|  |
| --- |
| **Analysis of biopsy of breast tissue**\_  Versie: 1 |

**Jorian Borst (574013)**

**Studiejaar: 3**

**Datum: 12 April 2023**

**INHOUDSOPGAVE**

[1 Inleiding 3](#_Toc116052662)

[2 De bedrijfsbeschrijving en -context 4](#_Toc116052663)

[2.1 Bedrijfsbeschrijving 4](#_Toc116052664)

[2.2 Bedrijfscontext 4](#_Toc116052665)

[3 Beroepsproducten 5](#_Toc116052666)

[3.1 Traineeship rapportage 5](#_Toc116052667)

[3.1.1 Beroepsproduct 5](#_Toc116052668)

[3.1.2 Totstandkoming 5](#_Toc116052669)

[3.1.3 Verschillen 6](#_Toc116052670)

[4 Reflectie 6](#_Toc116052671)

[5 conclusies en aanbevelingen 6](#_Toc116052672)

[6 Literatuur- en bronnenlijst 6](#_Toc116052673)

[References 6](#_Toc116052674)

# Inleiding

# Analyse van de data

*Analyze* the data using descriptive statistics. I am aware that you just finalized an assignment about data analysis, but remember no application of machine learning is ever possible without properly analyzing the data and its characteristics. Therefor keep it as simple as possible but informative to give you an impression of the data.  
  
Minimum requirements concerning your analyses:

1. Produce all the basic statistical values to give insight in central tendency of the data (mean, median, mode, range, variance, standard deviation, percentiles)
2. Check for statistically significant changes in feature means based on the binary dependent variable (malignant/benign)
3. Determine covariation and correlation between the descriptive features
4. Visualize the data: bars/histograms, scatterplots, box plots (use ggplot)
5. Note: ***no*** ***data pre-processing***, “just” getting to know the data.

# Data Preparation

*Prepare*your *data* set for further ML processing (for tasks 3, 4 and 5, possibly in a different way per model alternative). Some pre-processing steps you may consider to apply: normalizing continuous data, binning continuous data and factorizing (fi the dependent variable), engineering new values from old values, splitting the overall data sets in training, test and validation sets. **General task information**You have to experiment and describe what exactly you did to the data, why you considered doing it and subsequently perform suitable model fitting and validation (see below). Of course you have to compare the alternative models within the class of models (for task 3 = two logistic regression based variants, task 4 = two information based variants, so decision tree variants, task 5 = two similarity based variants, so k-NN variants).  
  
To give you one possible ***acceptable*** scenario for one of the tasks: “perform logistic regression using all the independent features at hand without performing any special data pre-processing and perform logistic regression dropping a number of features on the basis of possible correlation between the independent features themselves and for instance raising the feature value to for instance power 2 (= kwadrateren van een feature waarde)”.

So two modeling and validation trajectories per model type are required.  
  
Must: compare models using the validation sets/procedures!

Tip: it is wise to think about setting seed values! You compare model variants (see task 7). If you want a fair comparison, you have to be sure the models are trained and validated using the same sets.

# logistic regression

**:** *Create two well tested predictive models* using logistic regression.

# decision tree

*Create two well tested predictive models* using a *decision tree* algorithm.

# nearest neighbor

*Create two well tested predictive models* using a *nearest neighbor* algorithm.

# Analyse van de resultaten

Analyze the results/differences between the task 3-4-5 models in an statistically honest way using resampling techniques like cross validation/bagging when applicable and tunes the model hyperparameters.

