CHATEAU DE
LEELANAU
LEELANAU

Agenda

-- Data Analysis & Feature Engineering

-- Modeling & Metric Evaluation

Data Overview

- **The data is Vinho Verde* red wine samples from Portugal.**

- **The goal of this project is to determine wine quality based on the chemical properties**

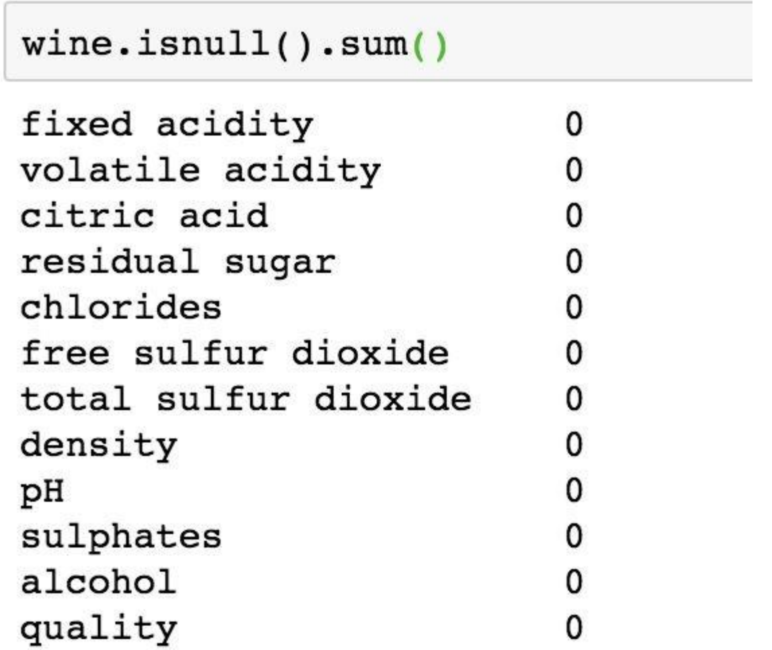
  - Input variable: based on physicochemical tests

  - Output variable: based on sensory data, median of at least 3 evaluations made by wine experts

*Portuguese wine that originated in the historic Minho province in the far north of the country

Features

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- PH
- Sulphates
- Alcohol

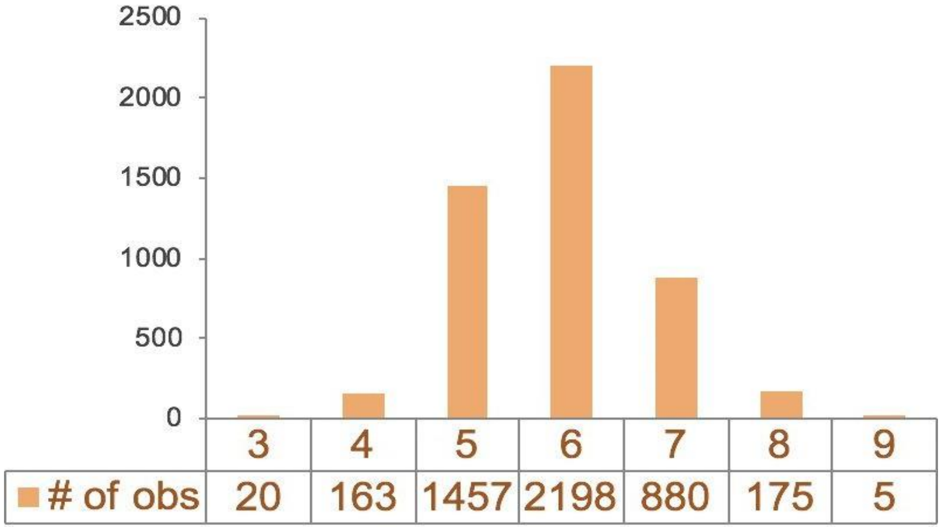


```
wine.isnull().sum()
```

```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
```

Label Distribution

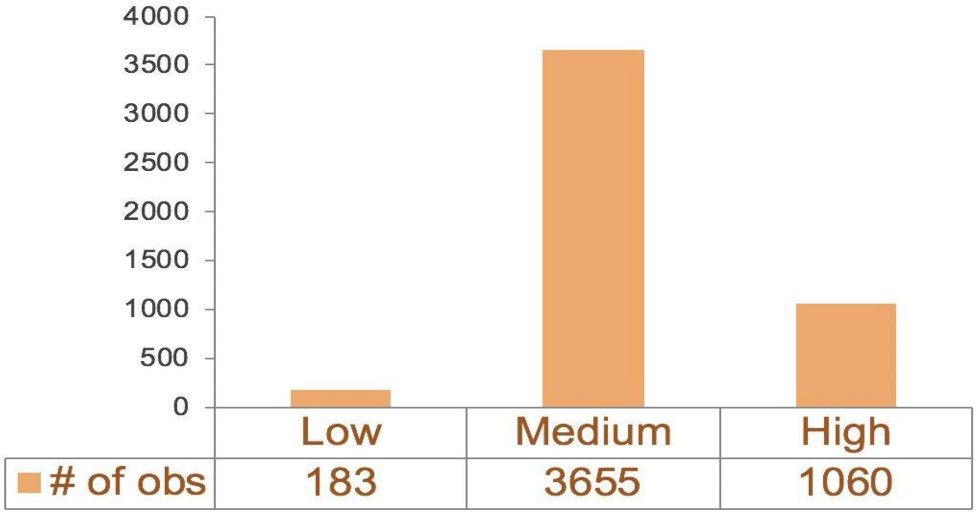
- Quality is represented by scores ranging from 0 to 10
- 0 is the worst and 10 is the best



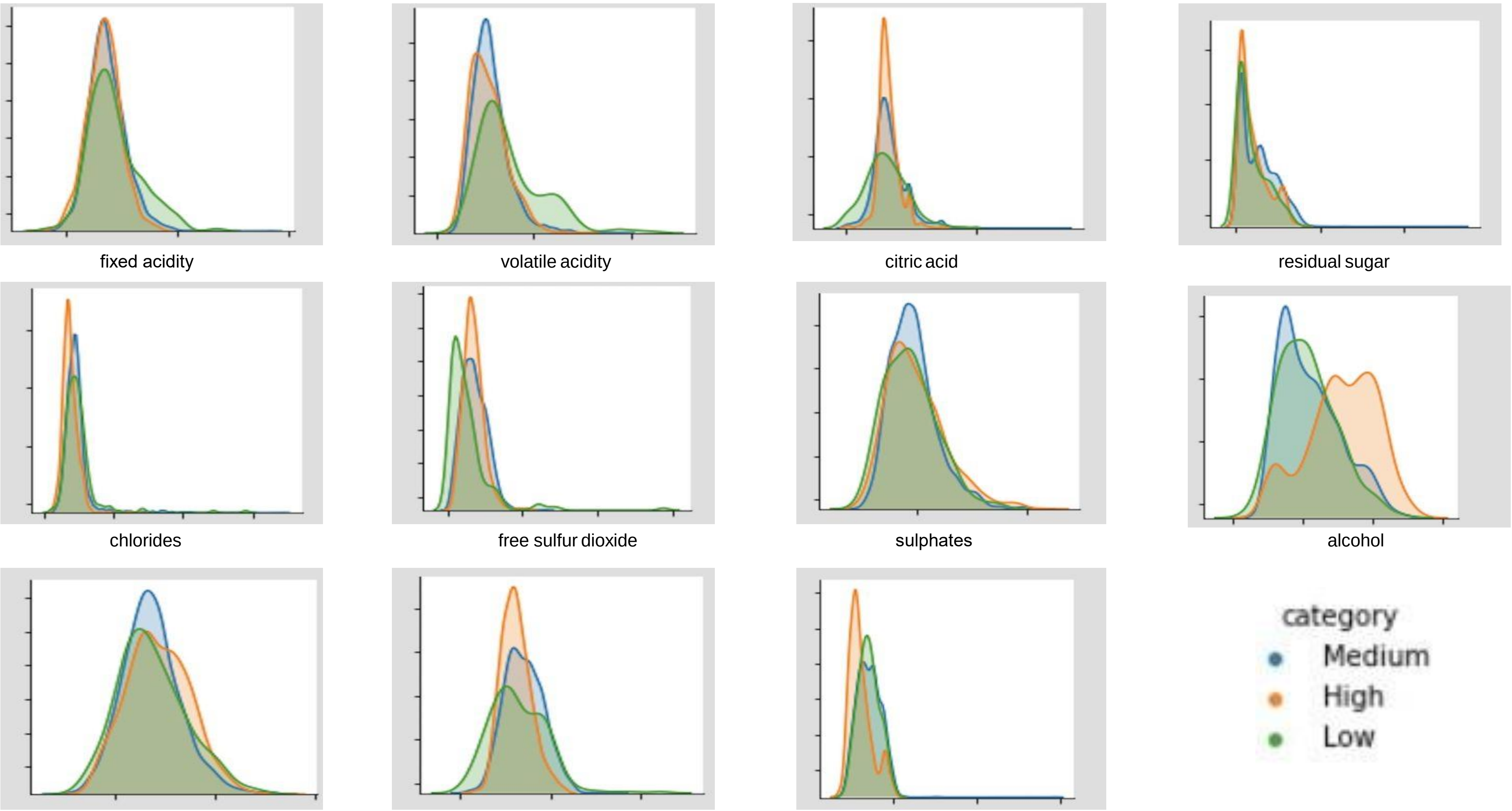
	3	4	5	6	7	8	9
■ # of obs	20	163	1457	2198	880	175	5

Labels and Encoding

- Binning:
  - score under 5 → “Low”
  - score above 6 → “High”
  - score of 5 and 6 → “Medium”



	Low	Medium	High
■ # of obs	183	3655	1060



fixed acidity

volatile acidity

citric acid

residual sugar

chlorides

free sulfur dioxide

sulphates

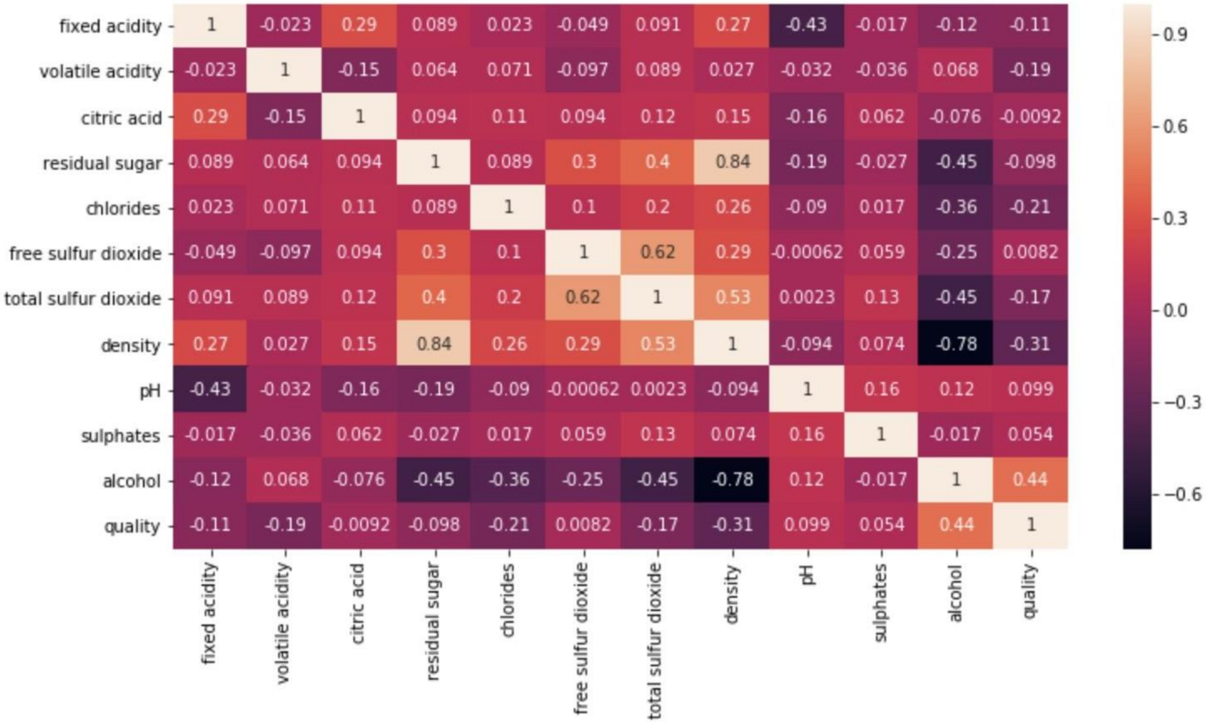
alcohol

PH

total sulfur dioxide

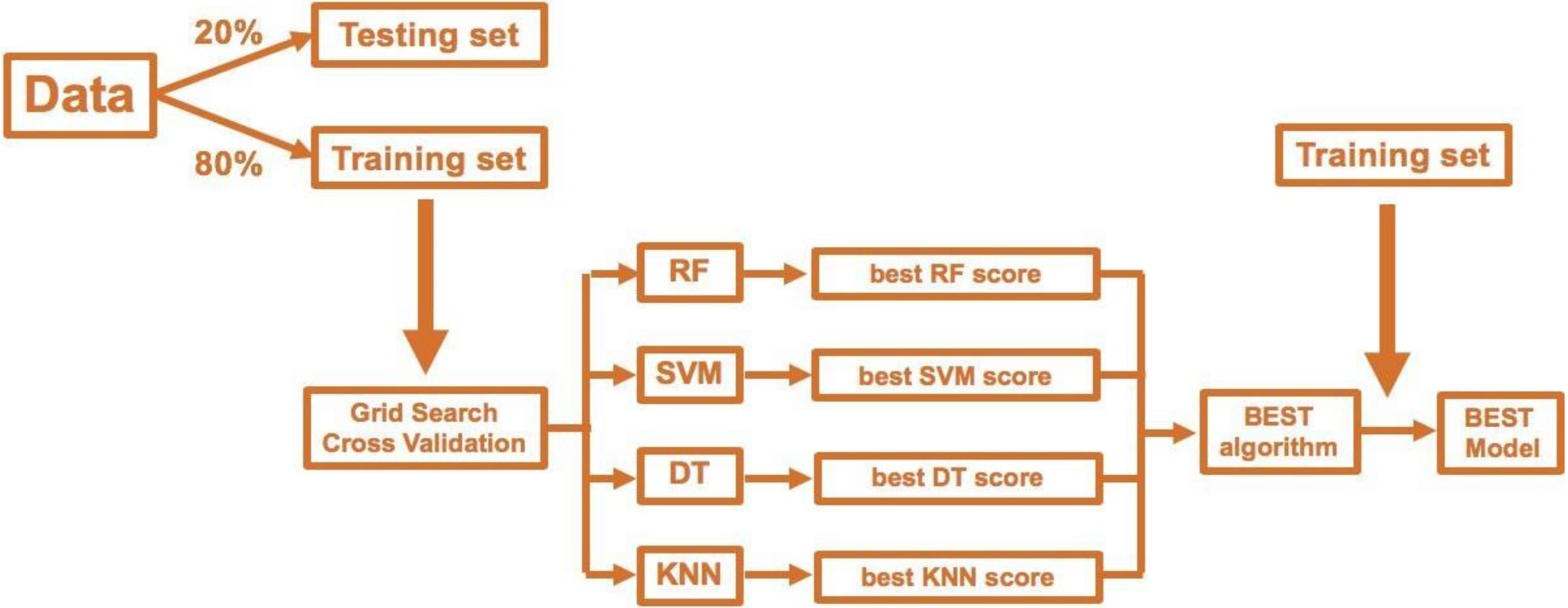
density

Variable Correlation Matrix



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.023	0.29	0.089	0.023	-0.049	0.091	0.27	-0.43	-0.017	-0.12	-0.11
volatile acidity	-0.023	1	-0.15	0.064	0.071	-0.097	0.089	0.027	-0.032	-0.036	0.068	-0.19
citric acid	0.29	-0.15	1	0.094	0.11	0.094	0.12	0.15	-0.16	0.062	-0.076	-0.0092
residual sugar	0.089	0.064	0.094	1	0.089	0.3	0.4	0.84	-0.19	-0.027	-0.45	-0.098
chlorides	0.023	0.071	0.11	0.089	1	0.1	0.2	0.26	-0.09	0.017	-0.36	-0.21
free sulfur dioxide	-0.049	-0.097	0.094	0.3	0.1	1	0.62	0.29	-0.00062	0.059	-0.25	0.0082
total sulfur dioxide	0.091	0.089	0.12	0.4	0.2	0.62	1	0.53	0.0023	0.13	-0.45	-0.17
density	0.27	0.027	0.15	0.84	0.26	0.29	0.53	1	-0.094	0.074	-0.78	-0.31
pH	-0.43	-0.032	-0.16	-0.19	-0.09	-0.00062	0.0023	-0.094	1	0.16	0.12	0.099
sulphates	-0.017	-0.036	0.062	-0.027	0.017	0.059	0.13	0.074	0.16	1	-0.017	0.054
alcohol	-0.12	0.068	-0.076	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.017	1	0.44
quality	-0.11	-0.19	-0.0092	-0.098	-0.21	0.0082	-0.17	-0.31	0.099	0.054	0.44	1

Modeling Process



Data

20%

Testing set

80%

Training set

Grid Search Cross Validation

RF

best RF score

SVM

best SVM score

DT

best DT score

KNN

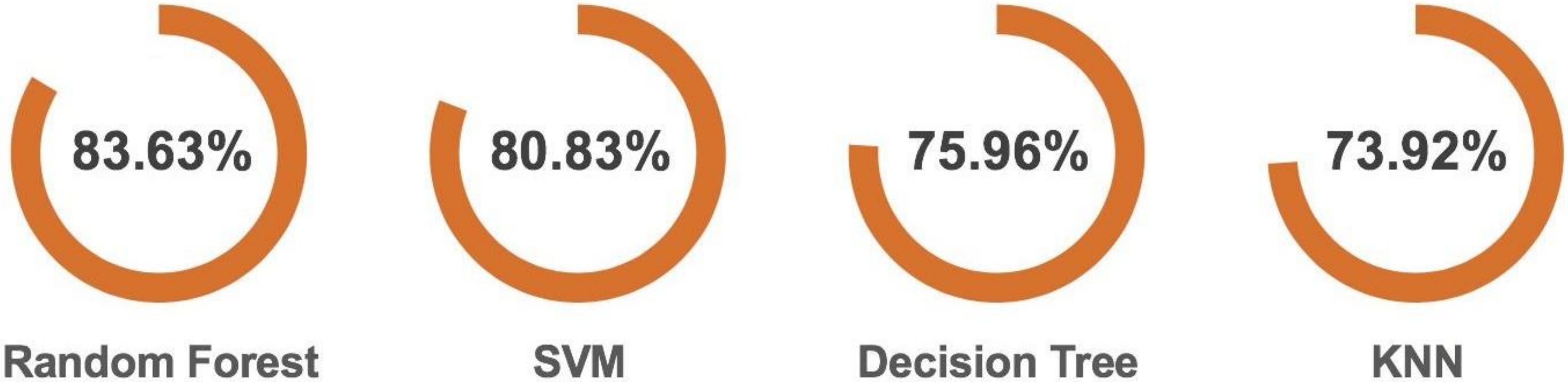
best KNN score

Training set

BEST algorithm

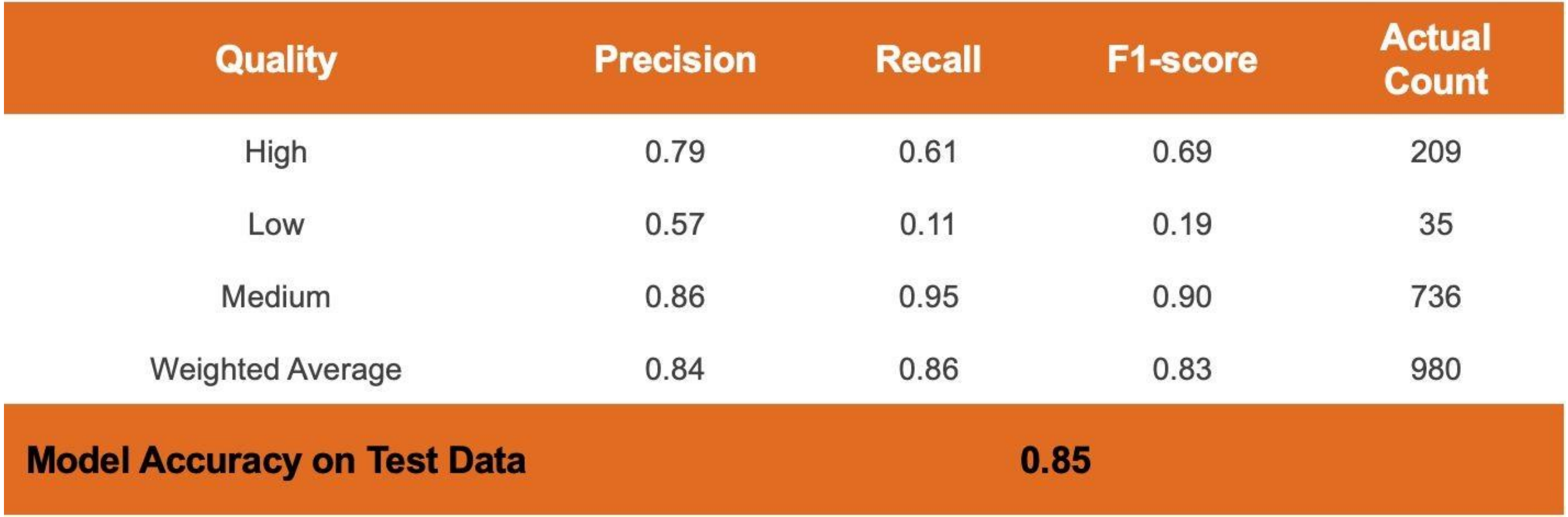
BEST Model

Validation Accuracy



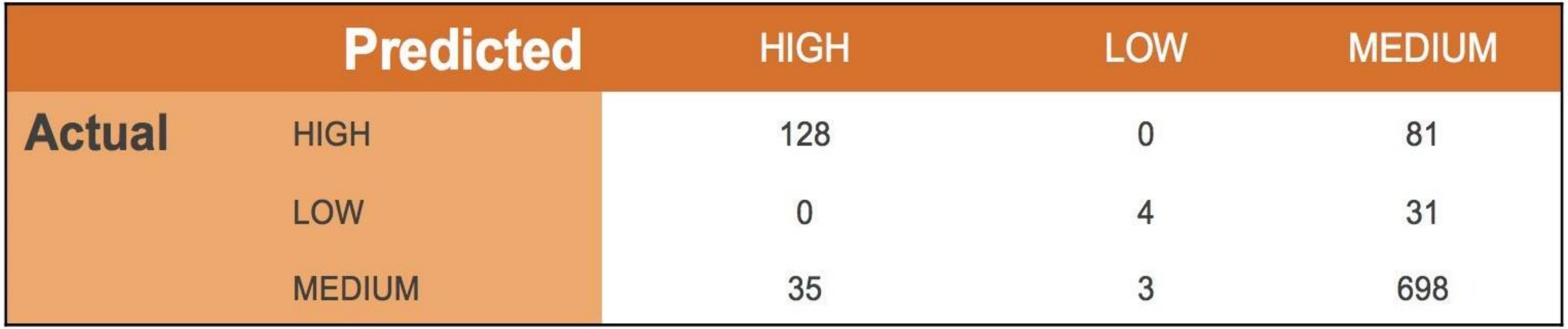
Random Forest performs the best on validation

RF on Testing Accuracy



Quality	Precision	Recall	F1-score	Actual Count
High	0.79	0.61	0.69	209
Low	0.57	0.11	0.19	35
Medium	0.86	0.95	0.90	736
Weighted Average	0.84	0.86	0.83	980
Model Accuracy on Test Data		**0.85**		

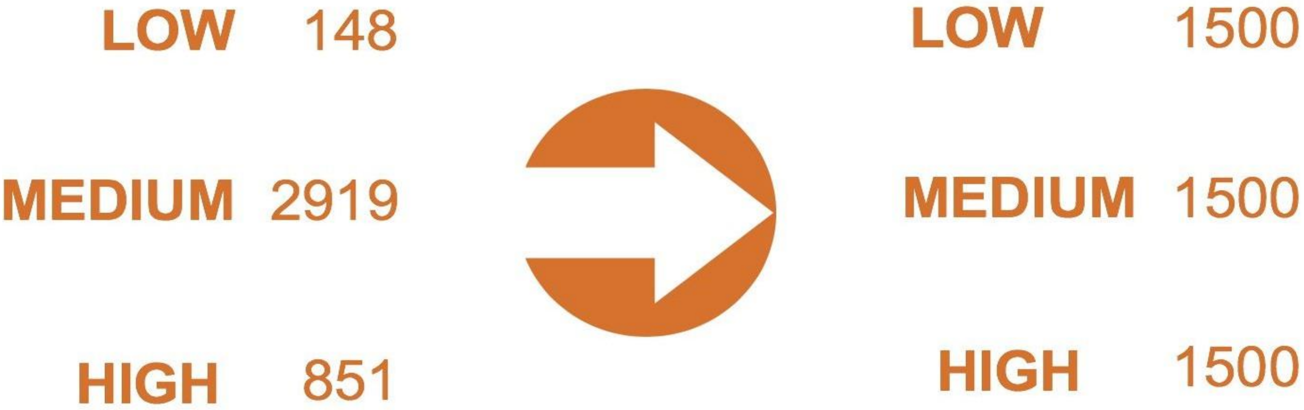
RF Confusion Matrix



	Predicted	HIGH	LOW	MEDIUM
Actual	HIGH	128	0	81
	LOW	0	4	31
	MEDIUM	35	3	698

Resample

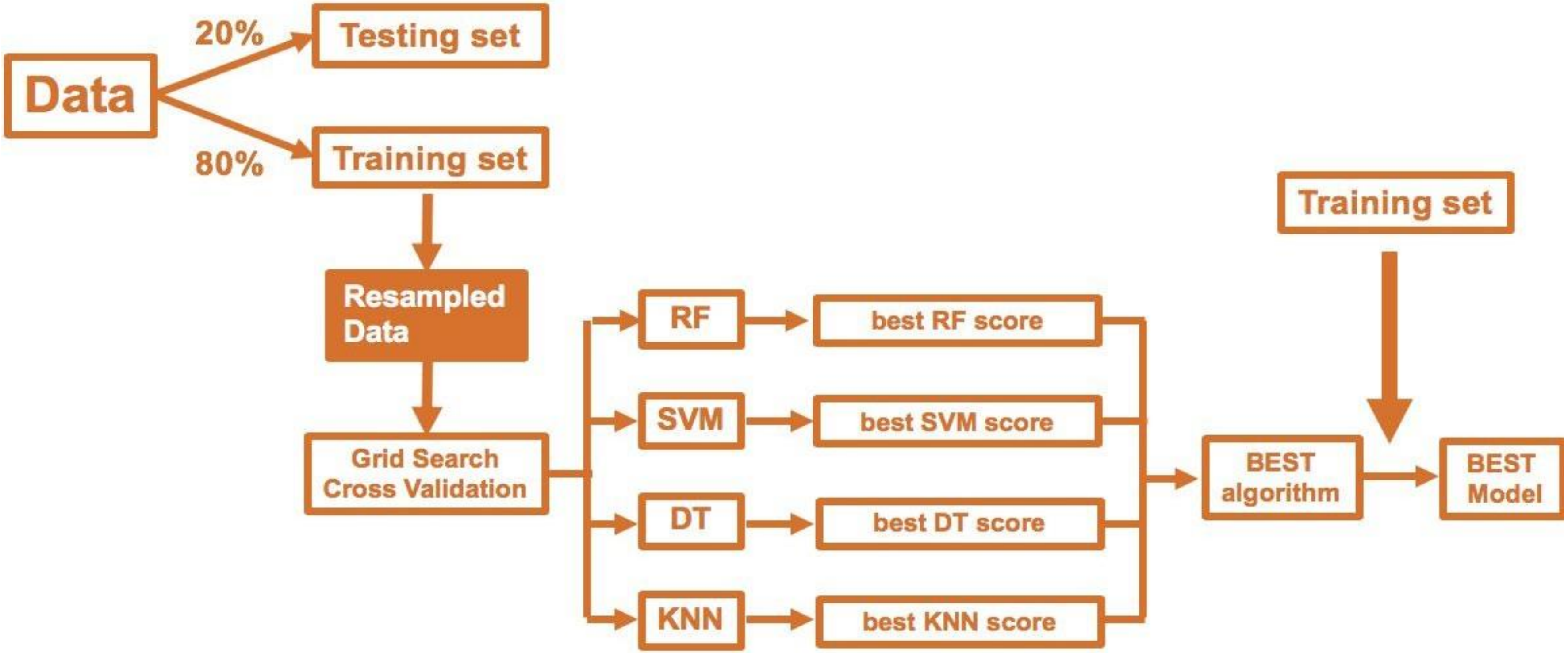
The data is significantly imbalanced, so we decided to resample the training data to improve accuracy for low and high



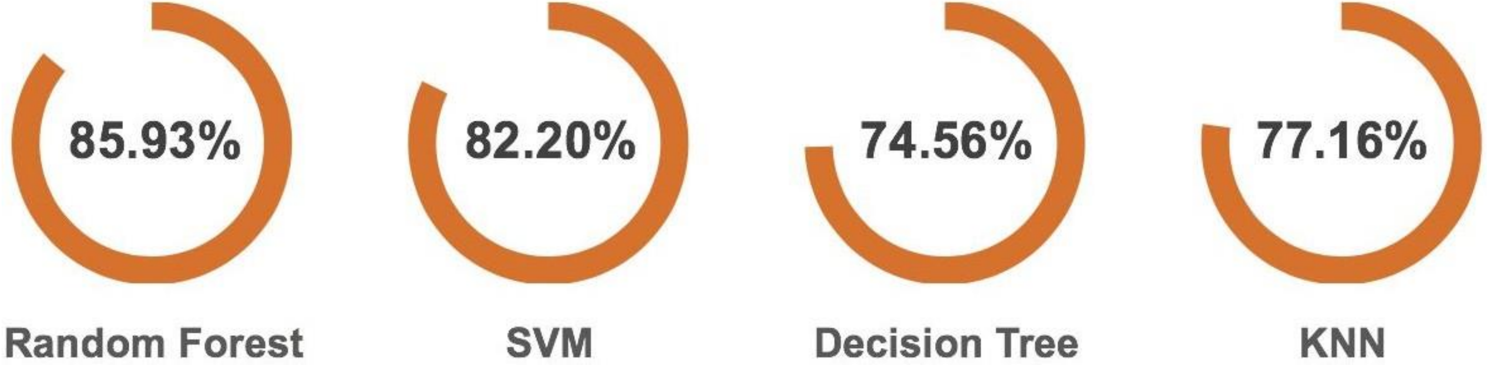
LOW	148
MEDIUM	2919
HIGH	851

LOW	1500
MEDIUM	1500
HIGH	1500

Similar Process



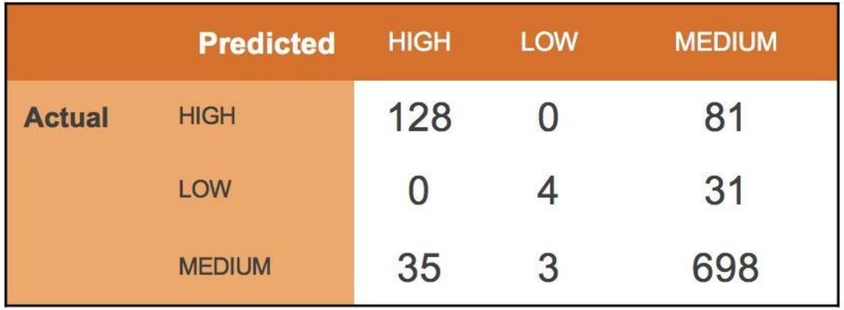
Validation Accuracy



Random Forest is still the best.

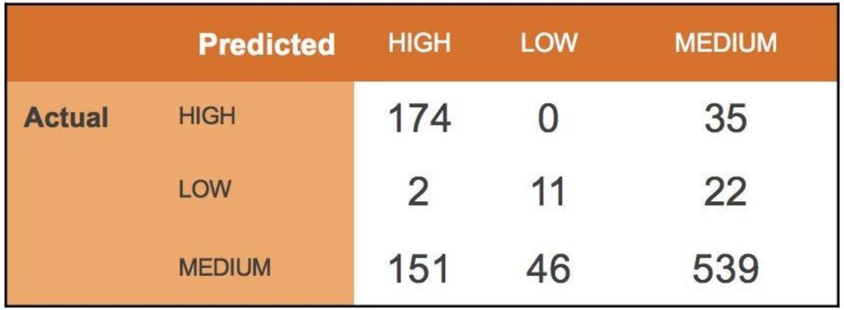
RF Confusion Matrix

RF acc on test data BEFORE resampling: 84.69%



	Predicted	HIGH	LOW	MEDIUM
Actual	HIGH	128	0	81
	LOW	0	4	31
	MEDIUM	35	3	698

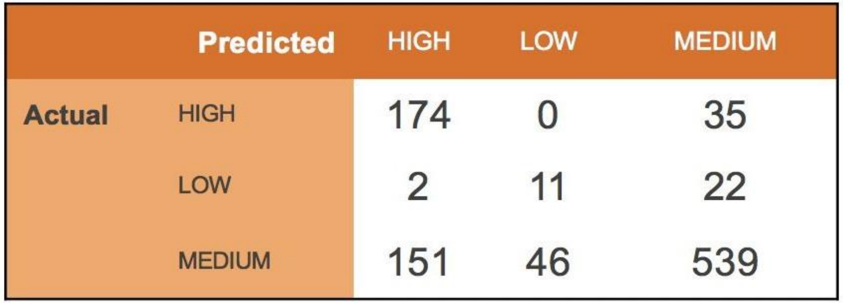
RF acc on test data AFTER resampling: 73.87%



	Predicted	HIGH	LOW	MEDIUM
Actual	HIGH	174	0	35
	LOW	2	11	22
	MEDIUM	151	46	539

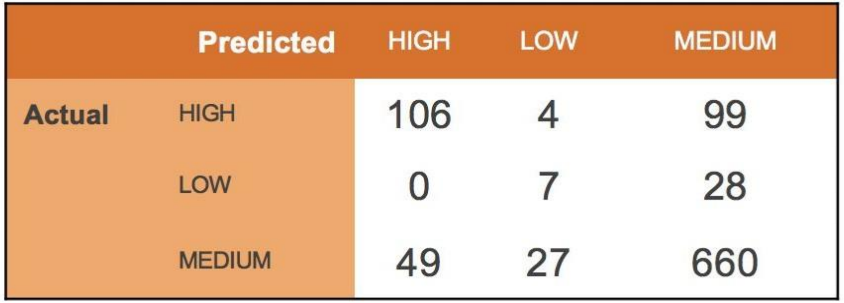
RF vs. SVM

RF acc on test data AFTER resampling: 73.87%



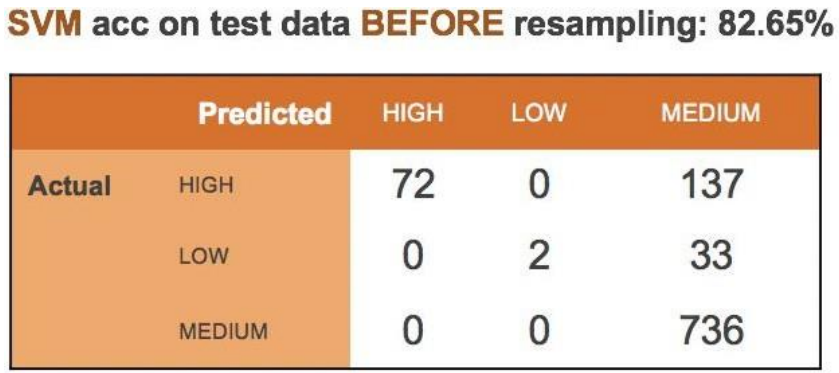
	Predicted	HIGH	LOW	MEDIUM
Actual	HIGH	174	0	35
	LOW	2	11	22
	MEDIUM	151	46	539

SVM acc on test data AFTER resampling: 78.88%



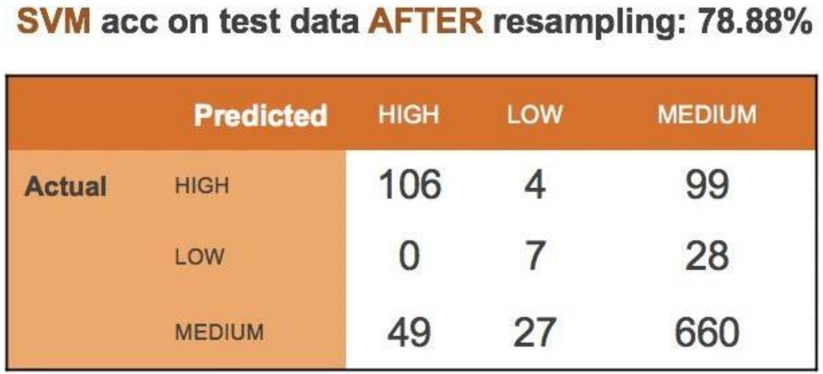
	Predicted	HIGH	LOW	MEDIUM
Actual	HIGH	106	4	99
	LOW	0	7	28
	MEDIUM	49	27	660

SVM Confusion Matrix



SVM acc on test data **BEFORE** resampling: 82.65%

	Predicted	HIGH	LOW	MEDIUM
Actual	HIGH	72	0	137
	LOW	0	2	33
	MEDIUM	0	0	736



SVM acc on test data **AFTER** resampling: 78.88%

	Predicted	HIGH	LOW	MEDIUM
Actual	HIGH	106	4	99
	LOW	0	7	28
	MEDIUM	49	27	660