Finding Risk Factors for Long-Term Sickness Absence Using Classification Trees

INA LUNDSTRÖM

Master of Science Thesis
Stockholm, Sweden 2013

Finding Risk Factors for Long-Term Sickness Absence Using Classification Trees

INA LUNDSTRÖM

Master's Thesis in Mathematical Statistics (30 ECTS credits)
Master Programme in Mathematics (120 credits)
Royal Institute of Technology year 2013
Supervisor at HealthWatch was Dan Hasson
Supervisor at KTH was Timo Koski
Examiner was Timo Koski

TRITA-MAT-E 2013:51
ISRN-KTH/MAT/E--13/51-SE

Royal Institute of Technology
School of Engineering Sciences

KTH SCI
SE-100 44 Stockholm, Sweden

URL: www.kth.se/sci

Abstract

In this thesis a model for predicting if someone has an over-risk for long-term sickness absence during the forthcoming year is developed. The model is a classification tree that classifies objects as having high or low risk for long-term sickness absence based on their answers on the HealthWatch form. The HealthWatch form is a questionnare about health consisting of eleven questions, such as "How do you feel right now?", "How did you sleep last night?", "How is your job satisfaction right now?" etc. As a measure on risk for long-term sickness absence, the Oldenburg Burnout Inventory and a scale for performance based self-esteem are used.

Separate models are made for men and for women. The model for women shows good enough performance on a test set for being acceptable as a general model and can be used for prediction. Some conclusions can also be drawn from the additional information given by the classification tree; workload and work atmosphere do not seem to contribute a lot to an increased risk for long-term sickness absence, while job satisfaction seems to be one of the most important factors.

The model for men performs poorly on a test set, and therefore it is not adivsable to use it for prediction or to draw other conclusions from it.

Contents

1 Introduction

1.1 Background

In this thesis the technique of classification trees is used to make a model that shows if someone is likely to have an increased risk of long-term sickness absence during the forthcoming year based on their answers on eleven questions concerning self-rated health and the psychosocial work environment.

A study conducted by Hallsten et al [2] showed that persons with high levels of exhaustion and disengagement and a high performance-based self-esteem (PBSE) have an over-risk of long-term sickness absence, the odds ratio for those persons is 2.84 with 95% confidence interval $1.61 - 5.01$ [2, p. 187].

In this thesis the Oldenburg Burnout Inventory (OLBI) is used as a measure on exhaustion and disengagement. OLBI is a questionnare with both positively and negatively framed questions that try to capture exhaustion and disengagement, which are the two core dimensions of burnout [1]. The questionnare consists of five questions related to exhaustion and five questions related to disengagement. The questions are answered on a scale that ranges from 1 to 5 and the average of the score on the exhaustion-questions and the average of the score on the disengagement-questions are computed. A person has high levels of exhaustion and disengagement if their averages on those two sets of questions are larger than 2.6 and 2.75 respectively. OLBI results in an index 1 and 0, where 1 means high exhaustion and disengagment and 0 means either low levels of both or low levels of one of them. For more information about OLBI and burnout, see [5] and [1].

Performance-based self-esteem is described in [4] as self-esteem that depends on how well one performs in roles that are important for one's self-realisation; "self-esteem primarily built on accomplishments and "doing" rather than on "being" or "having" are called performance-based self-esteem" [4, p. 5]. For a more detailed explanation and discussion about PBSE and its role in the burnout process, see [4].

As a measure on high PBSE the Pbse-scale was used. It consists of four questions with five response alternatives with end-points labeled "Fully disagree" and "Fully agree". The response scale is translated into the numbers 1-5 and the score on the Pbse-scale is the mean of the scores on the four questions. PBSE was considered high if the score was larger than the median score of the data set under study. There exist other ways of setting the limit for high PBSE, but since the median was used in the study [2] by Hallsten et al it was used for this thesis as well.

The eleven questions about health that are used as input variables in the model are the questions of the HealthWatch form. "HealthWatch currently consists of surveys for assessment of the phsychosocial work environment and occupational health interventions that can be used at the individual, group and organizational levels." [6, p 222]. The main part of HealthWatch is the form consisting of the eleven questions about one's health. These questions will be referred to as the HW11-questions or just HW11 from now on.

The HW11-questions are:	Label:
How do you feel right now?	Srh (self-rated health)
How did you sleep last night?	Sleep
How is your ability to concentrate right now?	Concentration
How stressed do you feel right now?	Stress
What is your energy level right now?	Energy
Do you have control over your life right now?	Control
How satisfied are you with your social life right now?	Social
How efficient are you at work right now?	Efficiency
How is your job satisfaction right now?	Workjoy
How high is your work load right now?	Workload
How is the job atmosphere right now?	Workatm

All of the questions are answered on a scale ranging from 0 to 100, but the person who is answering a question sees a scale with endpoints labeled "very bad" and "very good" (or equivalent statements depending on the question).

The outline of this thesis is as follows. In section 2 classification trees will be introduced and the mathematical theory behind it will be presented.
Some other mathematical concepts that were used in building the models of this thesis will also be presented.
After the sections of mathematical theory follows section 3 describing the method and data that were used. Finally the results are presented followed by a discussion about possible intrepretations and conclusions.

Terminology and definitions will be introduced when needed.

1.2 Classification Trees

Classification trees is a nonparametric statistical method based on recursive partitioning. The basic idea is predicting membership of objects in the classes of a categorical dependent varibale from their measurements on some predictor variables.
For example, the dependent variable could be health status with the two classes healthy and sick, where healthy could mean not suffering from a specific illness and sick could mean suffering from this specific illness. The predictor variables could be different symptoms and variables related to the patients' backgrounds (age, gender, education etc.). A classification tree for this problem could look like in Figure 1.

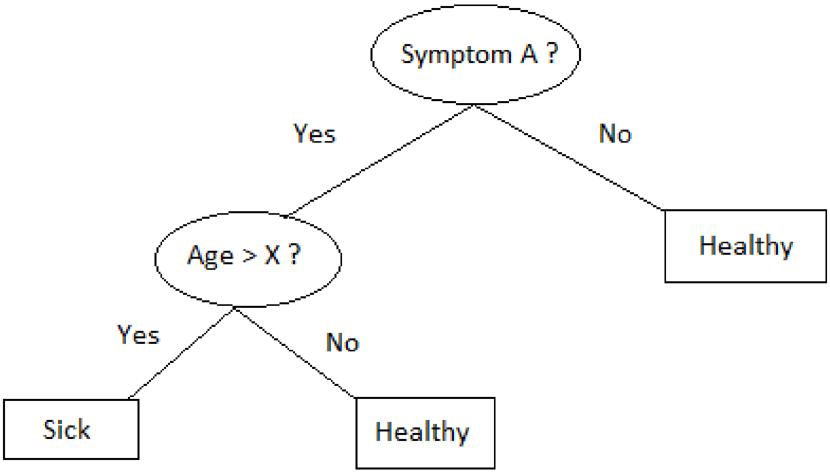

Figure 1: A hypothetical classification tree.

Some of the benefits with classification trees are that they are relatively simple to understand and that the results are easy to interpret without any deeper knowledge in mathematics.

The procedure for interpreting results is always the same. Start at the first node. Answer the question and choose branch and next node accordingly. Continue down the tree until a terminal node is reached. A terminal node is a node that does not split and that shows class membership. Each different path leading to a terminal node gives a combination of factors leading to a class.

The results of the classification tree in Figure 1 are interpreted as follows. If a patient is suffering from symptom A and is older than X he is sick. If he is suffering from symptom A and is younger than X he is healthy. If the pateint is not suffering from symptom A he is healthy.

Classification trees can be used both for prediction and for interference. Apart from the actual classification tree, information about importance and similarity of variables is also provided by the model, which can be used for analysing relationships between variables.

In this thesis the dependent variable has the two classes "over-risk for long-term sickness absence" denoted "1" and "no over-risk for long-term sickness absence" denoted "0". The HW11-questions are the predictor variables. Since the dependent variable only has two classes this is a binary classification problem.

Both prediction and interference are of interest in this thesis. It is desirable to be able to tell if someone is likely to have an over-risk for long-term sickness absence based on how they answer the HealthWatch form. Discerning relationships between the HW11-questions is also of interest because it can provide more information about how these different variables affect our health.

2 Theoretical background

2.1 Growing a Classification Tree

A classification tree is constructed using a *learning sample*, denoted by $\mathcal{L}$. A learning sample consists of measurement data on N objects together with their actual classification:

$$\mathcal{L} = \{(\mathbf{x}_1, j_1), ..., (\mathbf{x}_N, j_N)\}$$

where $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,n})$ is a vector containing measurements on each of n predictor variables and $j_i \in \mathcal{C}$, where $\mathcal{C} = \{1, 2, ..., J\}$ is the set of all J different classes.

Growing a classification tree is basically about three things:

- Selecting the splits.
- Assigning a class to each terminal node.
- Finding the right sized tree.

Definitions
Parent node and Daughter nodes: The daughter nodes of a parent node are the two nodes the parent node splits into.
Root node: The first/top node of the tree, i.e. the node that has no parent node. A classification tree has exactly one root node.
Terminal node: A node that does not split, i.e. that has no daughter nodes. Shows class membership. A classification tree has at least two terminal nodes and there can be more than one terminal node of the same class. Terminal nodes are denoted by $\tilde{T}$

2.2 Selecting splits

Selecting splits means selecting which criterion should be used to split a given node, i.e. which predictor variable should be split and based on what criterion should it be split. The basic idea of how to select splits is very simple. At each node, calculate the best split for each variable. Among all these best splits, choose the best one and use it as splitting criterion.

Number of possible splits

The number of possible splitting criteria for each variable depends on if the variable is categorical or ordinal. For a categorical variable with M distinct categories this number at a given node is $2^{M-1} - 1$.
For an ordinal variable the number of possible splitting criteria at a given node is one fewer than the number of its distinctly observed values. If for example age is a predictor variable and its observed values are all integers between 20 and 40, its possible splitting criteria are $x \leq 21, x \leq 22, ..., x \leq 40$.

Goodness of split criterion

To pick out the best split, a criterion for the goodness of a split is needed. The fundamental idea is to select each split so that the data in the daughter nodes are purer than the data in their parent node. A node that is perfectly pure, or homogeneous, contains objects of only one class. A node that is perfectly impure contains equal proportions of objects of each class, i.e. all classes are equally mixed.

Definitions
Node proportions The node proportions of node t, $p(j|t)$, $j = 1, 2, ..., J$, is the proportion of the objects in node t belonging to class j, so that $p(1|t) + ... + p(J|t) = 1$
Impurity measure An impurity measure is a measure $i(t)$ of the impurity of node t and it is defined as a nonnegative function ϕ of the node proportions $p(1|t), ..., p(J|t)$, $\phi(p(1|t), ..., p(J|t))$ such that

$$\phi(\frac{1}{J}, \frac{1}{J}, ..., \frac{1}{J}) = maximum$$

$$\phi(1, 0, ..., 0) = 0,\ \phi(0, 1, ..., 0) = 0,\ \ldots\ , \phi(0, 0, ..., 1) = 0$$

ϕ is a symmetric function of $p(1|t), ..., p(J|t)$

For node t, suppose that there is a possible split s with daughter nodes t_L and t_R and that a proportion p_L of the objects in t goes into node t_L and a proportion p_R goes into node t_R.
Then the goodness of the split is defined to be the decrease in impurity

$$\Delta i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

So the best split is the split with the larger $\Delta i(s, t)$.

At node t the best split is chosen as follows. For variable i, the best split s_i^* is obtained from

$$\Delta i(s_i^*, t) = \max_{s_i \in S_i} \Delta i(s_i, t)$$

where S_i is the set of all possible splits of variable i at node t.
The best split s^{**} is then obtained from

$$\Delta i(s^{**}, t) = \max_{s_i^* \in S^*} \Delta i(s_i^*, t)$$

where $S^* = \{s_1^*, s_2^*, ..., s_n^*\}$ is the set of the best splits of all n variables. s^{**} is the better split of the best splits.

Node impurity functions

Two commonly used node impurity functions when the dependet variable is categorical are:

$$\text{Entropy function: } i(t) = -\sum_j p(j|t)\log p(j|t)$$

Gini index: $i(t) = 1 - \sum_j p^2(j|t)$

In a binary classification problem they reduce to:

Entropy function: $i(t) = -p \cdot \log p - (1-p) \cdot \log(1-p)$

Gini index: $i(t) = 2p \cdot (1-p)$

There is not much difference between these two node impurity functions in practice and the properties of the final tree does not depend much on which impurity function that was used [3, Ch 2]. But there is a drawback with the Gini index, it favors end-cut splits [3, Ch 11]. End-cut splits are splits that are extremely unbalanced in size, one of the p_L and p_R is very close to zero or one. Because of this problem the entropy impurity function was used for the models in this thesis.

2.3 Assigning a class to each terminal node

Suppose a tree T with terminal nodes $\tilde{T}$ has been constructed. Every terminal node is assigned a class by the use of a *class assignment rule*, which is defined as:
Class assignment rule (Definition 2.9 in [3])
A class assignment rule assigns a class $j \in \{1, ..., J\}$ to every terminal node $t \in \tilde{T}$. The class assigned to node $t \in \tilde{T}$ is denoted by $j(t)$.

The following rule is used to assign a class to each terminal node:

$$j^*(t) = i_0 \text{ where } i_0 \text{ minimizes } \sum_j C(i|j)p(j|t)$$

where $C(i|j)$ is misclassification cost, defined in Section 2.4 below.

2.4 Finding the right sized tree

Getting the right sized tree is a crucial part of the tree-growing procedure. The most apparent approach might be to have a criterion for when to stop splitting nodes further. But then there is a big risk that the final tree will be too large or too small. A tree that is too large usually classifies the data in the learning set very well but does not work well as a general model. If the tree is too small some descriptive information in the data might be unused. Therefore another approach is used to obtain the optimally sized tree. The idea is to first grow a very large tree T_{max} and then to prune it upward in the right way. How this is done is described in the subsequent sections.

Misclassification cost

In the process of growing a tree some account is always taken to the cost of misclassifying objects. A perfect tree would always make correct classifications,

but this is not possible in practice. A tree can indeed be grown so far that the terminal nodes consist of objects of precisly one class, but the performance of this tree on data other than the learning set will most likely be very poor.
A tree that works well as a general model will always make some misclassifications. In most cases, especially in medicine, it is considered more serious to misclassify some objects than others. For example, the consequences of misclassifying an ill person as healthy is often much worse than misclassifying a healthy person as ill. The misclassification cost is defined as follows:

Misclassification cost (Definition 2.12 in [3])
$C(i|j)$ is the cost of misclassifying a class j object as a class i object and satisfies:

$$C(i|j) \geq 0, i \neq j$$

$$C(i|j) = 0, i = j$$

If all misclassifications that can occur are considerd equally bad the misclassification cost equals one for all i, j.
The upper bound for the misclassification cost is set as:

$$C(i|j) \leq \frac{\text{number of class-}i\text{ objects} + \text{number of class-}j\text{ objects}}{\text{number of class-}j\text{ objects}}$$

There exists no correct mathematical way of choosing which misclassification cost to use in a model. The usual way of choosing the "right" misclassification cost is to produce models for different misclassification costs and compute parameters of the performance of the model. The parameters that were used in this thesis are accuracy, precision, sensitivity and specificity. For binary classification, they are defined as:

$$\text{accuracy} = \frac{\text{number of true positives } + \text{ number of true negatives}}{\text{total number of tested objects}}$$

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of true positives } + \text{ number of false positives}}$$

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives } + \text{ number of false negatives}}$$

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives } + \text{ number of false positives}}$$

Where
True positive = a class-1 object that is classified as a class-1 object
True negative = a class-0 object that is classified as a class-0 object
False positive = a class-0 object that is classified as a class-1 object
False negative = a class-1 object that is classified as a class-0 object

Accuracy can be interpreted as probability of correct classification. Sensitivity

is interpreted as probability of classifying a class-1 object correctly and specificity is probability of classifying a class-0 object correctly.
Precision can be thought of as follows: if a person who in fact has an over-risk for long-term sickness absence answers the HW11-questions let's say every day during one week, high precision means that the model will classify the person correctly as class 1 (almost) everyday. If the precision is low the model will classify the person correctly as class 1 some days and incorrectly as class 0 some days.

Resubstitution estimate

An important parameter of a classification tree T is the probability of misclassification, $R^*(T)$. A simple, though inaccurate, estimate of $R^*(T)$ is the *resubstitution estimate*, $R(T)$. It is defined as:

$$R(T) = \frac{\text{the number of incorrectly classified objects}}{\text{the total number of classified objects}}$$

when the data in the learning set is classified by the tree.

Pruning combined with an improved estimation of $R^*(T)$ is the key to finding the right sized tree.

2.4.1 Pruning

Definitions
Descendant, ancestor: A node t' lower down on the tree is called a descendant or ancestor of a higher node t if there is a path leading from t to t'.
Branch: A branch T_t of T with root node $t \in T$ is the node t and all its descendants in T.

Then pruning a branch T_t from a tree T means deleting from T all descendants of t, that is removing all of T_t except for its root node t, which becomes a terminal node. The resulting tree is denoted by $T - T_t$ and is a subtree of T, denoted by $T - T_t \prec T$. This is illustrated in Figure 2

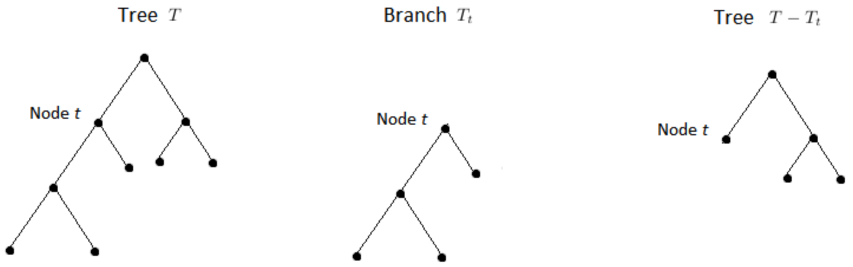

Figure 2: Pruning

In this thesis minimal cost-complexity pruning is used and therefore this type of pruning will be described in more detail.

2.4.2 Minimal Cost-complexity Pruning

Cost-complexity (Definition 3.5 in [3])
For any subtree $T \prec T_{max}$, define its complexity as $|\tilde{T}|$, the number of terminal nodes in T. Let $\alpha \geq 0$ be a real number called the complexity parameter and define the cost-complexity $R_\alpha(T)$ measure as

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

The idea is then to for each value of α find the subtree $T(\alpha)$ of T_{max} that minimizes the cost-complexity $R_\alpha(T)$.
α can be thought of as a penalty per terminal node. When α is small the penalty for having a big tree with many terminal nodes will be small so $T(\alpha)$ will be large. When α increases $T(\alpha)$ becomes smaller, i.e. it will have fewer terminal nodes, until α is so large that $T(\alpha)$ will be the root node only.
α runs through a continuum of values, but since there is a finite number of subtrees of T_{max} this pruning process will produce a finite decreasing sequence of subtrees $T_1 \succ T_2 \succ ... \succ \{t_1\}$ with fewer and fewer terminal nodes until the subtree consisting of only the root node t_1 is reached.
The next and final step in finding the right sized tree is to choose which one of these subtrees is the optimal one. To pick the best subtree a better estimate of $R^*(T)$ is needed. If the resubstitution estimate $R(T_K)$ would be used as a criterion, the largest tree T_1 would be selected. With a better estimate of the misclassification rate $R^*(T)$ the best subtree could be the one that minimizes $\hat{R}(T_K)$.
There are two commonly used methods for better estimation of $R^*(T)$; using an independent test sample and cross-validation. In this thesis cross-validation was used because it uses data more efficiently than a test sample does and it also gives useful information about the stability of the tree structure. The drawback is that it is more computationally expensive, but this did not turn out to be a problem.

2.4.3 Cross-validation

In V-fold cross-validation the learning set $\mathcal{L}$ is divided by random selection into V subsets $\mathcal{L}_v,\ v = 1, 2, .., V$. Each subset contains the same number of ojbects as nearly as possible. Then the v:th learning set $\mathcal{L}^{(v)}$ is the learning set without the ojbects in $\mathcal{L}_v$, i.e.

$$\mathcal{L}^{(v)} = \mathcal{L} - \mathcal{L}_v$$

Now V trees are grown, one on each learning set $\mathcal{L}^{(v)}$, together with the main tree grown using the whole learning set $\mathcal{L}$. All trees are then pruned as described above using cost-complexity pruning. For each value of α, $T(\alpha)$ and $T^{(v)}(\alpha)$ are the corresponding minimal cost-complexity subtrees of T_{max} and $T^{(v)}_{max}$. For each v, the set $\mathcal{L}_v$ can be used as an independent test sample for the tree $T^{(v)}(\alpha)$ since it was constructed without using any of the ojbects in $\mathcal{L}_v$.
Now for each v classify $\mathcal{L}_v$ according to the tree $T^{(v)}_{max}$. Fix α and for every value of v, i, j, define

$$N^v_{ij} = \text{the number of class } j \text{ objects in } \mathcal{L}_v \text{ classified as } i \text{ by } T^{(v)}(\alpha)$$

and the total number of j objects incorrectly classified as i is defined as:

$$N_{ij} = N_{ij}^1 + N_{ij}^2 + ... + N_{ij}^V$$

Now the idea is that $T^{(v)}(\alpha)$ should have about the same accuracy as $T(\alpha)$ if V is large, because then they are constructed using almost the same number of objects. The next step is to estimate $Q^*(i|j)$, the probability that a j object is classified as an i object, for $T(\alpha)$ as

$$Q^{CV}(i|j) = \frac{N_{ij}}{N_j}$$

and then set

$$R^{CV}(T(\alpha)) = \frac{1}{N} \sum_{i,j} C(i|j) N_{ij}$$

Remember that even though α varies continuously, the minimal cost-complexity trees grown on $\mathcal{L}$ are equal to T_K for $\alpha_k \leq \alpha < \alpha_{k+1}$. Put $\alpha_k' = \sqrt{\alpha_k \alpha_{k+1}}$. Then put

$$R^{CV}(T_k) = R^{CV}(T(\alpha_k'))$$

i.e. $R^{CV}(T_k)$ is gotten by classifying the test sets $\mathcal{L}_v$ according to the trees $T^{(v)}(\alpha_k')$. Now the rule for selecting the optimal tree is:

Select as right sized tree the tree T_{k_0} such that

$$R^{CV}(T_{k_0}) = \min_k R^{CV}(T_k)$$

and then use $R^{CV}(T_{k_0})$ as an estimate of $R^*(T_k)$

The position of the minimum of $R^{CV}(T_k)$ may be unstable, small changes in parameter values can cause large changes in $|\tilde{T}_k^{CV}|$ for the tree that minimizes $\hat{R}^{CV}$. To reduce this instability the 1 SE rule can be used for selecting the right sized tree.

The 1 SE Rule(Definition 3.18 in [3])
Define T_{k_0} by

$$\hat{R}(T_{k_0}) = min_k \hat{R}(T_k)$$

Then the tree selected is T_{k_1}, where k_1 is the maximum k satisfying

$$\hat{R}(T_{k_1}) \leq \hat{R}(T_{k_0}) + SE(\hat{R}(T_{k_0}))$$

where $SE(\hat{R}(T_{k_0}))$ is the standard error of $\hat{R}(T_{k_0})$.

2.4.4 Pruning in practice

It might sound as if pruning is a very complex step of the tree growing procedure, but in practice it is very easy and straight forward. In this thesis, the software environment R was used. R provides a table and a plot of the cross-validation results, called the cp table and the cp plot. The cp table contains one row for each tree that was grown during the cross-validation, i.e. one row per complexity parameter. The table has five columns, cp, nsplit, rel error, xerror, xstd. cp is a scaled version of the complexity parameter, nsplit is the number of splits in the tree, rel error is a scaled verson of the resubstitution estimate $R(T_K)$, xerror is a scaled version of the cross-validation estimate $R^{CV}(T_K)$ and xstd is the standard error of $R^{CV}(T_K)$. The columns rel error and xerror are always scaled so that the first node has an error of 1. The complexity parameter is scaled similarily. An example of a cp table is in Table 1. As is expected, the number of splits increases as the complexity parameter decreases.

The cp plot is a graphical illustration of the cross-validation results. An example is in Figure 3. The cross-validation estimate $R^{CV}(T_K)$ is on the y-axis and the complexity parameter is on the x-axis. The standard errors of the cross-validation estimates $R^{CV}(T_K)$ are shown by the vertical lines that go through the points. The dotted line shows the upper limit for the standard error of the minimal $R^{CV}(T_K)$.

The 1 SE-rule is now very easily implemented. Either look at the cp plot or at the cp table. Find the tree with minimal $R^{CV}(T_K)$, $R^{CV}(T_{K_0})$. Check if there is a smaller tree with $R^{CV}(T_K)$ within the range of the standard error of $R^{CV}(T_{K_0})$. If there is such a tree, then this is the optimally pruned tree. If the tree with $R^{CV}(T_{K_0})$ is itself the smallest tree within the range of its standard error, then this is the optimally pruned tree.

In the example in Figure 3 it is easily seen that the second point corresponds to the minimal $R^{CV}(T_K)$. Since there is no other point to the left of this point (which means a smaller tree) that lies within its standard error, the minimal point itself corresponds to the optimally pruned tree.

cp	nsplit	rel error	xerror	xstd
0.523810	0	1.00000	1.00000	0.174284
0.071429	1	0.47619	0.52381	0.091374
0.057143	2	0.40476	0.57143	0.096820
0.021429	4	0.29048	0.61429	0.115054
0.001000	6	0.24762	0.63333	0.119206

Table 1: cp-table

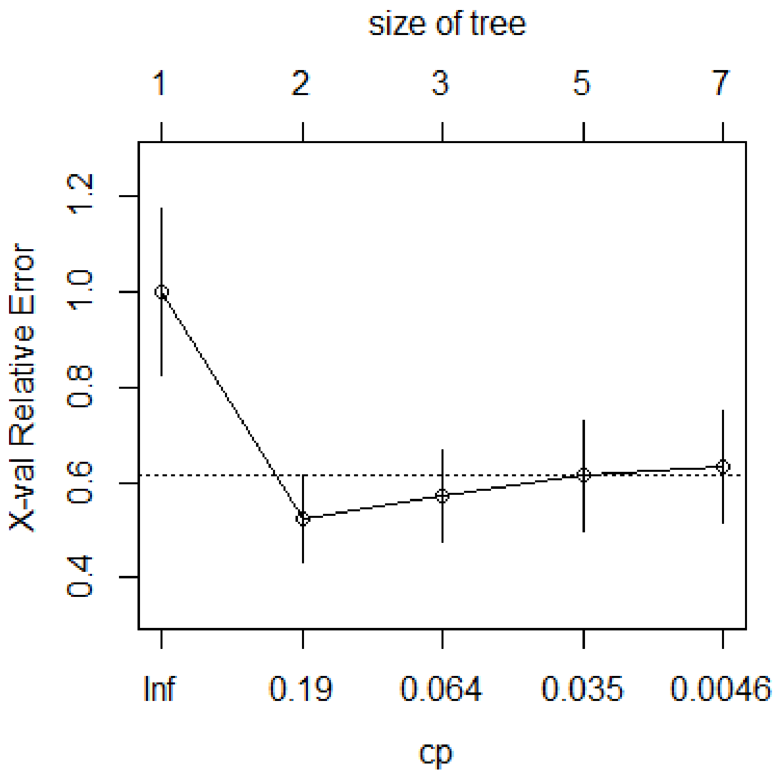

Figure 3: cp-plot

2.5 Missing data

Another benefit of classification trees is that they handle missing data in a good way. For example, if a tree contains splits on five different variables and an object that is to be classified by the tree has measurements on only four of them, the tree will use a so called *surrogate split* on the missing variable.
The idea with surrogate splits is to find a split of another variable that will send objects to the left and to the right as similar as possible as the real split of the variable that is missing.
For a formal definition of surrogate splits, let x_m be any variable and S_m the set of all splits on x_m. Let $\bar{S_m}$ be the set of all splits complementary to S_m. At a given node t, let s^{**} be the best split of t into t_L and t_R. For any split s_m in $S_m \cup \bar{S_m}$ that splits the node t into t'_L and t'_R, define $N_j(LL)$ as the number of objects in t that both s^{**} and s_m send to the left, that is that go to $t_L \cap t'_L$.
The probability that an object goes to $t_L \cap t'_L$ is estimated by

$$p(t_L \cap t'_L) = \sum_j \frac{N_j(LL)}{N}$$

Then the estimated probability $p_{LL}(s^{**}, s_m)$ that both s^{**} and s_m send an object in t to the left is

$$p_{LL}(s^{**}, s_m) = \frac{p(t_L \cap t'_L)}{p(t)}$$

where $p(t)$ is the resubstitution estimate for the probability that any object will fall into node t defined as $\sum_j \frac{N_j(t)}{N}$. $p_{RR}(s^{**}, s_m)$ is defined in the same way.
The probability that the split s_m predicts the split s^{**} correctly is then estimated by

$$p(s^{**}, s_m) = p_{LL}(s^{**}, s_m) + p_{RR}(s^{**}, s_m)$$

Now a formal definition of a surrogate split can be given:

Surrogate split (Definition 5.7 in [3]):
A split $\tilde{s_m} \in S_m \cup \bar{S_m}$ is called a surrogate split on x_m for s^{**} if

$$p(s^{**}, \tilde{s_m}) = \max_{s_m} p(s^{**}, s_m)$$

where the maximum is over $S_m \cup \bar{S_m}$.

The surrogate split $\tilde{s_m}$ can be interpreted as the split on x_m that most accuractely predicts the action of s^{**} [3].
If an object has missing data on both the primary split and on the best surrogate split, the second best surrogate split is used. If data on the second best surrogate split also is missing, the third best surrogate split is used and so on.

3 Method and Data

3.1 Learning set

Anonymized dummy data on answers on the HW11-questions were provided as a basis for the statistical analyses. The data were transformed into monthly

means; each observation in the learning set was the mean of answers from a calendar month. This was done because it is believed that a person's answers on the OLBI questionnare depend not only on how the person felt the time when he was answering (and at the same time answered the HW11-questions). It is believed that exhaustion and disengagement, which OLBI measures, come gradually. This would mean that if a person answers HW11 differently when he has high levels of exhaustion and disengagement, difference would also be detectable on earlier answers on HW11. It is belived that the answers on HW11 will be different in the nearest future as well.
Since the method treats the data cross-sectionally, the monthly means of HW11 were used as observations of the predictor variables. The corresponding observation of the dependent variable was constructed using the answers of OLBI and PBS from the same month. All observations in the learning set had complete measurements on all HW11-questions, the OLBI- and PBS questionnares.

Two separate models were made for men and women, so the learning set was divided by gender. The learning set of men consisted of 3043 observations, 88 of which belonged to class 1 (2.9%). The median score on PBSE was 3.25 for this dataset.
The learning set of women consisted of 2246 observations, 80 of which belonged to class 1 (3.6%). The median score on PBSE was 3.25 for this dataset as well.

3.2 Test set

For testing the model on a dataset other than the learning set that was used for creating the model, a separate test set was used. The test set consisted of data from a previously conducted scientific study [7], including the HW11-questions and the OLBI and PBS questionnares.
The test set of men consisted of 128 observations, 4 of which belonged to class 1 (3.1%). The test set of women consisted of 191 observations, 8 of which belonged to class 1 (4.2%). The median scores on PBSE was 2.75 for men and 3.25 for women.
All observations had complete answers on OLBI and PBS. Some observations had missing answers one some of the HW11-questions, but they were possible to classify using surrogate splits.

3.3 Creating the models

The models were built using the software environment R. 10-fold cross-validation (which is the default in R) was used for pruning. Before the results are presented, the method for choosing mislclassification cost is described.

Choosing misclassification cost

It was believed that misclassifying a class-1 object as a class-0 object is worse than the opposite misclassification. Therefore the misclassification cost $C(1|0)$ was set to 1 and the $C(0|1)$ was to be determined.

Possible values of $C(0|1)$ were

$$0 \leq C(0|1) \leq \frac{\text{total number of objects}}{\text{number of class-1 objects}}$$

In the learning set for men, the upper limit was equal to $\frac{3043}{88} = 34.6$. For women the upper limit was $\frac{2246}{80} = 28.1$.
To determine which misclassification cost to use in the model, a tree for each possible integer misclassification cost was produced. It seemed to be sufficient to try only with integer values of the misclassification cost because the outcome did not vary a lot for values close to each other, as can be seen in Figure 4, Figure 5 and in Tables 8-9 in Section 5.1 of the appendix.
First the classification tree was grown and pruned. Then its accuracy, precision, sensitivity and specificity were computed.
For women, 28 different trees were grown. Their values of accuracy, precision, sensitivity and specificity are shown in Figure 4, where accuracy and precision are plotted together and sensitivity and specificity are plotted together. For men, 34 trees were grown. Their accuracy, precision, sensitivity and specificity are in Figure 5.

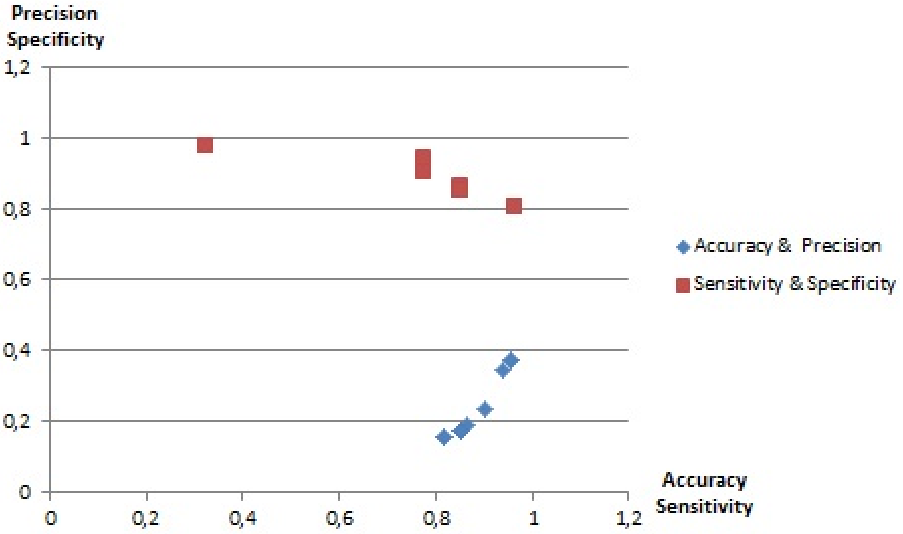

Figure 4: Accuracy, precision, sensitivity and specificity for different misclassification costs in the models for women.

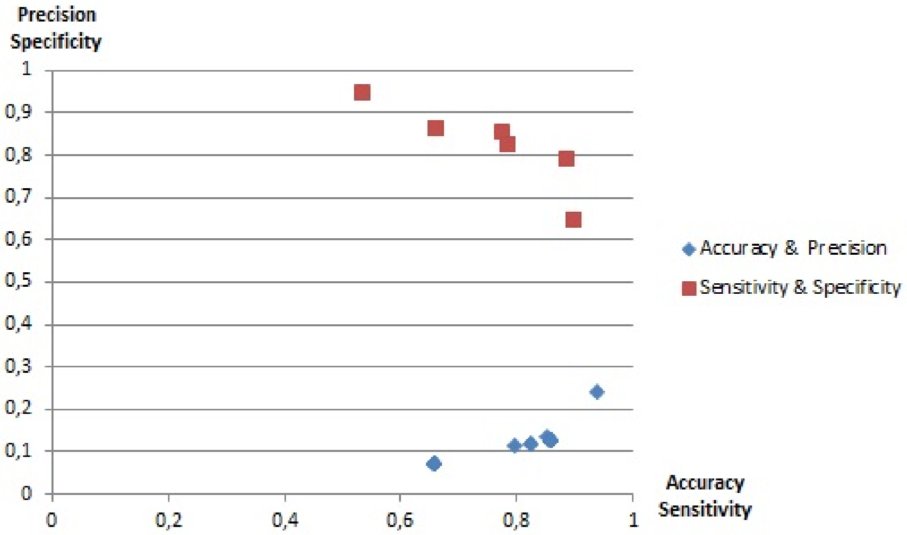

Figure 5: Accuracy, precision, sensitivity and specificity for different misclassification costs in the models for men.

Note that the parameters are plotted pairwise just for illustration. When they were compared for determining which misclassification cost gives the best model, they were compared as four-tuples (four parameters per tree). The actual values of the parameters for different misclassification costs are in Tables 8-9 in Section 5.1 of the appendix.

Sensitivity was believed to be the most important parameter because it was considered more important to be able to correctly classify class-1 objects than class-0 objects. So the tree with the higher sensitivity was consider the better and misclassification cost was chosen based on the criterion that the sensitivity is to be as high as possible.
Therefore misclassification cost 21 was chosen for women, it corresponded to the tree with the maximum sensitivity (0.9625).
Misclassification cost 33 was chosen for the model for men. It corresponded to the model with the second best sensitivity (0.8864). The highest sensitivity was 0.8977, but the values of the remaining parameters were improved a lot if the sensitivity decreased from 0.8977 to 0.8864, see Table 2. The improvement of the other parameters was considered to compensate for the small decrease in sensitivity.

Misclassification cost	Accuracy	Precision	Sensitivity	Specificity
24	0.6569	0.0709	0.8977	0.6497
33	0.7946	0.1126	0.8864	0.7919

Table 2: Comparison between the two best models for men

4 Results

In this official version of the thesis the numerical values of the splits in Figures 6, 8, 9 and 11 are not the real values. The real values are confidential and are substituted by randomly chosen numbers for this version.

4.1 Model for women

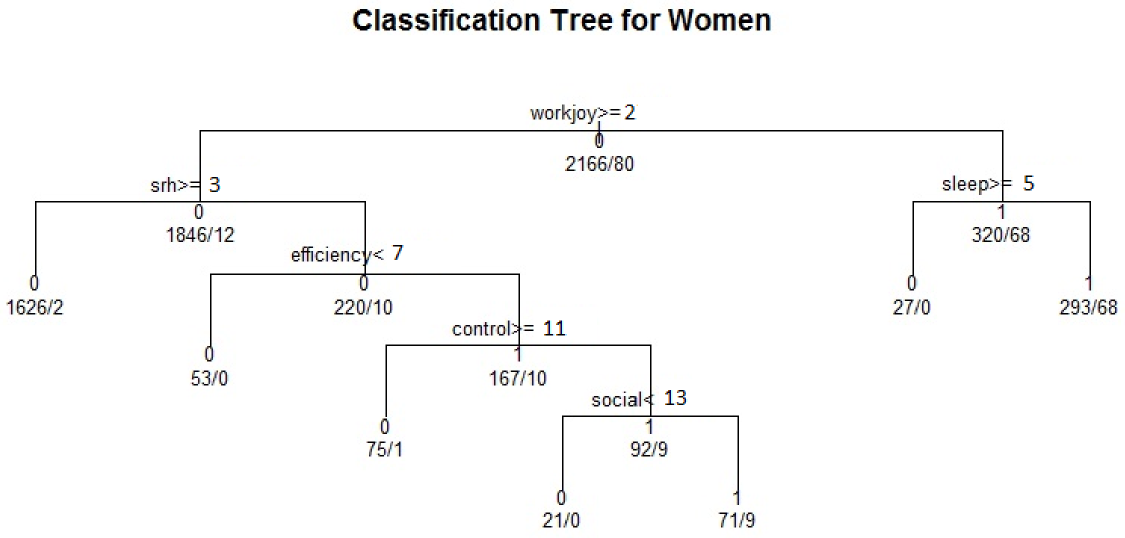

Figure 6: The full classification tree for women before pruning.

The full classification tree before pruning is in Figure 6 above. By inspection of the cp table, Table 3, and the cp plot, Figure 7, the optimally pruned subtree could easily be obtained as the second tree in the sequence of trees produced in the cross validation. This classification tree is in Figure 8 below.

cp	nsplit	rel error	xerror	xstd
0.659524	0	1.00000	1.00000	0.109794
0.019246	1	0.34048	0.34048	0.044009
0.016071	4	0.28274	0.39821	0.047432
0.012500	5	0.26667	0.39702	0.048972
0.010000	6	0.25417	0.42143	0.053389

Table 3: cp-table for women before pruning

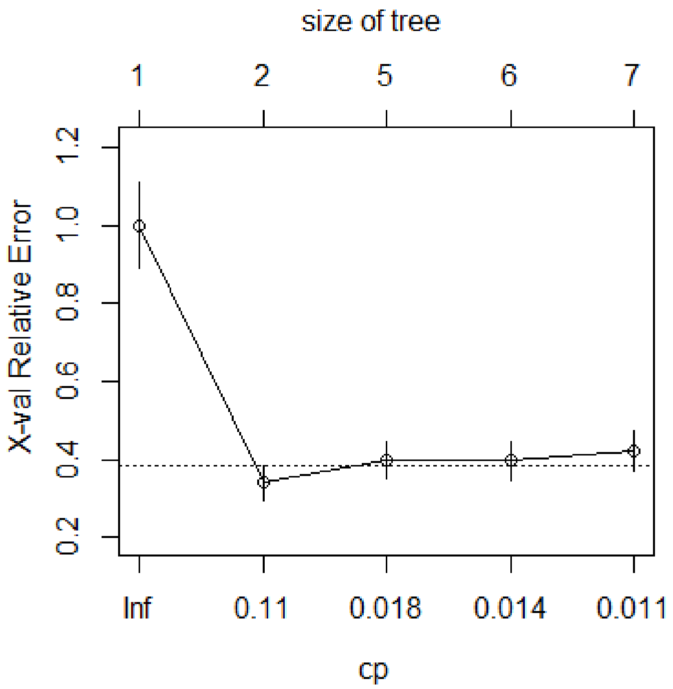

Figure 7: The cp-plot for women.

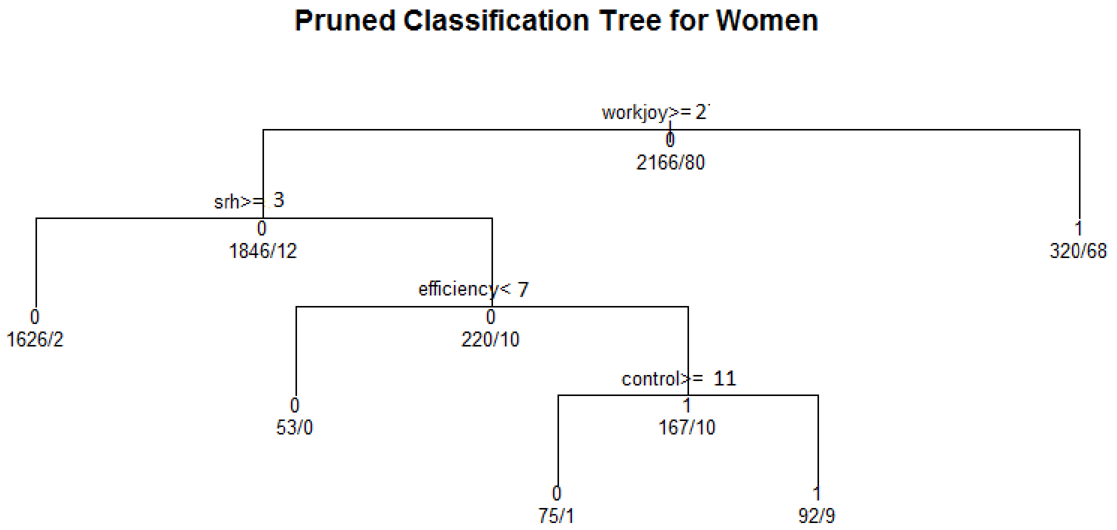

Figure 8: The optimally pruned subtree for women.

Since there are five terminal nodes in the tree, there are five different combinations of answers that lead to a classification. Two of the combinations will lead to class 1, i.e. over-risk for long-term sickness absence. Those combinations are:

- Workjoy less than or equal to 2

- Workjoy greater than 2, srh less than or equal to 3, efficiency greater than 7 and control less than 11.

This means that a woman with an answer on workjoy that is less than or equal to 2 will be classified as having an over-risk for long-term sickness absence no matter how she answered the other questions. If she has answered that her workjoy is greater than 2, her answers on srh, efficiency and control must also be taken into account in order to determine if she has an over-risk for long-term sickness absence or not.
There are three combinations of answers that lead to class 0, i.e. no over-risk for long-term sickness absence. Those are:

- Workjoy is greater than 2 and srh is greater than or equal to 3

- Workjoy is greater than 2, srh is less than or equal to 3 and efficiency is less than 7

- Workjoy is greater than 2, srh is less than or equal to 3, efficiency is greater than 7 and control is greater than or equal to 11

If there is no answer for one or more of the five variables in the tree, then surrogate splits are used for classification. *The lists of surrogate splits and primary splits are not included in this official version of the thesis.* The list of primary splits for a node is list of the five best splits of the node and it shows how much each split improves the purity of the tree. The first split in this list is the split that is used in the tree, i.e. the split that gives the largest improvement of purity. The second split in the list is the split with the second best improvement of purity, and so on. Remember that only the best split, which is shown in the tree as well, is used for classification. The lists of primary splits are used for furhter analysis of the variables. Together with the lists of surrogate splits, they can be used for analyzing importness of variables.

Test results

When the observations in the test set and learning set were classified by the model, the following results were obtained:

	Accuracy	Precision	Sensitivity	Specificity
Learning set:	0.8152	0.1575	0.9625	0.8098
Test set:	0.7435	0.1273	0.8750	0.7377

Table 4: Test results for women

The real values of the percentage numbers in the discussion below are confidential. In this version of the thesis they are randomly set to 19. The values of the parameters accuracy, precision, sensitivity and specificity are lower for the test set than for the learning set, but the test results are still good and are considered acceptable. The model seems to work well for data other than the learning set. Since the model is acceptable, it may be used for prediction and some conclusions can be drawn from analyzing the model and the tables of splits.

Examining the lists of surrogate and primary splits, one can see that workload never appears, neither in the tree nor in the lists of primary and surrogate splits. Therefore one might suspect that work load does not contribute to an icreased risk for long-term sickness absence.

Another variable that probably does not affect the risk of long-term sickness absence is work atmosphere. This variable is not in the tree, but it appears in the lists of surrogate and primary splits. Though it appears of low importance, it is the fourth best surrogate split to workjoy and it is the fifth best split in the node where control is the best split.

A variable that is not in the tree, but anyway seems to be a very important one for over-risk of long-term sickness absence is concentration. It is the first surrogate split both to efficiency and to workjoy and it also appears as the second best split in the list of primary splits to control, where control is 19% better.

Some conclusions can also be drawn regarding srh and workjoy. Both seem to be very important variables for an over-risk for long-term sickness absence. First of all, both appear as variables in the tree. They both also appear as surrogate splits more than once and in the lists of primary splits when they are not the best splits themselves. There also seems to be some kind of correlation between them. When workjoy is the best split, srh is the second best and agrees to 19% with workjoy as a surrogate split. When srh is the best split, workjoy is the fourth best split and agrees to 19% with srh as a surrogate split. Both workjoy and srh agree to 19% with efficiency and to 19% with control.

Workjoy seems to be one of the most important factors contributing to an over-risk for long-term sickness absence. This is motivated by the discussion above; it appears with high importance both in the model, as surrogate split and in the lists of primary splits. Another strong indicator of that workjoy is very important is the first split of the tree, which says that whenever workjoy is less than 2 there is an over-risk for long-term sickness absence.

4.2 Model for men

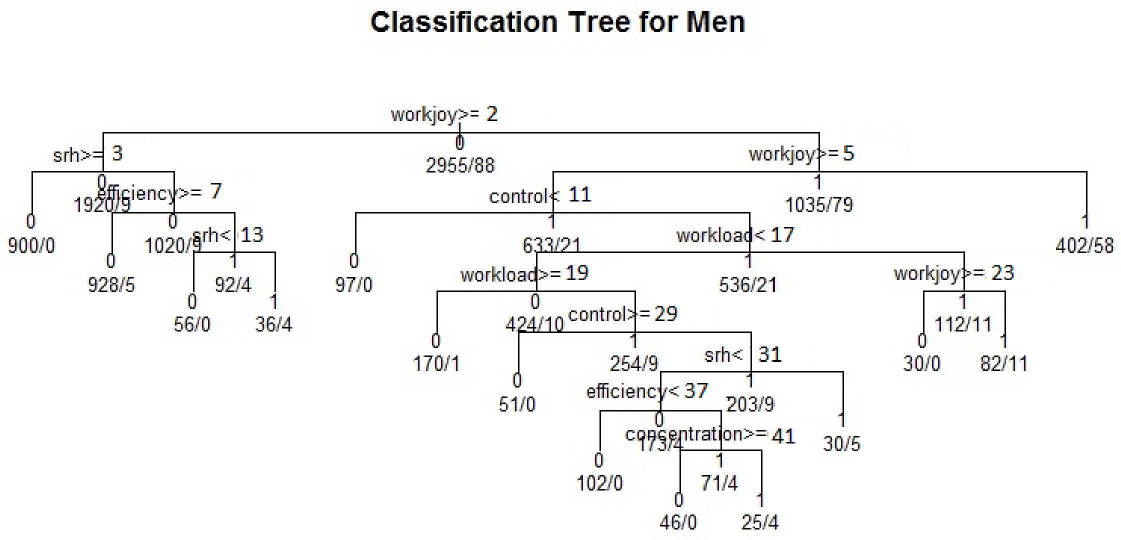

Figure 9: The full classification tree for men before pruning.

The full classification tree before pruning is in Figure 9 above. By inspection of the cp table, Table 5, and the cp plot, Figure 10, the optimally pruned subtree could easily be obtained as the second tree in the sequence of trees produced in the cross validation. This classification tree is in Figure 11 below.

cp	nsplit	rel error	xerror	xstd
0.541322	0	1.00000	1.07989	0.092215
0.016701	1	0.45868	0.46866	0.036625
0.015840	8	0.32541	0.51618	0.058916
0.011019	9	0.30957	0.52066	0.058916
0.010331	12	0.27652	0.53822	0.061004
0.010000	13	0.26618	0.53857	0.061004

Table 5: cp-table for men before pruning

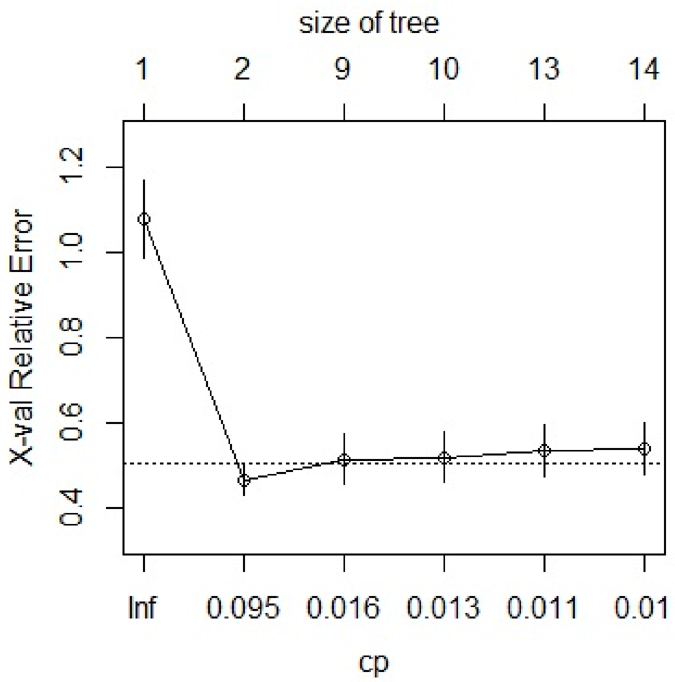

Figure 10: The cp-plot for men.

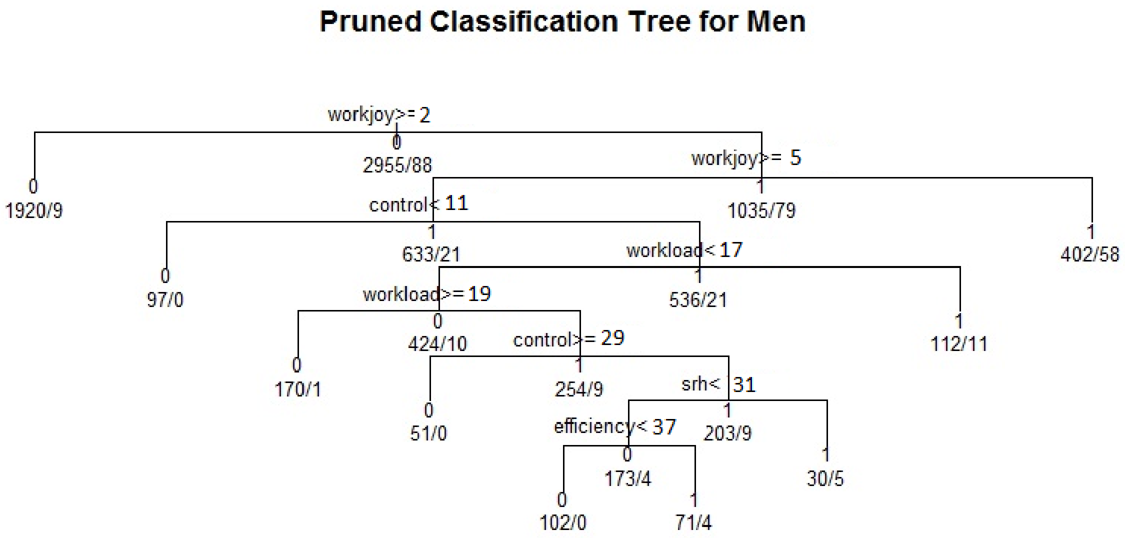

Figure 11: The optimally pruned subtree for men.

The tree has nine terminal nodes. Four of them correspond to class 1. The combinations that lead to class 1 are:

- workjoy less than 2
- workjoy between 2 and 5, control greater than 11 and workload greater than 17.

- workjoy between 2 and 5, control between 11 and 29, workload less than 17 and srh greater than 31
- workjoy between 2 and 5, control between 11 and 29, workload less than 17, srh less than 31 and efficiency greater than 37

If there is no answer for one or more of the five variables in the tree, then surrogate splits are used for classification. *The lists of surrogate splits contain confidential data and are not included in this official version of the report.*

Test results

When the observations in the test set and learning set were classified by the model, the following results were obtained:

	Accuracy	Precision	Sensitivity	Specificity
Learning set:	0.7946	0.1125	0.8864	0.7919
Test set:	0.8516	0.0588	0.2500	0.8710

Table 6: Test results for men

The sensitivity for the model was very low when the model was tested on the test set. A sensitivity of 25% can not be considered accepable, especially not when sensitivity was considered the most important one of these parameters.
One idea for why the sensitivity became so low in the model for men but not in the model for women is that the median score on PBSE coincided for the learning and test set for women, but they were different in the datasets of men. For women, both datasets had median score 3.25, but for men the median score was 3.25 in the learning set and 2.75 in the test set.
Accuracy, precision, sensitivity and specificity were computed with the limit of high PBSE set to 3.25 instead of 2.75 in the test set. The results are in Table 7 below. There is a small improvement of sensitivity, but not enough for considering the model acceptable as a general model.

	Accuracy	Precision	Sensitivity	Specificity
Learning set:	0.7946	0.1125	0.8864	0.7919
Test set, PBSE 2.75:	0.8516	0.0588	0.2500	0.8710
Test set, PBSE 3.25	0.8594	0.0588	0.3333	0.8720

Table 7: Test results for men

Since the model for men does not work well as a general model, no conclusions can be drawn from it and it not recommended to use it for prediction.

5 Appendix

5.1 Tables of accuracy, precision, sensitivity and specificity

Misclassification cost	Accuracy	Precision	Sensitivity	Specificity
2 (only root node)	-	-	-	-
3	0.9564	0.3714	0.3250	0.9797
4	0.9399	0.3464	0.7750	0.9460
5	0.9564	0.3714	0.3250	0.9797
6	0.8522	0.1753	0.8500	0.8523
7	0.8522	0.1753	0.8500	0.8523
8	0.9016	0.2340	0.7750	0.9063
9	0.8522	0.1753	0.8500	0.8523
10	0.8522	0.1753	0.8500	0.8523
11	0.8522	0.1753	0.8500	0.8523
12	0.8522	0.1753	0.8500	0.8523
13	0.8522	0.1753	0.8500	0.8523
14	0.8522	0.1753	0.8500	0.8523
15	0.8642	0.1884	0.8500	0.8647
16	0.8522	0.1753	0.8500	0.8523
17	0.8522	0.1753	0.8500	0.8523
18	0.8522	0.1753	0.8500	0.8523
19	0.8522	0.1753	0.8500	0.8523
20	0.8522	0.1753	0.8500	0.8523
21	**0.8152**	**0.1575**	**0.9625**	**0.8098**
22	0.8522	0.1753	0.8500	0.8523
23	0.8152	0.1575	0.9625	0.8098
24	0.8522	0.1753	0.8500	0.8523
25	0.8522	0.1753	0.8500	0.8523
26	0.8522	0.1753	0.8500	0.8523
27	0.8522	0.1753	0.8500	0.8523
28	0.8522	0.1753	0.8500	0.8523

Table 8: Table of accuracy, precision, sensitivity and specificity for different misclassification costs in the model for women.

Misclassification cost	Accuracy	Precision	Sensitivity	Specificity
2 (only root node)	-	-	-	-
3 (only root node)	-	-	-	-
4 (only root node)	-	-	-	-
5 (only root node)	-	-	-	-
6 (only root node)	-	-	-	-
7	0.9376	0.2398	0.5341	0.9496
8 (only root node)	-	-	-	-
9	0.8580	0.1261	0.6591	0.8640
10	0.8580	0.1261	0.6591	0.8640
11	0.8580	0.1261	0.6591	0.8640
12	0.8580	0.1261	0.6591	0.8640
13	0.8580	0.1261	0.6591	0.8640
14	0.8521	0.1365	0.7727	0.8545
15	0.8580	0.1261	0.6591	0.8640
16	0.8580	0.1261	0.6591	0.8640
17	0.8580	0.1261	0.6591	0.8640
18	0.8580	0.1261	0.6591	0.8640
19	0.8580	0.1261	0.6591	0.8640
20	0.8580	0.1261	0.6591	0.8640
21	0.8580	0.1261	0.6591	0.8640
22	0.8580	0.1261	0.6591	0.8640
23	0.8248	0.1184	0.7841	0.8261
24	0.6569	0.0709	0.8977	0.6497
25	0.6569	0.0709	0.8977	0.6497
26	0.6569	0.0709	0.8977	0.6497
27	0.6569	0.0709	0.8977	0.6497
28	0.8248	0.1184	0.7841	0.8261
29	0.6569	0.0709	0.8977	0.6497
30	0.6569	0.0709	0.8977	0.6497
31	0.6569	0.0709	0.8977	0.6497
32	0.6569	0.0709	0.8977	0.6497
33	**0.7946**	**0.1125**	**0.8864**	**0.7919**
34	0.6569	0.0709	0.8977	0.6497

Table 9: Table of accuracy, precision, sensitivity and specificity for different misclassification costs in the model for men.

6 References

[1] Maslasch et al. *Job Burnout*
[2] Hallsten L. et al. (2009). *Job burnout and job wornout as risk factors for longterm sickness absence*
[3] Breiman et al. (1984). *Classification and Regression Trees.* Boca Raton, Florida: Chapman & Hall CRC
[4] Hallsten L. et al. (2005). *Performance-based self-esteem, A driving force in burnout processes and its assessment*
[5] Bakker and Demerouti (2007). *Measurement of Burnout and Engagement*
[6] Bauer G. F. and G. J. Jenny (eds.)(2013). *Salutogenic organizations and change the concepts behind organizational health intervention research.* Dordrecht, New York: Springer
[7] Hasson D. et al. (2013). *Acute Stress Induces Hyperacusis in Women with High Levels of Emotional Exhaustion.* PloS one 8(1): e52945.

TRITA-MAT-E 2013:51
ISRN-KTH/MAT/E—13/51-SE

www.kth.se