Data mining project

Alicja Czajkowska

magda leszczynska

Paulina Luty

Anna Skowron

HDSDA

SEP2

Contents

[**Problem Statement** 2](#_Toc511629353)

[**Data set description** 2](#_Toc511629354)

[**Tools, Data pre-processing and formatting** 2](#_Toc511629355)

[**Models used** 2](#_Toc511629356)

[**Construction of data mining model** 2](#_Toc511629357)

[**Decision** **Tree** 3](#_Toc511629358)

[**Clustering** 3](#_Toc511629359)

[**Conclusions** 3](#_Toc511629360)

[**References** 3](#_Toc511629361)

# **Problem Statement**

Sdfsdf

# **Data set description**

We chose “Wine quality - white” as a data set for our project. Data set is available to download as a .csv file on <http://www3.dsi.uminho.pt/pcortez/wine/> [Cortez et al., 2009][[1]](#footnote-1). Data set consists of 4899 rows and the following 12 columns:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| fixed acidity | - | real number | - | amount of |
| volatile acidity | - | real number | - | amount of |
| citric acid | - | real number | - | amount of |
| residual sugar | - | real number | - | amount of residual sugar |
| chlorides | - | real number | - | amount of |
| free sulfur dioxide | - | real number | - | amount of |
| total sulfur dioxide | - | real number | - | amount of |
| density | - | real number | - | amount of |
| pH | - | real number | - | amount of |
| sulphates | - | real number | - | amount of |
| alcohol | - | real number | - | amount of |
| quality | - | real number | - | amount of scale 0 - 10 |

First 11 attributes are the input and 12th - “quality” is an output attribute. There is no missing attributes.

# **Tools, Data pre-processing and formatting**

We will use Excel, RapidMiner, R and Python in our project.

First, we checked the data set for NANs and duplications.

# **Models used**

We will use decision trees, clustering and association models.

Decision tree is a classification method and is also a predictive method. Classification methods use existing data to create a model that will allow to classify new data. In our project we will use decision tree to predict xxxxxx.

Clustering is a descriptive technique that finds groups of observations (clusters) that share similar characteristics in a data set.

# **Construction of data mining model**

xxxx

## **Decision** **Tree**

xxx

## **Clustering**

Xxx

# **Conclusions**

# **References**

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from[[2]](#footnote-2) physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

2. Sources

Created by: Paulo Cortez (Univ. Minho), António Cerdeira, Fernando Almeida, Telmo Matos and José Reis (CVRVV) @ 2009

1. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

   Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems>, Elsevier, 47(4):547-553. ISSN: 0167-9236. Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016 [Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf [bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)