

# Analysis of the dataset *Credit Card Default*

Gaspare Ferraro (520549)

Javad Khalili (546677)

Mario Matovic (583449)

December 11, 2018



University of Pisa

Exam: Data Mining

Year: 2018/2019

Instructors: Dino Pedreschi, Anna Monreale, Riccardo Guidotti

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Data Understanding</b>	<b>3</b>
2.1	Data semantics . . . . .	3
2.2	Distribution of the variables and statistics . . . . .	3
2.3	Assessing data quality . . . . .	3
2.4	Variables transformations . . . . .	3
2.5	Correlations and redundant variables . . . . .	3
<b>3</b>	<b>Clustering</b>	<b>4</b>
3.1	K-means . . . . .	4
3.1.1	Choice of attributes and distance function . . . . .	4
3.1.2	Choice of the best value of k . . . . .	4
3.1.3	Cluster analysis . . . . .	4
3.2	DBSCAN . . . . .	4
3.2.1	Choice of attributes and distance function . . . . .	4
3.2.2	Study of the clustering parameters . . . . .	4
3.2.3	Characterization and interpretation of the obtained clusters . . . . .	4
3.3	Hierarchical clustering . . . . .	4
3.3.1	Choice of attributes and distance function . . . . .	4
3.3.2	Discussion of dendograms using different algorithms . . . . .	4
3.4	Evaluation of clustering approaches and comparison of the clustering obtained . . . . .	4
<b>4</b>	<b>Association Rules Mining</b>	<b>5</b>
4.1	Frequent patterns extraction with different parameters . . . . .	5
4.2	Discussion of the most interesting frequent patterns . . . . .	5
4.3	Association rules extraction with different values of confidence . . . . .	5
4.4	Discussion of the most interesting rules . . . . .	5
4.5	Use the most meaningful rules to replace missing values . . . . .	5
4.6	Use the most meaningful rules to predict credit card defaults . . . . .	5
<b>5</b>	<b>Classification</b>	<b>6</b>
5.1	Choice of attributes for the decision trees . . . . .	6
5.2	Decision trees interpretation and validation with test and training set . . . . .	6
5.3	Discussion of the best prediction model . . . . .	6
<b>6</b>	<b>Conclusion</b>	<b>7</b>

# Chapter 1

## Introduction

## Chapter 2

# Data Understanding

2.1 Data semantics

2.2 Distribution of the variables and statistics

2.3 Assessing data quality

2.4 Variables transformations

2.5 Correlations and redundant variables

## Chapter 3

# Clustering

### 3.1 K-means

3.1.1 Choice of attributes and distance function

3.1.2 Choise of the best value of k

3.1.3 Cluster analysis

### 3.2 DBSCAN

3.2.1 Choice of attributes and distance function

3.2.2 Study of the clustering parameters

3.2.3 Characterization and interpretation of the obtained clusters

### 3.3 Hierarchical clustering

3.3.1 Choice of attributes and distance function

3.3.2 Discussion of dendograms using different algorithms

3.4 Evaluation of clustering approaches and comparison of the clustering obtained

## Chapter 4

# Association Rules Mining

- 4.1 Frequent patterns extraction with different parameters
- 4.2 Discussion of the most interesting frequent patterns
- 4.3 Association rules extraction with different values of confidence
- 4.4 Discussion of the most interesting rules
- 4.5 Use the most meaningful rules to replace missing values
- 4.6 Use the most meaningful rules to predict credit card defaults

## Chapter 5

# Classification

- 5.1 Choice of attributes for the decision trees
- 5.2 Decision trees interpretation and validation with test and training set
- 5.3 Discussion of the best prediction model

## Chapter 6

## Conclusion