Big Data Coursework

Mateusz Hawrot S1633727

[mhawro200@caledonian.ac.uk](mailto:mhawro200@caledonian.ac.uk)

Table of Contents

[1 Data Pre-processing and Analysis 3](#_Toc26881913)

[1.1 Introduction 3](#_Toc26881914)

[1.2 Dataset description 3](#_Toc26881915)

[1.3 Data Analysis 6](#_Toc26881916)

[1.3.1 Data Types 6](#_Toc26881917)

[1.3.2 Check for duplicates 7](#_Toc26881918)

[1.3.3 Data Visualisation 8](#_Toc26881919)

[2 Data Tuning and Testing 12](#_Toc26881920)

# Data Pre-processing and Analysis

## Introduction

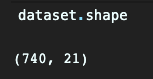
The purpose of the project is to implement supervised Machine Learning algorithm in practical data analysis work. The open-source dataset was selected, and number of operations were performed on it. The dataset needs to be taken into deep analyse and pre-processing. Afterwards, the testing and training model needs to be created and description provided. The chosen supervised machine learning needs to be implemented and the tuned the model in order to increase the accuracy. All the coursework task needs to be justified and explained.

## Dataset description

The dataset chosen for the coursework purpose was created with the records of absenteeism at work between 2007 and 2010 at courier company in Brazil.

The source of the data is:

The data analysis in python indicates the set stores 740 instances and 21 Attributes.



A screenshot of a video game

Description automatically generatedBy the use of *.head()* method all the attributes were displayed.

The deep exploration of the set allows to understand all of the attributes and provides the description. The following information were discovered during dataset exploration.

|  |  |  |
| --- | --- | --- |
| **No** | **Attribute Name** | **Information** |
| 1 | ID | Individual Identification |
| 2 | Reason for absence | Indicated the reason for absence which is separated into 28 categories. The absence attested by the International Code of Diseases (ICD). This will be explained in separated table |
| 3 | Month of absence |  |
| 4 | Day of the work | Monday = 2 | Tuesday = 3 | Wednesday = 4 | Thursday = 5 | Friday = 6 |
| 5 | Season | Summer = 1 |autumn = 2 | winter = 3 | spring = 4 |
| 6 | Transportation Expense | Expense in Euros |
| 7 | Distance from Residence to Work | Distance in kilometres |
| 8 | Service Time | Service time in years |
| 9 | Age | The age |
| 10 | Workload Average/day | - |
| 11 | Hit Target | - |
| 12 | Disciplinary failure | Yes = 1 | No = 0 |
| 13 | Education | High school = 1 | Graduate = 2 | Postgrad = 3 | Master or Doctor = 4 |
| 14 | Son | Number of children |
| 15 | Social drinker | Yes = 1 | No = 0 |
| 16 | Social smoker | Yes = 1 | No = 0 |
| 17 | Pet | Number of pets |
| 18 | Weight | In kg |
| 19 | Height | In cm |
| 20 | BMI | Body Mass Index |
| 21 | Absenteeism | Tim in hours (the target) |

The second attribute which is *Reason for absence* specifies the reason why each person was absent and indicates the type of disease. The following table explain each of them:

|  |  |
| --- | --- |
| **ID** | **Description** |
| 1 | Certain infectious and parasitic diseases |
| 2 | Neoplasms |
| 3 | Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism |
| 4 | Endocrine, nutritional and metabolic diseases |
| 5 | Mental and behavioral disorders |
| 6 | Diseases of the nervous system |
| 7 | Diseases of the eye and adnexa |
| 8 | Diseases of the ear and mastoid process |
| 9 | Diseases of the circulatory system |
| 10 | Diseases of the respiratory system |
| 11 | Diseases of the digestive system |
| 12 | Diseases of the skin and subcutaneous tissue |
| 13 | Diseases of the musculoskeletal system and connective tissue |
| 14 | Diseases of the genitourinary system |
| 15 | Pregnancy, childbirth and the puerperium |
| 16 | Certain conditions originating in the perinatal period |
| 17 | Congenital malformations, deformations and chromosomal abnormalities |
| 18 | Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified |
| 19 | Injury, poisoning and certain other consequences of external causes |
| 20 | External causes of morbidity and mortality |
| 21 | Factors influencing health status and contact with health services. |
| 22 | patient follow-up |
| 23 | Medical consultation |
| 24 | Blood donation |
| 25 | Laboratory examination |
| 26 | Unjustified absence |
| 27 | Physiotherapy |
| 28 | Dental Consultation |

## Data Analysis

This part of the coursework takes to analyse the dataset and perform operations such as checking for missing values, checking for duplicates, investigating the type of data and performing all needed operation to make the data ready for building a model and testing and training part.

### Data Types

The first check performed on the dataset was to distinguish the type of data that they are stored in. Usually, data are categorized into 4 basic types from a ML perspective: numerical data, categorical data, time series data and text. (Zhang, 2018).

The quick check performed in python script indicated that all data are integers, therefore all of them are matching the numerical data.

A screenshot of a cell phone

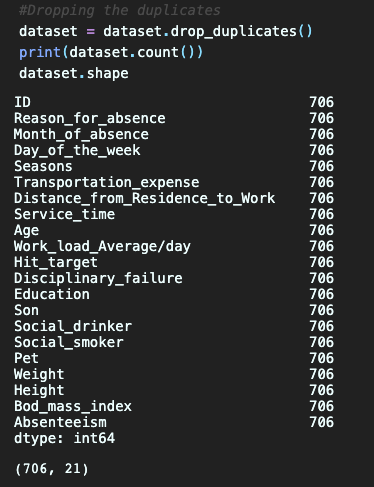
Description automatically generated

However, the data examination performed before indicates that *Reason for absence* are categorical data where a number represents a value (String).

### Check for duplicates and missing values

A black sign with white text

Description automatically generatedBy performing the duplicate check on the dataset, we can reduce the set size and remove unnecessary data. To check for duplicated the following python script need to be run:

It indicated that 34 rows contain duplicated data. Therefore, they should be dropped by *drop\_duplicates()* method.

The screenshot shown above illustrated that duplicates were dropped and the dataset size after that operation was reduced to (706, 21).

Missing data occurs in most of the data sets and can have significant effect on the final conclusion. Therefore, they should be handled in the preprocessing. Usually, for a missing value that the percentage is less than 5% the solution is to drop the values from the analysis.

The check for the missing values performed on the dataset shows that there are not missing values.

A screenshot of a cell phone

Description automatically generated

However, it does not mean that the data are ready to be modelled and tested. The summary statistics should be printed in order to examine:

* Standard deviation
* Mean of the values
* Minimum of the values
* Maximum of the values
* Count number of non\_NA/null observation.

A screenshot of a cell phone

Description automatically generatedTherefore, the *dataset.describe()* method was used to print the statistic for the attributes.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

### Data Visualisation

In order to better understand the data, the visualisation may be performed on the dataset. It is a part of exploring the data where a bar chart of heatmap may be generated and indicate the important correlation for between values in the dataset.

#### Histogram

The histogram is a part of function available in matplotlib and represents the distribution of data by forming it along the range. Finally, the bars are drawn to represent the data visually. (Seaborn, 2019)

The histogram drawn for coursework dataset represents the reason for absence distributed from the most popular till the less one.

A screenshot of a cell phone

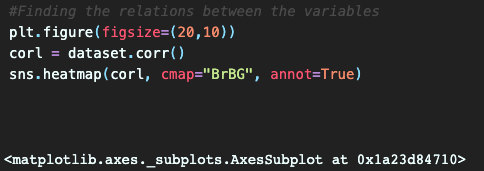
Description automatically generated

The histogram shows that the most popular reason of absence was the number 23. Comparing it to the data from the table provide in the beginning of this document it shows that number 23 relates to “Patient follow-up”. Therefore, the most popular absence reason was “Patient follow-up” where the second most popular is “Dental Consultation”

#### Heatmap

The purpose of drawing a heatmap is to distinguish the correlation between two features. It may indicate the features that may have the biggest impact on the result (the target).

Therefore, the heatmap for the coursework dataset was drawn using the following script:



A screenshot of a cell phone

Description automatically generatedAs a result of this the following heatmap was plotted and populated by the features.

The conclusion taken for the heatmap are:

* They may be high correlation between Weight and BMI – which is quite obvious
* The correlation between BMI Service time equals 0.49
* Most of the correlation may be negative correlation.

#### Scatterplot

Scatterplot help to find the correlation between the two variables. They are plotted together, and the diagram is generated which finally indicates if the correlation is negative, positive or zero.

By plotting the features of *Weight* and *Age* together, following scatterplot was generated.

A picture containing sky

Description automatically generated

According to me, the correlation between this two is difficult to distinguish. I may be somewhere between positive and zero correlation.

A screenshot of a computer

Description automatically generatedI have check another correlation between *Service time* and *Body Mass Index* which show quite low correlation between those two values.

# Data Tuning and Testing