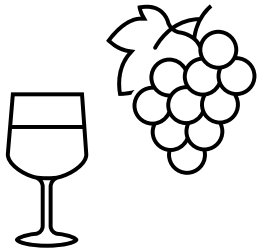
Data Mining – Team MINEHeim

Wine Prediction



UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

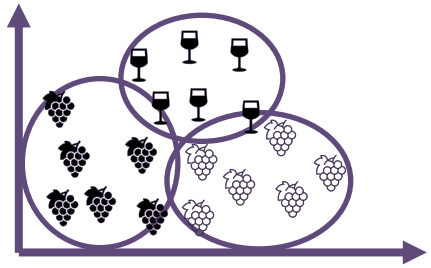
Agenda

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Research Problem

Data & Pre-Processing

Data Mining



Conclusion

Research Problem

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

PROBLEM

How can you determine the quality of wine without actually drinking it?
What makes a great wine … actually great?

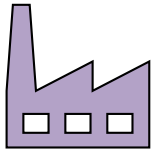
- Which factors/components influence the quality and taste?

GOAL

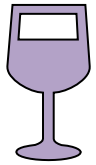
Objectively predict and classify the quality of wine based on its ingredients, components and features.

- e.g. sulphates, residual sugar, alcohol, pH-Value, type, …

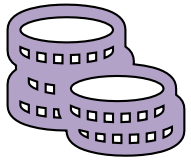
BUSINESS VALUE



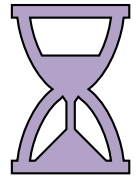
Wineries & Dealers



Old Wine



Reduce Costs



Save Time

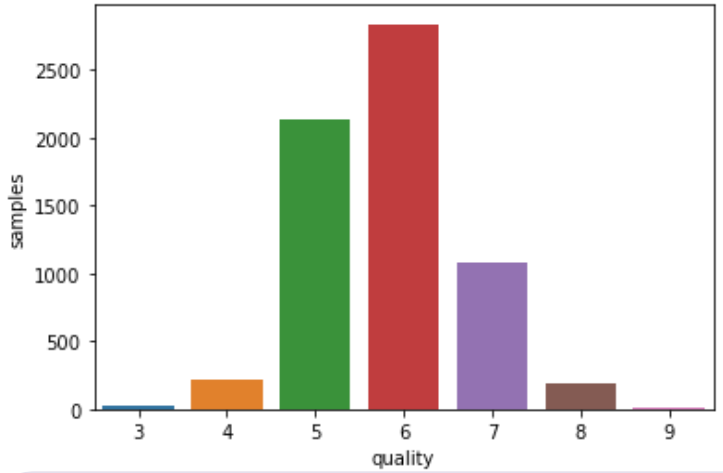
Single Source of Truth

Data Structure

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Data set related to red and white variants of Portuguese wine "Vinho Verde"

6497 samples, already gathered



13 features, mainly physiochemical properties e.g., pH-value, alcohol percentage etc.

Quality feature ranges from 0 to 10, assessed by wine tastings

Highly unbalanced data: lots of mediocre wines, few bad and excellent wines

Data & Pre-Processing

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Duplicate Deletion and Replacing Null-Values

- No duplicates
- 34 rows with missing values were deleted

Column Preprocessing

- Normalization of numerical values
- Encoding of categorical values

Outlier Deletion

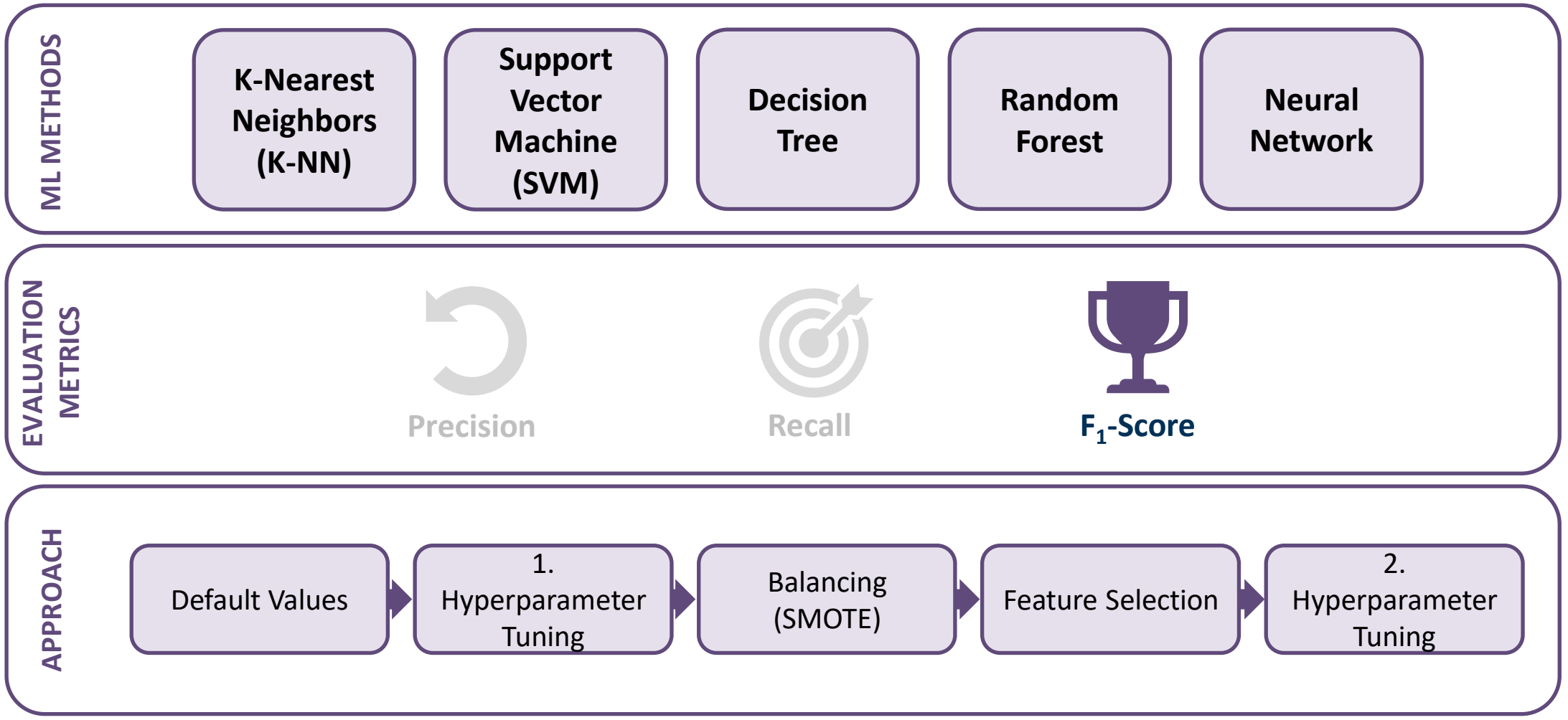
- Deleting 57 outliers on the basis of a 90 % confidence interval

Data Separation and Splitting

- Separation of the feature and target variable
- Train and test split

Data Mining

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

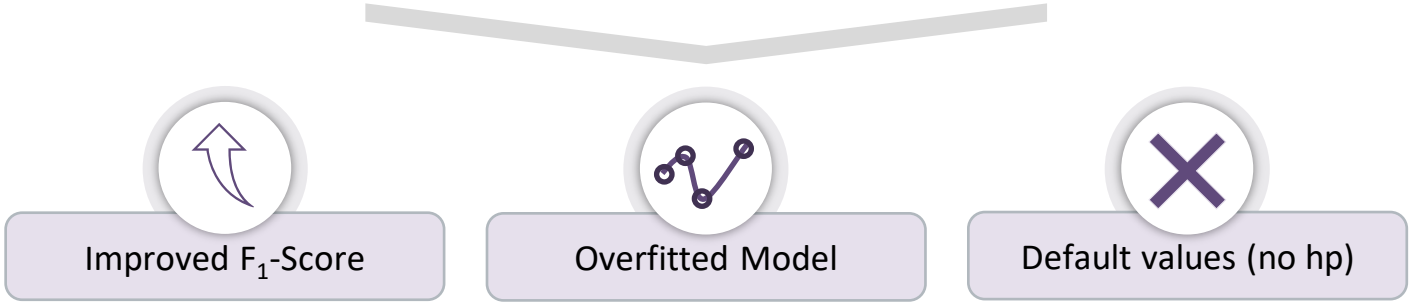


K-Nearest Neighbors (K-NN)

F_1-Score Overview

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

TRAIN	63.1 %	63.1 %	100 %	100 %	100 %
	Initial State	1. Hyperparameter Tuning	With Oversampling	With Feature Selection	2. Hyperparmeter Tuning
TEST	53.2 %	53.2 %	58.6 %	60 %	60 %
	Initial State	1. Hyperparameter Tuning	With Oversampling	With Feature Selection	2. Hyperparmeter Tuning



Support Vector Machine (SVM)

F_1-Score Overview

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

	Initial State	1. Hyperparameter Tuning	With Oversampling	With Feature Selection	2. Hyperparmeter Tuning
TRAIN	38.4 %	72.0 %	88.5 %	86.4 %*	93.5 %
TEST	41.2 %	57.5 %	56.5 %	56.1 %*	57.6 %

* 2 models were evaluated, the better one is listed here. Performance decreased, therefore not included in next step.

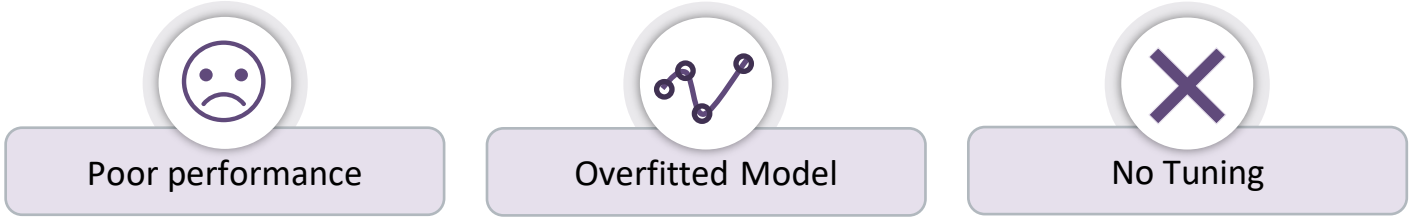
- Improved F_1-Score
- Overfit Tendency of Model
- No Feature Selection

Decision Tree

F_1-Score Overview

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

	Initial State	1. Hyperparameter Tuning	With Oversampling	With Feature Selection	2. Hyperparmeter Tuning
TRAIN	100 %	75.9%	100 %	100 %	85.8 %
TEST	58.7 %	53.8 %	55.7 %	57.5 %	51.8%

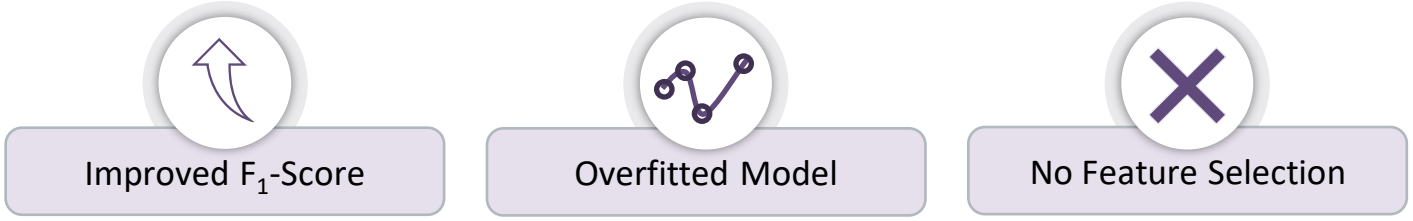


Random Forest

F_1-Score Overview

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

TRAIN	100 %	58.9 %	93.5 %	93.9 %	96.6 %
	Initial State	1. Hyperparameter Tuning	With Oversampling	With Feature Selection	2. Hyperparmeter Tuning
TEST	55 %	47.8 %	70.8 %	70.5 %	70.9 %
	Initial State	1. Hyperparameter Tuning	With Oversampling	With Feature Selection	2. Hyperparmeter Tuning



Neural Network

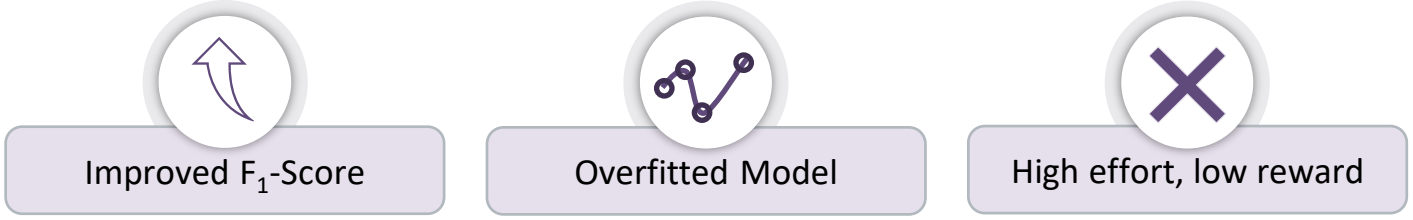
F_1-Score Overview

Legend

Batch	Layers
Epochs	Neurons

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

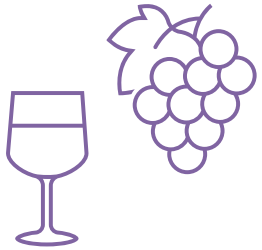
VALIDATION	38.3 %	72.8 %	76.8 %	80.9 %	84.6 %
	No Balancing				
	64 / 3	64 / 4	64 / 5	16 / 7	16 / 8
	30 / 100	30 / 300	30 / 400	40 / 1000	80 / 2048
TEST	41.4 %	51.9 %	49.9 %	55.7 %	56.6 %
	No Balancing				
	64 / 3	64 / 4	64 / 5	16 / 7	16 / 8
	30 / 100	30 / 300	30 / 400	40 / 1000	80 / 2048



Evaluation & Conclusion

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Business Use Case



Predict **wine quality** with the help of contents.

Results

KNN

as baseline did a reasonable job

Decision Tree

mediocre performance on its own

Random Forest

performed well as an aggregation of decision trees

SVM

also performed well as a classification algorithm

Neural Network

Efforts to create and were not rewarded with good predictions ➔ more resources, better experience needed

Overfitted Model

Random Forest

Other Features besides chemical characterisics might be important for the wine quality