Statistics and Data Analysis first project

Maria Lavoura (up201908426)
Nuno Gomes (up199300242)

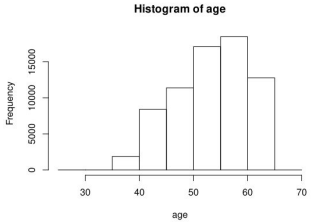
Masters in Data Science

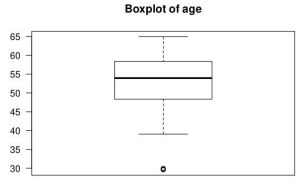
U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

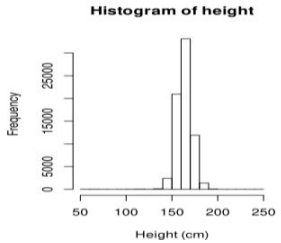
Dataset

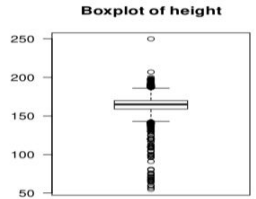
Description	Input type	Name	Type
Age	Objective	age	int (days, converted to years)
Gender	Objective	gender	1: women 2: men
Height	Objective	height	int (cm, converted to m)
Weight	Objective	weight	float (kg)
Systolic blood pressure	Examination	aphi	int
Diastolic blood pressure	Examination	aplo	int
Cholesterol	Examination	choles	1: normal 2: above normal 3: well above normal
Glucose	Examination	gluc	1: normal 2: above normal 3: well above normal
Smoking	Subjective	smoke	binary
Alcohol intake	Subjective	alco	binary
Physical activity	Subjective	active	binary
Cardiovascular disease	Target	cardio	binary

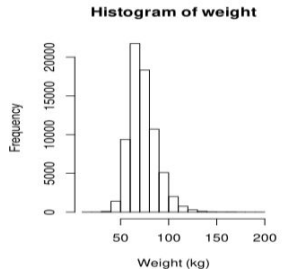
Quantitative variables: distributions and outliers

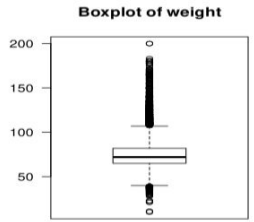


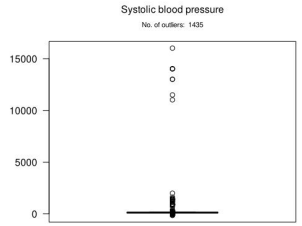


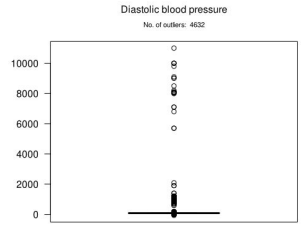


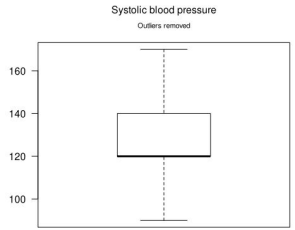


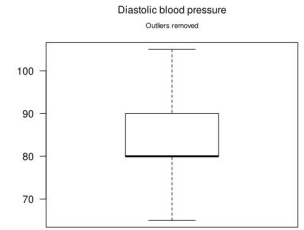




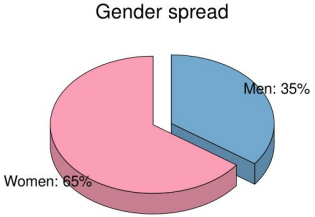


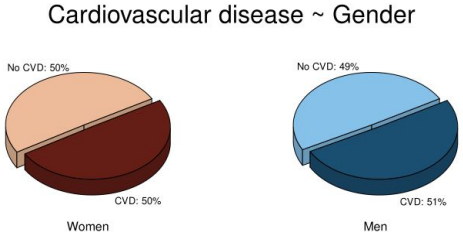


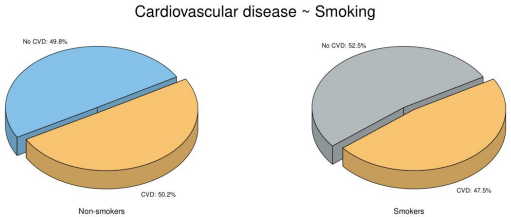


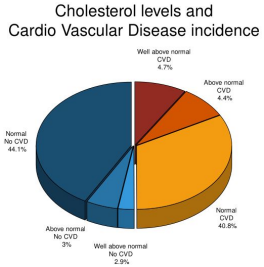


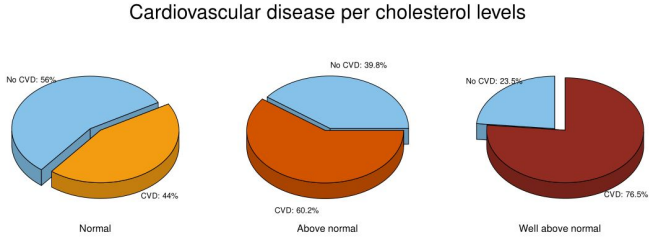
Qualitative variables: distributions

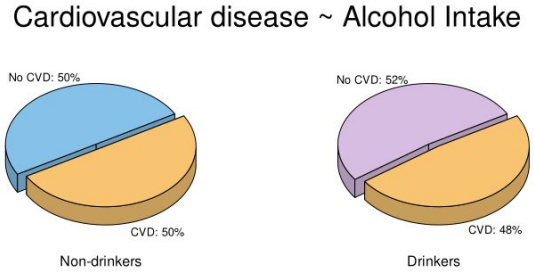


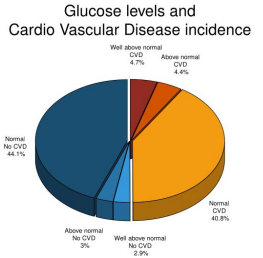


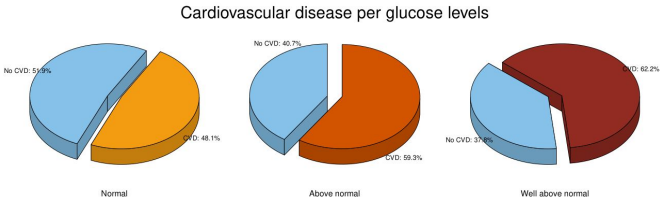


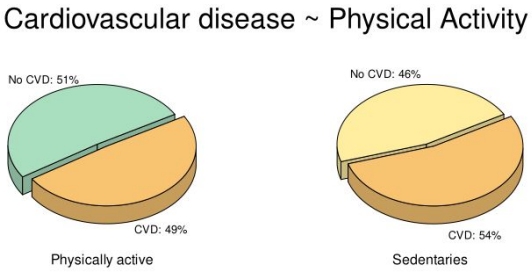




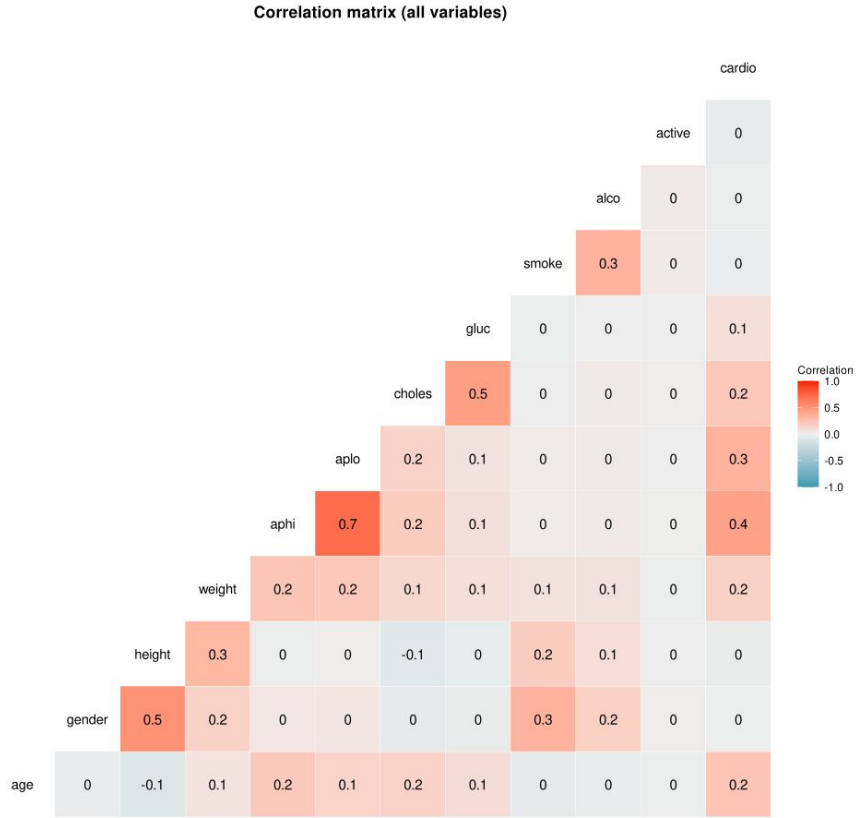


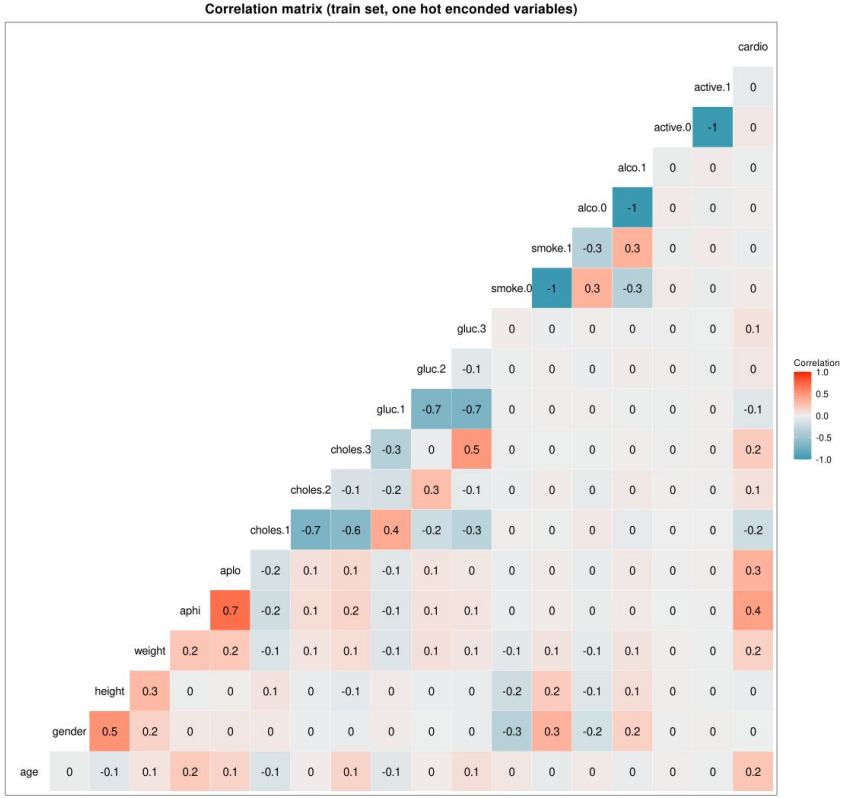






Correlations





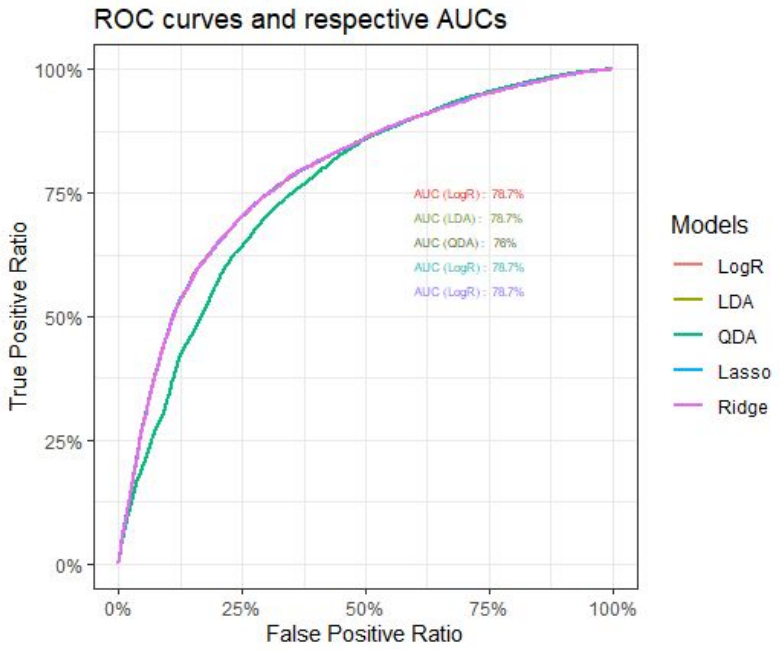
Final dataset - Linear Regression

- Statistical significant variables
- Complete model *vs* Removing **gender** and **height**

Modeling – Parameter tuning

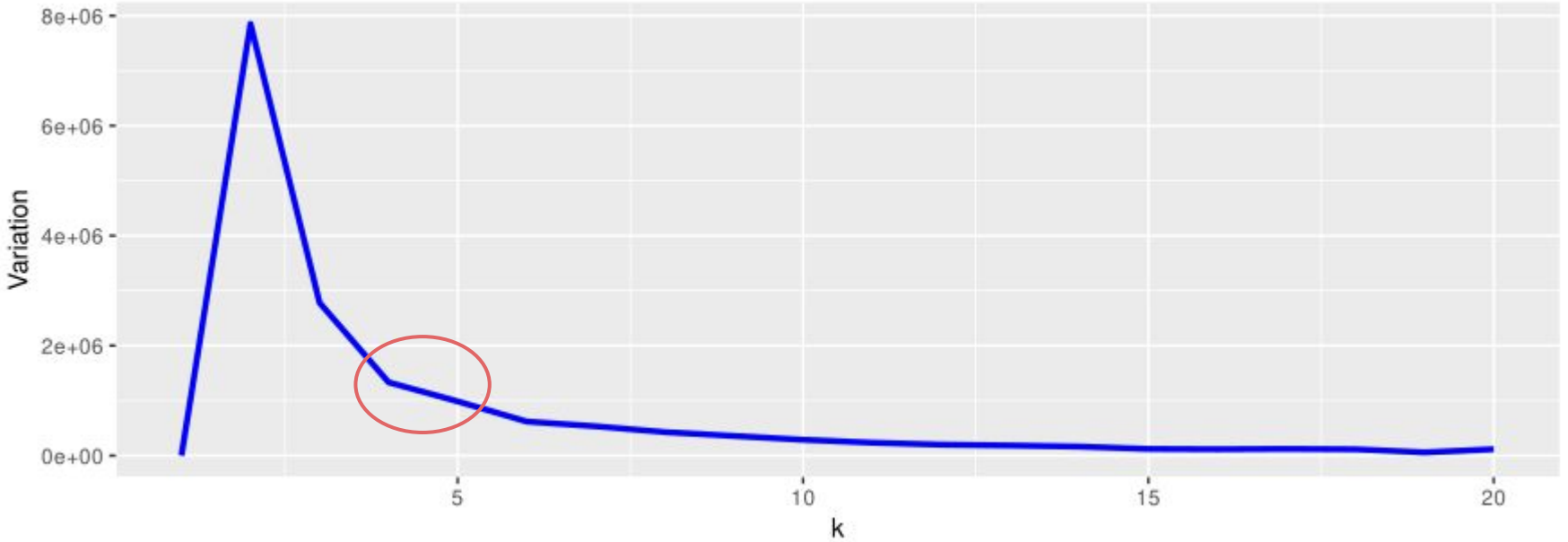
Model	Complete model		- gender and height	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.
LogR	0.705	0.703	0.704	0.702
LDA	0.723	0.724	0.723	0.726
QDA	0.687	0.684	0.688	0.687
Lasso	0.724	0.723	0.724	0.723
Ridge	0.721	0.723	0.722	0.725

ROC curves


ROC curves and respective AUCs
100%
75%
50%
25%
0%
True Positive Ratio
0%
25%
50%
75%
100%
False Positive Ratio
AUC (LogR) : 78.7%
AUC (LDA) : 78.7%
AUC (QDA) : 76%
AUC (LogR) : 78.7%
AUC (LogR) : 78.7%
Models
LogR
LDA
QDA
Lasso
Ridge

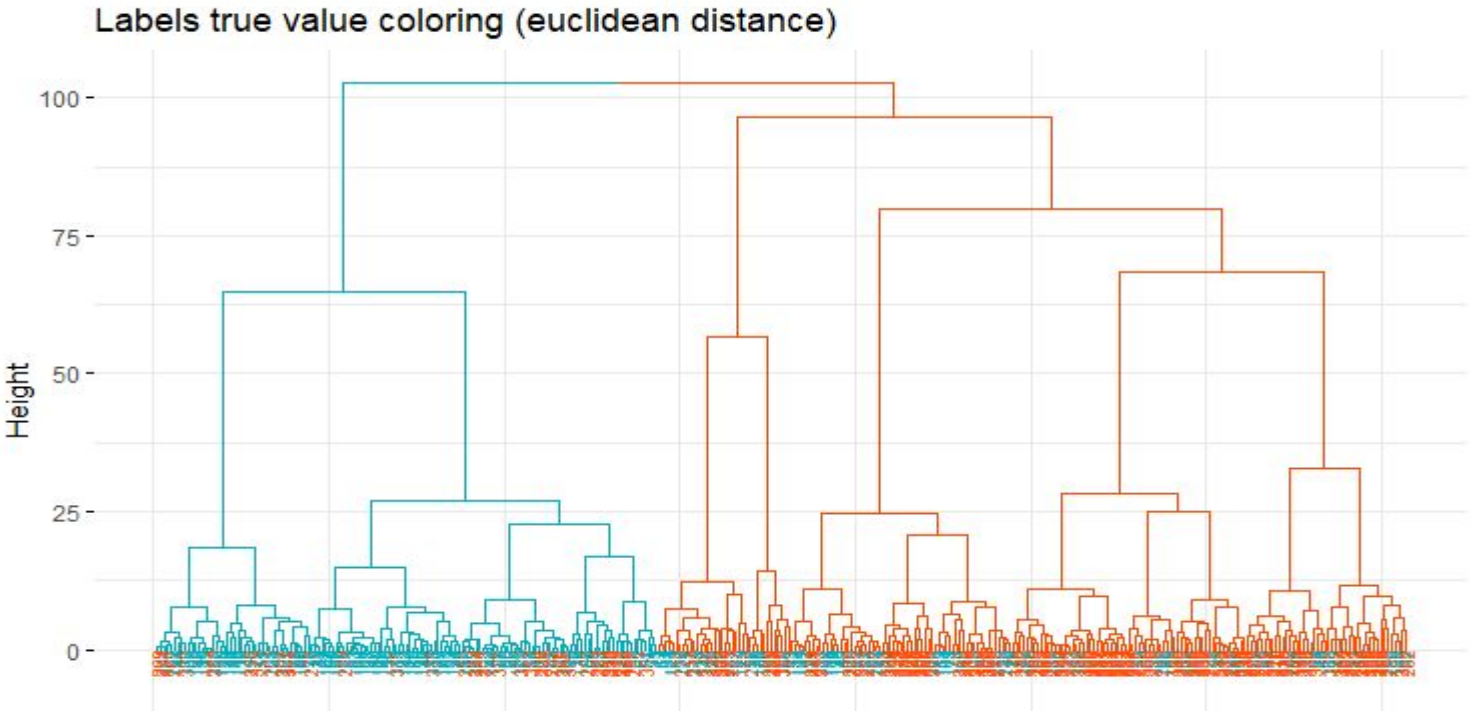
Modeling - Parameter tuning

K-Means Clustering



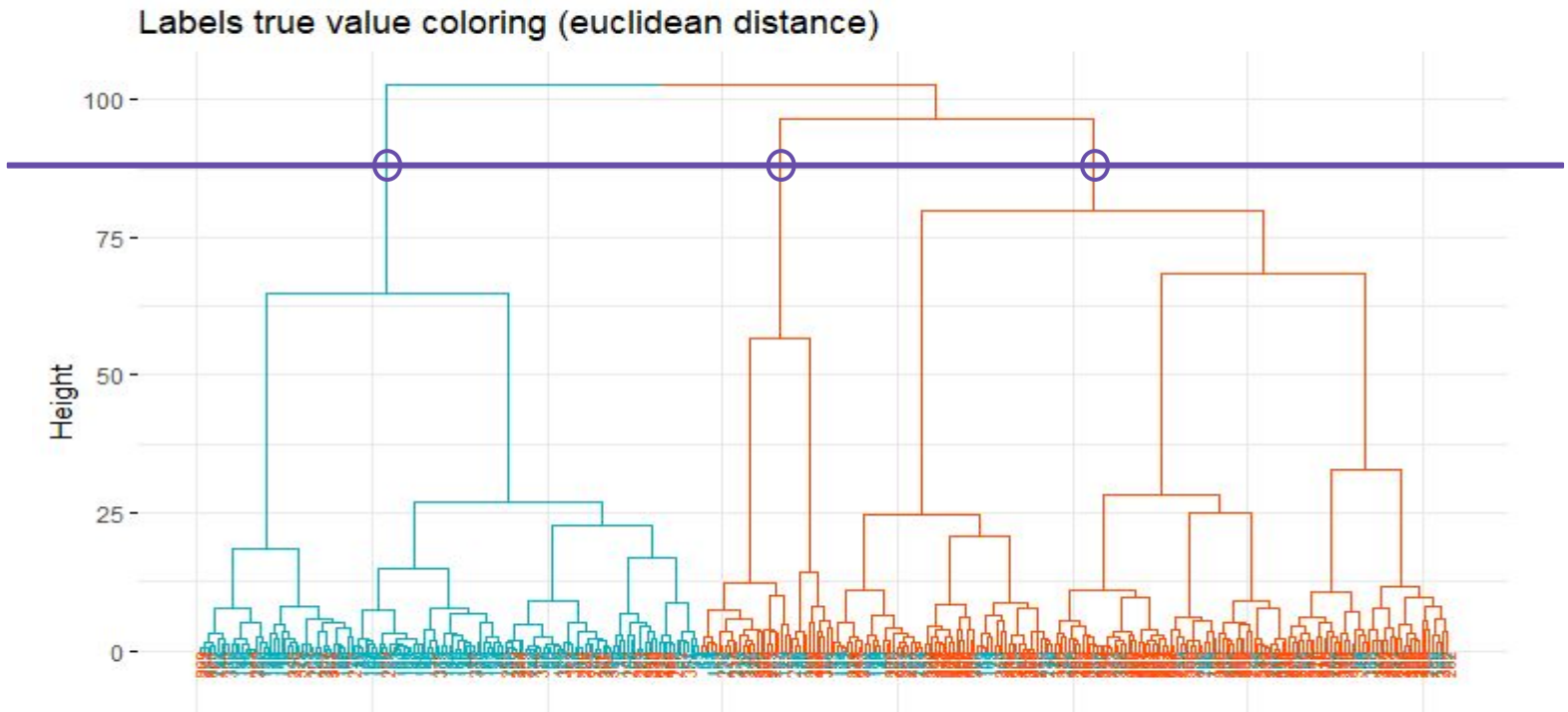
Modeling - Parameter tuning

Hierarchical Clustering - Euclidean distance



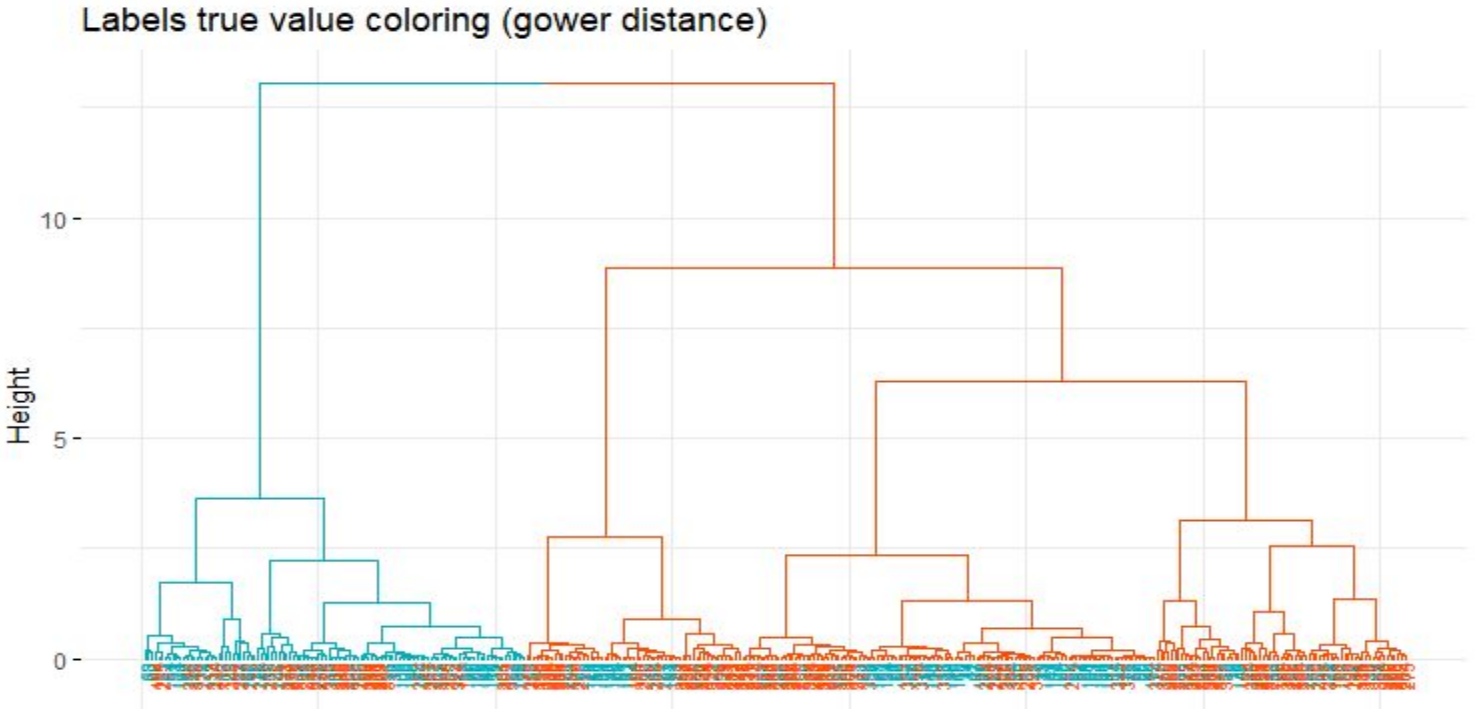
Modeling - Parameter tuning

Hierarchical Clustering - Euclidean distance



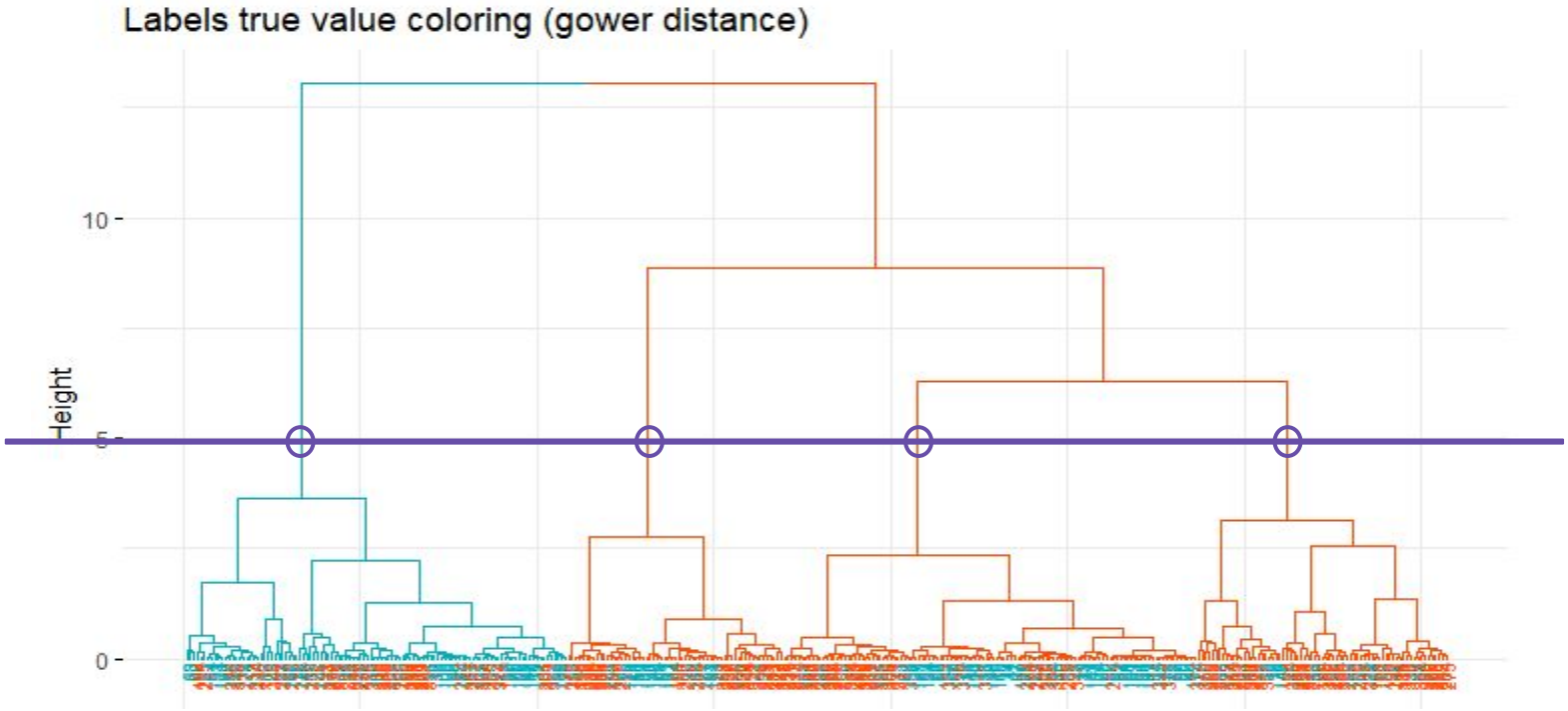
Modeling - Parameter tuning

Hierarchical Clustering - Gowers distance

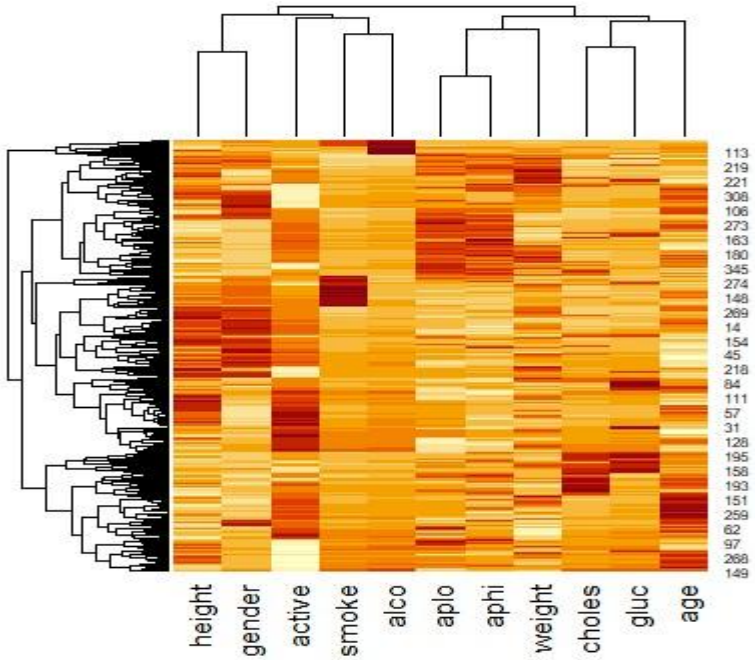


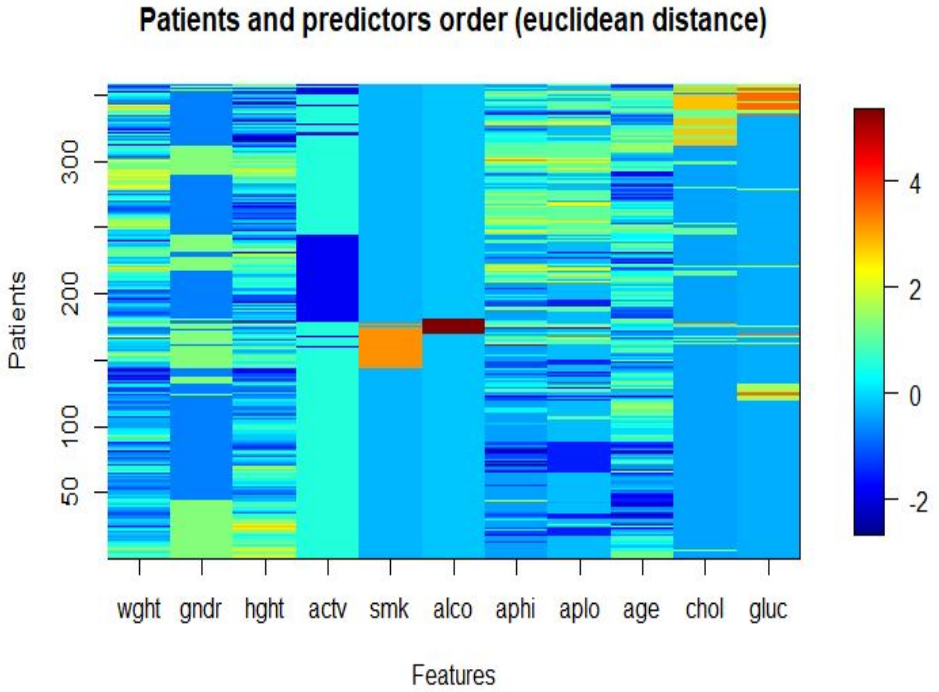
Modeling - Parameter tuning

Hierarchical Clustering - Gowers distance



Hierarchical Clustering





Conclusion

- Challegenig EDA
- Similar performance. QDA performed the worse and LDA the best
- Future work: Elastic-net comparison with Ridge and Lasso