Income mobility within Italian households: an exploratory study on children's income decile prediction

Group 3

Davide Bacigalupi,
Matteo Dalle Luche,
Gabriele Molé,
Camilla Pelosi

Project work for
Statistical Learning and Large Data, Module I
May 17th, 2023

Sant'Anna
School of Advanced Studies – Pisa

Introduction

- Goal: to understand the influence of family background on the economic position of children within households.

- This study analyzes the **Italian Income Registry**, data provided by ISTAT

- **Emergent patterns** from the data: **household composition**, **parental education**, and **disposable income** play a significant role in determining future economic opportunities.

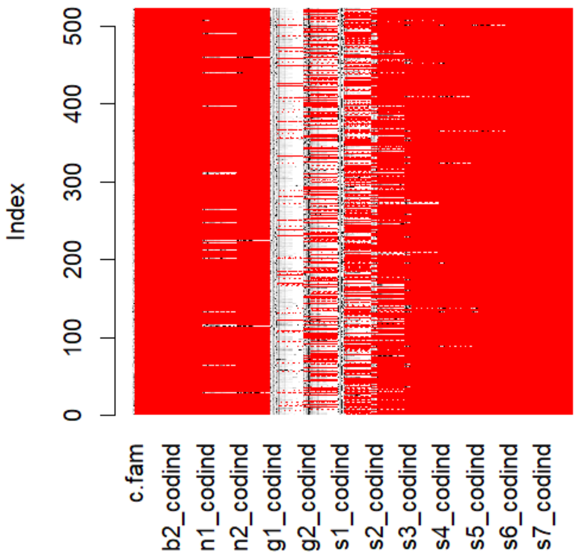
Methodology

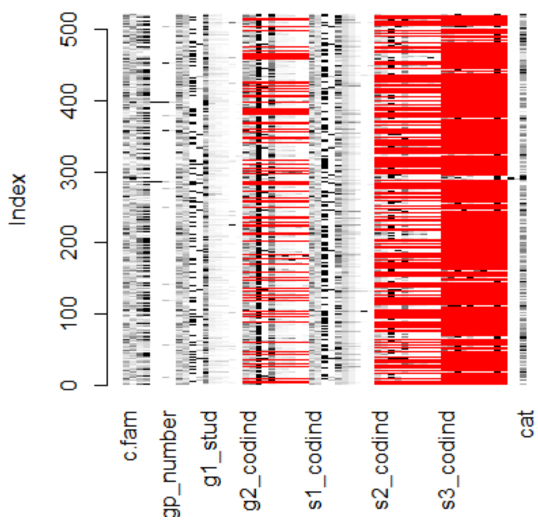
- Clustering analysis reveals distinct **grouping structures** in Italian households.
- Principal Component Analysis (PCA) identifies **variables that explain most of the** dataset's **variability**.
- PCA results guide the **selection of predictors** for **classifying income deciles** of children within households.
- Quadratic Discriminant Analysis (QDA) and K-Nearest Neighbors (KNN) classifiers are used to predict income deciles.
- Caveat: our **predicted classes** are **ordinal**, requiring alternative approach to standard classification methodologies
- Cross-validation is used to tune the models and evaluate their precision.

Data Transformation

- Individual-level data is transformed into a family-level dataset.
- Roles within the household are inferred from age and sex structure.
- The dataset excludes single-individual households, households with no significant age difference between members, and households with nine or more individuals.

Clustering





- Dealing with NAs
- Summarizing the information on great Grandparents, grandparents and sons (n>3)
- Imputation of 0 income to existing individual without income
- Average income for each category of the period 2015-2020.
- Education imputation based on age
- Subdivision in 6 types of families

Clustering

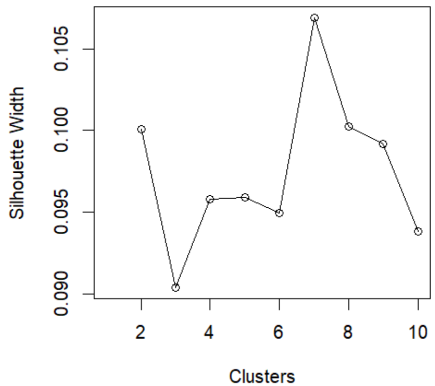
- Gower distance to deal with the mix of categorical and continuous variables

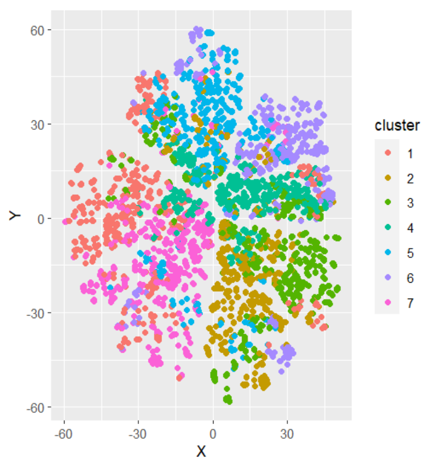
$$\text{Gowerdistance}(x_1, x_2) = 1 - \left(\frac{1}{p} \sum_{j=1}^{p} s_j(x_1, x_2) \right)$$

- K – Means and Hierarchical Clustering (complete linkage) for each type of family
- Average Silhouette and WSS to evaluate the performance
- Generally low values in Silhouette

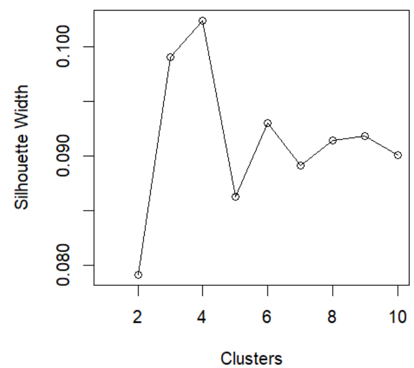
Clustering

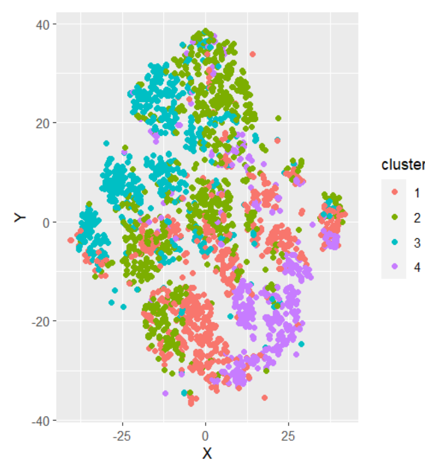
Single parent with only-child



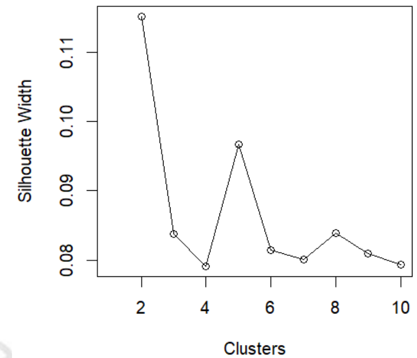


Single parent with two children





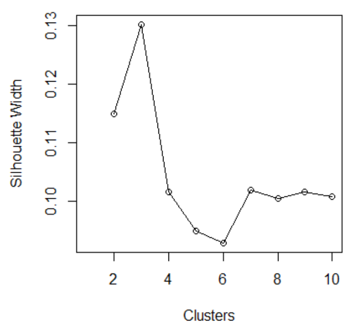
Single parent with three or more children

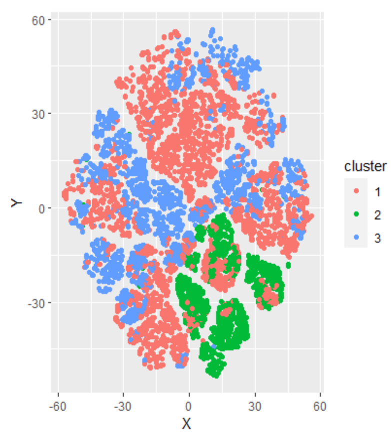


Sant'Anna
School of Advanced Studies – Pisa

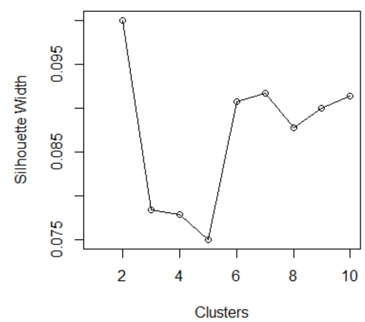
Clustering

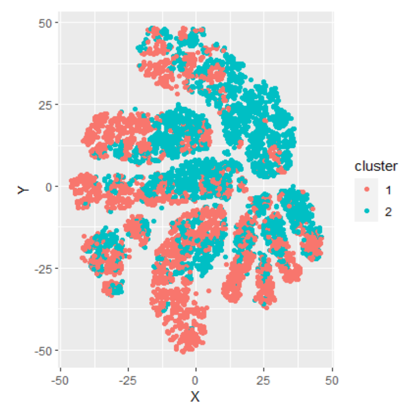
Two parents with only-child



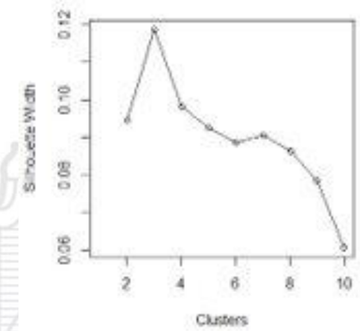


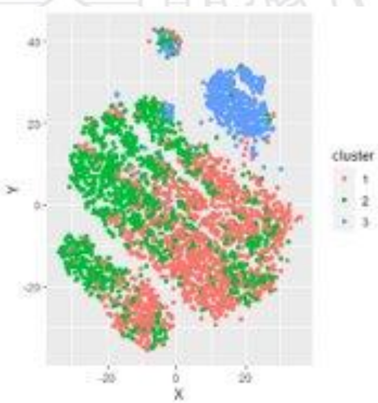
Two parents with two children





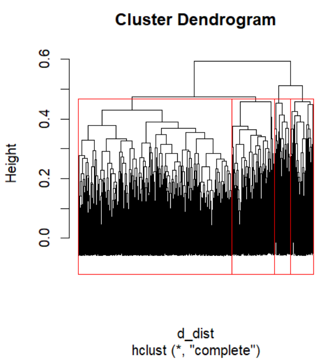
Two parents with three or more children

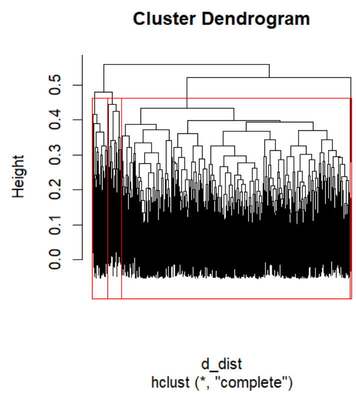


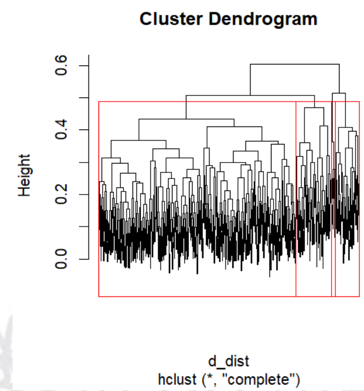


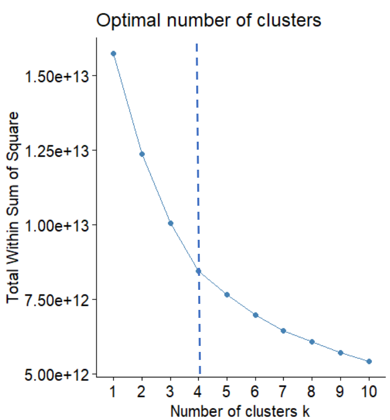
Clustering

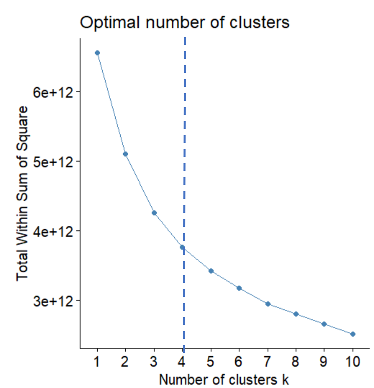
Single parent with only-child	Single parent with two children	Single parent with three or more children

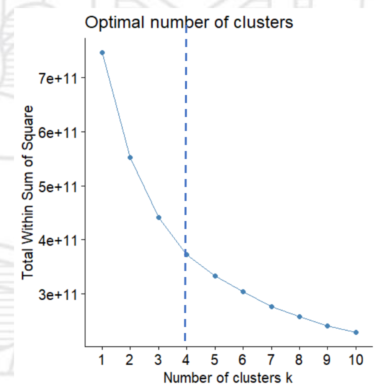








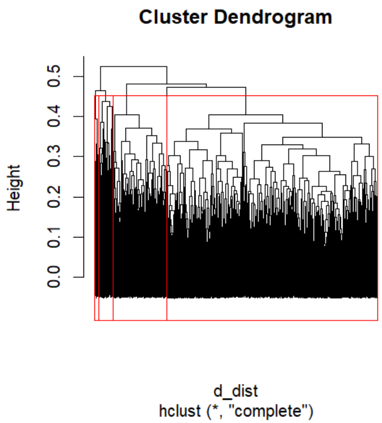


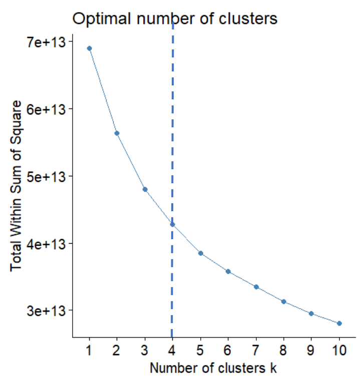


Sant'Anna
School of Advanced Studies – Pisa

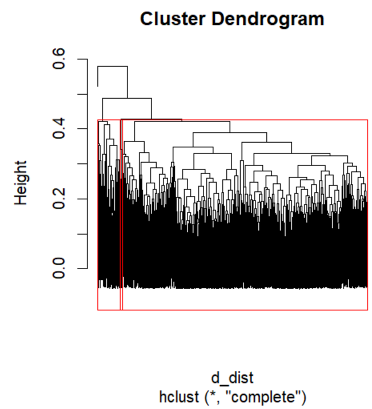
Clustering

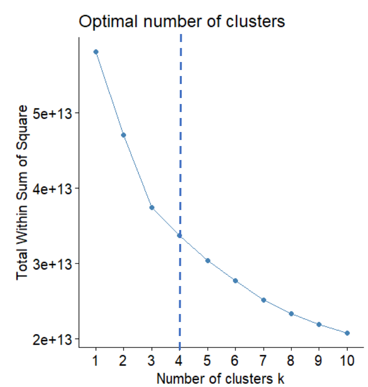
Two parents with only-child



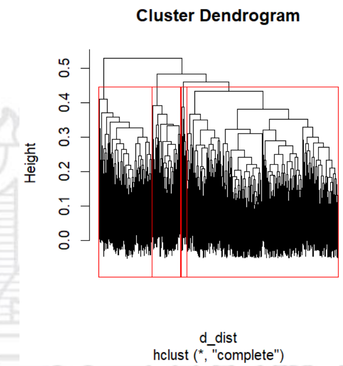


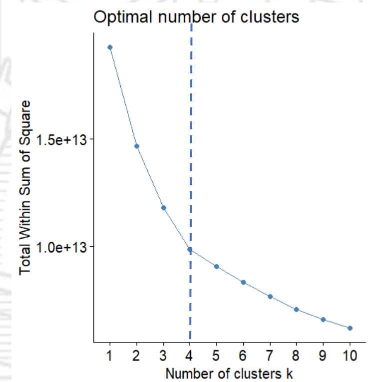
Two parents with two children





Two parents with three or more children





Classification outline

- The general idea behind classification is to **predict categorical responses** based on selected features.

- Our case: Predicting the income bracket of children based on selected variables.

Features selection: step 1

- ***zone***,
- ***any_nonni***, ***onlychild***, ***n_sons***,
- ***gYD***, (selecting only overall income, despite its breakdown)
- ***educ***, ***s_stud***,
- ***s_agediff***, ***s_cittad***, ***s_sex***

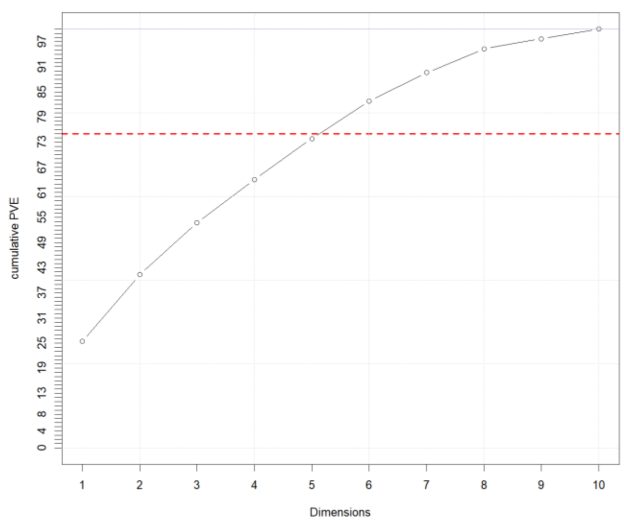
Features selection: PCA

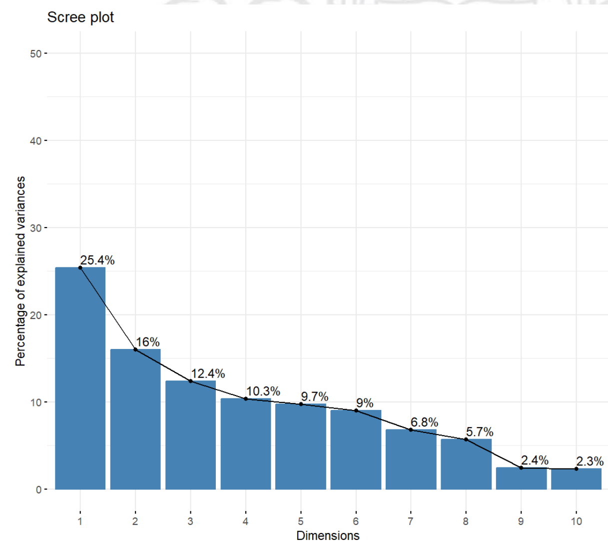
Objective:

- de-noise the data
- select the most relevant features for classification, carving out relevant variability

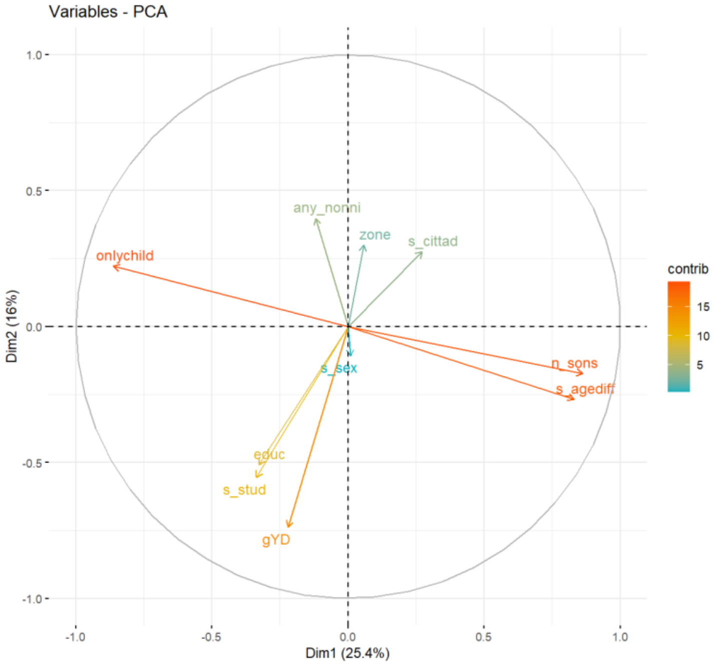
Fixed threshold: **80%** in CPVE

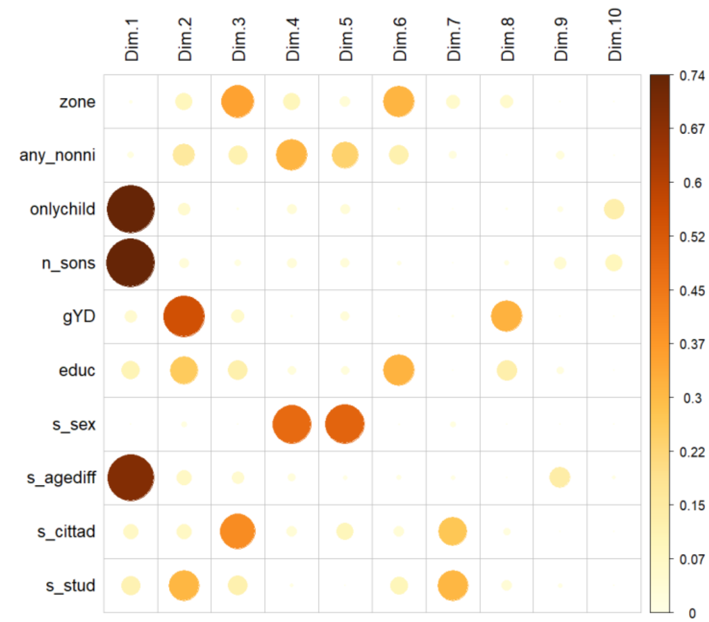
Elbow and CPVE threshold are reached on the **sixth principal component**





Features selection: PCA





Factor loadings return the most explanatory variables along the first two principal components:

- First principal component: explained by familiar composition (***onlychild***, ***n_sons***, ***agediff***)
- Second principal component: explained by economic and education background (***educ***, ***gYD***, ***s_stud***)

Classification: initial stages

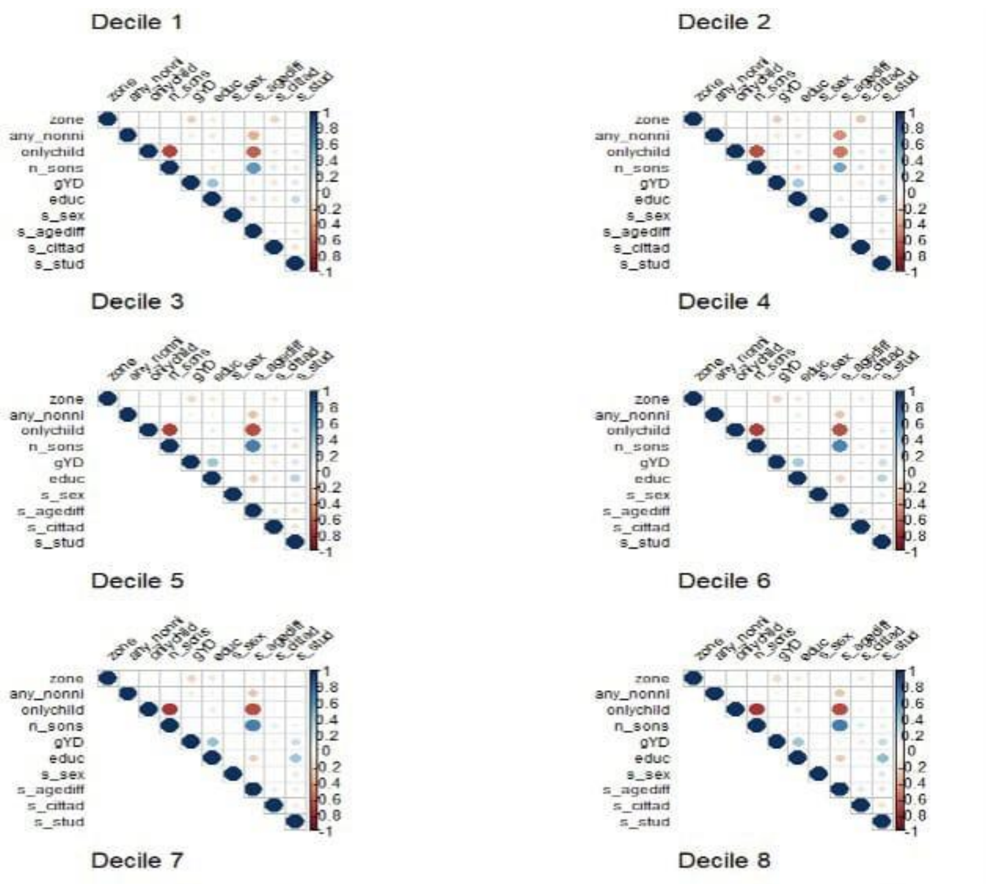
•Logistic Regression not suitable due to 10 ordinal classes.

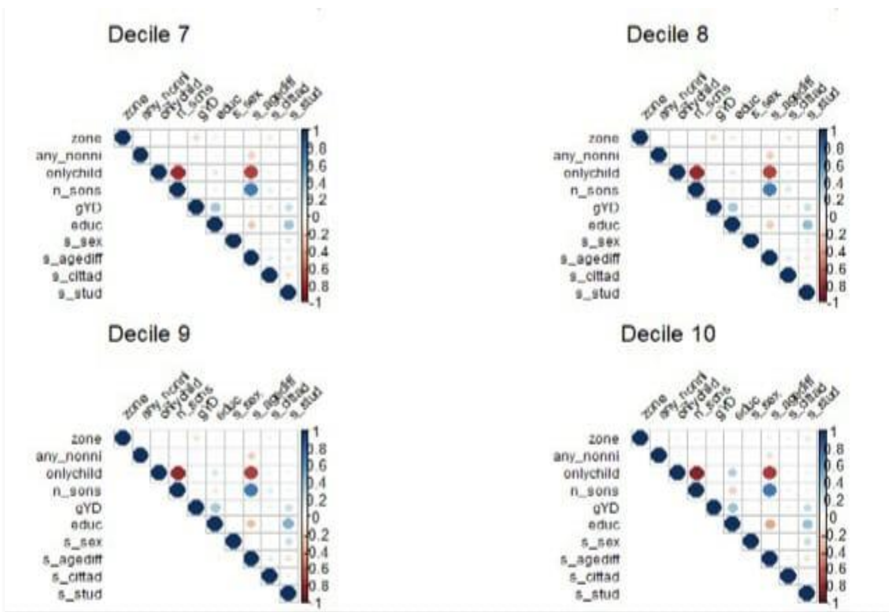
•Quadratic Discriminant Analysis (QDA) vs Linear Discriminant Analysis (LDA).

•Non-normal distribution of features, trended variances, different covariances across classes, large sample: we went for QDA.

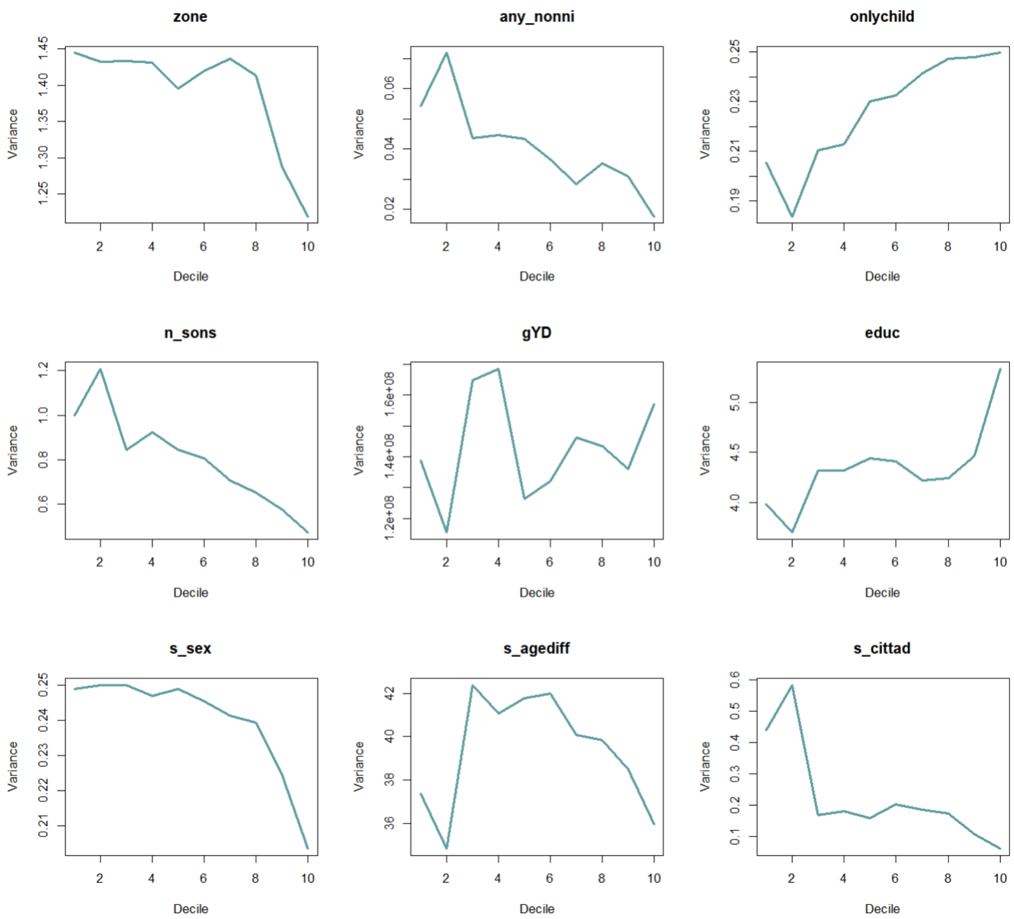
•K-Nearest Neighbors (KNN) algorithm also tested with low k values (1-10)

Variance-covariance structure

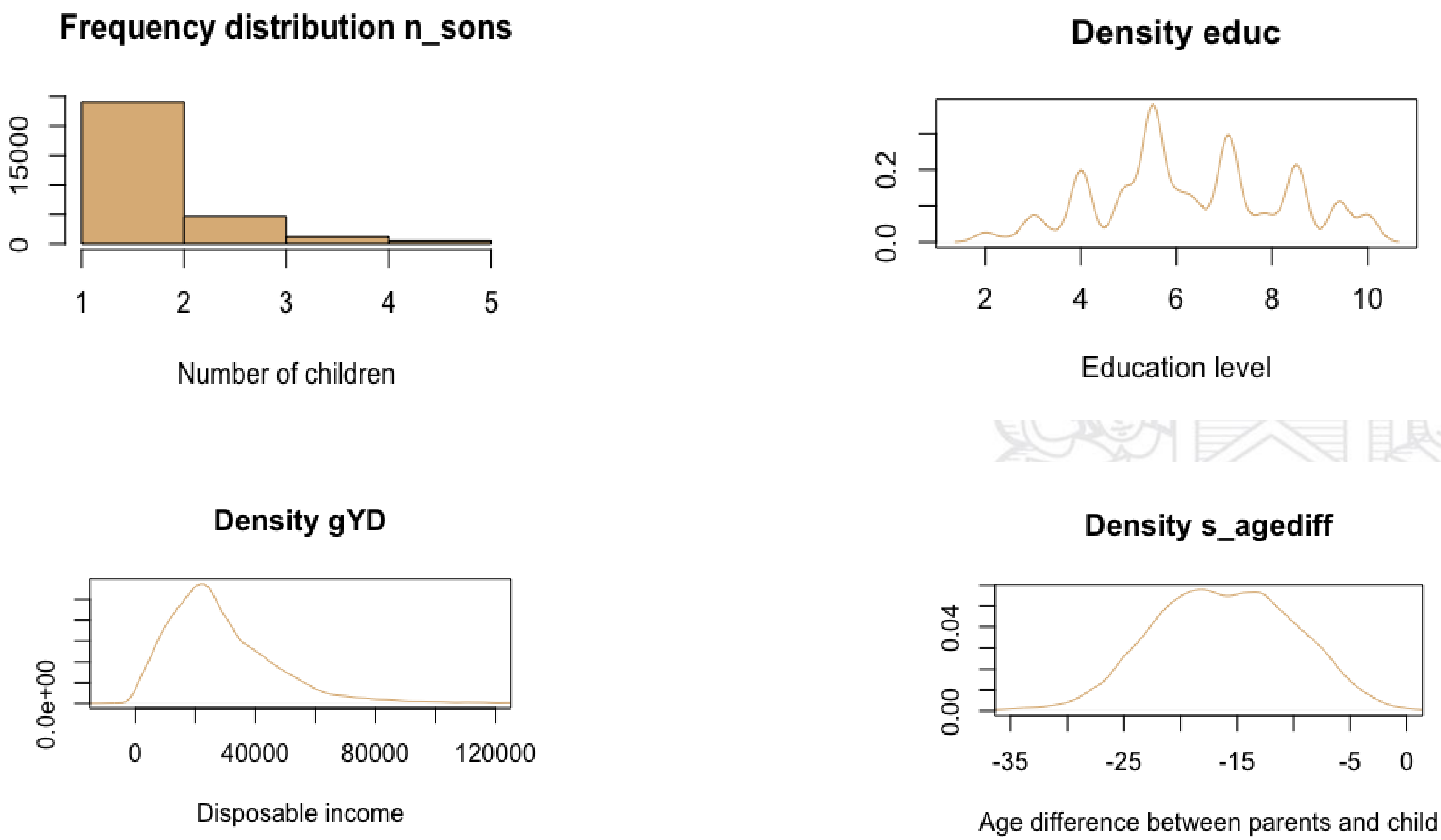




Variance for each class



Non-normal distribution of features



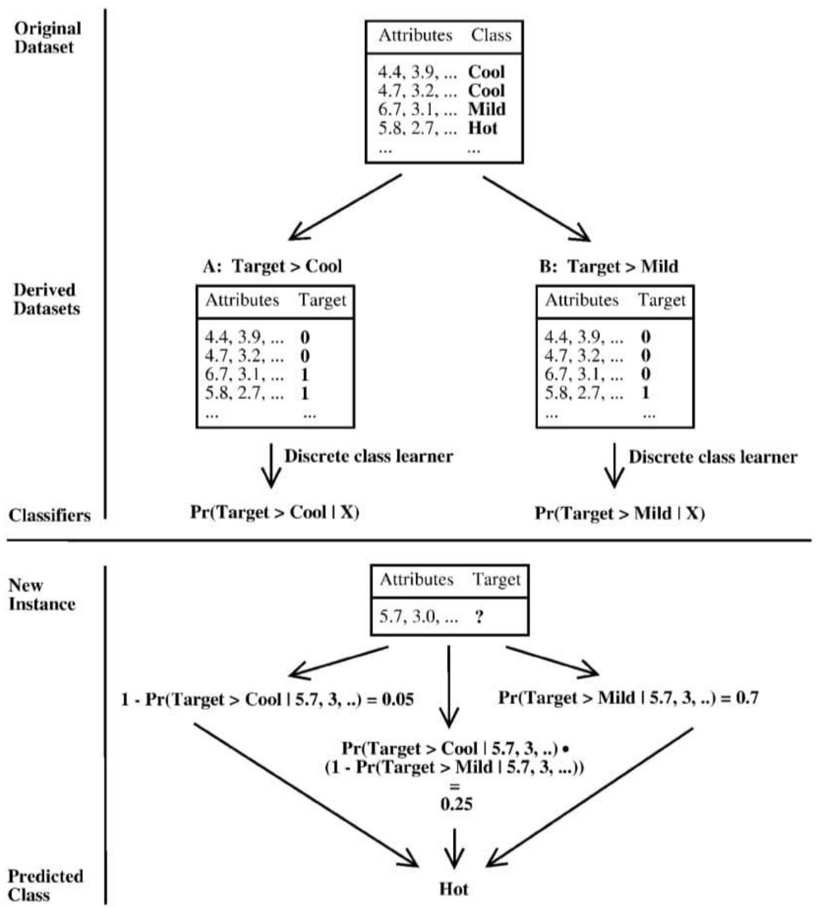
Dealing with Ordinal Classes

- Observations: Ten classes not independent, algorithms tend to classify extreme values.
- Objective: Minimize distance between predicted and observed classes considering interdependence.
- Conducted extensive literature review for classification methods considering ordinal labels.
- Goal: Minimize distance between predicted and observed classes for accurate classification.

Frank and Hall (2001) Approach

- Constructed binary datasets (above/below income decile) for K-1 datasets.
- Applied Frank and Hall (2001) method to all proposed classifiers, including logistic regression.
- Commonly used metrics for validating classification algorithms with ordinal outcomes:
    - Accuracy
    - Mean Squared Error (MSE)
    - Mean Absolute Error (MAE)

Frank and Hall (2001) Approach



Quantile		Q> 1	Q> 2	Q> 3	Q> 4	Q> 5	Q> 6	Q> 7	Q> 8	Q> 9
9		1	1	1	1	1	1	1	1	0
6		1	1	1	1	1	0	0	0	0
2		1	0	0	0	0	0	0	0	0
7		1	1	1	1	1	1	0	0	0
4		1	1	1	0	0	0	0	0	0
2	→	1	0	0	0	0	0	0	0	0
9		1	1	1	1	1	1	1	1	0
1		0	0	0	0	0	0	0	0	0
4		1	1	1	0	0	0	0	0	0
2		1	0	0	0	0	0	0	0	0

Parameter Optimisation

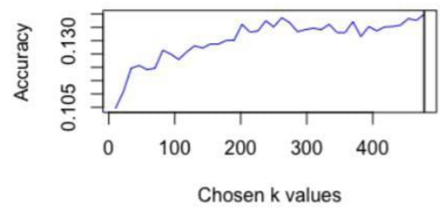
•Optimization focused on k value for KNN classifier.

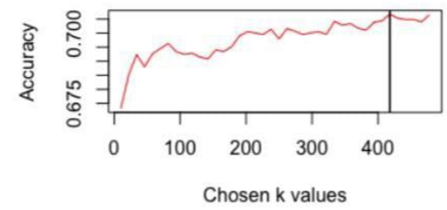
•Monotonically increasing metrics (accuracy, MSE, and MAE) in k range 1-10.

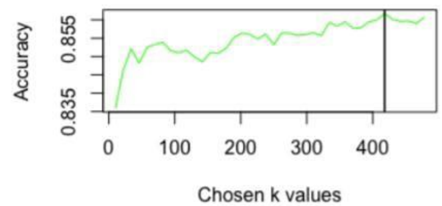
•Determined optimal k value using a loop and evaluated performance metrics (k=418).

•Peak values for (1-MSE%) and (1-MAE%) with different k values.

•Selected k = 418 to balance the two metrics.







Sant'Anna
School of Advanced Studies – Pisa

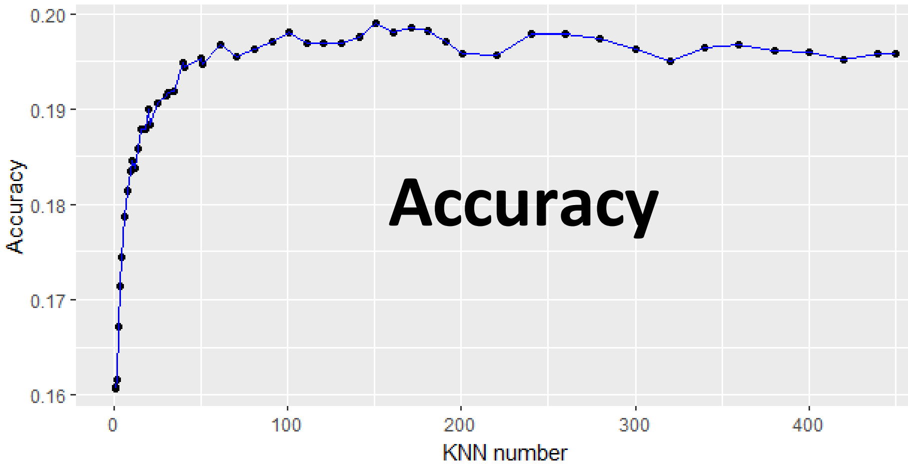
Classification results

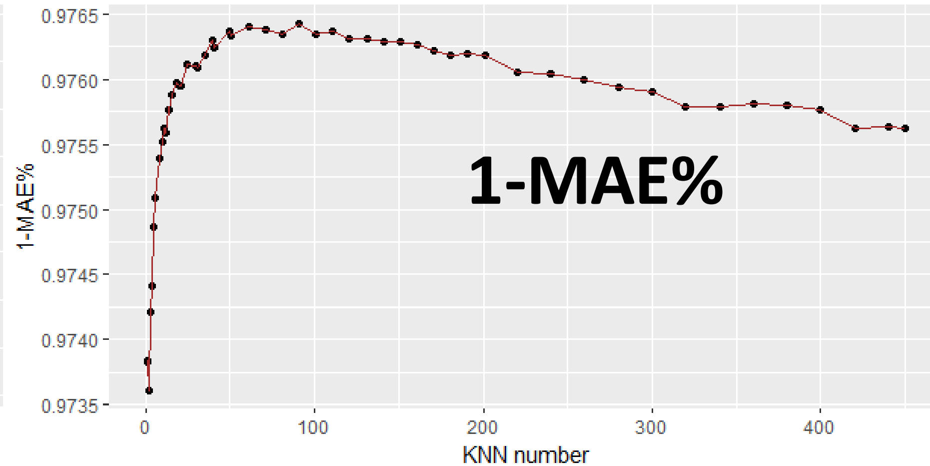
	Accuracy	(1-MAE%)	(1-MSE%)
QDA (Ordinal)	0.173	0.720	0.865
Logistic (Ordinal)	0.191	0.751	0.891
KNN418 (Ordinal)	0.202	0.758	0.895
QDA (Standard)	0.183	0.728	0.871
KNN418 (Standard)	0.176	0.730	0.875

Cross-validation

- Evaluation of classification results using 5-fold cross validation
- No use of LOOCV because of too many observations
- Comparison of "standard" approach with Frank and Hall (2001)
- Very low values of Accuracy overall (max. 20%)
- Nevertheless huge amount of time required for computations!

Cross-validation on "standard approach" classifiers





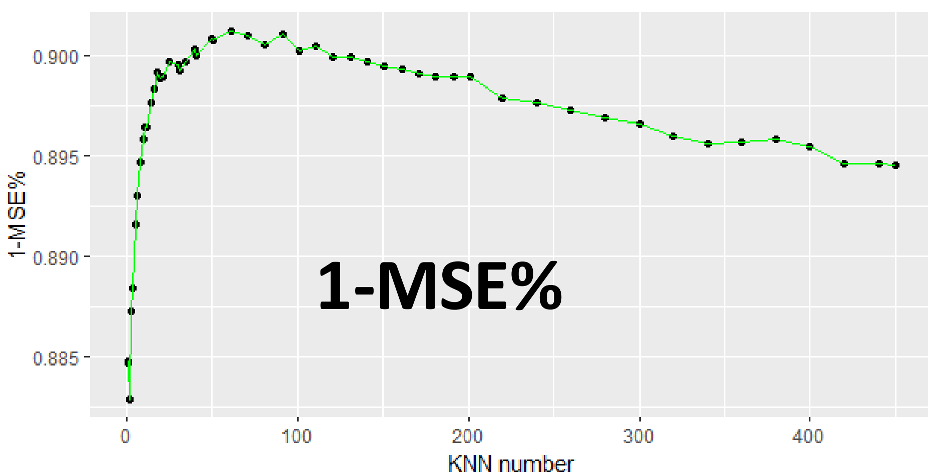
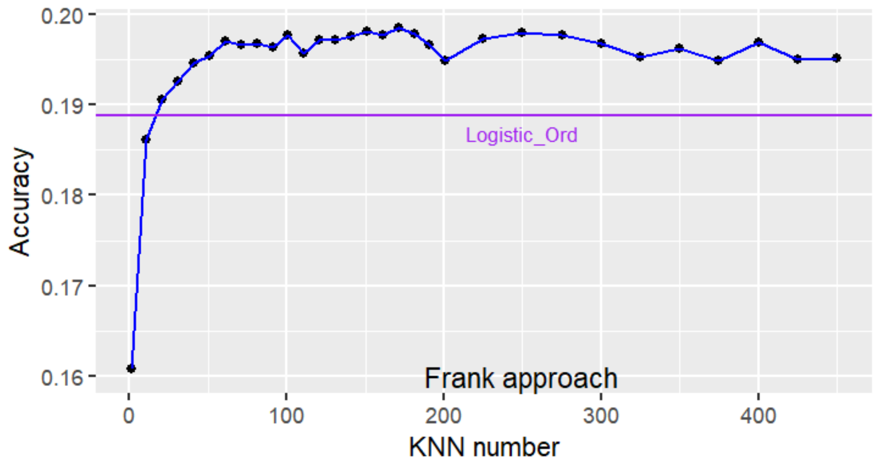
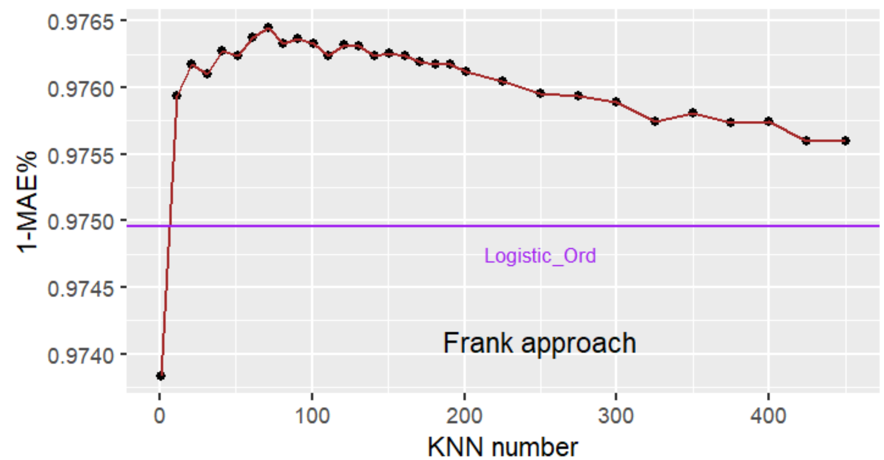


Table 3 QDA, standard approach

	Accuracy	(1-MAE%)	(1-MSE%)
QDA	0.1021570	0.9605275	0.7698306

Cross-validation on Frank (2001) classifiers





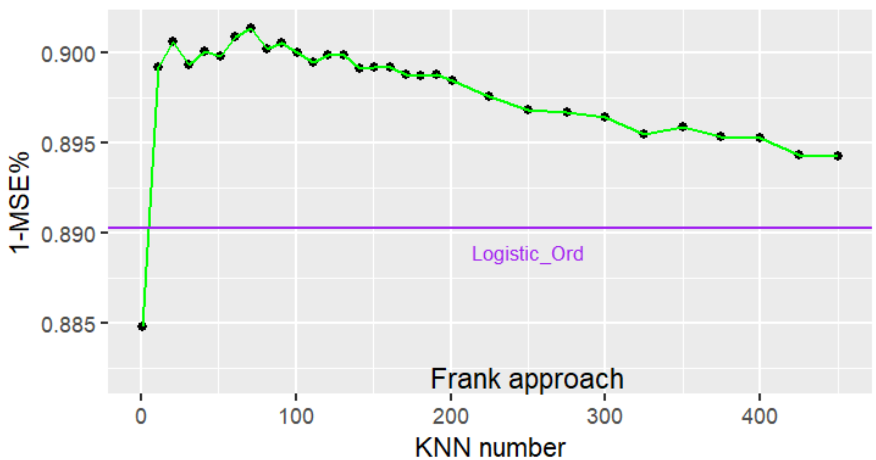
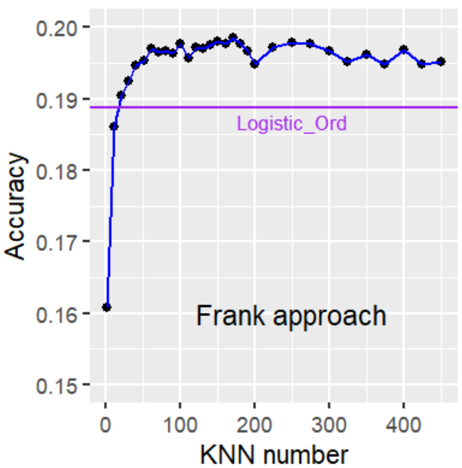


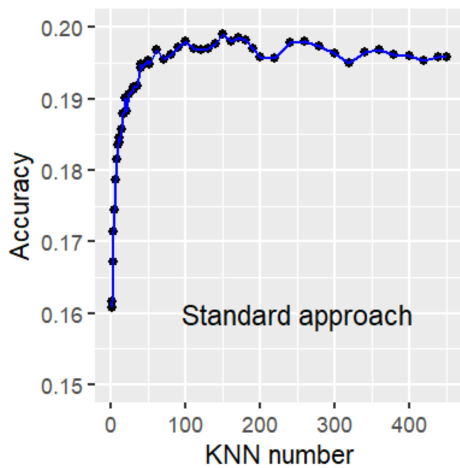
Table 7 QDA and Logistic, Frank (2001)

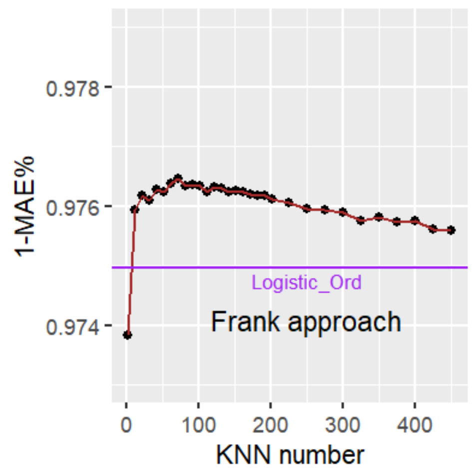
	Accuracy	(1-MAE%)	(1-MSE%)
QDA_Ord	0.1748672	0.9719164	0.8644483
Logistic_Ord	0.1888079	0.9749563	0.8902296

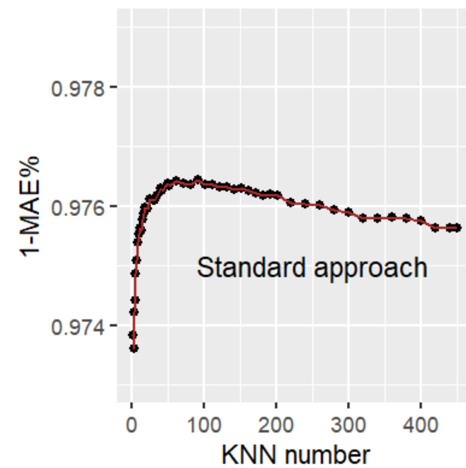
Sant'Anna
School of Advanced Studies – Pisa

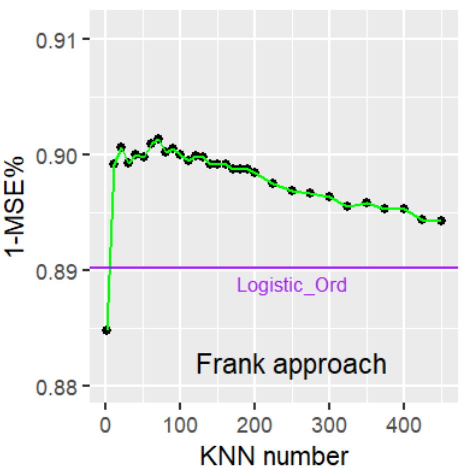
Overall comparison and final results











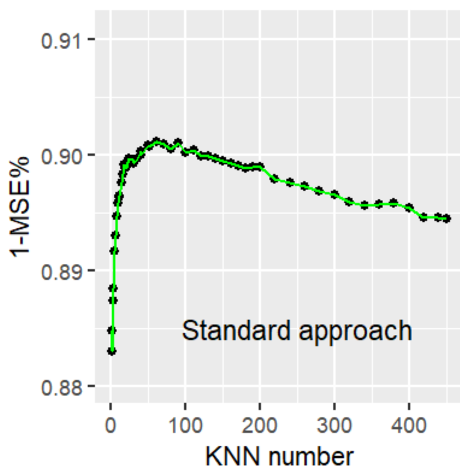


Table 11 Overall results

Classifier	Accuracy	(1-MAE%)	(1-MSE%)
QDA_norm	0.1021570	0.9605275	0.7698306
KNN_norm, k=151	0.1990245	0.9762911	0.8994522
QDA_Ord	0.1748672	0.9719164	0.8644483
Logistic_Ord	0.1888079	0.9749563	0.8902296
KNN_ord, k=71	0.1965528	0.9764499	0.9013756

Conclusion and future prospects

- Literature on ordinal classes is sparse and developing: beyond regression approaches to look at income mobility
- Capital income data are not available and not tracked, unlike in other EU countries: further investigate functional inequality
- Wealth (and estate) is measured only through surveys, but is a very relevant predictor to account for familiar background
- Possible policy implementation: pre-distributive policies should break down such patterns, to "harm" our performance

References

- Carrizosa, E., Restrepo, M. G., & Morales, D. R. (2021). On clustering categories of categorical predictors in generalized linear models. Expert Systems with Applications, 182, 115245.
- Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. Biometrics, 27(4), 857–871.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer.
- Preud'Homme, G., Duarte, K., Dalleau, K., Lacomblez, C., Bresso, E., Smaïl-Tabbone, M., ... & Girerd, N. (2021). Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. Scientific reports, 11(1), 1-14.31
- Frank, E., Hall, M. (2001). A Simple Approach to Ordinal Classification. In: De Raedt, L., Flach, P. (eds) Machine Learning: ECML 2001. ECML 2001. Lecture Notes in Computer Science(), vol 2167. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44795-4_13
- Cardoso, Jaime & Sousa, Ricardo. (2011). Measuring the Performance of Ordinal Classification.. IJPRAI. 25. 1173-1195. 10.1142/S0218001411009093.