MAP-i PhD Program

Adaptive Business Intelligence - ABI

Task 2 ABI Project Proposal

Group Members:

1 Name: Nuno Simões **email: up201801786@fc.up.pt**

2 Name: Vanessa Silva **email: vanessa.silva@dcc.fc.up.pt**

Aim (what do you intend to do? business impact)**:**

The chosen dataset, which will be further discussed in the next question, is related to absenteeism at work. In this work we intend to analyze the data collected on the referred thematic - absenteeism with a view to two main objectives. The first is related to the prediction of absenteeism at work, where we intend to initially simulate the techniques of the paper related to the dataset ("Application of a neuro fuzzy network in prediction of absenteeism at work", authors: A. Martiniano, R. P. Ferreira e R. J. Sassi, C. Affonso), and later test other algorithms and also carry out transformations/selections of attributes for this predictive analysis. The second objective is to try to use the dataset in an optimization study, an example would be to allocate costs for absences to work, take on a given team of employees and try to optimize the hiring of new employees in order to minimize abstention. We will use statistical analysis and Data Mining techniques to extract useful knowledge of the dataset and try to identify patterns of behavior of absentee employees from non absentee employees. Analyzing all the data we will try to arrive at a standard or a conclusion regarding the assiduity profile of each employee.

This subject has relevance, especially for the employer, because depending on some attributes one can try to predict the attendance of an employee. Taking this in attention, this point is directly related to the business or financial theme of a company, since absenteeism represents loss of productivity and profit. In this way, knowledge of the reasons that lead to employee absenteeism is fundamental to the tasks of management, planning and organization of companies.

ABI Dataset(s) selected (give details about name, domain, source, main characteristics, etc...)**:**

The working dataset is present at: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work#. The dataset is related to a brasilian courier company. The dataset name is "Absenteeism at work - Part I" and as the name suggest is related to the absenteeism in the employees of the referred company.

One of the characteristics is that these data were collected from July 2007 to July 2010, so the range of values turns out to be comprehensive. The creators of it are: Andrea Martiniano, Ricardo Pinto Ferreira and Renato Jose Sassi.
This dataset has 21 attributes and 740 instances, the "Reason for absence" attribute seems to be the most relevant. This taking into account the various characteristics associated with this attribute. These are also further elaborated by the authors in a document accompanying the dataset. The remaining attributes are directly related to absenteeism (such as duration, date, season, ...), and others directly related to the employee (such as physical characteristics, habits, ...). The data set does not contain missing values and refers to categorical, integer, and real attributes.

ABI Computational Tool(s) selected (give details about name, source, characteristics, etc...)**:**
The tools that we predict to use are essentially R (using various packages of data analysis and data visualization), and R Markdown to aid in the development of a tutorial.

Expected Results/Outcome (what do you believe you can achieve with this Project?)**:**
We believe that we will be able to discover relevant information on the statistics of the collected data, in order to draw relevant conclusions for a good management of the company. We also want to achieve the prediction results for the article related to the dataset as well as try to improve those results.
We intend to find a profile for absentee employees and non-absentee employees so that we can try a profile to predict the presence or absent of an employee.
Finally, we hope to introduce an optimization approach to this data set, showing that it is possible and beneficial for hiring new employees.