CS36110: Machine Learning - Assignment 1
Employee Absenteeism

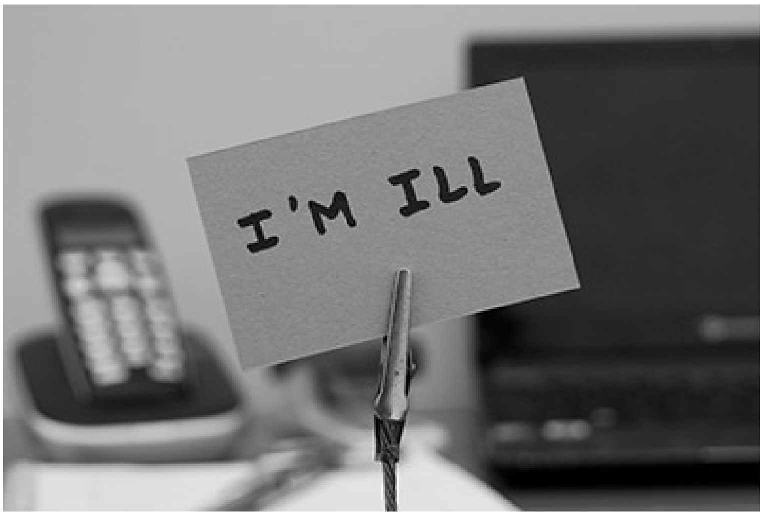
Release: 14th October 2019
Hand-in: 25th November 2019 @ 12:00pm
Feedback: 16th December 2019

This is the first and only assignment for CS36110, and comprises 50% of the total mark for the module. It will be assessed according to the Department's assessment criteria for essays. In particular, marks will take account of understanding of the problem, challenge of the chosen task, completion of the task and quality of the presented material. Other marks will cover knowledge of the lecture material, justification for the choices made, and quality of analysis.

Please submit your work via TurnItIn on Blackboard **before 12pm on Tuesday, 25th Nov. 2019**.

1 Introduction

This assignment is based upon a dataset relating to employee absenteeism. Employee Absenteeism is the absence of an employee from work. Its a significant problem for almost all employers. The data used in this assignment was gathered from a South African manufacturing company that was experiencing a loss of productivity as a result of employees being absent from their place of work. The company in question would like to gain a better understanding of questions such as: "What changes should we make to reduce the number of absences?" or "How will absenteeism affect productivity (and thus profit margins) if the same levels of employee absenteeism continue?"



(*Image: tint.com)*

The dataset essentially describes a number of attributes related to employees including: *no. of children*, *body weight*, whether the person is a *social smoker* or drinks alcohol, etc. - a full description of all of the attributes is provided in Appendix 1. The aim is to use the data characteristics in order to gain an insight into the reasons for employee absences and thus help to address the problems mentioned previously.

The data to be used in this assignment consists of data objects (or instances), with each object representing one event of absence from work. The dataset, `employee_abs.arff`, contains a total of 677 data objects and has 16 features (or attributes) including the decision class label. The decision feature can take four different values:

A, B, C or D, each representing a period of absence from a few hours to multiple days. The goal is to predict the length of absence based on the values of the conditional features.
The features of the dataset include information such as the `Reason_for_absence` (this is a medical coding, or `awol` representing an unauthorised absence), `Day_of_the_week` (the day on which the absence occurred) and `YearsService` (no. of years the employee has work at the company), etc. Further detail regarding the features of the data are provided in Appendix 1 of this document. It is important to note that some of the feature-values in the data are missing and these are encoded in the dataset as '?' (further discussion regarding this particular issue in section 3). The dataset can be used directly to perform experiments in WEKA[1].

2 General Guidelines

In the following tasks you are asked to run experiments, and analyse the results using **WEKA (version 3.8.3)**. If you are working on a machine outside of the departmental network (e.g. your laptop), **please ensure that you download and install this exact version. This is very important as your results need to be reproducible on a departmental machine.**

In many cases, the WEKA Explorer allows you to modify the different parameters of the learners you might use. You are welcome (and encouraged) to change these. However, you should have a good justification and explain clearly in your report if/why you are doing this. You will have to work out some details in applying WEKA Explorer, including definitions of certain terminologies (such as 'cross-validation' as referred to above). The `employee_abs.arff` file that you will be using is available from Blackboard.

3 Tasks

All of the experimentation and analysis for this assignment will be carried out using WEKA *Explorer*. The results must be generated using cross-validation (CV). These results can then be used to evaluate the performance of the classifier learners. You should use 10-fold cross-validation (the default setting) for your experimentation. You may use another setting for this, however, if you do so, you must have good well-argued reasoning for doing so.

1. Load the `employee_abs.arff` dataset into WEKA. Using the *Classify* tab in WEKA Explorer, train and evaluate different classifiers on the dataset using 10-fold CV. You should use both Naive Bayes (`NaiveBayes`) and C4.5 Decision Tree (`J48`), although you are not limited to these specific classifier learners. If you use others however, you <u>must</u> understand and explain how they work. If you change any of the tuneable parameters for the classifiers you must also provide a clear explanation and justification for doing so. **(30%)**
   (a) Briefly describe the classifiers you have used and explain and discuss the advantages of each. (10/30)
   (b) Compare the results **(using percent correct, error rate or any other metric you consider appropriate)** for the different classifiers you have chosen. What do you notice? (12/30)
   (c) What is a reasonable baseline against which to compare the classification performance results and why? (8/30)

2. In the introduction section, it is mentioned that some object-feature values have '?' assigned to them. It is important to deal with these if you believe that they have an effect on your results. **(30%)**
   (a) One way of addressing this might be to use one of the filters provided in WEKA to replace such values with something else. You could also do some manual manipulation of the data. In such cases, you should check the data carefully, use your best judgement and **clearly** state any assumptions you make when performing any subsequent experiments. (10/30)
   (b) Explain in your report whether you think a method of replacing missing values is appropriate in the case of the `employee_abs.arff` dataset. Also, you should indicate how many values were replaced and for which features (10/30)
   (c) Using whatever method you deem appropriate in part a) above to deal with missing values, re-train and evaluate the classifiers that you have been using in Task 1 with 10-fold cross-validation. Compare the parameters of the model learned by the classifiers for the modified and unmodified datasets. Discuss any differences and comment on performance. (10/30)

   Note: if you chose to apply a filter, or indeed if you choose to manually change the dataset, performing either of these steps will result in a modified dataset.

[1] WEKA: Waikato-Environment-for-Knowledge-Analysis `http://www.cs.waikato.ac.nz/ml/weka/`

3. Using WEKA, remove the following features from the dataset: `Month_of_absence`, `Day_of_the_week`, `Season`, `CommuteDistance`, `YearsService`, `WorkLoad`, `Education`, `Smoker`. This now leaves you with a reduced dimensionality dataset. **(20%)**
   (a) Compare the parameters of the models learned by the `J48` classifier with those from Task 1. What can you say about the structures of the trees generated for this dataset and the original analysed in Task 1? (12/20)
   (b) Discuss why you think this classification problem is similar or different from the one in Task 1. (8/20)

In your report, it is important to carefully summarise your findings and analysis for the three previous tasks and state clearly which approaches you have used and which you consider are suitable and which are not.

4 Submission and Marking

You are required to submit a report in `.pdf` format via TurnItIn on Blackboard. You should aim to keep your answers concise, while conveying the important information. **A report of 1,600 words (+10%) (excluding references, footnotes and tables)** is appropriate for this. You should also include the TurnItIn word count in your document so that it can be verified. Please do not exceed the word limit as this will delay the marking process and incur a penalty. The following marking scheme will be used to mark your submission:

- Task 1 - Describe the classifiers you have used and analysis of the results (as described previously). (30%)
- Task 2 - Discuss and investigate the options you chose for dealing with missing data (as described previously). (30%)
- Task 3 - Compare and contrast the results for the `J48` tree-based classifier for the two datasets. (20%)
- Report: readability, correct formatting, layout and proper referencing. (20%)

You should aim to keep your answers concise, while conveying the important information.

Plagiarism: *One of the dangers of this assignment is the temptation to use paragraphs from web documents or papers that you have read. Please resist this temptation and do not do this. Otherwise, you will be heavily penalised. The report should be completely in your own words. If it is appropriate and absolutely necessary to include sentences and materials from elsewhere, then they should be clearly indicated as quotes, and references should be cited. Please do not share your findings or show your analysis or report to other students.*

N. Mac Parthaláin, 10-2019.

Appendix 1. - Descriptions of the Features in the employee_abs.arff dataset.

A brief description of the features in the order they appear:

1. Reason_for_absence - This indicates the reason given for the employees absence. It can take one of 27 different values. The value awol indicates that this is an unauthorised absence. The remaining 26 are related to medical conditions or doctor/hospital visits:

   abnor - Congenital malformations, deformations and chromosomal abnormalities
   blood - Diseases of the blood and blood-forming organs
   circ - Diseases of the circulatory system
   digest - Diseases of the digestive system
   dentist - Dental visit or consultation
   ear - Diseases of the ear
   ext - External causes of morbidity and mortality
   endo - Endocrine, nutritional and metabolic diseases
   eye - Diseases of the eye
   GP - Doctors visit or consultation
   GU - Genitourinary infections or conditions
   injur - Physical injury or trauma
   inf - infectious or parasitic diseases
   hosp - Hospital visit or consultation
   malform - Congenital malformations, deformations and chromosomal abnormalities
   mental - Mental and behavioural disorders
   muscle - Diseases of the musculoskeletal system and connective tissue
   neo - Neoplasm
   nerv - Diseases of the nervous system
   perinat - Complications relating to perinatal factors
   preg - Pregnancy or childbirth related complications
   physio - Physiotherapy session or consultation
   resp - Diseases of the respitory system
   skin - Epidemiological conditions or diseases
   scan - Hospital or clinic imaging visit.
   status - Factors related to healthcare access

2. Month_of_absence - the month of the year in which the absence occurred (1 = Jan, 12 = Dec)

3. Day_of_the_week - day of the week on which the absence occurred.

4. Season - which season of the of the year the absence occurred in (spring, summer, autumn or winter).

5. TravelCost - the cost of travelling to place of work from employee's home (real-valued).

6. CommuteDistance - the distance from employee's home to place of work (real-valued).

7. YearsService - number of years service as an employee (integer valued).

8. WorkLoad - employee workload as an average across the company (real-valued).

9. Target - employee performance metric (real-valued [0,100]).

10. Education - Level of education achieved by the employee:
    1 = no formal education/secondary school
    2 = undergrad degree
    3 = postgrad masters
    4 = PhD/MBA.

11. Children - number of children in the employee's household.

12. Alcohol - whether the employee is a social drinker of alcohol, (binary: Y = yes, N = no).

13. Smoker - whether the employee is a social smoker (binary: Y = yes, N = no).

14. NoOfPets - the number of pets at the employees home (integer valued).

15. `BodyWeight` - body weight in kg (real-valued).

16. `Absent` - length of absence:

    `A` = 0-1 days
    `B` = 1-2 days
    `C` = 2-3 days
    `D` = 4 or more days