Statistical learning:
supervised and unsupervised analysis.

Chiara Saini

Index

Abstract

The following analysis applies supervised learning tecniques on the 1994 Census bureau dataset and unsupervised learning techniques on a mall's customers dataset.
The supervised learning algorithms were K nearest neighbor, logistic regression, tree prediction, tree predictor with Gini index, Naïve Bayes, and random forest. All the algorithms were compared according to their accuracies.
The unsupervised learning algorithm used was k-means. In order to choose the number of k, the elbow method, the average silhouette method, and the gap statistic were used.

1. Supervised learning

1.1 Data Exploration

The data used in this analysis were extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker. The dataset is composed of 15 attributes and 32.561 individual records. The attributes are below explained:

Age: continuous and numeric variable, it tells the age of the individuals taken into consideration for the analysis, it goes from a minimum of 17 to a maximum of 90, the mean and the median are respectively 38,58 and 37;

Work class: categorical variable**,** describes the type of employer of the individuals. It is divided as represented in *table 1;*

Table 1: Number of individuals per work class

Work class	N° of individuals
Government	4.289
Private	22.286
Self-employed	3.573
Other	14

Final weight: continuous and numeric variable, the weight assigned by the US census bureau;

Education: categorical variable, level of education of each individual. It is divided as represented in *table 2*;

Table 2: Number of individuals per education level

Education	N° of individuals
Assoc-acdm	1.008
Assoc-voc	1.307
Bachelors	5.044
Doctorate	375
HS-grad	9.840
Masters	1.627
not-HS-grad	3.741
Prof-school	542
Some-college	6.678

Education numeric: continuous and numeric variable, represents the number of years of education. It goes from a minimum of 1 to a maximum of 16, the mean and the median are respectively 10,08 and 10;

Marital status: categorical variable describes the marital status of the individual. In *table 3***,** it is possible to see a representation of how this attribute is divided;

Table 3: Number of individuals per marital status

Marital status	N° of individuals
Married	14.456
Not married	5.980
Never married	9.726

Occupation: categorical variable tells the occupation of the individual. In *table 4,* it is possible to see a representation of how this attribute is divided;

Table 4: Number of individuals per occupation

Occupation	N° of individuals
Armed-Forces	9
Office	7.713
Professional	4.038
Sales	3.584
Service	4.911
Worker	9.907

Relationship: categorical variable, describes the current relationship. In *table 5***,** it is possible to see a representation of how this attribute is divided;

Table 5: Number of individuals per relationship status

Relationship	N° of individuals
Husband	13.193
Not in family	8.305
Other relative	981
Own child	5.068
Unmarried	3.446
Wife	1.568

Race: categorical variable tells the race of the individual. In *table 6***,** it is possible to see a representation of how this attribute is divided;

Table 6: Number of individuals per race

Race	N° of individuals
American Indian, Eskimo	31
Asian Pacific, Islander	1.039
Black	3.124
White	27.816
Other	271

Sex: categorical variable, tells the sex of the individual (*table 7*);

Table 7: Number of individuals per sex

Sex	N° of individuals
Female	10.771
Male	21.790

Capital gain: continuous and numeric variable, represents the capital gain made by the individual, goes from a minimum of 0 to a maximum of 99.999. The mean and the median are respectively 1.078 and 0;

Capital loss: continuous and numeric variable, represents the capital loss made by the individual, goes from a minimum of 0 to a maximum of 4.356. The mean and the median are respectively 87,3 and 0;

Hours per week: continuous and numeric variable, represents the working hours per week of the individual, goes from a minimum of 1 to a maximum of 99. The mean and the median are respectively 40,44 and 40;

Native country: categorical variable, tells the native county of origin of the individuals (*table 8*);

Table 8:Number of individuals per native country

Native country	N° of individuals
North America	27.720
South America	1.230
Asia	634
Europe	493
Other	85

Income: categorical variable, represents the income level of the individuals. It can be at least 50.000$ or bigger than 50.000$.

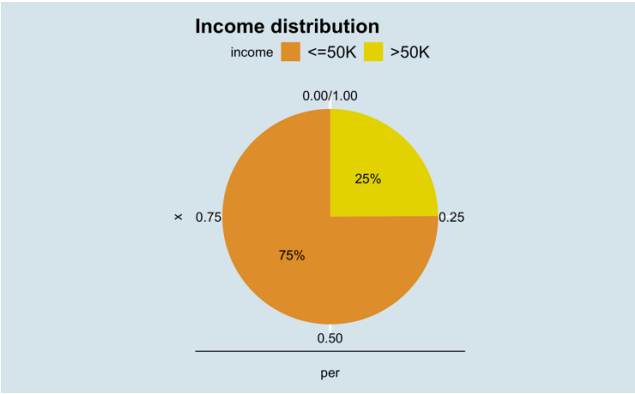
Note that some attributes were grouped to guarantee a better analysis.
In *table 9***,** it is possible to see the summary of the entire dataset.

Table 9: Summary of the dataset

Attribute	vars	n	mean	sd	min	max	range	se
age	1	30162	38.437902	13.1346648	17	90	73	0.0756291
workclass	2	30162	-	-	-	-	-	-
fnlwgt	3	30162	189793.834	105652.972	13769	1484705	1470936	608.347389
education	4	30162	-	-	-	-	-	-
education.num	5	30162	10.1213116	2.54999492	1	16	15	0.01468281
marital.status	6	30162	-	-	-	-	-	-
occupation	7	30162	-	-	-	-	-	-
relationship	8	30162	-	-	-	-	-	-
race	9	30162	-	-	-	-	-	-
sex	10	30162	-	-	-	-	-	-
capital.gain	11	30162	1092.00786	7406.3465	0	99999	99999	42.6455734
capital.loss	12	30162	88.3724886	404.29837	0	4356	4356	2.32794075
hours.per.week	13	30162	40.931238	11.9799842	1	99	98	0.06898047
-tive.country	14	30162	-	-	-	-	-	-
income	15	30162	-	-	-	-	-	-

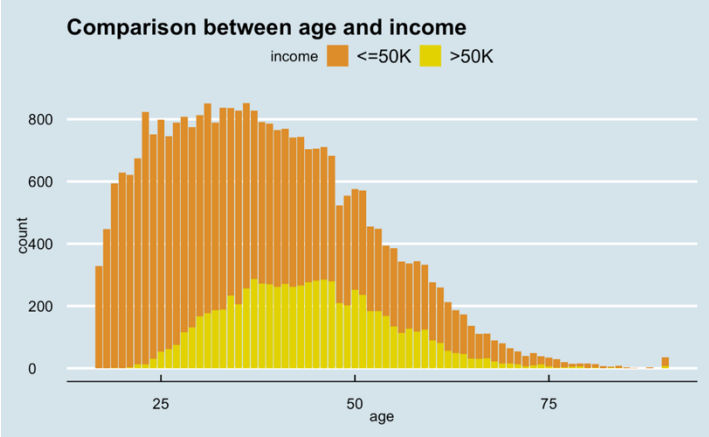
In *figure 1***,** it is possible to see the income distribution. The label “<=50K” represents income that is at least 50.000$ (75% of the total) and the label “>50K” represents income higher than 50.000$ (25% of the total). Income is the target variable and is unbalanced.

Figure 1: Income distribution



Below, the most relevant variables are analyzed in more depth regarding income.
In *figure 2*, it is possible to see the comparison between age and income. The age distribution is skewed to the left; there is a higher concentration of young adults. Nonetheless, the highest number of people with an income above $50k is between thirty and fifty years old.

Figure 2: Comparison between age and income



Given what is displayed in *figure 3*, the highest percentage of individuals work in the private sector. This is also the sector with the highest number of people earning more than $50k per year. At the same time, most of $50k earning people work in an office (see *figure 4*).

Figure 3:Comparison between work class and income

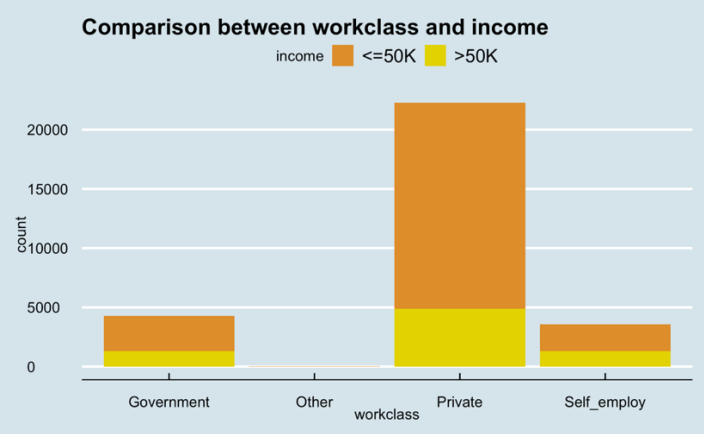
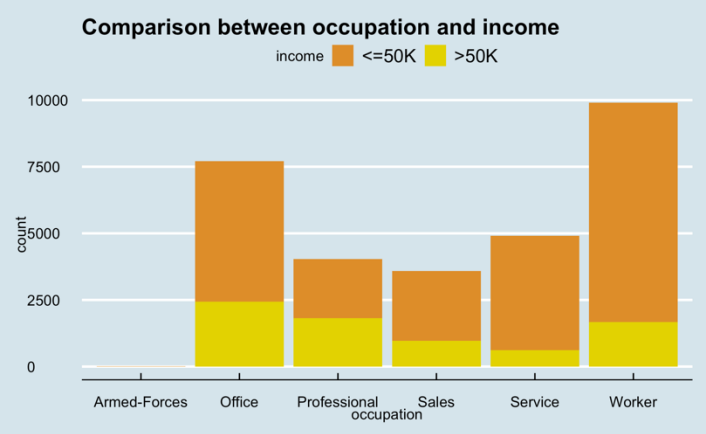


Figure 4: Comparison between occupation and income



Most people are high school graduates, and most $ 50k-earning people have a bachelor's degree. Most individuals studied for nine years, and the most significant number of $50k earning people studied for thirteen years (see *Figures 5 and 6)*.

Figure 5: Comparison between ezducation and income

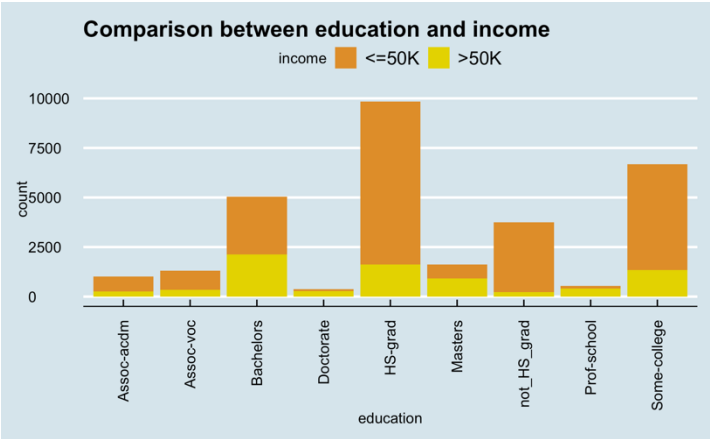
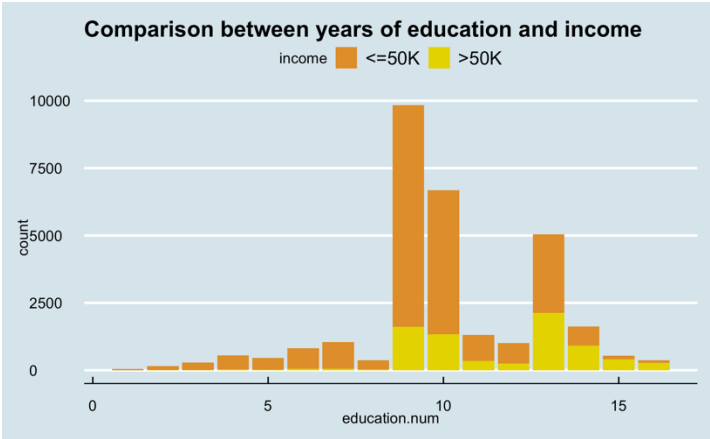


Figure 6: Comparison between years of education and income



In *figure 7*, it is possible to see the comparison between income and marital status. Clearly, people earning more than $50k are mostly married. The relationship status that favors a higher income than $50k is a husband.

Figure 7: Comparison between marital status and income

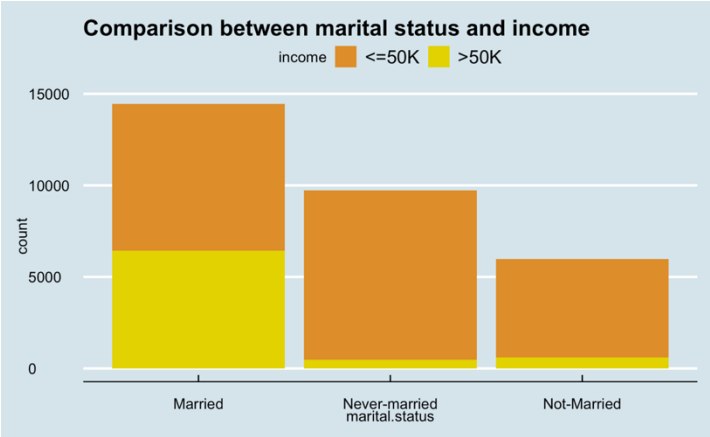
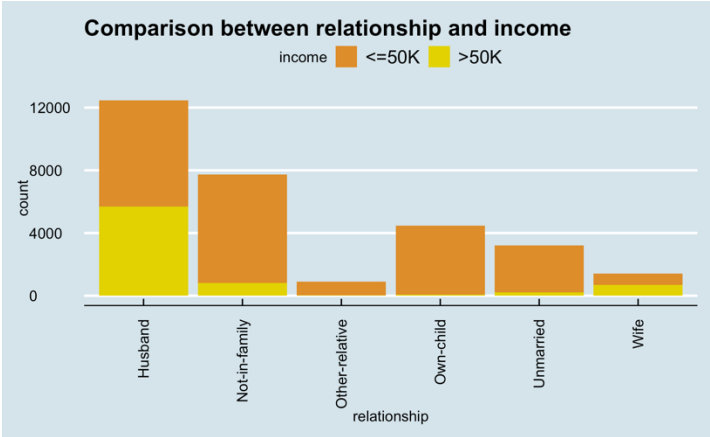
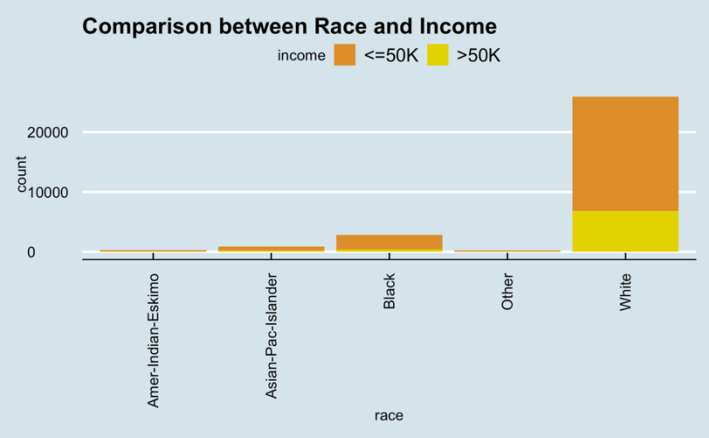


Figure 8: Comparison between relationship and income



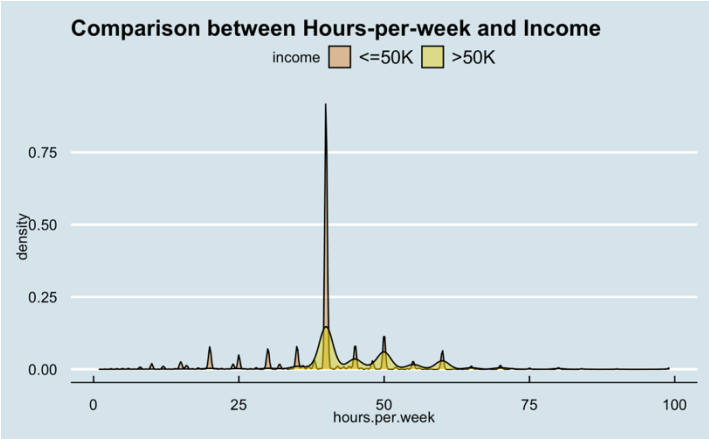
White people are more likely to earn more than $50k per year than people of other races, as shown in *figure 9*.

Figure 9: Comparison between race and income



As shown in *picture 10,* most people work the same amount of hours per week. This variable does not seem to influence the amount of income remarkably.

Figure 10: Comparison between hours of work per week and income



1.2 Splitting the data

In order to perform the analysis, the data is split into train and test. The training data take 70% of the information, and the remaining 30% is assigned to the test. A resampling was also done in order to balance the data. Both under-sampling and oversampling techniques were used.

1.3 K-nn

The first algorithm applied to predict the individuals' income is K nearest neighbor algorithm. K took different values in the analysis: 20, 10, and 5. In *image 11*, it is possible to see the results.

Figure 11: Confusion matricies of K-nn

K=20

```
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  6794 2197
     >50K     38   20

               Accuracy : 0.753
                 95% CI : (0.744, 0.7619)
    No Information Rate : 0.755
    P-Value [Acc > NIR] : 0.675

                  Kappa : 0.0052

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.009021
            Specificity : 0.994438
         Pos Pred Value : 0.344828
         Neg Pred Value : 0.755645
             Prevalence : 0.244999
         Detection Rate : 0.002210
   Detection Prevalence : 0.006410
      Balanced Accuracy : 0.501730

       'Positive' Class : >50K
```

K=10

```
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  6573 2124
     >50K    259   93

               Accuracy : 0.7367
                 95% CI : (0.7275, 0.7457)
    No Information Rate : 0.755
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0056

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.04195
            Specificity : 0.96209
         Pos Pred Value : 0.26420
         Neg Pred Value : 0.75578
             Prevalence : 0.24500
         Detection Rate : 0.01028
   Detection Prevalence : 0.03890
      Balanced Accuracy : 0.50202

       'Positive' Class : >50K
```

K=5

```
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  6203 2018
     >50K    629  199

               Accuracy : 0.7075
                 95% CI : (0.698, 0.7168)
    No Information Rate : 0.755
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0029

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.08976
            Specificity : 0.90793
         Pos Pred Value : 0.24034
         Neg Pred Value : 0.75453
             Prevalence : 0.24500
         Detection Rate : 0.02199
   Detection Prevalence : 0.09150
      Balanced Accuracy : 0.49885

       'Positive' Class : >50K
```

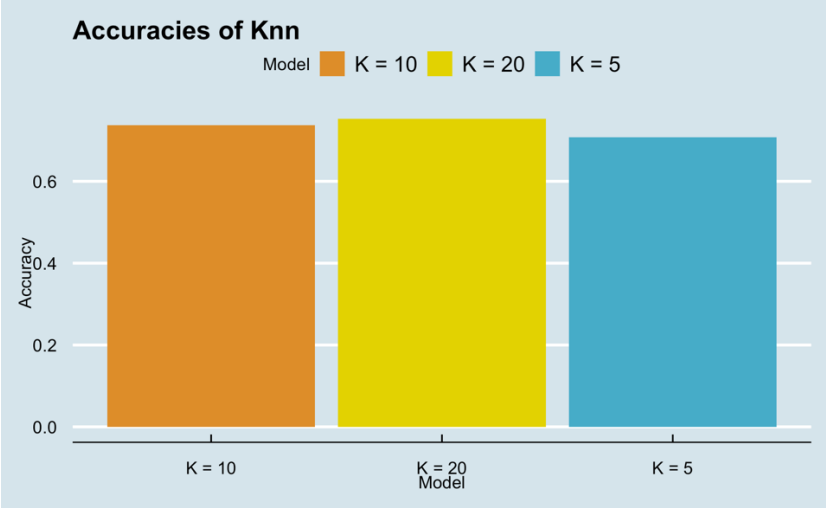
- **K = 20**: the model correctly predicted that 6.794 individuals earn less than 50k per year and 20 individuals earn more than 50k per year. The accuracy of the prediction is 0,753; the sensitivity is 0.009021 (the number of correct positive predictions divided by the total number of positives, an ideal model would have sensitivity = 1.0); the specificity is 0.994438 (the number of correct negative predictions divided by the total number of negatives, an ideal model would have specificity = 1.0);

- **K = 10**: the model correctly predicted that 6.573 individuals earn less than 50k per year (worse prediction compared to K = 20) and 93 individuals earn more than 50k per year (better prediction compared to K = 20). The accuracy of the prediction is 0,7367. This result is worse than the previous one; the sensitivity is 0,04195, which is better than the previous result; the specificity is 0,96209, a worse result than the previous;

- **K = 5**: the model correctly predicted that 6.203 individuals earn less than 50k per year (worse prediction compared to K = 10) and 199 individuals earn more than 50k per year (better prediction compared to K = 10). The accuracy of the prediction is 0,7075. This result is worse than the previous one; the sensitivity is 0,08976, which is better than the previous result; the specificity is 0,90793, a worse result than the previous;

In *figure 12*, it is possible to see the actual accuracies of the three K-nn approaches analyzed. The most accurate result was performed by K = 20.

Figure 12: Accuracies of K-nn



1.4 Logistic regression

Logistic regression is the second algorithm used in this analysis. Logistic regression is used in the case of binary classification. In this case, the outcome can be that the individuals earn more than $50k or less. In *picture 13***,** it is possible to see the summary of the logistic regression model.
Firstly, it is possible to see the formula of the model and some measure of deviance of the residuals. Under these results, the coefficients and their relative measures are explained. The intercept tells the log-odds of income being less than 50K with respect to the other variables.
For every attribute where p>0.05, the null hypothesis cannot be rejected[1].

[1] The null hypothesis for the logistic regression states that none of the predictor variables have a statistically significant relationship with the response variable, y (in this case income).

Figure 13: Summary of logistic regression

```
glm(formula = income ~ ., family = binomial(logit), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2102  -0.5411   0.0000   0.6404   2.7521

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -1.695e+01  1.779e+02  -0.095  0.92409
age                             3.029e-02  1.868e-03  16.217  < 2e-16 ***
workclassOther                 -1.267e+01  1.205e+02  -0.105  0.91622
workclassPrivate                7.546e-02  5.715e-02   1.321  0.18666
workclassSelf_employ           -2.118e-01  7.392e-02  -2.866  0.00416 **
fnlwgt                          1.036e-06  1.869e-07   5.542 2.99e-08 ***
educationAssoc-voc              2.173e-01  1.408e-01   1.543  0.12283
educationBachelors              5.552e-01  1.205e-01   4.608 4.07e-06 ***
educationDoctorate              6.107e-01  2.844e-01   2.147  0.03177 *
educationHS-grad                7.211e-02  1.800e-01   0.401  0.68874
educationMasters                7.249e-01  1.630e-01   4.446 8.75e-06 ***
educationnot_HS_grad           -2.218e-01  3.298e-01  -0.672  0.50133
educationProf-school            1.291e+00  2.474e-01   5.220 1.79e-07 ***
educationSome-college           2.978e-01  1.455e-01   2.046  0.04074 *
education.num                   1.962e-01  4.825e-02   4.066 4.77e-05 ***
marital.statusNever-married    -1.358e+00  1.704e-01  -7.971 1.58e-15 ***
marital.statusNot-Married      -7.136e-01  1.708e-01  -4.178 2.94e-05 ***
occupationOffice                1.091e+01  1.779e+02   0.061  0.95110
occupationProfessional          1.079e+01  1.779e+02   0.061  0.95162
occupationSales                 1.063e+01  1.779e+02   0.060  0.95234
occupationService               1.037e+01  1.779e+02   0.058  0.95351
occupationWorker                1.016e+01  1.779e+02   0.057  0.95444
relationshipNot-in-family      -8.769e-01  1.696e-01  -5.172 2.32e-07 ***
relationshipOther-relative     -1.253e+00  2.073e-01  -6.048 1.47e-09 ***
relationshipOwn-child          -1.792e+00  1.907e-01  -9.397  < 2e-16 ***
relationshipUnmarried          -1.088e+00  1.862e-01  -5.840 5.21e-09 ***
relationshipWife                1.259e+00  1.071e-01  11.761  < 2e-16 ***
raceAsian-Pac-Islander          6.073e-01  2.736e-01   2.220  0.02645 *
raceBlack                      -2.489e-01  2.275e-01  -1.094  0.27403
raceOther                      -8.546e-01  3.607e-01  -2.369  0.01783 *
raceWhite                       7.302e-02  2.145e-01   0.340  0.73360
sexMale                         1.026e+00  7.582e-02  13.538  < 2e-16 ***
capital.gain                    3.017e-04  1.209e-05  24.948  < 2e-16 ***
capital.loss                    7.318e-04  4.732e-05  15.464  < 2e-16 ***
hours.per.week                  3.621e-02  1.873e-03  19.328  < 2e-16 ***
native.countryEurope            7.127e-01  2.417e-01   2.948  0.00320 **
native.countryN.A               6.099e-01  2.000e-01   3.049  0.00229 **
native.countryRemaining_count -3.675e-01  4.153e-01  -0.885  0.37630
native.countryS.A               7.049e-02  2.388e-01   0.295  0.76781
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29268  on 21112  degrees of freedom
Residual deviance: 16768  on 21074  degrees of freedom
AIC: 16846

Number of Fisher Scoring iterations: 11
```

The p-value for the overall Chi-Square statistic of the logistic model is 0. Since $p<0.05$, the null hypothesis is rejected. Therefore there is a statistically significant relationship between the combination of the attributes and income. In *figure 14*, it is possible to see the confusion matrix that derives from the logistic regression.

Figure 14: Confusion matrix of logistic regression

```
      FALSE TRUE
<=50K  5401 1431
>50K    389 1828
```

In *table 10*, it is possible to see the evaluation measures deriving from the logistic model.

Table 10: Evaluation measures of logistic regression

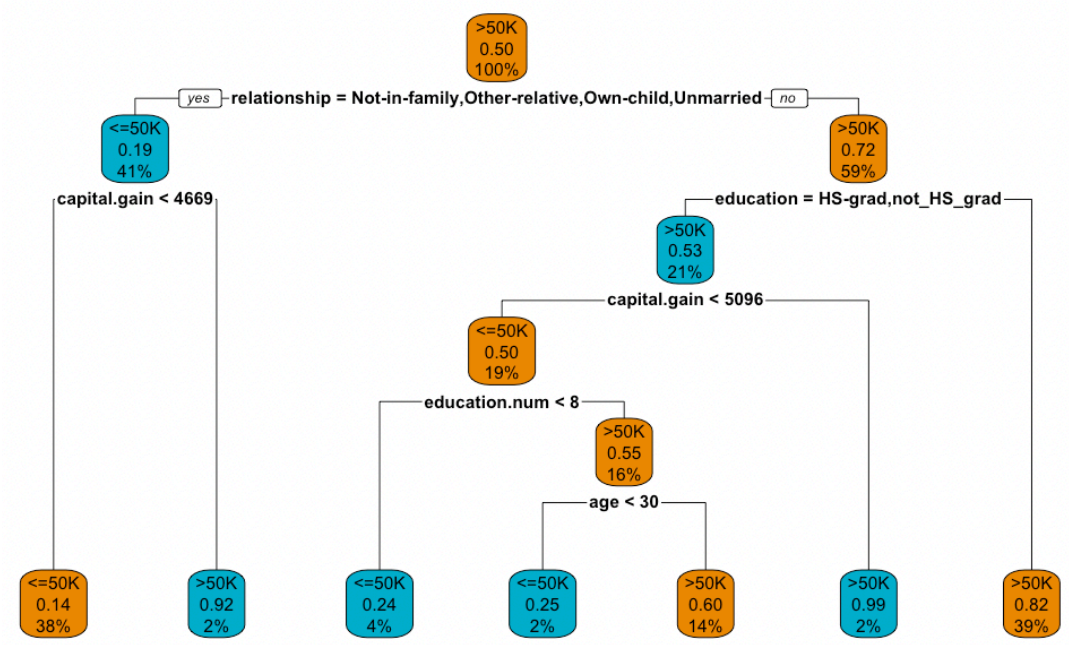
Accuracy	Sensitivity	Specificity	Precision	Error rate
0.7988728	0.5609083	0.9328152	0.8245377	0.2011272

Predicts the income of individuals with 0.7988728 accuracy. The sensitivity is 0.5609083, the specificity is 0.9328152, and the error rate is 0.2011272.

1.5 Tree predictor

The tree predictor was also used to predict the income of the individuals. The tree predictor can be used for both classification and regression problems. The target variable income was regressed on the entire dataset; the method used was class. The root of the tree selected by the algorithm was relationship status. In *figure 16*, it is possible to see the development of the tree.

Figure 15: Tree



In *figure 17*, it is possible to see that the accuracy of this model reaches 0,7716, the sensitivity is 0,7422, and the specificity is 0,8620. The model correctly specified that 5.071 individuals have less than $50k as income per year, and 1.1911 individuals have more than $50k as income per year. The precision is 0,9430909 (the closer is to 1, the better the prediction), and the F1 score (harmonic mean of precision and sensitivity, the closer to 1, the better the prediction) is 0,8306987.

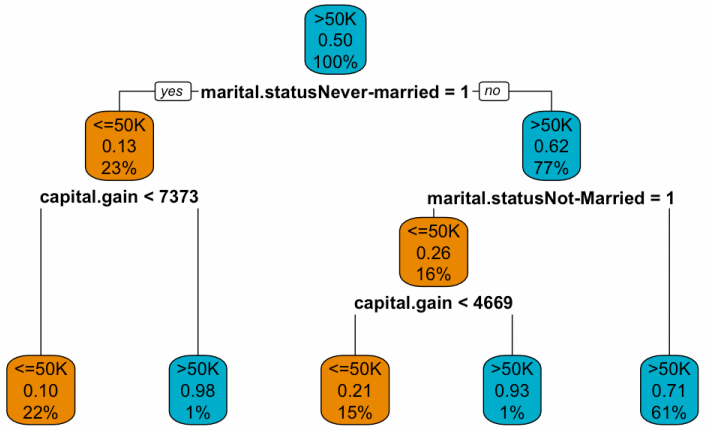
Figure 16: Confusion matrix of tree predictor

```
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  5071  306
     >50K   1761 1911

               Accuracy : 0.7716
                 95% CI : (0.7628, 0.7802)
    No Information Rate : 0.755
    P-Value [Acc > NIR] : 0.0001163

                  Kappa : 0.4946

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7422
            Specificity : 0.8620
         Pos Pred Value : 0.9431
         Neg Pred Value : 0.5204
             Prevalence : 0.7550
         Detection Rate : 0.5604
   Detection Prevalence : 0.5942
      Balanced Accuracy : 0.8021

       'Positive' Class : <=50K
```

1.6 Gini index

In *figure 18*, it is possible to see the division of the tree predictor using the Gini index approach. The entry node is marital status.

Figure 17: Tree with Gini index



In *figure 19*, it is possible to see that the accuracy of this model reaches 0,7091, the sensitivity is 0,6464, and the specificity is 0,9026. The model correctly specified that 4.416 individuals have less than $50k as income per year, and 2.001 individuals have more than $50k as income per year. The precision is 0,9533679, and the F1 score is 0,7704117.

Figure 18: Confusion matrix of tree predictor with Gini index

```
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  4416  216
     >50K   2416 2001

               Accuracy : 0.7091
                 95% CI : (0.6997, 0.7185)
    No Information Rate : 0.755
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4111

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6464
            Specificity : 0.9026
         Pos Pred Value : 0.9534
         Neg Pred Value : 0.4530
             Prevalence : 0.7550
         Detection Rate : 0.4880
   Detection Prevalence : 0.5119
      Balanced Accuracy : 0.7745

       'Positive' Class : <=50K
```

1.7 Naïve Bayes

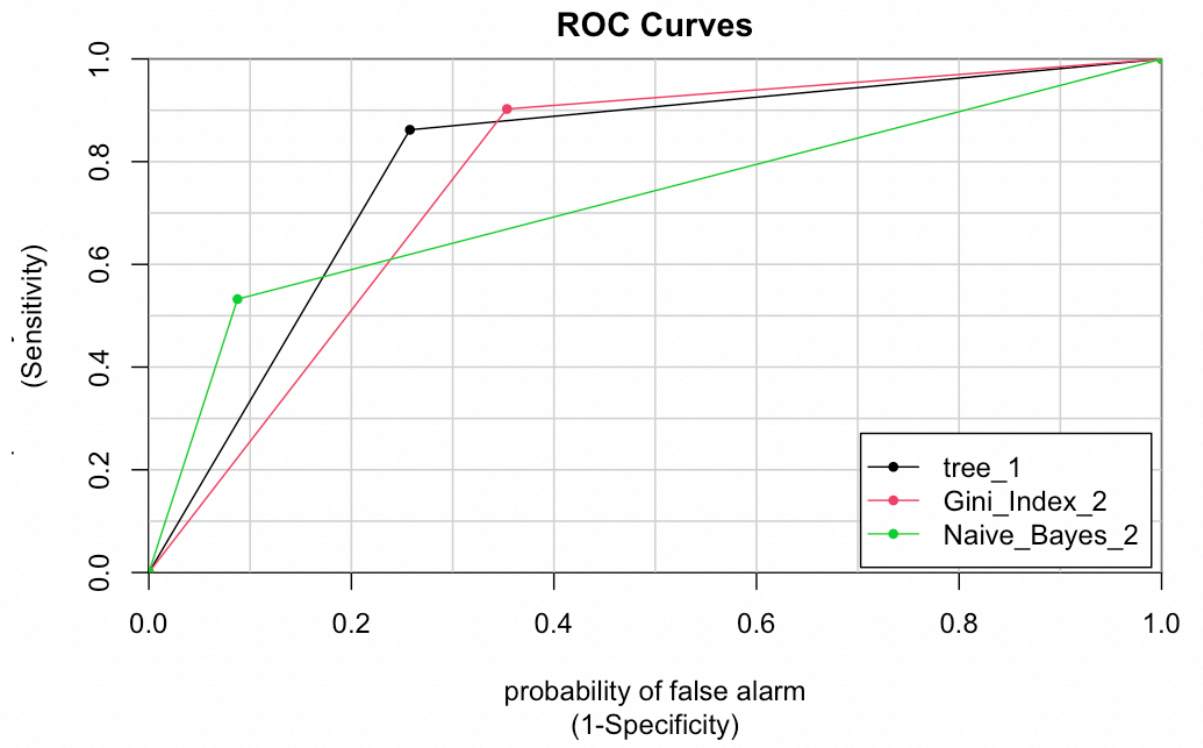
Naïve Bayes classifier is used next. This algorithm uses Bayesian probability. In *figure 20*, it is possible to see that the accuracy of this model reaches 0,8193, the sensitivity is 0,9125, and the specificity is 0,5323. The model correctly specified that 6,234 individuals have less than $50k as income per year, and 1.180individuals have more than $50k as income per year. The precision is 0,8573786, and the F1 score is 0,8840672. Except for specificity value, this model has the best results, as seen in *figure 21*.

Figure 19: Confusion matrix of Naive Bayes

```
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  6234 1037
     >50K    598 1180

               Accuracy : 0.8193
                 95% CI : (0.8112, 0.8272)
    No Information Rate : 0.755
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4766

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9125
            Specificity : 0.5323
         Pos Pred Value : 0.8574
         Neg Pred Value : 0.6637
             Prevalence : 0.7550
         Detection Rate : 0.6889
   Detection Prevalence : 0.8035
      Balanced Accuracy : 0.7224

       'Positive' Class : <=50K
```

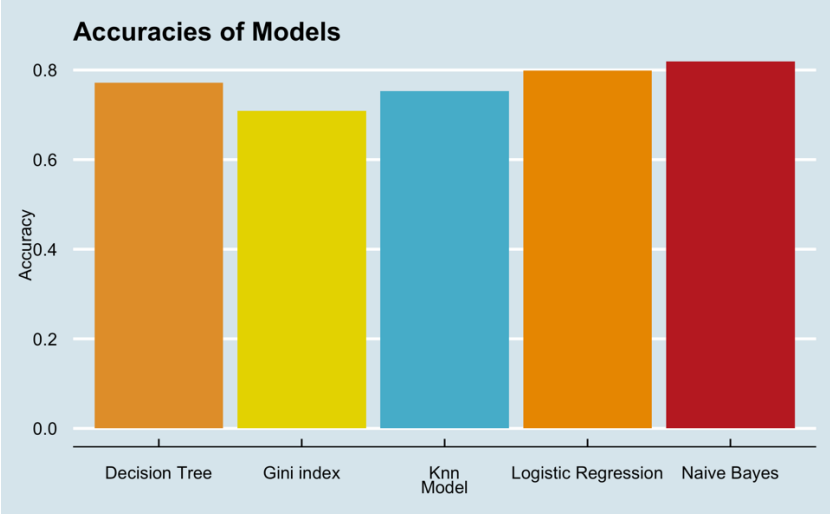
By looking at the ROC curves (figure 21) for the tree predictor, the best algorithm between Gini Index and the Naïve Bayes, is the last one.

Figure 20: ROC curves



In figure 22, it is possible to see the comparison between the accuracies of the models analyzed since now. Naive Bayes was the best-performing algorithm.

Figure 21: Accuracies of models



1.8 Random forest

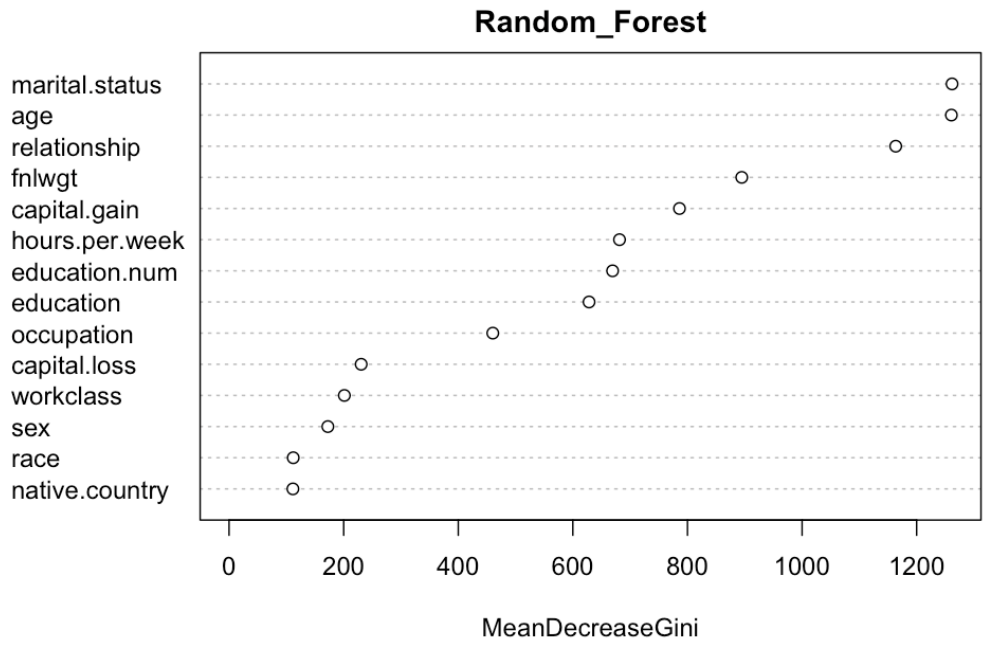
Finally, the last algorithm exploited is the random forest. In *figure 22*, it is possible to see that the accuracy of this model reaches 0,8181, the sensitivity is 0,8318, and the specificity is 0,8137. The model correctly specified that 5.559 individuals have less than $50k as income per year, and 1.844 individuals have more than $50k as income per year. The Naïve Bayes model has a higher accuracy..

Figure 22: Confusion matrix of random forest

```
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  5559  373
     >50K   1273 1844

               Accuracy : 0.8181
                 95% CI : (0.81, 0.826)
    No Information Rate : 0.755
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5676

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8318
            Specificity : 0.8137
         Pos Pred Value : 0.5916
         Neg Pred Value : 0.9371
             Prevalence : 0.2450
         Detection Rate : 0.2038
   Detection Prevalence : 0.3445
      Balanced Accuracy : 0.8227

       'Positive' Class : >50K
```

In *plot 23*, it is possible to see the relevance of each attribute on income determination. Marital status and age are the most relevant attribute, followed by relationship. The least essential attributes are race and native country.

Figure 23: Relevance of variables in random forest



2. Unsupervised learning

2.1 Data Exploration

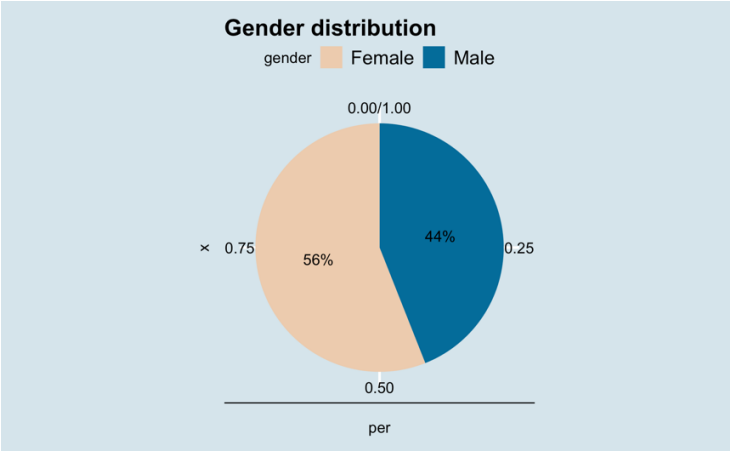
This dataset contains basic information about 200 customers of a mall. The recorded information are ID, age, gender, income, and spending score. There are no missing values. In *table 11*, it is possible to see the summary of the data.

Table 11: Summary of the dataset

Attributes	Vars	N°	Mean	Standard deviation	Min	Max	Range	Standard error
Customer ID	1	200	100.5	57.8791845	1	200	199	4.09267639
Gender	2	200	-	-	-	-	-	-
Age	3	200	38.85	13.9690073	18	70	52	0.98775798
Annual income	4	200	60.56	26.2647212	15	137	122	1.85719624
Spending score	5	200	50.2	25.8235217	1	99	98	1.82599873

In *figure 24*, it is possible to see the gender distribution. As displayed in the graph, female customers represent 56% of total customers.

Figure 24: Gender distribution



In *figure 25*, the density of age by gender is described. Most of the mall customers are females in their thirties and fifties. Male customers are primarily in their thirties. *Figure 26* shows that the most frequent buyers are between thirty and forty years old.

Figure 25: Density of age by gender

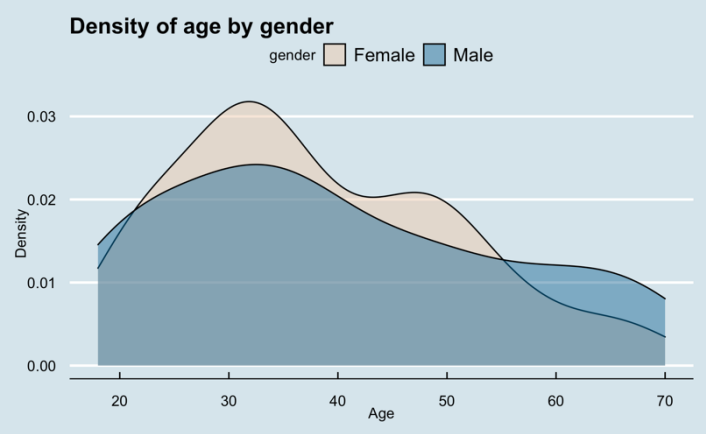


Figure 26: Age analysis

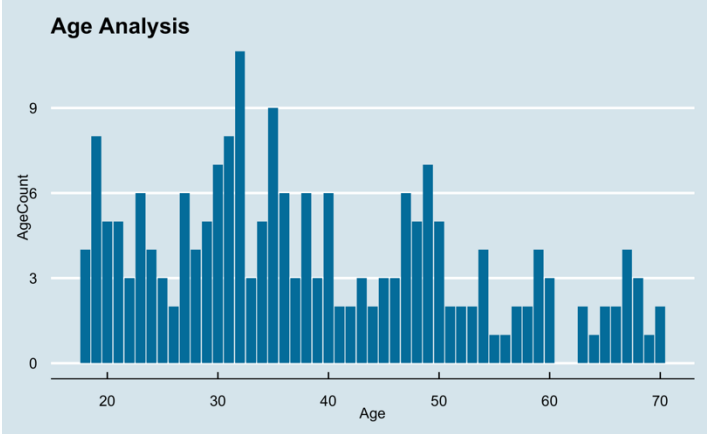
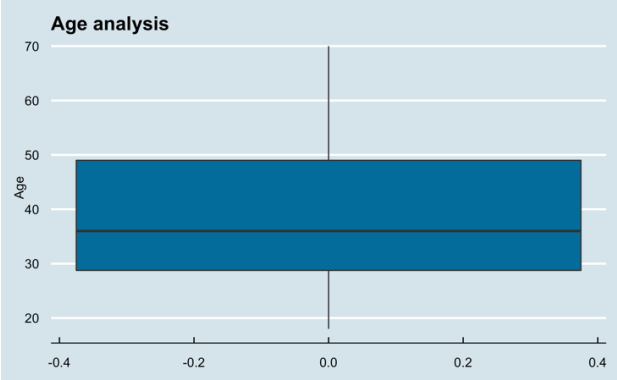


Figure 27 shows that there are no significant outliers in the age attribute.

Figure 27: Age analysis boxplot



Figures 5 and 6 show, respectively, the income and spending score density. Most customers have an annual income between 50 and 100 and a spending score of 50.

Figure 28: Income density

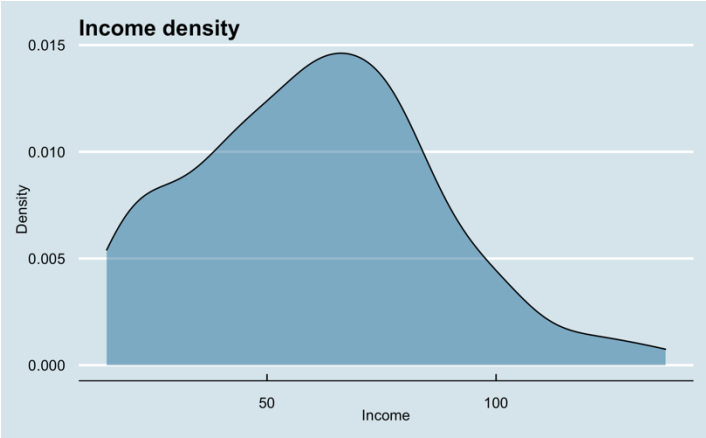
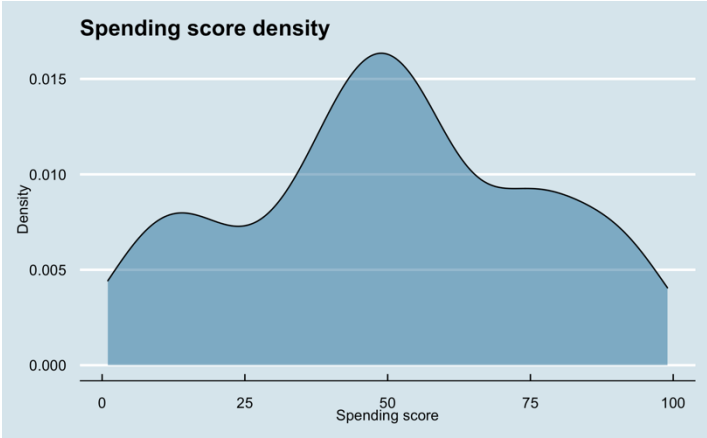
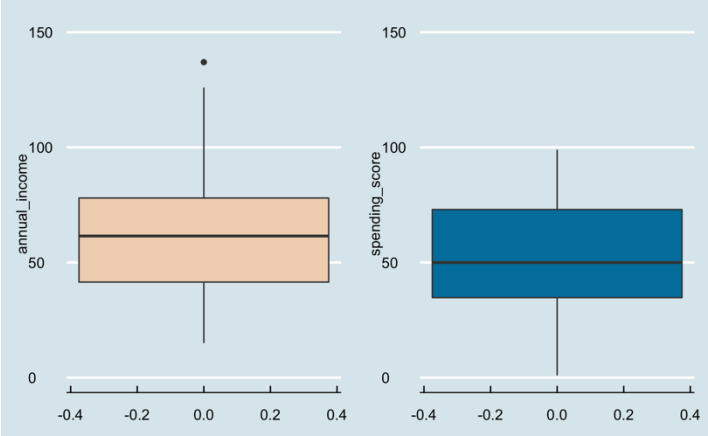


Figure 29: Spending score density



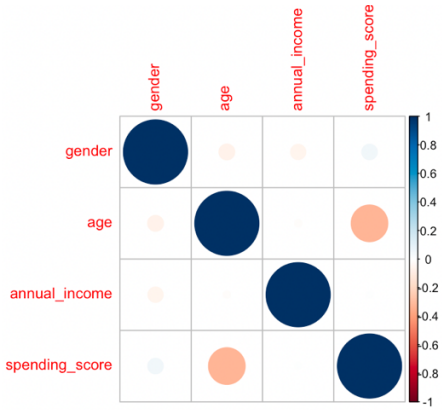
From *figure 30* it is possible to see that there are no significant outliers in income and spending score.

Figure 30: Income and spending score boxplot



In *figure 31* it is possible to see the correlation matrix of the attributes. The most interesting value is between age and spending score, that are negatively correlated.

Figure 31: Correlation matrix



In order to proceed with the analysis, the data were normalized.

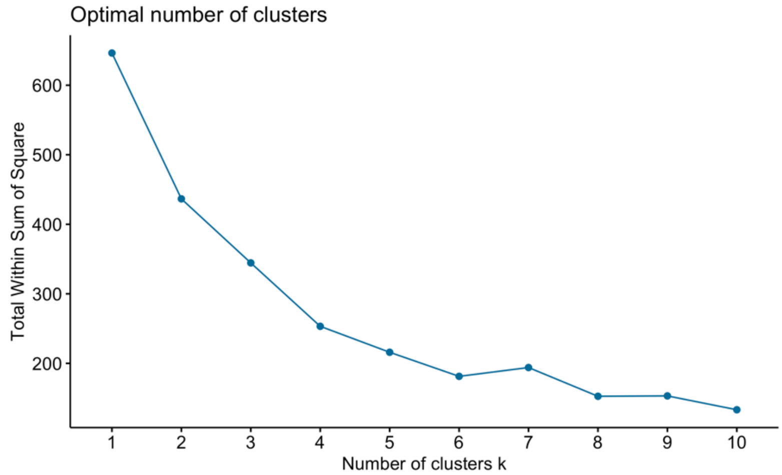
2.2 K-means

K-means belongs to the family of the unsupervised algorithm, it cluster data based on the similarity of groups. The number of K clusters is decided using different methods

2.2.1 Elbow method

The first method used is the elbow method, represented in *figure 32*. The optimal number of clusters suggested by this method is 4 or 5.

Figure 32: Elbow chart



In *figure 33***,** it is possible to see the outcome of applying the K-means algorithm to the dataset with K = 4. The size of the four clusters is 24, 47, 39, and 90. Some important measures are the cluster means, which tells the center of each cluster, and the within-sum of squares by clusters (a measure of the variability of the observations within each cluster, a cluster with a small sum of squares is more compact). In this case, the within-sum is 53,5%.

Figure 33: K-means, K = 4

```
K-means clustering with 4 clusters of sizes 24, 47, 39, 90

Cluster means:
     gender         age annual_income spending_score
1 0.6250000  0.30603463   -1.27934603    -1.08880063
2 0.5744681  1.20182469   -0.23518320    -0.05223672
3 0.4871795  0.07314728    0.97250467    -1.19429976
4 0.5666667 -0.74092595    0.04255814     0.83515591

Clustering vector:
  [1] 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 2 4 1 4 1 4 2 4 1 1 2 4 4 2 2 2 2 2 4 2 2 4 2 2 2 4 2 2 4 4 2 2 2 2 2 4 2
 [78] 2 4 2 2 4 2 2 4 2 2 4 4 2 2 4 2 2 4 4 2 4 2 4 4 2 2 4 2 4 2 2 2 2 2 4 3 4 4 4 2 2 2 2 4 3 4 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4
[155] 3 4 3 4 3 4 2 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4

Within cluster sum of squares by cluster:
[1]  34.08911  38.14601  56.13351 171.92923
 (between_SS / total_SS =  53.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"

 1  2  3  4
24 47 39 90
```

Figure 34 shows the output of K-means with K = 5 applied to the customer dataset. The five classes are composed of 47, 40, 20, 54, and 39 individuals. In this case, the within-sum is 66,5%.

Figure 34: K-means, K = 5

```
K-means clustering with 5 clusters of sizes 47, 40, 20, 54, 39

Cluster means:
     gender         age annual_income spending_score
1 0.5744681  1.20182469    -0.2351832    -0.05223672
2 0.5500000 -0.42773261     0.9724070     1.21304137
3 0.6000000  0.52974416    -1.2872781    -1.23337167
4 0.5925926 -0.97822376    -0.7411999     0.46627028
5 0.4871795  0.07314728     0.9725047    -1.19429976

Clustering vector:
  [1] 4 4 3 4 4 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 1 4 3 4 3 4 1 4 4 4 1 4 4 1 1 1 1 1 4 1 1 4 1 1 1 4 1 1 4 4 1 1 1 1 1 4 1
 [78] 1 4 1 1 4 1 1 4 1 1 4 4 1 1 4 1 1 4 4 1 4 1 4 4 1 1 4 1 4 1 1 1 1 1 1 4 5 4 4 4 1 1 1 1 4 5 2 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2
[155] 5 2 5 2 5 2 1 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2

Within cluster sum of squares by cluster:
[1] 38.14601 33.81544 23.38760 64.89376 56.13351
 (between_SS / total_SS =  66.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"

 1  2  3  4  5
47 40 20 54 39
```

2.2.2 Average silhouette method

As shown in *figure 35*, the average silhouette method suggests that the optimal number of clusters should be 6.

Figure 35: Average Silouette chart

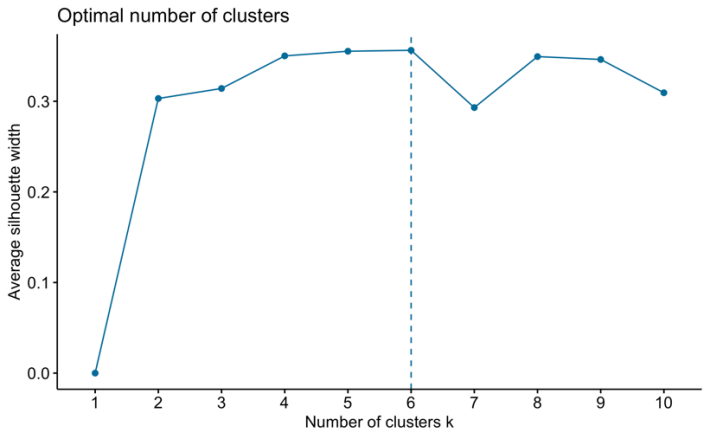


Figure 36 displays the output of K-means with K = 6. In this case, the six classes are composed of 30, 20, 54, 10, 39, and 47 individuals. In this case, the within-sum is 68,2%.

Figure 36: K-means, K = 6

```
K-means clustering with 6 clusters of sizes 30, 20, 54, 10, 39, 47

Cluster means:
     gender         age annual_income spending_score
1 0.5333333 -0.41162553     0.6728925     1.20691001
2 0.6000000  0.52974416    -1.2872781    -1.23337167
3 0.5925926 -0.97822376    -0.7411999     0.46627028
4 0.6000000 -0.47605387     1.8709508     1.23143545
5 0.4871795  0.07314728     0.9725047    -1.19429976
6 0.5744681  1.20182469    -0.2351832    -0.05223672

Clustering vector:
  [1] 3 3 2 3 3 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 6 3 2 3 2 3 6 3 3 3 6 3 3 6 6 6 6 6 3 6 6 3 6 6 6 3 6 6 3 3 6 6 6 6 6 3 6
 [78] 6 3 6 6 3 6 6 3 6 6 3 3 6 6 3 6 6 3 3 6 3 6 3 3 6 6 3 6 3 6 6 6 6 6 3 5 3 3 3 6 6 6 6 3 5 1 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1
[155] 5 1 5 1 5 1 6 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4 5 4

Within cluster sum of squares by cluster:
[1] 16.41985 23.38760 64.89376  6.56153 56.13351 38.14601
 (between_SS / total_SS =  68.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"

 1  2  3  4  5  6
30 20 54 10 39 47
```

2.2.3 Gap statistics

The gap statistics is the last method used to decide the number of K clusters for K-means. In *figure 37*, it is possible to see that the gap statistics suggest K = 2 as the optimal k number.

Figure 37: Gap statistics chart

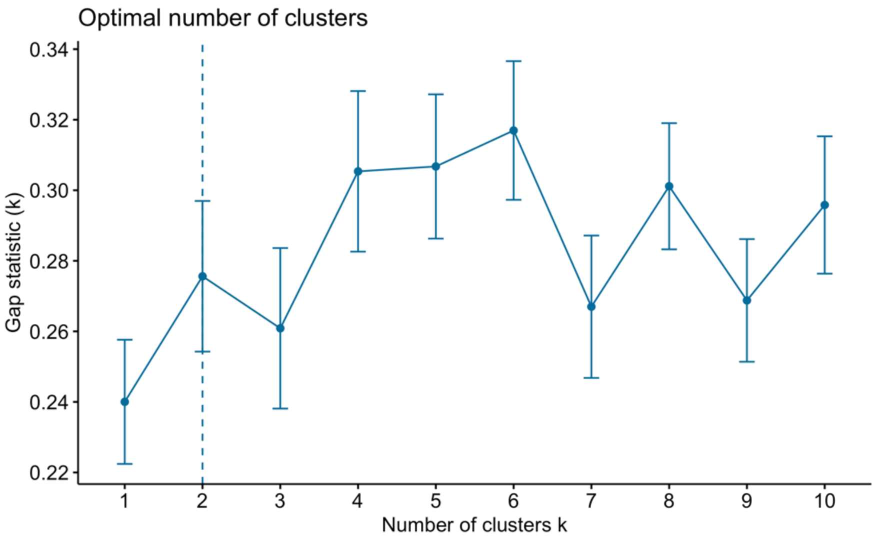


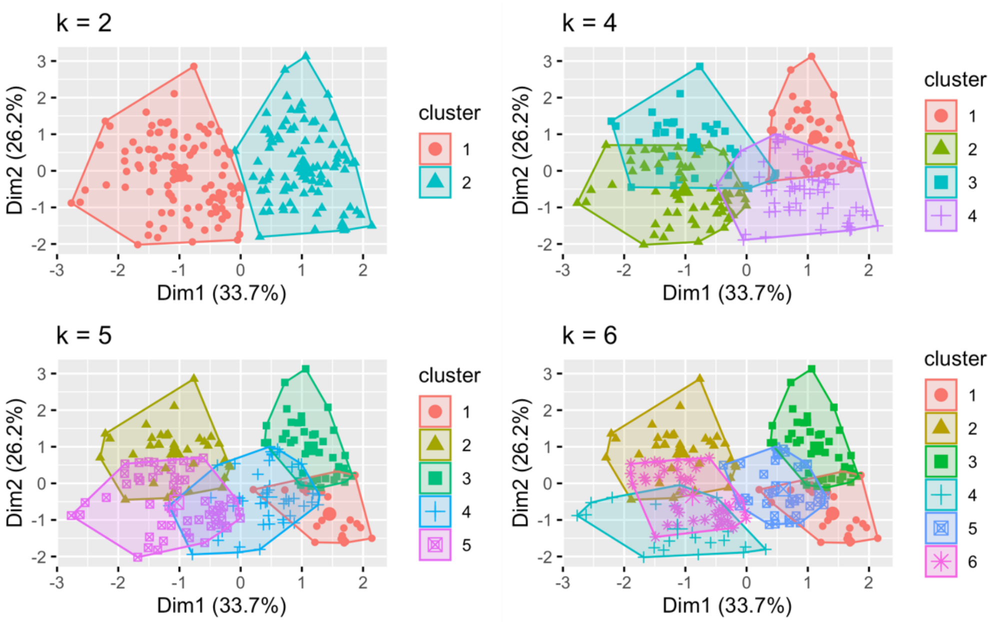
Figure 38 shows the consequences of applying K-means with K = 2 on the dataset. The size of the two classes is 103 and 97. The within-sum is 32,4%.

Figure 38: K-means, K = 2

```
K-means clustering with 2 clusters of sizes 103, 97

Cluster means:
     gender        age annual_income spending_score
1 0.5339806  0.7071480  -0.002469258     -0.6976405
2 0.5876289 -0.7508891   0.002621995      0.7407935

Clustering vector:
  [1] 2 2 1 2 2 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 2 2 1 2 2 1 1 1 1 1 2 1 1 2 1 1 1 2 1 1 2 2 1 1 1 1 1 2 1
 [78] 1 2 1 1 2 1 1 2 1 1 2 2 1 1 2 1 1 2 2 1 2 1 2 2 1 1 2 1 2 1 1 1 1 1 2 1 2 2 2 1 1 1 1 2 1 2 2 2 2 1 2 1 2 1 2 2 2 1 2 1 2 1 2 1 2 2 2 1 2 1 2 1 2 1 2 1 2
[155] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2

Within cluster sum of squares by cluster:
[1] 243.3800 193.1955
 (between_SS / total_SS =  32.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"

  1   2
103  97
```

Figure 39 shows a comparison between the division in clusters using K-means with K =2, K=4, K=5, and K=6. It is worth saying that this is a representation in two dimensions, while the clustering is done in four dimensions. Therefore the mere graphical representation of the clusters is misleading. *Figure 38* shows the consequences of applying K-means with K = 2 on the dataset. The size of the two classes is 103 and 97. The within-sum is 32,4%.

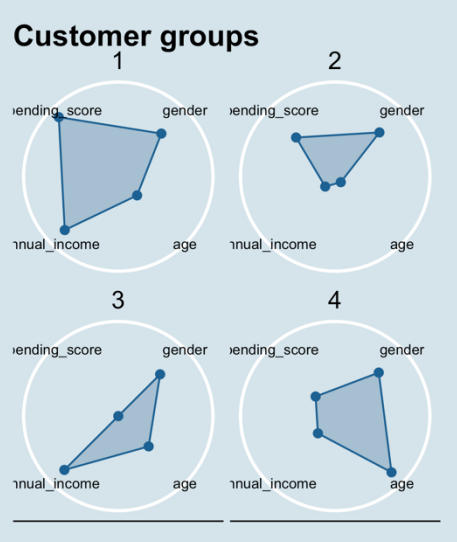
Figure 39: Visualization of k-means clusters



2.3 Results

Customer divisions in 4 clusters (*see figure 40*):

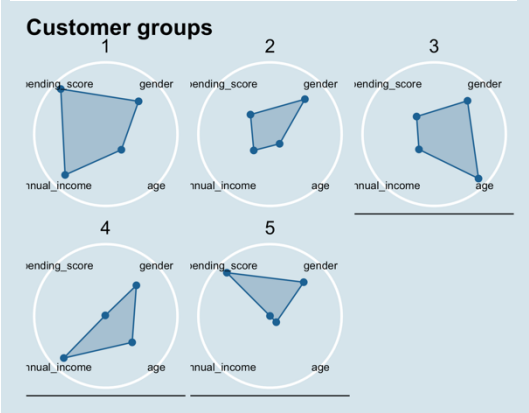
Figure 40: Customers divided in four groups



1. The first cluster contains the youngest analyzed individuals, 18 years old, both males and females. Their annual income is meager, but they have a high spending score. The mall should implement purchasing opportunities and stimulate the demand of this group.
2. The second cluster is composed of young adults around their thirties. They have a high income and a high spending score. These are gold-type customers; The mall should increase their purchasing opportunities.
3. The third group comprises mature people with low income and low spending scores. This type of clientele is of little interest to the mall because they bring little to no gain.
4. The fourth group is composed of middle-aged people. They have a high income but do not seem interested in purchasing at the mall.

Customer division in 5 clusters (*see figure 41*):

Figure 41: Customers divided in five groups

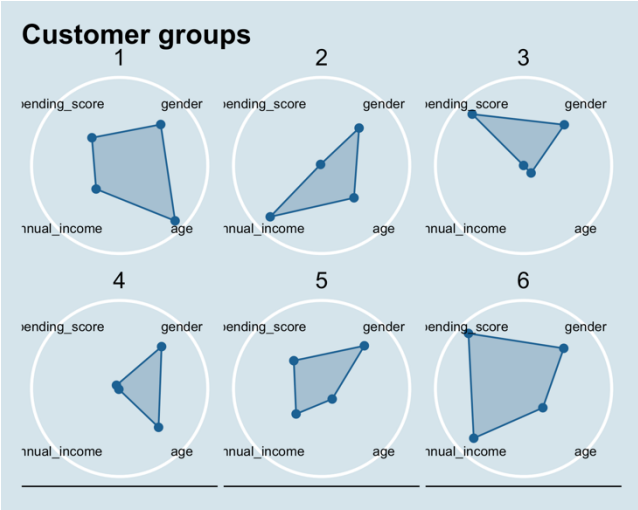


1. The first cluster is composed of young adults that have both high income and high spending scores. These are very profitable types of customers. Therefore the mall should increase their purchase opportunities.
2. The second cluster comprises young adults who earn a modest income and tend to spend moderately. The mall should try to stimulate the demand of purchases of these individuals, perhaps by offering more advantageous prices.
3. The third group has a similar income and spending score to the second cluster. The main difference is that it is composed of mature people.
4. The fourth cluster comprises middle-aged people with high income but no spending score. They are not interested in purchasing at the mall.
5. The fifth cluster is composed of young people with low income but high spending scores. These are very profitable types of customers. The mall should increase their purchase opportunities.

Customer division in 6 groups (*see figure 42*):

1. The first cluster comprises mature people who earn a modest income and tend to spend moderately. The mall should try to stimulate the demand of purchases of these individuals, perhaps by offering more advantageous prices.

Figure 42: Customers divided in six groups



2. The second cluster comprises middle-aged people with high income but no spending score. They are not interested in purchasing at the mall.
3. The third cluster is composed of young people with low income but high spending scores. The mall should concentrate on these customers and increase their purchase opportunities.
4. The fourth group is composed of middle-aged that have no income and no spending score. The mall has no interest in these types of customers.
5. The fifth cluster is composed of young people that have moderate income and moderate spending scores. The mall should stimulate the demand of these types of customers to earn more from them.
6. The sixth cluster comprises young adults with high income and high spending scores. This is the most profitable group of customers for the mall.