Income Prediction from the Adult Dataset

DAVID MORÁN

JOEL CANTERO

XAVIER TIMONEDA

19/06/2019

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Table of Contents

Description of the problem

- **Main objectives:**
  - Full **multivariate analysis**
  - Build a **classifier** that **discriminates** people based on its **income**
    - Two categories: **more** and **less** than **50k dollars** per year
- **Very interesting dataset**:
  - Very **interpretable**
  - **Large number** of both **features** and **instances**
  - Mixture of **numerical** and **categorical** variables
  - Great **opportunity** to apply all the **techniques** learnt in **MVA**!
- It seems reasonable that the income is quite **related** to those **socioeconomical** features
- **Good models** are expected!

Available Data

- **Numerical variables**
  - Age, Fnlwgt, EducationNum, CapitalGain, CapitalLoss, WorkingHours

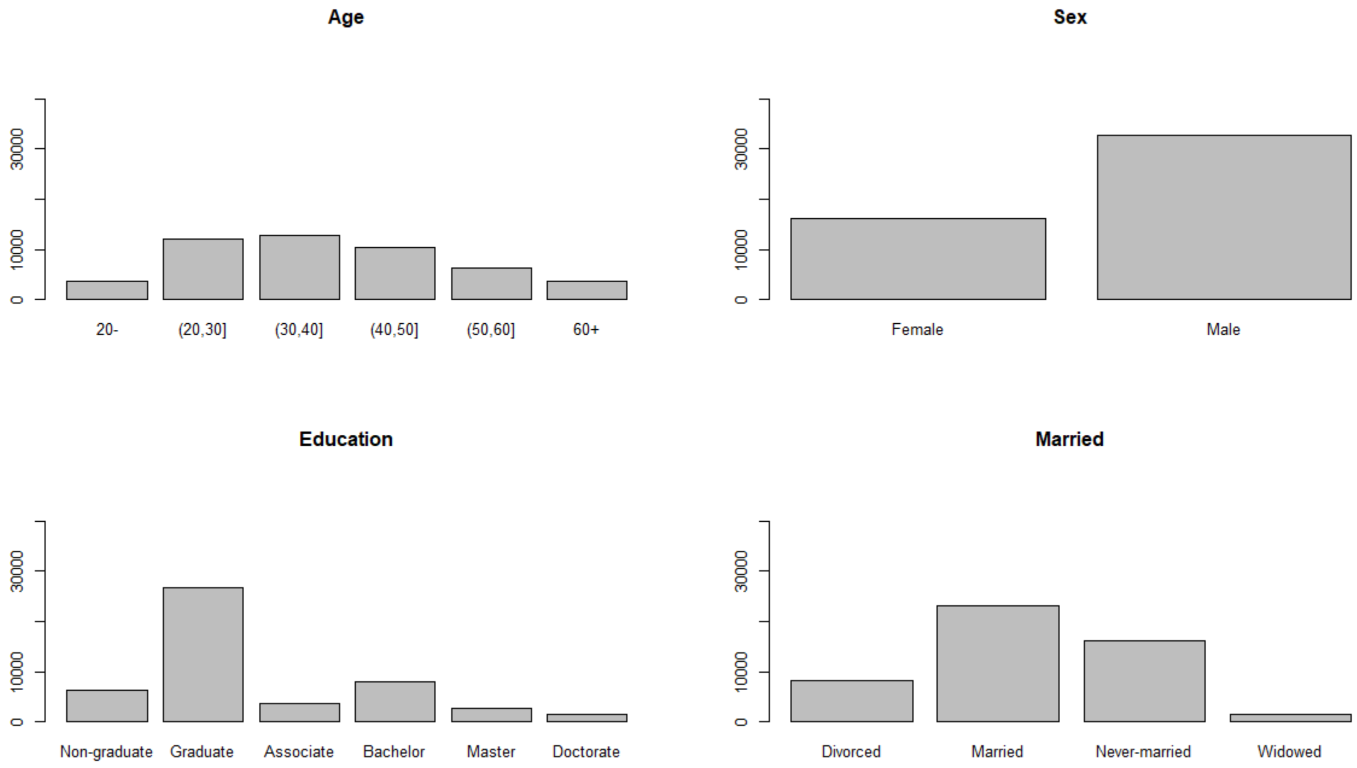
- **Categorical variables**
  - Work, Education, MaritalStatus, Occupation, Relationship, Race, Sex, NativeCountry

- Some of the variables are uninformative or redundant
- Lots of modalities for the categorical variables
- 7.4% of instances have missing values!

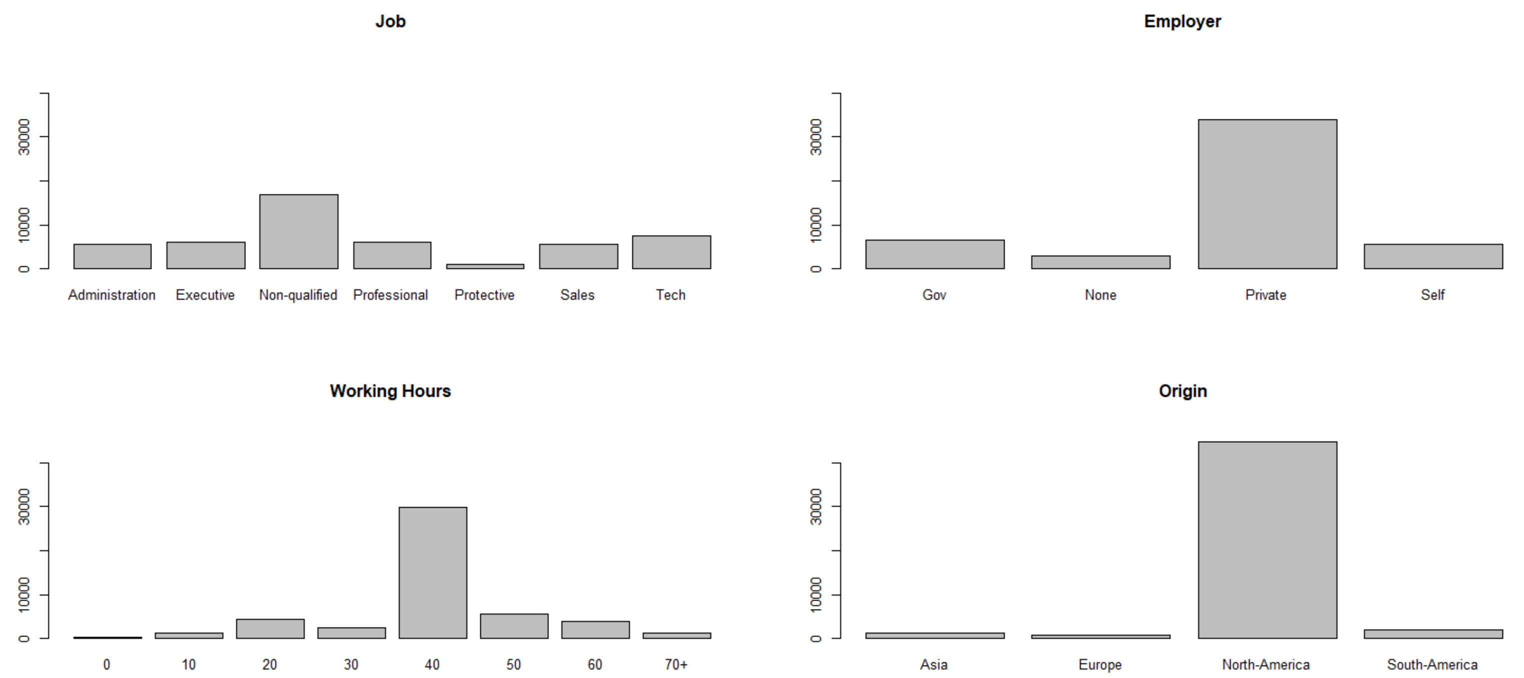
Pre-process

- Redundant and uninformative variables removed
  - Fnlwgt, EducationNum, Relationship
- Numerical variables discretized
  - Age, WorkingHours, Capital (Gain + Loss)
- Categorical variables simplified
  - Work, Education, MaritalStatus, Occupation, Race, Sex, NativeCountry
- Some missing values have been explained
- The rest of missing values have been imputated

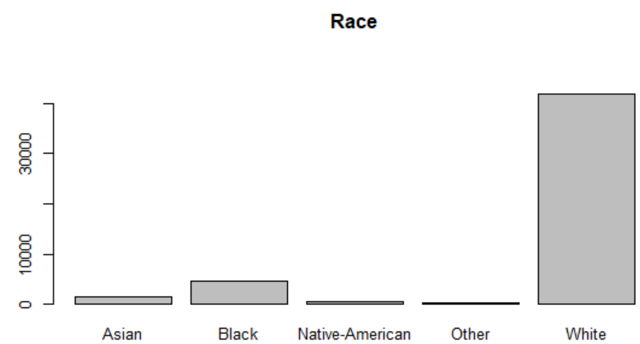
Data summary

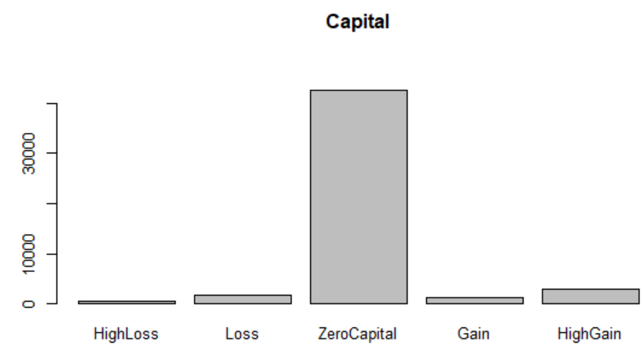


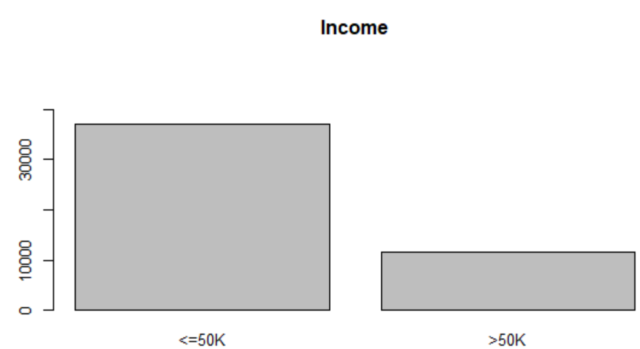
Data summary



Data summary



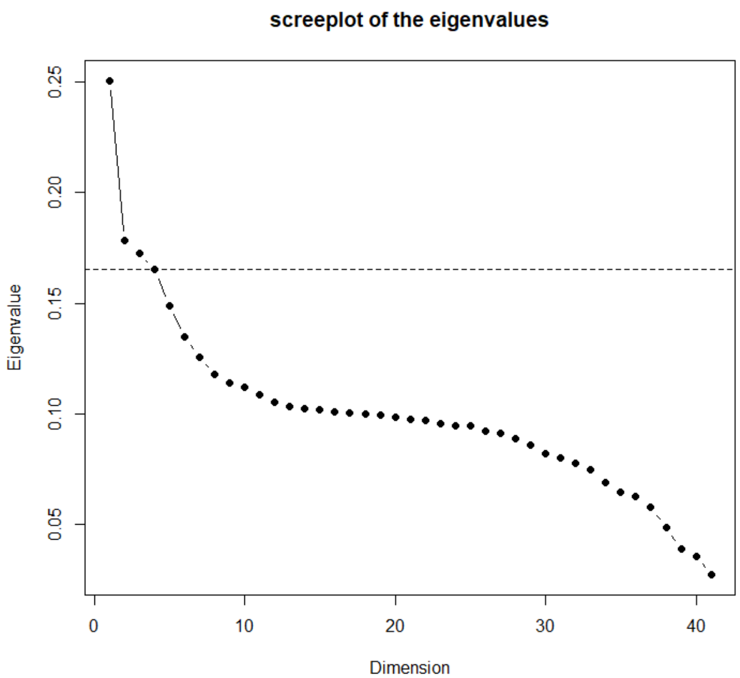


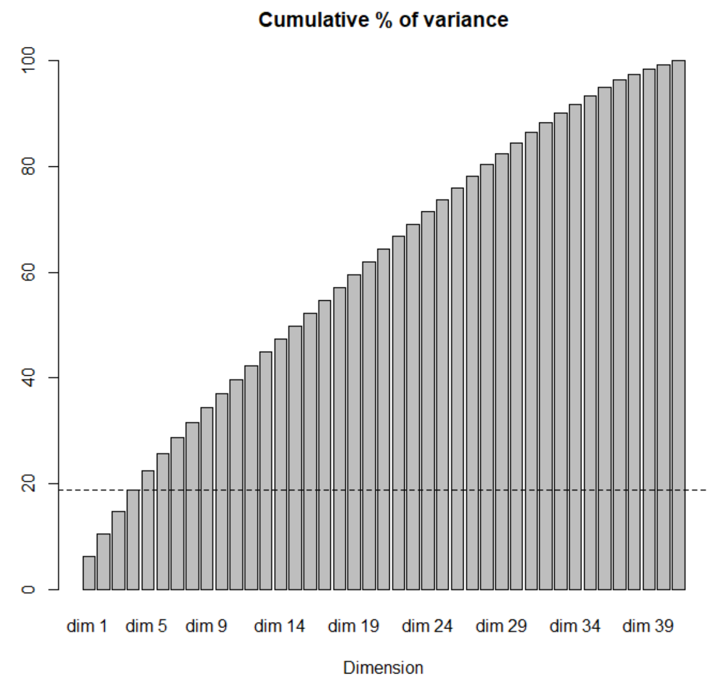


Validation protocol

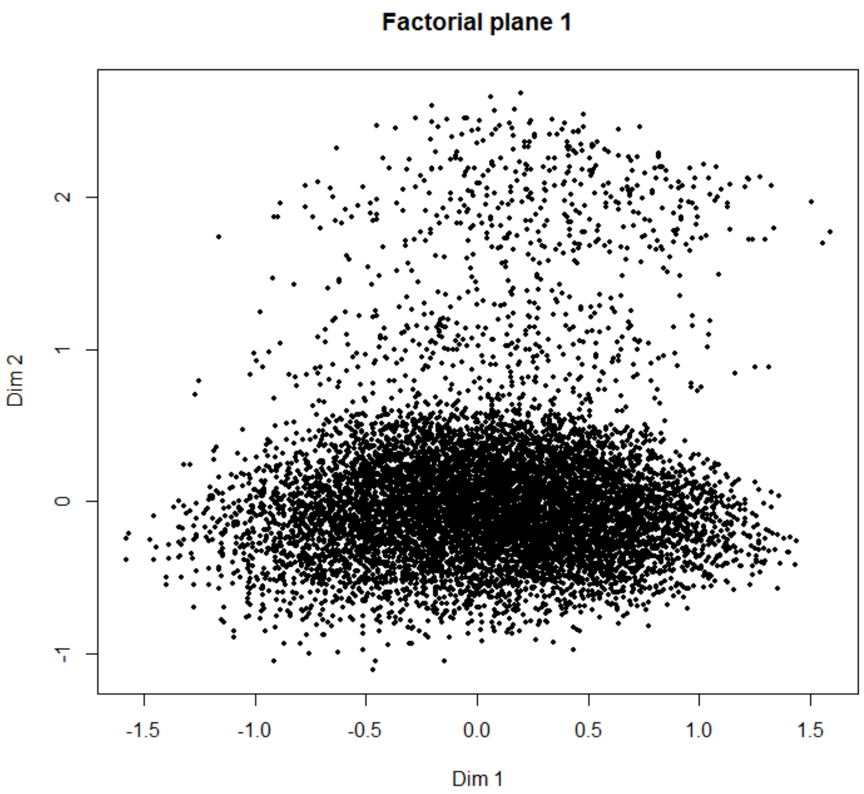
- Necessary in order to validate the results of the analysis
- High amount of data → **train** and **test** splits (50% each!)
- Reduce data from 48,842 data samples to two sets of 24,421 data samples each
- Allows us to use some time-consuming techniques in the analysis

Visualization (MCA)

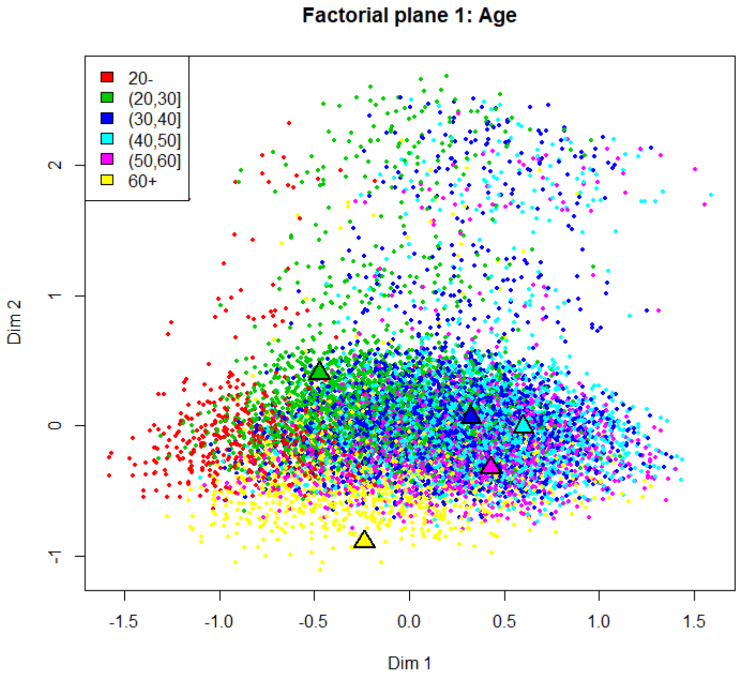


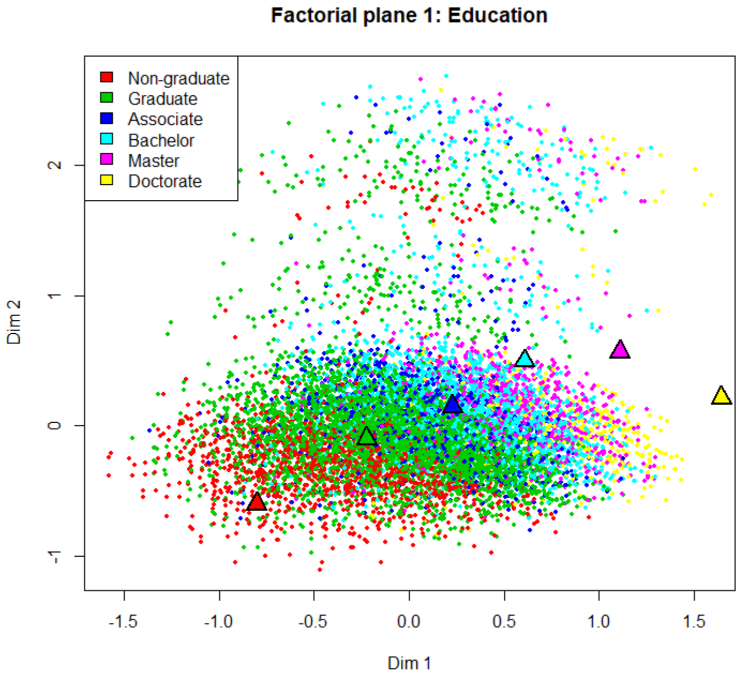


First factorial plane

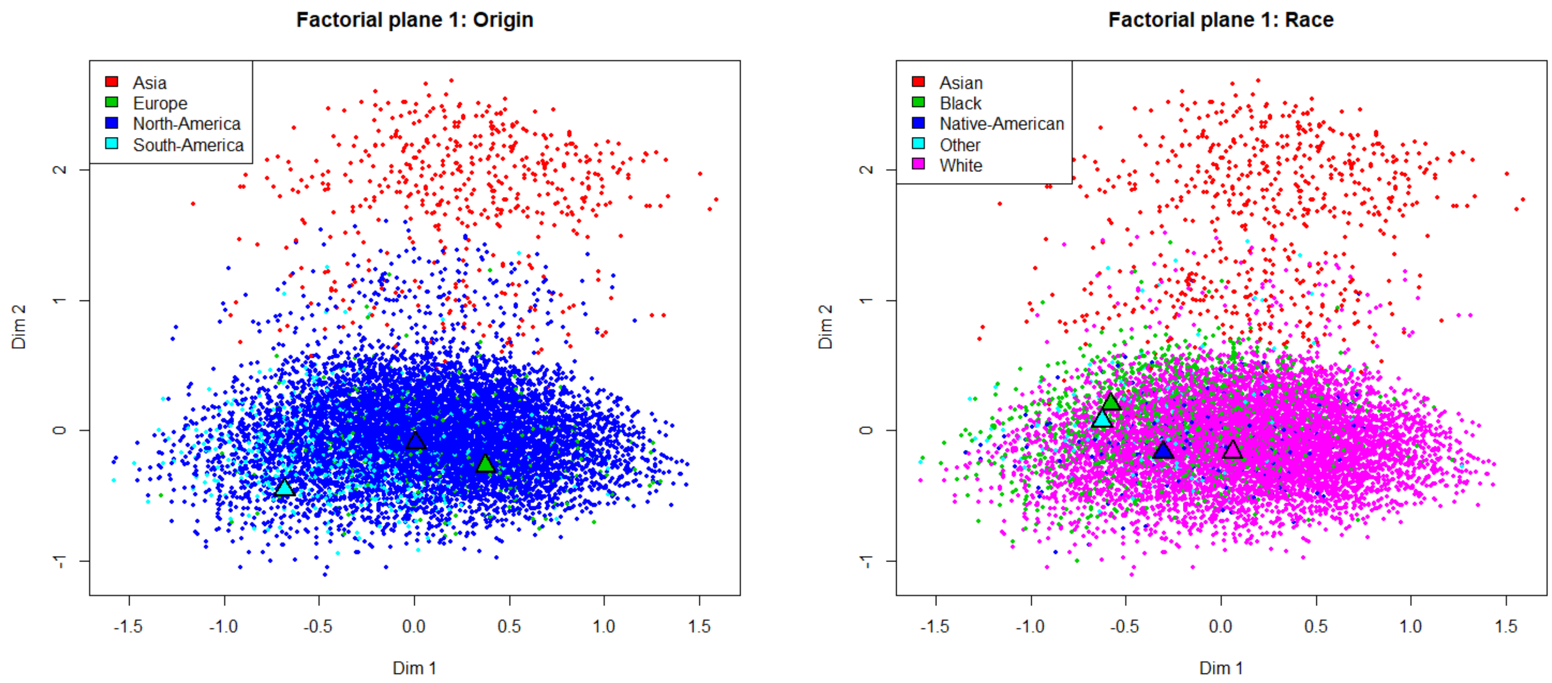


First factorial plane

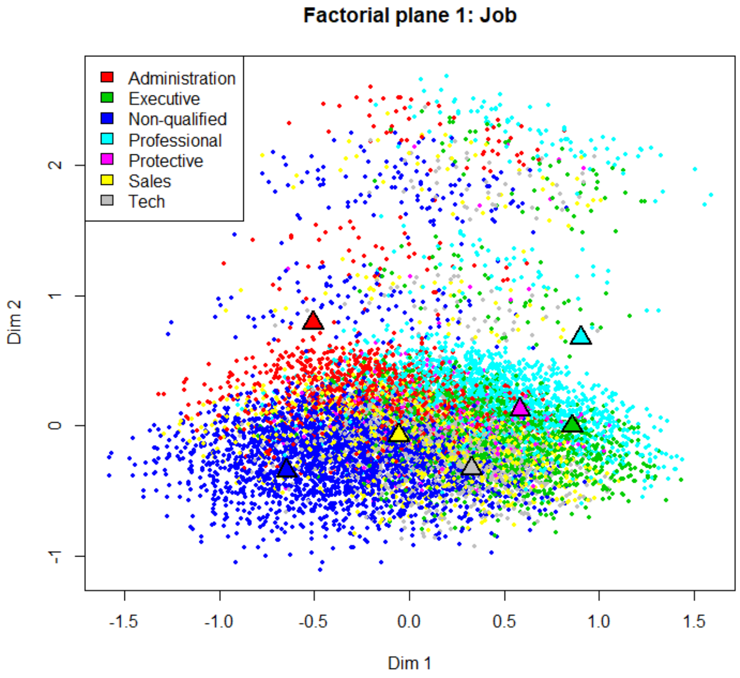


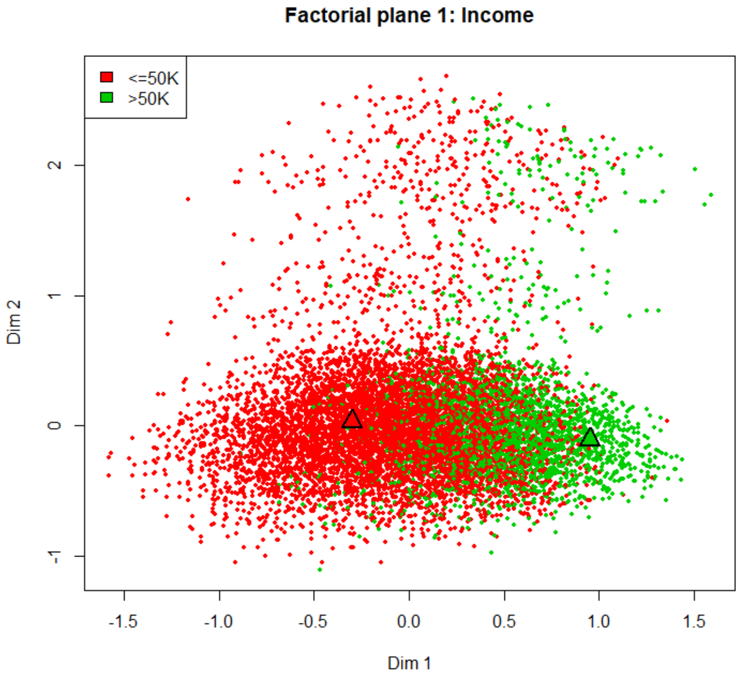


First factorial plane



First factorial plane





First component

- The First component is a latent variable that measures **how successful a certain individual is**
- Main modalities:
  - Age: between 30 and 60
  - Education: Bachelor or higher
  - Married: Married
  - Job: Executive, Professional or Protective
  - Working Hours: 40 or more
  - Income: More than 50K per year
- Better education and job is very correlated with Income!

Significant variables:

Variable	R2
Age	0.4227
Education	0.3191
Married	0.4004
Job	0.3939
WorkingHours	0.3404
Income	0.2860

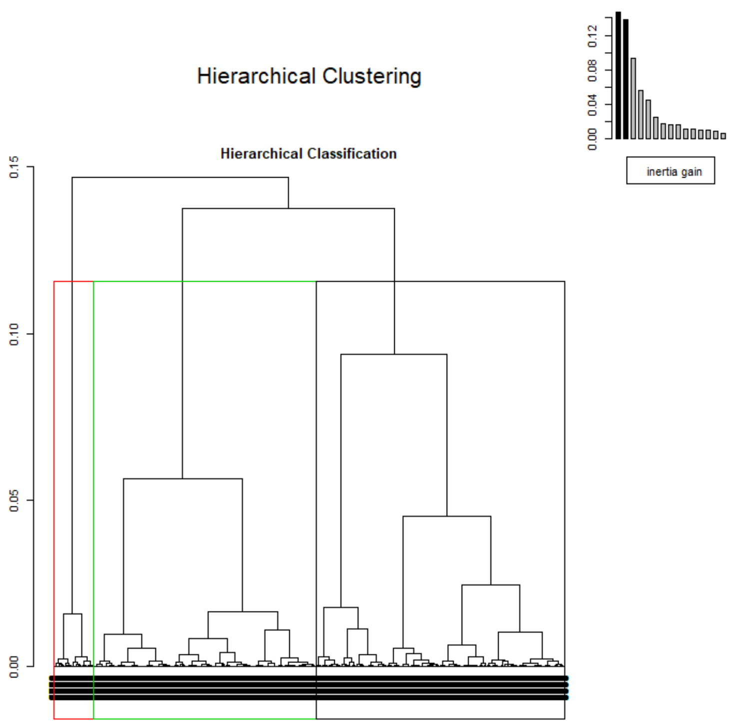
Second component

- The Second component is a latent variable that measures **if a certain individual is Asian, whether being born there or being descendant of Asians**
- Main modalities:
  - Origin: Asia
  - Race: Asian
- Cultural differences between Europeans, North-Americans and South-Americans is not as big as the difference with all of them and the Asian people

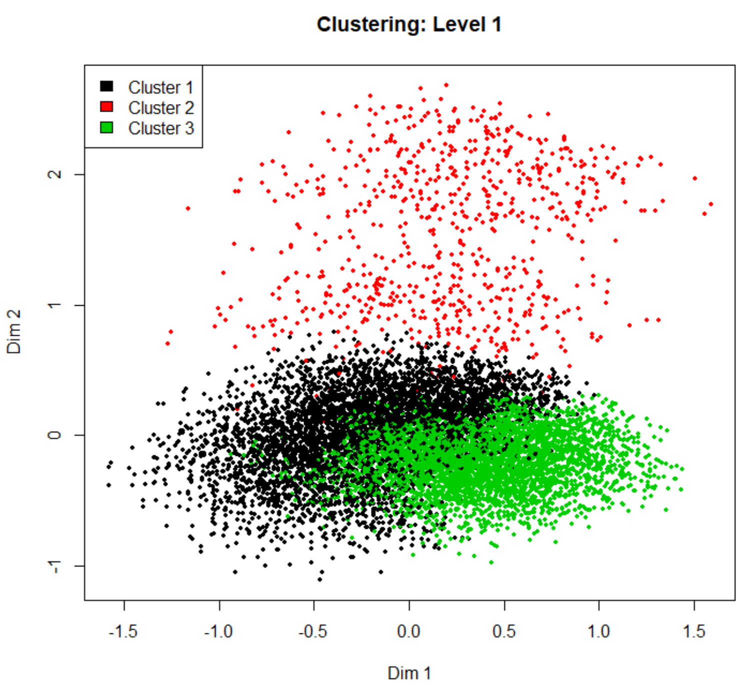
Significant variables:

Variable	R2
Origin	0.4782
Race	0.5022

Clustering

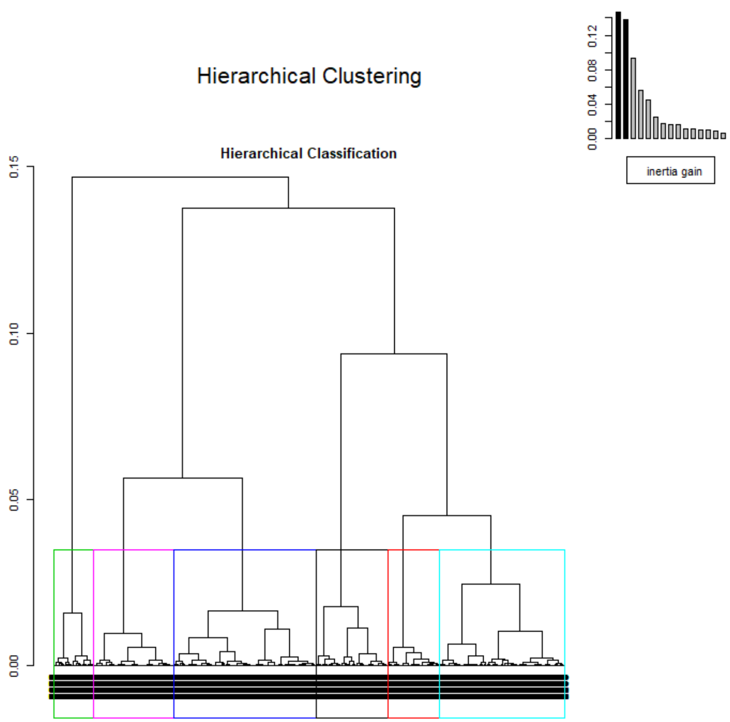


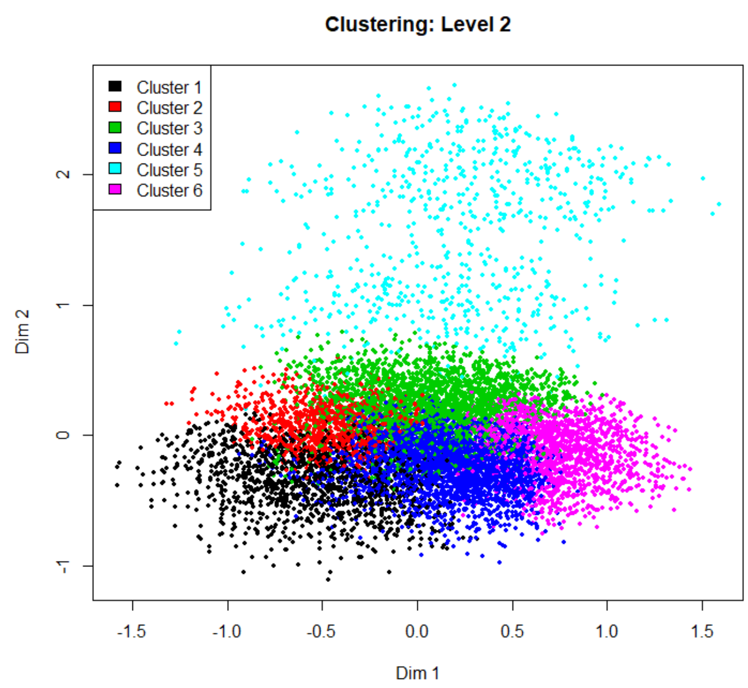
Hierarchical Clustering
Hierarchical Classification
inertia gain
0.00 0.04 0.08 0.12
0.00 0.05 0.10 0.15



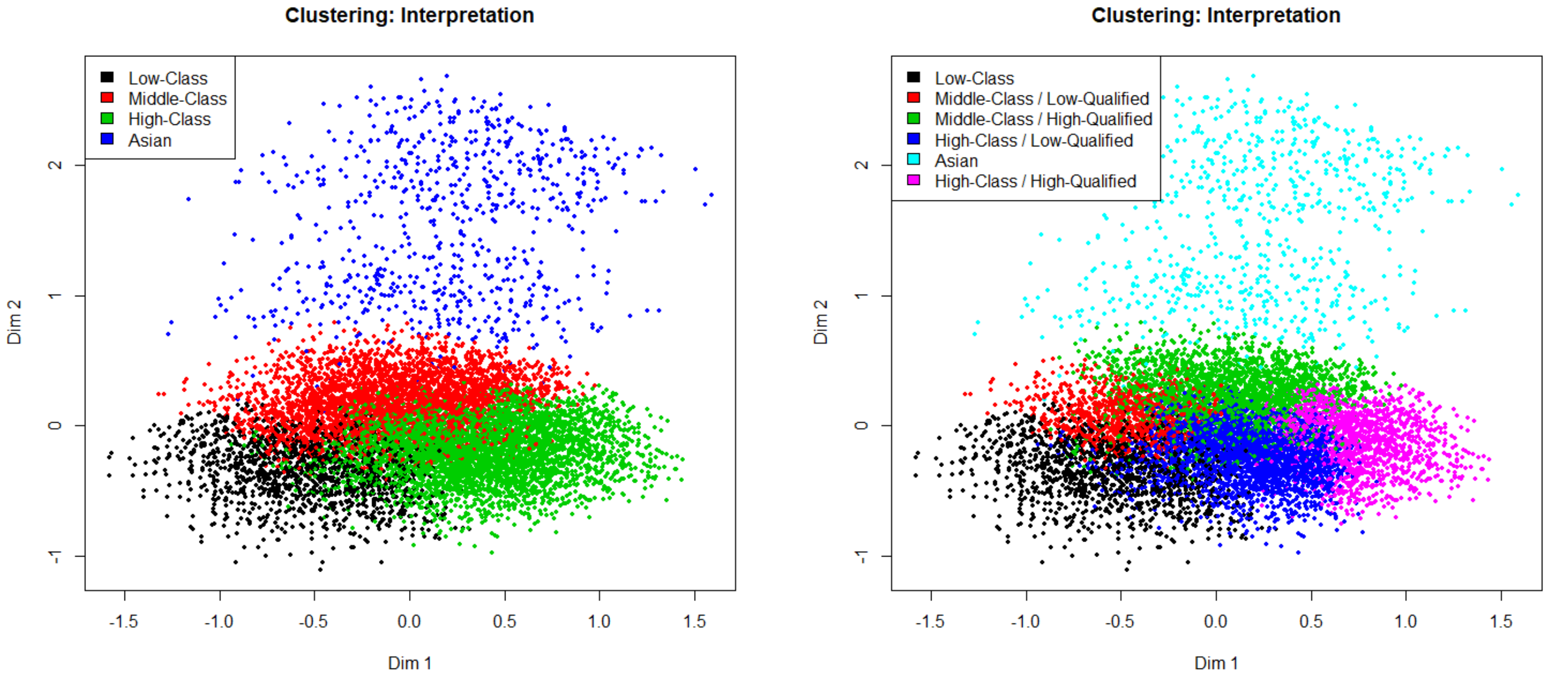
Clustering: Level 1
Cluster 1
Cluster 2
Cluster 3
Dim 1
Dim 2
-1.5 -1.0 -0.5 0.0 0.5 1.0 1.5
-1 0 1 2

Clustering


Hierarchical Clustering
Hierarchical Classification
inertia gain
0.00
0.04
0.08
0.12
0.00
0.05
0.10
0.15


Clustering: Level 2
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Dim 1
Dim 2
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
-1
0
1
2

Clustering Interpretation



Sample validation

1) Chi-squared test (individual tests)

Variable	P-value	Same?
Age	0.642	Yes
Sex	0.840	Yes
Education	0.124	Yes
Married	0.763	Yes
Job	0.328	Yes
Employer	0.136	Yes
Working Hours	0.880	Yes
Origin	0.144	Yes
Race	0.628	Yes
Capital	0.806	Yes
Income	0.656	Yes

Individually, all variables in both sets have the **same distribution**

2) Chi-squared test (single test)

In order to check homogeneity on the distribution of all the variables all together, we performed a Chi-squared test over the combined contingency table of all the variables

P-value: 0.8668

Globally, we accept that both sets have the **same distribution**

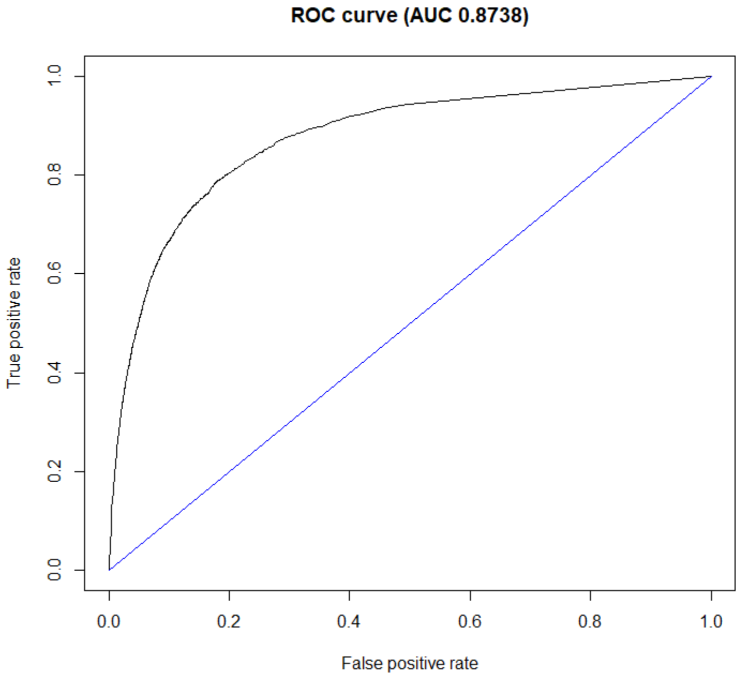
Modelling: Random Forest

- Prediction model: **Random Forest**
  - Very robust
  - Handles categorical data
  - Proven successful in many challenging tasks
- A Random Forest is an **ensemble** of several **Decision Trees**
  - Greater predictive accuracy
  - Less prone to overfitting
- In this Project, an ensemble of **1000 trees** is considered
- The rest of the parameters are optimized automatically

Modelling: Results

Confusion Matrix		Predicted	
		<=50K	>50K
Actual	<=50K	17322	1277
	>50K	2404	3418

Metric	Value
Accuracy	0.8494
Precision	0.7280
Recall	0.5879
AUC	0.8738



Conclusions

- Very interesting information about people in the USA in 1994
- Lots of variables and modalities
- Missing data is very interpretable
- Latent variables related to social status and origin culture
- Quite good model for Income classification
- In general, all the results were very interpretable
- Lots of MVA techniques have been applied

Thank you!

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB