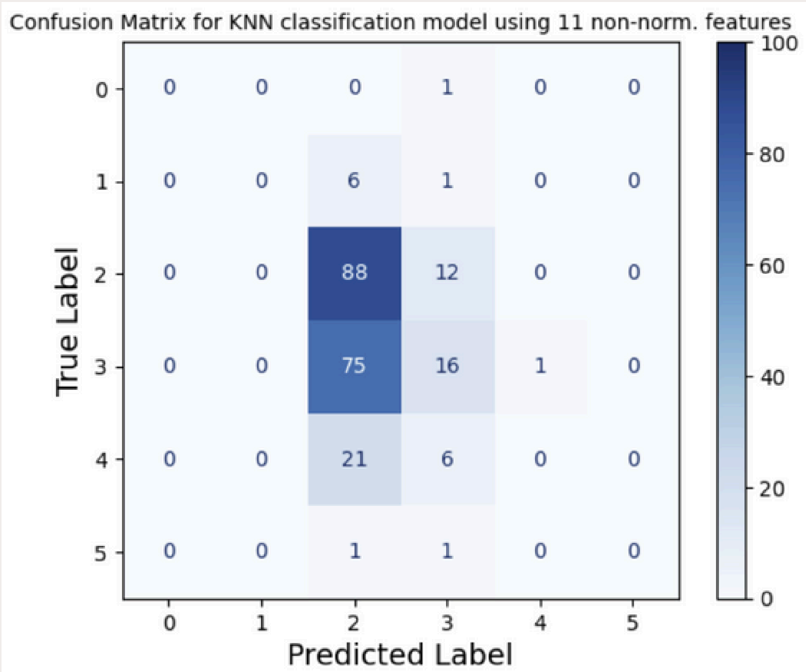
KNN: 45.4% -> **62.0%** -> **61.6% > Tree: 31.8%**

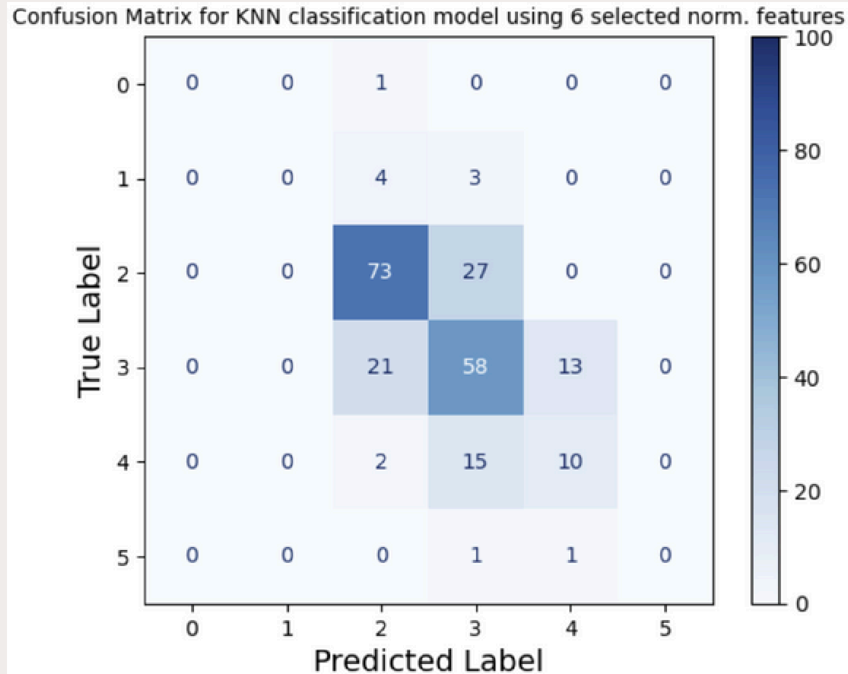
-> Logistic regression model results in best prediction!

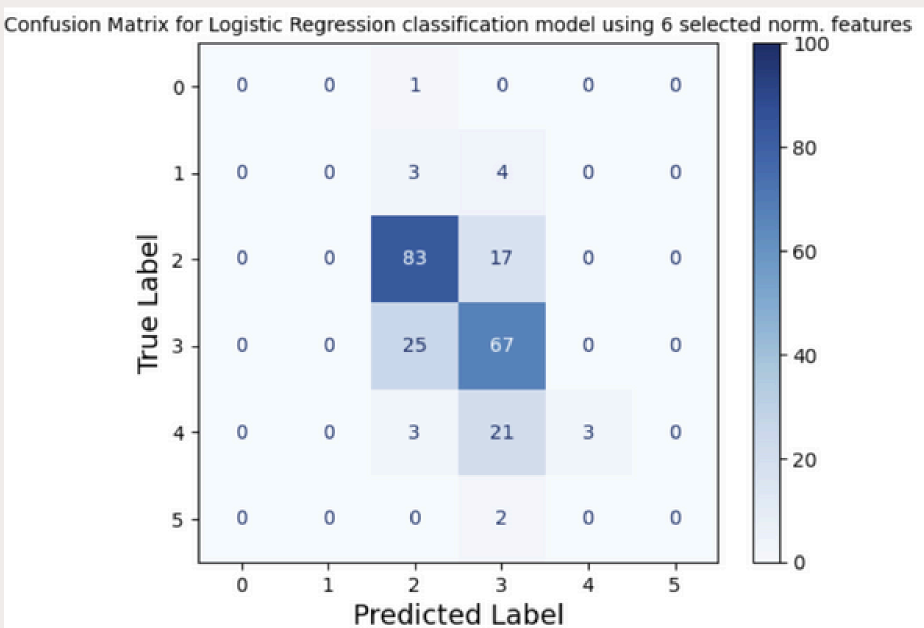


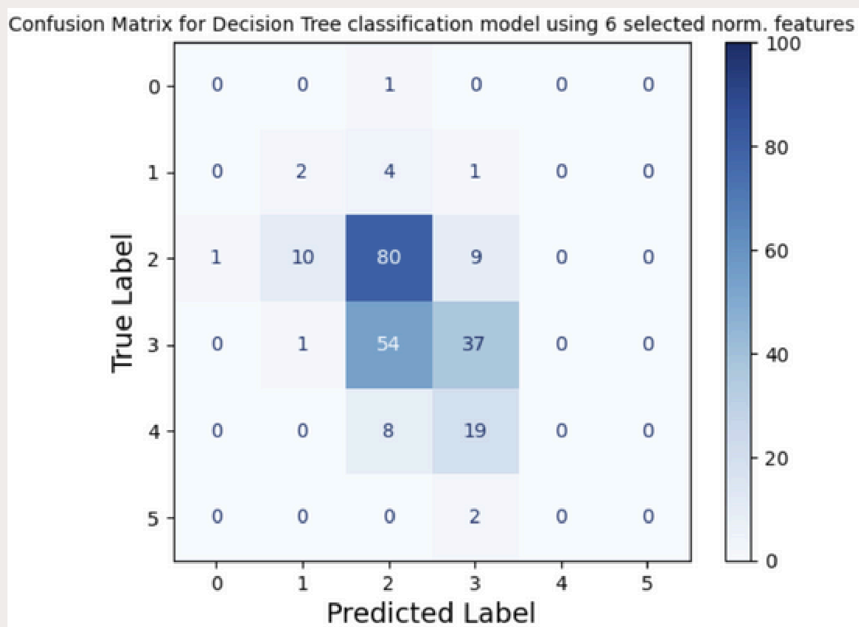
Model Evaluation

CONFUSION MATRICES









ML Model Optimization

Hyperparameters = settings controlling training process

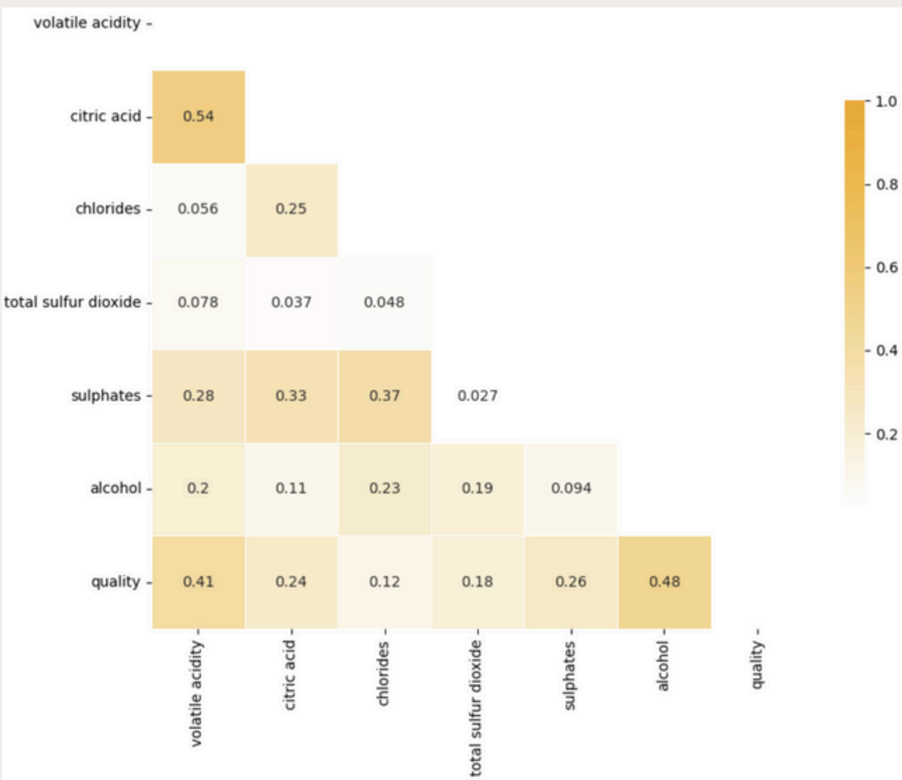
Hyperparameter tuning techniques employed:

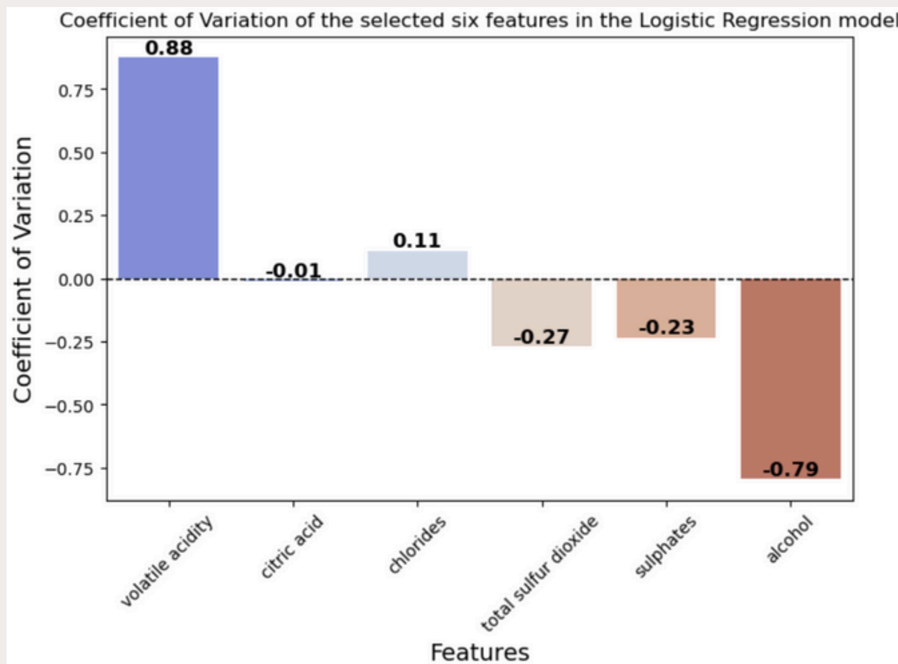
- Grid Search (not covered)
- Cross-Validation (not covered)

- KNN: n_neighbors=20 (low complexity)
- LogReg: default
- DecTree: max_depth=5 (high complexity)
- Bagging: n_estimators=100, max_samples = 500
- RandForest: n_estimators=100, max_depth=20
- GradBoost: max_depth=20, n_estimators=100
- AdaptBoost: n_estimators=50, learning_rate=1.0

Key Findings & Insights

- MinMax normalization improved prediction
- Model accuracy: **LogReg > 4x Ensemble > KNN > Tree**
- Feature effectiveness: correlation with wine quality
  %alc > vol.ac. > sulph. > cit.ac. > tot. sulf. diox. > chlor.
- LogReg model Coeff.Var:
  vol.ac. > %alc > tot. sulf. diox. > sulph. > chlor. > cit.ac.





Real-World Application & Impact

Application of Wine Quality Prediction:

- Make informed decision when buying wine
- Get drunk using good quality wine

Ethical Considerations & Limitations:

- higher prevalence of wine quality 5-7
- no normal distribution of wine qualities
- low quality wines may turn sour in shelves

Challenges & Learnings

Challenges faced:

- depression cause working alone -> drink wine
- developing ideas -> drink wine
- getting info & advice -> ask ChatGPT
- time management

Key learnings:

- Chillax! Focus!
- You can do it if you really want,
- but you must try...

Future Work & Improvements

Future Work & Expansion of Project:

- Add wine names
- Add distributers

-> facilitate access for users

- Increase number of wines listed
- Increase number of features
- Model optimization applying hyperparameter tuning techniques

-> increase accuracy of prediction

Now you know what to consider when you want to get drunk in style!

Thank you !

Timo

Ironhack Bootcamp
Februray 2025