**Big Data Coursework 2019-20**

**Big Data (MHI222956/MHI225101)**

Mateusz Hawrot S1633727

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

\

**“I declare that all work submitted for this coursework is the work of Mateusz Hawrot alone unless stated otherwise.”**

Table of Contents

[1 Data Pre-processing and Analysis 3](#_Toc26961042)

[1.1 Introduction 3](#_Toc26961043)

[1.2 Problem Description 3](#_Toc26961044)

[1.3 Data Analysis 6](#_Toc26961045)

[1.3.1 Data Types 6](#_Toc26961046)

[1.3.2 Check for duplicates and missing values 7](#_Toc26961047)

[1.3.3 Data Visualisation 9](#_Toc26961048)

[2 Model, Tune and Test 14](#_Toc26961049)

[2.1 Building a model, training, tuning and testing 14](#_Toc26961050)

[3 Discussion 18](#_Toc26961051)

[4 Bibliography 20](#_Toc26961052)

# 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.

All the scripts implementation for coursework purpose are included in zip file which was submitted through Turnitin.

## Problem Description

The report purpose is to analyse the dataset of Absenteeism at work. It will require to implement a data analysis algorithm that will process the data and distinguish what has an impact on absenteeism at work. The algorithm chosen for this project is a Decision Tree.

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 dataset was accessed from the following link:

https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

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

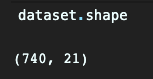


Figure 1 Number of attributes and instances

A screenshot of a video game

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

Figure 2 First 10 attributes

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

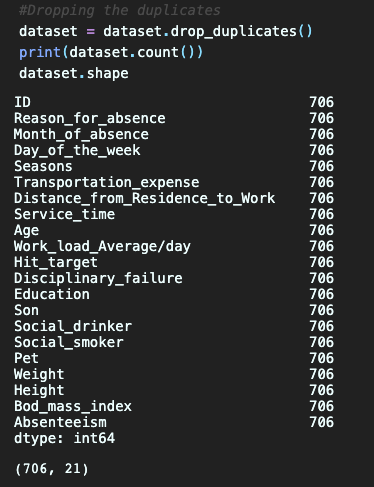
Description automatically generated

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

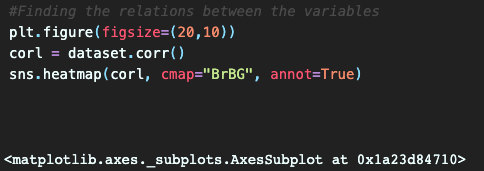
Description automatically generated

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 data gathered, 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 generatedSecond correlation between *Service time* and *Body Mass Index* was plotted and the results shows quite low correlation between those two values. This correlation may be assumed as zero correlation.

#### Storing processed dataset

Once the raw dataset went through the pre-processing procedures, the final version of the dataset will be store into new file and will be ready for implementation of machine learning algorithm.

In order to store the data into CSV file, the *to\_csv* method was run.

Performing the following script, the dataset was store into *preProcessedData.csv*

A picture containing object

Description automatically generated

# Model, Tune and Test

The first part of constructing the model is to decide which supervised or unsupervised machine learning algorithm will be performed on dataset. For the coursework purpose the Decision Tree algorithm will be performed on the dataset.

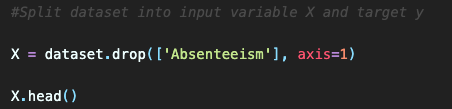
Decision Trees are versatile Machine Learning algorithms which can perform classification and regression task. They are capable of fitting complex dataset. (Geron, 2019). A decision tree is drawn upside down with its root on the top, and its branches/edges underneath. One of the advantageous of this machine leering method is the fact that decision tree can handle numerical and categorical data with multiple output problem. Moreover, they are simple to understand and visualise. Decision tree can determine the best attribute of the dataset which is a big advantage.

The information gained above has driven me into implementation of decision tree supervised machine learning into my dataset. Therefore, the Decision Tree Classifier will be used in the next steps of the project.

## Building a model, training, tuning and testing

The first step of this process is to import modules and load the dataset which was preprocessed before. Determining the shape of the model indicates that it has 706 attributes and 21 features. The reason for that is the data went through the cleaning process.

The next step is to prepare the data for the training. Therefore, they need to be separated into two variables, where inside the first one all the features will be loaded apart from the target.  
In this case, the *Absenteeism* is a target, thus it needs to be dropped from the X variable.



A screenshot of a cell phone

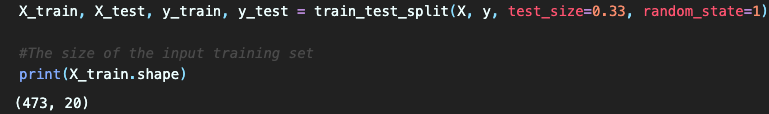
Description automatically generatedThe target will be loaded into a second variable which will prepare the set of data for the training. The *Absenteeism* needs to be loaded in e loaded in *y* variable. To test whether the attributes were loaded successfully, the first 10 of them were printed.

The operation of splitting data as variable X and target y prepared the data for inserting them into training and tasting. To accomplish this, the *train\_test\_split* method imported from *sklearn.model\_selection* wil be used. What it does is essentially splitting arrays or matrices into random train and subsets. The following parameters can be set and modified:

* **test\_size** – specifies the percentage of data includes into the test split
* **train\_size** – specifies the percentage of data includes into the train split
* **random\_state** – species the seed used by the random number generator
* **shuffle** – boolen – whether to shuffle or not the data

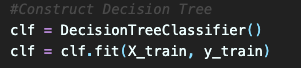
For the coursework purpose the test\_size is set to 0.33 which is an equivalent to 33% and random stated to 1.

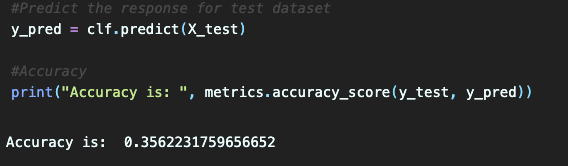
The operation shown below splits the data and prints the shape of the train variable.



It indicates that 473 attributers are used for the train.

Once the data are separated, the Decision Tree is constructed, and data are fitted into the classifier. This process analyses the data in supervised machine learning and splits them accordingly to a certain parameter.

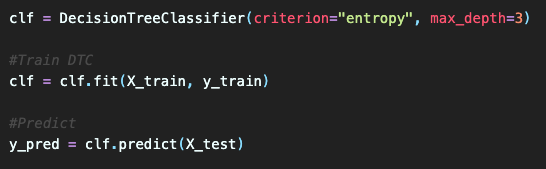


Once the data went through the analyse process, the response for the dataset should be predicted and assigned into the variable. Using the metrics, the test data and predicted response is compared in order to determine the accuracy score.

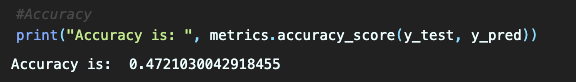
The response from the metrics indicates that the accuracy of the model is 36%. This is not a high score that may satisfy as the model may be accurate once per three attempts.

Therefore, model shall be tuned. Tuning in machine learning is a process where model’s performance is maximized without overfitting or creating too high variance. (Anderson, 2017).

The screenshot shown below describe the tuning process where the classifier criterion is assigned to *entropy*. The data are fit into the Decision Tree Classifier again the the prediction is evaluated.



The evaluation of model accuracy by metric accuracy score indicates that the model accuracy has risen into 47%.



This indicates a huge improvement of the model where the correctness is nearly every second time.

A screenshot of a cell phone

Description automatically generatedIn order to visualise the Decision Tree, the implementation of the graphviz is used. The attributes are assigned as decision tree features and the visual decision tree is generated. The following script indicates the process of creating the visualisation.

As a result, the following visualisation was generated:

A screenshot of a cell phone

Description automatically generated

# Discussion

The accomplishment of the coursework helped me to understand the process of supervised machine learning performed on the dataset. The processes taken in the coursework were analysing and processing the data, construct the model, tune it and testing. Those steps allowed me to come to the following conclusion.

The pre-processing of the dataset reduced the instances from the 740 to 706 by removing the duplicates. By the further analyse of the data I have realised that some of the attributes contains minimum value. However, I did not deal with that in account of the fact that some of the data are represented as 0 = False and 1 = True (This is specified in the model description).

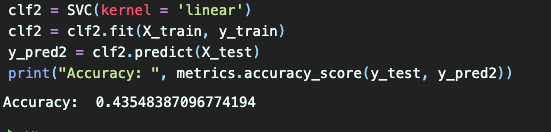
The correlation determined by scatteroplot and heatmap illustrate that the correlations are quite poor and they are close to zero correlation.

For the machine learning algorithm, I have chosen the Decision Tree. However, during the entire process I have also tried implementation of the following algorithms:

**Linear Regression** – the results given by that algorithm were worse than the Decision Tree what is indicated by calculating the MAE, MSE, and RMSR. The results are shown below

A screenshot of a cell phone

Description automatically generated

**SVM** – again, the accuracy given after tuning the model was 43%, what is slightly lower than in Decision Tree.

To sum up, in total I have implemented three different supervised machine learning algorithms. This was not required by the coursework, but I wanted to determine whether different algorithms may increase the performance and accuracy.

The final accuracy of decision tree is equal to 47%. In comparison with and without the tuning the different is significant. However, I am not satisfied by the result which is lower than 50%. I come up with two facts that may have impact on the accuracy of the model:

* The dataset was difficult to process, where the *Reason\_for\_absense* was represented by numerical value. Moreover, some of the atributes were set to Boolean (1 = True, 0 = False). Therefore, I could have changed the dataset to different one. However, I decided to work on the difficult data and try to learn something new rather than search for simple dataset.
* Another factor that may have impact on the dataset is pre-processing. I think that if advanced knowledge of the pre-processing data would be implemented the final result may or may not be better. Due to lack of the time I could not dive into advanced data pre-processing.

# Bibliography

Anderson, B., 2017. *Tuning Machine Learning Models.* [Online]   
Available at: https://riskspan.com/news-insight-blog/tuning-machine-learning-models/  
[Accessed 2 12 2019].

Geron, A., 2019. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow.* Second Edition ed. s.l.:O'Reilly.

Seaborn, 2019. *Visualizing the distribution of a dataset.* [Online]   
Available at: https://seaborn.pydata.org/tutorial/distributions.html  
[Accessed 1 12 2019].

Zhang, A., 2018. *Data Types From A Machine Learning Perspective With Examples.* [Online]   
Available at: https://towardsdatascience.com/data-types-from-a-machine-learning-perspective-with-examples-111ac679e8bc  
[Accessed 1 11 2019].