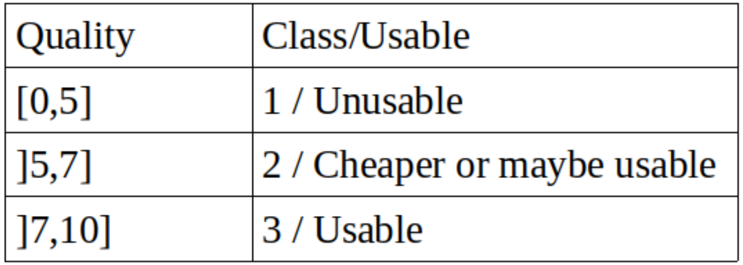
Jonne Kurokallio

Python case / Pre-assignment

Data

- This dataset is public available for research. The details are described in [Cortez et al., 2009]. Please include this citation if you plan to use this database:
  - P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.
- Wine database with thousands of wine samples with 11 features such as pH and acidity.
- Qualities evaluated by tree experts
- **Problem: determine usability of future samples.**
- It is unnecessary to divide the wine into 10 categories, so I divide it into three classes based on it’s usability.

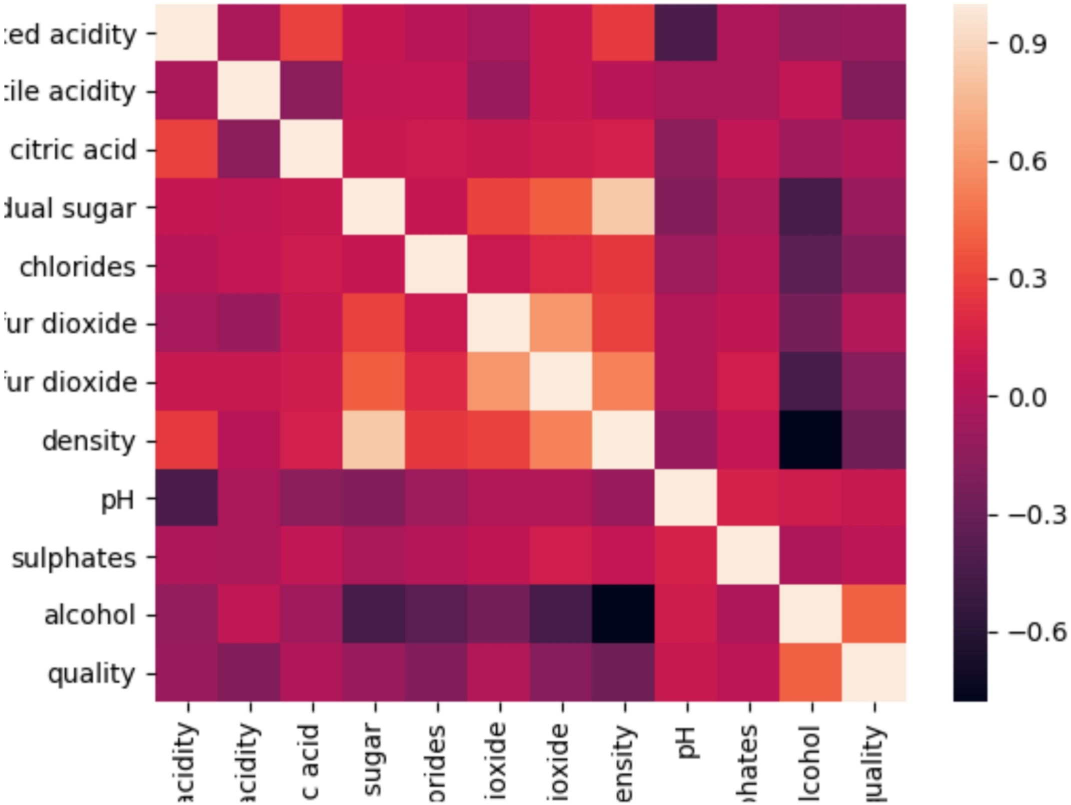


Quality	Class/Usable
[0,5]	1 / Unusable
]5,7]	2 / Cheaper or maybe usable
]7,10]	3 / Usable

Preprocessing the data

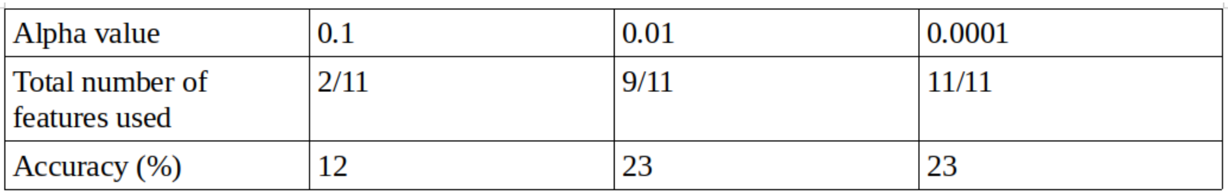
- Processed with Pandas
- Scaling the features
  - Scaling subtracts the mean value of the observation and then divides it by the unit variance of the observation
  - This eliminates the dominating (if exist) feature values
- Checking linearity
- Feature selection
  - 1. with Lasso before trying models
  - 2. with ExtraTreeClassifier before trying the models
  - 3. with Recursive Feature Elimination after training the model to see if this makes any difference

Linear equivalence



Feature selection

- Lasso

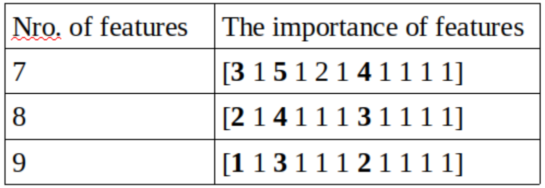


Alpha value	0.1	0.01	0.0001
Total number of features used	2/11	9/11	11/11
Accuracy (%)	12	23	23

- ExtraTreeClassifier

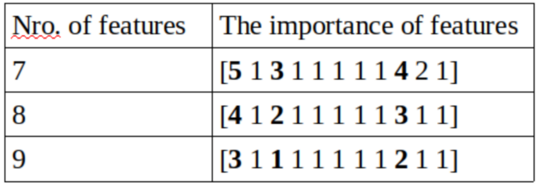
[0.07217551 0.10953069 0.0845641 0.08082555 0.0829807 0.09360738 0.08148188 0.07966577 0.07944639 0.07345181 0.16227022]

- Recursive Feature Elimination
  - SVM- linear



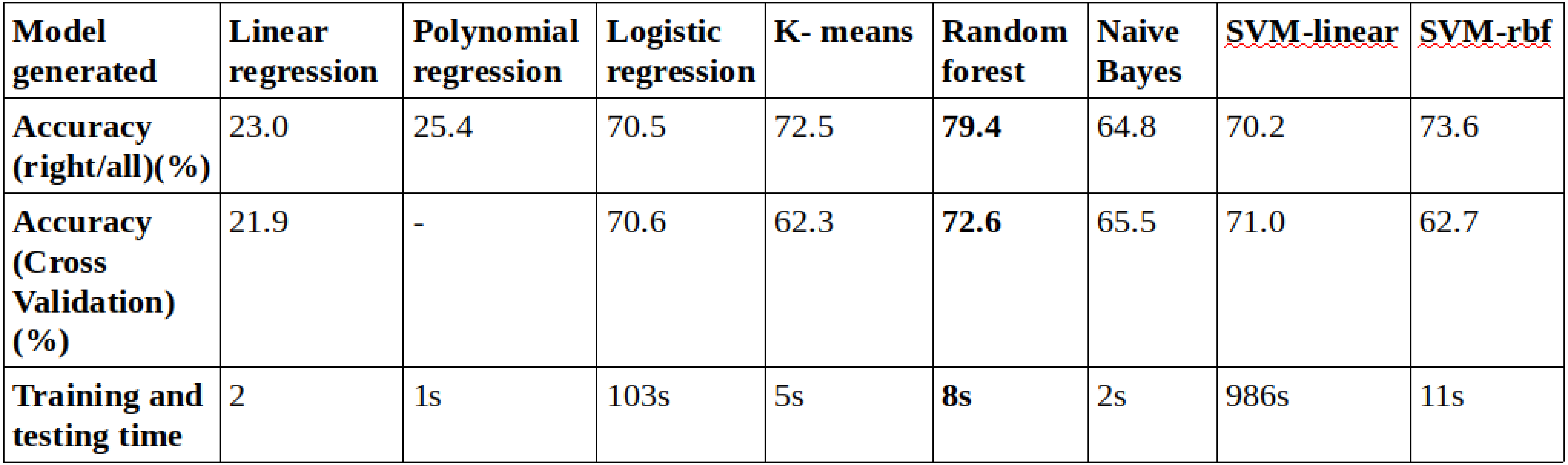
Nro. of features	The importance of features
7	[**3** 1 **5** 1 2 1 **4** 1 1 1 1]
8	[**2** 1 **4** 1 1 1 **3** 1 1 1 1]
9	[**1** 1 **3** 1 1 1 **2** 1 1 1 1]

  - Random Forest



Nro. of features	The importance of features
7	[**5** 1 **3** 1 1 1 1 1 **4** 2 1]
8	[**4** 1 **2** 1 1 1 1 1 **3** 1 1]
9	[**3** 1 **1** 1 1 1 1 1 **2** 1 1]

Trained models without feature selection



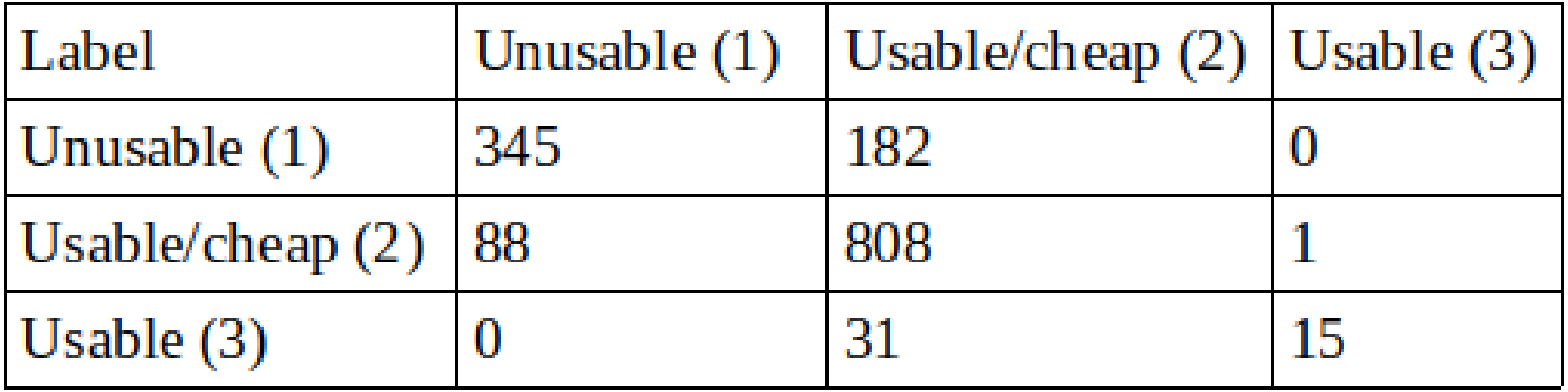
Model generated	**Linear regression**	**Polynomial regression**	**Logistic regression**	**K- means**	**Random forest**	**Naive Bayes**	**SVM-linear**	**SVM-rbf**
Accuracy (right/all)(%)	23.0	25.4	70.5	72.5	**79.4**	64.8	70.2	73.6
Accuracy (Cross Validation)(%)	21.9	-	70.6	62.3	**72.6**	65.5	71.0	62.7
Training and testing time	2	1s	103s	5s	**8s**	2s	986s	11s

RF and SVM-linear with feature selection

Model	Random forest	SVM-linear
Accuracy (right/all)(%)	78.8	69.8
Accuracy (Cross Validation) (%)	72.0	71.0
Training and testing time	7s	1478s

Analysis of the results

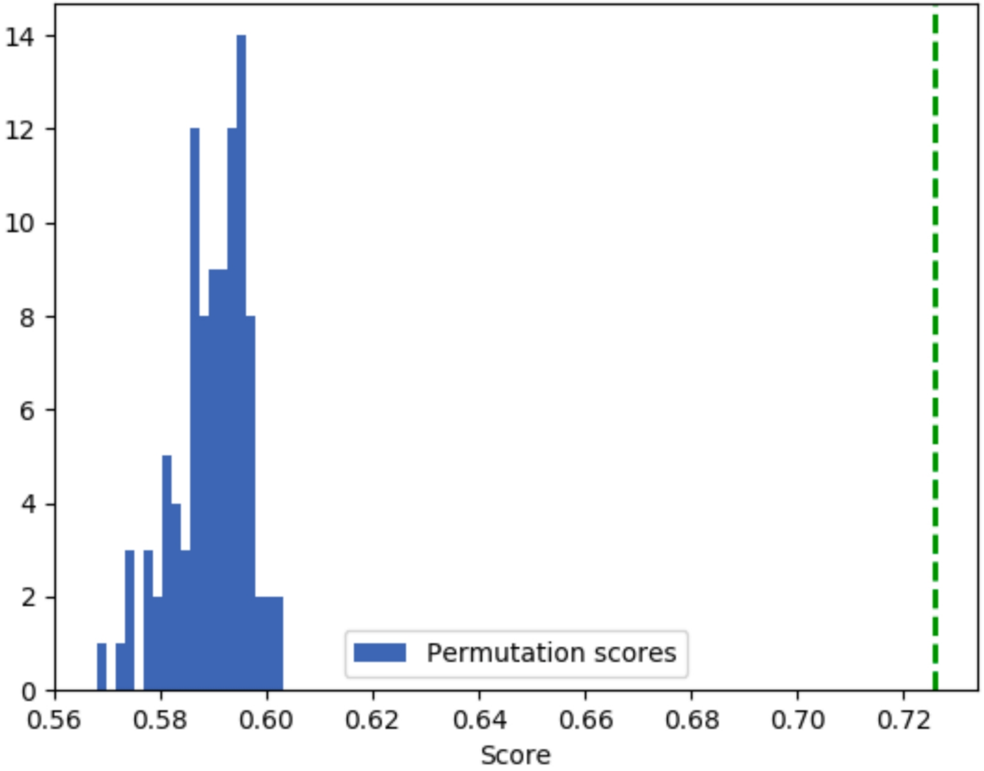
- Confusion matrix (Random Forest with out Feature selection)



Label	Unusable (1)	Usable/cheap (2)	Usable (3)
Unusable (1)	345	182	0
Usable/cheap (2)	88	808	1
Usable (3)	0	31	15

Permutation analysis

- Random Forest wi
- p.value = 0.009



Future work

- More study on the feature selection
  - Better and more accurate results with fewer features
- Try different scaling method
- Tune in more parameters