Homework 2 - Part 1 and Part 2

Part 1.

Completing Homework 1 provided me with valuable insights into the practical challenges and nuances of building machine learning models. One key takeaway was the importance of **data preprocessing and exploration**. Initially, I underestimated the impact of missing values, inconsistent data formatting, and scaling issues, but through feedback, I recognized how essential it is to carefully examine and clean the data before modeling. This step ensures the reliability of results and prevents misleading interpretations.

Another lesson I learned was about the **importance of model evaluation techniques**. In the first assignment, I primarily relied on basic accuracy as a performance metric. However, I now understand the limitations of accuracy, especially in the presence of imbalanced datasets. I realized that relying solely on accuracy could mask poor performance in minority classes. As a result, I have learned to consider more comprehensive metrics like **AUC, precision, recall, and F1-score for** a more nuanced assessment.

Additionally, the feedback helped me appreciate the role of **model interpretability**. It's not enough to build a model that performs well; it's equally important to understand how the model makes its predictions. This awareness will guide my future model selection — especially in real-world scenarios where explainability is crucial, such as in healthcare or finance.

Lastly, I learned the value of **documenting my thought process**. Whether it's choosing a specific model, handling outliers, or selecting a performance metric, articulating the reasoning behind my choices is just as important as the technical implementation. This not only improves communication with peers and stakeholders but also enhances my own critical thinking.

In summary, Homework 1 helped me grow both technically and analytically. Moving forward, I will pay greater attention to **data quality, evaluation metrics, interpretability, and clear communication**, all of which are essential components of successful data science projects.

Part 2: Model Card

Property	Decision Tree	Naive Bayes	K-Nearest Neighbor	Logistic Regression	SVM
Parametric/Non-parametric	Non-parametric	Parametric	Non-parametric	Parametric	Non-parametric
Input	Both	Both	Both	Both	Both
Output	Discrete or continuous	Discrete	Discrete	Discrete	Discrete
Handle Missing Value	No	No	No	No	No
Model Representation	Tree structure	Probabilistic (Bayes rule)	Lazy learner (memory-based)	Linear equation	Hyperplane
Model Parameters	Depth, Split criteria	Priors, Likelihoods	Number of neighbors (k)	Coefficients, Intercept	Kernel type, C, gamma
Make the Model More Complex	Increase depth	Add features	Increase k	Add interaction terms	Use non-linear kernel (RBF)
Make the Model Less Complex	Prune tree	Remove features	Decrease k	Regularization (L1/L2)	Regularization, linear kernel
Interpretability/Transparency	High	Medium	Low to Medium	High	Low

Part 3.

YData Profiling Report

Overview Variables Interactions Correlations Missing values Sample Duplicate rows

Overview

Brought to you by YData

Overview Alerts 4 Reproduction

Dataset statistics

Number of variables	4
Number of observations	571
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	13
Duplicate rows (%)	2.3%
Total size in memory	42.8 KiB
Average record size in memory	76.8 B

Variable types

Numeric	3
Categorical	1

YData Profiling Report

Overview Variables Interactions Correlations Missing values Sample Duplicate rows

Variables

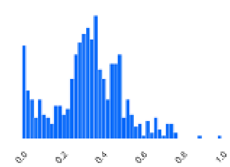
Select Columns

citric acid

Real number (ℝ)

Zeros

Distinct	74	**Minimum**	0
Distinct (%)	13.0%	**Maximum**	1
Missing	0	**Zeros**	25
Missing (%)	0.0%	**Zeros (%)**	4.4%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.32467601	**Memory size**	4.6 KiB



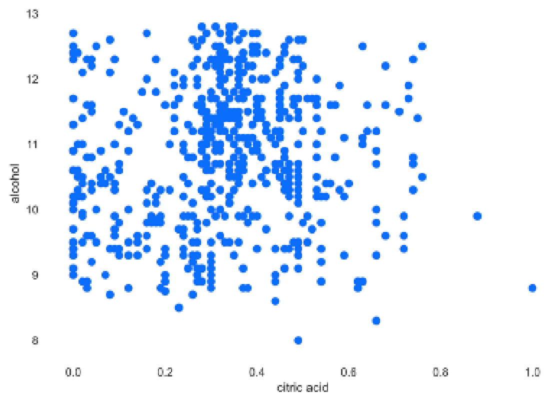
More details

YData Profiling Report   Overview Variables Interactions Correlations Missing values Sample Duplicate rows

Interactions

citric acid | sulphates | alcohol

alcohol | citric acid | sulphates

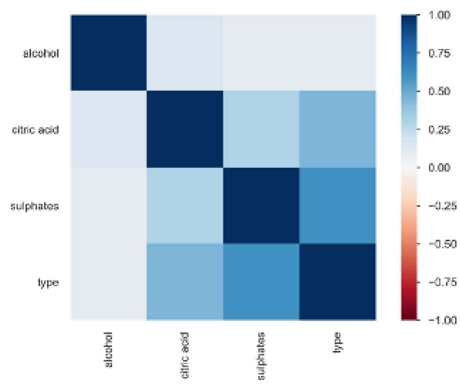


YData Profiling Report   Overview Variables Interactions Correlations Missing values Sample Duplicate rows

Correlations

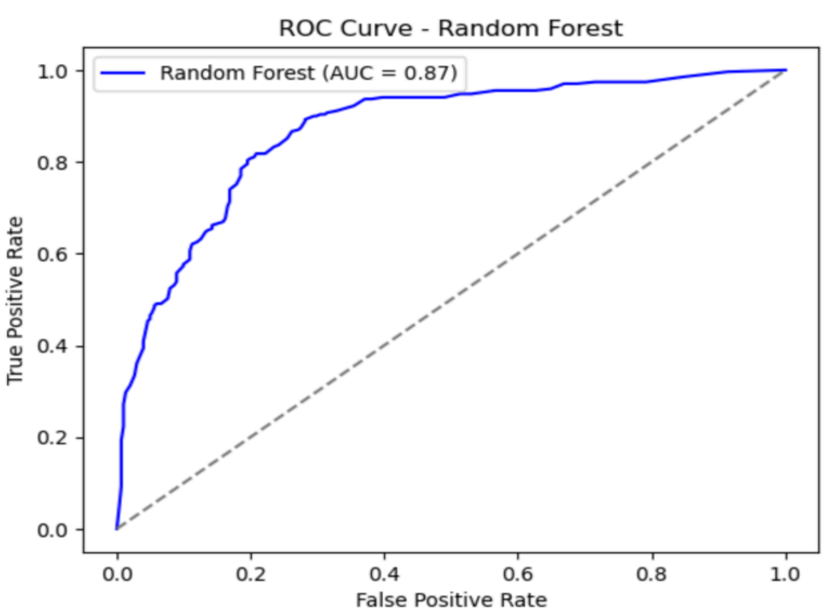
Auto

Heatmap | Table



```
Baseline Accuracy: 52.89
Baseline AUC: 0.5

Model: Logistic Regression
Accuracy: 79.51
AUC: 0.8709470937246117

Model: Naive Bayes
Accuracy: 82.14
AUC: 0.882481104901647

Model: Decision Tree
Accuracy: 76.18
AUC: 0.7628819025579162

Model: SVM–Linear
Accuracy: 78.98
AUC: 0.8710332602969053

Model: SVM–RBF
Accuracy: 82.31
AUC: 0.9098820748910608

Model: Random Forest
Accuracy: 80.04
AUC: 0.8660909919003422
```

Model	Baseline	Logistic Regression	Naive Bayes	Decision Tree	SVM - Linear	SVM - RBF	Random Forest
AUC	0.50	0.87	0.88248	0.76021	0.87130	0.91059	0.86945
Accuracy	52.89%	79.51%	82.14%	75.83%	78.98%	82.31%	80.04%



4.

```
   citric acid  sulphates  alcohol type
0         0.24       0.52      9.4  low
1         0.49       0.56      9.4  low
2         0.66       0.73     10.0  low
3         0.32       0.77     10.0  low
4         0.38       0.82     10.0  low
Accuracy: 81.25
AUC Score: 0.9455
```

Accuracy : 81.25

AUC score : 0.9455

5. I would recommend using a Decision Tree or Logistic Regression model.

These models are much easier to understand and explain. For example:

A Decision Tree shows a clear set of “if-then” rules (like “if alcohol > 10%, then likely red”), which is great for helping wine experts see how the model is making its decisions.

Logistic Regression gives simple weights to each feature, so experts can see exactly which chemical properties (like sulphates or alcohol) are influencing the prediction the most.

Models like SVM or Random Forest might be more accurate in some cases, but they work more like black boxes — they make decisions without giving much insight into why.

So for wine-tasting experts who want to understand the reasoning behind predictions, simpler and more transparent models like Decision Trees and Logistic Regression are the better choice.