# "Vinho Verde" Wine Quality Classification

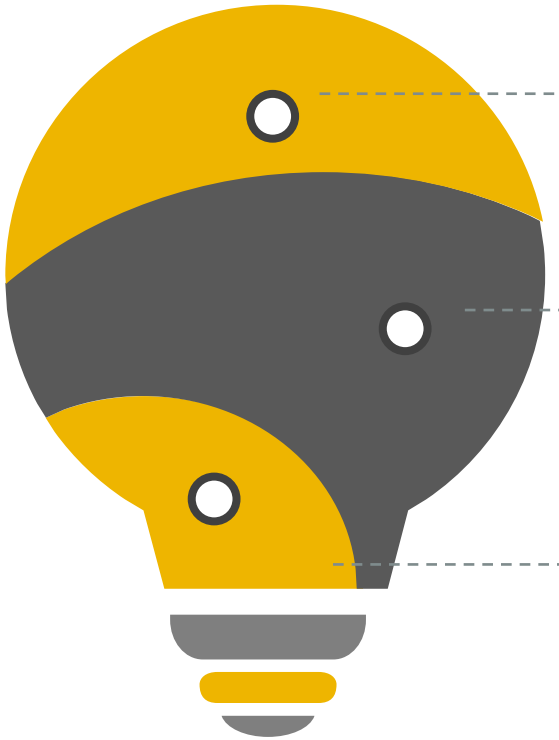
**Group 9**
Meng Wang
Nicholas Yang
Rush Bhardwaj
Yi Cai

Intro to Business Analytics Final Project

# Agenda

- Business Understanding
- Data Exploration & Preparation
- Model Building
- Model Evaluation
- Deployment & Conclusion

# BUSINESS UNDERSTANDING



## BUSINESS CONTEXT

- U.S. wine consumption has grown since 1993 ①

## BUSINESS PROBLEM

- Ambigious classification of wine

## BUSINESS GOALS

- Classification
- Process and quality control
- Pricing
- Distribution

# Data Exploration

Explore the dependent variables

## Data Source and Description

- The Red wine quality data includes a dataset of Portuguese "Vinho Verde" wine sample. The dataset has 1599 instances.
- This dataset is originally from UCI machine learning repository on both white and red wine samples (UCI Machine Learning Repository: Wine Quality Dataset. (n.d.)). We only focus on the red wine dataset.
- Only physicochemical condition of the wine are considered

## Dependent Variables

Different scales: Need standardize for k-NN model

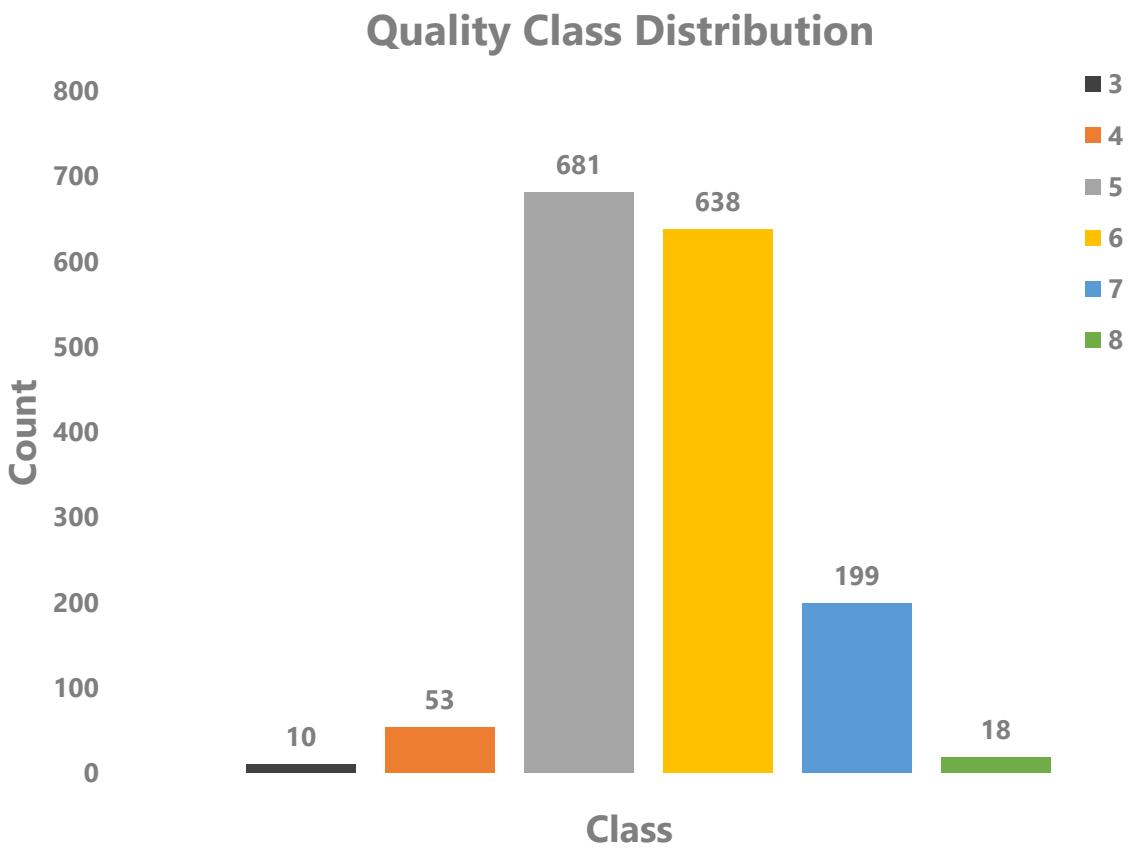
No missing values

No specific outliers need to be considered

| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol |
|---|---|---|---|---|---|---|---|---|---|---|---|
| max | 15.90 | 1.58 | 1.00 | 15.5 | 0.61 | 72.00 | 289.00 | 1.00 | 4.01 | 2.00 | 14.90 |
| mean | 8.32 | 0.53 | 0.27 | 2.53 | 0.09 | 15.87 | 46.47 | 0.99 | 3.31 | 0.66 | 10.42 |
| std | 1.74 | 0.18 | 0.19 | 1.40 | 0.05 | 10.46 | 32.90 | 0.01 | 0.15 | 0.17 | 1.07 |
| min | 4.60 | 0.12 | 0.00 | 0.90 | 0.01 | 1.00 | 6.00 | 0.99 | 2.74 | 0.33 | 8.40 |

# Data Exploration

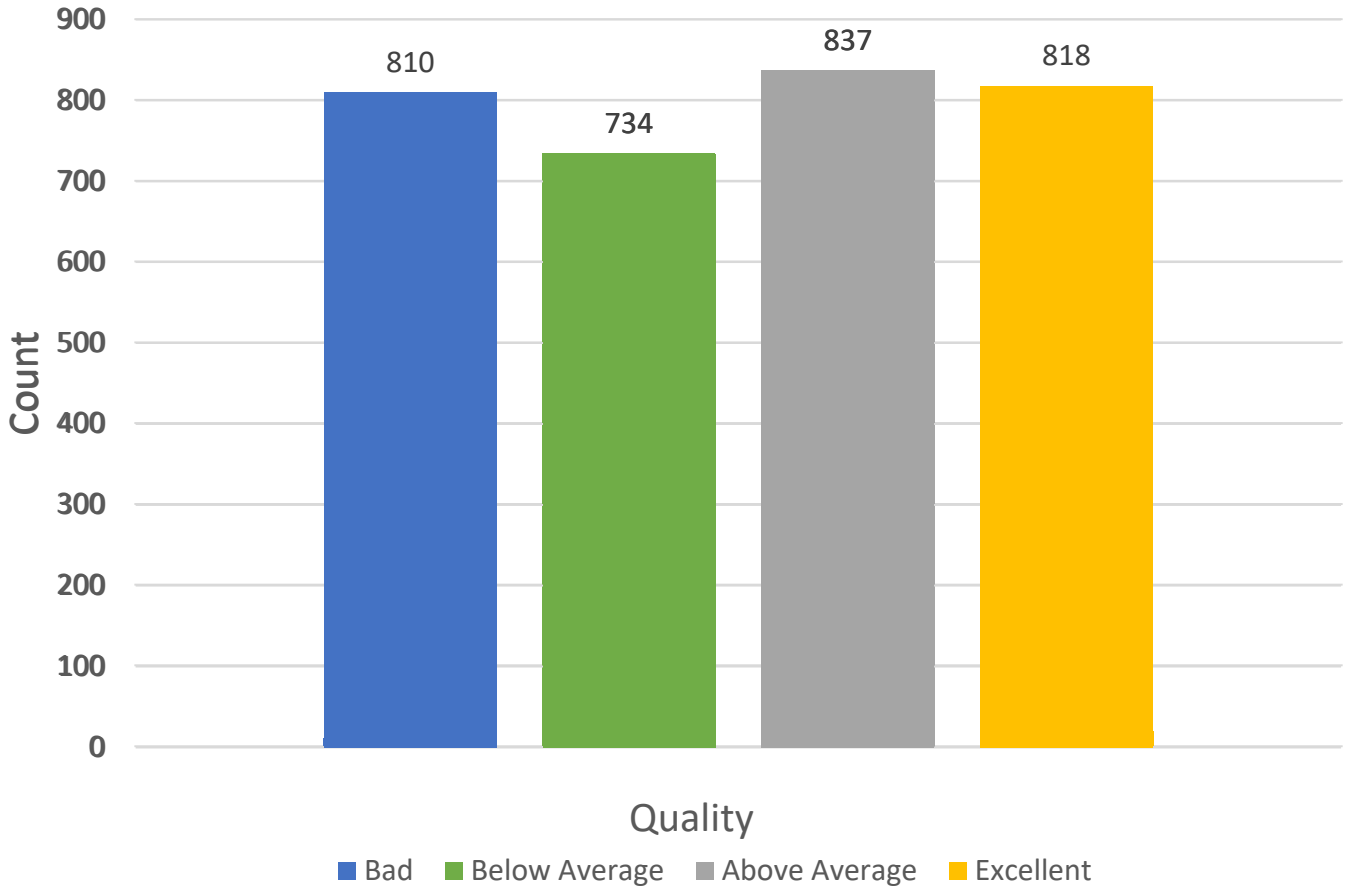
Explore the target variable- Wine Quality

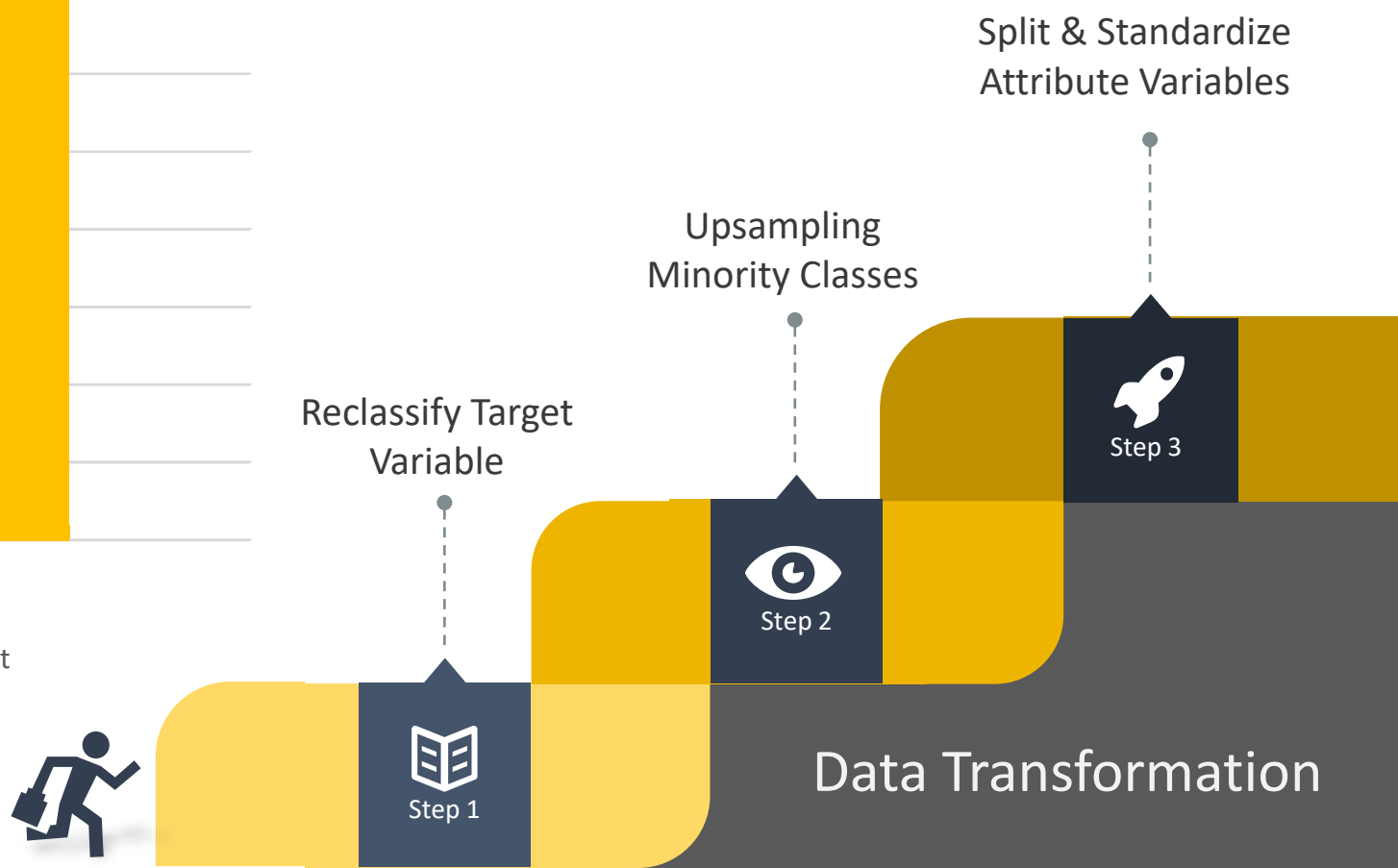



## Data Description

- The dataset split the wine quality into 10 classes ranging from 1 to 10
- The dataset is very unbalanced: no records on class 1, 2, 9, 10
- The most prevalent class is class 5 with 681 instances, which accounts for 0.43% of the whole dataset

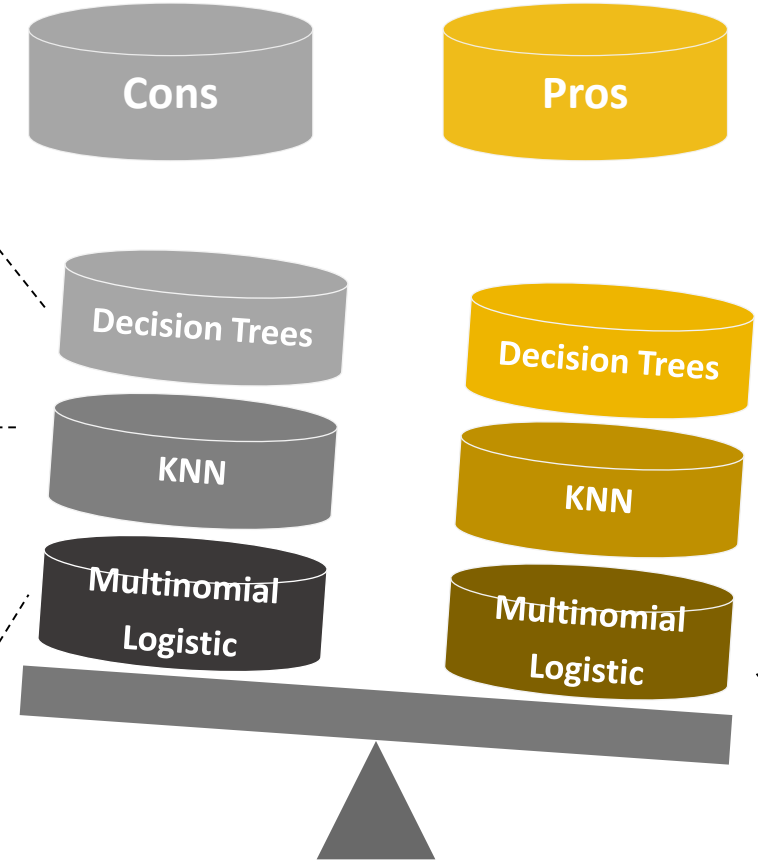
# Data Preparation



# Classification Methods

Decision Making in Classifying Wine Quality: Multi-class Target Attribute




1. Sampling Errors
2. Locally Optimization (Overfit)

1. Computational Cost
2. Data Storage
3. Lazy Learner

1. Hard to understand
2. Discrete class only

1. Easy to interpret & understand
2. Big data size not required
3. Continuous Output

1. Easy to interpret
2. Capture interactive relationships without model building
3. Non-linear

1. Benchmark
2. Data Transformation not required

**Model Comparison**

# Grid Search

Find the best parameters for each model

## Decision Trees

- Class weight: balanced
- criterion: ‘gini’
- Max depth: 10
- Max leaf nodes: 50

**83%**
GS
Accuracy

## K-NN

- N neighbors: 5
- weights: ‘distance‘

**86%**
GS
Accuracy

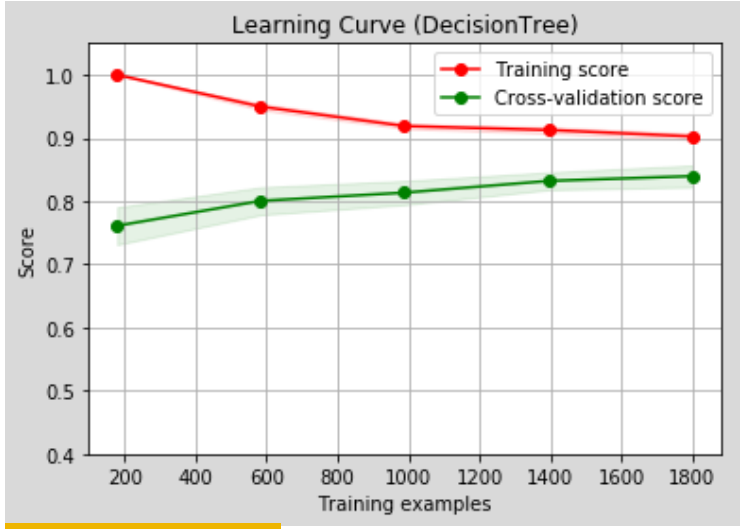
## Multinomial Logistic

- C=100000
- Solver: “newton-cg”

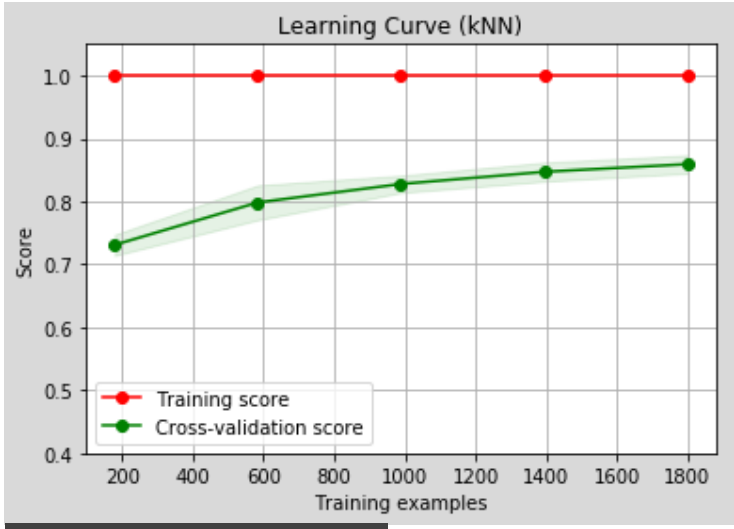
**77%**
GS
Accuracy

# Training Evaluation

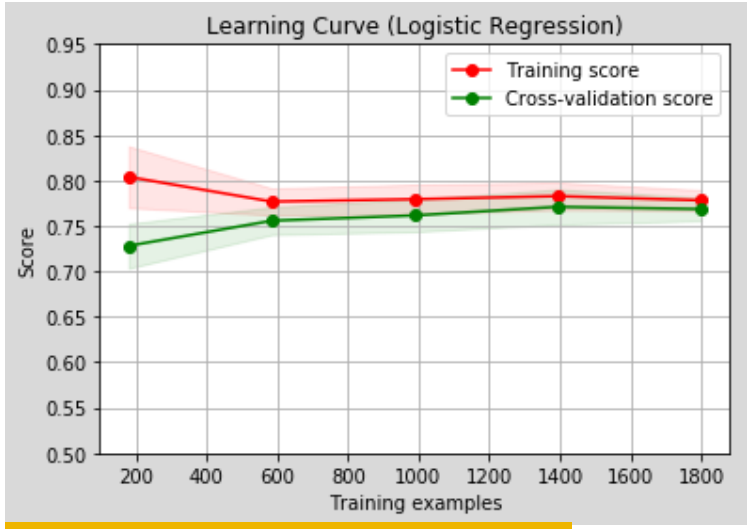
Accuracy Comparison and Overfitting Prevention




Decision Trees




K-nearest Neighbors




Multinomial Logistic Regression

| | Decision Trees | K-nearest Neighbors | Multinomial Logistic Regression |
|---|---|---|---|
| In-sample Accuracy | 89% | 100% | 77% |
| CV Score | 84% | 87% | 77% |

# Model Evaluation

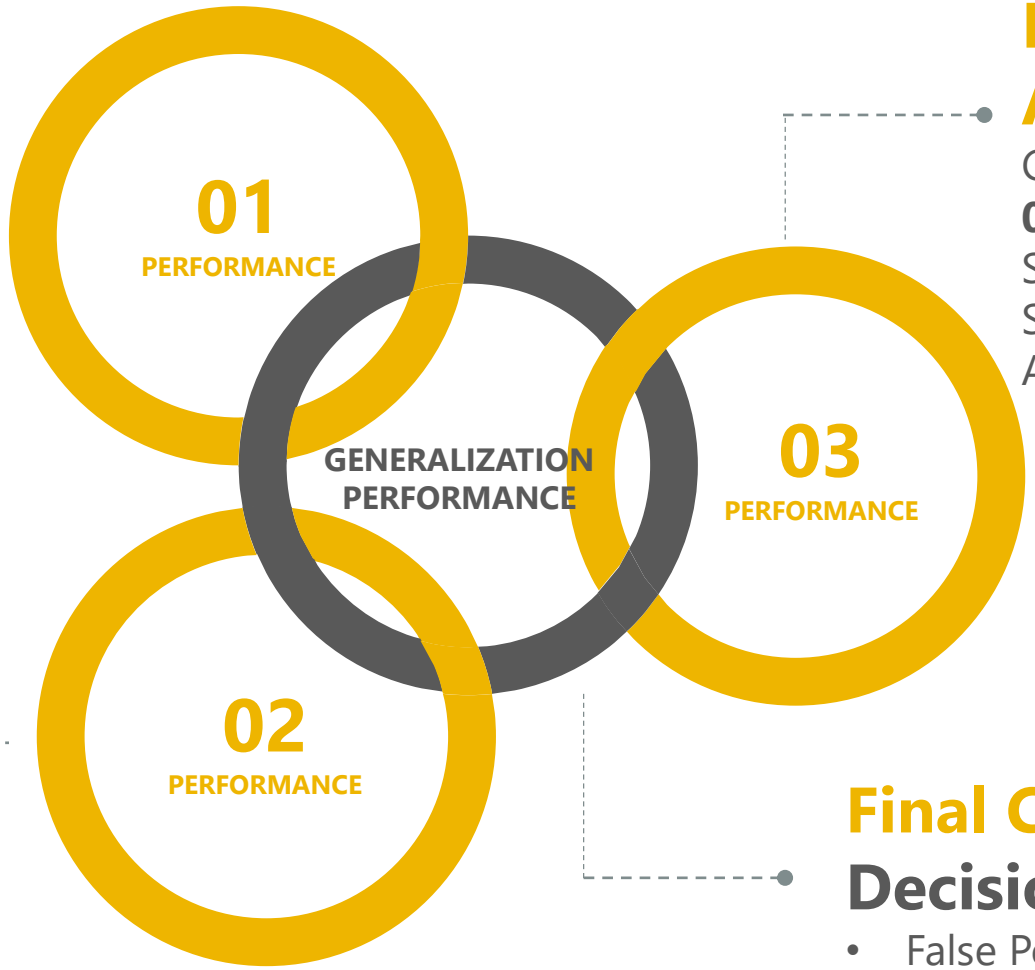
Machine Learning Perspective

## Decision Tree Classifier

CV Accuracy (Entire Data):
**0.839 +/- 0.027**
SV Accuracy (out-of-sample): **0.85**
SV Accuracy (in-sample): **0.89**
AUC: **0.84 (+/- 0.01)**

## Multinomial Logistic Regression

CV Accuracy (Entire Data):
**0.767 +/- 0.031**
SV Accuracy (out-of-sample): **0.77**
SV Accuracy (in-sample): **0.79**
AUC: **0.77 (+/- 0.03)**




## KNN Algorithm

CV Accuracy (Entire Data):
**0.820 +/- 0.025**
SV Accuracy (out-of-sample): **0.88**
SV Accuracy (in-sample): **1.00**
AUC: **0.84 (+/- 0.02)**

## Final Choice:

## Decision Tree Classifier

- False Positive rate for Excellent: 0
- F1 Score: 0.85

# Model Evaluation

Business Perspective

## Wine Fermentation

- Better quality control by adjusting physicochemical ingredients
- Measure: Production **Defect** Rate

## Testing, Aging & Bottling

- Ease the stress of designing aging (storage) and bottling plans
- Measure: **Average** Wine Quality Score

## Marketing

- Develop marketing strategies targeting wines with very unique physicochemical attributes and quality scores
- Measure: Cost and **Benefit** of the marketing campaign

## Pricing

- More specific/segmented pricing strategy
- Measure: Revenue/Profit **Changing** Rate

## Distribution Channels

- Different classes of wines will be distributed through different channels
- Measure: Channels' Satisfaction Rate/**Retention** Rate

# Deployment & Conclusion

## Cost & Benefit Analysis

| NORMALIZATION | true Below Average | true Above Average | true Excellent | true Bad |
|---|---|---|---|---|
| pred. Below Average | 0.17 | 0.07 | 0.00 | 0.00 |
| pred. Above Average | 0.06 | 0.17 | 0.00 | 0.00 |
| pred. Excellent | 0.00 | 0.02 | 0.26 | 0.00 |
| pred. Bad | 0.00 | 0.00 | 0.00 | 0.25 |

| | true Below Average | true Above Average | true Excellent | true Bad |
|---|---|---|---|---|
| pred. Below Average | **$0.66** | -$0.57 | $0.00 | $0.00 |
| pred. Above Average | -$0.47 | **$4.62** | $0.00 | $0.00 |
| pred. Excellent | $0.00 | -$0.14 | **$28.38** | $0.00 |
| pred. Bad | -$0.04 | -$0.02 | $0.00 | **-$0.77** |

$$Estimated\ Profit = \$31.66$$

*Assumptions: Please check appendix page X

01

02

03

04

| | true Below Average | true Above Average | true Excellent | true Bad |
|---|---|---|---|---|
| pred. Below Average | 532.00 | 227.00 | 0.00 | 0.00 |
| pred. Above Average | 186.00 | 547.00 | 0.00 | 0.00 |
| pred. Excellent | 1.00 | 57.00 | 818.00 | 0.00 |
| pred. Bad | 15.00 | 6.00 | 0.00 | 810.00 |

| COST/BENEFIT INFORMATION | true Below Average | true Above Average | true Excellent | true Bad |
|---|---|---|---|---|
| pred. Below Average | **$3.99** | -$8.00 | -$8.00 | -$8.00 |
| pred. Above Average | -$8.00 | **$26.99** | -$8.00 | -$8.00 |
| pred. Excellent | -$8.00 | -$8.00 | **$111.00** | -$8.00 |
| pred. Bad | -$8.00 | -$8.00 | $8.00 | **-$3.03** |

# Thank You.

Q & A

# Appendix

Assumption, data and links

Assumptions:

- Cost of manufacturing a bottle of wine is $8.00. ③
- All bottles of wine have a volume capacity of 750 ml (standard).
- Barefoot Cabernet Sauvignon – Bad quality wine - 4.97 ②
- Menage a trois silkred wine - Below average quality wine - 11.99 ②
- Scattered Peaks Cabernet Sauvignon 2015 – Above average quality wine - $34.99 ②
- Robert Mondavi To Kalon Vineyard Reserve Cabernet Sauvignon 2014 – Excellent quality wine - $119 ②


① https://www.svb.com/globalassets/library/uploadedfiles/content/trends_and_insights/reports/wine_report/svb-2018-wine-report.pdf

② https://www.wine.com/

③ http://eckraus.com/recipes-guides/wine-making-faqs/

# Appendix

## Expected profit calculation

| | true Below Average | true Above Average | true Excellent | true Bad |
|---|---|---|---|---|
| pred. Below Average | 532.00 | 227.00 | 0.00 | 0.00 |
| pred. Above Average | 186.00 | 547.00 | 0.00 | 0.00 |
| pred. Excellent | 1.00 | 57.00 | 818.00 | 0.00 |
| pred. Bad | 15.00 | 6.00 | 0.00 | 810.00 |

| | |
|---|---|
| Cost of manufacturing | $8.00 |

| | |
|---|---|
| Barefoot Cabernet Sauvignon | $4.97 |
| Menage a trois silkred wine | $11.99 |
| Scattered Peaks Cabernet Sauvignon 2015 | $34.99 |
| Robert Mondavi To Kalon Vineyard Reserve Cabernet Sauvignon 2014 | $119.00 |

| NORMALIZATION | true Below Average | true Above Average | true Excellent | true Bad |
|---|---|---|---|---|
| pred. Below Average | 0.17 | 0.07 | 0.00 | 0.00 |
| pred. Above Average | 0.06 | 0.17 | 0.00 | 0.00 |
| pred. Excellent | 0.00 | 0.02 | 0.26 | 0.00 |
| pred. Bad | 0.00 | 0.00 | 0.00 | 0.25 |

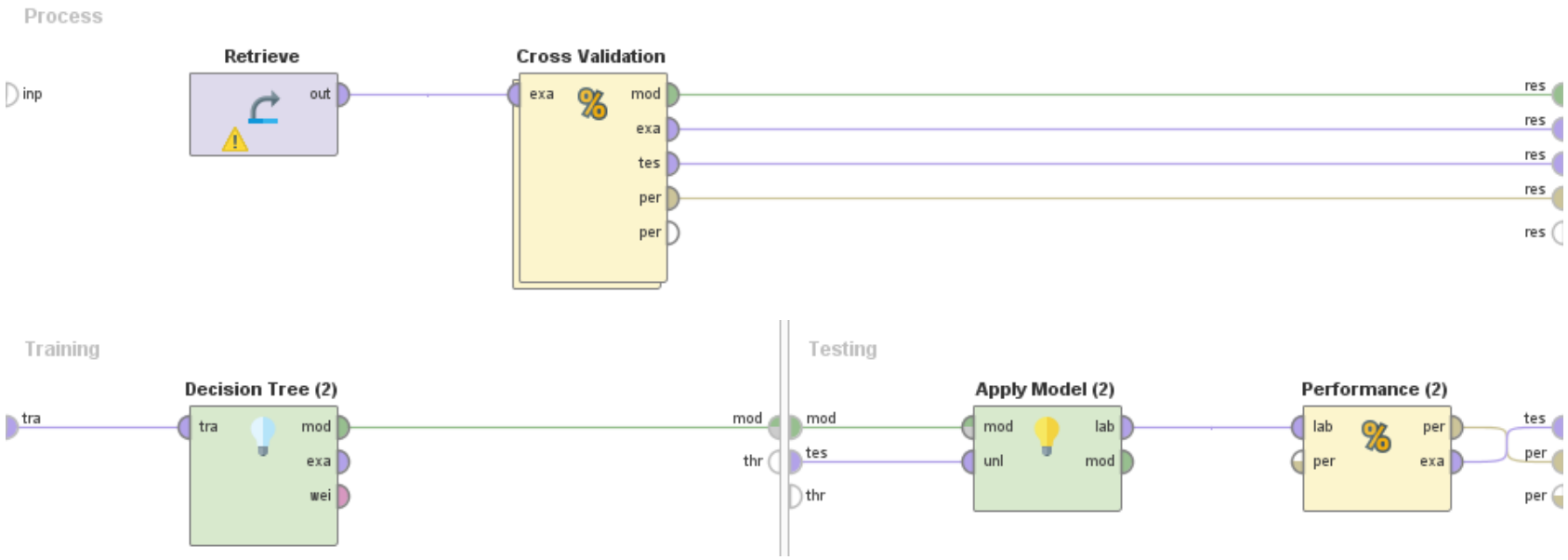
| COST/BENEFIT INFORMATION | true Below Average | true Above Average | true Excellent | true Bad |
|---|---|---|---|---|
| pred. Below Average | **$3.99** | -$8.00 | -$8.00 | -$8.00 |
| pred. Above Average | -$8.00 | **$26.99** | -$8.00 | -$8.00 |
| pred. Excellent | -$8.00 | -$8.00 | **$111.00** | -$8.00 |
| pred. Bad | -$8.00 | -$8.00 | $8.00 | **-$3.03** |

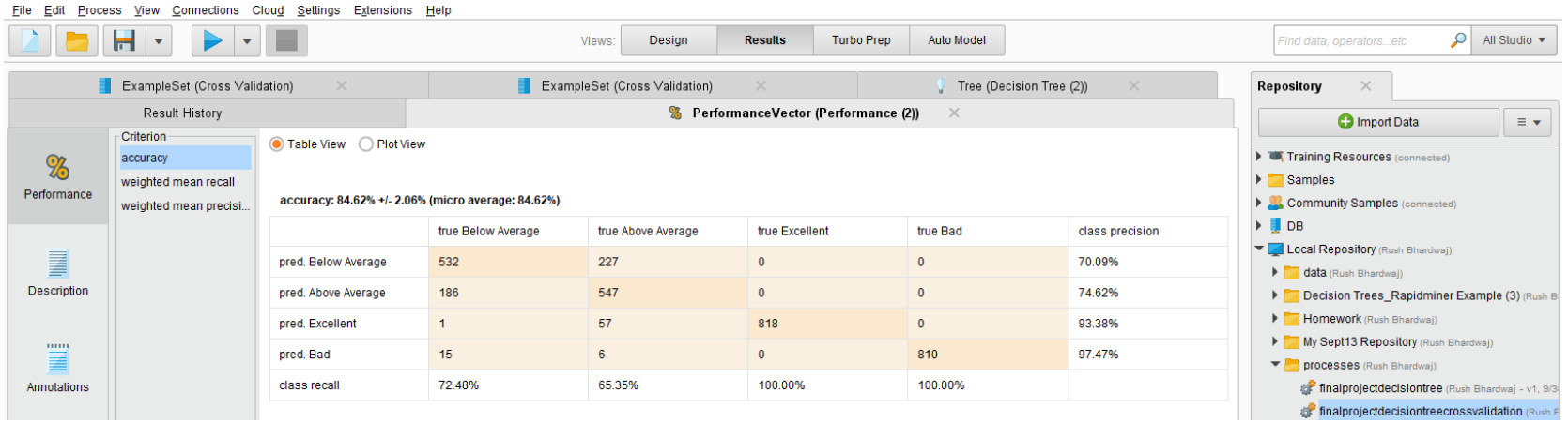
| | true Below Average | true Above Average | true Excellent | true Bad |
|---|---|---|---|---|
| pred. Below Average | **$0.66** | -$0.57 | $0.00 | $0.00 |
| pred. Above Average | -$0.47 | **$4.62** | $0.00 | $0.00 |
| pred. Excellent | $0.00 | -$0.14 | **$28.38** | $0.00 |
| pred. Bad | -$0.04 | -$0.02 | $0.00 | **-$0.77** |

| | |
|---|---|
| EV | $31.66 |

# Appendix

## Rapidminer – Decision tree

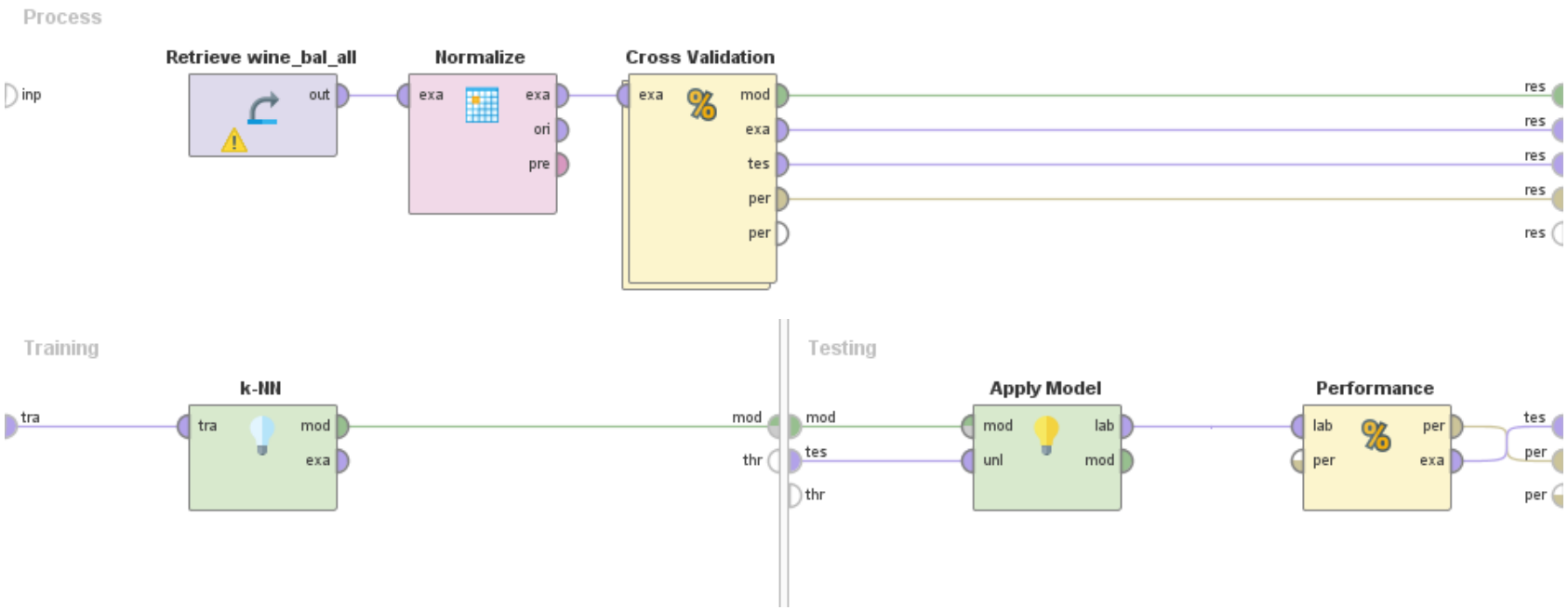


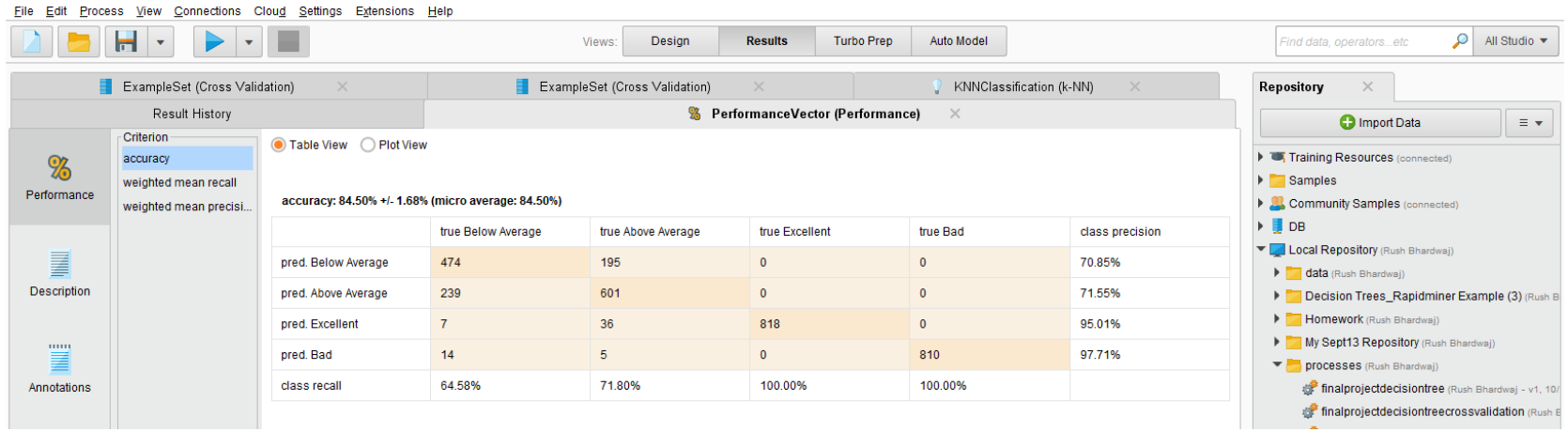




| | true Below Average | true Above Average | true Excellent | true Bad | class precision |
|---|---|---|---|---|---|
| pred. Below Average | 532 | 227 | 0 | 0 | 70.09% |
| pred. Above Average | 186 | 547 | 0 | 0 | 74.62% |
| pred. Excellent | 1 | 57 | 818 | 0 | 93.38% |
| pred. Bad | 15 | 6 | 0 | 810 | 97.47% |
| class recall | 72.48% | 65.35% | 100.00% | 100.00% | |

# Appendix

## Rapidminer – knn classification

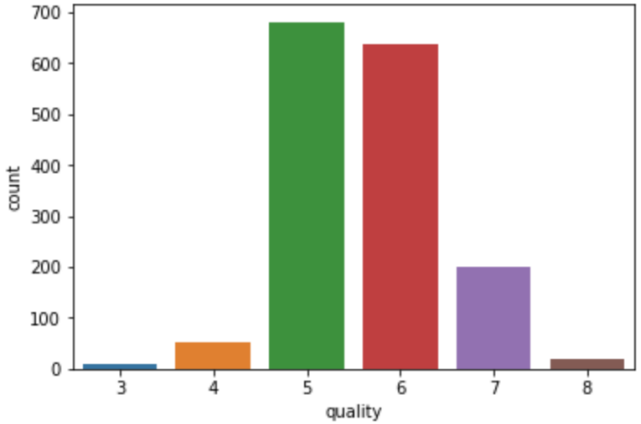


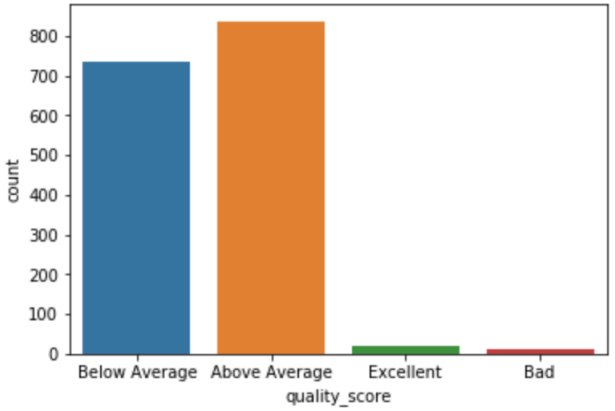



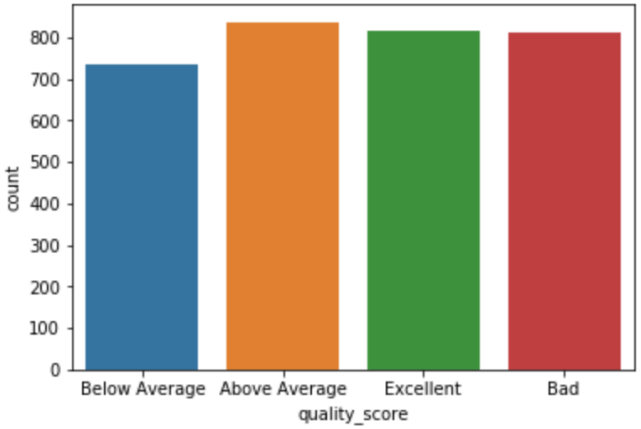

| | true Below Average | true Above Average | true Excellent | true Bad | class precision |
|---|---|---|---|---|---|
| pred. Below Average | 474 | 195 | 0 | 0 | 70.85% |
| pred. Above Average | 239 | 601 | 0 | 0 | 71.55% |
| pred. Excellent | 7 | 36 | 818 | 0 | 95.01% |
| pred. Bad | 14 | 5 | 0 | 810 | 97.71% |
| class recall | 64.58% | 71.80% | 100.00% | 100.00% | |

# Appendix

## Wine quality



# Appendix

knn classification - accuracy_score, f1_score, cohen_kappa_score, confusion_matrix,classification_report

```
Accuracy (out-of-sample): 0.89
Accuracy (in-sample): 1.00
F1 score (out-of-sample):  0.879825402863144
F1 score (in-sample)    :  1.0
Kappa score (out-of-sample):  0.8463317980432359
Kappa score (in-sample)    :  1.0
[[171   3  26   9]
 [  0 203   0   0]
 [ 48   4 129   2]
 [  0   0   0 205]]
               precision    recall  f1-score   support

          Bad       0.78      0.82      0.80       209
Below Average       0.97      1.00      0.98       203
Above Average       0.83      0.70      0.76       183
    Excellent       0.95      1.00      0.97       205

  avg / total       0.88      0.89      0.88       800
```

# Appendix

Logistic regression- accuracy_score, f1_score, cohen_kappa_score, confusion_matrix,classification_report

```
Accuracy (out-of-sample): 0.76
Accuracy (in-sample): 0.77
F1 score (out-of-sample):  0.7485365951853591
F1 score (in-sample)    :  0.7620678568229352
Kappa score (out-of-sample):  0.6810513615850577
Kappa score (in-sample)    :  0.6962421428679098
[[115   8  46  40]
 [  0 203   0   0]
 [ 46  25 105   7]
 [ 19   0   0 186]]
               precision    recall  f1-score   support

Above Average     0.6389    0.5502    0.5913       209
          Bad     0.8602    1.0000    0.9248       203
Below Average     0.6954    0.5738    0.6287       183
    Excellent     0.7983    0.9073    0.8493       205

  avg / total     0.7488    0.7612    0.7506       800
```

# Appendix

Decision tree - accuracy_score, f1_score, cohen_kappa_score, confusion_matrix,classification_report

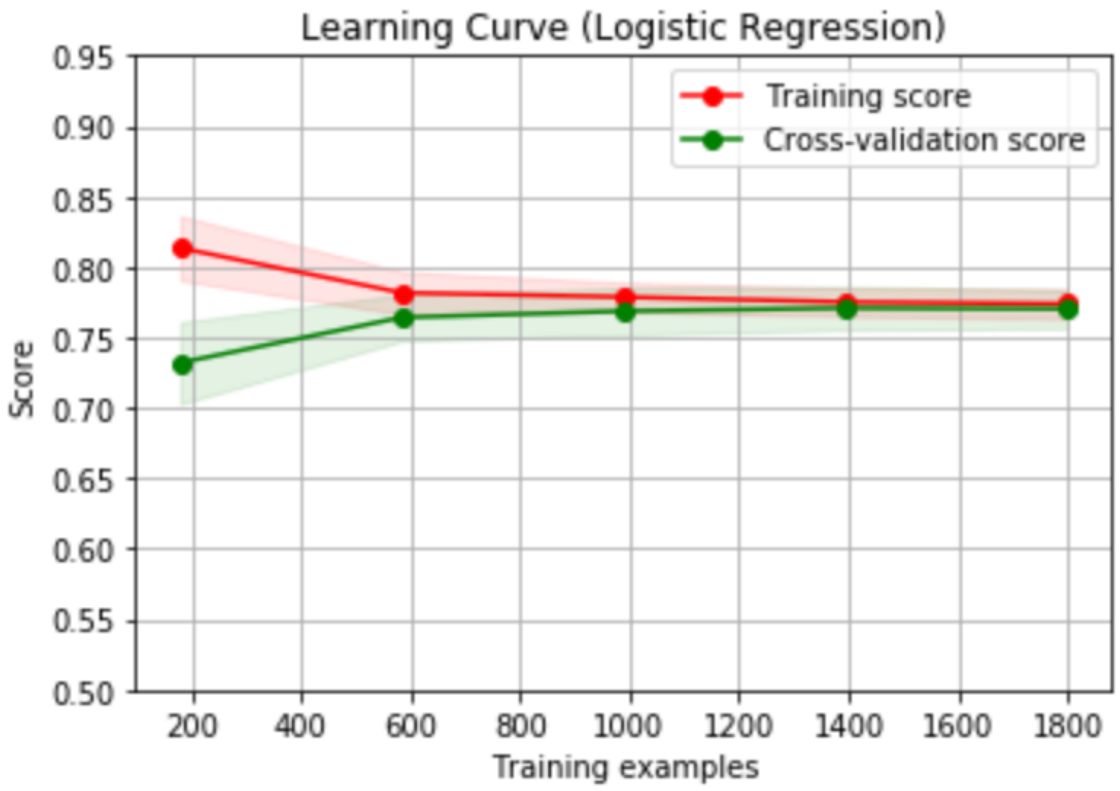
```
Accuracy (out-of-sample): 0.84
Accuracy (in-sample): 0.88
F1 score (out-of-sample):  0.8306195642132188
F1 score (in-sample)    :  0.8811848744045158
Kappa score (out-of-sample):  0.7852800306267244
Kappa score (in-sample)    :  0.846778401837019
[[117   5  72  15]
 [  0 203   0   0]
 [ 32   5 146   0]
 [  0   0   0 205]]
              precision    recall  f1-score   support

Above Average     0.7852    0.5598    0.6536       209
          Bad     0.9531    1.0000    0.9760       203
Below Average     0.6697    0.7978    0.7282       183
    Excellent     0.9318    1.0000    0.9647       205

  avg / total     0.8390    0.8387    0.8322       800
```

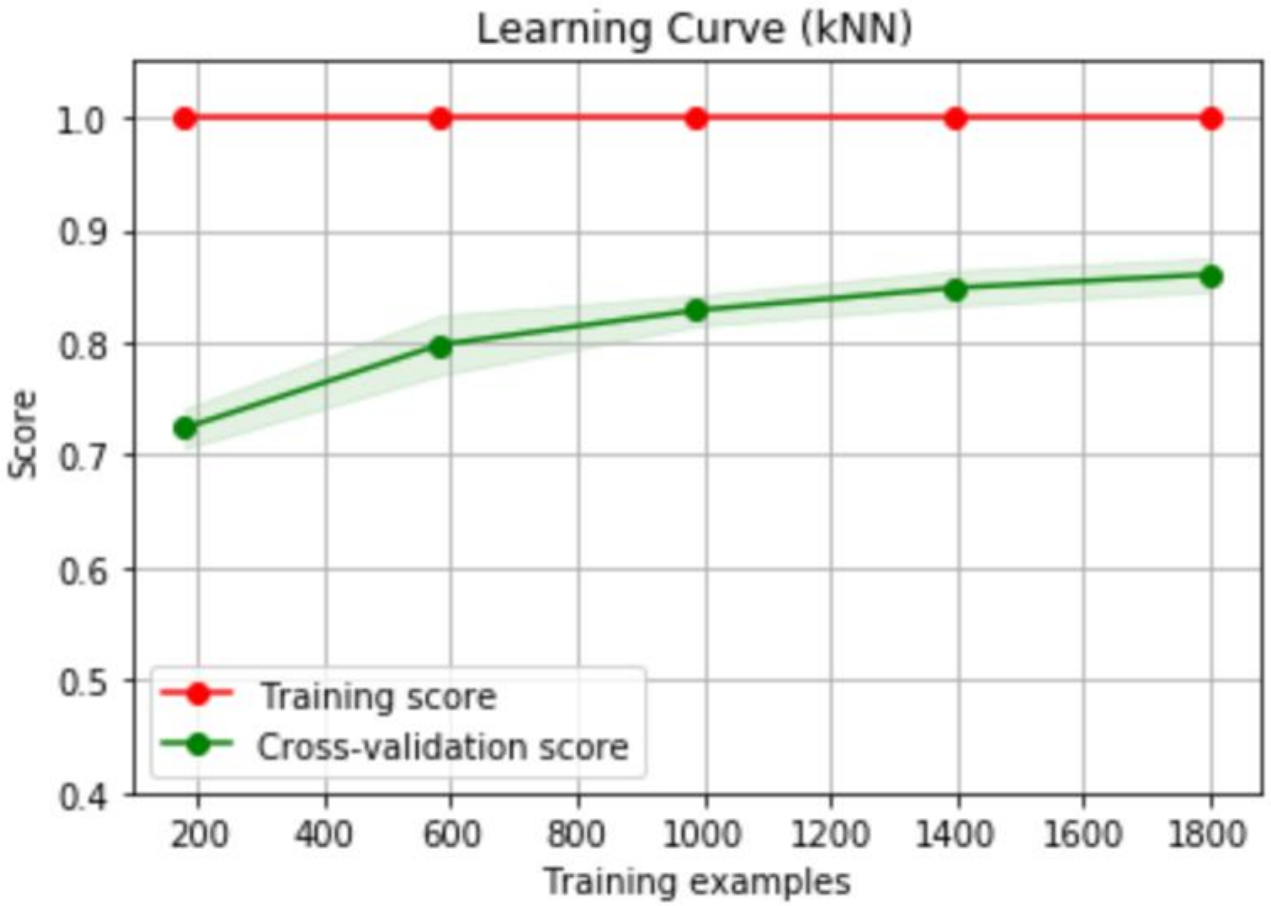
# Appendix

Learning curve – logistic regression



# Appendix

Learning curve – knn classification



# Appendix

Learning curve – logistic regression