

# Multivariate analysis - D2 report

Andreja Andrejic

David Candela

Nayara Costa

Julie Oppedal

Alexandros Tremopoulos

## 1 Working plan

### 1.1 Task division

Task	Andreja	David	Nayara	Julie	Alexandros
Report	X	X	X	X	X
Data cleaning				X	X
Descriptive analysis			X	X	
PCA	X	X	X		
ACM and MFA	X	X		X	
Associations rules			X		X
Clustering	X				X
Profiling		X		X	
Discriminant analysis	X		X		
Decision trees		X			X
Presentation preparation	X	X	X	X	X

Table 1: Task division

### 1.2 Risk management

- A team member leaves the group  
**How to prevent:** All the tasks need to have more than one member assigned to them.  
**How to manage:** Re-assign the tasks to compensate workload.
- A team member can't keep up with his/her task  
**How to prevent:** More than one member in task.  
**How to manage:** Extra member gets involved and re-assign.
- Previous task not completed fully  
**How to prevent:** After each task, communicate and discuss the outcomes  
**How to manage:** Update the previous task and check with the group
- Inconsistent data through tasks  
**How to prevent:** Common database  
**How to manage:** Server or drive backups
- A team member has problems with oral presentations  
**How to prevent:** Ensure communication within the group to detect the team members that will need assistance  
**How to manage:** Divide the oral presentation accordingly

### 1.3 Grantt diagram

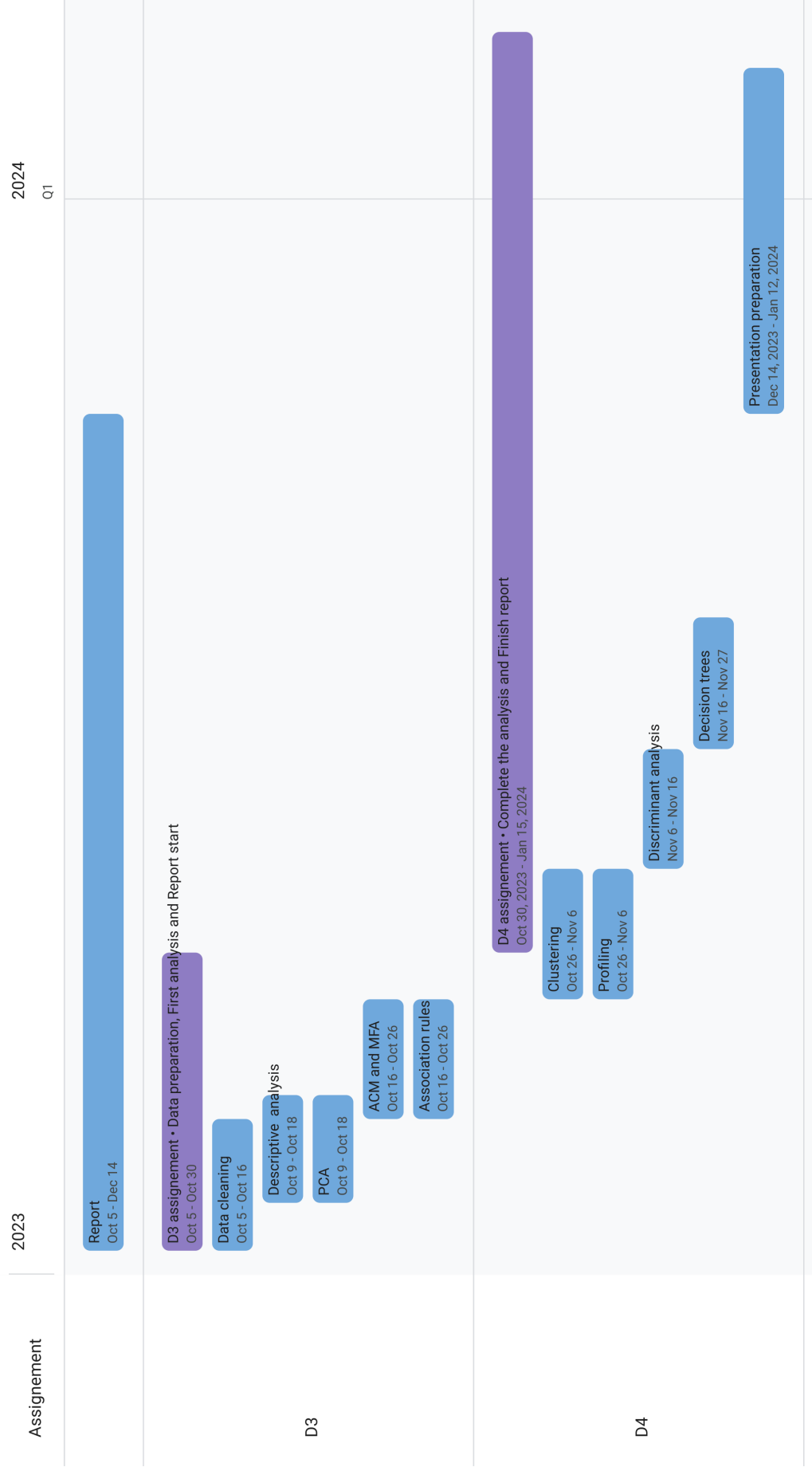


Figure 1: Grantt diagram

## 2 Metadata

### 2.1 Dataset information

**Dataset:** [Bank Marketing](#)<sup>1</sup>

**How to download:** The files can be downloaded directly through this [link](#) or going through the [Bank Marketing](#) link and pressing the ‘Download’ button. This will download a zip folder called ‘bank+marketing’. When the contents are extracted, there is a zip folder called ‘bank’. Inside this folder, there is the file ‘bank.csv’ that we are using. ‘bank.csv’ contains 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

**Description:** The Bank Marketing dataset compiles call information from a Portuguese banking institution with the objective of predicting whether a client will subscribe to a term deposit. This classification task relies on various categorical features, such as the client’s education level, occupation, and marital status. Additionally, numerical features like the client’s age and average account balance are considered in the analysis. For more information visit the [Bank Marketing](#)<sup>1</sup> web page.

### 2.2 Contents

The datasets contains a total of 4521 records with seventeen different variables. The dataset’s variables are the following:

---

<sup>1</sup><https://archive.ics.uci.edu/dataset/222/bank+marketing>

Full name	Variable name	Description	Type	Units	Modalities	Missing markers	Missing percentage
Age	age	Client's age	Integer	years			0
Job	job	Client's job	Categorical		'admin.', 'unemployed', 'management', ... <sup>a</sup>	'unknown'	0.85
Marital status	marital	Client's marital status	Categorical		'married', 'divorced' <sup>b</sup> , 'single'		0
Educational level	education	Client's level of education	Categorical		'secondary', 'primary', 'tertiary', 'unknown'	'unknown'	4.13
Credit default	default	Does the client have credit in default?	Binary		'yes', 'no'		0
Account balance	balance	Client's average account balance, in euros	Integer	euros			0
Has housing loan	housing	Does the client have a housing loan?	Binary		'yes', 'no'		0
Has personal loan	loan	Does the client have any other kind of loan?	Binary		'yes', 'no'		0
Contact method	contact	Method used to communicate with the client	Categorical		'telephone', 'cellular', 'unknown'	'unknown'	29.7
Last contact	day	Day of the month of the last contact	Date	days			0
Last contact duration	month	Month of the last contact	Date	months	'jan', 'feb', 'mar', ..., 'nov', 'dec'		0
Total times contacted	duration	Duration of the last contact with the client	Integer	seconds			0
Days since last contact	campaign	Number of times the client was contacted during the campaign	Integer				0
Previous total contacts	pdays	Time since the client was last contacted from a previous campaign	Integer	days		'-1' <sup>c</sup>	81.9
Previous result	previous	Number of times the client was contacted previous to the campaign	Integer				0
	poutcome	Outcome of the previous marketing campaign	Categorical		'other', 'failure', 'success', 'unknown'	'unknown'	81.9
Has subscribed	y	<b>Outcome of this marketing campaign</b>	Binary		'yes', 'no'		0

Table 2: Variables metadata

<sup>a</sup>Complete job list: 'admin.', 'unemployed', 'management', 'housemaid', 'entrepreneur', 'student', 'blue-collar', 'self-employed', 'retired', 'technician', 'services', 'unknown'

<sup>b</sup>'divorced' means divorced or widowed

<sup>c</sup>The client has never been contacted previously