PART 2
Programming for Data Analytics 1

Riko Wolhüter

Total Score: 25 %   ⊖ Medium risk

Submission UUID: 4bdd7f91-b6d7-5129-88c9-3c9b1ed7ce98

Total Number of Reports	Highest Match	Average Match	Submitted on	Average Word Count
1	25 % Safe assign test.docx	25 %	01/06/23 21:05 GMT+2	3,112 Highest: Safe assign test.docx

⊖ **Attachment 1**   25 %

Word Count: 3,112
Source: Safe assign test.docx

Global database (8)   **11 %**

① Student paper	⑫ Student paper	⑩ Student paper
⑭ Student paper	⑪ Student paper	③ Student paper
⑤ Student paper	⑬ Student paper	

Institutional database (5)   **10 %**

⑨ Student paper	④ My paper	⑯ Student paper
⑥ Student paper	⑮ Student paper	

Internet (3)   **4 %**

② everwoodbiocostruzioni	⑦ h2o	⑧ ehu

Scholarly journals & publications (1)   **0 %**

⑰ ProQuest document

Top sources (3)

⑨ Student paper	5 %	④ My paper	3 %	① Student paper	3 %

Contents

Background

In a scenario when a person is going on a dinner date and they have to pick up a nice wine, we don't always know what wine is the best at its price point. The dataset we will be analyzing was set up to predict the quality of wine dependent on multiple other physicochemical features. This is a classification problem therefor we will be training and testing our models using K-NN and Na ve Bayes Classifiers, we will also use evaluation metrics to determine the best classification algorithm to use. We will be determining the class label for the quality of wine. We will be performing feature engineering and cross validation to get the most accurate models. ④ The main question we want to answer is: "What is the quality of wine given the wine characteristics?" To answer the question, the python programming language will be used to pre-process the data, generate graphs, and perform analysis. The features that we will be looking at are: 1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

12 - quality (score between 0 and 10) By looking at the variables and their relationships to one another, hopefully we can train a model to predict the quality of wine. Classification

In machine learning, classifcation refers to the process of categorizing data into classes or categories based on the features provided. It invloves supervised learning where you train a model on a labeled dataset, each dataset instance is associated with a class or category. We use classification to teach a model to accurately predict the class of unseen data based on the patterns and relationships it learnt from the training data. The model then uses the input variables and assigns it to one of the predefined classes. There are multiple classification algorithms we can use such as: ① decision trees, logistic regression, support vector machines (SVM), k-nearest neighbors (KNN), random forests, Na ve bayes and neural networks. These algorithms learn from the labeled data and then builds a model around it. The performance of a classification model can be evaluated using metrics such as: ⑤ accuracy, precision, recall, F1 score, and area under the ROC curve (AUC-ROC). We use certain metrics dependent on the problem were trying to solve. ⑥ [1] Feature Engineering

Feature Engineering is the process of converting raw data into a set of meaningful features that we can use to train our models. The process involves selecting, extracting, and manipulating the characteristics from the input data so that we can improve the performance and accuracy of our model. Most often raw data contains some irrelevant information and some useful information, so feature engineering aims to highlight the patterns and relationships within the data. By creating informative features, the machine learning algorithms we are using make more accurate predictions and classifications. In feature engineering, we first choose the most relevant features from our dataset so that we can remove noise, reduce dimensionality and to focus on the most informative features. Next, we create new features by transforming the existing features, we do this by using mathematical operations to extract information that could be useful in our model. After that we modify the scale, distribution, and representation of the features to improve the learning process. We can use techniques such as Normalization to transform the features. For the next step if necessary, we encode some features, where we convert categorical or textual features into numerical representations so that the machine learning algorithms can understand the data better. Lastly, we create new features by combining different attributes. [2] [3] Evaluation Metrics

Evaluation metrics is used in machine learning to assess the performance of a machine learning model. Quantitative metrics are measured to evaluate how well

the model is performing in tasks such as classification, regression, or clustering. Evaluation metrics allows you to compare different models with one another or to compare the variations in the same model. They are compared using the accuracies measured and comparing the predictions to the actual ground truth values in the dataset. ③ The evaluation metrics used in classification: Accuracy: The correctly classified instances. Precision: The true positive values out of the total predicted positives. F1 score: The mean of the Precision metric and the Sensitivity metric. Sensitivity: The true positive values out of the total actual positives. AUC-ROC: The classifiers ability to distinguish between classes. Confusion Matrix: ⑦ counts of true positives, true negatives, false positives, and false negatives. We choose certain metrics for the purpose of the problem that we are faced with in the analysis. ② [4] Model selection and Hyperparameter tuning

For the model selection process, we chose the best model architecture and the best classification algorithm to solve our data analysis problem. The purpose of model selection is to identify the model that generalizes well and performs the best. We use evaluation metrics to find the best model, the evaluation is often performed by using a training set and then evaluating it on a validation set. The model that performs the best on the validation set is selected as our final model. For hyperparameter tuning, we find the optimal values for the hyperparameters of a chosen model. The parameters are not learned from the data instead they are set before the machine learning process begins. The parameters control the behavior and performance of the model. We find the combination of hyperparameter values that results in the best performing model. We can use techniques such as: grid search, random search, or Bayesian optimization to perform hyperparameter tuning. Using these techniques we can find the combination of hyperparameters that yield the best performance, these are then chosen for our model. [5] Data set chosen for classification

The dataset was selected for classification, for predicting the quality of wine based on the independent variables. The dataset includes 11 physiochemical features that are both objective and measurable. The features encompass a range of different chemical properties such as acidity levels, residual sugar, chlorides, sulfur dioxide, density, pH, sulphates, and alcohol content. The features are quantitatively measurable, making them suitable for classification algorithms. There is a single output variable in the dataset that we seek to predict which is the quality score of the wine. The quality score is a continuous variable that ranges from 0-10. The quality score is the overall sensory perception of the wine. For the dataset to be used for classification we would need to insert the quality scores into classes or determine a threshold to distinguish between the different quality levels. The quality score is continuous so we would have to convert it into discrete class labels to perform classification. We could create classes such as "low quality", "medium quality" and "high quality" depending on the range of the quality scores. Therefor we would be able to frame this problem as a classification task, we would then be able to predict the class label (quality level) based on the input variables. The dataset is labeled, therefor each instance has a corresponding quality score, this makes the problem suitable for supervised learning, so we can train a classification model using the input variables as predictors and use the quality of wine as the target variable. Using the dataset and various classification algorithms we can develop a model that has learnt about the relationship between the physiochemical properties of the wine and the quality of the wine. We can use the trained model to predict the quality level of the new unlabeled instances. Analysis and Answer to Question

We will be performing prediction analysis on the dataset, determining the quality of wine. The data analysis process we will be following goes as follows: Data processing – We start the analysis process by examining if there are any outliers, missing values or strange values in the dataset. We clean the data by handling the outliers and missing values in the dataset. We normalize and scale the inputs so that the values have the same range. Exploratory Data Analysis (EDA) – We generate some visualizations, examine the distribution of the variables and examine the relationship between the variables so that we can identify patterns and dependencies in our data. Feature selection – Using various techniques such as correlation analysis, we select the most important and relevant features in the dataset so that the analysis is performed more efficiently and so that we get more accurate models.

Model selection – We choose our classification algorithms that we think will be suitable for our dataset, most importantly we will be using K-Nearest-Neighbor and Na ve Bayes. Train-Test Split – We will split our dataset into a training and testing set, the training set will be used to train our model and the testing set will be used to evaluate the performance of the trained model. We will split the model 70%-30%. Model training – We will use the classification algorithms that we picked out to fit them onto our training data. The algorithms will be taught on the patterns and relationships between the input and output variables. Model Evaluation – We will use metrics for classification to evaluate the performance of our trained model. Hyperparameter tuning – We perform hyperparameter tuning techniques to improve the performance of our model, techniques such as grid search. Model validation – We validate our final model on the test set to assess if the model can generalize well. We also check for overfitting once that is done, we will adjust the model accordingly. Prediction – Once we are happy with the trained model, we can predict the quality of wine given the physicochemical input variables. The main question the analysis will answer is what the quality of a wine is given certain input variables, we will also answer what the most efficient classification algorithm is for the dataset that we are using. ④ [6] Process used for the data analysis

Classification algorithms

⑧ Na ve Bayes classifier: Training – This is a probabilistic type of machine learning algorithm, which means that during the training phase our model estimate the probability distribution of each feature for each class. The algorithm calculates the prior probability of each class based on the frequency of occurrence in the training data. The features are conditionally independent, given the classes. It then builds a model based on the probabilities. Testing – For testing, the algorithm uses the model we trained to classify new unseen instances. The algorithm will calculate the posterior probability of each class and then select the class with the highest probability as the predicted class. The algorithm assumes that the new features are independent and thus calculates the probability of each class. ③ K-Nearest Neighbors: Training – This algorithm uses the training data as the model itself. The algorithm will store the features' values, labels and corresponding class labels. Testing – The algorithm will calculate the position of a new instance and all instances around it using a distance metric. It selects the K nearest neighbors closest to the new instance. The class label used of the K nearest neighbors is assigned to the predicted new instance. [7] Evaluation metrics

The most common metric that we will be using to evaluate the classification algorithms in our analysis is the accuracy score. The accuracy score measures the percentage of correctly predicted instances out of all the instances.

First, we use our trained model to make predictions, we compare the predicted labels with the true labels in the test set, we calculate the accuracy score by

dividing the number of correct predictions made by the total number of instances in the test set. The accuracy score will be represented as a value between 0 and 1 or we can represent the value as a percentage. We will then interpret the accuracy score, higher the score the better indication we will get that the model is performing well. We can use additional evaluation metrics like F1 score, recall etc. so that we can get a complete picture of the model's performance. [4] Changing Training and Testing splits

For tuning the accuracy of the predictions being made by our machine learning model we will be changing the training/test splits. By changing the training/test splits, we are essentially changing the composition of the data being used for testing of our model. This will change the estimated accuracy of the model, a larger test will provide more reliable estimates of the model's accuracy. The training set contributes to building model that generalizes well to unseen data. Changing the training/test splits will alter the distribution in the data. Finding the right balance between training and testing sets we can avoid overfitting and underfitting the model. Changing the training and testing splits can introduce variability in our model's performance, by experimenting with the different splits we can find the find the best variability and stability of the model's performance. We want our model to be fair and to reduce bias when making predictions, by overrepresenting or underrepresenting our groups and classes this can lead to biased predictions. The training/testing splits have to represent the overall distribution of the data. We can use a technique known as cross-validation to mitigate the effects of training/testing splits. For the technique we divide the data into multiple folds and perform testing and training on different combinations of these folds. We will yield a more robust evaluation of the model's performance averaging the results across the splits. [8] Evaluation and improvement of our model

For evaluating and improving on our trained model we follow a process: Firstly, we divide our dataset into three parts: ⑨ the training set, the validation set and the testing set. We will use the validation set to tune the hyperparameters and to monitor the performance of the training set during training. Next, we will choose the appropriate evaluation metrics for the classification algorithms that we are using. We will train a baseline model on the training set using the algorithm that we want to use. We will evaluate the baseline model on the validation set using our evaluation metrics, we get an initial understanding of the areas that need to be worked on in the model. We will fine tune the model by adjusting the hyperparameters, this will optimize the evaluation metrics on the validation set. Now, we will retrain the model using the optimal hyperparameters on the combined validation and training sets. Once the hyperparameters are sorted out, we will evaluate the retrained model on the test set, the evaluation will provide us with unbiased estimates of the model's performance. We will examine the predictions made on the test set to grasp the strengths and weaknesses of the model. If there are misclassified samples or recurring errors, we can further improve the model by performing error analysis. Error analysis will involve: collecting more data for the dataset, modifying the model, augmenting the data or trying various other techniques. [9]

References

Classification: [1] simplilearn. 2023. ⑩ Classification in Machine Learning: What it is & Classification Models. ⑨ [ONLINE] Available at: ⑩ https://www.simplilearn.com/tutorials/machine-learning-tutorial/classification-in-machine-learning. [Accessed 1 June 2023]. Feature Engineering: [2] towardsdatascience. 2021. ⑪ What is Feature Engineering — Importance, Tools and Techniques for Machine Learning. ⑨ [ONLINE] Available at: ⑫ https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10. [Accessed 1 June 2023]. [3] analyticsvidhya. 2022. Introduction to Feature Engineering – Everything You Need to Know!. ⑨ [ONLINE] Available at: ⑬ https://www.analyticsvidhya.com/blog/2021/10/a-beginners-guide-to-feature-engineering-everything-you-need-to-know/. [Accessed 1 June 2023]. Evaluation metrics: [4] analyticsvidhya. 2021. ① Metrics to Evaluate your Classification Model to take the right decisions. ⑨ [ONLINE] Available at: ⑭ https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/. [Accessed 1 June 2023]. ② Model selection and hyperparameter tuning: [5] towardsdatascience. 2020. Machine Learning: ② Model Selection and Hyperparameter Tuning. ⑨ [ONLINE] Available at: https://towardsdatascience.com/machine-learning-model-selection-and-hyperparameter-tuning-736158357dc4. [Accessed 1 June 2023]. ④ Data analysis process: [6] simplilearn. 2023. ⑮ What is Data Analysis? ⑯ Methods, Process and Types Explained. ⑨ [ONLINE] Available at: ⑯ https://www.simplilearn.com/data-analysis-methods-process-types-article. [Accessed 1 June 2023]. Classification algorithms: [7] monkeylearn. 2023. 5 Types of Classification Algorithms in Machine Learning. ⑨ [ONLINE] Available at: https://monkeylearn.com/blog/classification-algorithms/. [Accessed 1 June 2023]. ⑰ Train Test split: [8] v7labs. 2021. ② Train Test Validation Split: How To & Best Practices [2023]. ⑨ [ONLINE] Available at: https://www.v7labs.com/blog/train-validation-test-set. [Accessed 1 June 2023]. Evaluation and improving model: [9] techtarget. 2021. ⑨ 3 ways to evaluate and improve machine learning models. [ONLINE] Available at: https://www.techtarget.com/searchenterpriseai/post/3-ways-to-evaluate-and-improve-machine-learning-models. [Accessed 1 June 2023].

Source Matches (38)	
① Student paper	100 %

Submitted paper	Original source
Table of contents.	Table of Contents

2 http://pmuu.everwoodbiocostruzioni.it/hyperparameter-tuning-sklearn.html ⊡	79 %

Submitted paper	Original source
Model selection and Hyperparameter tuning...	model selection, model tuning and hyperparameter tuning

3 Student paper	71 %

Submitted paper	Original source
Evaluation metrics for Process.	Evaluation Metrics for Classification

4 My paper	83 %

Submitted paper	Original source
The main question we want to answer is: "What is the quality of wine given the wine characteristics?" To answer the question, the python programming language will be used to pre-process the data, generate graphs, and perform analysis. The features that we will be looking at are:	The main question we want answering is To answer the question, the python programming language will be used to pre-process the data, generate graphs, and perform analysis using supervised algorithms The features that we will be looking at are

1 Student paper	63 %

Submitted paper	Original source
decision trees, logistic regression, support vector machines (SVM), k-nearest neighbors (KNN), random forests, Na ve bayes and neural networks.	Examples of supervised learning algorithms include linear regression, logistic regression, decision trees, random forests, support vector machines, and neural networks

5 Student paper	76 %

Submitted paper	Original source
accuracy, precision, recall, F1 score, and area under the ROC curve (AUC-ROC).	Learning Rate Weight Decay Accuracy (%) Precision (%) Recall (%) F1-Score Area Under ROC Curve

6 Student paper	72 %

Submitted paper	Original source
[1] Feature Engineering	The feature engineering of the

3 Student paper	80 %

Submitted paper	Original source
The evaluation metrics used in classification:	Evaluation Metrics for Classification

7 http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/scorers.html ⊡	67 %

Submitted paper	Original source
counts of true positives, true negatives, false positives, and false negatives.	The MCC metric combines the true positives, false positives, true negatives, and false negatives using the equation described below

2 http://pmuu.everwoodbiocostruzioni.it/hyperparameter-tuning-sklearn.html ⊡	72 %

Submitted paper	Original source
[4] Model selection and Hyperparameter tuning	model selection, model tuning and hyperparameter tuning

(4) My paper	64 %

Submitted paper	Original source
[6] Process used for the data analysis	Process of data analysis

(8) https://addi.ehu.es:443/bitstream/handle/10810/43895/TFM_PAULA_REYERO_14_05_2020.pdf	87 %

Submitted paper	Original source
Na ve Bayes classifier:	• Na ve Bayes Classifier (NB)

(3) Student paper	86 %

Submitted paper	Original source
K-Nearest Neighbors:	K-Nearest Neighbors Model

(9) Student paper	80 %

Submitted paper	Original source
the training set, the validation set and the testing set.	the training set and the validation set

(10) Student paper	83 %

Submitted paper	Original source
Classification in Machine Learning:	tutorial/classification-in-machine-learning

(9) Student paper	100 %

Submitted paper	Original source
[ONLINE] Available at:	[online] Available at

(10) Student paper	100 %

Submitted paper	Original source
https://www.simplilearn.com/tutorials/machine-learning-tutorial/classification-in-machine-learning.	https://www.simplilearn.com/tutorials/machine-learning-tutorial/classification-in-machine-learning

(11) Student paper	100 %

Submitted paper	Original source
What is Feature Engineering — Importance, Tools and Techniques for Machine Learning.	What is feature engineering - importance, tools and techniques for machine learning

(9) Student paper	100 %

Submitted paper	Original source
[ONLINE] Available at:	[online] Available at

(12) Student paper		**85 %**
Submitted paper	**Original source**	
https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10.	Retrieved May 4, 2022, from https://towardsdatascience.com/ what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10	

(9) Student paper		**100 %**
Submitted paper	**Original source**	
[ONLINE] Available at:	[online] Available at	

(13) Student paper		**70 %**
Submitted paper	**Original source**	
https://www.analyticsvidhya.com/blog/2021/10/a-beginners-guide-to-feature-engineering-everything-you-need-to-know/.	https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-feature-engineering	

(1) Student paper		**82 %**
Submitted paper	**Original source**	
Metrics to Evaluate your Classification Model to take the right decisions.	(2022) Metrics to evaluate your classification model to take the right decisions, Analytics Vidhya	

(9) Student paper		**100 %**
Submitted paper	**Original source**	
[ONLINE] Available at:	[online] Available at	

(14) Student paper		**96 %**
Submitted paper	**Original source**	
https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/.	[3] https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your- classification-model-to-take-the-right-decisions/	

(2) http://pmuu.everwoodbiocostruzioni.it/hyperparameter-tuning-sklearn.html		**79 %**
Submitted paper	**Original source**	
Model selection and hyperparameter tuning:	model selection, model tuning and hyperparameter tuning	

(2) http://pmuu.everwoodbiocostruzioni.it/hyperparameter-tuning-sklearn.html		**79 %**
Submitted paper	**Original source**	
Model Selection and Hyperparameter Tuning.	model selection, model tuning and hyperparameter tuning	

(9) Student paper		**100 %**
Submitted paper	**Original source**	
[ONLINE] Available at:	[online] Available at	

(4) My paper		96 %
Submitted paper	**Original source**	
Data analysis process:	Process of data analysis	

(15) Student paper		100 %
Submitted paper	**Original source**	
What is Data Analysis?	What is data analysis	

(16) Student paper		100 %
Submitted paper	**Original source**	
Methods, Process and Types Explained.	Methods, Process and types explained	

(9) Student paper		100 %
Submitted paper	**Original source**	
[ONLINE] Available at:	[online] Available at	

(16) Student paper		100 %
Submitted paper	**Original source**	
https://www.simplilearn.com/data-analysis-methods-process-types-article.	https://www.simplilearn.com/data-analysis-methods-process-types-article	

(9) Student paper		100 %
Submitted paper	**Original source**	
[ONLINE] Available at:	[online] Available at	

(17) ProQuest document		100 %
Submitted paper	**Original source**	
Train Test split:	Train-test split	

(2) http://pmuu.everwoodbiocostruzioni.it/hyperparameter-tuning-sklearn.html		87 %
Submitted paper	**Original source**	
Train Test Validation Split:	Train-Validation Split	

(9) Student paper		100 %
Submitted paper	**Original source**	
[ONLINE] Available at:	[online] Available at	

(9) Student paper		96 %
Submitted paper	**Original source**	
3 ways to evaluate and improve machine learning models. [ONLINE] Available at: https://www.techtarget.com/searchenterpriseai/post/3-ways-to-evaluate-and-improve-machine-learning-models.	3 ways to evaluate and improve machine learning models [online] Available at https://www.techtarget.com/searchenterpriseai/post/3-ways-to-evaluate-and-improve-machine-learning-models [Accessed 6 Jun	