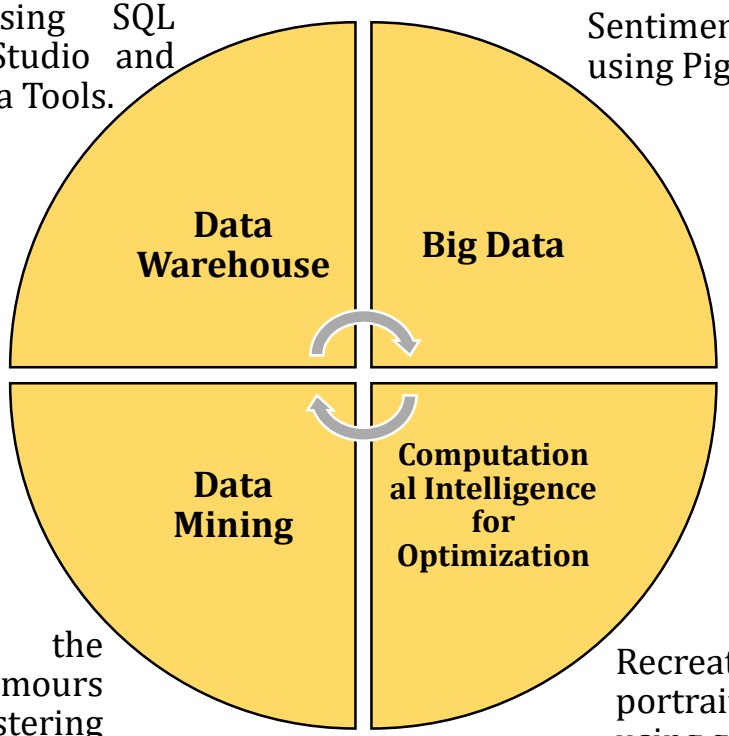
Sofia Jerónimo, 20th June

Overview of Data Science Projects

Building a Data warehouse using SQL Management Studio and SQL Server Data Tools.

Mini project in Sentimental analysis using Pig.

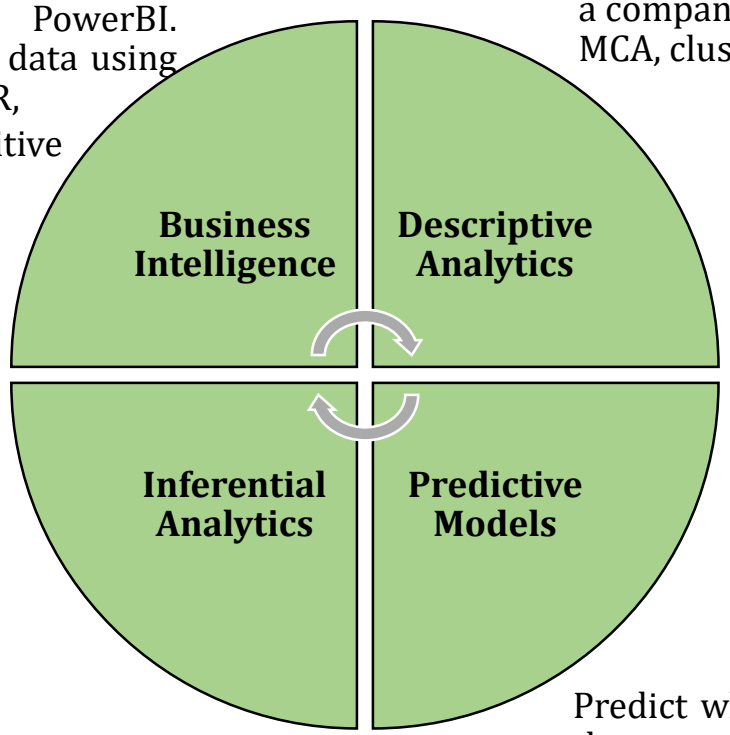


Pattern of the cancer tumours using R: Clustering techniques (Partitional and hierarchical clustering), decision tree.

Recreation of Mona Lisa portrait with triangles using genetic algorithm in Java.

Build a dashboard and reports in PowerBI. Extract twitter data using Twitter API in R, Microsoft cognitive services

Pattern of absenteeism in a company in R: PCA, MCA, clustering.



Predict who would accept the next marketing campaign in SAS.

Genetic Programming to predict human oral bioavailability of a new drug.

Pattern of absenteeism

Initial variables:

34 employees, 667 instances

Categorical: **ID, Reason for absence, Education,** Day of the week, Month, Seasons, **Social Smoker, Social Drinker**

Numerical: **Transportation expense**, **Distance from Residence to Work, Service Time, Age,** Hit target, **Children, Pet, Weight, Height, BMI**, Workload, Absenteeism time

- **Data Pre-processing:**
  - Checked inconsistencies
  - Added more variables: Freq. absence, Freq. failure, First start, categorical BMI, Bad habits
  - Feature selection: correlation analysis
  - Checked outliers
    - Regrouped some of the categories: the reasons for absence, education
    - Removed 4,5% observations

Pattern of absenteeism

Feature selection:

- Feature selection: Pearson's and Spearman's correlation

Service and Age are positively correlated (Pearson's correlation = 0.68 and Spearman = 0.78)

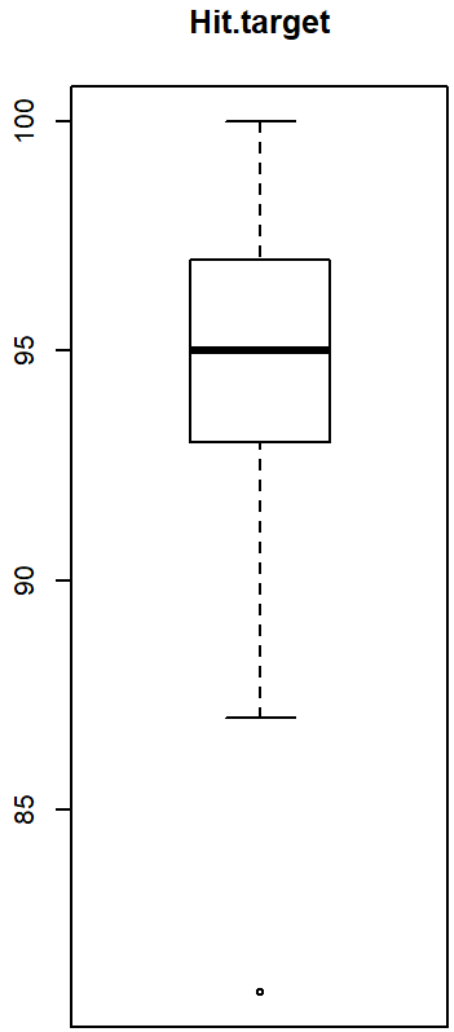
Age and First.start are positively correlated (Pearson's correlation = 0.70 and Spearman = 0.57)

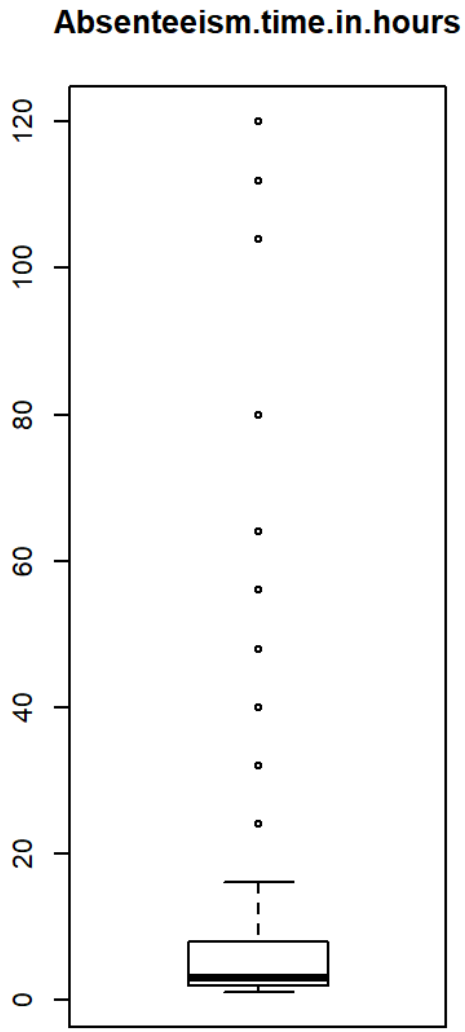
Weight and **BMI** are positively correlated (Pearson's correlation = 0.90 and Spearman = 0.88)

Pattern of absenteeism

Outliers treatment:

- Regrouped some of the categories: the reasons for absence, education
- Numeric: Hit-target<85% and Absenteeism time > 48 (4,1%)





Pattern of absenteeism

12 Variables: Freq.failure, Transportation.expense, Distance.from.Residence.to.Work, Service.Time, Hit.target, Son, Pet, Height, BMI, Freq.absence, Workload, First.Start

PCA: reduce number of variables

- Standardized variables (Z-score)

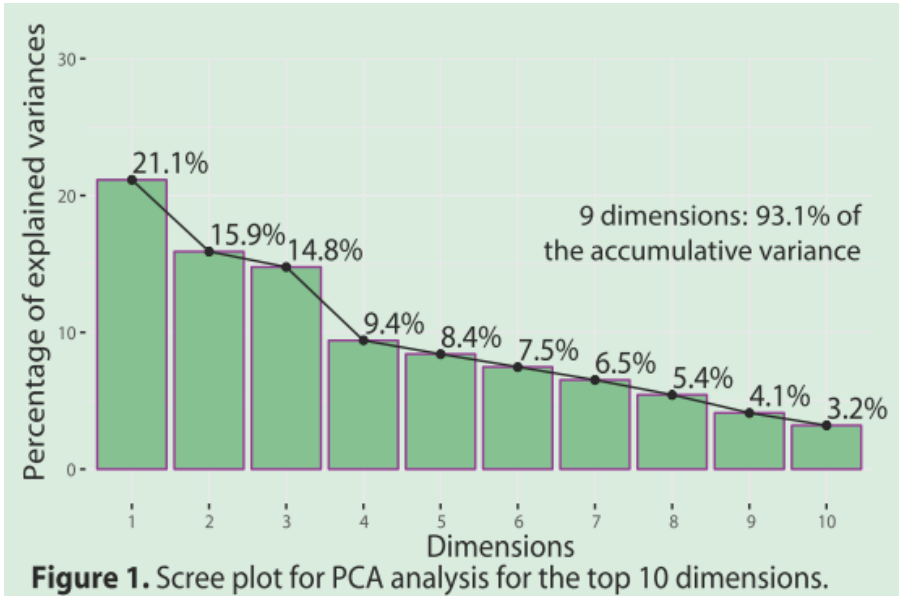


Figure 1. Scree plot for PCA analysis for the top 10 dimensions.

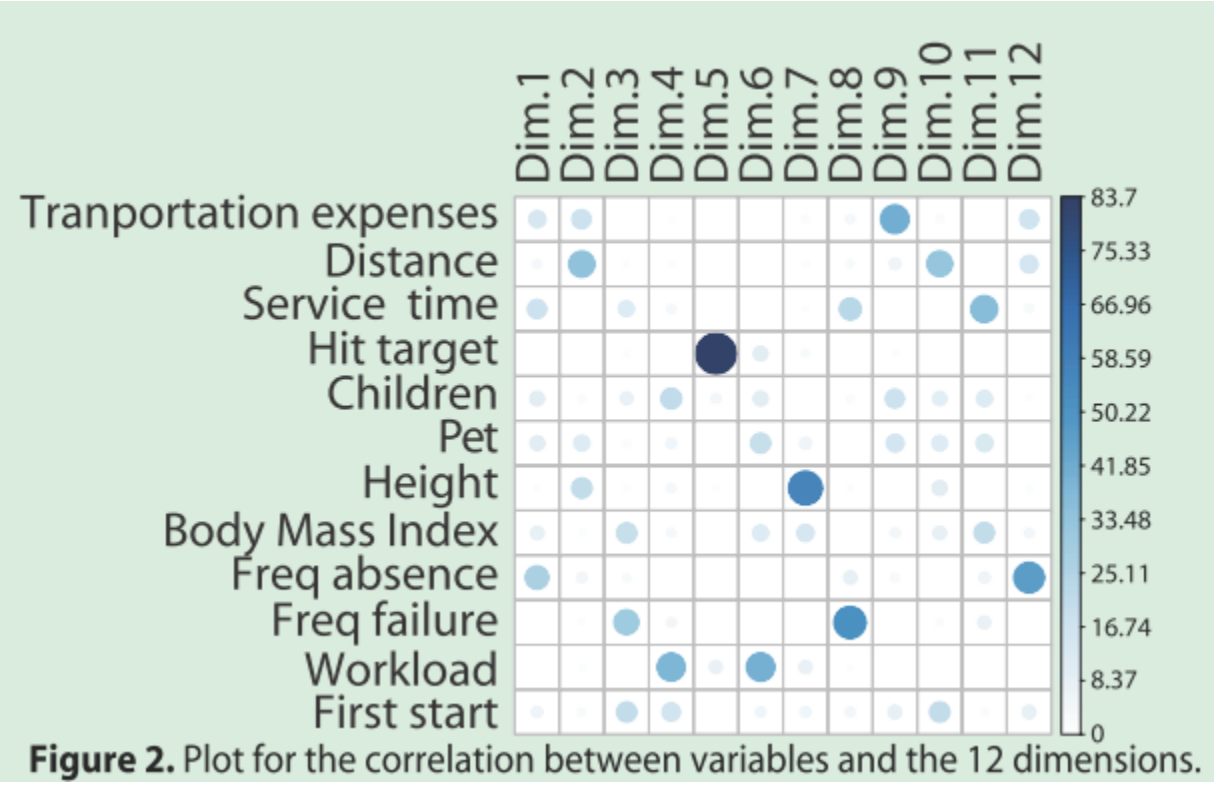


Figure 2. Plot for the correlation between variables and the 12 dimensions.

Pattern of absenteeism

Clustering: K-medoids

```
library(factoextra)
dSpearm=get_dist(matcomp9, method = "spearman")

clmSpearm=pamk(dSpearm, k=3, criterion="asw", usepam=TRUE, scaling=FALSE, alpha=0.001, diss=TRUE, critout=FALSE, ns=10, seed=NULL)
```

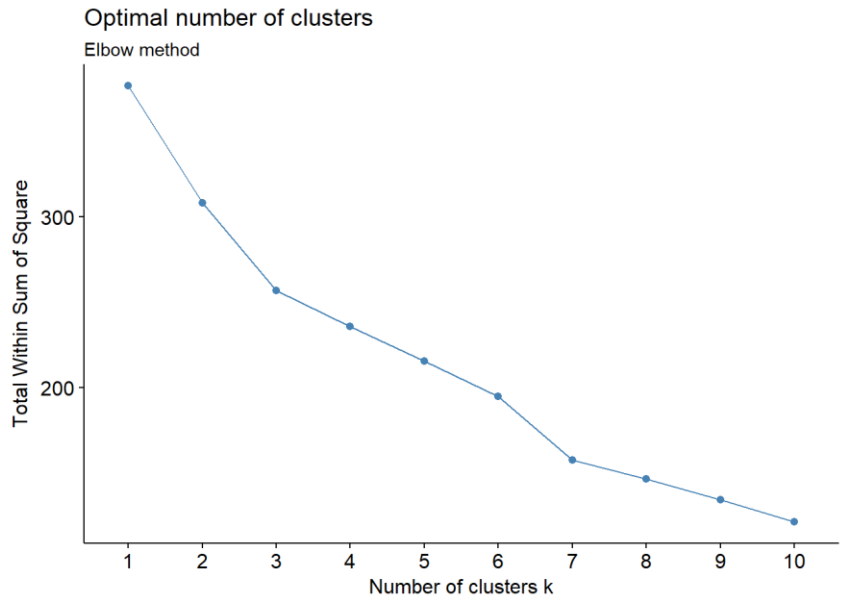


Table 1. Centroids of the 15 variables for the 3 clusters and the significance test.

Variables	Cluster 1	Cluster 2	Cluster 3	P-value
Transportation expense	252.72	195.40	209.71	<2e-16
Distance	39.03	18.41	32.65	<2e-16
Service time	13.26	11.95	12.19	0.00317
Age	36.16	37.40	34.46	8.21e-06
Hit target	94.30	95.63	95.12	9.46e-06
Children	1.37	0.71	0.87	2.78e-11
Pet	0.89	0.27	1.14	1.63e-12
Weight	82.33	77.16	76.01	1.01e-07
Height	170.15	174.75	171.25	<2e-16
Absent hours	5.78	5.53	5.43	0.833
Body mass index	28.45	25.21	25.90	<2e-16
Freq.absence	54.38	29.46	60.37	<2e-16
Freq.failure	1.79	0.85	1.32	1.26e-12
Workload	4.16	4.33	5.17	<2e-16
First start	22.89	25.45	22.28	3.21e-14

Pattern of absenteeism

Table 2. Frequency of each reason for the 3 clusters.

	Variables	Cluster 1	Cluster 2	Cluster 3
Reasons	Accompanying person	21	7	10
	Dental consultation	46	31	30
	Diagnosis, donation and vaccination	10	21	9
	Diseases	42	78	55
	Injury, poisoning	16	10	8
	Medical consultation	52	51	44
	Physiotheraphy	23	14	31
	Pregnancy, childbirth, perinatal complications	1	4	1
	Symptons and abnormal exams	7	9	4
	Unjustified	20	10	2
	Total	**238**	**235**	**194**

Table 3. Frequency of the absenteeism hour for the 3 clusters.

	Variables	Cluster 1	Cluster 2	Cluster 3
Absent time	1 hour	25	30	31
	2 hours	35	64	58
	3-7 hours	77	54	45
	>8 hours	101	88	60
	Total	**238**	**235**	**194**

Pattern of absenteeism

- **MCA:** dependence between categorical variables and clustering
- **11 Variables (m):** Seasons, Pet, Children, Freq.failure, Reason for absence, Freq. absence, Bad habits, Absenteeism time (discrete), Body mass, Day of the week, BMI.
- Total nº levels=47, 47-m=47-11=37, (1/37)*100=2.70%

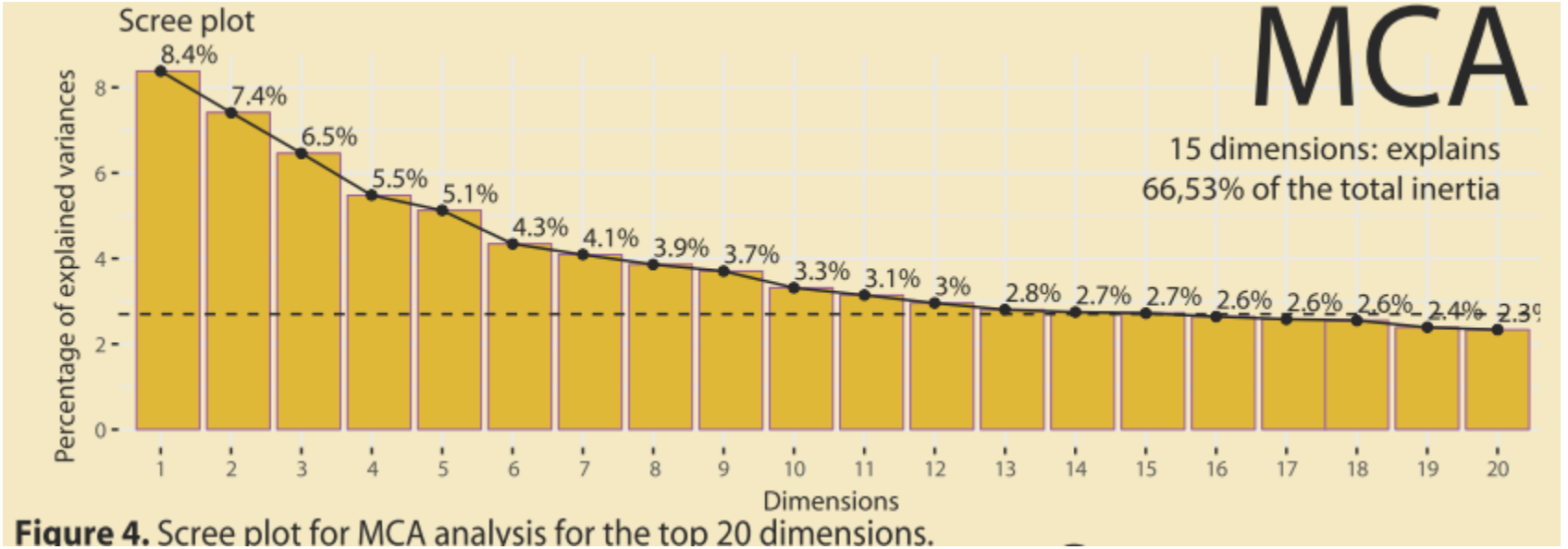


Figure 4. Scree plot for MCA analysis for the top 20 dimensions.

Pattern of absenteeism

- Clustering: HCPC

```
res.mca <- MCA
(MCAdata2, ncp=15,
graph=TRUE)
res.hcpc =
HCPC(res.mca,nb.clust=4)
```

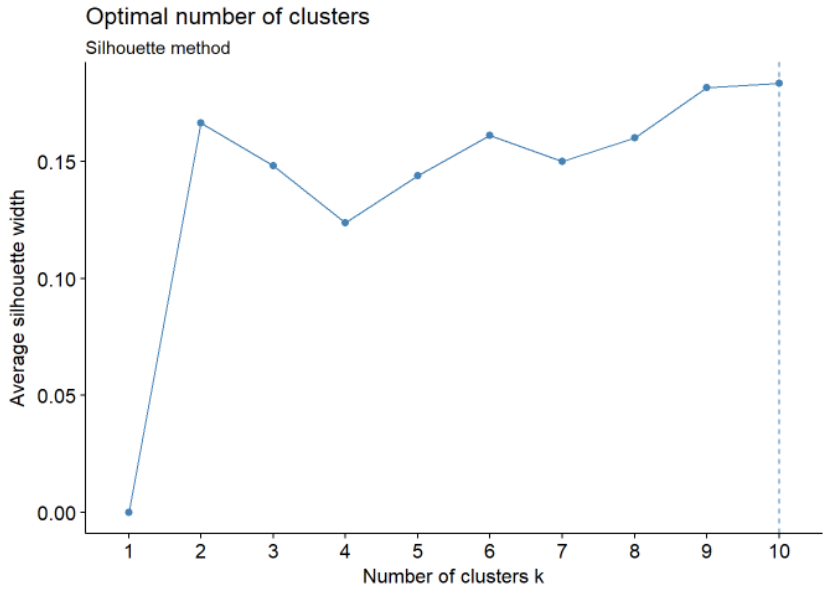


Table 6. Centroids of the 15 variables for the 4 clusters

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Transportation expense	186.52	224.90	248.57	191.28
Distance	47.22	26.36	29.88	17.99
Service time	17.14	9.36	12.62	10.05
Age	37.79	28.40	38.47	34.04
Hit target	95.46	94.99	94.76	95.11
Children	0.00	1.01	1.85	0.19
Pet	0.00	2.00	1.01	0.23
Weight	85.73	69.51	79.75	75.44
Height	170.69	168.99	172.92	173.14
Absenteeism time	4.26	3.29	7.14	4.86
Body mass index	29.48	24.34	26.67	25.12
Freq.absence	96.88	72.35	26.18	35.50
Freq.failure	1.00	1.92	1.84	0.33
Workload	4.34	4.72	4.56	4.47

Pattern of absenteeism

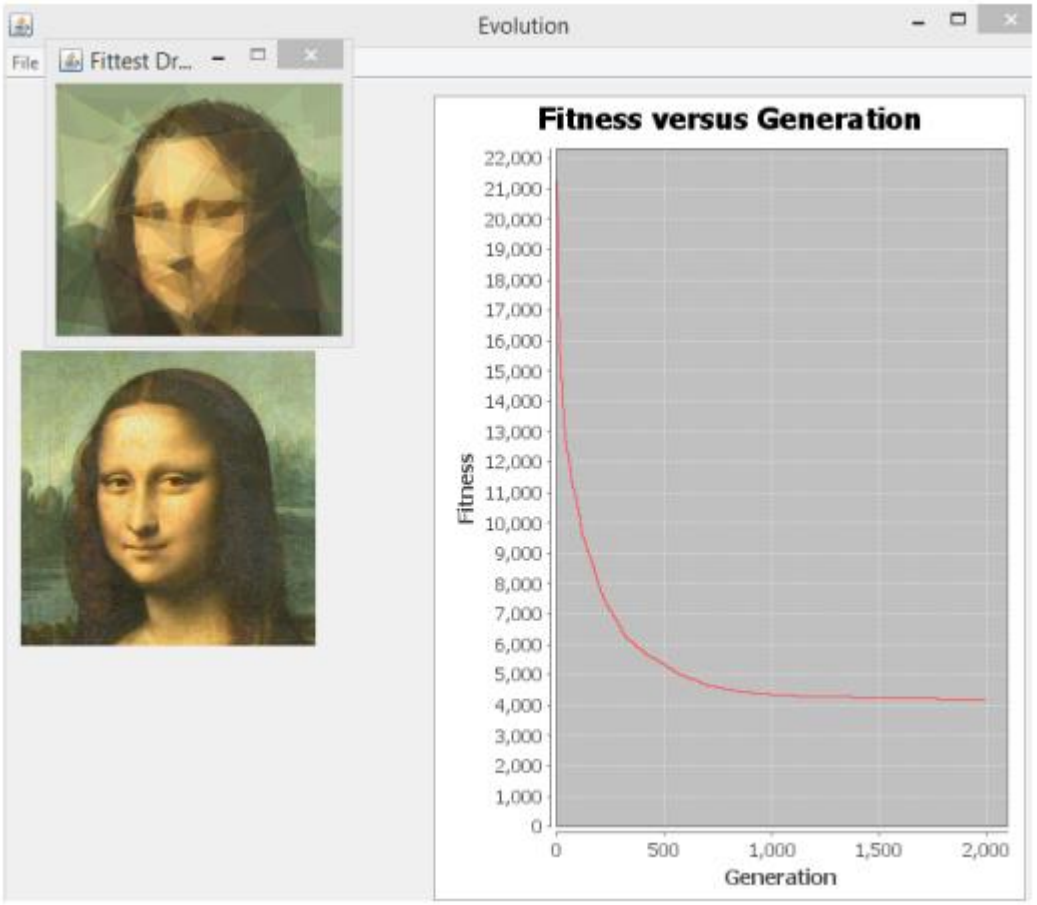
Table 5. Frequency of each reason for the 4 clusters.

	Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Reasons	Accompanying person	0	1	32	5
	Dental consultation	39	10	45	13
	Diagnosis, donation and vaccination	5	6	18	11
	Diseases	23	16	79	57
	Injury, poisoning	2	3	25	4
	Medical consultation	18	36	60	33
	Physiotheraphy	38	4	0	26
	Pregnancy, childbirth, perinatal complications	0	0	5	1
	Symptons and abnormal exams	2	1	12	5
	Unjustified	1	0	24	7
	Total	**128**	**77**	**300**	**162**

Table 4. Frequency of the absenteeism hour for the 4 clusters.

	Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Absent time	1 hour	21	13	34	18
	2 hours	34	26	47	50
	3-7 hours	49	26	61	39
	>8 hours	24	12	158	55
	Total	**128**	**77**	**300**	**162**

Mona Lisa optimization problem



Minimization problem.

Each solution has 100 triangles.

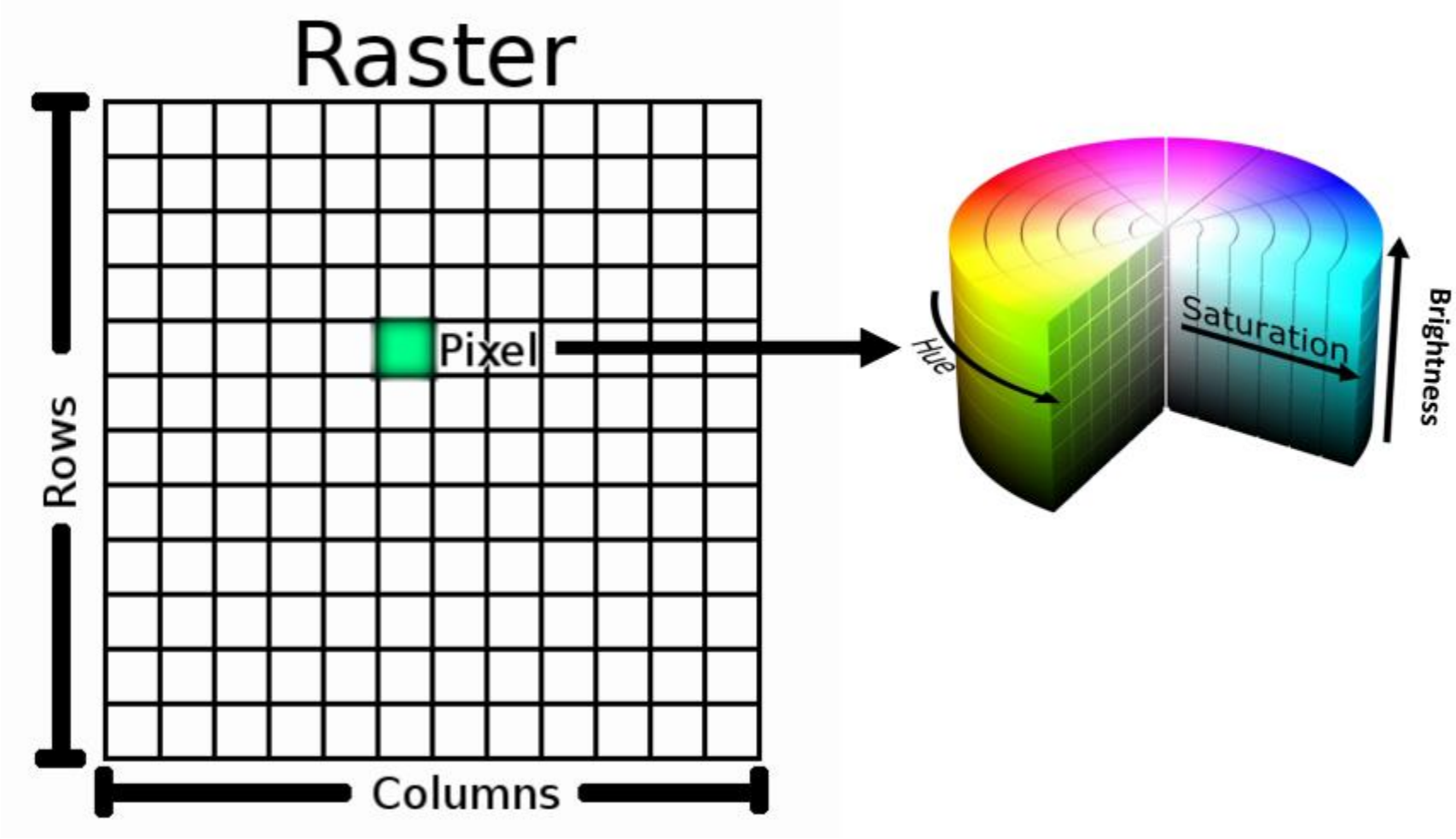
2000 generations

Population size: 25

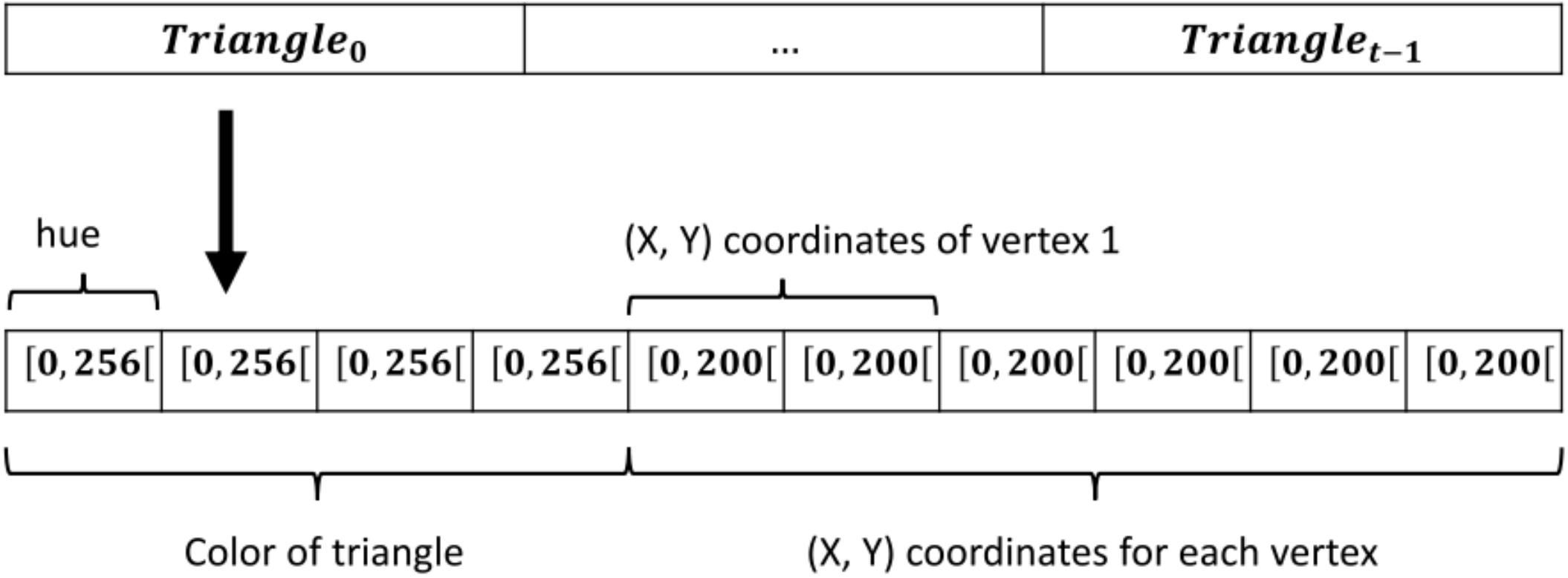
Fitness: Eucledian distance between our solution and the target image.

The triangles in a solution represent raster (one pixeled image) and it has the same size as the target raster.

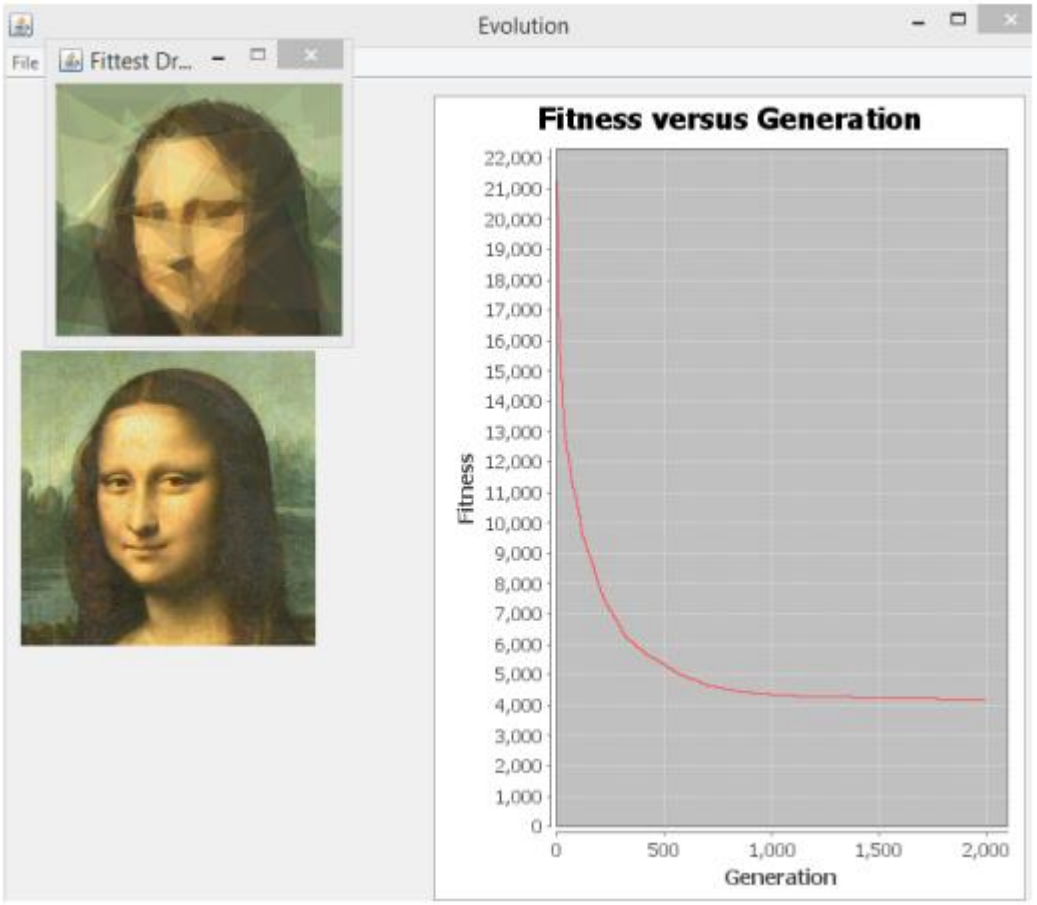
Mona Lisa using Java



Mona Lisa using Java



Mona Lisa using Java



Initialization: parameter 'GoodInitialization' to start in an already good solution from a previous run

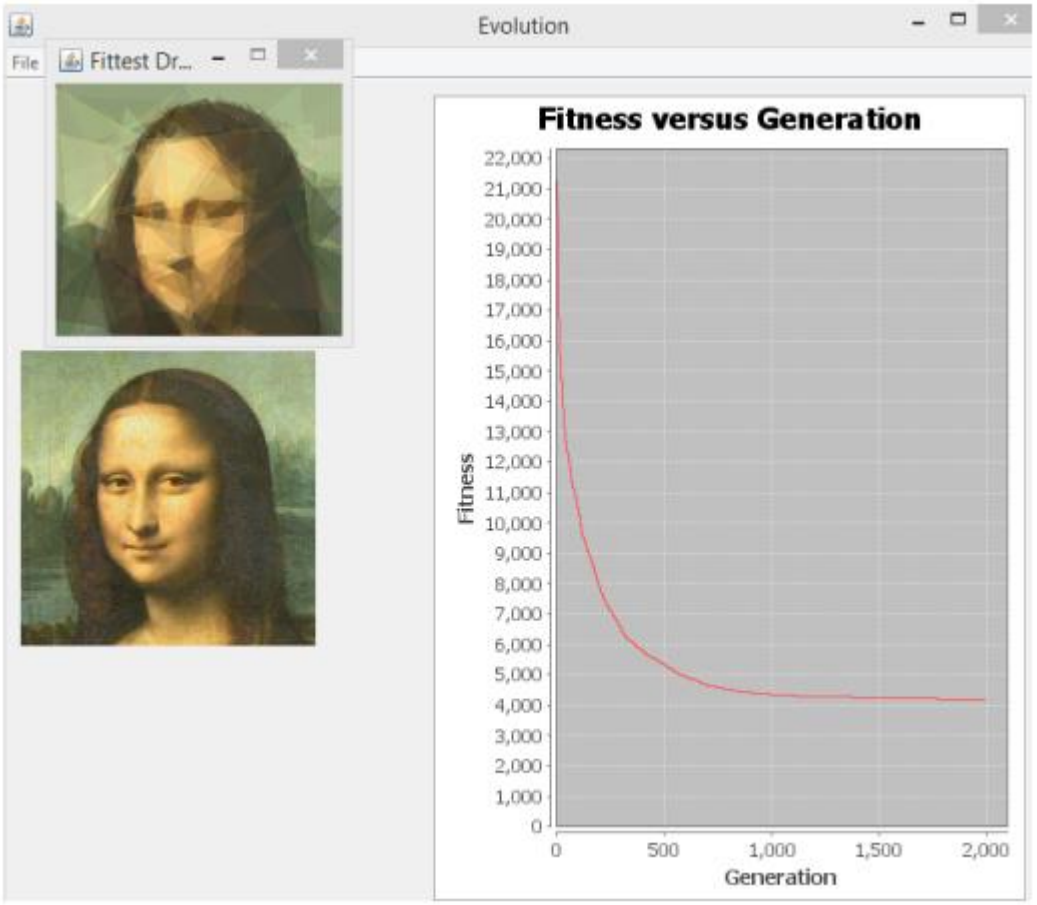
Selection: Tournament, Roulette

Presence/Absence of Elitism

Crossovers: Single point, Two point, Average point, K-Point

Mutation: Standard, Box mutation

Mona Lisa using Java



Best solution returned at 2000th generation has the following characteristics:

Fitness: 4193

Tournament selection

Standard mutation

With elite

Single point crossover