INCOME PREDICTION FROM THE ADULT DATASET

Practical Work

David Morán
Joel Cantero
Xavi Timoneda

18/06/2019
MVA
MIRI–DS

Index

1. Description of the problem and available data

1.1. Description of the problem

In this project, we will be working with the Adult dataset[1], in which we have census data from the United States. The main objective of this project is to perform a full multivariate analysis and build a classifier that allows us to discriminate between people in that census that earn more and less than 50k dollars in one year.

This data was extracted from the census bureau database[2] by Barry Becker in 1994, and had a little preprocessing done. This data has some sociodemographic attributes about people living in the US in that year, and it is expected that there exists some correlation between these features and the amount of money they earn per year.

This data is very unbalanced, since only 23.93% of the elements of this dataset are members of the minor class (>50K). Therefore, it is not acceptable to have any error rate higher than this 23.93%. In fact, some of the previous studies[3] on this dataset suggest that it is quite easy to achieve an error rate around 15%.

It is a very interesting dataset from the multivariate analysis point of view, for several reasons. First, it is a very interpretable dataset, since all variables correspond to features everybody understands easily, like age, studies or job. Second, it is a dataset with a very big number of both features and instances, so there will probably not be many problems with the normality assumptions that are usually made. And finally, the dataset provides a mixture of numerical and categorical variables with several missing values, which leads to a great opportunity to apply all the techniques learnt in the multivariate analysis subject.

As for the classification task, it seems reasonable that the fact that one wins more or less than 50K dollars per year is quite related to those socioeconomical information, which should let to quite good models. On the other side, this information only makes sense in the big scale, since there might be certain individuals which share lots of characteristics but they have different incomes. Therefore, it will probably exist a limit on how good any model can perform on this task, since there is probably not enough information available in order to surpass that gap.

All the project, including data analysis, plots and modelling, will be done in the R programming language. The R code will be attached to this document in separate files.

[1] https://archive.ics.uci.edu/ml/datasets/adult
[2] http://www.census.gov
[3] https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names

1.2. Available Data

First of all, the data was obtained through the UCI website in two different files: "`adultPre.train`" and "`adultPre.test`". This files were loaded into the R system and combined into a single data frame in order to perform a general analysis. Since the fields were unnamed, some names were given to them according to its description in the website or a quick sight at the values they contain. This is a summary of the data:

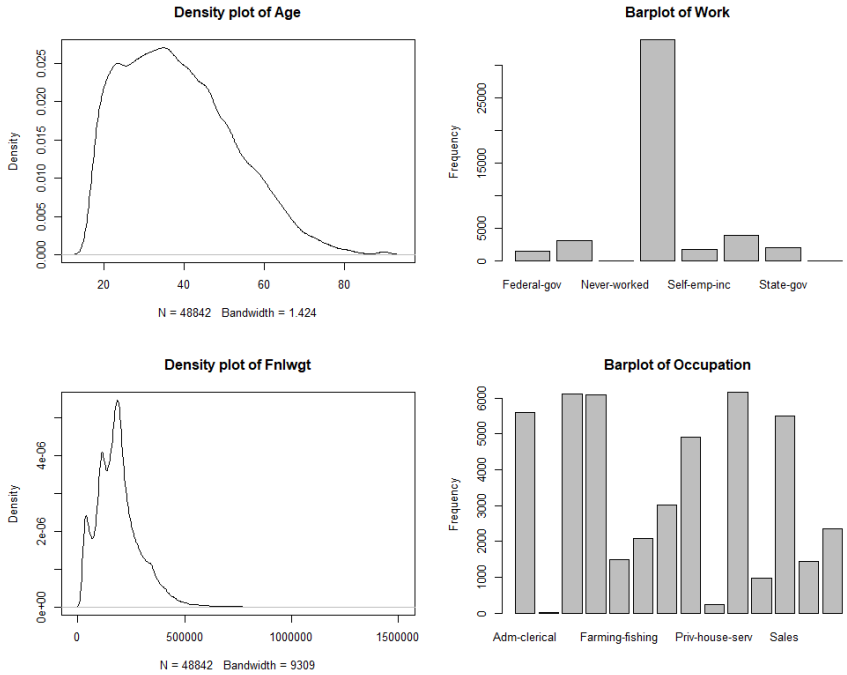
Field	Type	Values
Age	Numerical	[17,90]
Work	Categorical	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Fnlwgt	Numerical	[12285, 1490400]
Education	Categorical	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
EducationNum	Numerical	[1,16]
MaritalStatus	Categorical	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Categorical	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Relationship	Categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race	Categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex	Categorical	Female, Male
CapitalGain	Numerical	[0,99999]
CapitalLoss	Numerical	[0,4356]
WorkingHours	Numerical	[1,99]
NativeCountry	Categorical	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
Income	Categorical	<=50K, >50K

The first thing we noticed was that unknown values were labelled as "`?`", so we proceed to replace it everywhere with `<NA>` values. Then, all factor levels were updated so as to reflect this changes and the disappearing of the "`?`" value.

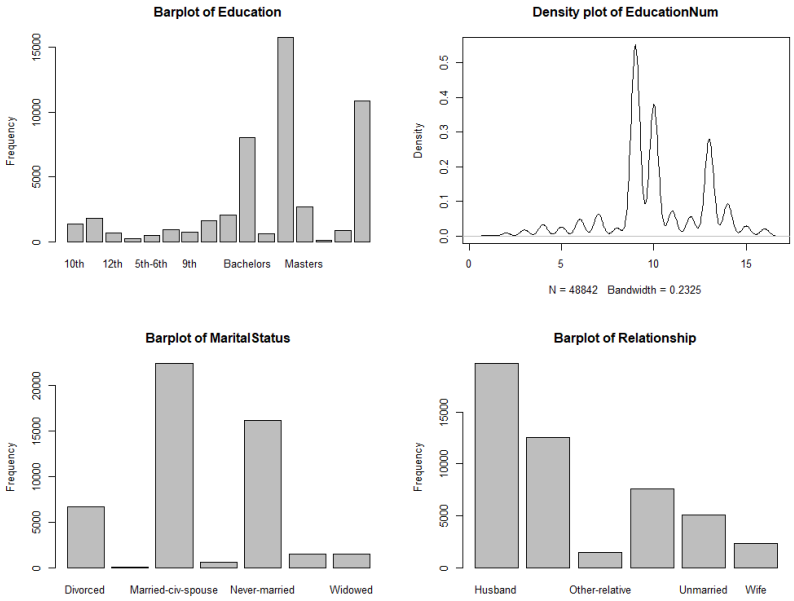
Then, we performed a first statistical analysis as a summary of the data before pre-processing. We also did a missing data analysis.

1.2.1. Univariate analysis

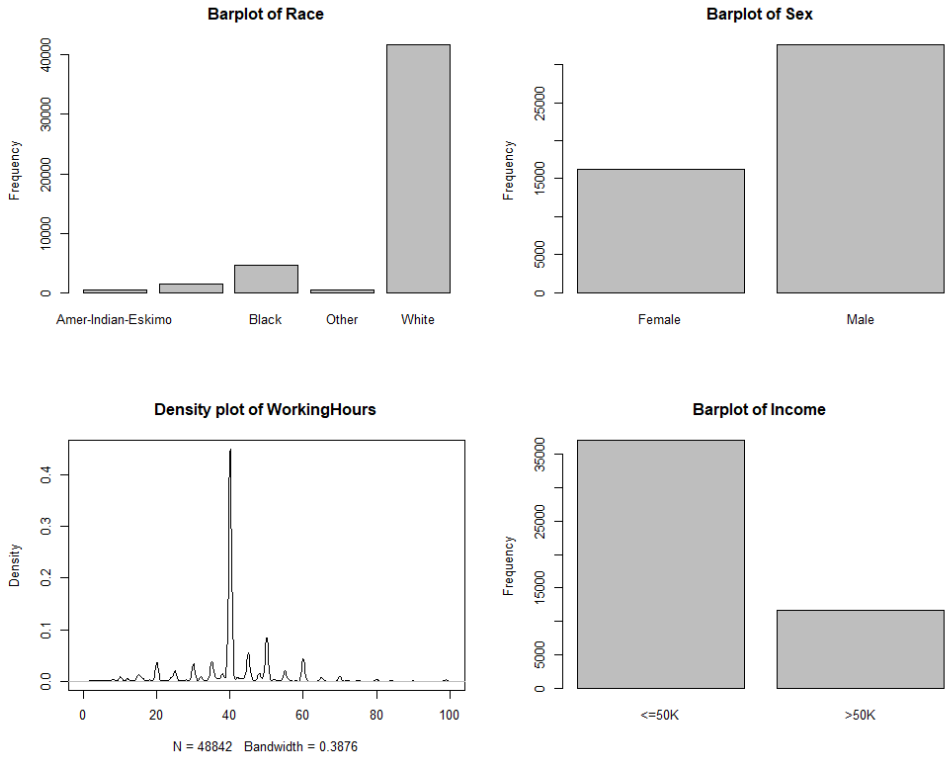
For the univariate analysis, density plots were obtained for the numerical variables and bar plots were generated for the categorical variables. These are the results obtained:



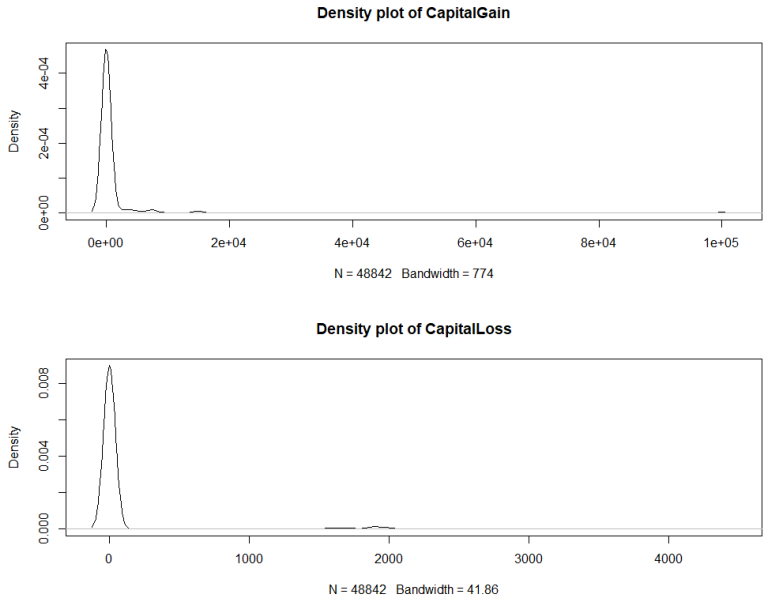
Here, we can see usual values for the Age, with a clear normal tendency. As for both Work and Occupation, there are some values with very low representation, and thus it might be interesting to perform some grouping and reduce the number of levels. Fnlwgt also seems sort of normal, so there are no problems there.



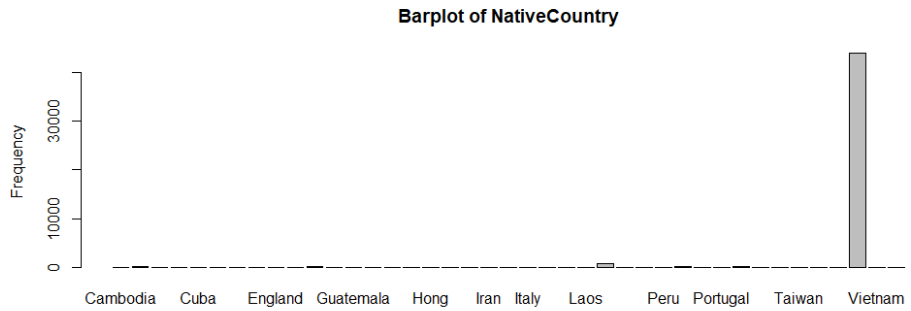
Here, we can see that, in fact, the levels of Educations correspond to the integer values of EducationNum, unsorted. The second one is probably the sorting order of the first one, so they both might be combined in a single field. As for MaritatlStatus and Relationship there are too many levels, so they also should be grouped.

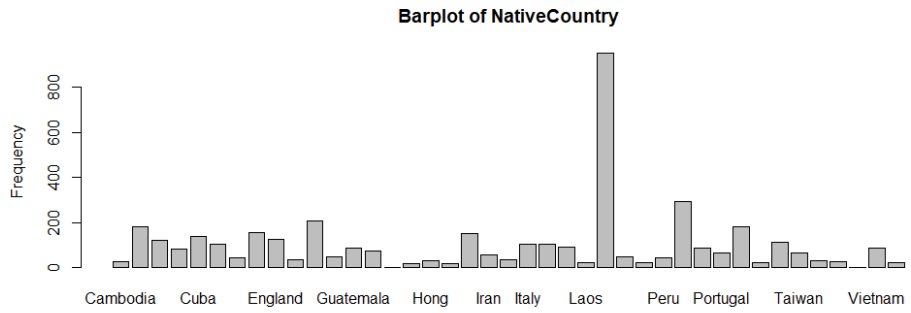


Here, we can see that over 95% of the people are either white race or black race. Therefore, most of the other levels should be combined. No problems for Sex, except for a certain unexpected unbalance in the male-female proportion. If we look at the density plot of WorkingHours, we can see that the value 40 is, by far, the most common. If we consider all the values except this one, we can see a certain tendency to normality centred around these 40 hours, but most of the cases are centred around values that are multiple of 5 (40, 35, 30…).



Here, we see once again that there is a value which appears in the majority of the cases: In this case, is the 0 value. Apart from this zero value, we see that small gains are more common than medium gains, and that medium losses are higher than smaller or bigger ones.



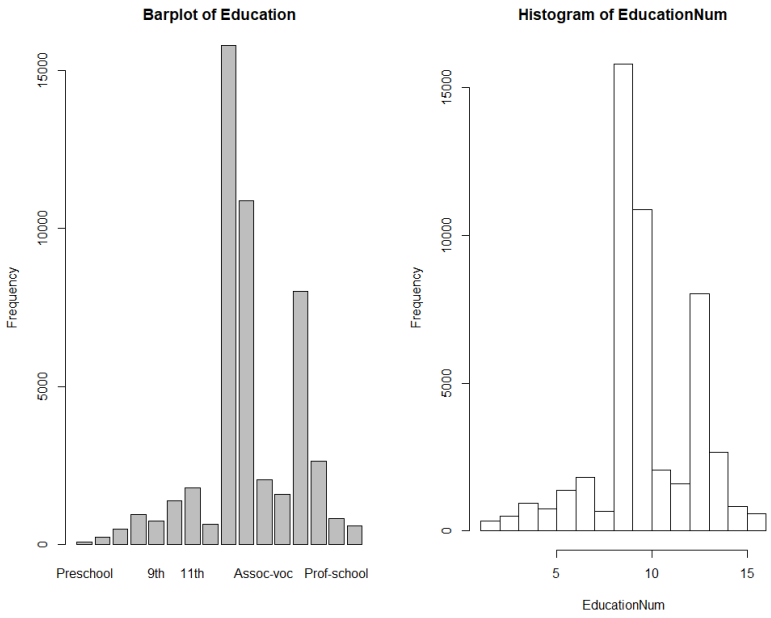


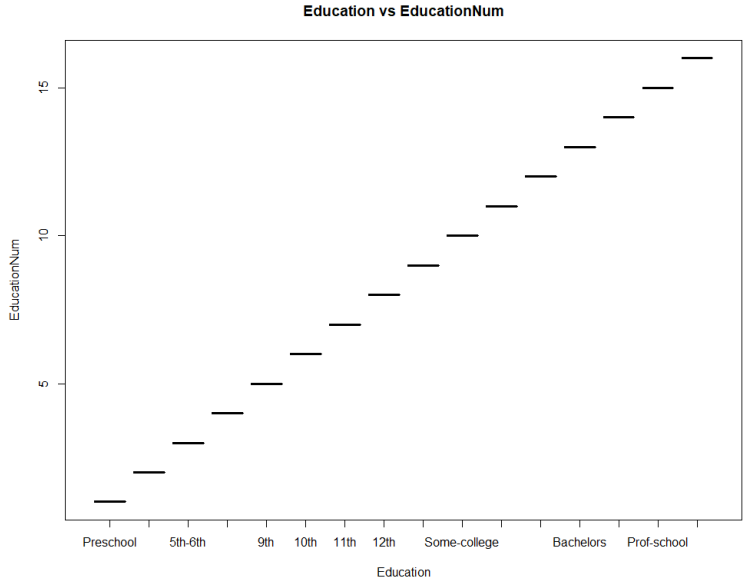
When analysing the NativeCountry field, we noticed that almost 90% of the people in this dataset are from the United States. If we plot the native country without the USA, we can see that Mexico is the second one, and all the others are more or less equally distributed. In fact, each of the other countries are at most of 400 people, which is less than 1% each. Only Mexico has around 2% of the people. Therefore, this levels should be highly blocked, for instance by continent or big geographical area. Otherwise, the analysis will be very complicated and too much specific for each native country.

1.2.2. Bivariate analysis

For the bivariate analysis, we are not able to perform all binary combinations, since for d features there are $\frac{d(d-1)}{2}$ binary combinations of features. In our case, that is 105 combinations! Therefore, only the most interesting pairs are analysed, the ones we suspect there might be some interaction.

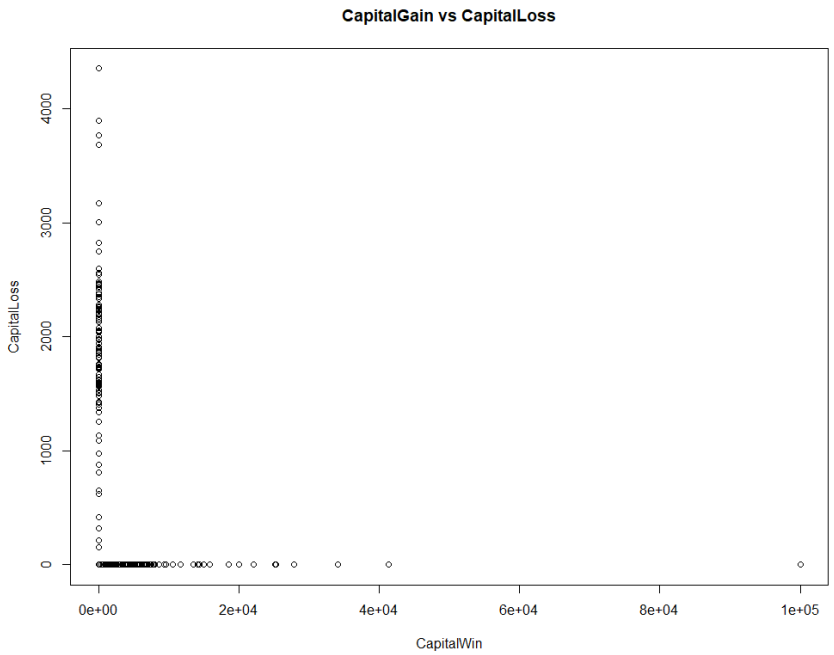
The first thing we did was to compare Education with EducationNum. With this purpose, we sorted the levels of Education by its corresponding number in EducationNum. Then, we built again both univariate plots and its combined plot, so we obtained the following:





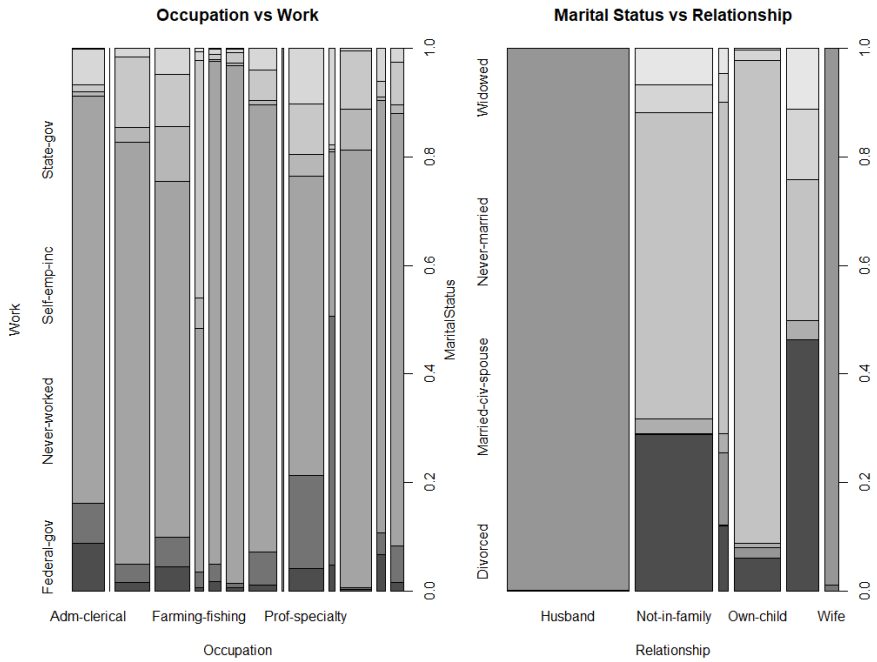
If we compare the bar plot and the histogram, we can now clearly see that they are absolutely equal. Also, in the Education vs EducationNum plot (which is in fact a boxplot), we can see that there is a perfect correlation between each Education level and its corresponding EducationNum. Therefore, the field EducationNum must be removed and Education must be considered as a sorted factor.

Then, we proceed to analyse both capital features. If we plot CapitalGain vs CapitalLoss, we obtain the following plot:



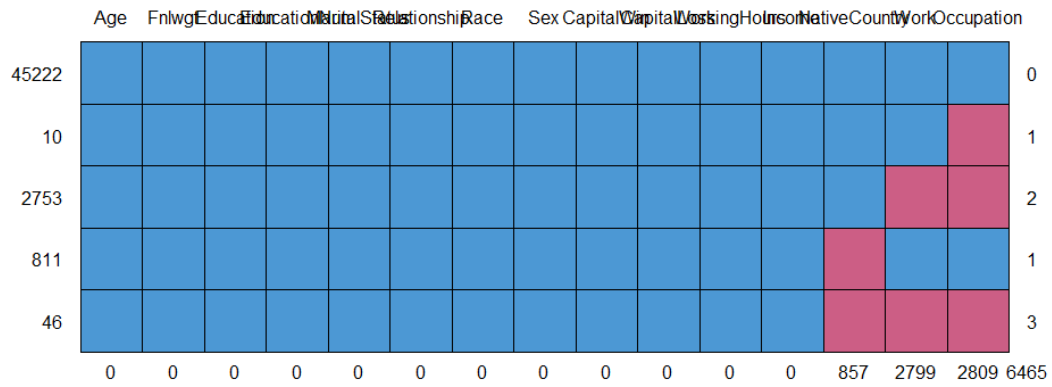
Here, we can see that there is no point which has both positive CapitalGain and CapitalLoss. In other words, for each instance of the dataset, one can have either some gain or some loss, but not both. This can be used to combine both features into a single one, with positive value for the gains and negative values for the losses.

Finally, we also tried to see if there was any kind of relation between Work and Occupation, or MaritalStatus and Relationship. In both cases we found that the huge number of levels did not allow us to perform any kind of analysis, so we suggested reducing the number of levels before trying this analysis again.



1.2.3. Missing data analysis

To perform some missing data analysis, the first thing we did was to do a casuistic occurrence analysis of the missing values. It can be seen in the following plot:



Here, we can see that only NativeCountry, Work and Occupation have missing values. There are three cases in which Occupation is missing: Alone, together with Work and together with Work and NativeCountry. Since the three together missing is quite rare (only 46 cases), we will assume that NativeCountry missings are independent of Work and Occupation ones.

If we look at all the 10 cases in which only Occupation is missing, we can clearly see that its work is always "`Never-worked`". This might mean that there is some people which have no particular occupation, since they were never able to work anywhere. In this cases, we can state that this "`No-occupation`" might be considered "`Non-qualified`" for the sake of simplicity.

If we analyse the cases with missing Work and Occupation, it seems that people who never worked and do not have an occupation might have not answered this question. Therefore, in these cases we may change this pair of missing values of Work + Occupation into "`None`" + "`Non-qualified`". Since we assumed that missing values in Work + Occupation and in NativeCountry are independent, we can apply the same treatment to the cases with three missing values.

As for the remaining NativeCountry missing values, imputation might be performed in order to fill the instances that do not have any country assigned. This imputation can be done by deducing the original native country using the rest of the dataset.

2. Data pre-process

After performing the data analysis of the data, we applied all the simplification suggestions we had based on the data exploration. It is basically a data selection, cleaning formatting process, reconstructing data.

2.1. Pre-process

First, some categorical variables are blocked into a lower number of modalities, numerical variables are discretized and some of the variables are removed.

Field	Old value	New value	Field	Old value	New value
Work	Federal-gov	Gov	**Occupation**	Adm-clerical	Administration
	Local-gov			Exec-managerial	Executive
	State-gov			Prof-specialty	Professional
	Without-pay	None		Armed-Forces	Non-qualified
	Never-worked			Farming-fishing	
	<NA>			Handlers-cleaners	
	Self-emp-inc	Self		Machine-op-inspct	
	Self-emp-not-inc			Other-service	
	Private	Private		Priv-house-serv	
Education	Preschool	Non-graduate		Transport-moving	
	1st-4th			<NA>	
	5th-6th			Craft-repair	Tech
	7th-8th			Tech-support	
	9th			Protective-serv	Protective
	10th			Professional	Professional
	11th			Sales	Sales
	12th		**Race**	White	White
	HS-grad	Graduate		Black	Black
	Some-college			Amer-Indian-Eskimo	Native-American
	Assoc-voc	Associate		Asian-Pac-Islander	Asian
	Assoc-acdm			Other	Other
	Bachelors	Bachelor	**Age**	(-∞,20]	20-
	Masters	Master		(20,30]	(20,30]
	Prof-school	Doctorate		(30,40]	(30,40]
	Doctorate			(40,50]	(40,50]
MaritalStatus	Divorced	Divorced		(50,60]	(50,60]
	Separated			(65,∞)	60+
	Never-married	Never-married	**WorkingHours**	[0,5)	0
	Widowed	Widowed		[5,15)	10
	Married-spouse-absent	Married		[15,25)	20
	Married-AF-spouse			[25,35)	30
	Married-civ-spouse			[35,45)	40
CapitalLoss	[0, 500]	ZeroCapital		[45,55)	50
	(500, 2000]	Loss		[55,65)	60
	(2000,∞)	HighLoss		[65,∞)	70+
CapitalGain	[0, 500]	ZeroCapital	**Relationship**	–	Removed
	(500, 4000]	Gain	**Fnlwgt**	–	Removed
	(4000,∞)	HighGain			

It is important to notice that the variables **CapitalGain** and **CapitalLoss** are merged together into a single **Capital** variable.

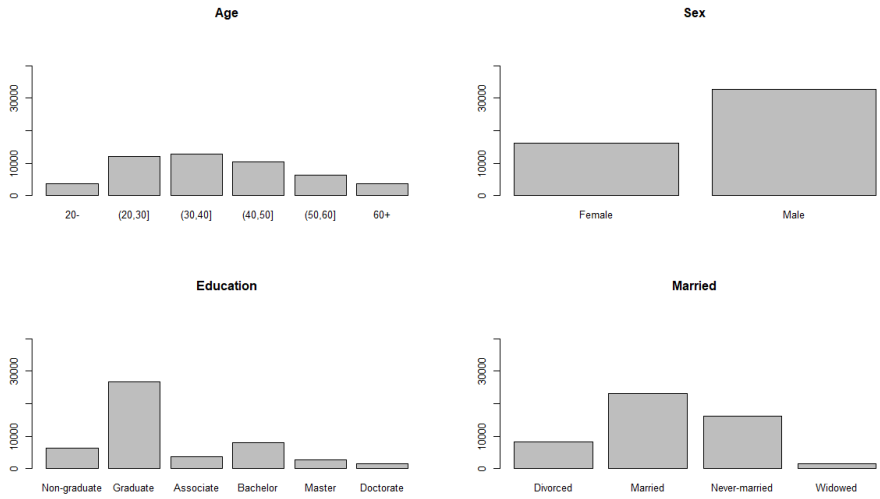
Due to the huge number of countries, this feature transformation is summarized in a different table:

Old value	New value	Old value	New value
United-States	North-America	Cuba	South-America
Outlying-US		Dominican-Republic	
Canada		Ecuador	
Cambodia	Asia	El-Salvador	
China		Guatemala	
Columbia		Haiti	
Hong		Honduras	
Laos		Jamaica	
Philippines		Mexico	
South		Nicaragua	
Taiwan		Peru	
Thailand		Puerto-Rico	
Vietnam		Trinadad&Tobago	
Japan		England	Europe
Iran		France	
India		Germany	
		Greece	
		Holand-Netherlands	
		Hungary	
		Ireland	
		Italy	
		Poland	
		Portugal	
		Scotland	
		Yugoslavia	

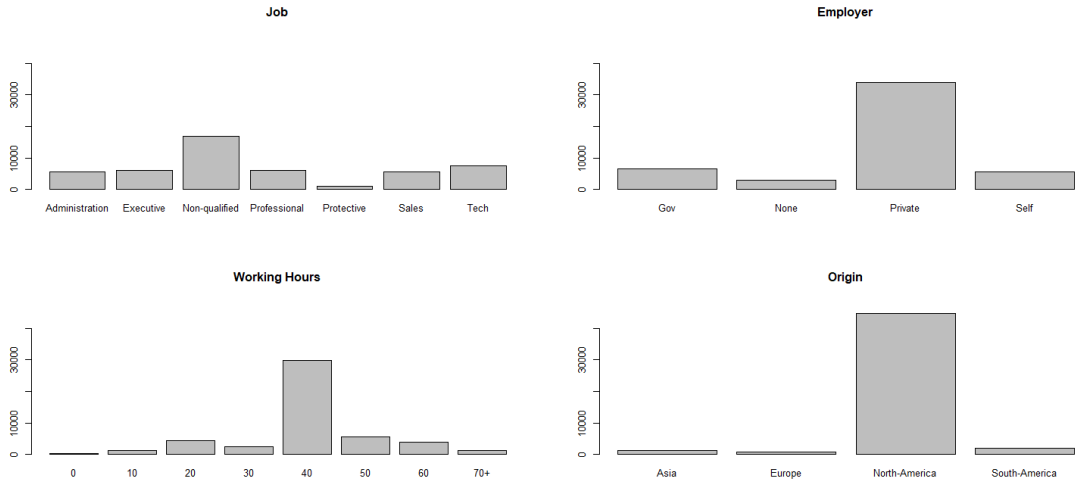
2.2. New data Summary

Since the previous pre-processing modified significantly the previous analysis, it is interesting to repeat it with the new values of the dataset. Also, since the aim of this project is to perform income prediction, the relation of this variable with the rest of variables is also studied.

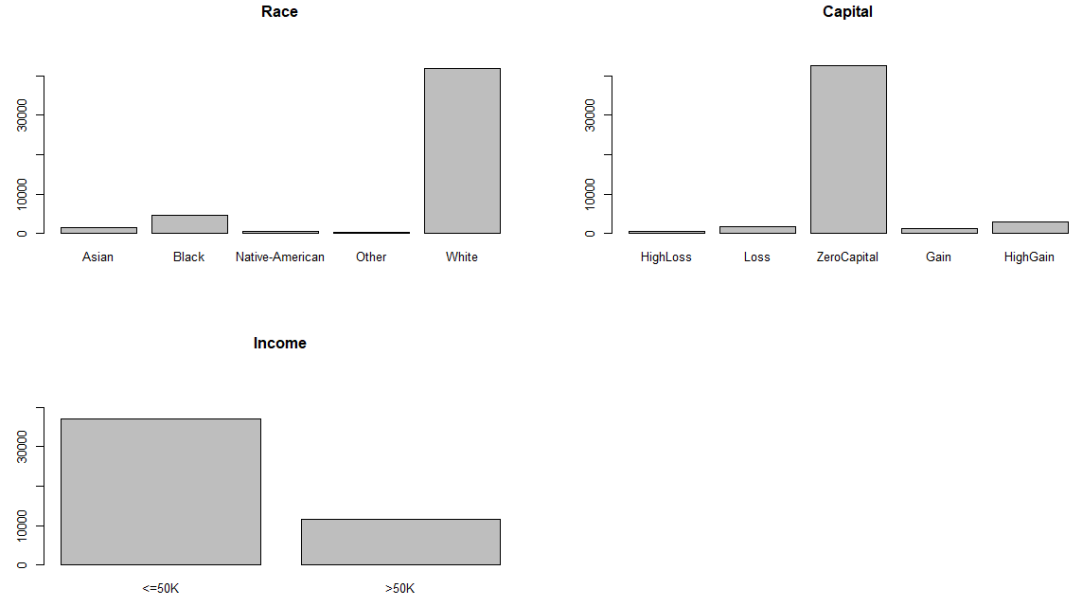
2.2.1. Univariate analysis



In this plot, we can see that normality is still observable at the Age variable. Also, the levels of Education are more compact now, and each class has more members. The same can be said about the Married modalities, which are more interpretable now.



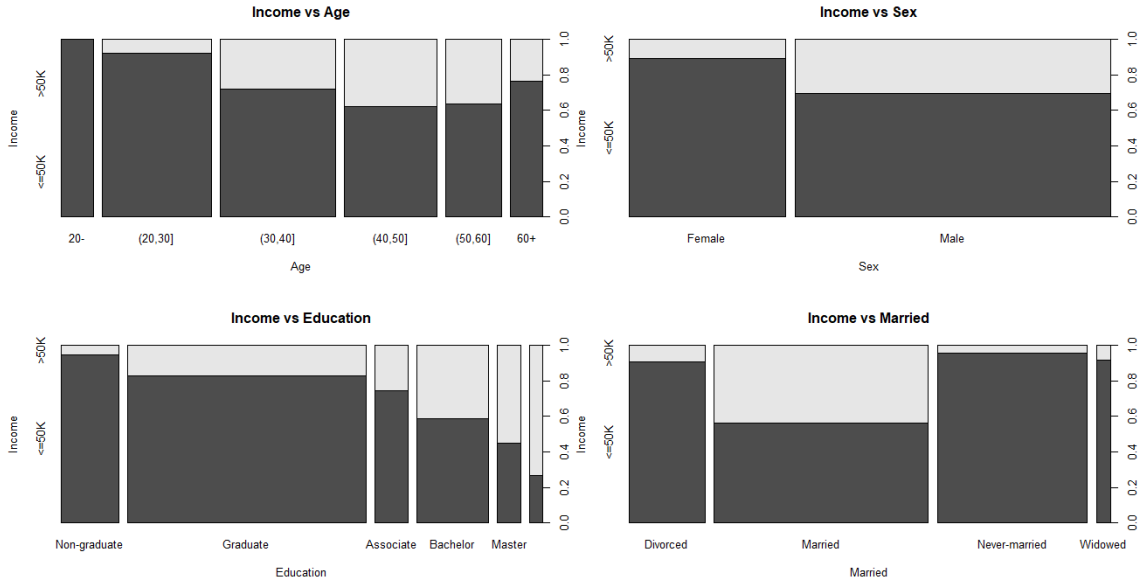
In this plot, we can see that the Job variable is more balanced now, and so is the Employer variable. Also, we keep the normality in the working hours (except for that huge amount of people around 40 hours) and the Origin of the people is more interpretable now.



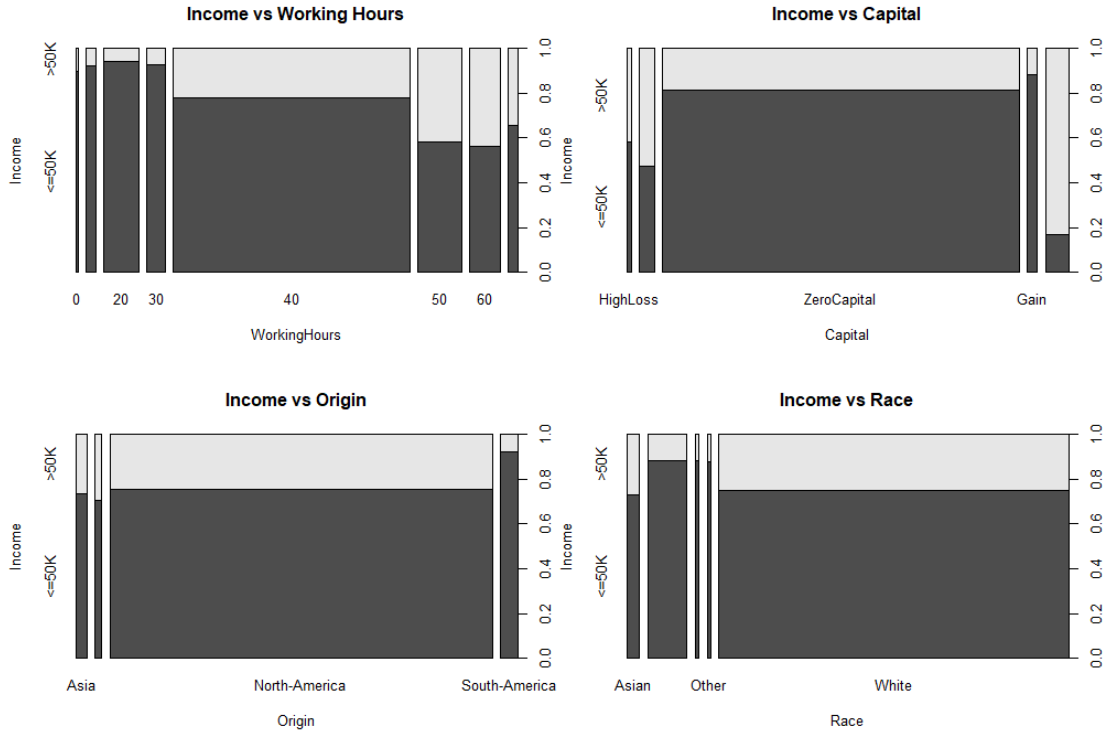
In this plot, we can see that the Race distribution has not changed, even though the names have been changed to a more clarifying ones. Also, it can be seen that Capital has the huge majority of the people having zero capital, and few people holding some gains and losses.

2.2.2. Bivariate analysis

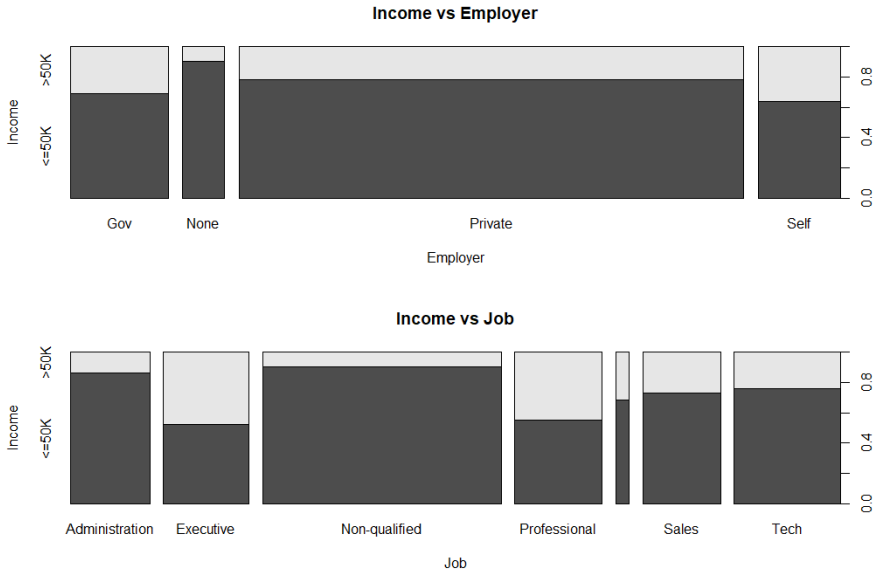
This bivariate analysis is oriented towards the income prediction. This means that all variables have been plotted against the Income variable to get an insight of its relation.



In this plot, we can see that people between 40 and 60 have better chance of achieving higher income. Also, males have a slightly bigger proportion of samples achieving >50K Income. As for Education, it is clear that there is a strong correlation between the level of studies and the probability of earning more than 50K per year. Also, married people are more likely to achieve higher income that any kind of single people.



In this plot, we can see that working more hours increases the chances of earning >50K, but just up to 60 hours. This might be due to high executives working lots of hours, but people who work more than 70 hours a week are less likely to be executives and hence they earn less money. As for the Capital, it seems than people with any kind of movement of capital is more likely to have higher income. This may be due to the fact that only people who earns lots of money are able to perform capital movements of any kind. When we look at the origin of the people, it does not seem a major factor in terms of Income, even though people from South-America tend to earn less than the other origins. The same happens with the Race, being in this case the Black people who tend to have lower incomes.



In these plots, we can see that the type of employer is only important if you do not have any, and that professional and Executive jobs tend to earn more money than the other jobs. In particular, the Non-qualified are the people who tend to have lower Income.

2.2.3. Outlier detection

It is interesting to talk about outlier detection, even though this topic is not addressed in this work. This is due to all our variables being categorical, and hence being no concept of *outlier* other than a certain combination of variable values being less likely to happen. And, in our case, that would be considered relevant information, instead of an outlier.

In any case, discretization of the continuous variables helps with dealing with the outlier problem, since variables that have big (positive or negative) values get counted as the further category. For instance, ages of 99 got into category 70+, and hence being treated as if they were, for instance, a 75. This can be considered therefore some kind of outlier correction.

3. Validation Protocol

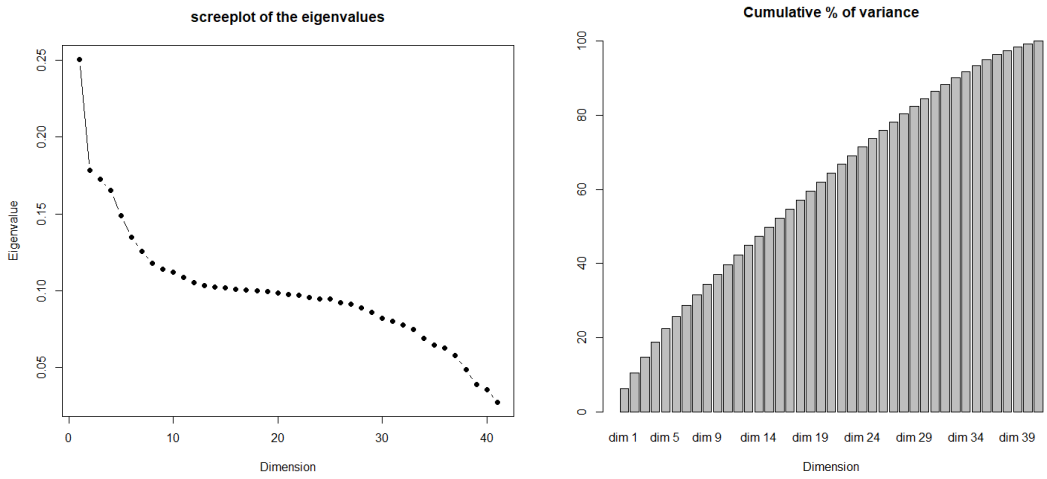
In order to validate the results of the multivariate analysis, a validation protocol must be defined. In this project, since the amount of available data is very high, we decided to split our dataset into a training and a test sets. Since we want to use some expensive methods in the multivariate analysis, like hierarchical clustering, we do not want to hold a very big dataset. Therefore, it was decided that each split would hold exactly 50% of the data, reducing the total dataset from 48842 data samples to two different data sets holding 24421 data samples each. The partition was performed randomly using the `sample` function, and having previously set a seed of `1234` so as to make all the results reproducible.

4. Visualization

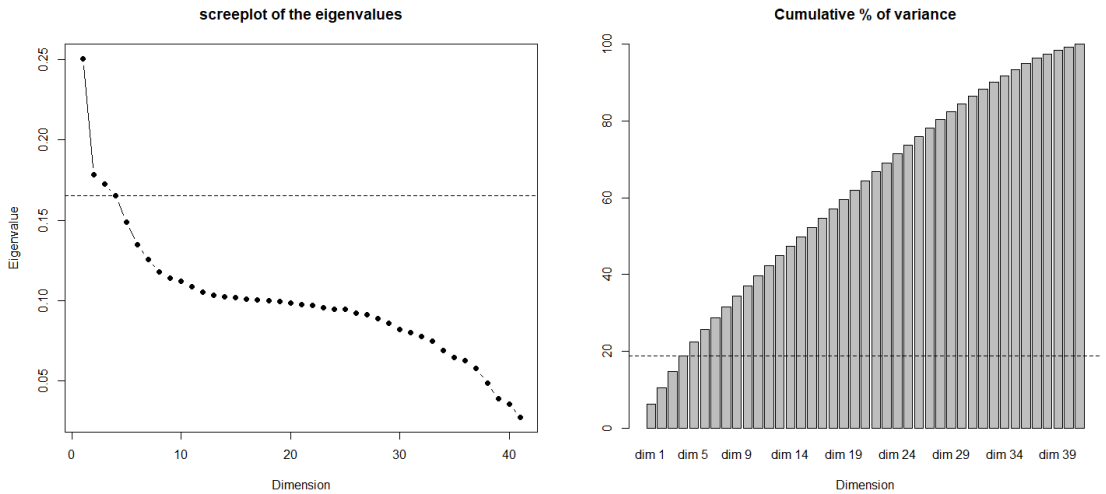
Since all the variables in our dataset are categorical, MCA was performed over the data to perform a multivariate analysis and proper visualization. In order to perform such MCA, test data was taken as supplementary, whereas train data was used to perform the MCA. Also, the Income variable was also provided as supplementary, because since it is the target variable of this problem, we want to see if there is any relation of this variable with the other ones.

First, it is important to notice than the provided, pre-processed dataset spans a 41 dimensional space from all its variables and modalities. Therefore, it is important to recall that the dimensionality of the MCA space is very big, and that it makes sense to pick just a reduced number of dimensions to perform the analysis.

Then, the application of MCA to the Adult dataset resulted in the following dimensions:



As it can be seen on the right plot, the cumulative percentage of variance grows quite uniformly, since the differences on the eigenvalues are not very big. Therefore, a huge number of dimensions would be required to keep a significant part of the variance. Therefore, we decided to keep just the most informative components, based on the elbow rule. Therefore, only **four** dimensions are kept for the rest of the procedure. The results after the cut can be seen in the following plots:

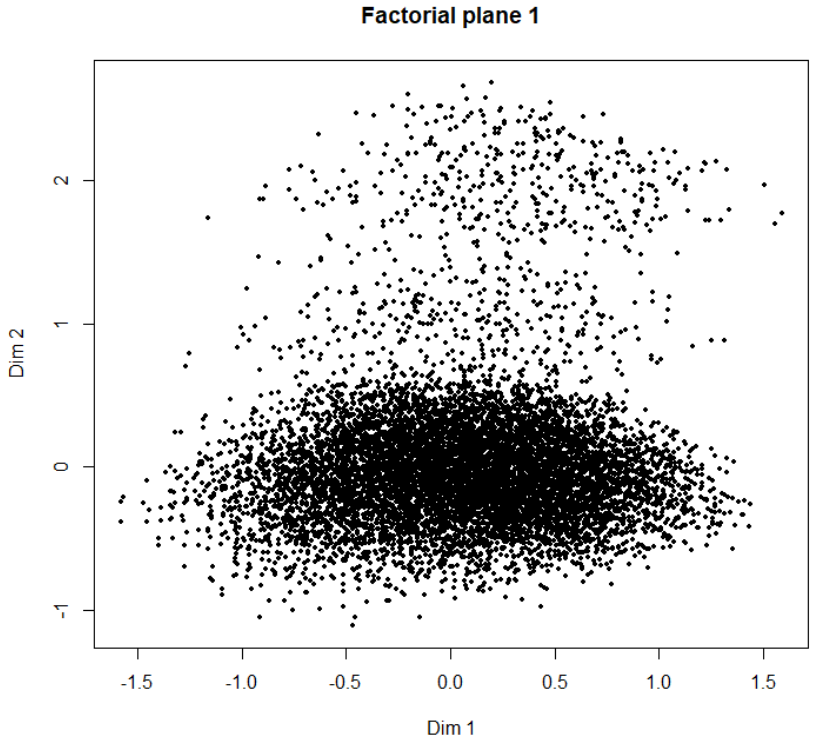


It is important to take into account that only 18.70% of the variance is kept within this four dimensions, with is less than a fifth of the total variance. This might seem like a very low value, but it has to be taken into account that this is achieved with less than 10% of the components. Also, the rest of the components quickly tend to retain almost no information, and hence they might be no interpretable at all.

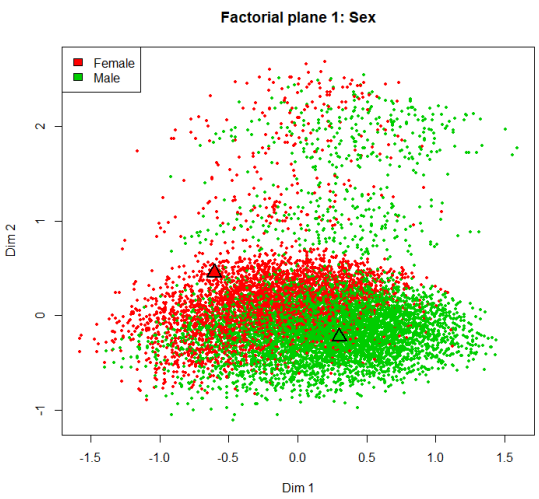
Then, those four components are grouped into two factorial planes. In order to perform a proper visualization of the data, it has been decided to plot, for all variables, all the data samples in the factorial plane, coloured according to its corresponding modality in that variable.

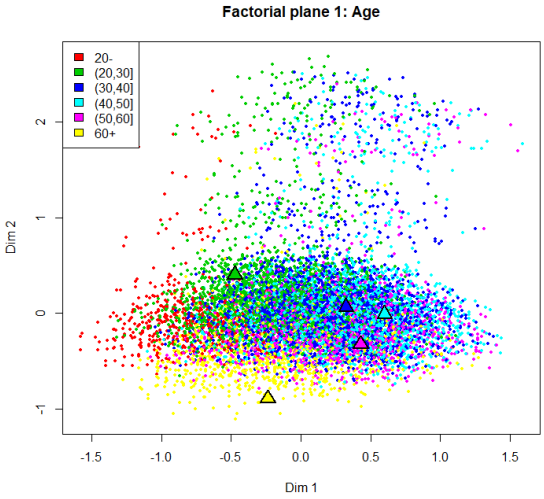
4.1. First Factorial Plane

The results of the first factorial plane can be seen in the following plots:

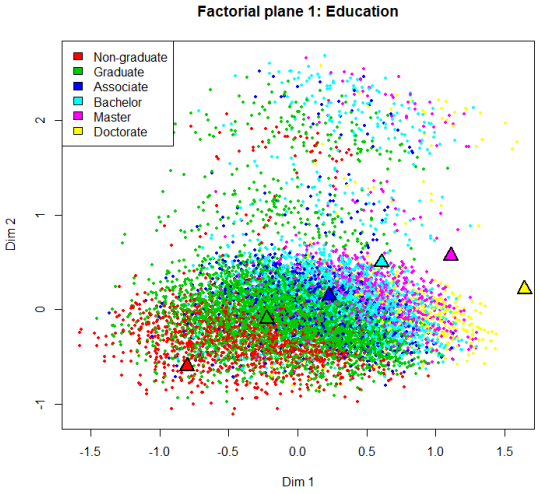


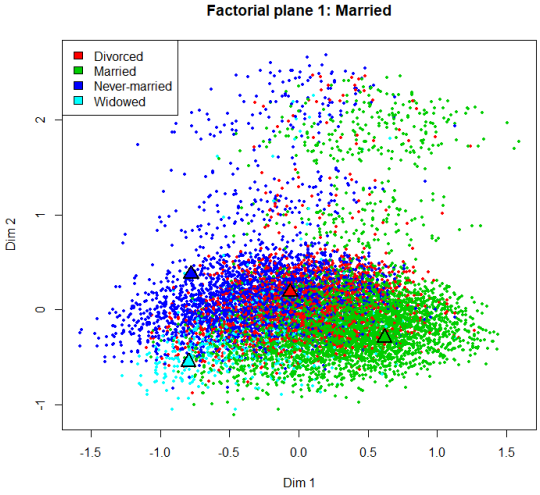
A first visual inspection on the factorial plane leads to a clear identification of three clouds of points: There is a very big, very dense cloud at the bottom, a very sparse cloud at the top, and an even sparser cloud of points between the other two. Its interpretation will be considered in the latent variables section.



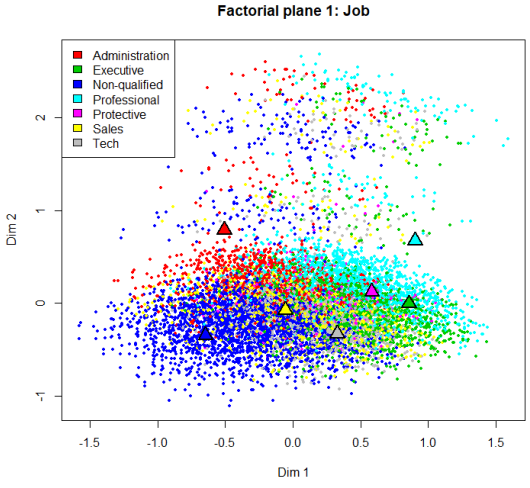


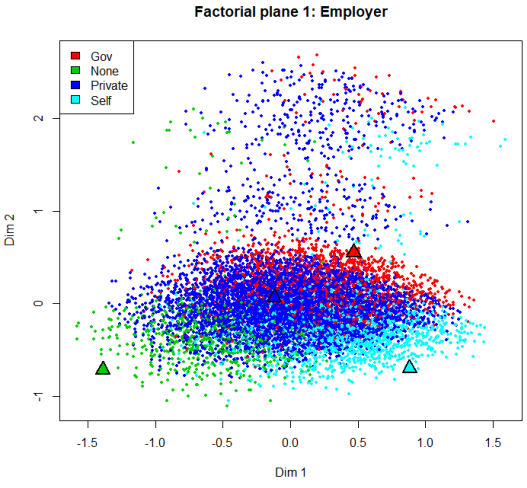
If we look at the Sex plot, it seems that males are mostly grouped at the bottom right of each of the clouds, and females are distributed among the upper left corner of each of the clouds. In the Age plot, younger people (less than 30) are at the left of the clouds, whereas people from 30 to 60 are at the right and people over 60 are at the bottom of the clouds.



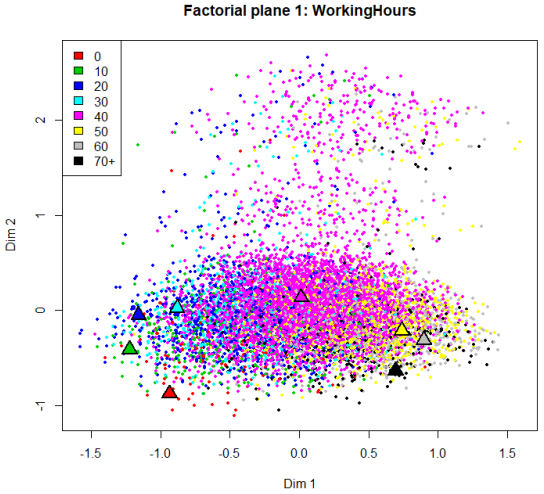


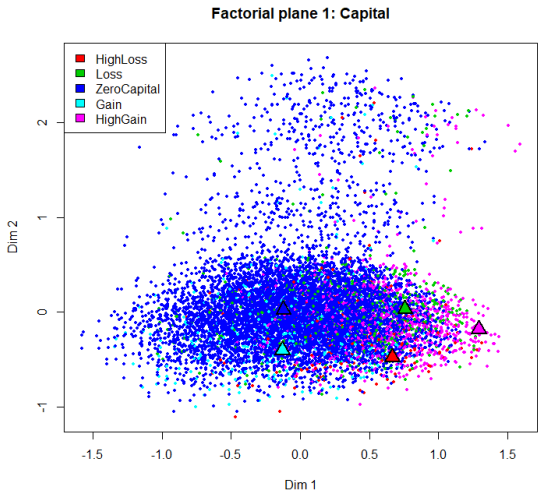
If we look at the Education plot, we can see that the Education levels are sorted from left to right, with a tendency to move up in the higher studies. As for the Married plot, it clearly separates the four categories, being Never-married and Widowed in the left, Married in the right and Divorced in the middle.



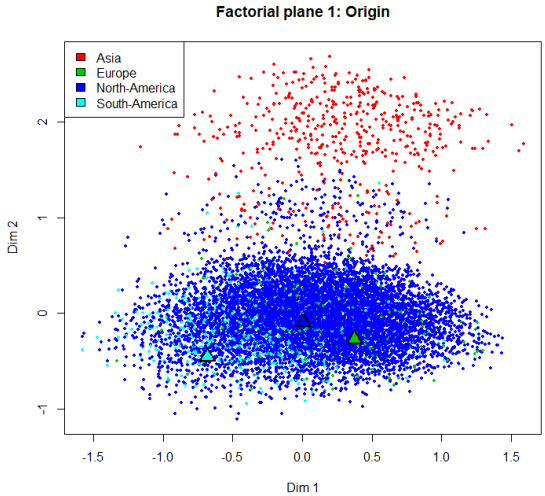


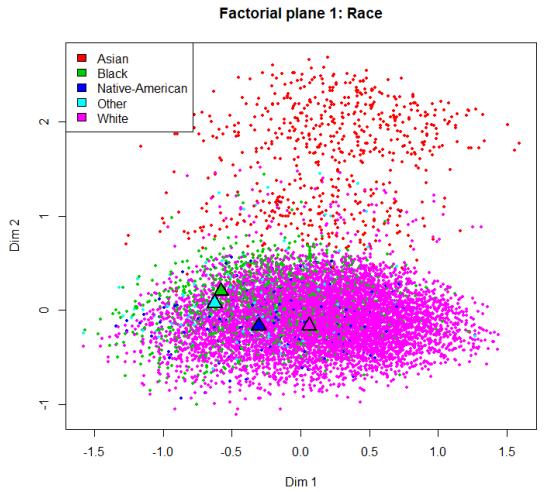
The distribution of jobs is more or less correlated with the education of the individuals, since the more qualified jobs tend to be at the right and the less qualified at the left. It is important to notice than the Non-qualified jobs are the ones placed more at the left. The Employer plot is very consistent with the Job plot, being the governmental very close to the Administration jobs at the top right, and the None Employer highly overlapped with the Non-qualified at the bottom left. Self-employees are placed at the bottom left, overlapping with almost all kinds of jobs. (except for the Administration).





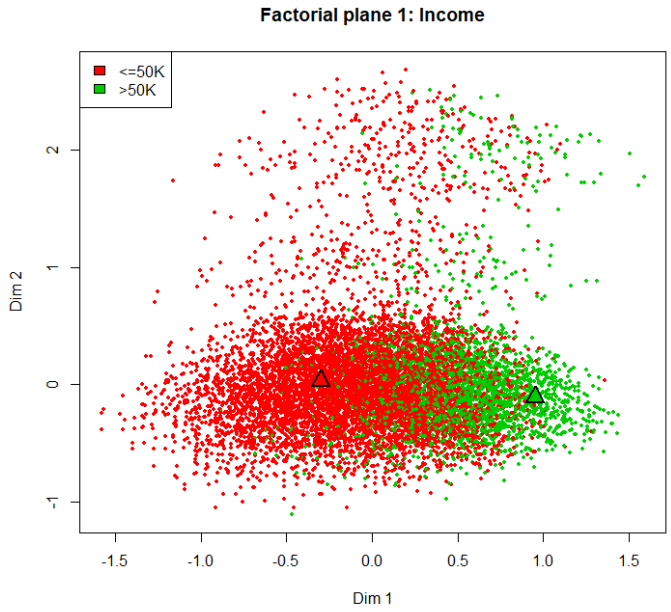
As for the working hours, we can see that there is a huge mass of instances working 40 hours per week, which is the mandatory full-time amount of hours to work. On the other side, people working more hours are distributed to the right of the clouds, and people working less hours are distributed to the left of the cloud. In terms of Capital gains and losses, the vast majority of the instances have zero capital, and are placed at the centre of the cloud. Only instances with losses or high gains are placed to the right of the cloud.





In these two plots, we can finally understand the meaning of the three differentiated clouds, which internally have all the same structure so far on. In terms of Origin, the upper cloud is made of people coming from Asia, whereas the bottom cloud is made of people from North-America, South-America or Europe. The middle cloud is mixed with people from North-America and Asia. If we compare it with the Race, we can see more or less the same: White, Black and Native-American are placed in the bottom cloud, whereas Asian people are placed in the upper cloud. In the middle cloud, there is a mix of White and Asian people.

If we look carefully at the data samples placed in the middle cloud, we can notice that, in that cloud, all the people from Asia has not Asian race, and all the Asian people are not from Asia. This means that the middle cloud contains all the people for which the Origin and the Race is not "consistent" in the sense that it is not assigned according to the rest of the population. People being there might be, for instance, American people descending from Asian people, or people from Asia descending from Europeans, or this kind of casuistry.

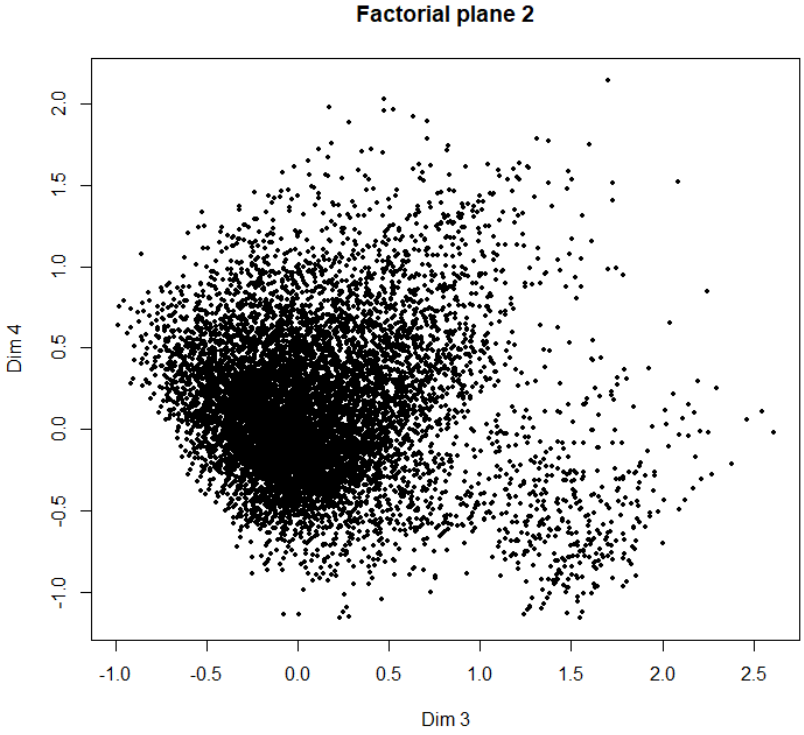


Finally, we have the Income plot. This plot is very interesting, since the Income was provided to the MCA as a supplementary variable, and hence the projection is not aware of any of the information this variable provides. Even though, we can clearly see that people earning more than 50K dollars tend to get to the right of all clouds, and the rest of the people gets grouped towards the left. It is important to recall there is a huge area of overlapping, since the central part of the cloud contains most of the data samples and has people of the two categories.

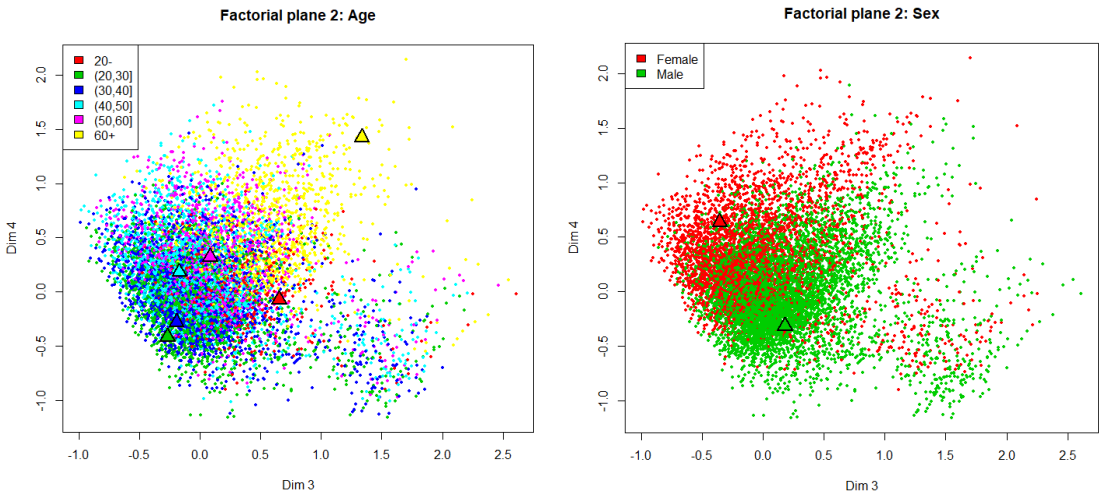
After reviewing all the plots of the first factorial plane, we can conclude that it is representing quite well all the variables of the dataset, but grouping at the right the modalities that tend to earn better salaries (work more hours, better jobs, more studies…) and grouping at the left the ones that earn lower salaries (work few or any hours, less studies, non-qualified jobs and/or not currently working…). Also, people coming from Asia is separated from the rest of the Origins, with Asian people at the top and the rest at the bottom. In the middle remain those instances which have non-consistent Origin and Race, like Asian people born in North-America and White people born in Asia.

4.2. Second Factorial Plane

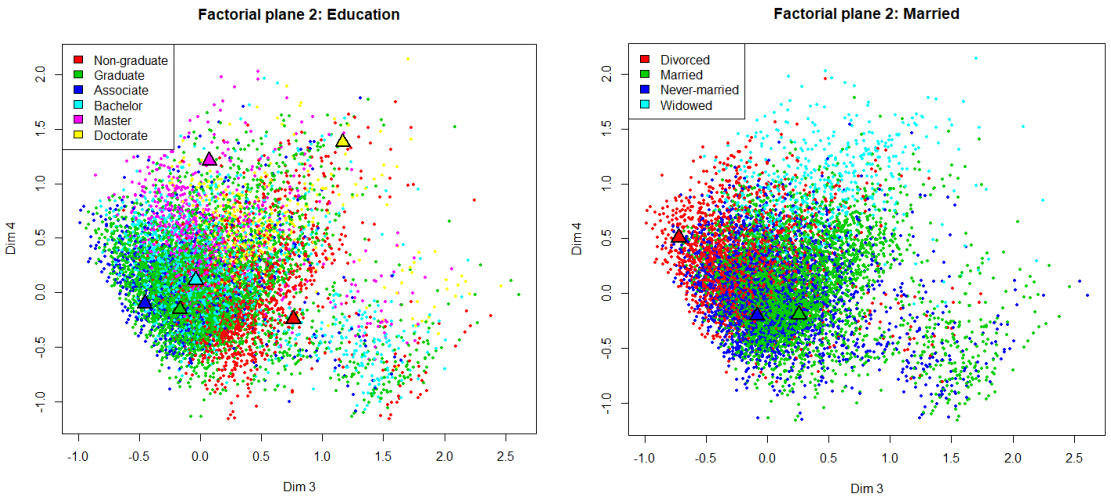
The results of the second factorial plane can be seen in the following plots:



A first visual inspection on the factorial plane leads to a clear identification of two clouds of points: There is a very big, very dense cloud at the top left and a sparser cloud of points at the bottom right. Its interpretation will be considered in the latent variables section.

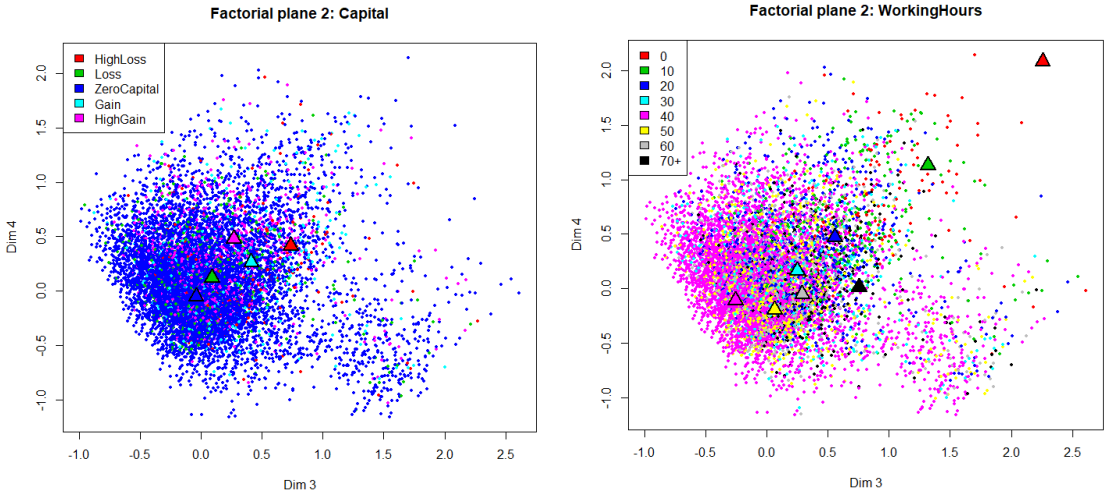


In the Age plot, we can see that Age levels are highly overlapped, except for the 60+ modality which is at the top-right corner of the cloud. In the Sex plot, we can still see a great separation between Males and Females, although they are more overlapped than in the first factorial plane.

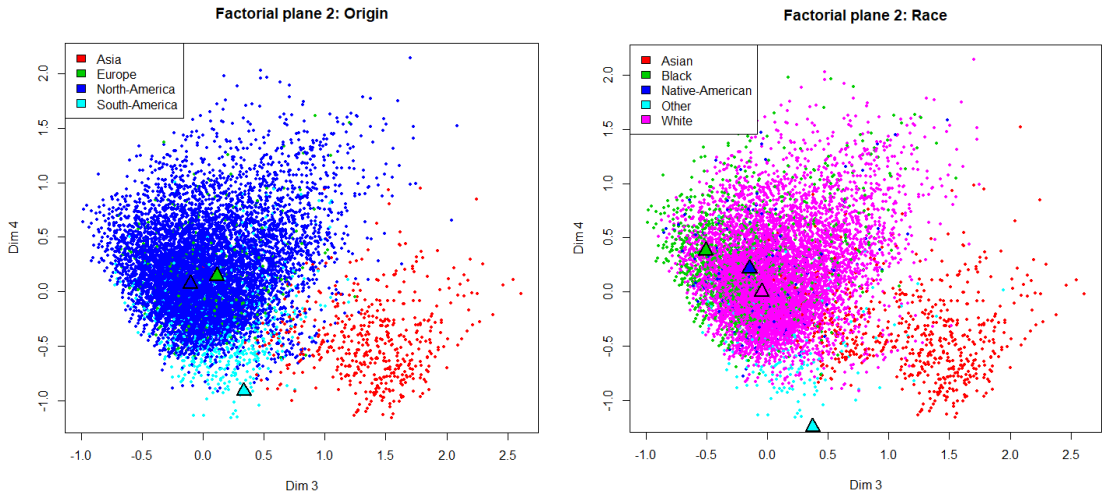


Education is again highly overlapped, even though Non-graduate can be more distinguished than the rest of the modalities. Married modalities can clearly be distinguished, although they appear quite overlapped again.

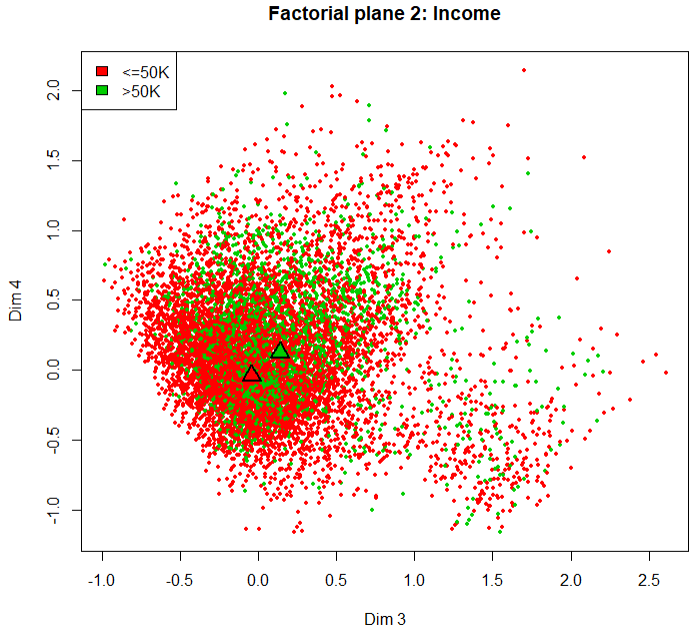
Again, both Employer and Job are highly overlapped, with some modalities practically indistinguishable.



Capital and Working Hours appear completely indistinguishable in this factorial plane.



It seems that the Origin is the most represented variable in this factorial plane. Asian and South-American people are almost completely separated from European and North-American people, which appear overlapped. Race is also very clear in this plot, with Black, White and Native-American completely overlapped and Asian and Others separated from them.



Finally, we have the Income plot. It is very different from the first factorial plane Income plot, since in this case both income categories do completely overlap.

After reviewing all the plots of the second factorial plane, we can conclude that it does not represent all the variables as well as the first factorial plane. It stills allows us to differentiate between some particular modalities not very well represented in the first factorial plane (like Age over 70 or non-graduate Education), but it is non-representative of most of the variables. It is only interesting in terms of Origin and Race, in which it complements the information provided by the first factorial plane.

5. Latent Concepts

The latent concept interpretation is performed in terms of the latent variables obtained from the MCA analysis. In this analysis, we will consider the 4 selected components of the MCA performed on the training data, and we will try to interpret and explain them.

For all of the components, only variables with R2 coefficient greater than 0.25 are considered significant. Hence, only the modalities of the significant variables are shown.

5.1. First component

When the first component is analysed, we found the following significant variables:

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

Then, for each of those variables, we can check how its modalities distribute:

Variable	Modality	Value
Age	20-	-0.8008
	(20,30]	-0.1411
	(30,40]	0.2577
	(40,50]	0.3964
	(50,60]	0.3103
	60+	-0.0225
Education	Non-graduate	-0.6156
	Graduate	-0.3260
	Associate	-0.0996
	Bachelor	0.0904
	Master	0.3416
	Doctorate	0.6094
Married	Divorced	0.0966
	Married	0.4375
	Never-married	-0.2621
	Widowed	-0.2720
Job	Administration	-0.3585
	Executive	0.3245
	Non-qualified	-0.4282
	Professional	0.3479
	Protective	0.1876
	Sales	-0.1324
	Tech	0.0590

Variable	Modality	Value
WorkingHours	0	-0.3531
	10	-0.4966
	20	-0.4644
	30	-0.3253
	40	0.1232
	50	0.4867
	60	0.5658
	70+	0.4636
Income	<=50K	-0.3133
	>50K	0.3133

Here, we can see that ages between 30 and 60 have the greatest values in this variable, and they are the ages in which typically people achieve its greatest professional success. This is consistent with the highest levels of education holding greater values in this axis.

In this variable, married people achieve the better punctuation. It is important to recall that married people can share expenses and therefore are more likely to have greater income all together. Also, the best qualified jobs get higher punctuations in this variable.

As for the working hours, the values of the variable are clearly proportional to the number of worked hours. It seems logical to think that people working more hours (like executives and professionals) have higher incomes. And finally, the greater income modality has positive values on this axis, which is consistent with all the previous observations.

Therefore, we can conclude that the first component is a latent variable that measures **how successful a certain individual is**, based mainly in its age, education, marital status, current job, amount of working hours and income.

5.2. Second component

When the second component is analysed, we found the following significant variables:

Variable	R2
Origin	0.4782
Race	0.5022

Then, for each of those variables, we can check how its modalities distribute:

Variable	Modality	Value
Origin	Asia	1.4391
	Europe	-0.4778
	North-America	-0.4043
	South-America	-0.5570
Race	Asian	1.3346
	Black	-0.2427
	Native-American	-0.3986
	White	-0.3968

Here, we can see that people either coming from Asia or people being of Asian race score higher punctuations on this axis, whereas the rest of origins and races score virtually the same. Notice than people with is both Asian from Asia score the maximum punctuation, whereas there might be a mixed group of Asian people not coming from Asia or people from Asia not being Asian which will score a lower (but still positive) value. All the other combinations of origin and race will score virtually the same.

Therefore, we can conclude that the second component is a latent variable that measures **if a certain individual is Asian, whether being born there or being descendant of Asians**. This might make a difference with the rest of origins and races, since they have a much more different culture and way of work than the rest of the countries and races. It is also the case than the cultural differences between Europeans, North-Americans and South-Americans is not as big as the difference with all of them and the Asian people.

5.3. Third component

When the third component is analysed, we found the following significant variables:

Variable	R2
Origin	0.2864
Race	0.2856

Then, for each of those variables, we can check how its modalities distribute:

Variable	Modality	Value
Origin	Asia	0.9870
	Europe	-0.3292
	North-America	-0.4203
	South-America	-0.2373
Race	Asian	0.9910
	Black	-0.4244
	Native-American	-0.2749
	White	-0.2329

Here, we can see that the table is almost identical to the second component, except for a variation of the distribution of the non-asian modalities: now South-America and Black have higher values than the other non-asian modalities. Anyway, all the magnitudes of these variables are much lower in this axis than they were on the second component.

Therefore, we can conclude that the third component is not a latent variable by itself, but it is more likely a complement to the second component. They together may define a factorial plane which effectively represents **all the information about the origin of each individual**.

5.4. Fourth component

When the fourth component is analysed, we found the following significant variables:

Variable	R2
Married	0.2541
Job	0.3249

Then, for each of those variables, we can check how its modalities distribute:

Variable	Modality	Value
Married	Divorced	-0.0481
	Married	-0.3339
	Never-married	-0.3373
	Widowed	0.7194
Job	Administration	0.1555
	Executive	0.0170
	Non-qualified	-0.0780
	Professional	0.3758
	Protective	0.1344
	Sales	-0.1772
	Tech	-0.4276

In this variable, Widowed people achieve the better punctuation, whereas Divorced people get an almost 0 punctuation and both Married and Never-married get virtually the same punctuation. This may be trying to separate people who receive external economical help (like widows and some of the divorced people) and people who do not. It makes sense for divorced people to become neutral, since half of the divorced people might receive some economical help from the other partner, whereas the other half are the ones who pay that help. This makes that group to get, in average, neutral on this axis.

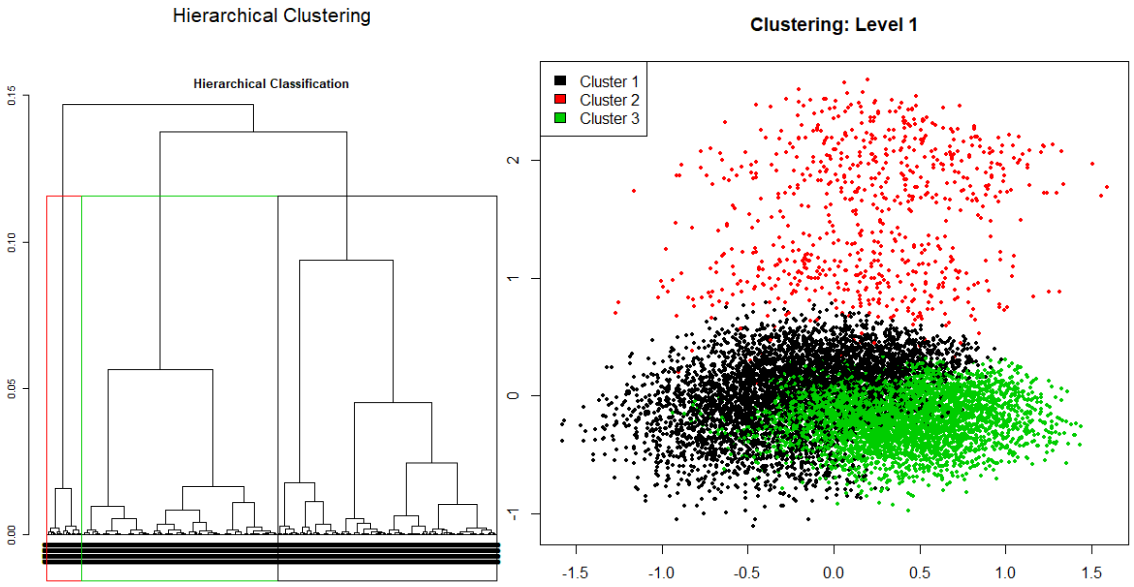
Also, Professional jobs achieve better punctuation, whereas Administration and Protective get moderate punctuations, Executive and Non-qualified are almost 0, Sales is moderately negative and tech is highly negative. This might be trying to separate the kinds of job to whereas they are more likely to receive any help from the state in terms of economical subvention.

Therefore, we can conclude that the fourth component is a latent variable that measures **how likely a certain individual is to receive any economical help**, based mainly in its marital status and current job.

6. Clustering

After the **visualization** and the **latent variable analysis**, clustering has been performed in the dataset. In a certain way, it can be considered a **latent class analysis**, since the obtained clusters are expected to correspond to some latent classes present in the dataset.

Clustering will be performed as follows: First, a Hierarchical clustering will be performed, obtaining a cluster tree. From that tree, two levels of clusters will be extracted, and they will be processed from down to top. In the bottom level of the clustering, k-means consolidation is applied after the hierarchical clustering. When performed, the hierarchical clustering resulted in the following trees:



The first level resulted into three clusters, one of them containing our previously discovered "Asian" clouds, and the other two resulting from the split of the "USA" cloud. It achieved a Calinski-Harabasz index of 7200. The second level divided the "USA" cloud into even more clusters, thus obtaining more specific classes from that group of people. It achieved a Calinski-Harabasz index of 8154.

7. Clustering Interpretation

In order to perform **profiling** over the found clusters to interpret them, we will see which variables do outstand in each class. First, we will interpret the upper clustering, and then will expand the information about its sub-clusters. Particularly, since cluster 2 in the first cut is exactly the same as cluster 5 in the second cut, it will not be described again.

Then, for the first cut, we can list for each cluster which are the most relevant variables.

Variable	Cluster 1	Cluster 2	Cluster 3
Age	Less than 30 or more than 60	Between 20 and 40	Between 30 and 60
Sex	Female		Male
Education	Non-graduate, Graduate	Bachelor, Master, Doctorate	Associate, Doctorate
Married	Never-married, Divorced, Widowed	Never-married, Married	Married
Job	Administration, Gov, Non-qualified, Professional	Professional	Tech, Executive, Sales, Protective
Employer	None		Self, Private
Working Hours	Less than 30	40	More than 50
Origin	North-America	Asia	North-America, South-America, Europe
Race	Black, Native-American	Asian	White
Capital	Zero		High Gain, Loss, High Loss
Income	<=50K		>50K

Here, we can clearly see that cluster 2 corresponds to the **Asian** individuals that got higher punctuation in the Asian latent variable, since they are the only cluster containing both Asia and Asian, and graphically it can be seen that they compound the two upper clouds of points. Asian people tend to get higher studies and professional jobs, together with regular working hours and. On average, they are between 20 and 40 and they do not get divorced nor widowed.

As for cluster 1, it corresponds to the **Low-Class** profile, mainly holding younger and older people with more difficulties to get a job. Also, lower level of studies shows up, and worse jobs appear. They typically work less than 30 hours, live alone and have very low incomes.

Cluster 3, on the other side, corresponds to the **High-Class** profile, mainly holding white man between 30 and 60, with high level of studies and working lots of hours in well-paid jobs. They are usually married and have a high income level.

For the second cut, we can also list for each cluster which are the most relevant variables. First, let's look at the clusters that arise from the **Low-Class** cluster:

Variable	Cluster 1	Cluster 2	Cluster 3
Age	Less than 20 or more than 60	Less than 30	Between 20 and 50
Sex	Female	Female	Female
Education	Non-graduate	Graduate	Bachelor, Associate, Master
Married	Never-married, Widowed	Never-married	Divorced, Widowed, Never-married
Job	Non-qualified	Non-qualified, Sales, Administration	Professional, Administration, Executive, Protective
Employer	None	Private	Gov
Working Hours	Less than 30	Between 20 and 30	40
Origin	South-America	North-America, South-America	North-America
Race	Black	Black, Other	Black
Capital	Zero, Gain	Zero	Zero
Income	<=50K	<=50K	<=50K

Cluster number 1 corresponds to the **lowest class** of the society, single people with no studies nor qualified jobs which work less than 30 hours. It includes both people with less than 20 years old or more than 60 years old. It also includes mainly people from South-America.

Cluster 2 corresponds to **Middle-Class low-qualified sector**, holding people with a bit more studies and better jobs, but still not the best ones. Working hours are higher on average, and age is lower than 30. It contains both North-American and South-American people.

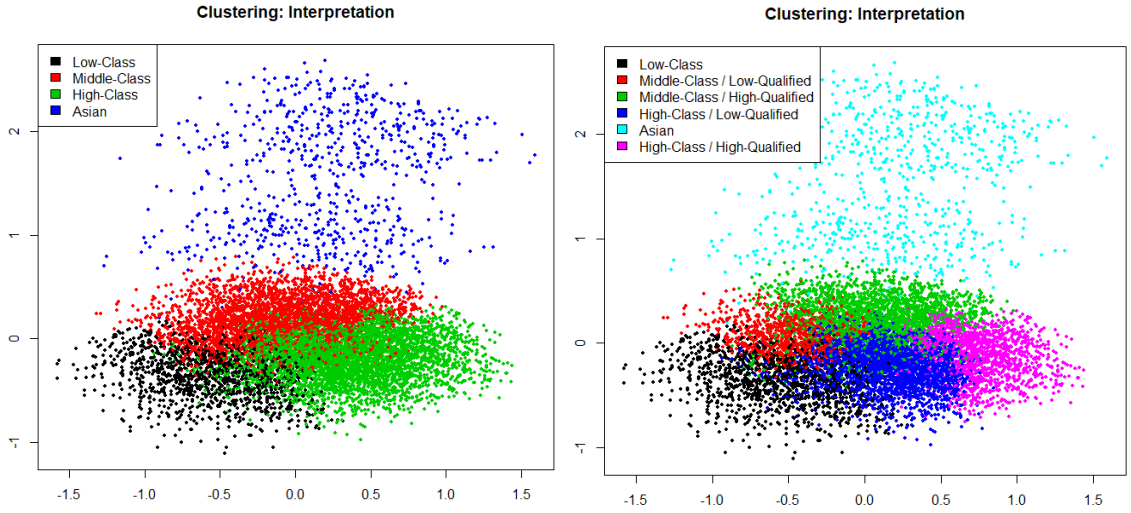
Cluster 3 corresponds to **Middle-Class high-qualified sector**, holding people with higher studies working in much better jobs for the government. Age is more uniformly divided between 20 and 50, with an average of 40 working hours per week. It is mainly made of North-American people.

And now, let's look at the clusters that arise from the **High-Class** cluster:

Variable	Cluster 4	Cluster 6
Age	Between 30 and 60	Between 30 and 60
Sex	Male	Male
Education	Non-graduate, Graduate, Associate	Bachelor, Master, Doctorate
Married	Married	Married
Job	Tech, Sales, Non-qualified	Professional, Executive, Protective
Employer	Private, Self	Self, Gov
Working Hours	More than 40	More than 50
Origin	South-America, Europe	North-America, Europe
Race	White, Other	White
Capital	Zero, Gain	High Gain, Loss, High Loss
Income	>=50K	>=50K

Cluster 4 corresponds to the **High-Class low-qualified**, where people achieve regular jobs but high incomes. Cluster 6 corresponds to the **High-Class high-qualified**, which usually have the better level of studies and jobs. In fact, these two clusters are much more similar than any other pair of clusters, since its main features are basically the same.

Based on the clustering tree and in the interpretation of the clusters, we can conclude that the most significant group is the **Asian** group, which even if it is very small, is very different from the rest of the population. Apart from that, population can be divided into **High-Class**, **Middle-Class** and **Low-Class**, based on its education, job and income. Both High-Class and Middle-Class can be divided depending on the **quality of their jobs**.



8. Sample validation

In order to generalize the previous and following results properly, it is important to check if the sample taken becomes from the same distribution as the whole population. Since this is statistically impossible, we are going to check if both **train** and **test** splits have the same distribution, that is, if they belong to the same population.

In order to test if both splits have the same distribution, we have to look for the differences of the test sample respect to the training one. In this case, since all variables are categorical, two tests are carried out:

1) Chi-squared test over all variables, individually.

2) Chi-squared test over all variables, on a single test.

The first test consisted on applying a chi-squared test over the contingency table of each of the variables. The variations of the distribution of the contingency values are assumed to be significant only if its p-value is less than 0.05, otherwise they are considered to hold the same distribution. The results are shown in the following table:

Variable	p-value	Same Distribution?
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

Here, we can see that all the variables passed the test. This means that, individually, all the variables of the data set have the same distribution in the train and in the test sets.

In order to check if this homogeneity on the distribution is valid for all the variables all together, we performed a test over the combined contingency table of all the variables. Since the p-value of the test turned out to be 0.868, we accept that both sets have the same distribution.

Therefore, we can conclude that the test individuals can be taken as a sample of the training ones, since both train and test splits follow the same distribution.

9. Modelling

In order to predict the Income of a certain individual based on the rest of its features, a statistical model will be built from the available training data. Then, this model will be evaluated on the test data, which has never been seen during the training.

As for the model, **Random Forest** is chosen, due to being a very robust technique which has been proved successful in many challenging tasks. It is also a very useful technique in this case, since in this problem all the variables are categorical and there are not many models which are able to handle categorical variables directly.

A Random Forest is an ensemble of several **Decision Trees**, which are a well-known statistical model that has also been studied for lots of years. Since Decision Trees tend to overfit the data, Random Forest can compensate that and lead to more near-optimal accuracies. In this project, an ensemble of 1000 trees is considered, whereas the rest of the parameters are optimized automatically.

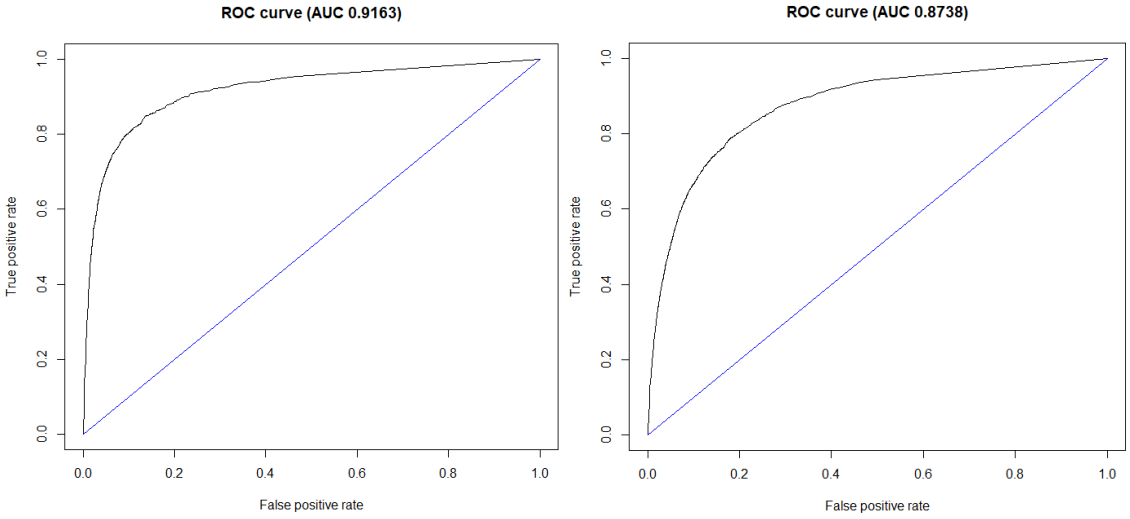
The model is therefore trained with the training set of data, whereas the test set is provided as supplementary. After training, the following confusion matrices are obtained for both sets, together with its corresponding metrics:

Train Set		Predicted	
		<=50K	>50K
Actual	<=50K	17259	1297
	>50K	2424	3441

Test Set		Predicted	
		<=50K	>50K
Actual	<=50K	17322	1277
	>50K	2404	3418

	Train Set	Test Set
Accuracy	0.8893	0.8494
Precision	0.8291	0.7280
Recall	0.6791	0.5879
AUC	0.9163	0.8738

Also, ROC curves for both sets can be computed (train at the left, test at the right):



The previous confusion matrices can be directly compared, since the train sample size is the same as the test sample size. In this case, we see that the model works better for the majority class in the test sample than in the train sample, whereas for the minority class it works the opposite way.

In general, we can state that the model works better in train sample than in the test sample, since its accuracy is higher in the first sample (89%) than in the second sample (85%). This is somewhat expected, since the training algorithm can actually see the training sample when training, whereas the test sample was unknown to him.

Even though, the accuracy difference is only of around 4%. This is not considered to be a problem of **overfitting**, since the difference is not that big in a problem in which we are talking about a lot of data and a ≈15% test error.

Precision and Recall have also been observed to be higher in the train set (83% / 68%) than in the test set (73% / 59%). This means that the model is more confident for the positive class in the train set than it is in the test set. This goes according to the confusion matrices observed, in which the model performed better for the positive class in the train than in the test set.

AUCs are quite similar in both cases, with only a 4% gap between them (91% on train, 87% on test). This is approximately the same gap that exist between the train and test accuracies, and hence it is consistent with the previously obtained results. AUC measures around 0.9 are actually very good, and mean that the classifier is in fact quite informative (an AUC of 0.5 would mean that the classifier is complete uninformative and is not working at all).

This can be confirmed with the plots, in which the ideal curve would be a straight line from (0, 0) to (0, 1) and then to (1, 1). Therefore, we can conclude that our curves are quite good, since its shape is near to the squared one and far away from the diagonal one (which would be a non-informative classifier).

In general, the obtained results can be considered to be quite good. The proposed task on this dataset is actually quite difficult, with lots of exceptions and contradictory information (people with very similar features might actually belong to different classes). Therefore, 85% accuracy seems a good result, and all the other metrics agree with that.

The classifier model seems quite stable, and very robust to noisy samples. Also, it is not very affected by the fact that the dataset is unbalanced, since the behaviour on the negative class is neither that bad.

As a conclusion, we can state that the classified performed quite well given the difficulty of the task, and that its results are quite reliable and stable.

10. Conclusions

After performing all the proposed tasks over the Adult dataset, we could obtain very interesting information about the population of the United States in 1994. First, we have been able to analyse individually all its features, and perform a significant aggregation of the different modalities available, which were in fact a bit too much detailed.

This cleaning over the input data allowed us to properly treat all the missing values: Some of them have been interpreted and given a proper value, whereas the rest of them have been imputated by a specialized algorithm.

In order to test our further insights, the dataset was properly divided into a train and a test partitions. The train partition was used to perform the exploration, whereas the test dataset was used to validate the knowledge extracted from that exploration.

First, the dataset was properly visualized, by performing a projection into the first two factorial planes. This was done by keeping the first four components from the MCA analysis over the dataset.

Then, two kinds of latent analysis were performed. The first, known as **latent variable analysis**, aims to find interesting latent variables in the dataset, like the obtained **successfulness**, **asian** and **economical help** variables. The second, commonly known as **clustering** (although it can be considered a **latent class analysis**), aims to find interesting classes or categories in the dataset, like the obtained **high-class**, **middle-class**, **low-class** and **asian**. Also, it is interesting to notice that the cluster analysis revealed that both **middle-class** and **high-class** can be divided into further subclasses, based on its jobs and the level of studies.

After that, in order to prove that the previous analyses can be generalized to the general population, some statistical tests were performed in order to check if the train and the test sets share the same distribution, and hence they could be considered to belong to the same population. As it was shown, they can effectively be considered to belong to the same population.

Finally, a model was trained to perform the proposed task. The chosen model was a **Random Forest**, because it has very founded theoretical properties, works specially well with categorical data and is able to work with lots of data as we had in our dataset. The trained model achieved very good on the test dataset, given the difficulty of the task and the amount of data available.

Therefore, all the required tasks have been performed successfully, providing us a lot of information and insights from the Adult dataset. It has also been very interesting to work with a dataset so interpretable, because we have learned to work with a more real-world oriented project. It has also helped us a lot in understanding the techniques and algorithms used in the multivariate analysis, and how to take them into practice. It was a very interesting project we enjoyed a lot and in which we all learned lots of interesting concepts we will surely be using in our future life.