**Big Five Personality Test**

**Apendizaje de Máquinas – 2020-I**

Mateo González ***a***

Alejandro Uribe ***b***

***a***1037627601; ***b***1037627601

June 2020

**Table of Content**

[1. Introduction 1](#_Toc43509574)

[2. Data Mining Process 1](#_Toc43509575)

[2.1. Business Understanding 1](#_Toc43509576)

[2.2. Data Understanding 2](#_Toc43509577)

[2.3. Data Preparation 3](#_Toc43509578)

[2.4. Modelling 3](#_Toc43509579)

[2.4.1. Choice of Algorithm 3](#_Toc43509580)

[2.4.2. Training 3](#_Toc43509581)

[2.5. Evaluation 3](#_Toc43509582)

[2.5.1. Choice of Metrics 3](#_Toc43509583)

[2.5.2. Validation 3](#_Toc43509584)

[3. Conclusion 3](#_Toc43509585)

[4. References 3](#_Toc43509586)

**Table of Figures**

[Figure 1. CRISP-DM Methodology. 1](#_Toc43509527)

[Figure 2. Project schedule. 2](#_Toc43509528)

**List of Tables**

[Table 1. Traits description. 2](#_Toc43509615)

# Introduction

Big Five…

# Data Mining Process

*CRoss-Industry Standard Process for Data Mining* (CRISP-DM) is the most commonly used methodology for analytics, data mining and analytics projects which provides an overview of the life cycle of a data mining project. CRISP-DM is a robust industry standard process which guides the analyst through a set of phases which cover all possible data mining situations [1]. The Figure 1 shows the CRISP-DM Methodology.

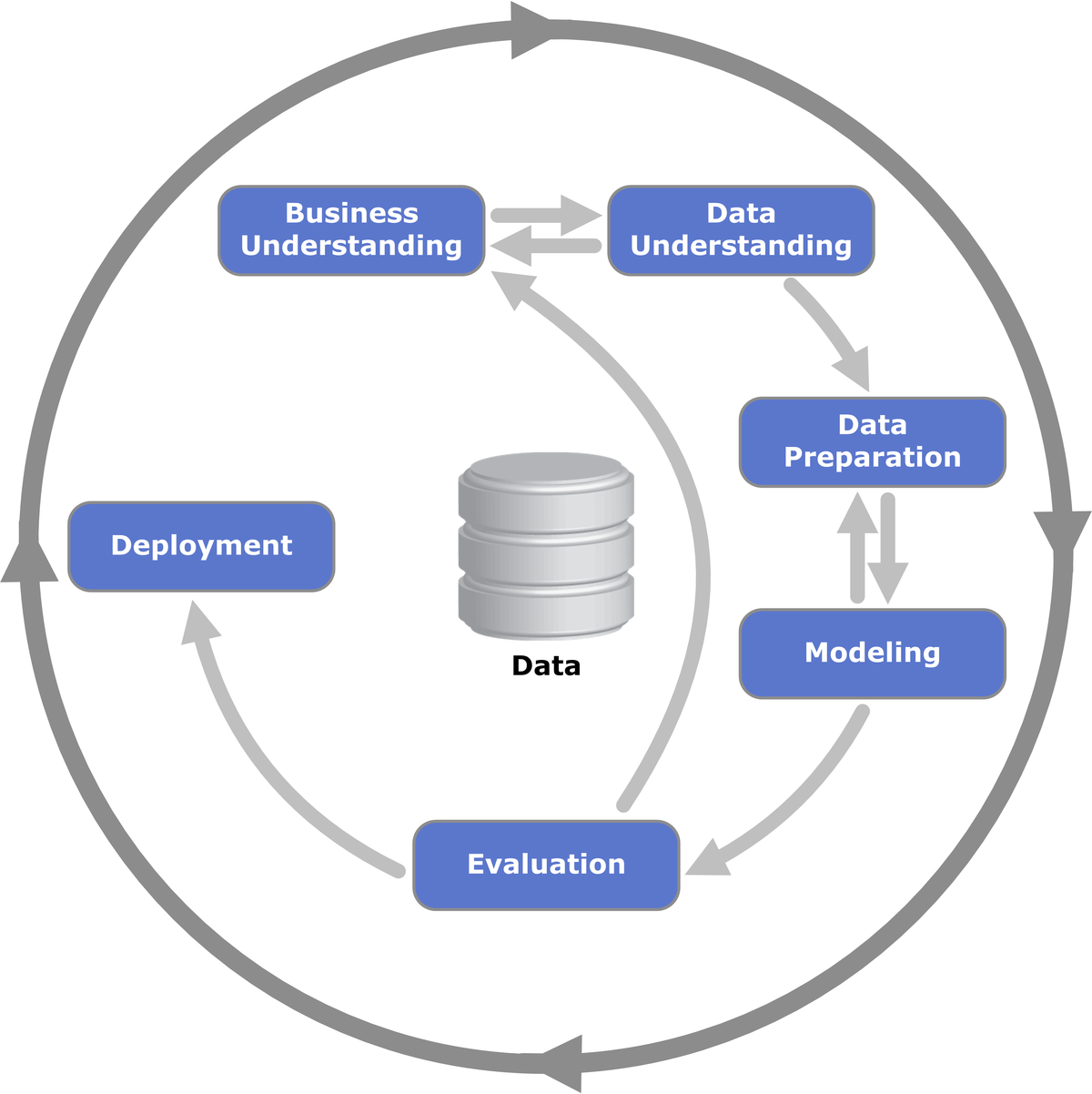


Figure 1. CRISP-DM Methodology.

Source: [2].

CRISP-DM methodology phases will be followed in order to build a solution for the current problem.

## Business Understanding

The first phase is the most important in the lifecycle of the project, the analysts can set expectations and success from the thorough understanding of the business requirements.

The problem at hand arises from the need to group the people according to their personality, thus simplifying the understanding of human mind. The “Big Five Personality Test” have widespread since its release in XX century due to its scalability and ease of distribution.

The optimal solution is a model able to classify the trait with max value of the test, given the answers of this one. These traits are described in Table 1.

Table 1. Traits description.

|  |  |  |
| --- | --- | --- |
| Trait | Tag | Description |
| Extroversion | E | **High scores:** People who tend to be social  **Low scores:** People who prefer to work alone in projects |
| Agreeableness | A | **High scores:** Politically correct people  **Low Scores:** Direct people |
| Conscientiousness | C | **High Scores:** People who tend to follow the rules and prefer order  **Low Scores:** People who tend to be disorganized |
| Neuroticism | N | How emotional the person can be |
| Openness to Experience | O | **High Scores:** People who tend to “dream with their eyes open”  **Low Scores:** People who tend to “have their feet on the ground” |

The Figure 2 shows the schedule followed by the team to carry out the project.

A screenshot of a cell phone

Description automatically generated

Figure 2. Project schedule.

Aquí habría que definir un posible error dado el poder de un sicólogo experto para poder diferenciar (¿?)

Discuss and document possible Machine Learning and data mining methods suitable for the solution by assessing possible tools, algorithms, and techniques. <---- listar los posibles modelos

## Data Understanding

This phase involves collecting the data, describing its attributes and performing the data exploratory analysis. Firstly, the data can be found in open data sources, e.g., Kaggle [3] or can be purchased in third party companies. Secondly, an initial analysis on the data will provided the very first insights about it such as its nature, volume, features and relationships. The Table 2 summarizes the characteristics of the data:

Table 2. Summary of the data.

|  |  |
| --- | --- |
| Characteristic | Number |
| Records | 964573 |
| Total Features | 112 |
| Numeric Features | 107 |
| Categorical Features | 4 |
| Datetime Features | 1 |

* Explore, describe, and visualize data attributes
* Select data and attributes subsets that seem most important for the problem
* Extensive analysis to find correlations and associations and test hypotheses
* Note missing data points if any

## Data Preparation

This phase can represent 80% of the spent in the project, is aimed to detect the errors and inconsistencies in the dataset, these values can be either ignored or replaced.

## Modelling

Here comes the “fun part”, since there is a wide range of algorithms or techniques which can apply for the given problem. Multiple algorithms will be tried

### Choice of Algorithm

### Training

## Evaluation

### Choice of Metrics

### Validation

# Conclusion

­­

# References

[1] M. F. Hornick, E. Marcadé, and S. Venkayala, “Chapter 3 - Data Mining Process,” in *The Morgan Kaufmann Series in Data Management Systems*, M. F. Hornick, E. Marcadé, and S. B. T.-J. D. M. Venkayala, Eds. Burlington: Morgan Kaufmann, 2007, pp. 51–83.

[2] “File:CRISP-DM Process Diagram.png - Wikimedia Commons.” .

[3] B. Tunguz, L. Petar, and M. Akdag, “Big Five Personality Test | 1M Answers to 50 personality items, and technical information.” [Online]. Available: https://www.kaggle.com/tunguz/big-five-personality-test. [Accessed: 20-Jun-2020].