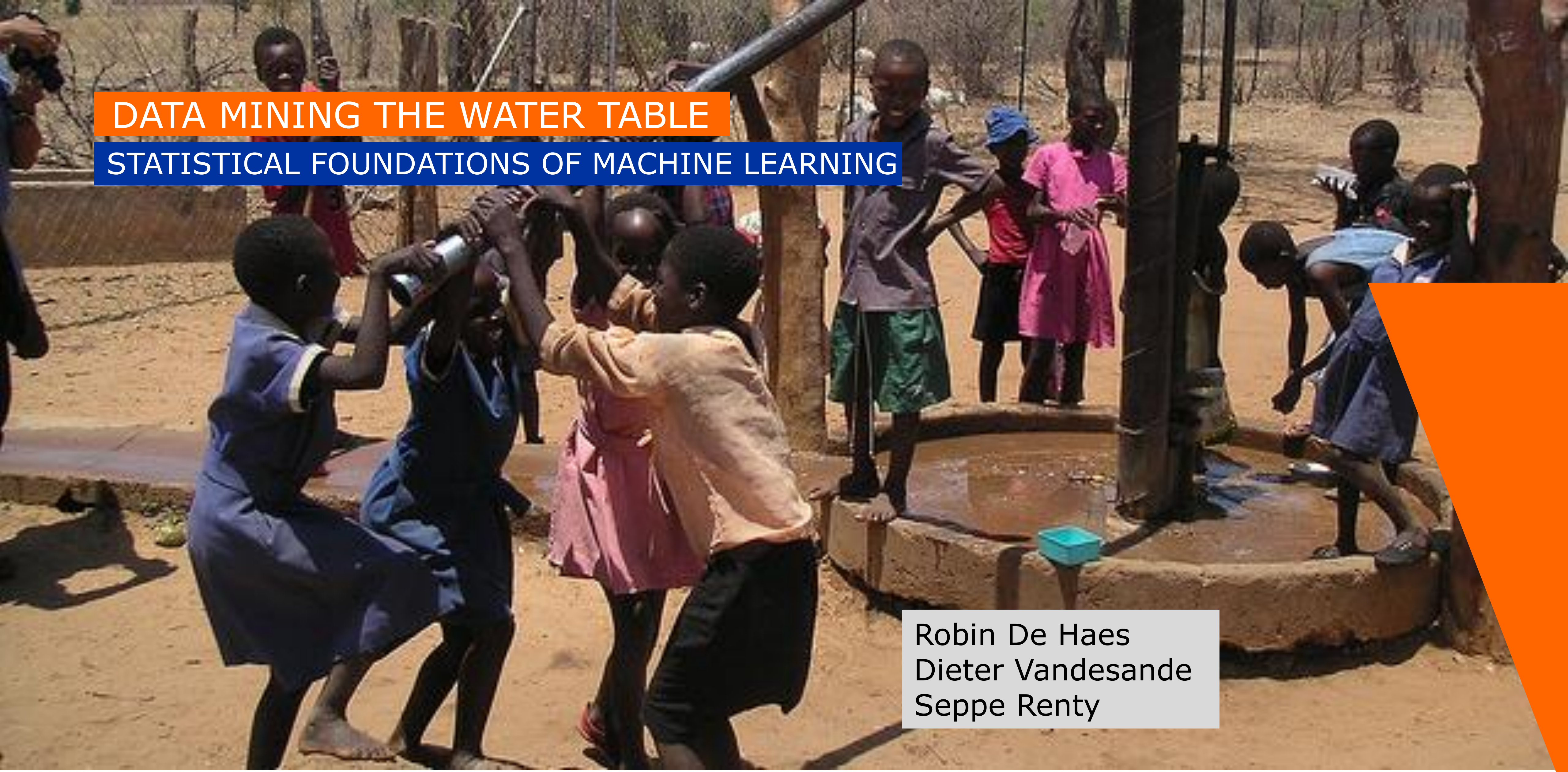


DATA MINING THE WATER TABLE

STATISTICAL FOUNDATIONS OF MACHINE LEARNING

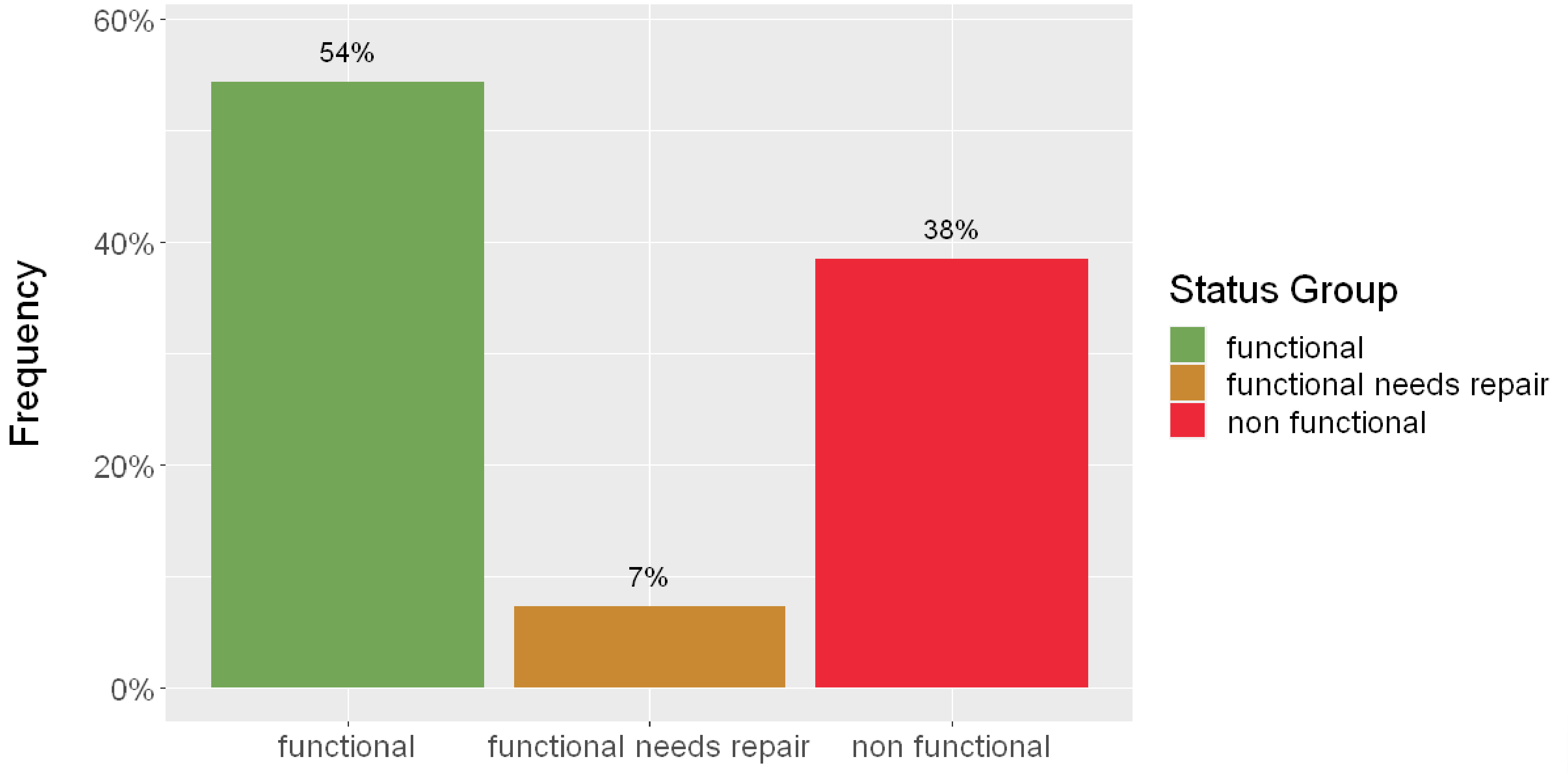
Robin De Haes
Dieter Vandesande
Seppe Renty

VRIJE UNIVERSITEIT BRUSSEL

DATA EXPLORATION

STATUS GROUP

TARGET LABELS



FEATURES

amount_tsh	gps_height	latitude	basin	region_code
date_recorded	waterpoint_type_group	wpt_name	subvillage	district_code
funder	management_group	num_private	region	lga
longitude	extraction_type_group	id	management	installer
quantity	extraction_type_class	source	source_type	source_class
permit	extraction_type	water_quality	quality_group	recorded_by
construction_year	scheme_management	payment_type	quantity_group	scheme_name
ward	population	public_meeting	payment	waterpoint_type

FEATURES
NUMERICAL

amount_tsh	gps_height	latitude	basin	region_code
date_recorded	waterpoint_type_group	wpt_name	subvillage	district_code
funder	management_group	num_private	region	lga
longitude	extraction_type_group	id	management	installer
quantity	extraction_type_class	source	source_type	source_class
permit	extraction_type	water_quality	quality_group	recorded_by
construction_year	scheme_management	payment_type	quantity_group	scheme_name
ward	population	public_meeting	payment	waterpoint_type

FEATURES
CATEGORICAL

amount_tsh	gps_height	latitude	basin	region_code
date_recorded	waterpoint_type_group	wpt_name	subvillage	district_code
funder	management_group	num_private	region	lga
longitude	extraction_type_group	id	management	installer
quantity	extraction_type_class	source	source_type	source_class
permit	extraction_type	water_quality	quality_group	recorded_by
construction_year	scheme_management	payment_type	quantity_group	scheme_name
ward	population	public_meeting	payment	waterpoint_type

FEATURES

REDUNDANCY

amount_tsh	gps_height	latitude	basin	region_code
date_recorded	waterpoint_type_group	wpt_name	subvillage	district_code
funder	management_group	num_private	region	lga
longitude	extraction_type_group	id	management	installer
quantity	extraction_type_class	source	source_type	source_class
permit	extraction_type	water_quality	quality_group	recorded_by
construction_year	scheme_management	payment_type	quantity_group	scheme_name
ward	population	public_meeting	payment	waterpoint_type

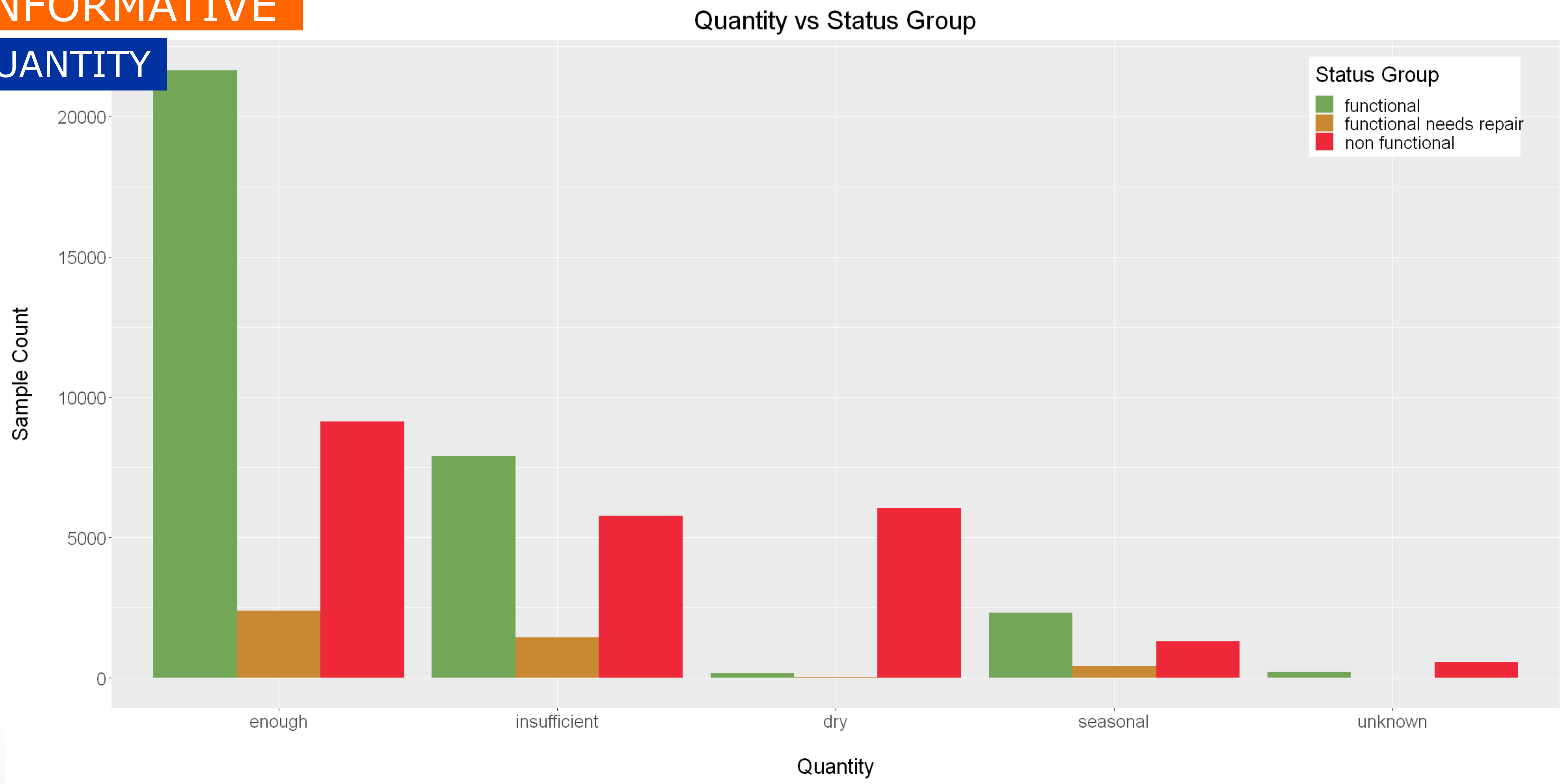
FEATURES

REDUNDANCY

amount_tsh	gps_height	latitude	basin	region_code
date_recorded	waterpoint_type_group	wpt_name	subvillage	district_code
funder	management_group	num_private	region	lga
longitude	extraction_type_group	id	management	installer
quantity	extraction_type_class	source	source_type	source_class
permit	extraction_type	water_quality	quality_group	recorded_by
construction_year	scheme_management	payment_type	quantity_group	scheme_name
ward	population	public_meeting	payment	waterpoint_type

INFORMATIVE

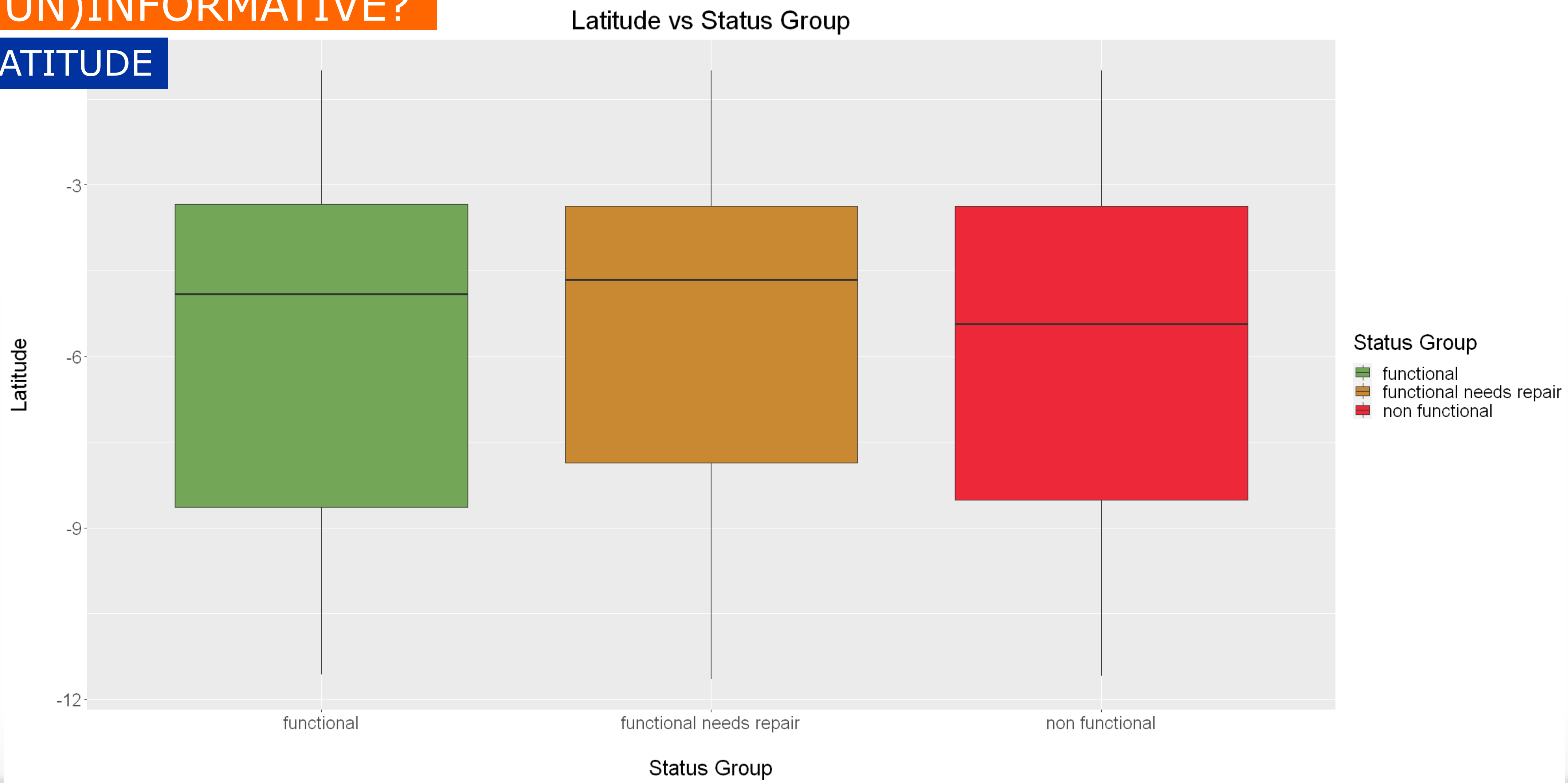
QUANTITY



VRIJE UNIVERSITEIT BRUSSEL

(UN)INFORMATIVE?

LATITUDE

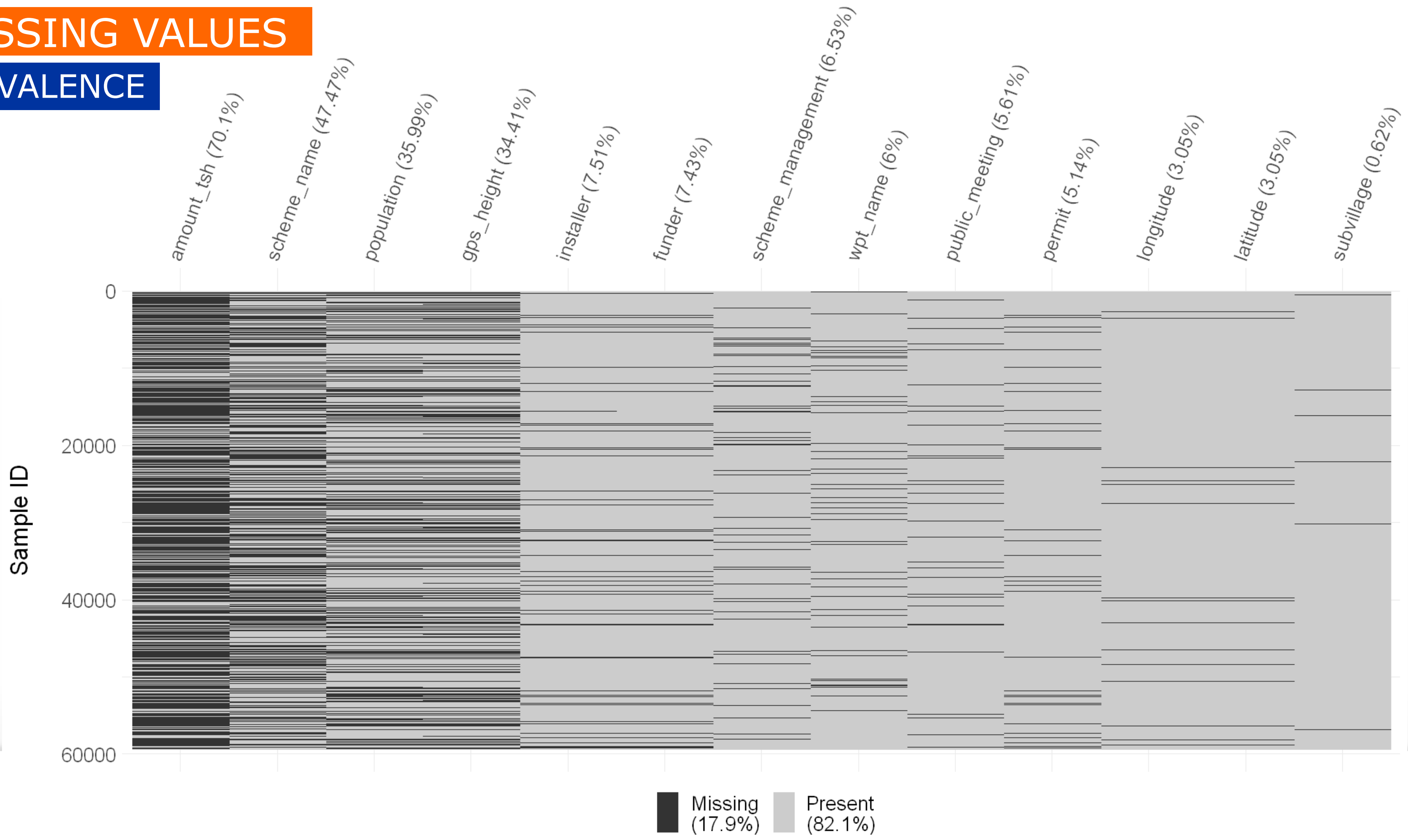


VRIJE UNIVERSITEIT BRUSSEL

DATA PREPROCESSING

MISSING VALUES

PREVALENCE



MISSING VALUE IMPUTATION

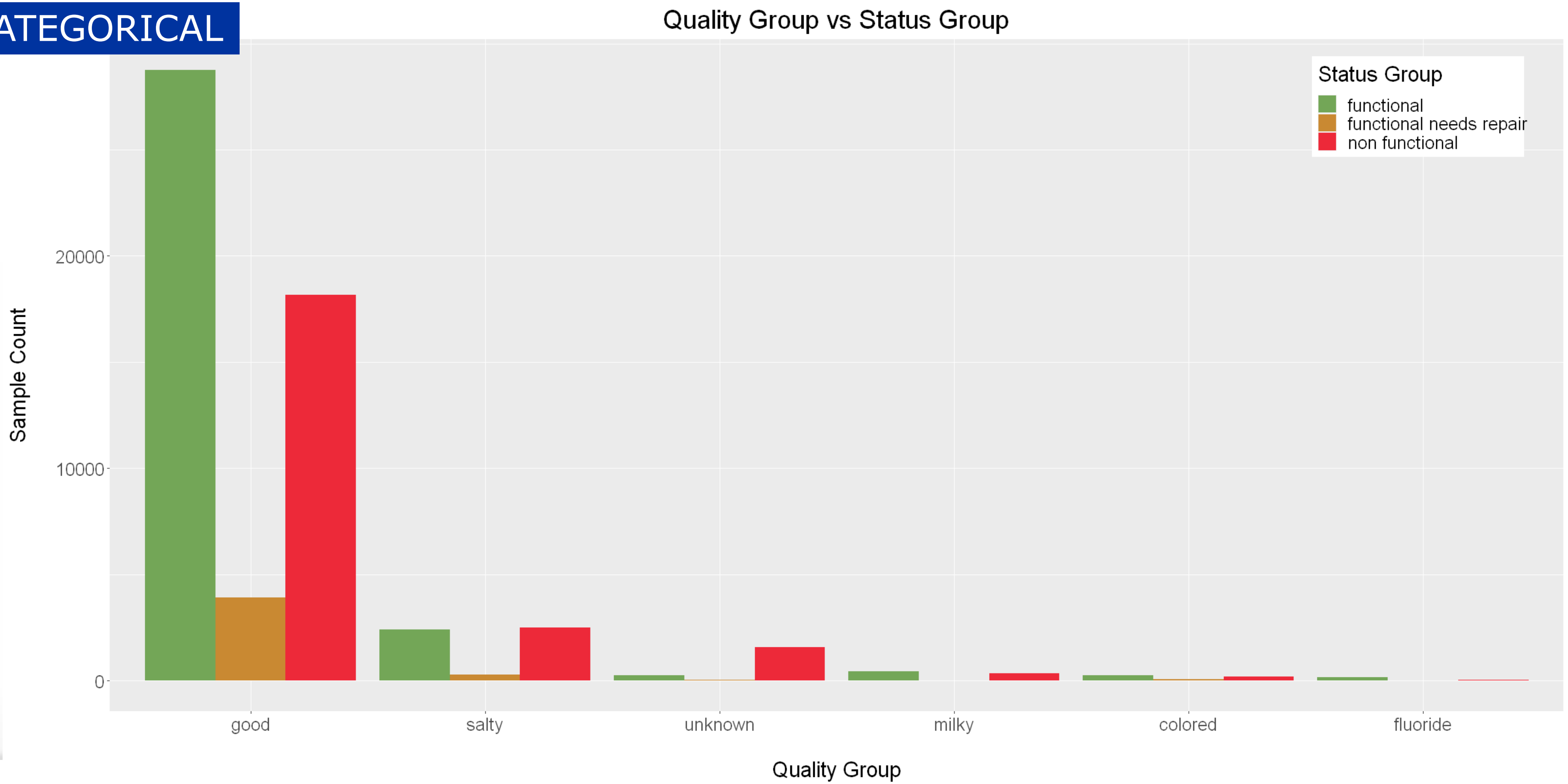
CATEGORICAL

- empty string ('')
- -
- 0
- not known
- unknown
- none
- no

→ unknown

MISSING VALUE IMPUTATION

CATEGORICAL



MISSING VALUE IMPUTATION

NUMERICAL

0 → Invalid?

→ Median

→ Geographic mean

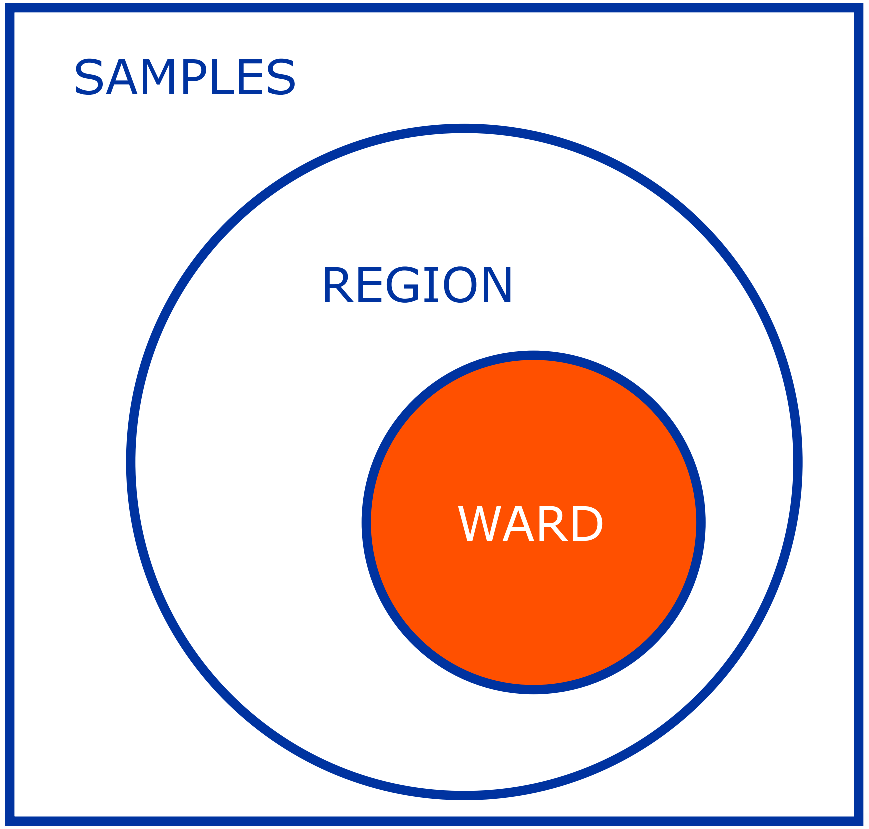
MISSING VALUE IMPUTATION

NUMERICAL

0 → Invalid?

→ Median

→ Geographic mean



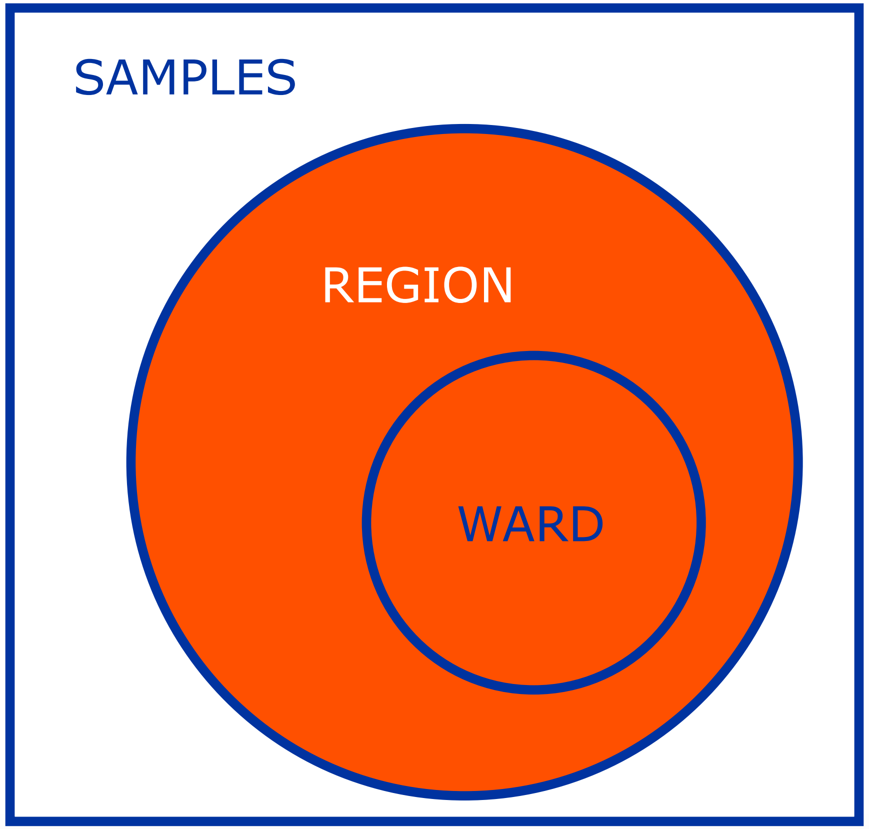
MISSING VALUE IMPUTATION

NUMERICAL

0 → Invalid?

→ Median

→ Geographic mean



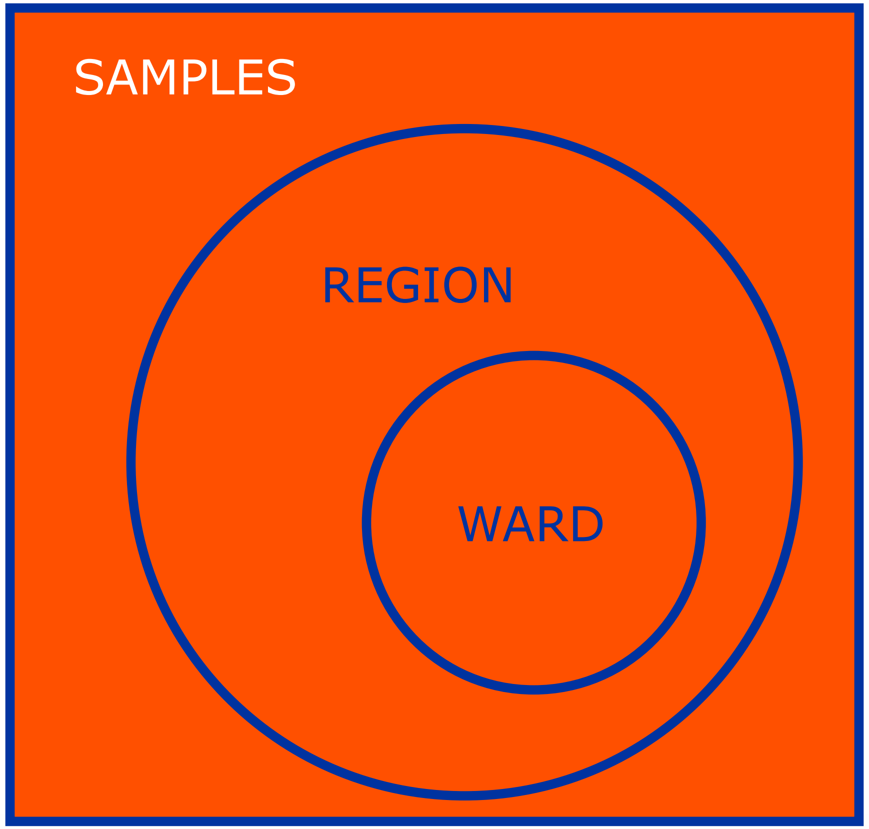
MISSING VALUE IMPUTATION

NUMERICAL

0 → Invalid?

→ Median

→ Geographic mean



FEATURE ENGINEERING
CREATION & MODIFICATION

Creation

- pump_age
  = date_recorded[year] - construction_year
- season
  = seasonal binning of date_recorded[month]

Modification

- Manual splits/merges
- Low-frequency merging (< 1%)

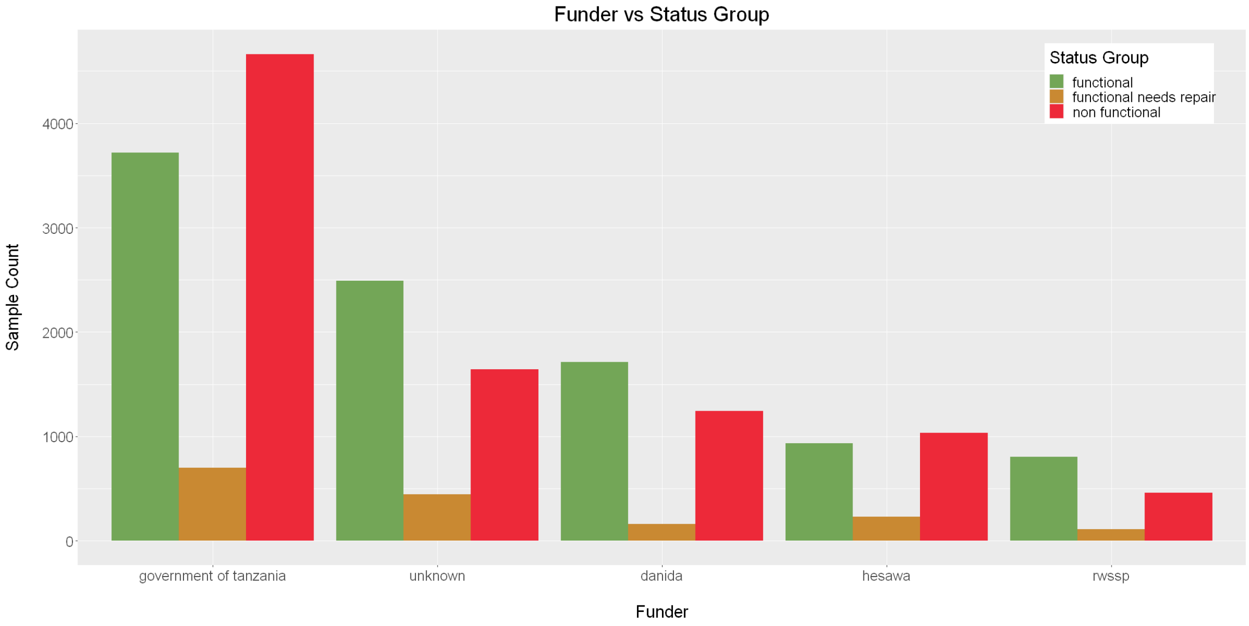
FEATURE ENGINEERING
CREATION & MODIFICATION

Creation

- pump_age
  = date_recorded[year] - construction_year
- season
  = seasonal binning of date_recorded[month]

Modification

- Manual splits/merges
- Low-frequency merging (< 1%)



FEATURE SELECTION
FILTER

Manual selection

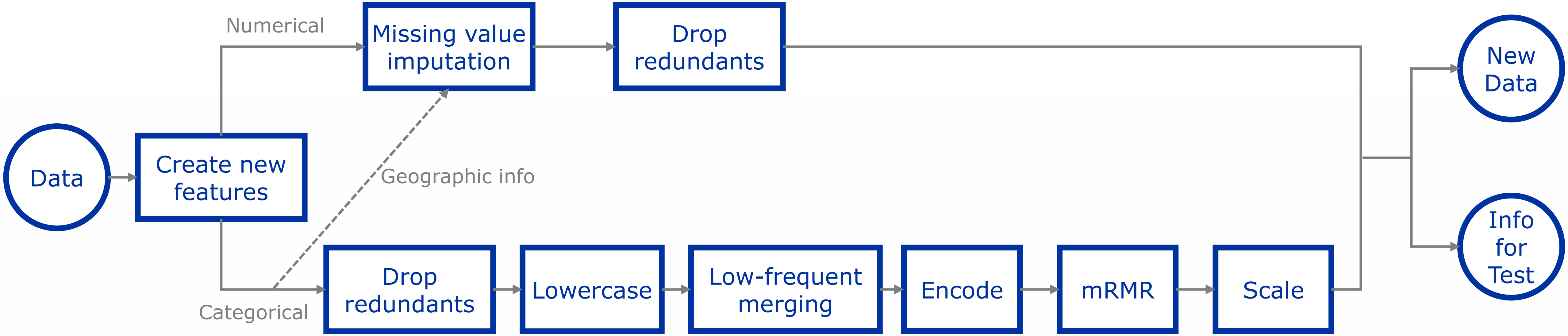
- Dropping redundant features

Automatic selection

- One-hot encoded features
- Maximum Relevancy Minimum Redundancy (80 features)

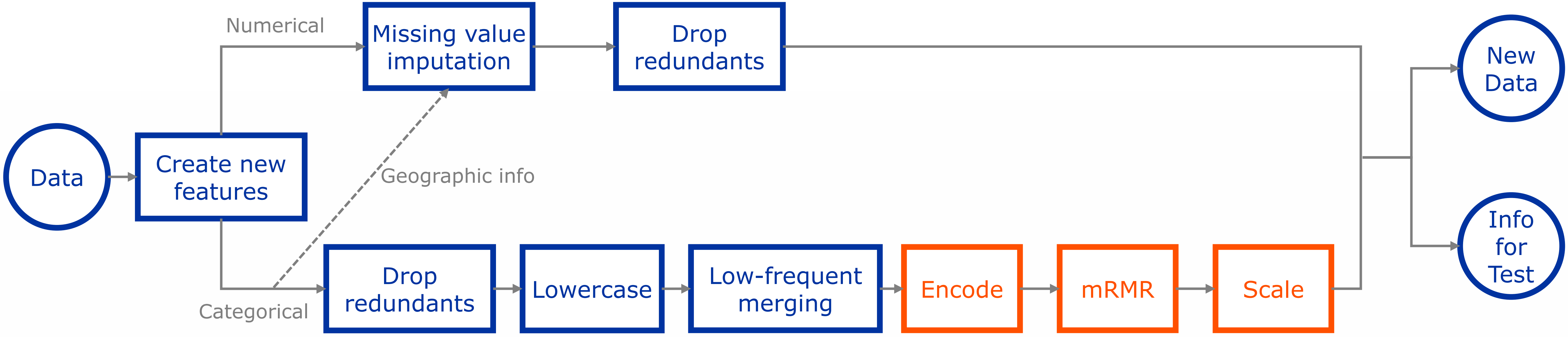
PREPROCESSING PIPELINE

TRAINING DATA



PREPROCESSING PIPELINE

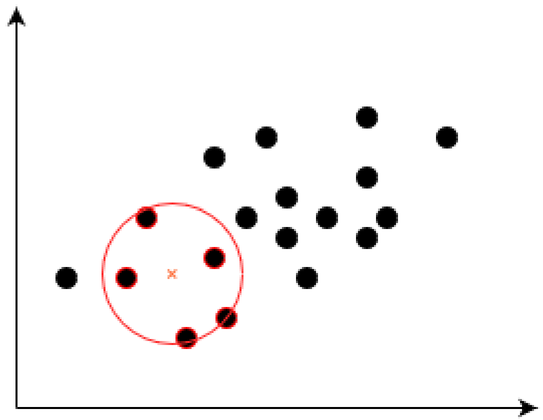
OPTIONAL SETTINGS



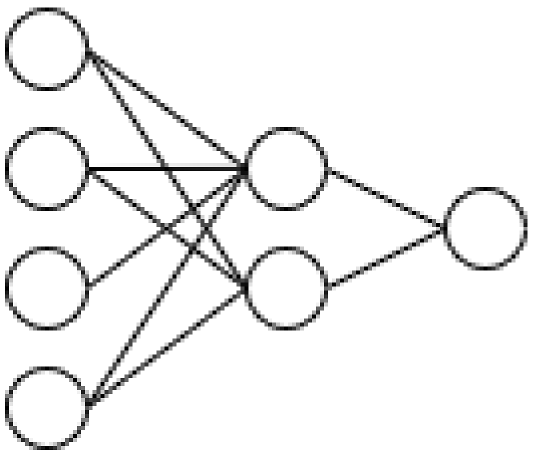
TUNING MODELS

TUNING MODELS

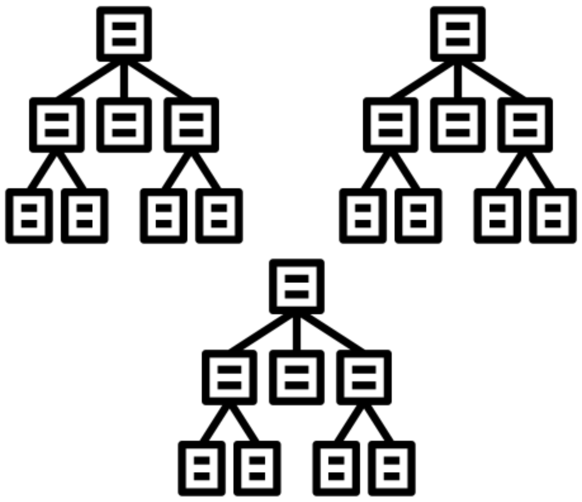
KNN



NNET



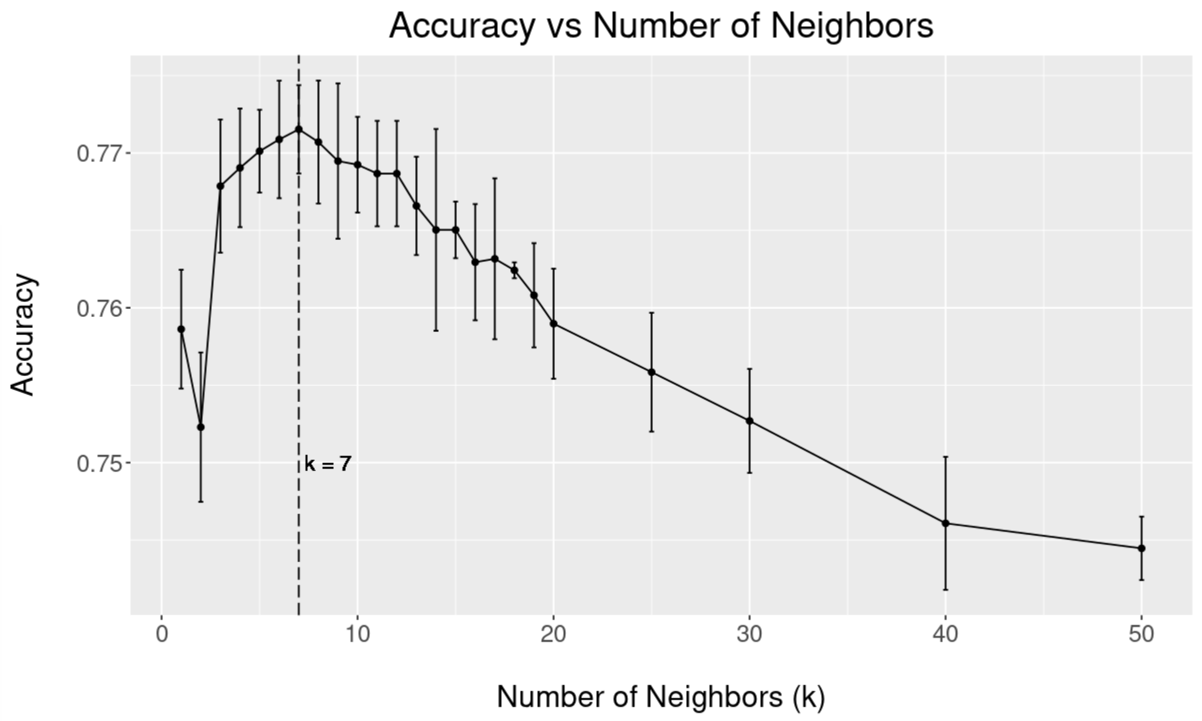
RANDOM FORESTS



KNN

HYPERPARAMETER TUNING

- **Preprocessing used:**
  - **One Hot Encoding**
  - **Scaling**
  - **mRMR**
- **K = number of neighbours**
- **Distance = Euclidean**



KNN

MODEL ANALYSIS

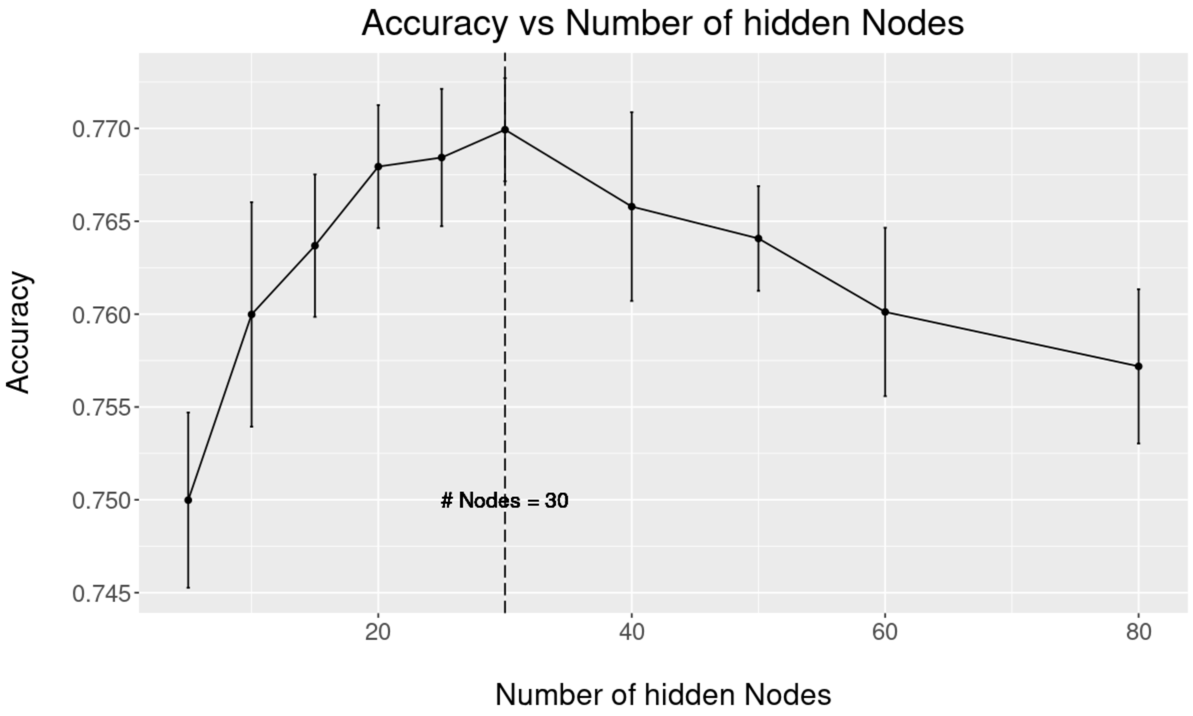
Accuracy of 77.5 % +/- 0.2 %

Prediction ->	Functional	Functional Needs repair	Non Functional
Functional	87.3%	2.2%	10.5%
Functional Needs repair	53.7%	28.8%	17.2%
Non Functional	26.2%	1.9%	71.9%

NNET

HYPERPARAMETER TUNING

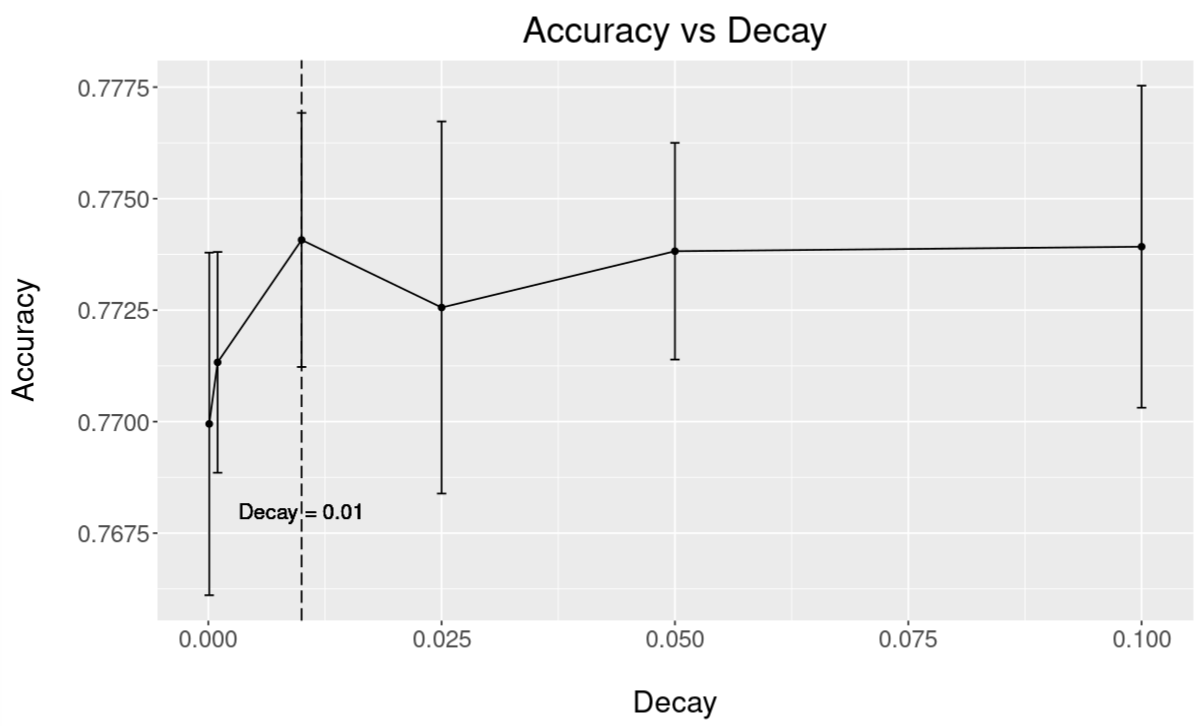
- **Preprocessing used:**
  - **One Hot Encoding**
  - **Scaling**
  - **mRMR**
- **Size of the hidden layer**
- **Decay**



NNET

HYPERPARAMETER TUNING

- **Preprocessing used:**
  - **One Hot Encoding**
  - **Scaling**
  - **mRMR**
- **Size of the hidden layer**
- **Decay**



NNET

MODEL ANALYSIS

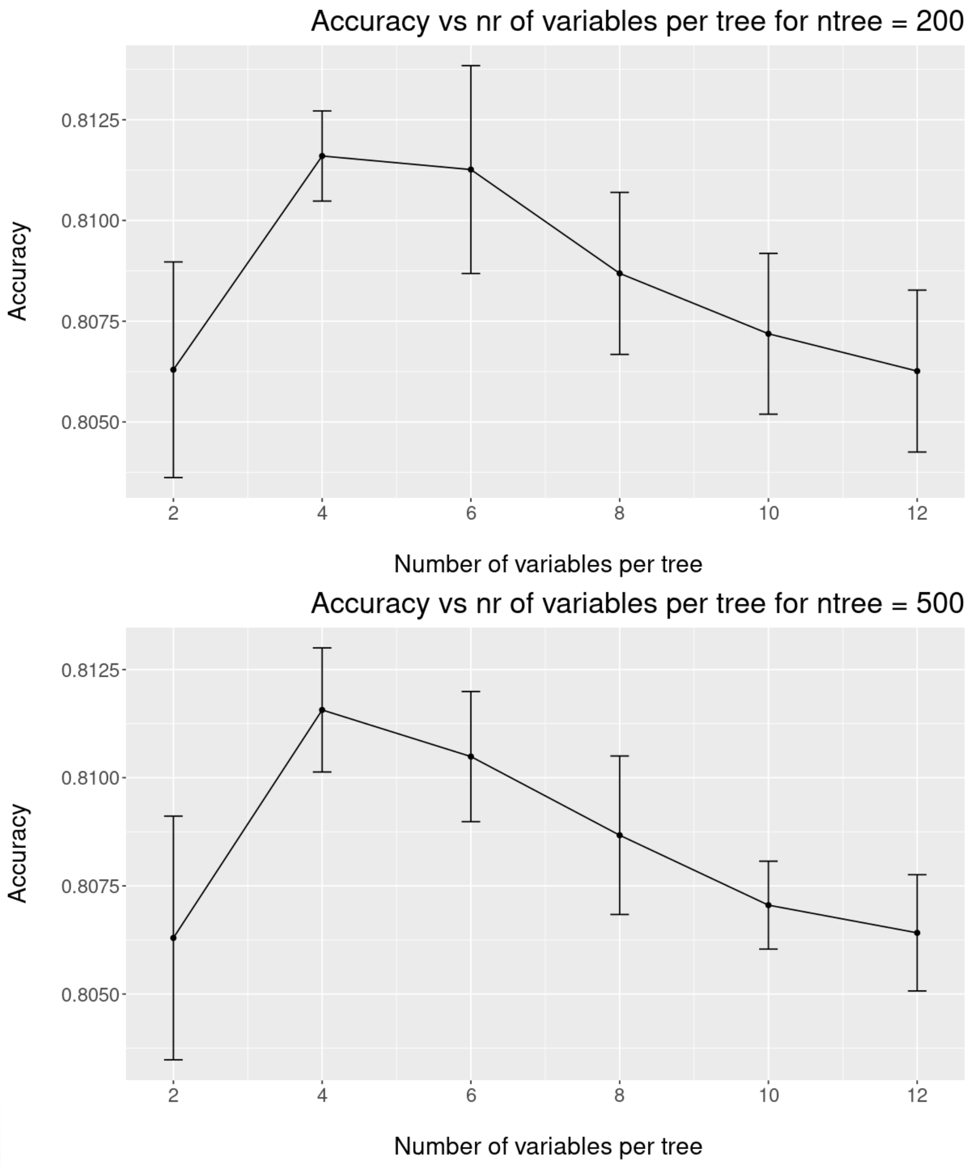
Accuracy of 77.4 % +/- 0.4 %

Prediction ->	Functional	Functional Needs repair	Non Functional
Functional	86.8%	2.2%	11%
Functional Needs repair	54.9%	26.2%	18.9%
Non Functional	25.0%	1.8%	73.2%

RANDOM FOREST

HYPERPARAMETER TUNING

- **Preprocessing needed:**
  - **Categorical Data**
- **Number of variables per tree**
- **Number of trees per forest**

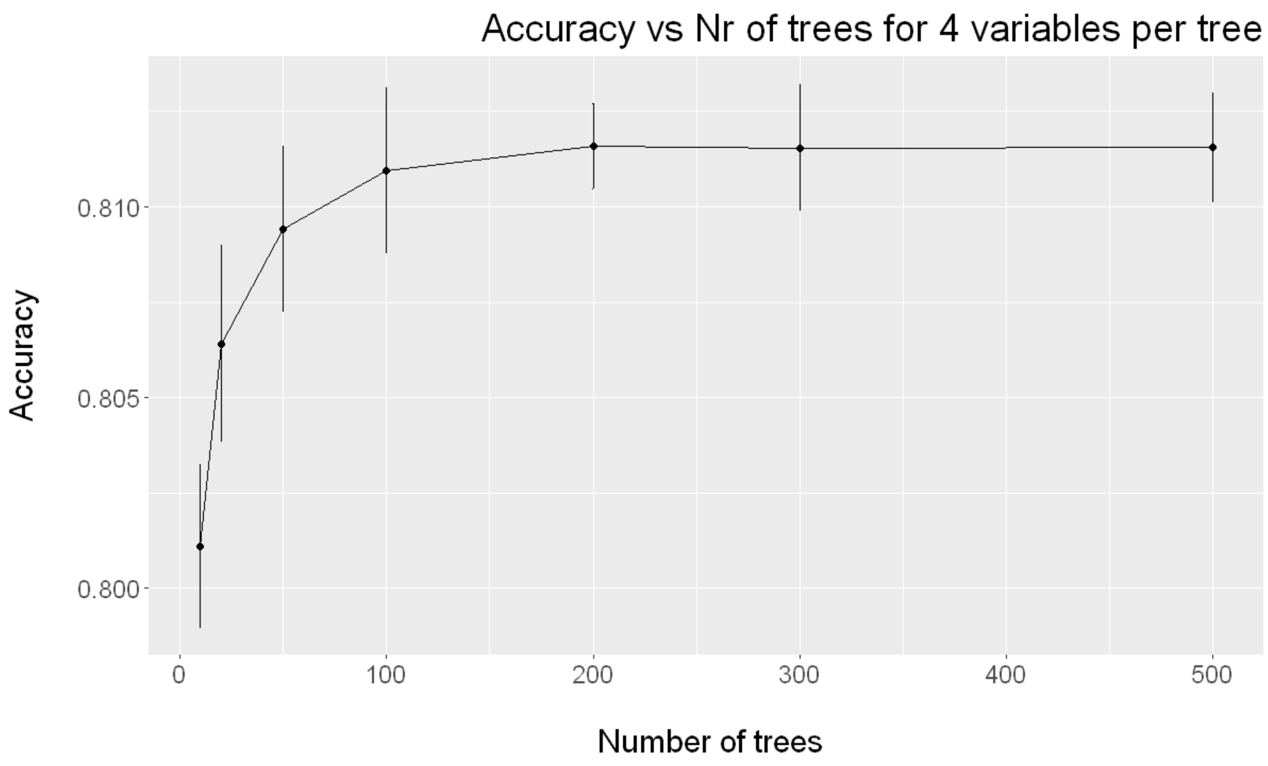


VRIJE UNIVERSITEIT BRUSSEL

RANDOM FOREST

HYPERPARAMETER TUNING

- **Preprocessing needed:**
  - **Categorical Data**
- **Number of variables per tree**
- **Number of trees per forest**



RANDOM FOREST

MODEL ANALYSIS

Accuracy of 81.2 % +/- 0.1 %

Prediction ->	Functional	Functional Needs repair	Non Functional
Functional	90.0%	2.0%	8.0%
Functional Needs repair	53.4%	32.4%	14.2%
Non Functional	21.0%	1.3%	77.6%

COMPARISON OF ALL THE TUNED MODELS

Model	Accuracy +/- SD
ElasticNet	73.8 +/- 0.6
Decision Tree	78.7 +/- 0.4
Random Forest	**81.2 +/- 0.1**
kNN	77.5 +/- 0.3
SVM	77.9 +/- 0.5
Neural Network	77.4 +/- 0.4

Woohoo! We processed your submission!

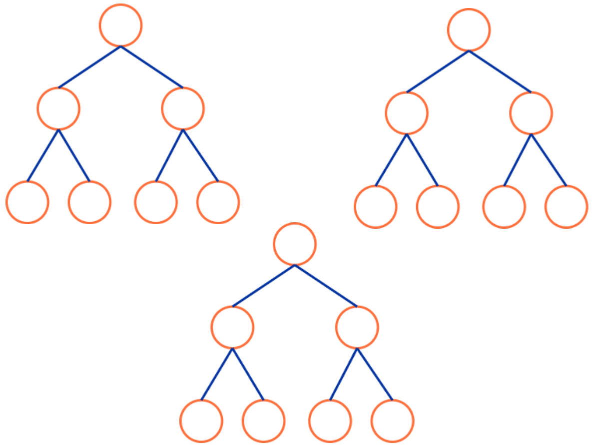
Your score for this submission is:

0.8172

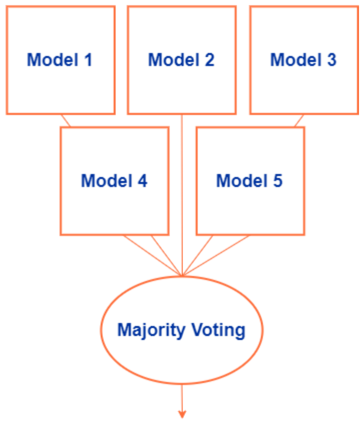
ALTERNATIVE MODELS

ALTERNATIVE MODELS

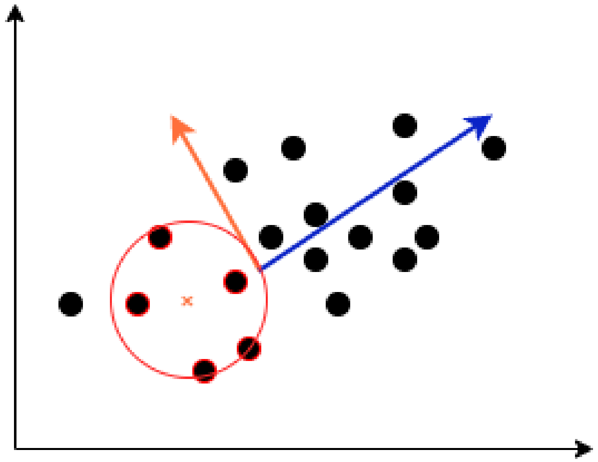
Gradient Boosting Trees



Ensemble of models



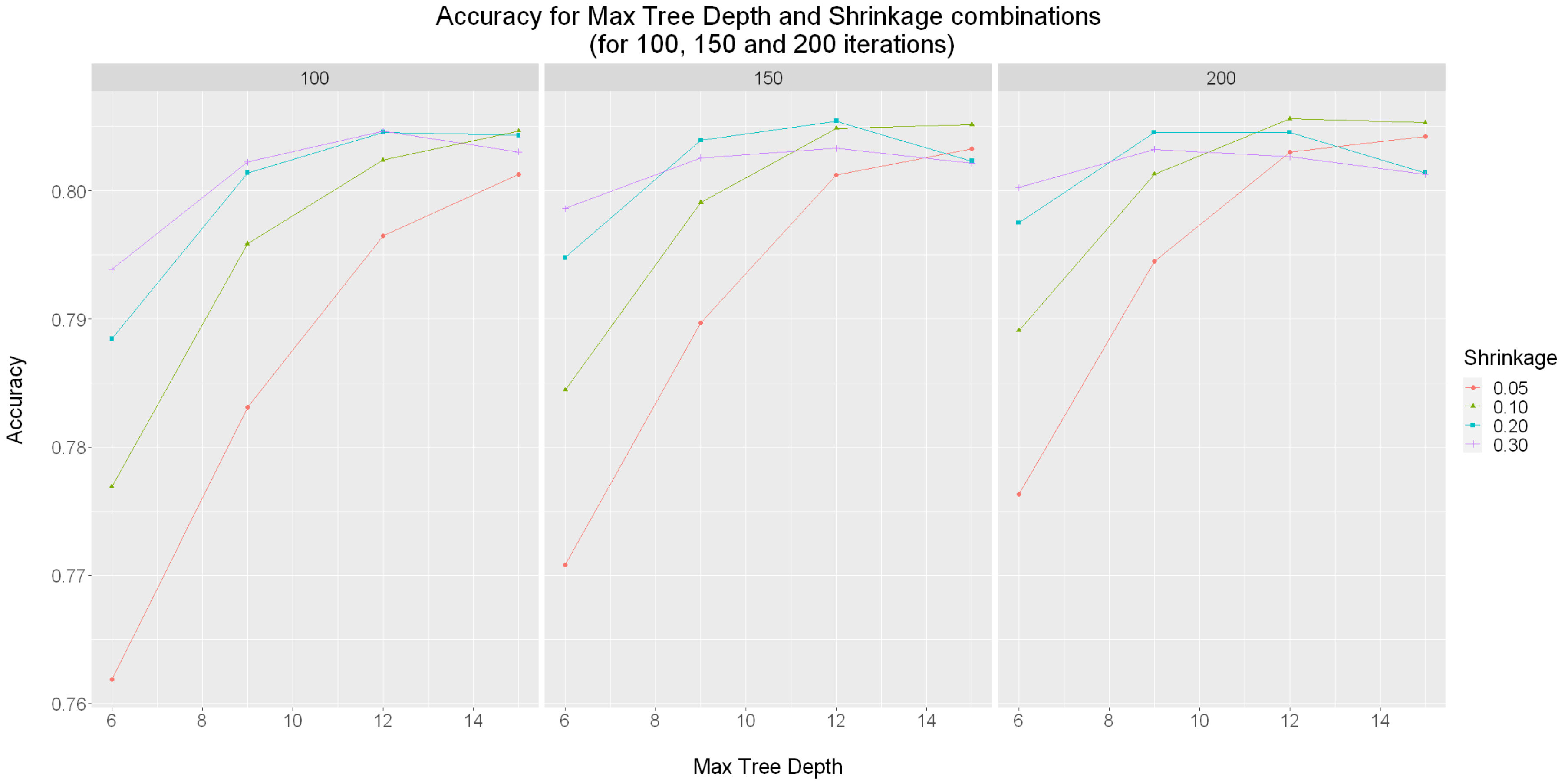
KNN with PCA



GRADIENT BOOSTING TREES

HYPERPARAMETER TUNING

- **Nrounds**
- **Eta**
- **Max depth**
- Min child weight
- Colsample Bytree



GRADIENT BOOSTING TREES

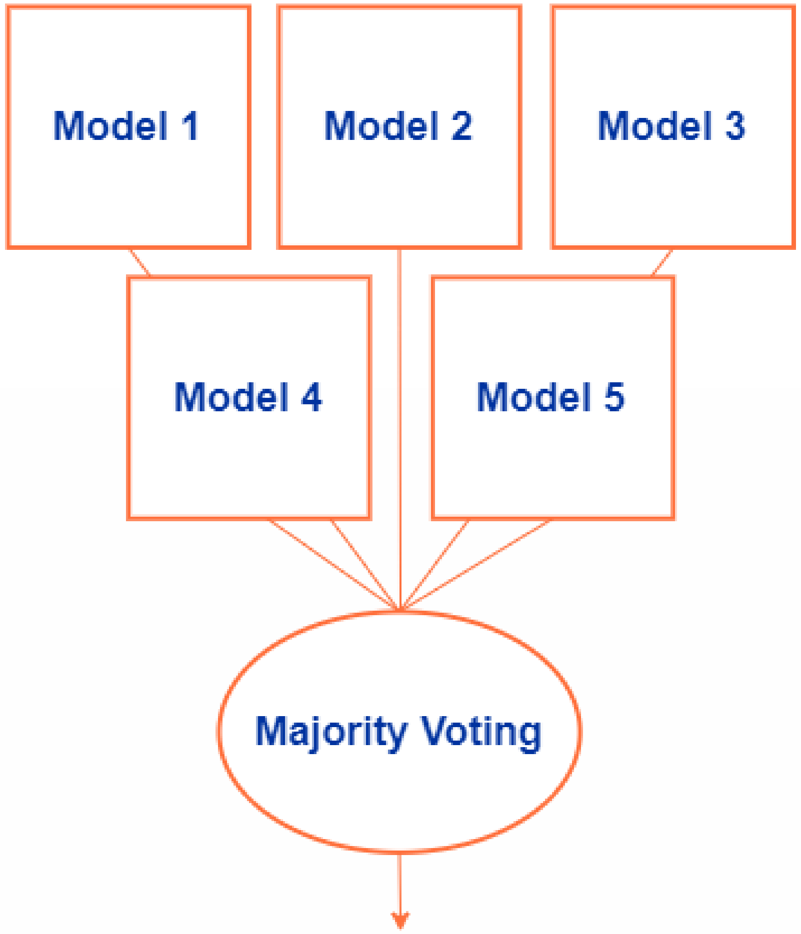
MODEL ANALYSIS

Accuracy of 81.4 % +/- 0.2 %

Prediction ->	Functional	Functional Needs repair	Non Functional
Functional	91.4%	1.6%	7.1%
Functional Needs repair	56.1%	29.9%	14.1%
Non Functional	21.9%	1%	77.1%

ENSEMBLE OF MODELS

1. **Gradient Boosting Trees**
2. **Gradient Boosting / Random Forest**
3. **Gradient Boosting + Balancer**



ENSEMBLE OF BOOSTING TREES

Gradient Boosting Trees

Learner	Accuracy
XGBTree 1	81.34 +/- 0.40 %
XGBTree 2	81.29 +/- 0.38 %
XGBTree 3	81.34 +/- 0.44 %
XGBTree 4	81.30 +/- 0.42 %
XGBTree 5	81.36 +/- 0.40 %
Ensemble	**81.35 +/- 0.38 %**

Gradient Boosting & Random Forest

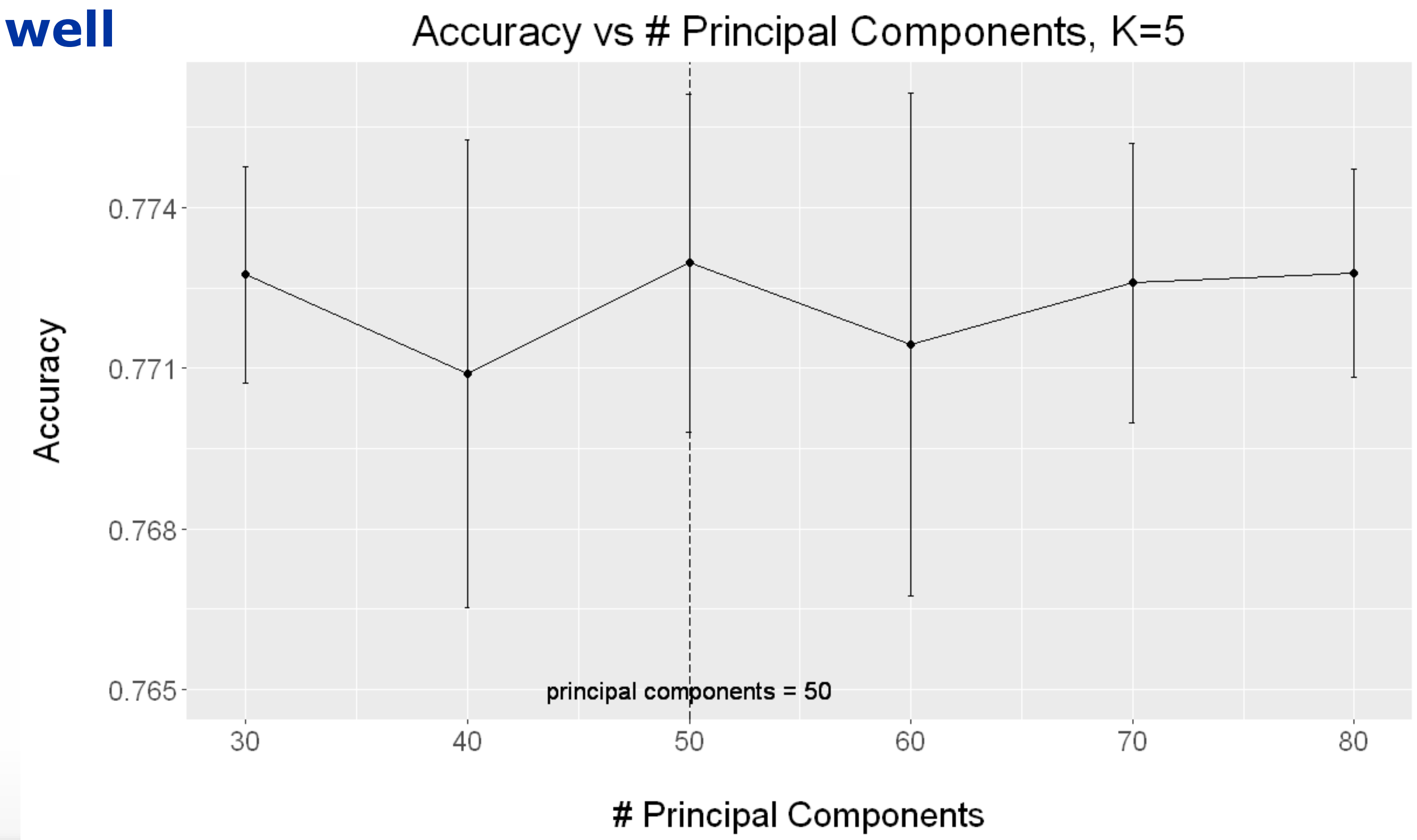
Learner	Accuracy
XGBTree 1	81.34 +/- 0.40 %
XGBTree 2	81.29 +/- 0.38 %
RF 1	81.13 +/- 0.33 %
RF 2	81.15 +/- 0.37 %
Ensemble	**81.37 +/- 0.36 %**

Gradient Boosting + Balanced Tree

Learner	Accuracy
XGBTree 1	81.32 +/- 0.34 %
XGBTree 2	81.38 +/- 0.46 %
XGBTree 3	81.27 +/- 0.46 %
XGBTree 4	81.31 +/- 0.40 %
XGBTree (Bal.)	74.73 +/- 1.04 %
Ensemble	**81.37 +/- 0.43 %**

KNN WITH PCA

- **KNN simple model, performed well**
- **Curse of high dimensionality**
- **PCA**



CONCLUSION

- **Extensive Feature Analysis, a lot of redundant data**
- **Tree based models performed best but no free lunch.**
- **Class imbalance had a big effect, bad performance on *functional needs repair* class**
- **Ensemble as final model**

Woohoo! We processed your submission!

Your score for this submission is:

0.8180