UNIVERSITA' DEGLI STUDI DI MILANO
Facoltà di Medicina e Chirurgia
Dipartimento di Scienze Cliniche e di Comunità
Sezione di Statistica Medica e Biometria "Giulio A. Maccacaro"

Dottorato di Ricerca in Statistica Biomedica

XXVIII Ciclo – Settore scientifico disciplinare MED/01

A.A. 2014/2015

ORDINAL CLASSIFICATION TREES: METHODS AND APPLICATION

Dott. **Alessandra Lugo**
Matricola R10100

Tutor
Prof. **Adriano Decarli**

Coordinatore del Dottorato
Prof. **Adriano Decarli**

INDEX

ABSTRACT

Introduction to classification trees

Tree-based methods are non parametric regression methods belonging to a group of techniques called “recursive partitioning”. They recursively partition the feature space, which includes all the predictors, into a set of nested rectangular areas. The main objective of classification trees is to obtain subgroups of observations (nodes) which should be more homogeneous as possible in terms of the values of the response variable. A quantitative measure of the extent of a node homogeneity is the notion of node purity (or impurity). A node is completely pure if all the observations in it belong to the same category of the outcome, and its impurity is zero. On the other hand, a node is completely impure if the probability of selecting a subject belonging to a category is the same for all the categories of the outcome. In the nominal classification setting, there are mainly three impurity functions (∅): i) misclassification error rate, ii) Gini index, and iii) cross-entropy, or deviance. However, when the response to be predicted is ordinal, an ordinal splitting method is usually preferred. The most frequently used ordinal impurity function is the generalized Gini function:

$$i_{GG}(t) = \sum_{k} \sum_{k \neq l} C(w_k|w_l)p(w_k\,|t)p(w_l|t)$$

And the decrease in node impurity with the split s in node t is given by:

$$\Delta Imp_{GG}(t,s) = p(t)i_{GG}(t) - p(t_R)i_{GG}(t_R) - p(t_L)i_{GG}(t_L)$$

The split, among all the possible binary splits for a given node, resulting in the largest value of $\Delta Imp_{GG}(t,s)$ is selected.

The process of splitting continues in each node until some stop condition is reached, and a large tree T0 is built. However, a very large tree may overfit data. Overfitting refers to the fact that a classifier adapts too closely to the training dataset and it may fit random variation beside discovering the systematic components of the structure present in the population. Overfitting leads to poor test performance when applied to the validation set. Thus, a common strategy is

to eliminate those parts of a classifier that are likely to overfit the training data. This process is called pruning, and consists of eliminating branches that do not add information to prediction accuracy. Classification tree analysis uses a method of cost-complexity pruning. This approach balances the complexity (i.e. the number of predictors and terminal nodes) of the sub-tree and the overall misclassification rate:

$$R_\alpha(T) = R(T) + \alpha \cdot card(T)$$

Two predictive performance measures to be used when Y is an ordinal outcome are the total number of misclassified observations $\mathrm{R_{mr}(T)}$ and the total misclassification cost $\mathrm{R_{mc}(T)}$. The first one assigns to the observations within a node the modal class of the outcome, while the second one assigns the median class.

In ordinal classification settings, the performance of the tree is evaluated through measures of association between the observed and the predicted response, taking into account the ordinal classification of the outcome. Thus, Gamma statistics and Somers' d measure can be used.

With classification trees we are able to examine complex interactions among risk factors that do not need to be pre-specified a priori. Moreover, we will likely be able to identify the most important risk (or protective) factors among various predictors, and we have the possibility to identify ideal cut-offs of continuous variables, according to some pre-specified criteria. On the other hand, since this is a data-driven method, a drawback is that small changes in the data can result in a very different series of splits, making interpretation instable.

Methods

To conduct the analyses on ordinal classification tree method, I used the set of controls of various case-control studies conducted in six Italian provinces between 1991 and 2008. Overall, 7750 subjects were considered in the present analyses. Controls were individuals with no history of cancer admitted to the same hospitals of cases for acute, non-neoplastic, conditions, unrelated to diseases or to conditions linked to the cancer in study. Predictors were food groups, related to the subjects' dietary habits during the 2 years before hospitalization, assessed through a validated and reproducible food frequency questionnaire, which included information on weekly consumption of 78 foods and beverages.

Two different types of analyses were performed to evaluate the performance of classification trees methodology in predicting the category of total energy intake (kcal/day): single tree analysis and resampling analysis (100 different pairs of training and validation sets have been considered to overcome sampling variability). I compared five different scenarios, four in the context of ordinal classification trees (generalized Gini impurity function) and one in the context of nominal classification trees (Gini impurity measure). In the ordinal context, each scenario was a combination of the splitting function (absolute or quadratic misclassification cost) and the predictive performance measure (misclassification error rate/mode or misclassification cost/median). Also classification trees to predict the daily consumption (grams) of red meat and processed meat was performed.

Results

The most important predictor for energy intake was bread consumption. Indeed, this predictor resulted as the first split in each of the five scenarios, with a threshold of 16.4 portions/week. Other predictors common to all the five scenarios were desserts and red meat intake.

The best agreement for the prediction of energy intake was observed using an ordinal classification tree with a quadratic misclassification cost and misclassification cost (median value) as the predictive performance measure. Performance of tree built using the modal value was only slightly lower. This findings were consistent both in the single-tree and in the resampling analysis. In the single-tree analysis, the values of Somers' d measure ranged between 0.489 and 0.534, and those of Gamma measure ranged between 0.717 and 0.651. In general, the ranking of the scenarios according to their predictive performance was the same using the two ordinal measures of association. In the resampling analysis, Friedman's test rejected the global equality hypothesis across the five models (p<0.001), both considering Somers'd and Gamma values.

In the application on red meat and processed meat intake, it emerged that important predictors for red meat consumption were total intake of sweets (first split) and bread consumption, while important predictors for processed meat intake were the consumption of eggs, bread and sweets. In particular, subjects eating less than 1 egg per week and less than 2 portions of bread per day were classified as having small consumption (<25 g/day) of processed meat. On the other hand, individuals eating more than 1 egg per week and more than 30 portions of sweets per week were predicted to have a great (≥50 g/day) consumption of processed meat. Another analysis put in evidence that the intake of red meat and processed meat were related. In fact, processed meat intake was the first split when considering red meat as the response variable, and, accordingly, processed meat intake is the first predictor that categorized red meat intake.

Discussion

The comparison between five different methods put in evidence that, in case of ordinal outcome, adequate ordinal methods should be preferred. According to the prediction accuracy between various ordinal models, it emerged that models with quadratic misclassification cost had better predictive power, in particular when median was used to assign outcome classes. A good predictive performance was also observed with quadratic misclassification cost and modal values.

Classification trees allowed to identify profiles of consumers of red meat and processed meat. Individuals eating high quantities of processed meat (≥50 g/day) also ate eggs and large portions of sweets. Moreover, through classification trees analysis, we confirmed that red meat and processed meat consumption were strongly related.

LIST OF TABLES AND FIGURES

Tables

Figures

PREFACE

My PhD programme focused on the study of classification tree methodology, and the statistical and computational methods related to them. In particular, during the second year of my PhD in Biomedical Statistics, I introduced recursive partitioning techniques, including classification and regression trees. During the third year, I focused my research on ordinal classification trees, that is classification trees in case of ordinal response variables.

Recently (January 2014), a paper by Navarro Silvera and colleagues was published in the peer-reviewed journal Annals of Epidemiology (Navarro Silvera et al 2014). In this manuscript, authors investigated the role of various dietary and lifestyle factors, as predictors of oesophageal, gastric cardia and other (noncardia) gastric cancers, and the putative interactions between available risk factors in determining the onset of these tumours. Navarro Silvera and colleagues used the classification tree approach in order to better understand which of the correlated explanatory variables appeared to be the most important for risk stratification, and examined multilevel interactions involving these same variables.

Following the example of Navarro Silvera and colleagues, we decided to apply classification trees to Italian case-control studies on cancer risk. With this method we are able to examine complex interactions among risk factors that do not need to be pre-specified a priori (like in regression models). In particular, we investigated the association between dietary factors and lifestyle habits and OCP, breast and prostatic cancers risk. Moreover, with this method we are likely able to identify the most important risk (or protective) factors among various predictors, and we have the possibility to identify ideal cut-offs of continuous variables, according to some pre-specified criteria. Indeed, a usual choice for threshold is that of the percentile distributions of continuous variables in the population of controls, or is a cut-off arbitrarily chosen by researchers. We want to verify whether usual thresholds are confirmed with the present analyses.

Usually, in epidemiology, the outcome is a dichotomous variable (i.e., in case-control studies). However, the interest of researchers may also be the identification of profiles of some specific populations (e.g., current smokers, alcohol drinkers, fruit and vegetable consumers) according to specific characteristics including demographic and socio-economic variables, lifestyle habits, and dietary behaviours. In these cases, response variable may be an ordinal categorical variable (e.g., smoking of <15, 15-24, ≥25 cigs/day; light, intermediate, heavy drinking; low, medium, high weekly consumption of vegetables), and a dichotomisation of the outcome may led to a loss of information.

In 2008, Piccarreta pointed out that when the response to be predicted is ordinal, an ordinal splitting method is usually preferred (Piccarreta 2008). She also proposed an alternative ordinal impurity function, other than that proposed by Breiman (Breiman et al 1984), that does not require the assignment of misclassification costs.

Programmers have created various R packages in order to implement ordinal classification trees. In particular, rpartOrdinal (Archer 2010) and rpartScore (Galimberti et al 2012) can manage different impurity functions, misclassification cost functions and pruning methodologies.

I was inspired by a recent study conducted by a group of American researchers using classification tree methodology to evaluate occupational exposure, identifying a rule to be independent by expert-assessed exposure estimates, which are not transparent and systematic (Wheeler et al 2015). Wheeler and colleagues compared the performance of nominal and ordinal classification trees in predicting ordinal occupational diesel exhaust exposure estimates in a case-control study on bladder cancer. I was also inspired by a paper by Biesbroek and colleagues comparing the results of 4 different methodologies, including classification trees, in identifying dietary patterns and comparing them in the association with coronary artery disease and stroke risk in the EPIC-NL cohort (Biesbroek et al 2015).

The objectives of my second year of PhD were: i) to study decision trees, and in particular classification trees to be applied in case of binary outcome; ii) to apply classification trees methodology to real data using case-control study on cancer in order to investigate the association between dietary factors and lifestyle habits and this cancer; iii) to compare the performance in terms of reliability of predictions of classification trees and logistic regression models. Moreover, an introduction to methods proposed in the literature to overcome limitations of classification tree methodology were investigated.

The objectives of my third year of PhD were: i) to study ordinal classification trees, and corresponding various impurity functions, misclassification cost functions, and pruning methodologies; ii) to apply ordinal classification trees methodology to real data using the set of controls of various Italian case-control studies to analyse the profile of Italians according to total energy intake; iii) to compare the performance of nominal and ordinal classification trees in case of an ordinal response variable. Moreover, a comparison of various impurity measures and splitting functions in ordinal classification trees were also performed.

1. CLASSIFICATION TREES

1.1. Introduction to classification trees

Tree-based methods are non parametric regression methods belonging to a group of techniques called "recursive partitioning". They recursively partition the feature space, which includes all the predictors, into a set of nested rectangular areas, as shown in **Figure 1.1** (according to the decision rules built in Figure 1.2).

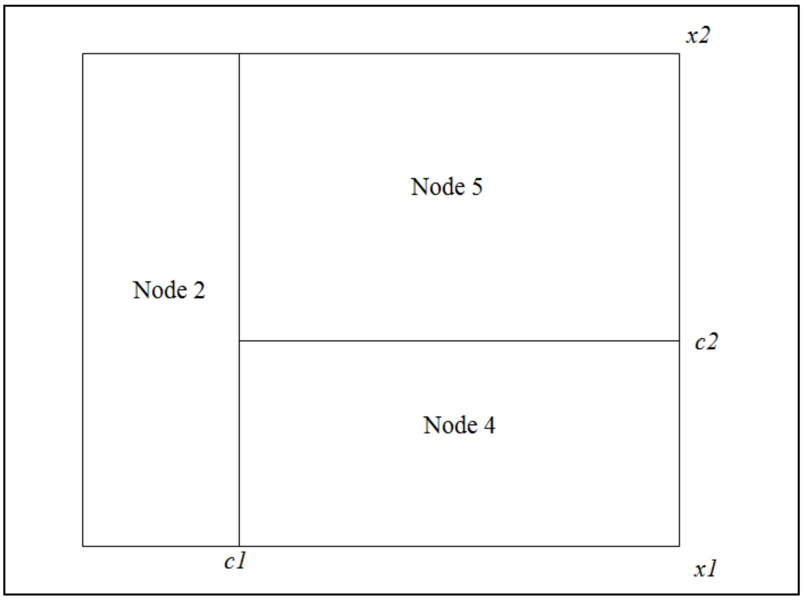

Figure 1.1. *Rectangular recursive partition of the feature space.*

The partition is developed in order to create groups of similar observations in terms of the values of the response variable.

Tree-based methods can be applied both to regression and classification problems. Regression trees are built in presence of a continuous quantitative outcome. On the other hand, classification trees are built when the outcome is a qualitative variable. It can be either a categorical nominal, categorical ordinal or a dichotomous variable.

According to the type of the outcome variable, the partition of the units is developed through a series of splitting rules. Splitting rules for nominal or dichotomous variables will be explained in S*ection 1.2*, while splitting rules for ordinal outcomes will be presented in *Section 2.1*. The aim is to build appropriate decision rules able to assign each observation to a specific group, clustering similar individuals. It is a common choice to realize binary splits, that is, to partition a group of observations in only two subgroups.

Classification and regression trees have been widely applied in different fields, including finance, engineering, astronomy, environmental research, psychology, marketing, etc. One of the field in which tree-based methods are used is biomedical research, for which classification is a central issue. Indeed, through tree-based methods, it is possible to construct diagnostic tests. In epidemiology, tree-based methods are applied to fulfill two specific aims:

i) Classification: detect high risk groups for a disease;

ii) Prediction: apply the model to independent new data and predict an outcome following the rules built in the tree.

From the prediction point of view, a good splitting rule is able to exactly predict the outcome value for each observation. Indeed, after the partition is completed, a constant value of the response variable is predicted within each area, thus, the observations within each group will be assigned a value of the response variable. This value changes according to the type of the outcome: in regression problems, the mean value of y is assigned within each group, in nominal classification problems the predicted value for a group of observations is the most frequent value of y in that area, while in ordinal classification problems the predicted value may be either the modal or the median value in each group of observations.

These methods are also known as decision trees methods because the set of splitting (or decision) rules used to segment the predictor space can be summarized in a tree. **Figure 1.2** represents an example of a classification tree and describes its components. It is based on the recursive partitioning of the feature space shown in Figure 1.1.

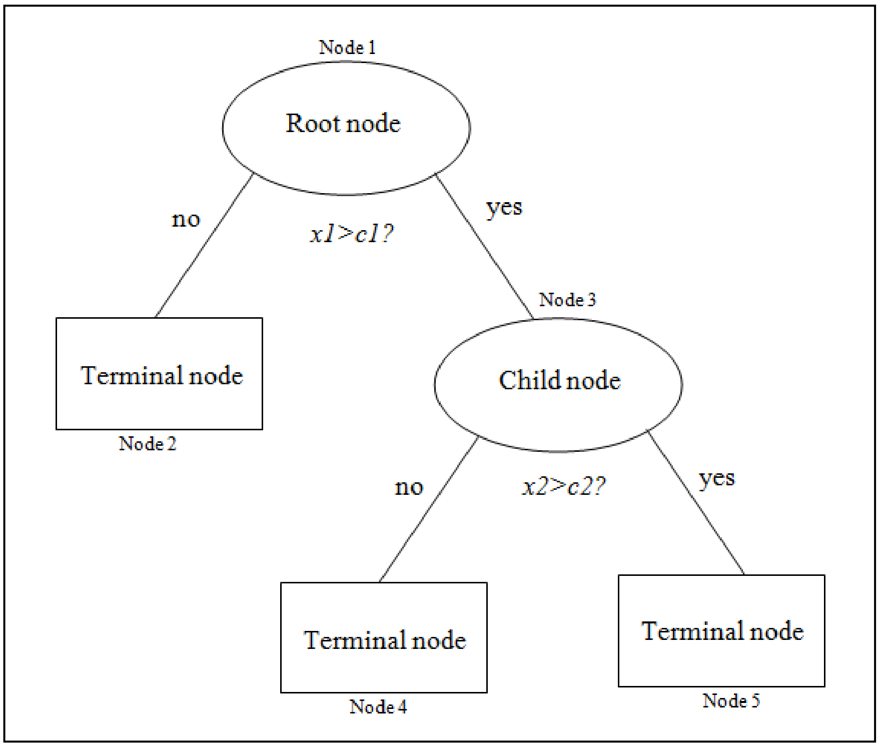

Figure 1.2. *Classification tree structure.*

In Figure 1.1 and Figure 1.2, the first split in the variable *x1* partition the entire sample, while the second split in the variable *x2* partition only those observations with a value of the variable *x1* higher than a certain cut-off *c1* (*x1>c1*).

A classification tree has the following elements:

- Nodes: every group of observations. They can be classified as a root node, parent nodes, child nodes or terminal nodes. In each node there is a specific proportion of subjects belonging to a category of the outcome.
- Root node: it is the starting point and it includes all the observations of the whole dataset. From the root the first split of the tree is realized. It is unique, so in each tree only one root node exists.

- Parent nodes: nodes where a split is generated. Since we are considering only binary partitioning, they are partitioned only into two child nodes.
- Child nodes: nodes generated by a split of a parent node.
- Terminal nodes: nodes that do not generate child nodes. They are at the end of a branch.

Usually, in figures, the root and the parent nodes are indicated by circles, while terminal nodes are represented by squares. Every node in a tree is a subset of the root node.

In the construction of classification trees, it is fundamental to know the outcome category for each observation, and this represents the main difference with other classification techniques, such as cluster analysis. Then, the model built will be fitted to a new set observations, since we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen data, for which we do not know the value of the outcome. The set of observations used to build the classification tree is called training (or learning) set, because it is used to “train” our model. The set of new unseen observations is called the validation (or test) set. Not always a new set of observations is available to test the performance of our model, thus resampling methods, such as cross-validation techniques, can be used to obtain additional information about the fitted model and to estimate the test error associated to it using a subset of the training observations.

Regression and classification trees methodologies have several advantages:

- Recursive partitioning techniques are non-parametric methods, and do not require a specified model. They are designed to model and to quantify the relationship between two sets of variables. Different parametric (or semi-parametric) methods (including linear regression for continuous data, logistic regression for binary data, mixed-effect regression for longitudinal data…) may not lead to an effective description of data when the underlying assumptions are not satisfied. In many cases, recursive partitioning provides a useful alternative to the parametric regression methods;

- Not only linear associations, but also nonlinear and even non monotone association rules, can be examined. They do not need to be specified in advance, but are determined in a data-driven way;

- Regression and classification trees can handle both qualitative and quantitative predictors. Qualitative predictors can be included without the creation of dummy variables, and thresholds do not need to be decided a-priori;

- Interactions are intrinsic in the model and do not need to be specified a-priori;

- They can handle missing values without problems. They consider missing values as another category of data, and we can also discover that observations with missing values for some measurements behave differently than those with non missing values;

- As that they can be graphically represented, they can also be easily interpreted even by non-experts;

- Decision trees more closely mirror human decision-making than other regression approaches do. They can be helpful in illness diagnosis and classification of high-risk subgroups for a disease.

On the other hand, disadvantages of these methods include:

- They do not have the same level of predictive accuracy as some of the other regression and classification approaches. But in some cases prediction using a tree may be preferred when the interpretability and graphic visualization are key issues;

- Their high variance. Often a small change in the data can result in a very different series of splits, making interpretation precarious. This high variability is due to the hierarchical nature of the methods. Indeed, errors that occur at the top of the tree, are propagated down to all splits below it. A solution can be to use a more stable split criterion, but instability is not completely solved in any way (bagging is an alternative to reduce variance).

1.2. Splitting rules in classification trees

The main objective of classification trees is to obtain nodes which should be more homogeneous as possible in terms of the values of the response variable. A quantitative measure of the extent of a node homogeneity is the notion of node purity (or impurity). A node is completely pure if all the observations in it belongs to the same category of the outcome, and its impurity is zero. On the other hand, a node is completely impure if the probability of selecting a case is the same of that of selecting a control. Thus, following the approach of impurity reduction, the aim of decision trees is the maximization of homogeneity within a node (minimization diversity or heterogeneity within a node) and maximization of heterogeneity between nodes.

When a classification trees is implemented, we consider to have a binary/categorical outcome (for simplicity we consider a dichotomous outcome) and a vector of predictors $x_1....x_p$. These predictors can be either continuous and categorical variables. In the phase of splitting, all allowable splits for all the predictors included are considered. Thus, in the phase of splitting, two important choices are taken: both the predictor and the threshold to minimize the heterogeneity within a node are selected. This is a joint choice.

The splitting procedure is an iterative procedure to obtain child nodes that are more "pure" than the parents nodes, and follows the scheme in **Figure 1.3**.

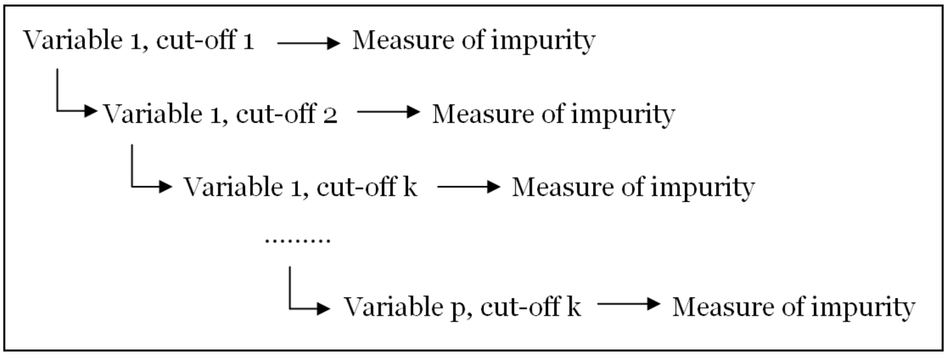
Variable 1, cut-off 1 → Measure of impurity
Variable 1, cut-off 2 → Measure of impurity
Variable 1, cut-off k → Measure of impurity
....
Variable p, cut-off k → Measure of impurity

***Figure 1.3**. Iterative procedure of splitting a tree.*

This process applies to the partition of any node, and the optimal choice is the one than that minimizes a certain measure of the impurity (∅). In each node the variable that is most strongly associated with the response variable (that produces the highest impurity reduction) is selected for the next split.

If the variable considered is a binary outcome, only two possible splits are taken into account, using the natural cut-off of the variable; if the predictor is a continuous variable or a ordinal categorical variable with k values, all the k-1 splits are taken into account; finally if the predictor is a nominal categorical variable with k levels, all the 2 $^{k-1}$ -1 allowable splits are considered.

In regression trees, where the outcome is a continuous variable, the residual sum of squares (RSS) computed as (1.1) is used as the impurity measure:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2 \tag{1.1}$$

In classification trees, the RSS cannot be used as a criterion for making binary splits. In the nominal classification setting, mainly three alternatives are present in order to estimate the impurity (or purity) of a node.

First of all let’s define $\hat{p}_{mk}$ as the proportion of observations in a node *m*, (the region R_m with N_m observations) belonging to the class k of the outcome:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_j} I(\, y_i = k) \tag{1.2}$$

We classify the observations in node *m* to class *k(m)* which represents the majority class in the node *m*:

$$k(m) = \arg\max_{\mathrm{k}} \hat{p}_{mk} \tag{1.3}$$

This means that we assign an observation in a given region to the most commonly occurring class in that region.
The three measures of impurity are:

i) Misclassification error rate (also called the Bayes error, or the minimum error)

This method is very simple and it is the proportion of observations in a region R_m that do not belong to the most frequent class.

$$E = \frac{1}{N_m} \sum_{i \in R_j} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)} \qquad (1.4)$$

However, this method is not sensitive enough for tree growing, and two other indexes are preferable.

ii) Gini index

$$G = \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{i=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \qquad (1.5)$$

It is a measure of the total variance across the k classes.

The Gini index ranges between 0 and 1. It reaches small values when all the $\hat{p}_{mk}$ are close to 0 or 1. A small value of this index means that all, or almost all, the observations in a node m belong to the same category k. Thus, the Gini index gives a measure of a “purity” of the node, and what we are looking for using this classification technique is pure regions. When Gini index is 0.5 then the maximum level of heterogeneity in a node is reached and the proportion of cases is the same of the proportion of controls.

iii) Cross-entropy, or deviance

$$D = - \sum_{i=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \qquad (1.6)$$

This measure ranges between 0 and 1, and, like the Gini index, reaches small values when all the $\hat{p}_{mk}$are close to 0 or 1. Thus also the cross-entropy gives a measure of the purity of the node. In case of maximum heterogeneity, the deviance value is 0.7.

The formulas presented above are in the general case of a categorical outcome variable with K categories, but, in the case of a binary outcome, the three impurity functions are simplified as follows:

i) $E = 1 - \max(p, 1 - p)$

ii) $G = 2p(1 - p)$

iii) $D = -p\log(p) - (1 - p)\log(1 - p)$

and they are represented in **Figure 1.4**.

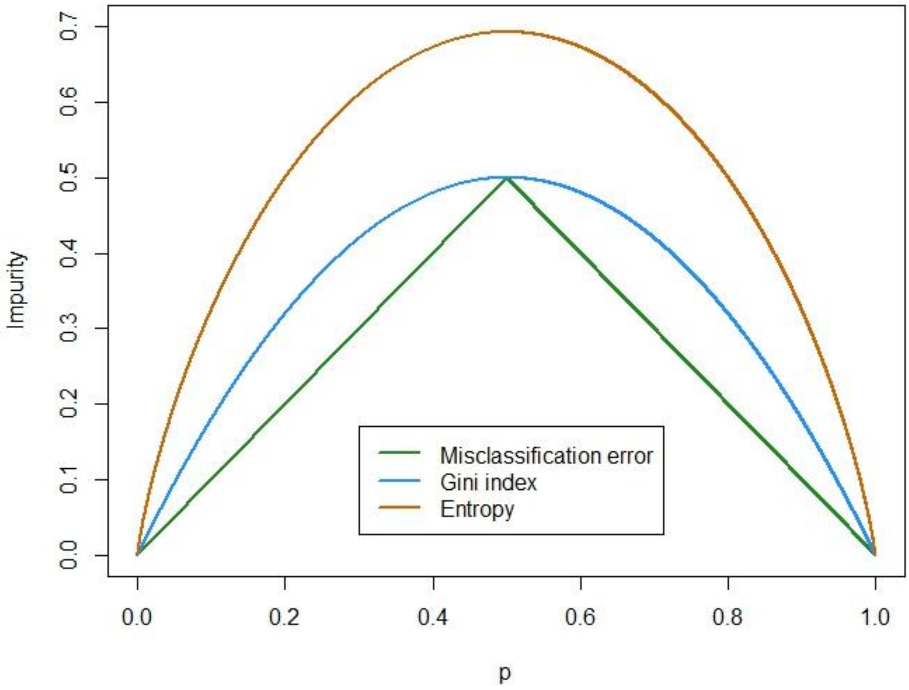

***Figure 1.4**. Node impurity measures for binary outcome classifications, as a function of p.*

From this figure, some properties of impurity functions ($\emptyset$) emerge:

- Concave shape;

- Positive functions: $\emptyset \geq 0$;

- Symmetric functions: $\emptyset(p) = \emptyset(1 - p)$;

- Reach the minimum when p is 0 or 1: $\emptyset(0) = \emptyset(1) < \emptyset(p)$.

Gini index and cross-entropy are more sensitive to changes in the node probability than the misclassification error, thus, when we grow the tree, it is a preferable choice to select these two measures to compute node purity. However, all these three methods can be used to guide the cost-complexity pruning.

Obviously, in steps after the first split of the root node, we need to weight the node impurity measures by the number N_{mL} and N_{mR} which indicates the number of observations in the left child node and in right child node, respectively, created by splitting the node m.

The construction of a tree may also be seen as a sort of variable selection procedure. Indeed, the procedure selects only the most important variables for the segmentation of the statistical units. It can happen that not all the initial predictors actively contribute to the definition of the final decision rule. It can also happen that a predictor is included more than one with different cut-offs.

1.3. Stopping and Pruning

The process shown in Figure 1.3 iteratively continues in each node until some stop condition is reached. "Natural" stopping condition may be: i) all leaf nodes are pure, and contain observations only of one category of the outcome (all cases or all controls). In this case no gain in purity can be further obtained; ii) only one subject is included in the node, and no further split is practicable. Otherwise, some stopping rules can be decided at the beginning of the process: iii) a given threshold for the minimum number of observations in a node (i.e. 5 subjects, or 1% of the sample size); iv) a given threshold for the minimum change in the impurity measure; v) a maximum number of allowable splits.

This process may lead to overfitting. Overfitting refers to the fact that a classifier adapts too closely to the training dataset and it may fit noisy instances and random variation that is present in the learning data due to random sampling, beside discovering the systematic components of the structure present in the population. Overfitting leads to poor test performance, thus, when a overfitted model is applied to the validation set, its performance and its predictive power will be limited.

In decision trees, a common strategy is to eliminate those parts of a classifier that are likely to overfit the training data. This process is called pruning, and can increase both the accuracy and the interpretability of the resulting classifier. The success of a pruning mechanism depends on its ability to distinguish noisy instances from predictive patterns in the training data. To prune the tree means to eliminate branches that do not add to the prediction accuracy after growing the tree.

In the pruning process, a very large tree T_0 is used. Then, selected branches of the initial large tree are eliminated from the end of the tree (bottom-up procedure), and a sequence of nested candidate sub-trees are generated. Our goal is to select the sub-tree leading to the lowest test error rate.

Classification tree analysis uses a method of cost-complexity pruning in which child nodes are pruned and the predicted misclassification cost is compared with the change in tree complexity, yielding a number of smaller nested trees.

A technique called cross-validation is used to provide unbiased estimates of the misclassification error rates in each of the candidate sub-trees, and to identify the most complex sub-tree that minimizes the cross-validated misclassification rate. In a 10-fold cross-validation, the dataset is randomly partitioned in 10 subsamples, each containing 10% of the subjects. A new classification tree is generated on the 90% of subjects, while the remaining 10% of the dataset is set aside. The misclassification error rate is calculated on this 10% of subjects as the validation sample. This process is repeated for all the 10 partitions, and an average of the 10 misclassification rates gives a 10-fold cross-validated and unbiased estimate of the overall misclassification for each candidate sub-tree. This approach balances the trade-off of complexity (i.e. the number of predictors and terminal nodes) in the final tree while saving to minimize the overall misclassification rate.

1.3.1. 1-SE rule

In order to choose the optimal size of the tree, researchers may prune the tree in correspondence to the minimum value of errors. However, when the minimum value of errors is considered to prune the tree, the position of this minimum value may be unstable, and small changes in the parameter values or in the random number generator used to create the k subsets for the cross-validation techniques may be cause large changes in the number of optimal splits. Thus, the optimal size of the tree can be determined with the 1-SE rule (Breiman et al 1984). In fact, 1-SE rule allow researchers choosing the optimal number of nodes, reducing the instability and choosing the simplest tree with a high accuracy.

Let's define T_{k0} the tree that minimizes the misclassification errors:

$$\hat{R}(T_{k0}) = min_k \hat{R}(T_k) \tag{1.7}$$

The 1-SE rule choose the tree T_{k1}, where k_1 is the maximum k satisfying the following:

$$\hat{R}(T_{k1}) \leq \hat{R}(T_{k0}) + SE(\hat{R}(T_{k0})) \tag{1.8}$$

1.4. Evaluation of performance

With the creation of a final classification tree, we can use the information provided by the terminal nodes to evaluate the performance of the tree and the accuracy of the predictions. Indeed, each terminal node is categorized as "case" or "control" according to the most frequent outcome category of the observations included in the terminal node. Therefore, each individual included in a terminal node is predicted as a case or a control according to this simple rule. A perfect prediction rule is the one that makes no mistakes in the prediction of the outcome category for each of the individual in the validation set.

A simple 2x2 confusion matrix can be built as in **Figure 1.5**, where columns represent the instances in an actual class, while rows represent the instances in a predicted class

TP represents true positive subjects, that is cases correctly classified as cases; TN are true negatives subjects, that is controls correctly classified as controls; FP are false positive subjects, that is controls wrongly predicted as cases; and FN are false negative subjects, that is cases wrongly predicted as controls.

ACTUAL PREDICTED	Case	Control
Case	TP	FP
Control	FN	TN

***Figure 1.5**. Confusion matrix.*

With these simple quantities, different accuracy measures can be computed in order to evaluate the performance of the tree model.

i) Accuracy (1-error rate)=$\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$

ii) Sensitivity=$\frac{\text{TP}}{\text{TP+FN}}$

iii) Specificity=$\frac{\text{TN}}{\text{TN+FP}}$

iv) Youden index=Sensitivity + Specificity – 1

v) Area under the ROC curve (Receiver Operating Characteristic): AUC

The first three measures range between 0 and 1, with high values representing high accuracy. The Youden index ranges between -1 and 1, with values smaller than 0 indicating the accuracy is scarce.

The ROC curve is a plot which illustrates the performance of a binary classifier as its discrimination threshold varies. The curve is created by plotting the true positive rate (sensitivity) on the y-axis, against the false positive rate (1-Specificity) on the x-axis, at various threshold settings.

An AUC near 0 is the situation represented by the blue line. Such a test does not have any diagnostic benefit. The ideal situation is represented by the green curve, and represents the situation in which both cases and controls are perfectly classified. In this case the AUC is 1. A possible real situation is the one represented by the red curve, with the AUC ranging between 0 and 1.

These parameters can also be used to perform model comparisons, to compare the performance of various models in the terms of accuracy of the predictions.

2. ORDINAL CLASSIFICATION TREES

Ordinal variables are variables that have two or more categories, with categories that can also be ordered or ranked. Suppose n independent observations to be classified. These observations are characterized by a vector of p predictors **x**, and by one of the J classes of the outcome variable. The proportion of subjects in each of the J classes within a node are called node proportions, that is $p(w_j|t)$ for j=1,.....,J such that $p(w_1|t) + p(w_2|t) + \cdots + p(w_J|t) = 1$. In *Chapter 2* this case will be explained in more detail.

2.1. Splitting rules

In classification trees, the split of a parent node is performed according to a measure called impurity function, which is a measure of heterogeneity in the node with respect to the outcome variable class. Thus, the optimal split is defined as the split providing the largest decrease in node impurity, resulting in increasingly more homogeneous nodes with respect to the outcome class.

2.1.1 Generalized Gini impurity function

The Gini criterion, the most commonly used within-node impurity measure in case of nominal response classification (Breiman et al 1984), has the following formula:

$$i_G(t) = \sum_k \sum_{k \neq l} p(w_k\,|t)p(w_l|t) \tag{2.1}$$

where $p(w_k|t)$ is the proportion of observations in the node t belonging to the class k (among the J possible classes) of the outcome.

This impurity measure does not take advantage of the additional information present when the response is ordinal. For this reason, the generalized Gini impurity function (Breiman et al 1984) should be considered:

$$i_{GG}(t) = \sum_{k} \sum_{k \neq l} C(w_k|w_l) p(w_k\,|t) p(w_l|t) \qquad (2.2)$$

The additional factor $C(w_k|w_l)$ is the misclassification cost of assigning an observation belonging to the class l of the outcome, to the class k.

For any binary split s of node t, units assigned to node t are partitioned into two child nodes, t_R and t_L. The decrease in node impurity with the split s in node t is given by:

$$\Delta Imp_{GG}(t,s) = p(t) i_{GG}(t) - p(t_R) i_{GG}(t_R) - p(t_L) i_{GG}(t_L) \qquad (2.3)$$

where p(t), p(t_R), and p(t_L) are the proportions of units assigned to nodes t, t_R and t_L, respectively. The split, among all the possible binary splits for a given node, resulting in the largest value of $\Delta Imp_{GG}(t,s)$ is selected:

$$s^*(t)=\arg\max_s \ \Delta Imp_{GG}(t,s) \qquad (2.4)$$

In the ordinal setting, $C(w_k|w_l)$ is a measure of dissimilarity between the actual and the assigned category, taking into account the ordinal nature of the response variable. Suppose to assign a set of increasing scores $\{s_1 < s_2 < \cdots < s_J\}$ to the ordered categories of the response Y. Variable misclassifications costs can be defined by considering suitable transformations of the absolute differences between pairs of scores. Possible choices of $C(w_k|w_l)$ can be the absolute difference $C(w_k|w_l) = |s_k - s_l|$ or the squared difference $C(w_k|w_l) = (s_k - s_l)^2$. In the first case, *function 2.2* becomes:

$$i_{GG1}(t) = \sum_{k} \sum_{k \neq l} |s_k - s_l| p(w_k\,|t) p(w_l|t) \qquad (2.5)$$

And in the second case, it becomes:

$$i_{GG2}(t) = \sum_{k} \sum_{k \neq l} (s_k - s_l)^2 p(w_k\,|t) p(w_l|t) \qquad (2.6)$$

With these functions, ordinal classification trees give more importance in the impurity function on observations that are incorrectly classified far from the true class.

The nominal setting is a special case of the nominal one, with $C(w_k|w_l)$ assuming only two possible values: $C(w_k|w_l) = 1$ when k≠j and the classification is incorrect, and $C(w_k|w_l) = 0$ when k=j and the classification is correct.

2.1.2 Ordered twoing method

The twoing criterion was introduced by Breiman and colleagues (Breiman et al 1984) and is an alternative method to evaluate a split. Let's call $C(t) = \{y_1, y_2, \dots, y_J\}$ the set of categories of the outcome Y in the node t. The twoing method proceeds by reformulating the outcome as a vector of dichotomous responses.

In the nominal case, the set $C(t)$ is divided in two subsets $C_1(t) = \{y_{g1}, y_{g2}, \dots, y_{gh}\}$ and $\overline{C_1}(t) = C \backslash C_1$. The binary response variable can be defined as:

$$Y^* = \begin{cases} 1 \;\; if\; Y \in \; C_1 \\ \\ 0 \;\; otherwise \end{cases} \tag{2.7}$$

The impurity within a node t, depending on the choice of $C_1(t)$, is measured as:

$$i_{NT}(t|C_1) = 2\, p(C_1|t)p(\overline{C_1}|t) \tag{2.8}$$

where $p(C_1|t) = \sum_{g:y_g \in C_1} p(w_g|t)$.

Given C_1, the decrease in node impurity provided by split s is:

$$\begin{aligned} \Delta Imp_{NT}(t,s|C_1) &= i_{NT}(t,C_1) - p(t_R)i_{NT}(t_R,C_1) - p(t_L)i_{NT}(t_L,C_1) \\ &= 2p(t_R)p(t_L)[p(C_1|t_R) - p(C_1|t_L)]^2 \end{aligned} \tag{2.9}$$

This formula allow us to choose the best split of the set of classes of the outcome C_1 through this criterion:

$$C_{1|s}\text{*=arg max}_{\text{C1}}\ \Delta Imp_{NT}(t,s|C_1) \tag{2.10}$$

Split s is evaluated through:

$$\Delta Imp_{NT}(t,s) = 2p(t_R)p(t_L)\left[p\left(C_{1|s}^*\middle|t_R\right) - p\left(C_{1|s}^*\middle|t_L\right)\right]^2 \tag{2.11}$$

The best split is:

$$\text{s*(t)=arg max}_{\text{s}}\ \Delta Imp_{NT}(t,s) \tag{2.12}$$

The nominal twoing method can be easily extended to the ordinal setting. While in the nominal case any subset of the categories of the outcome Y can be considered, in the ordinal case $C_j = \{y_1, y_2, \dots, y_j\}$ and $\overline{C_j} = \{y_{j+1}, y_{j+2}, \dots, y_J\}$. Thus, the binary response variable can be defined as:

$$Y^* = \begin{cases} 1 \ \ if\ Y \in\ C_j \\ \\ 0 \ \ otherwise \end{cases} \tag{2.13}$$

Recalling the formula of the nominal case:

$$\Delta Imp_{OT}(t,s|C_j) = p(t_R)p(t_L)\left[p\left(C_j\middle|t_R\right) - p\left(C_j\middle|t_L\right)\right]^2 \tag{2.14}$$

Since $p(C_j) = \sum_{k=1}^{j} p(w_k|t) = F\left(w_j\middle|t\right)$ is the cumulative distribution function (cdf) of Y evaluated in y_j, the formula can be also written as:

$$\Delta Imp_{OT}(t,s|C_j) = p(t_R)p(t_L)\left[F\left(w_j\middle|t_R\right) - F\left(w_j\middle|t_L\right)\right]^2 \tag{2.15}$$

For a given split, the class maximizing $\Delta Imp_{OT}(t,s|C_j)$ is $C_{j*|s}$ with

$$\text{j*=arg max}_{\text{g}}\ \Delta Imp_{OT}(t,s|C_j) \tag{2.16}$$

Hence, the ordered twoing criterion to evaluate a split is

$$\Delta Imp_{OT}(s,t) = p(t_R)p(t_L)max_j\left[F\left(w_j\middle|t_R\right) - F\left(w_j\middle|t_L\right)\right]^2 \tag{2.17}$$

2.1.3 Ordered impurity function

This is an ordinal impurity function for deriving an ordinal response classification tree based on a measure of nominal-ordinal association (Piccarreta 2001) that does not require the assignment of costs of misclassification. The formula is:

$$i_{OS}(t) = \sum_{j=1}^{J} F(w_j|t)[1 - F(w_j|t)] \tag{2.18}$$

where $F(w_j|t) = \sum_{k=1}^{j} p(w_k|t)$ (Archer 2010).

2.2 Pruning the tree

A crucial issue in classification tree methodology is the choice of the optimal tree size. In fact, a large tree built on the training set of data may fit the training data very well, but may poorly predict the testing set of data (overfitting).

One of the most popular techniques is the pruning procedure based on a cost-complexity pruning in which child nodes are pruned and the predicted misclassification cost is compared with the change in tree complexity, yielding a number of smaller nested trees. The choice of functional form for $R(T)$ depends on the nature of Y.

$$R_\alpha(T) = R(T) + \alpha \cdot card(T) \tag{2.19}$$

- $R(t)$is a measure of the predictive performance of the tree T;
- *card(T)* is a measure of the complexity of the tree, usually measured as the number of leaves or terminal nodes;
- α is a tuning parameter that controls the trade-off between the predictive performance and the tree complexity (α>0).

2.2.1 Misclassification rate and misclassification cost

Two predictive performance measures to be used when Y is and ordinal outcome are the total number of misclassified observations $R_{mr}(T)$ (Breiman et al 1984) and the total misclassification costs $R_{mc}(T)$ (Archer 2010). Suppose to assign a set of increasing scores $\{s_1 < s_2 < \cdots < s_J\}$ to the ordered categories of Y; the measures to predict the performance will be computed as follows:

$$R_{mr}(T) = \sum_{i=1}^{n} \left[1 - I_{\{s_i\}}(\hat{s}_{i,T})\right] \tag{2.20}$$

$$R_{mc}(T) = \sum_{i=1}^{n} \left|s_i - \hat{s}_{i,T}\right| \tag{2.21}$$

- s_i is the observed score for unit i;
- $\hat{s}_{i,T}$ is the predicted score for unit i according to the tree T;
- $I_{\{s_i\}}(\hat{s}_{i,T})$ is an indicator that indicates whether the prediction of the score for the unit i is the same as the real score. Thus, its value will be 1 if $s_i = \hat{s}_{i,T}$ and 0 otherwise. $R_{mr}(T)$ is a sum of all the observations that are classified incorrectly. On the other hand, $R_{mc}(T)$ is a sum over all the observations of the absolute difference between the observed score and the predicted score.

$\hat{s}_{i,T} = \hat{s}(t)$ for all units i in the terminal node t of the tree T, according to the splitting rules in the tree T. Thus, $\hat{s}(t)$ is the predicted value of Y within a node t. In particular, $\hat{s}(t)$ is given by the within-node modal score when $R_{mr}(T)$ is chosen, and by the within-node median score when $R_{mc}(T)$ is used.

2.3 Performance of the classifier

For nominal outcomes, the overall predictive performance of the classifier is often determined through the misclassification error rate, that is the number of misclassified observations (the prediction is wrong) on the total number of observations, or on accuracy measures, including sensibility, specificity, and

Youden index (see *Section 1.4*). For ordinal outcomes, it may be of more interest to estimate the predictive performance through some ordinal measures of association between the observed and the predicted response, taking into account the ordinal classification. Two of these measures are the gamma statistic and the Somers' d measure (Agresti 2002).

When X and Y are ordinal, a monotone trend association may be estimated. As the level of X increases, responses on Y tend to increase toward higher levels, or responses on Y tend to decrease toward lower levels. Some measures, including the gamma and the Somers'd statistics, are based on classifying each pair of subjects as concordant or discordant.

A pair is defined concordant if the subject ranking higher on X also ranks higher on Y; a pair is defined discordant if the subject ranking higher in X ranks lower on Y; a pair is tied if the subjects have the same classification on X and/or Y. P is the total number of concordant pairs and Q is the total number of discordant pairs. Moreover, we can compute the number of pairs of observations that are untied on X, regardless of their status on Y (X_u), or the number of pairs of observations that are untied on Y, regardless of their status on X (Y_u).

Table 2.1. *Hypothetical frequency distribution for two ordinal variables.*

	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$
$\mathbf{Y_1}$	a	b	c
$\mathbf{Y_2}$	d	e	f

Considering the example proposed in **Table 2.1**, important quantities to derive ordinal association measures are computed as follows:

- $P = a(e + f) + b(f)$ is the number of concordant pairs;
- $Q = c(d + e) + b(d)$ is the number of discordant pairs;
- $X_o = ad + be + cf$ is the number of pairs tied only X;

- $\mathrm{Y_o} = a(b+c) + bc + d(e+f) + ef$ is the number of pairs tied only Y;
- $\mathrm{X_u} = P + Q + \mathrm{Y_o}$ is the number of pairs untied on X regardless their status on Y ;
- $\mathrm{Y_u} = P + Q + \mathrm{X_o}$ is the number of pairs untied on Y regardless their status on X .

2.3.1 Gamma statistics

Gamma statistics is computed as follows:

$$\gamma = \frac{P-Q}{P+Q} \tag{2.22}$$

It is a symmetric measure of ordinal association between the observed and the predicted score. This means that it is unnecessary to identify one classification as the response variable. Gamma statistics range is $-1 \leq \gamma \leq +1$. Its symmetric nature implicate that a reversal in the category ordering of one variable causes a change in the sign of γ (Agresti 2002).

2.3.2 Somers'd measure

Unlike the gamma statistics, Somers'd measure is an asymmetric measure of ordinal association. Two different versions of this measure exists, and they differ for the denominator of the ratio:

$$d_{yx} = \frac{P-Q}{P+Q+\mathrm{Y_o}} \tag{2.23}$$

$$d_{xy} = \frac{P-Q}{P+Q+\mathrm{X_o}} \tag{2.24}$$

Somers'd measure range is $-1 \leq d \leq +1$ (Somers 1962).

2.3.3 Friedman nonparametric test

The Friedman test is the non-parametric alternative to the one-way ANOVA with repeated measures. It is used to test for differences between groups when the dependent variable being measured is ordinal. It can also be used for continuous data that has violated the assumptions necessary to run the one-way ANOVA with repeated measures (e.g., data that has marked deviations from normality). Friedman rank test is a test on median, and is used to determine whether c groups have been selected from populations having equal medians (M). The hypothesis is expressed as follows:

$$\begin{cases} H_0: M_1 = M_2 = \cdots = M_c \\ H_1: M_i \neq M_j \; for \; at \; least \; one \; i \neq j \; (i, j = 1, \ldots, c) \end{cases} \tag{2.25}$$

The test has the following formula:

$$F_R = \frac{12}{rc(c+1)} \sum_{j=1}^{c} R_{.j}^2 - 3r(c+1) \tag{2.26}$$

- R_{ij} is the rank (from 1 to c) associated with the *j*th group in the *i*th block. Data values have to be replaced in each of the r independent blocks with the corresponding ranks, assigning rank 1 to the smallest value in the block, and rank c to the largest. Tied values within a block have to be assigned the mean of the ranks that they would otherwise have been assigned.

- $R_{.j}^2$ is the square of the total ranks for group j (j=1,...,c)

- r is the number of blocks

- c is the number of groups

Approximated test statistic assume a chi-square distribution with c-1 degrees of freedom. Thus, for any selected level of significance α, the null hypothesis have to be rejected if the observed value of F_R is greater than χ_α^2 with c-1 degrees of freedom.

2.4 rpartScore

rpartScore is the R package implemented by Galimberti and colleagues to build classification trees for ordinal responses (Galimberti et al 2012). Previously, another R package, rpartOrdinal, was created for the same purpose by Archer (Archer 2010).

Authors implement the Somers'd measure to assess the performance of the classifier, instead of the gamma statistics used by Archer and colleagues in rpartOrdinal (Archer 2010). This choice is motivated by the fact that symmetric measures are not defined when the predicted score is constant for all the units (which happens when the optimal size of the tree is 1) (Galimberti et al 2012). In rpartScore the Somers'd measure untied with respect to the observed scores is implemented. Considering X as the observed scores and Y as the predicted scores, the measure is that reported in d_{yx}.

3. APPLICATION TO REAL DATA – NOMINAL CLASSIFICATION TREES

3.1 Case-control studies

To conduct the present analyses, I used data from 3 Italian case-control studies on various cancers. The studies protocols were approved by ethical committees of the hospitals involved according to the regulations at the time of the each study conduction, and all participants gave informed consent to participate.

3.1.1. Description of the case-control study on OCP cancer

The case-control study on oral cavity and pharyngeal (OCP) cancer was conducted between 1991 and 2010 in northern (the greater Milan area and the province of Pordenone) and central Italy (the provinces of Rome and Latina), and in the Swiss Canton of Vaud.

Cases were 1507 patients (1220 men and 287 women) under age 80 (median: 58 years, range: 22-79 years) with incident, histologically confirmed squamous cell cancer of the OCP (excluding cancer of the lip, salivary glands, and nasopharynx), admitted to the major teaching or general hospitals in the areas under investigation. Controls were 3849 (2619 men and 1230 women, median age: 58 years, range 19-82 years) with no previous history of cancer, admitted to the same hospitals as cases for acute, non-neoplastic conditions, unrelated to tobacco smoking, alcohol drinking, or long-term dietary modifications.

Among controls, 17% were admitted for traumas, 33% for other orthopaedic conditions, 25% for acute surgical conditions, and 25% for miscellaneous other illnesses, including eye, nose, ear, skin or dental disorders. The proportion of refusals of subjects approached for interview was <5% in Italy and about 15% in Switzerland.

Only participants with complete information about the considered predictors (socio-demographic variables, lifestyle habits and dietary indicators) were included in the present analyses. Thus, results are based on a sample of 1493 cases and 3816 controls (n=5309).

3.1.2. Description of the case-control study on breast cancer

The case-control study on breast cancer was a multicentric case–control study conducted from June 1991 to April 1998 in six Italian areas: the provinces of Pordenone and Gorizia, the greater Milan area, the urban area of Genoa, the province of Forli, the province of Latina, and the urban area of Naples.

Cases were 2569 women with incident, histologically confirmed breast cancer (median age 55, range 23–78 years) admitted to major teaching and general hospitals of the study areas. Controls were 2588 women (median age 55, range 19–79 years) with no history of cancer admitted to the same hospitals for acute, non-neoplastic, nongynecological conditions, unrelated to hormonal or digestive tract diseases or to conditions linked to diet. Among controls, 22% had traumas, 33% other orthopedic diseases such as low back pain or strains, 15% acute surgical conditions, 18% eye diseases, and 12% other miscellaneous diseases. Less than 4% of cases and controls approached for interview did not consent to participate.

Only participants with complete information about the considered predictors (socio-demographic variables, lifestyle habits and dietary indicators) were included in the present analyses. Thus, results are based on a sample of 2548 cases and 2545 controls (n=5093).

3.1.3. Description of the case-control study on prostatic cancer

The case-control study on prostatic cancer was a multicentric case–control study conducted between 1991 and 2002 in a network of 57 teaching and general hospitals in the greater Milan area, the provinces of Pordenone and Gorizia in northern Italy, the province of Latina in central Italy, and the urban area of Naples in southern Italy.

Cases were 1294 men (median age 65, range 46-74 years) admitted with incident, histologically confirmed prostate cancer to a network of hospitals in the areas under investigation. Controls were 1451 patients (median age 63, range 46-74 years) admitted to the same hospitals as cases for a wide spectrum of acute, nonmalignant conditions, unrelated to long-term modifications of diet. Among controls, 32% had nontraumatic orthopaedic disorders, 21% traumas, 17% surgical conditions, and 29% miscellaneous other illnesses, such as eye, ear, and skin disorders. Cases and controls were identified and questioned by centrally trained interviewers who regularly visited the departments of the selected hospitals and approached the patients eligible as cases or controls on the basis of the admission diagnosis reported in the clinical records. Of the subjects approached, only 3% of cases and 4% of controls refused to be interviewed.

Only participants with complete information about the considered predictors (socio-demographic variables, lifestyle habits and dietary indicators) were included in the present analyses. Thus, results are based on a sample of 1281 cases and 1438 controls (n=2719).

3.1.4. Data collection

Data were collected by trained interviewers who admitted a structured questionnaire to cases and controls during hospitalization. The questionnaire included information on socio-demographic characteristics, anthropometric measures, and selected lifestyle habits (including tobacco smoking and alcohol drinking). Subjects' dietary habits during the 2 years before cancer diagnosis (for cases) or hospitalization (for controls) were assessed through a validated (Decarli et al 1996) and reproducible (Franceschi et al 1993) food frequency questionnaire (FFQ).

The FFQ included information on weekly consumption of 78 foods and beverages, as well as a range of recipes, that are, the most common ones in the Italian and Swiss diet, grouped into seven sections:

i) bread and cereals dishes(first courses);
ii) meat and other main dishes (second courses);
iii) vegetables (side dishes);
iv) fruit;
v) sweets, desserts, and soft drinks;
vi) milk and hot beverages;
vii) alcoholic beverages.

Subjects were asked to indicate the average weekly frequency of consumption of each of the dietary items; occasional intake (lower than once a week, but at least once a month) was coded as 0.5 per week.

3.1.5. Variables

The variables of interest considered in the present thesis can be divided in 3 main groups (see **Table 3.1** for details on the variable names and the corresponding descriptions):

i) Socio-demographic characteristics

In this group I included the following variables (confounders): age, sex, centre and years of education.

ii) Lifestyle habits

Lifestyle habits considered in the present analyses were tobacco smoking and alcohol drinking. Variables of interest were: smoking status (current, ex- and never smokers), smoking consumption (number of cigarettes smoked per day), smoking duration (computed as the difference between age at the interview and age at smoking initiation for current smokers, and as the difference between age at quitting and age at starting smoking for ex-smokers), alcohol drinking status (current, ex- and never drinkers), and alcohol consumption (number of drinks per week, computed as the sum of beer, wine, and spirits). This variable was categorized in a 4-classes variable. The categories were: 0 drinks/week (abstainers), 0.1-7 drinks/week, 7.1-14 drinks/week, and ≥14.1 drinks/week.

iii) Food groups

Food items were combined in 25 food groups: milk and yogurt, coffee, decaffeinated coffee and tea, bread, pasta and rice, soups, eggs, poultry, red meat, liver, pork and processed meat, fish, cheese, potatoes, pulses (i.e., green peas, beans, lentils), leafy vegetables, fruiting vegetables, root vegetables, cruciferous vegetables, other vegetables, citrus fruits, other fruits, soft drinks and fruit juices, desserts, sugar and candies. The weekly intake for each group was calculated summing up the intake of each food item included in the group.

Table 3.2 provides the distribution of cases and controls, according to sex, age, centre, and other selected characteristics.

Table 3.3 provides the distribution of breast cancer cases and controls, according to age, centre, and other selected characteristics. **Table 3.4** provides average weekly consumption of various food groups among women with breast cancer and corresponding controls, separately. Women with breast cancer consumed bread and processed meat more frequently than controls, while the weekly consumption of white meat, fish, and root vegetables was more frequent among controls.

Table 3.5 provides the distribution of prostatic cancer cases and controls, according to age, centre, and other selected characteristics. **Table 3.6** provides average weekly consumption of various food groups among men with prostatic cancer and controls, separately. Milk, bread, fish, and cheese were more frequently consumed by men with prostatic cancer, while the weekly intake of liver and root vegetables was higher in controls.

3.2 Statistical analyses

I performed classification tree analyses to relate selected risk factors (described in **Table 3.1**) and their interactions with OCP, breast and prostatic cancer risk. Since these are case-control studies, the outcome is a dichotomous variable expressing whether a participant is a case or a control. The 34 predictors were either continuous or categorical variables, and node splits in continuous variables can occur at any non-predetermined value.

For OCP cancer, the analyses were implemented with the R software and two different packages performing recursive partitioning techniques were used: *rpart* and *tree* packages. I built the classification tree with all the 34 predictors. In the tree built with function in the *rpart* package, the Gini index was used as the method to maximize node separation (**model tc1**), while with the tree package, two different classification trees were built: the first one used the Gini index (**model tc2**), and the second one the deviance, or cross-entropy, method (**model tc3**) to reduce nodes impurity. For breast cancer analyses, two different models were performed using *rpart* and Gini index: the first one included all the 34 predictors (**model tb1**), and the second one included only the 25 predictors of food groups (**model tb2**). For prostatic cancer analyses, the classification tree with all the 34 predictors was built (**model tp1**).

Each of these trees was built through three steps: i) construction of a large classification tree using recursive partitioning to choose the predictors variables; iii) selection of an optimum-size tree from a nested sequence of smaller trees using the cross-validation technique and using the 1-SE rule (*Section 1.3.1*); and iii) pruning the trees and obtaining the final optimal classification tree.

Both for OCP cancer, and for breast and prostatic cancer case-control studies, the entire datasets were randomly divided into a learning (with 70% of observations) and a validation subset (30%), maintaining the same proportion of cases and controls in each of the two groups. The learning set was used to create the models and select the best classification tree, whose performance was evaluated on the validation set by means of different measures (accuracy, sensitivity, specificity, and Youden index).

For OCP cancer case-control study, the training dataset included 3716 observations (1045 cases and 2671 controls), and the validation dataset included 1593 women (448 cases and 1145 controls). For breast cancer case-control study, the training dataset included 3564 observations (1783 cases and 1781 controls), and the validation dataset included 1529 women (765 cases and 764 controls). For prostatic cancer case-control study, the training dataset included 1902 men (896 cases and 1006 controls), and the validation dataset included 817 men (385 cases and 432 controls).

***Table 3.1**. Names of the variables included in the models, and corresponding descriptions.*

Socio-demographic characteristics (k=4)	
Sex	Sex (2 categories)
V8	Age (continuous)
Center	Centre (7 categories)
V12	Years of education (continuous)

Lifestyle habits (k=5)	
Statusfum	Smoking status (3 categories)
Fum3	Number of cigs/day (continuous)
Duration	Smoking duration (ex and current smokers)
Statusalc	Alcohol drinking status (3 categories)
Alccat	Alcohol consumption (6 categories)

Food groups (k=25)	Weekly consumption of... (continuous)
Milk	Milk
Cafcap	Coffee
Tedec	Tea and decaffeinated coffee
Bre	Bread
Pas	Pasta and rice
Nsou	Soup
Egg	Egg
Pou	White meat
Redmeat	Red meat
Offals	Liver
Pork	Processed meat
Fish	Fish
Che	Cheese
Pot	Potatoes
Pul	Pulses
Leafy	Leafy vegetables
Fruiting	Fruiting vegetables
Root	Root vegetables
Cruc	Cruciferous vegetables
Othver	Other vegetables
Cfru	Citrus fruit
Fru	Other fruit
Sdrink	Soft drinks and fruit juices
Des	Desserts
Sug	Sugar and candies

***Table 3.2**. Distribution of 1493 incident cases of oral and pharyngeal cancer and 3816 controls according to centre, sex, age and other selected characteristics. Italy and Switzerland, 1991-2010.*

	Cases		**Controls**	
Characteristics	**N**	**%**	**N**	**%**
Centre				
Milan (1992-2009)	337	22.5	973	25.5
Pordenone (1991-1997)	492	33.0	1048	27.5
Rome and Latina (1994-97)	104	7.0	438	11.5
Switzerland (1992-2010)	560	37.5	1357	35.5
Sex				
Men	1209	81.0	2599	68.1
Women	284	19.0	1217	31.9
Age group (years)				
<50	285	19.1	964	25.3
50-59	573	38.4	1260	29.5
60-69	480	32.1	1209	31.6
≥70	155	10.4	517	13.6
Level of education (years)^				
<7	550	36.8	1274	33.4
7-11	461	30.9	1160	30.4
≥12	482	32.3	1382	36.2
Smoking status^				
Never smokers	174	11.7	1718	45.0
Ex-smokers	313	21.0	1016	26.6
Current smokers				
<15 cigarettes/day	163	10.9	431	11.3
15-24 cigarettes/day	409	27.4	507	13.3
≥25 cigarettes/day	434	29.0	144	3.8
Alcohol consumption (drinks per day)				
0	103	6.9	1117	29.3
1-7	711	47.6	2463	64.5
8-14	519	34.8	196	5.1
15-28	146	9.8	35	0.9
>28	14	0.9	5	0.1

***Table 3.3**. Distribution of 2548 incident cases of breast cancer and 2545 controls according to centre, age and other selected characteristics. Italy, 1991-1998.*

	Cases		Controls	
Characteristics	**N**	**%**	**N**	**%**
Centre				
Pordenone (1991-98)	1040	40.8	1002	39.4
Milan (1991-98)	582	22.8	615	24.2
Genova (1991-94)	290	11.4	309	12.1
Forlì (1992-94)	210	8.2	209	8.2
Naples (1991-94)	251	9.9	235	9.2
Rome and Latina (1992-94)	175	6.9	175	6.9
Age group (years)				
<35	85	3.3	135	5.3
35-44	380	14.9	326	12.8
45-54	764	30.0	686	27.0
55-64	794	31.2	792	31.1
≥65	525	20.6	606	23.8
Level of education (years)				
<7	1256	49.3	1558	61.2
7-11	713	28.0	640	25.2
≥12	579	22.7	347	13.6
Smoking status				
Never smokers	1675	65.7	1741	68.4
Ex-smokers	341	13.4	241	9.5
Current smokers				
<15 cigarettes/day	317	12.4	339	13.3
15-24 cigarettes/day	178	7.0	190	7.5
≥25 cigarettes/day	37	1.5	34	1.3
Alcohol consumption (drinks per day)				
0	1327	52.1	1398	54.9
1-7	1211	47.5	1138	44.7
8-14	9	0.3	4	0.2
>14	1	0.1	5	0.2

Table 3.4*. Mean weekly consumption of different food groups for cases of breast cancer and controls, separately. Italy, 1991-1998.*

Food groups	**Cases**		**Controls**		**p**
	Mean	**SD**	**Mean**	**SD**	
Milk	6.84	6.7	6.96	6.7	
Coffee	14.56	10.7	14.75	11.5	
Tea and decaffeinated coffee	2.62	4.8	2.89	5.7	
Bread	18.43	10.3	17.45	10.4	*
Pasta and rice	4.84	2.1	4.77	2.1	
Soup	2.28	1.8	2.31	1.9	
Egg	1.34	1.3	1.32	1.2	
White meat	1.96	1.3	2.07	1.4	*
Red meat	3.97	2.1	3.84	2.0	*
Liver	0.19	0.3	0.20	0.3	
Processed meat	2.72	2.0	2.56	2.0	*
Fish	1.72	1.1	1.83	1.2	*
Cheese	4.56	2.8	4.47	2.8	
Potatoes	1.64	1.2	1.70	1.3	
Pulses	1.66	1.2	1.65	1.2	
Leafy vegetables	5.30	3.2	5.39	3.2	
Fruiting vegetables	2.91	2.5	2.91	2.5	
Root vegetables	2.07	2.2	2.21	2.3	*
Cruciferous vegetables	0.43	0.6	0.45	0.6	
Citrus fruit	4.38	4.4	4.38	5.2	
Other fruit	13.88	9.0	14.07	10.6	
Soft drinks and fruit juices	2.13	6.1	2.00	5.6	
Desserts	6.33	6.1	6.11	6.9	
Sugar and candies	30.96	24.3	30.4.0	25.5	

* Statistically significant difference of consumption between cases and controls (α=0.05).

***Table 3.5**. Distribution of 1281 incident cases of breast cancer and 1438 controls according to centre, age and other selected characteristics. Italy, 1991-2002.*

	Cases		**Controls**	
Characteristics	**N**	**%**	**N**	**%**
Centre				
Pordenone (1991-2002)	894	69.8	948	65.9
Milan (1991-2002)	163	12.7	192	13.4
Naples (1993-2001)	127	9.9	184	12.8
Rome and Latina (1993-99)	97	7.6	114	7.9
Age group (years)				
<60	213	16.6	425	29.6
60-64	307	24.0	357	24.8
65-69	417	32.6	361	25.1
≥70	344	26.8	295	20.5
Level of education (years)				
<7	636	49.7	841	58.5
7-11	382	29.8	406	28.2
≥12	263	20.5	191	13.3
Smoking status				
Never smokers	365	28.5	342	23.7
Ex-smokers	648	50.6	684	47.6
Current smokers				
<15 cigarettes/day	106	8.3	168	11.7
15-24 cigarettes/day	131	10.2	180	12.5
≥25 cigarettes/day	31	2.4	64	4.5
Alcohol consumption (drinks per day)				
0	168	13.1	190	13.2
1-7	963	75.2	1031	71.7
8-14	137	10.7	191	13.3
>14	13	1.0	26	1.8

***Table 3.6**. Mean weekly consumption of different food groups for cases of prostatic cancer and controls, separately. Italy, 1991-2002.*

Food groups	**Cases**		**Controls**		**p**
	Mean	**SD**	**Mean**	**SD**	
Milk	6.62	6.8	5.49	6.2	*
Coffee	14.77	10.5	15.01	11.2	
Tea and decaffeinated coffee	2.83	6.2	2.41	5.1	
Bread	24.24	13.0	22.89	13.0	*
Pasta and rice	5.19	2.1	4.93	2.0	
Soup	2.20	1.7	2.26	1.7	
Egg	1.65	1.5	1.60	1.6	
White meat	1.80	1.3	1.72	1.3	
Red meat	4.16	2.3	4.14	2.3	
Liver	0.17	0.3	0.21	0.3	*
Processed meat	2.70	2.1	2.63	2.1	
Fish	1.81	1.1	1.69	1.1	*
Cheese	4.68	3.0	4.42	2.7	*
Potatoes	1.74	1.2	1.72	1.3	
Pulses	1.53	1.0	1.63	1.1	
Leafy vegetables	5.11	4.5	5.05	3.5	
Fruiting vegetables	1.46	1.5	1.46	1.4	
Root vegetables	1.74	1.8	1.91	2.3	*
Cruciferous vegetables	0.21	0.3	0.22	0.3	
Citrus fruit	1.99	3.0	1.89	3.0	
Other fruit	10.49	7.1	10.27	7.3	
Soft drinks and fruit juices	2.76	7.9	2.41	6.8	
Desserts	4.40	5.0	4.43	6.0	
Sugar and candies	36.00	29.2	34.16	26.7	

* Statistically significant difference of consumption between cases and controls (α=0.05).

3.3 Results OCP cancer

Model tc1 (classification tree with rpart package and Gini index)

The classification tree built on the training dataset is represented in **Figure 3.1**. The impurity measure used in the *rpart* package is the Gini index. In the figure, the predicted class of the outcome is displayed for each terminal node, and the printed splits indicate the left branches of the tree.

Figure 3.1 represents the pruned tree, to assure no overfitting in our data. In order to prune the data, we need to find the appropriate cost-complexity parameter, that is the parameter that minimize the error. In our data the cost-complexity parameter is 0.010 and corresponds to an ideal number of three splits. In this specific case the pruned tree is the same of the initial tree.

The root node includes all the 3716 subjects of the training set, 1045 cases and 2671 controls. The root node is therefore classified as "control" (the most frequent category), with an error of 0.28 (the proportion of cases misclassified as controls). The most important variable represented in the first split is the number of drinks per week. This predictor produces the highest improvement in nodes purity among all the predictors of the model. Individuals drinking less then around 5 (4.964) drinks per week create the terminal node 2. Those consuming an higher quantity of alcoholic beverages create the node 3. The node 3 further splits using smoking status as the predictor. Never and ex-smokers create the left node 6, while current smokers create the terminal node 7, and individuals in it are classified as cases. Never and ex-smokers are further split according to duration of smoking. Those having smoked less than 33 years (the whole sample of never smokers and a portion of ex-smokers) are classified as controls, while subjects in the other group are classified as cases. **Table 3.7** shows a description of terminal subgroups, and corresponding ORs for OCP cancer risk.

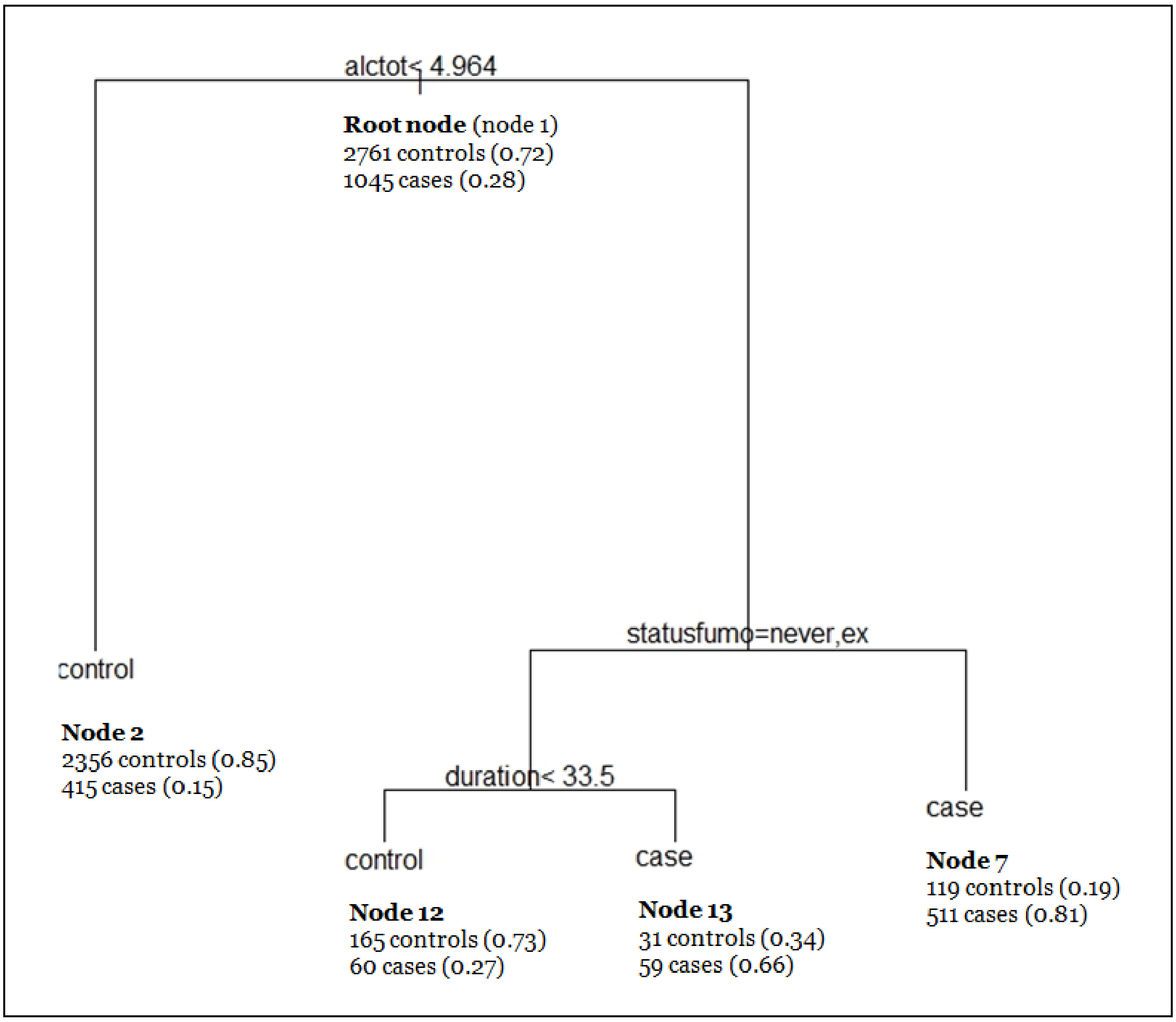

***Figure 3.1**. Classification tree tc1 on the learning set. For each terminal node, number and proportion of cases and controls, and the predicted class of the outcome are displayed. The printed splits indicate the left branches of the tree.*

Thus, the most important predictors of OCP cancer, according to this model, are alcohol drinking, smoking status and smoking duration. No dietary indicators are included as the most important predictors of OCP cancer.

***Table 3.7**. Description of terminal nodes of the classification tree tc1, and crude odds ratios (OR).*

Node	**Classification**	**Description**	**Crude ORs***
2	Control	< 5 drinks per week	1
7	Case	≥ 5 drinks per week, current smokers	24.4
12	Control	≥ 5 drinks per week, never smokers or ex-smokers, having smoked for less than 34 years	2.1
13	Case	≥ 5 drinks per week, ex-smokers, having smoked for more than 34 years	10.8

* The crude ORs for each terminal node were computed using subjects in terminal node 2 as the reference group.

In order to evaluate the performance of the model, we need to fit the model on a new set of observations, and then use the measures presented in *Section 1.4* to evaluate the accuracy of the predictions. The confusion matrix for the observations in the validation set, and performance evaluation are provided in **Table 3.8**.

***Table 3.8**. Confusion matrix and various measures of prediction accuracy on the validation set of model tc1.*

	Actual	
Predicted	**Case**	**Control**
Case	230	65
Control	218	1080

Measure	**Value**
Accuracy	0.82
Test error	0.18
Sensitivity	0.51
Specificity	0.94
Youden index	0.45

Model t1 has a good level of accuracy, indeed only 18% of individuals are misclassified. Specificity is very high (0.94), but sensitivity is quite low (0.51).

Model tc2 (classification tree with tree package and Gini index)

The classification tree built with the *tree* package, using the Gini index as the measure of impurity, is shown in Figure 2.4. This classification tree is obtained after a cost-complexity cross-validation procedure, in order to avoid overfitting and to identify the optimal number of terminal nodes of a tree. The cross-validation procedure is shown in **Figure 3.2**. The number of terminal nodes in the x-axis are related to the misclassification error rate in the y-axis. The highest value of the misclassification error rate is in correspondence to 1 terminal node. After that value the error is stable. We chose therefore 7 terminal nodes which is a good compromise between misclassification error and complexity and interpretability of the tree.

Differently from t1 model, in the t2 model two predictors belonging to the food groups category appear in the classification tree: sweeteners (osug) and milk (**Figure 3.3**). Other important predictors are alcohol consumption, smoking status and number of cigarettes smoked per day. Both node 2 and node 3 are split with alcohol consumption.

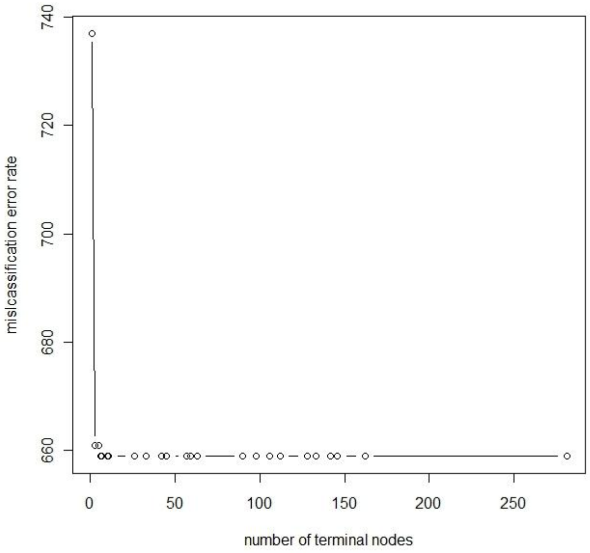

Figure 3.2*. Cross-validation procedure for tc2.*

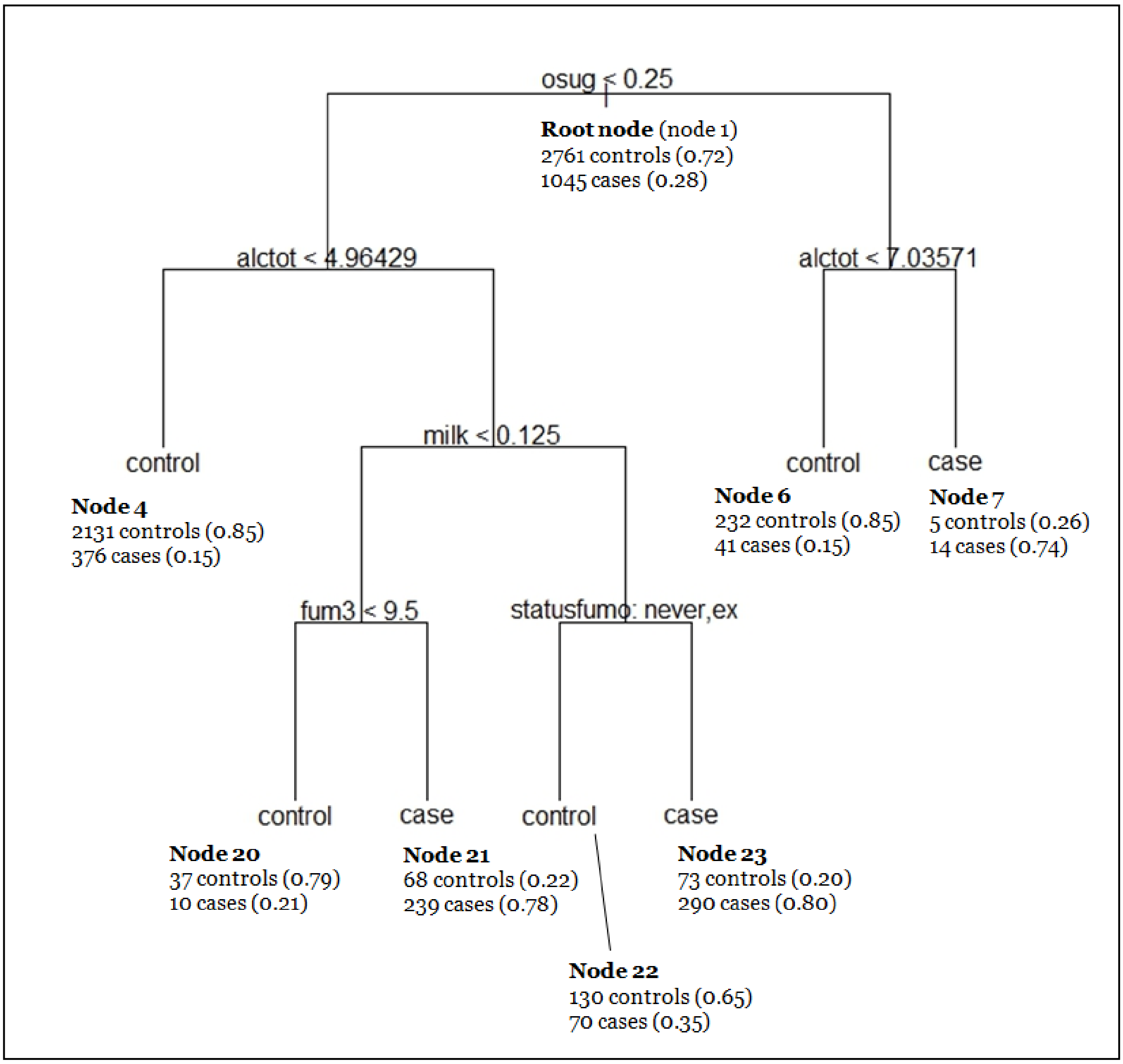

Figure 3.3. *Classification tree tc2 on the learning set.*

Table 3.9 shows the description of terminal nodes of the tree obtained with the t2 model, and the corresponding crude ORs. The subset at lower risk is the one taken as the reference category, composed by subjects consuming a low quantity of sweeteners and less than 5 drinks per day. High risk subgroups are subjects consuming an high quantity of sweeteners and more than 7 drink per week (terminal node number 7), subjects consuming a low quantity of sweeteners, more than 5 drinks per week, a low quantity of milk and are current or ex-smokers smoking (or having smoked) more than 10 cigarettes per day (terminal node 21), and subjects consuming a low quantity of sweeteners, more than 5 drinks per week, and are current smokers, independently by smoking consumption (terminal node 23).

Table 3.9. *Description of terminal nodes of the classification tree t2, and crude odds ratios (OR).*

Node	Classification	Description	Crude ORs*
4	Control	<0.25 sweeteners/week, <5 drinks/week	1
6	Control	≥0.25 sweeteners/week, <7 drinks/week	1.0
7	Case	≥0.25 sweeteners/week, ≥7 drinks/week	15.9
20	Control	<0.25 sweeteners/week, ≥5 drinks/week, <0.1 milk/week, <10 cigarettes/day	1.5
21	Case	<0.25 sweeteners/week, ≥5 drinks/week, <0.1 milk/week, ≥10 cigarettes/day	19.9
22	Control	<0.25 sweeteners/week, ≥5 drinks/week, ≥0.1 milk/week, never or ex-smokers	3.1
23	Case	<0.25 sweeteners/week, ≥5 drinks/week, ≥0.1 milk/week, current smokers	22.5

* The crude ORs for each terminal node were computed using subjects in terminal node 4 as the reference group.

As for the previous model, we build the confusion matrix to compute the measures of the accuracy of the predictions on the validation set (**Table 3.10**).

Table 3.10. *Confusion matrix and various measures of prediction accuracy on the validation set of model tc2.*

	Actual	
Predicted	**Case**	**Control**
Case	228	63
Control	220	1082

Measure	Value
Accuracy	0.82
Test error	0.18
Sensitivity	0.51
Specificity	0.94
Youden index	0.45

Model tc3 (classification tree with tree package and deviance)

The classification tree built on the training dataset using the *tree* package and the deviance impurity measure is represented in **Figure 3.4**. The most important predictor is alcoholic beverages consumption, and corresponds to the first split of the tree. Those drinking more than 5 drinks per week are further divided according to their smoking status. Never (and ex-) smokers are further divided according to smoking consumption, and those smoking more than 27 cigarettes per day are split according to bread consumption.

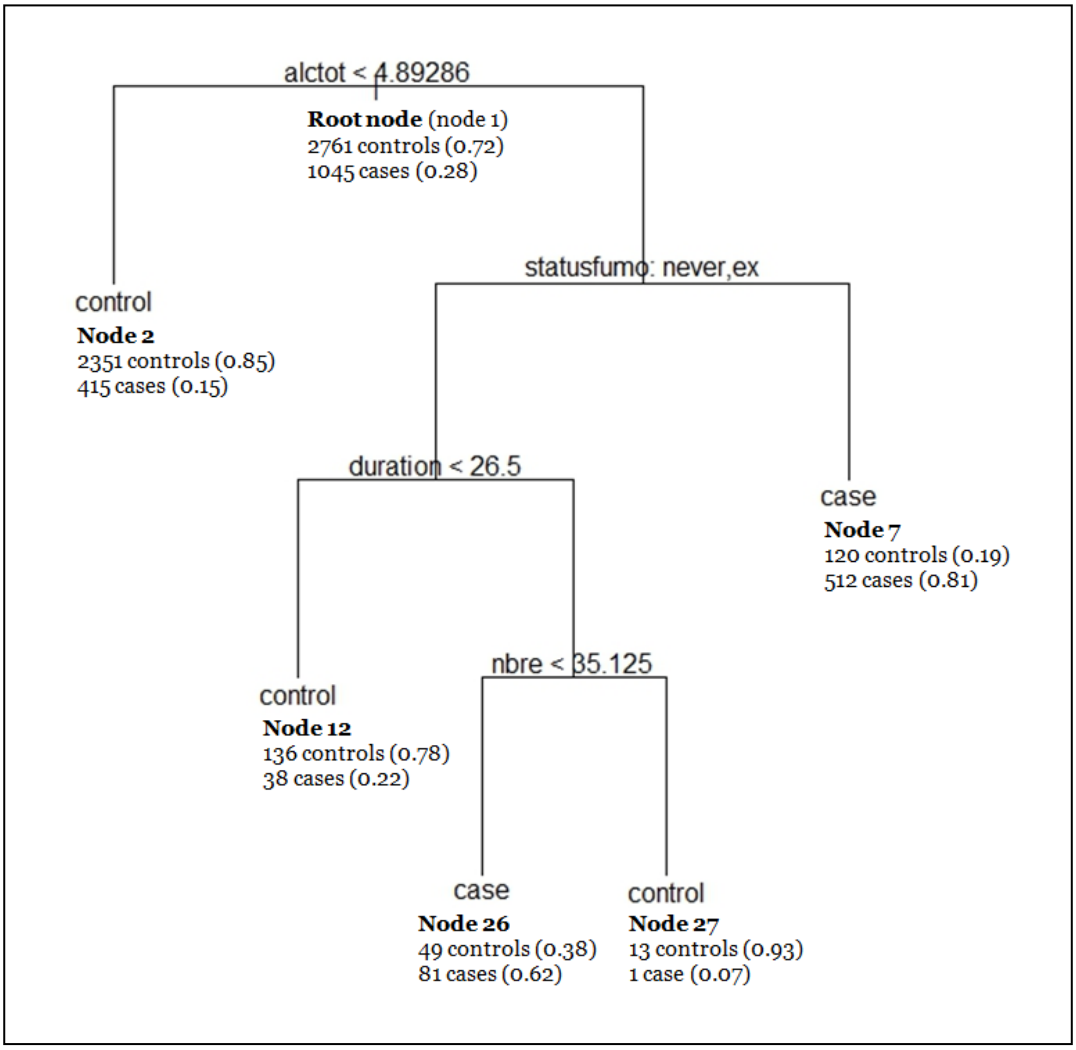

***Figure 3.4**. Classification tree with model tc3 on the learning set.*

In **Table 3.11**, the description of terminal nodes of the tree is provided. The reference category considered to compute crude ORs contains subjects drinking less than 5 drinks per week. The highest risk group is composed by subjects drinking more than 5 drinks per week and being current smokers (terminal node 7), while a low risk subgroup is composed by subjects drinking more than 5 drinks per week, never or ex-smokers having smoked for more than 27 years, consuming more than 35 portions of bread per week.

Table 3.11. *Description of terminal nodes of the classification tree tc3, and crude odds ratios (OR).*

Node	**Classification**	**Description**	**Crude ORs***
2	Control	<5 drinks/week	1
7	Case	≥5 drinks/week, current smoker	24.2
12	Control	≥5 drinks/week, never or ex-smoker, having smoked for <27 years	1.6
26	Case	≥5 drinks/week, never or ex-smoker, having smoked for ≥27 years, <35 breads/week	9.4
27	Control	≥5 drinks/week, never or ex-smoker, having smoked for ≥27 years, ≥35 bread/week	0.4

* The crude ORs for each terminal node were computed using subjects in terminal node 4 as the reference group.

Confusion matrix and accuracy measures of the predictions on the validation set are shown in **Table 3.12**. The performance of model tc3 is very similar to that of models tc1 and tc2, with a low test error, a high specificity and a discrete sensitivity. A sensitivity of 0.53 means that half of cases are correctly classified as cases by the classifiers (classification tree). The remaining portion of cases are predicted as control. Thus, the classifier fails in predicting half of the cases. Controls are almost all correctly classified as disease free subjects.

Table 3.12*. Confusion matrix and various measures of prediction accuracy on the validation set of model tc3.*

	Actual	
Predicted	**Case**	**Control**
Case	238	80
Control	210	1065

Measure	**Value**
Accuracy	0.82
Test error	0.18
Sensitivity	0.53
Specificity	0.93
Youden index	0.46

An overall comparison of the measures of the performance (accuracy, sensitivity, specificity and Youden index) of the different models on the validation set is provided in **Table 3.13**. All models had accuracy of 82%, sensitivity between 51% and 53%, specificity between 93% and 94%, and Youden index between 0.45-0.46.

Table 3.13. *Various measures of prediction accuracy on the validation set.*

	Measure			
Model	Accuracy	Sensitivity	Specificity	Youden index
Classification Tree (tc1)	0.82	0.51	0.94	0.45
Classification Tree (tc2)	0.82	0.51	0.94	0.45
Classification Tree (tc3)	0.82	0.53	0.93	0.46

3.4 Results breast cancer

Model tb1 (classification tree with rpart package and Gini index)

The classification tree built on the training dataset is represented in **Figure 3.5**. The impurity measure used in the *rpart* package is the Gini index. In the figure, the predicted class of the outcome is displayed for each terminal node, and the printed splits indicate the left branches of the tree.

Figure 3.5 represents the pruned tree, to assure no overfitting in our data. In order to prune the data, we need to find the appropriate cost-complexity parameter, that is the parameter that minimize the error. Using the 1-SE rule, in our data, the cost-complexity parameter is 0.010 and corresponds to an ideal number of two splits, and thus three terminal nodes. In this specific case the pruned tree is the same of the initial tree.

The root node includes all the 3564 subjects of the training set, 1783 cases of breast cancer and 1781 controls. The root node is therefore classified as "case" (the most frequent category), with a root node error of approximately 50%. The most important variable represented in the first split is the number of years of education (v12). This predictor, with the corresponding cut-off, produces the highest improvement in nodes purity among all the predictors of the model, using the Gini index as impurity measure. Individuals having studied less than 7.5 years in their lifetime create the terminal node 3. Those having studied more than 7.5 years create the node 2. Node 2 further splits using age as the predictor. Women with more than 34 years create the left node 4, and were classified as cases, while women with less than 34 years create the terminal node 5, and were classified as controls.

Table 3.14 shows a description of terminal subgroups, and corresponding ORs for breast cancer risk.

Table 3.14. *Description of terminal nodes of the classification tree tb1, and crude odds ratios (OR).*

Node	**Classification**	**Description**	**Crude ORs***
3	Control	<7.5 years of education	1
4	Case	≥7.5 years of education, ≥33.5 years of age	1.92
5	Control	≥7.5 years of education, <33.5 years of age	0.81

* The crude ORs for each terminal node were computed using subjects in terminal node 2 as the reference group.

Thus, the most important predictors of breast cancer, according to this model, are education and age. No dietary indicators are included among the most important predictors of breast cancer.

In order to evaluate the performance of the model, we need to fit the model on a new set of observations, the validation set, to evaluate the accuracy of the predictions. The confusion matrix for the observations in the validation set, and performance evaluation are provided in **Table 3.15**.

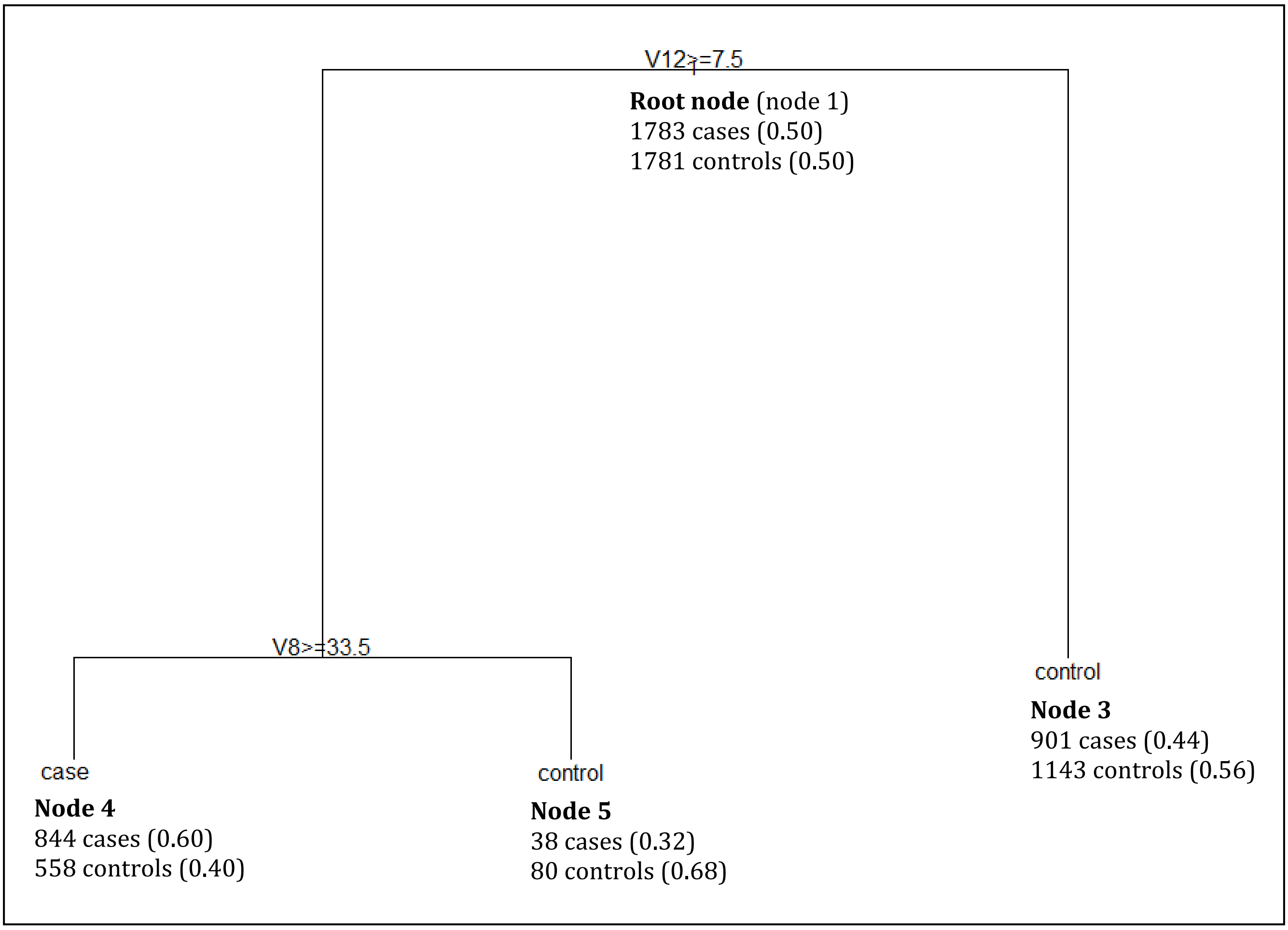

***Figure 3.5**. Classification tree tb1 on the learning set. For each terminal node, number and proportion of cases and controls, and the predicted class of the outcome are displayed. The printed splits indicate the left branches of the tree.*

***Table 3.15**. Confusion matrix and various measures of prediction accuracy on the validation set of model tb1.*

	Actual	
Predicted	**Case**	**Control**
Case	331	257
Control	434	507

Measure	**Value**
Accuracy	0.55
Test error	0.45
Sensitivity	0.57
Specificity	0.66
Youden index	0.23

Model tb1 does not have a good lever of accuracy (near 50%), and also sensitivity and specificity are near 0.50.

Model tb2 (classification tree with rpart package and Gini index, excluding socio-demographic characteristics)

I have conducted a further analysis excluding by the set of the predictors the socio-demographic variables (age and education), in order to verify whether some lifestyle or dietary habits may influence the risk of breast cancer.

Figure 3.6 represents the pruned tree, with an optimal number of four splits. The most important variable represented in the first split is smoking status. Ex-smokers create the terminal node 2. Current or never smokers create the node 3. Node 3 further splits according to the average weekly consumption of desserts. Women whose weekly consumption of desserts is <4.1 portions create the node 7, and were classified as controls. Women consuming more than 4.1 desserts per week create the child node 5, and were classified as controls. Node 5 further splits according to the consumption of fruit: women consuming less than 2 portions of fruit per day were classified as cases in node 12, while women with a daily consumption of fruit greater than 2 were classified as controls in node 13.

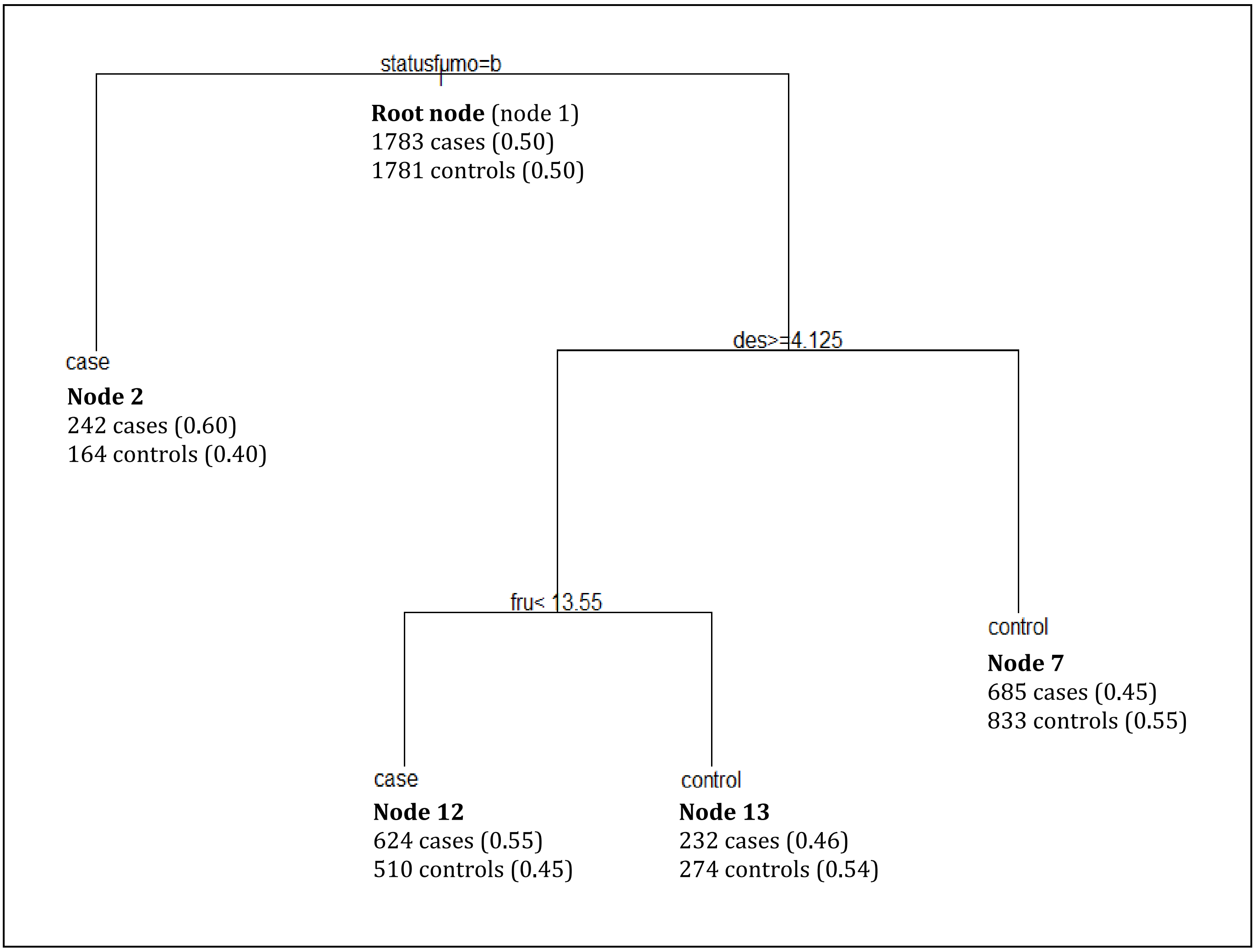

Figure 3.6. *Classification tree tb2 on the learning set.*

Table 3.16 shows a description of terminal subgroups, and corresponding ORs for breast cancer risk.

In **Table 3.17**, confusion matrix and prediction accuracy are shown.

***Table 3.16**. Description of terminal nodes of the classification tree tb2, and crude odds ratios (OR).*

Node	**Classification**	**Description**	**Crude ORs***
7	Control	Current or never smokers, desserts <4.1/week	1
2	Case	Ex-smokers	1.79
12	Case	Current or never smokers, desserts ≥4.1/week, fruit <13.5/week	1.49
13	Control	Current or never smokers, desserts ≥4.1/week, fruit ≥13.5/week	1.03

***Table 3.17**. Confusion matrix and various measures of prediction accuracy on the validation set of model tb2.*

	Actual	
Predicted	**Case**	**Control**
Case	347	313
Control	418	451

Measure	**Value**
Accuracy	0.52
Test error	0.48
Sensitivity	0.55
Specificity	0.59
Youden index	0.14
VPP	0.53
VPN	0.52

Also model tb2 does not have a good level of accuracy, sensitivity and specificity (near 50%).

3.5 Results prostatic cancer

Model tp1 (classification tree with rpart package and Gini index)

The classification tree built on the training dataset is represented in **Figure 3.7**. The optimal number of splits was set at 4.

The root node includes all the 1902 subjects of the training set, 896 cases of breast cancer and 1006 controls. The root node is therefore classified as "control". The most important variable represented in the first split is age. Men younger than 58.5 years old create the right node 3. Men having an age grater then 59 years create the left node 2. Node 2 further splits using the duration of smoking habit as the predictor. Men with a smoking duration <10.5 years (never smokers whose smoking duration is 0, or current or ex-smokers having smoked less than 10 years in their lifetime) create the terminal node 4, and are classified as cases of pancreatic cancer. Men having smoked for more than 10 years further split according to the consumption of bread. Men consuming more than 21 portions of bread per week (more than 3 per day) create the left terminal node 10 and are classified as cases, while men whose average weekly consumption of bread is lower than 21 create the right node 11 and are classified as controls.

Table 3.18 shows a description of terminal subgroups, and corresponding ORs for prostatic cancer risk. Thus, the most important predictors of prostatic cancer, according to this model, are age, smoking duration and bread consumption. No dietary indicators are included among the most important predictors of breast cancer.

The confusion matrix for the observations in the validation set, and performance evaluation are provided in **Table 3.19**.

Table 3.18*. Description of terminal nodes of the classification tree tp1, and crude odds ratios (OR).*

Node	**Classification**	**Description**	**Crude ORs***
3	Control	Age <58.5	1
4	Case	Age ≥58.5, never smokers, current or ex-smokers with smoking duration <10.5 years	3.02
10	Case	Age ≥58.5, current or ex-smokers with smoking duration≥10.5 years, bread consumption ≥21.1/week	3.42
11	Control	Age ≥58.5, current or ex-smokers with smoking duration≥10.5 years, bread consumption <21.1/week	1.48

* The crude ORs for each terminal node were computed using subjects in terminal node 3 as the reference group.

Table 3.19*. Confusion matrix and various measures of prediction accuracy on the validation set of model tp1.*

	Actual	
Predicted	**Case**	**Control**
Case	233	213
Control	152	219

Measure	**Value**
Accuracy	0.55
Test error	0.45
Sensitivity	0.61
Specificity	0.51
Youden index	0.11
VPP	0.52
VPN	0.59

Model tp1 does not have a good lever of accuracy (near 50%), and also sensitivity and specificity are near 0.50.

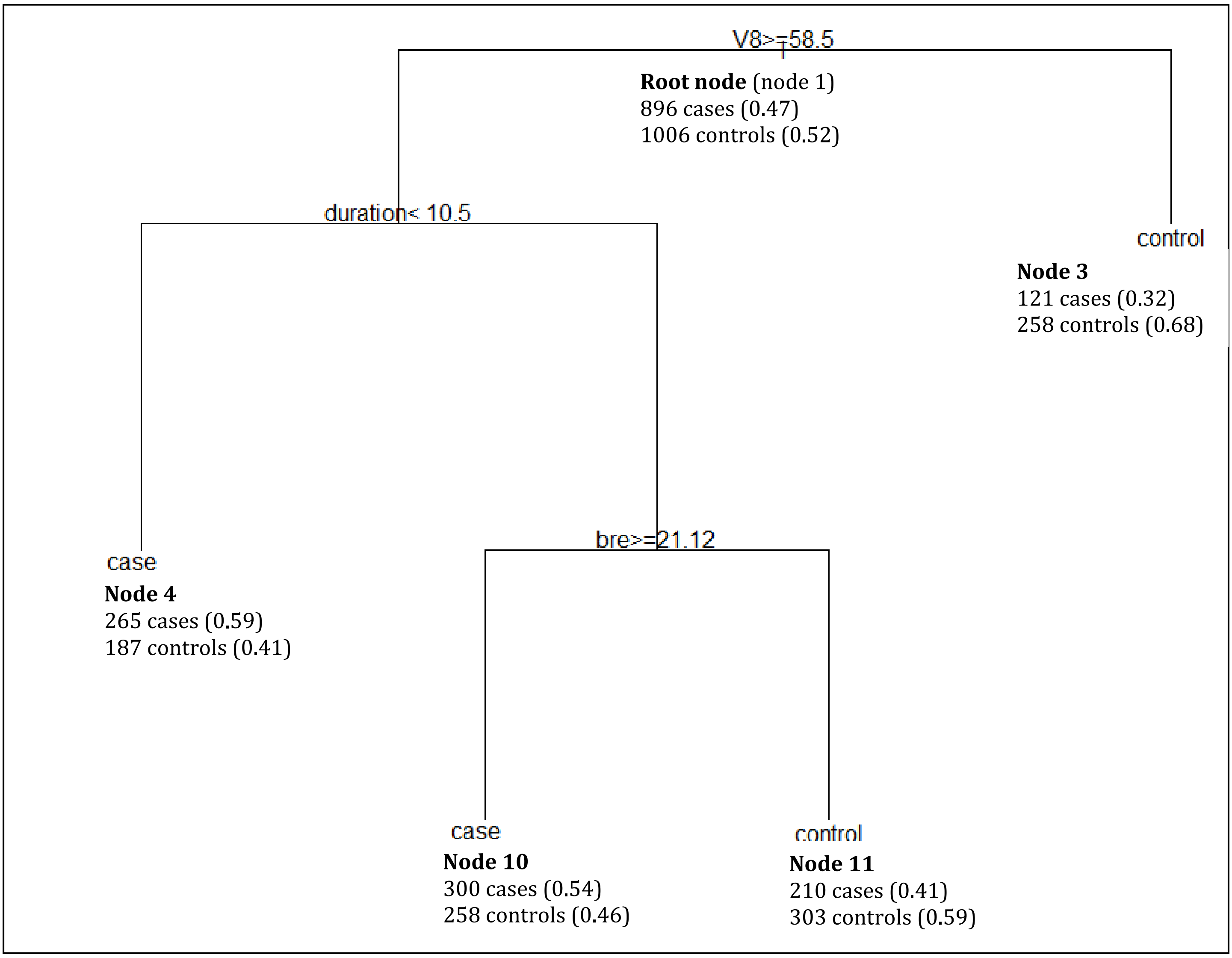

Figure 3.7. *Classification tree tp1 on the learning set.*

4. APPLICATION TO REAL DATA – ORDINAL CLASSIFICATION TREES

4.1 Set of controls

To conduct the present analyses, I used the set of controls of various case-control studies conducted in six Italian provinces (Pordenone, Milan, Genova, Forlì, Naples and Latina) between 1991 and 2008. Overall, 7750 subjects were considered in the present analysis. Controls were individuals (median age 59, range 17–82 years) with no history of cancer admitted to the same hospitals of cases for acute, non-neoplastic, conditions, unrelated to diseases or to conditions linked to the cancer in study.

One application of ordinal classification trees considered energy intake (kcal, in 4 categories using quartiles) as the ordinal outcome, and the 25 food groups shown in Table 3.1 as predictors. **Table 4.1** shows the distribution of controls divided according their intake of total energy, by centre, sex, and age.

The second application of ordinal classification trees considered red meat and processed meat intake (g/day, in 3 categories) as the ordinal outcomes, and food groups as predictors.

Table 4.1. *Distribution of 7750 controls divided according to their total energy intake (quartiles), by centre, sex, age and other selected characteristics. Italy, 1991-2008.*

Characteristics	<1820.1 kcal/day		1820.1-2253.5 kcal/day		2253.6-2776.0 kcal/day		≥2776.6 kcal/day	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
N	1938	-	1937	-	1938	-	1937	-
Centre								
Pordenone	505	26.1	716	37.0	793	40.9	919	47.4
Milan	936	48.3	676	34.9	548	28.3	422	21.8
Genova	81	4.2	129	6.7	172	8.9	251	13.0
Forlì	74	3.8	82	4.2	100	5.2	76	3.9
Naples	185	9.6	174	9.0	149	7.7	116	6.0
Latina	157	8.1	160	8.3	176	9.1	153	7.9
Sex								
Men	461	23.8	712	36.8	977	50.4	1354	69.9
Women	1477	76.2	1225	63.2	961	49.6	583	30.1
Age group (years)								
<50	361	18.6	429	22.2	487	25.1	526	27.2
50-54	230	11.9	274	14.2	273	14.1	296	15.3
55-59	312	16.1	266	13.7	329	17.0	315	16.3
60-64	335	17.3	337	17.4	307	15.8	317	16.4
65-69	384	19.8	335	17.3	286	14.8	284	14.7
≥70	316	16.3	296	15.3	256	13.2	199	10.3

4.2 Application energy intake

4.2.1 Statistical methods

Two different types of analyses were performed to evaluate the performance of classification trees methodology in predicting the category of total energy intake (kcal): single tree analysis and resampling analysis.

In both cases, I compared five different scenarios, four in the context of ordinal classification trees (i-iv) and one in the context of nominal classification trees (v). In the ordinal context, each scenario is a combination of the splitting function and the predictive performance measure. For each option it is specified the r package, the split function and the predictive performance measure:

i) rpartScore, generalized Gini impurity function with absolute misclassification cost function (*function 2.2*), misclassification cost (median score, *function 2.21*);

ii) rpartScore, generalized Gini impurity function with absolute misclassification cost (*function 2.2*), misclassification rate (modal score, *function 2.20*);

iii) rpartScore, generalized Gini impurity function with quadratic misclassification cost function (*function 2.2*), misclassification cost (median score, *function 2.21*);

iv) rpartScore, generalized Gini impurity function with quadratic misclassification cost function (*function 2.2*), misclassification rate (modal score, *function 2.20*);

v) rpart, Gini impurity function (*function 1.5*), misclassification error rate.

In single tree analysis, I used as the training set a subset composed by 70% of the observations, randomly chosen from the overall dataset (n=5422), and as the validation set a subset of 30% of the overall observations (n=2328). The training set was used to build the different classification trees, one for each different scenario, using 10-fold cross-validation and the 1-SE rule to prune the tree. The best classification trees obtained for each scenario in the learning set, were used to predict the category of the outcome of the units in the validation set. The performance of the five classification trees was evaluated on the validation set by means of two ordinal measure of association between the predicted and observed observations (Somers' d measure and Gamma statistics). In this analysis, I considered the version of the Somers' d measure with the denominator considering the observation untied in the observed score, implemented in rpartScore package (*function 2.23*). The strength of this procedure is that the performance of the classifier is evaluated on an independent dataset.

To account for sampling variability, the resampling analysis was performed. I generated 100 training (70% of the data) and evaluated model performance in 100 validation sets (30% of the data). Thus, 100 different pairs of training and validation sets have been considered, resulting in a total of 500 values of Somers' d (100 for each scenario). The global hypothesis of no difference in the agreement of various models was tested using Friedman's non-parametric rank test for repeated measurements in a randomized complete block design (Hollander et al 2014), treating each of the 100 training sets as a block. The test statistics and the corresponding asymptotic p-values were calculated using the *friedman_test* function in the *coin* R package (Horton et al 2008).

4.2.2 Single tree analysis

Results for the five scenarios are presented in **Figures 4.1-4.5**. In the left panels there are the relative cross-validated errors for various classification trees – cross validated total number of misclassifications for scenarios ii), iv) and v) and cross validated total misclassification costs for trees i) and ii). The 1-SE rule threshold used to choose the optimal size (number of splits/terminal nodes) of each classification tree is also shown in the left figures. In the right panel there are the illustrations of optimal classification trees for each scenario. In each classification tree, the predicted class in each terminal node according to specific predictive performance measure (modal or median value) is specified.

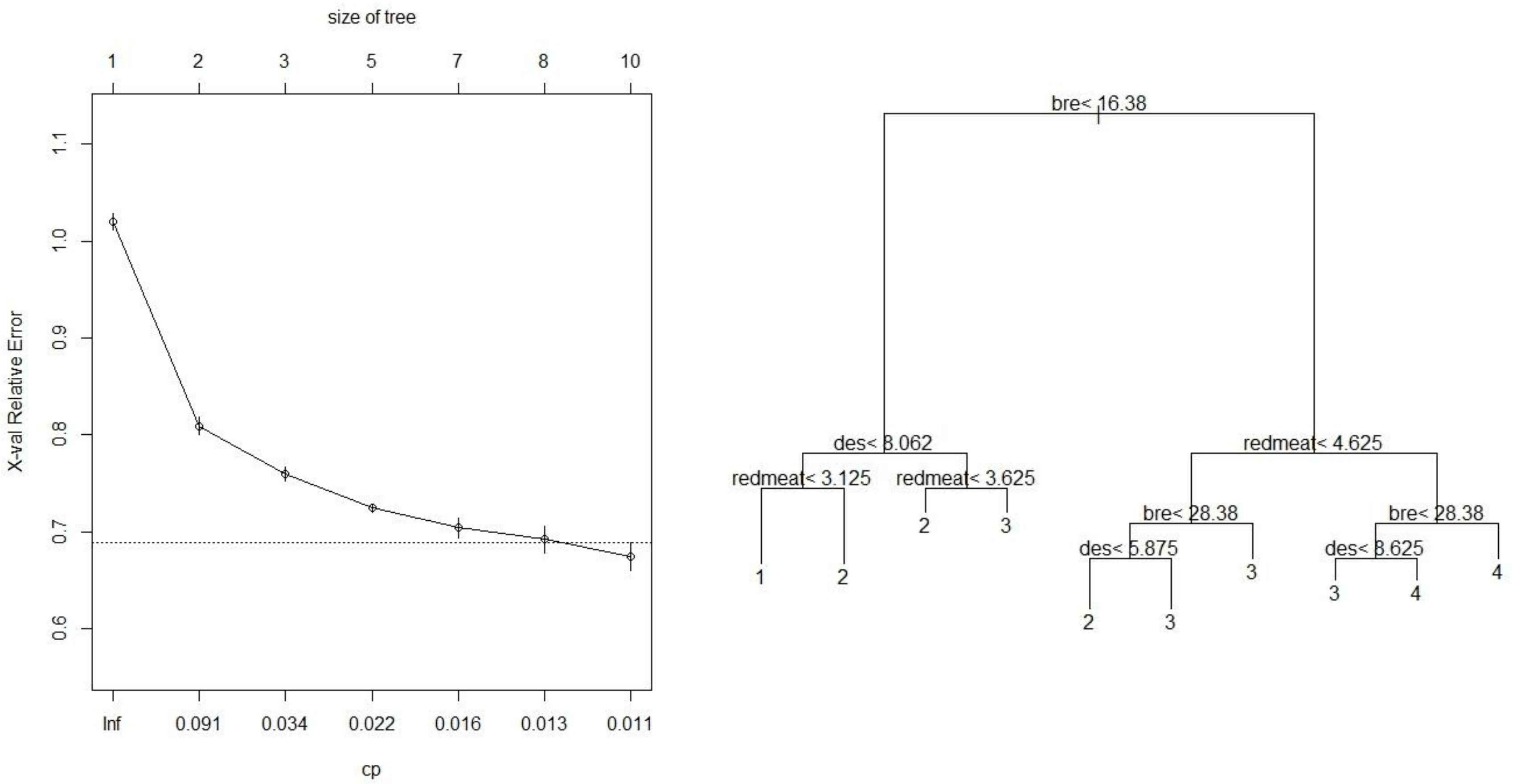

***Figure 4.1**. Results of scenario i).* $i_{GG1}(t)$ *split function and* $R_{mc}(T)$ *predictive performance measure.*

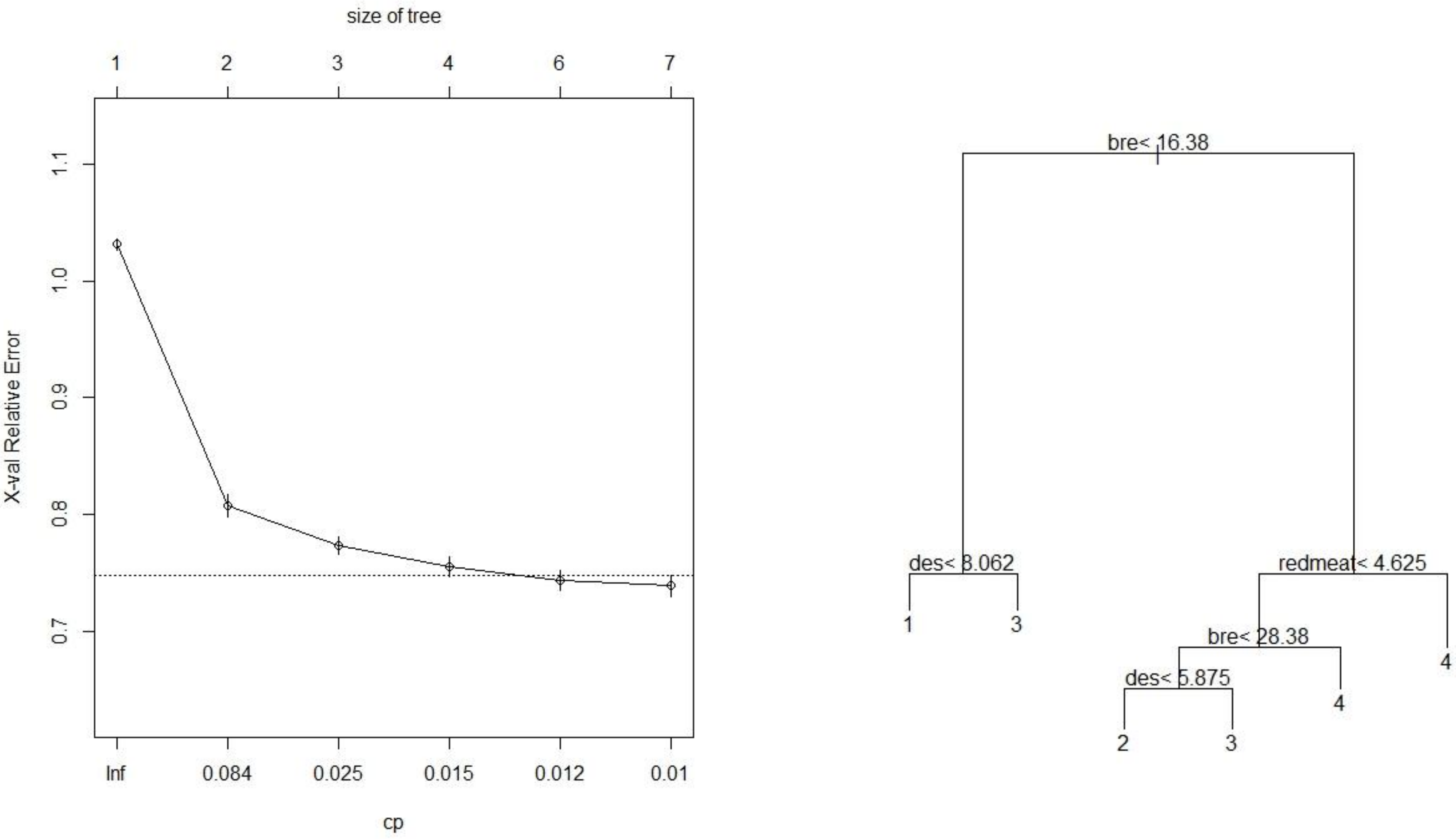

Figure 4.2. *Results of scenario ii).* $i_{GG1}(t)$ *split function and* $R_{mr}(T)$ *predictive performance measure.*

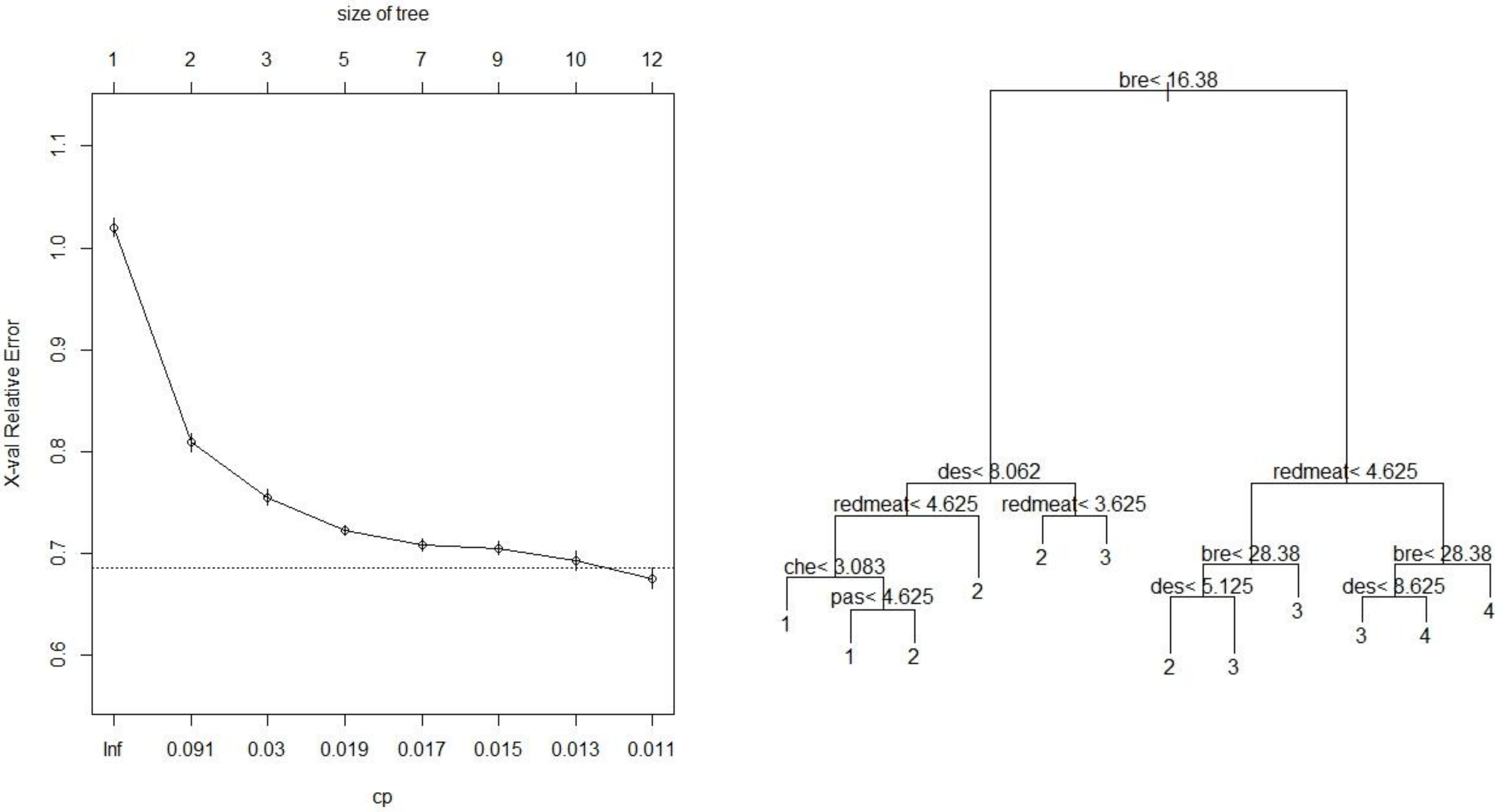

***Figure 4.3**. Results of scenario iii). $i_{GG2}(t)$ split function and $R_{mc}(T)$ predictive performance measure.*

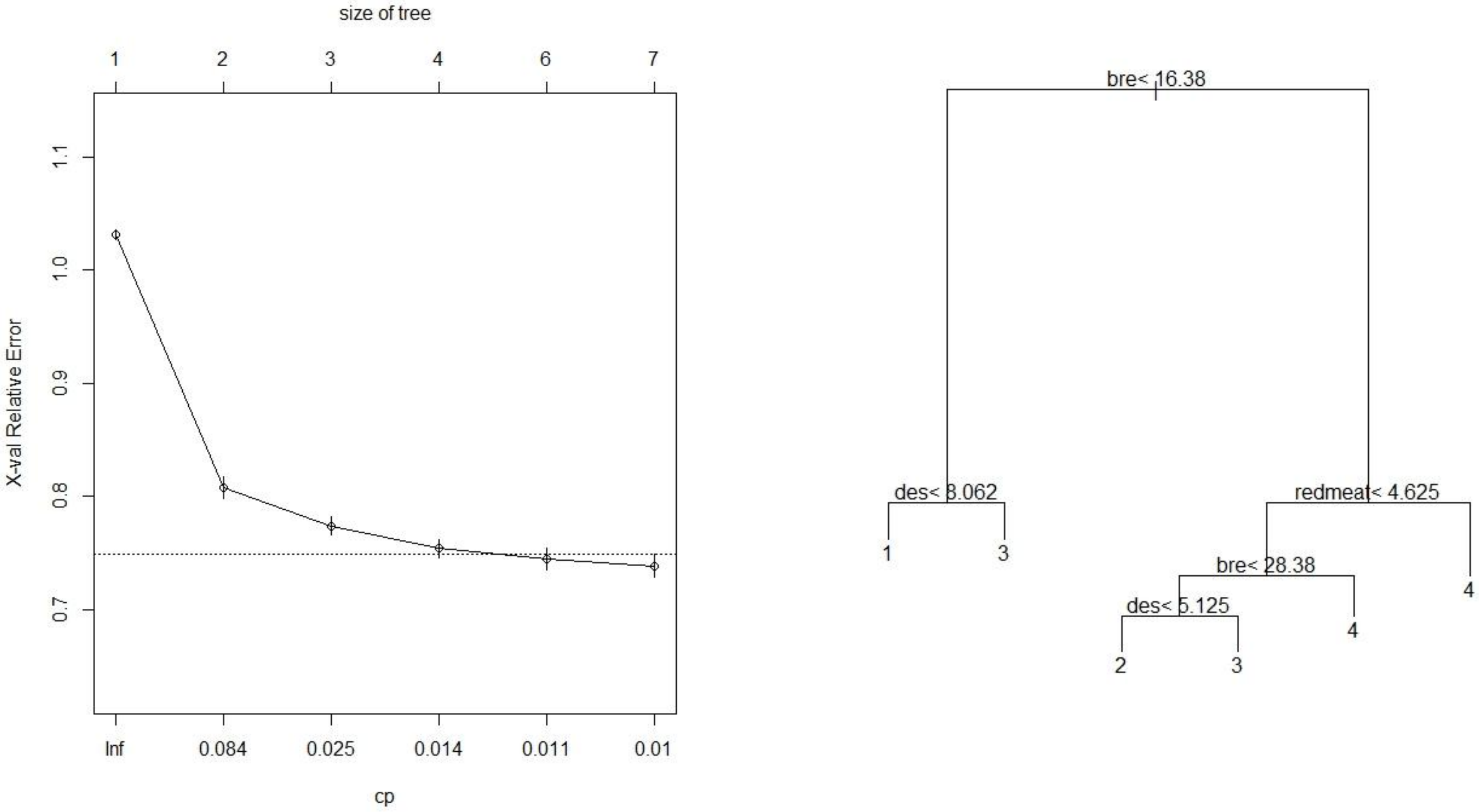

Figure 4.4. *Results of scenario iv). $i_{GG2}(t)$ split function and $R_{mr}(T)$ predictive performance measure.*

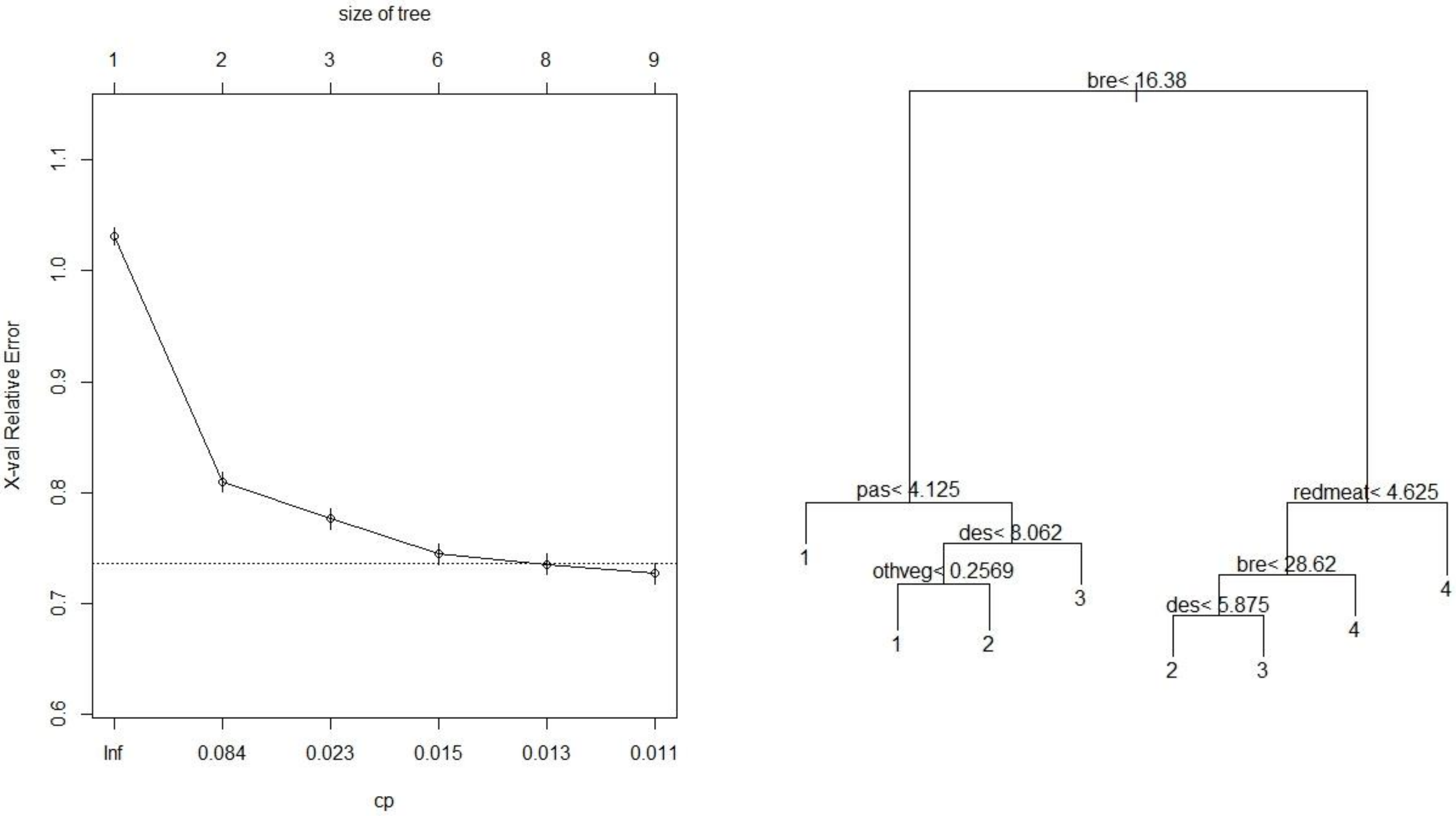

Figure 4.5*. Results of scenario v). Nominal classification tree.*

The most important predictor for energy intake appears to be bread consumption. Indeed, this predictor results as the first split in each of the five scenarios, with a threshold of 16.4 portions/week. Other predictors common to all the five scenarios are desserts and red meat intake. The ordinal classification tree obtained using a quadratic misclassification cost and the misclassification cost as the predictive performance measure (scenario iii, **Figure 4.3**) includes as important predictors also cheese and pasta consumption, while the nominal classification tree (scenario v, **Figure 4.5**) detects other vegetable consumption as an important predictor for energy intake.

Tables 4.2-4.6 provides the description of terminal nodes for the five scenarios, from the lowest to the highest predicted class of energy intake (approximated values for cut-off of predictors).

***Table 4.2**. Description of terminal nodes of the ordinal classification tree in scenario i.*

Node	Predicted Class	Description
8	1	Bread consumption <16/week, desserts <8/week, red meat <3/week
9	2	Bread consumption <16/week, desserts <8/week, red meat ≥3/week
10	2	Bread consumption <16/week, desserts ≥8/week, red meat <3.5/week
24	2	Bread consumption ≥16/week, red meat <4.5/week, bread<28/week, desserts <6/week
11	3	Bread consumption <16/week, desserts ≥8/week, red meat ≥3.5/week
25	3	Bread consumption ≥16/week, red meat <4.5/week, bread<28/week, desserts ≥6/week
13	3	Bread consumption ≥16/week, red meat <4.5/week, bread≥28/week
28	3	Bread consumption ≥16/week, red meat ≥4.5/week, bread<28/week, desserts <8/week
29	4	Bread consumption ≥16/week, red meat ≥4.5/week, bread<28/week, desserts ≥8/week
15	4	Bread consumption ≥16/week, red meat ≥4.5/week, bread≥28/week

***Table 4.3**. Description of terminal nodes of the ordinal classification tree in scenario ii.*

Node	Predicted Class	Description
4	1	Bread consumption <16/week, desserts <8/week
24	2	Bread consumption ≥16/week, red meat <4.5/week, bread <28/week, desserts <6/week
5	3	Bread consumption <16/week, desserts ≥8/week
25	3	Bread consumption ≥16/week, red meat <4.5/week, bread <28/week, desserts ≥6/week
13	4	Bread consumption ≥16/week, red meat <4.5/week, bread ≥28/week
7	4	Bread consumption ≥16/week, red meat ≥4.5/week

***Table 4.4**. Description of terminal nodes of the ordinal classification tree in scenario iii.*

Node	Predicted Class	Description
16	1	Bread consumption <16/week, desserts <8/week, red meat <4.5/week, cheese <3/week
34	1	Bread consumption <16/week, desserts <8/week, red meat <4.5/week, cheese ≥3/week, pasta <4.5/week
9	2	Bread consumption <16/week, desserts <8/week, red meat ≥4.5/week
35	2	Bread consumption <16/week, desserts <8/week, red meat <4.5/week, cheese ≥3/week, pasta ≥4.5/week
10	2	Bread consumption <16/week, desserts ≥8/week, red meat <3.5/week
24	2	Bread consumption ≥16/week, red meat <4.5/week, bread <28/week, desserts <5/week
11	3	Bread consumption <16/week, desserts ≥8/week, red meat ≥3.5/week
25	3	Bread consumption ≥16/week, red meat <4.5/week, bread <28/week, desserts ≥5/week
13	3	Bread consumption ≥16/week, red meat <4.5/week, bread ≥28/week
28	3	Bread consumption ≥16/week, red meat ≥4.5/week, bread <28/week, desserts <3.5/week
29	4	Bread consumption ≥16/week, red meat ≥4.5/week, bread <28/week, desserts ≥3.5/week
15	4	Bread consumption ≥16/week, red meat ≥4.5/week, bread ≥28/week

Table 4.5*. Description of terminal nodes of the ordinal classification tree in scenario iv.*

Node	**Predicted Class**	**Description**
4	1	Bread consumption <16/week, desserts <8/week
24	2	Bread consumption ≥16/week, red meat <4.5/week, bread <28/week, desserts <5/week
5	3	Bread consumption <16/week, desserts ≥8/week
25	3	Bread consumption ≥16/week, red meat <4.5/week, bread <28/week, desserts ≥5/week
13	4	Bread consumption ≥16/week, red meat <4.5/week, bread ≥28/week
7	4	Bread consumption ≥16/week, red meat ≥4.5/week

Table 4.6. *Description of terminal nodes of the nominal classification tree in scenario v.*

Node	**Predicted Class**	**Description**
4	1	Bread consumption <16/week, pasta <4/week
20	1	Bread consumption <16/week, pasta ≥4/week, desserts <8/week, other vegetables <0.3/week
21	2	Bread consumption <16/week, pasta ≥4/week, desserts <8/week, other vegetables ≥0.3/week
24	2	Bread consumption ≥16/week, red meat <4.5/week, bread <28/week, desserts<6/week
11	3	Bread consumption <16/week, pasta ≥4/week, desserts ≥8/week
25	3	Bread consumption ≥16/week, red meat <4.5/week, bread <28/week, desserts≥6/week
13	4	Bread consumption ≥16/week, red meat <4.5/week, bread ≥28/week
7	4	Bread consumption ≥16/week, red meat ≥4.5/week

Important predictors appear in the tree with multiple splits. I will describe scenario i as an example. The first case is the weekly intake of red meat. This predictor appears as important both in the left and in the right node built with the split on desserts (left branch). Here it is evident the "interaction" between desserts and red meat intake. In fact individuals with a "low" consumption of desserts and a "high" consumption of red meat are predicted in the same class of individuals with a "high" consumption of desserts and a "low" consumption of red meat. The second case is bread which appears in two consecutively splits in the right branch of the tree.

The change in the predictive performance – from misclassification cost to misclassification rate – in our example, leads to the selection of simpler optimal trees, passing from 10 to 7 terminal nodes in the case of an absolute misclassification cost, and from 12 to 6 in the case of a quadratic misclassification cost.

In this application, the choice of the splitting function has little effect on the tree topology, when comparing trees using misclassification rate as predictive performance measure. In fact, trees of scenario ii and iv differ only in the threshold used to split desserts in the right branch of the trees: 5.875 in tree of scenario ii and 5.125 in tree of scenario iv. Similarly, also trees obtained using misclassification cost to prune the tree are similar. The only difference is the presence of two additional splits in tree in scenario i as compared to the tree in scenario iii.

The difference in the predictive performance measures is visible in the predicted class of the same decision process. For example, individuals consuming more than 16 portions of bread per week, less than 4.5 portions of red meat per week, and more than 28 portions of bread per week have a predicted outcome class of 3 considering the misclassification cost, and thus a median score (scenario i and iii), and a class of 4 using the misclassification rate, and thus the modal score (scenario ii, iv and v).

Table 4.7 shows the predictive performance of the 5 different classification trees using two measures of ordinal association, in the single tree analysis.

***Table 4.7**. Single tree analysis. Somers' d and gamma values based on the validation set comparing the observed and the predicted estimated form the five scenarios of classification trees.*

Model	Somers' d measure	Gamma statistics
i) rpartScore $\mathrm{i}_{\mathrm{GG1}}(\mathrm{t})$ and $\mathrm{R}_{\mathrm{mc}}(\mathrm{T})$	0.527	0.714
ii) rpartScore $\mathrm{i}_{\mathrm{GG1}}(\mathrm{t})$ and $\mathrm{R}_{\mathrm{mr}}(\mathrm{T})$	0.489	0.651
iii) rpartScore $\mathrm{i}_{\mathrm{GG2}}(\mathrm{t})$ and $\mathrm{R}_{\mathrm{mc}}(\mathrm{T})$	0.534	0.717
iv) rpartScore $\mathrm{i}_{\mathrm{GG2}}(\mathrm{t})$ and $\mathrm{R}_{\mathrm{mr}}(\mathrm{T})$	0.489	0.652
v) rpart	0.514	0.661

In this single tree analysis, the best agreement for the prediction of energy intake, according to the values of Somers'd measure, was observed using an ordinal classification tree (rpartScore) with a quadratic misclassification cost and misclassification cost (median value) as the predictive performance measure (scenario iii), with scenario i and scenario v only slightly lower. The values of Somers' d measure ranged between 0.489 and 0.534, representing an intermediate agreement between the observed and the predicted values (maximum value is 1). The right column reported the valued of another measure of ordinal association, the Gamma statistics. These values are slightly higher as compared to those of Somers' d ones, ranging between 0.717 and 0.651. Also using this measure of association, the higher agreement was obtained in scenario iii, followed by scenario i. In general, the ranking of the scenarios according to their predictive performance is the same using the two ordinal measures of association.

The confusion matrix for the observations in the validation set (n=2328), for each of the five scenarios are provided in **Tables 4.8-4.12**.

Table 4.8. *Single tree analysis. Confusion matrix on the validation set of scenario i.*

	Observed			
Predicted	**I quartile**	**II quartile**	**III quartile**	**IV quartile**
I quartile	279	102	26	8
II quartile	257	303	209	104
III quartile	43	160	273	278
IV quartile	4	16	75	191

Table 4.9. *Single tree analysis. Confusion matrix on the validation set of scenario ii.*

	Observed			
Predicted	**I quartile**	**II quartile**	**III quartile**	**IV quartile**
I quartile	435	250	126	57
II quartile	64	116	65	30
III quartile	63	116	162	126
IV quartile	21	99	230	368

Table 4.10. *Single tree analysis. Confusion matrix on the validation set scenario iii.*

	Observed			
Predicted	**I quartile**	**II quartile**	**III quartile**	**IV quartile**
I quartile	301	104	28	9
II quartile	233	295	201	99
III quartile	45	166	279	282
IV quartile	4	16	75	191

Table 4.11. *Single tree analysis. Confusion matrix on the validation set scenario iv.*

	Observed			
Predicted	**I quartile**	**II quartile**	**III quartile**	**IV quartile**
I quartile	435	250	126	57
II quartile	62	110	59	26
III quartile	65	122	168	130
IV quartile	21	99	230	368

Table 4.12. *Single tree analysis. Confusion matrix on the validation set scenario v.*

	Observed			
Predicted	**I quartile**	**II quartile**	**III quartile**	**IV quartile**
I quartile	361	150	70	30
II quartile	164	235	150	72
III quartile	39	99	142	117
IV quartile	19	97	221	362

4.2.3 Resampling analysis

In the resampling analysis, whose details are provided in *Section 4.2*, similar patterns to the single tree analysis were observed in the mean and median Somers' d values comparing observed and predicted classes of energy intake (**Table 4.13** and **Table 4.14**). Using Somers'd measure as ordinal association index, scenario iii (rparScore, quadratic cost, misclassification cost) had the highest performance measure (mean=0.506 and median=0.508), followed by scenario i (rparScore, absolute cost, misclassification cost) that has a slightly lower performance (mean and median=0.505). They are followed by the nominal classification tree (mean=0.480 and median=0.478), and then by scenario ii (mean and median=0.470) and scenario iv (mean=0.465 and median=0.464).

Table 4.13*. Resampling analysis (100 trees). Somers' d values based on the validation set comparing the observed and the predicted class estimated form the five scenarios of classification trees.*

	Somers' d measure	
Model	**Mean (variance)**	**Median**
i) rpartScore $i_{GG1}(t)$ and $R_{mc}(T)$	0.505 (0.0004)	0.505
ii) rpartScore $i_{GG1}(t)$ and $R_{mr}(T)$	0.470 (0.0005)	0.470
iii) rpartScore $i_{GG2}(t)$ and $R_{mc}(T)$	0.506 (0.0004)	0.508
iv) rpartScore $i_{GG2}(t)$ and $R_{mr}(T)$	0.465 (0.0006)	0.464
v) rpart	0.480 (0.0004)	0.478

***Table 4.14**. Resampling analysis (100 trees). Somers' d values based on the validation set comparing the observed and the predicted class estimated form the five scenarios of classification trees.*

	Gamma	
Model	**Mean (variance)**	**Median**
i) rpartScore $\mathrm{i_{GG1}(t)}$ and $\mathrm{R_{mc}(T)}$	0.681 (0.0005)	0.681
ii) rpartScore $\mathrm{i_{GG1}(t)}$ and $\mathrm{R_{mr}(T)}$	0.645 (0.0005)	0.647
iii) rpartScore $\mathrm{i_{GG2}(t)}$ and $\mathrm{R_{mc}(T)}$	0.681 (0.0006)	0.682
iv) rpartScore $\mathrm{i_{GG2}(t)}$ and $\mathrm{R_{mr}(T)}$	0.648 (0.0005)	0.651
v) rpart	0.650 (0.0004)	0.647

Accordingly, using Gamma statistics measures, the highest performance was observed in scenario iii and scenario i (mean=0.681 and median=0.682 and 0.681, respectively). Nominal model, scenario ii and scenario iv have a lower performance, with their median values of gamma ranging between 0.651 and 0.647.

The above patterns are more clearly shown in the distributions of the Somers' d and Gamma metrics presented in **Figure 4.6** and **Figure 4.7**. For ordinal classification trees built with rpartScore using the misclassification cost as the predictive performance measure, both of the trees (with absolute or quadratic misclassification cost) have the highest performances. Nominal classification tree built with rpart had an intermediate predictive performance, while the two ordinal trees using the misclassification rate has the lowest predictive performance.

Friedman's test rejected the global equality hypothesis across the five models ($p<0.001$), both considering Somers'd and Gamma values. Thus, there is at least one scenario providing a higher predictive performance (higher median of the Somers' d or Gamma measure). This may be the model built with an ordinal classification tree methodology using the median as the predictive value.

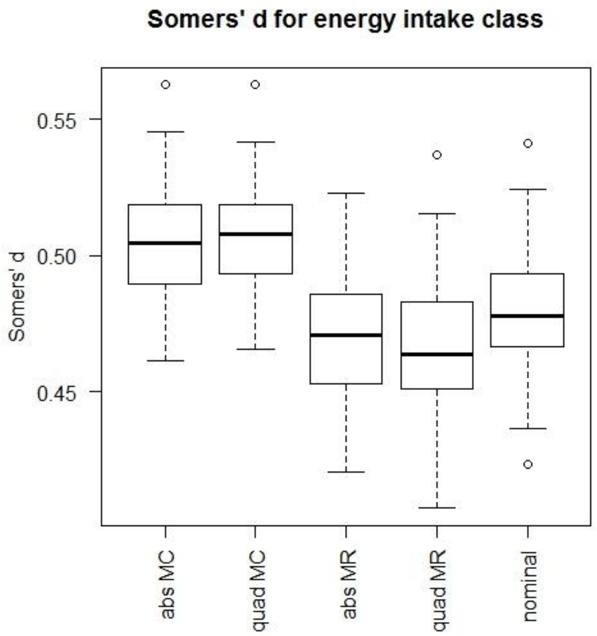

Figure 4.6 *Resampling analysis. Distribution of Somers' d comparing observed and predicted class of energy intake from the five scenarios of classification trees*

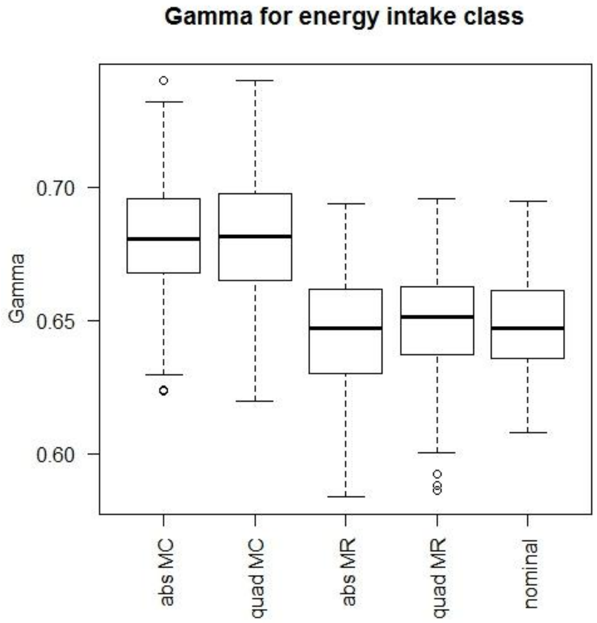

Figure 4.7 *Resampling analysis. Distribution of Gamma comparing observed and predicted class of energy intake from the five scenarios of classification trees*

4.3 Application red and processed meat

4.3.1 Statistical methods

Single tree analyses were performed to build classification trees to investigate whether red and processed meat consumption was associated with some specific dietary pattern. Response variables were red meat consumption (3 categories, <50 g/day, 50-99 g/day, ≥100 g/day), and processed meat intake (3 categories, <25 g/day, 25-49 g/day, ≥50 g/day). These variables were computed considering both the frequency (number of portions per week) and the portion. Small portions were 1/3 smaller than the mean portion and big portions were 1/3 larger than the mean one.

Since these 2 outcomes are categorical ordinal variable, I performed ordinal classification trees using *rpart* package. Generalized Gini impurity function with quadratic misclassification cost as split function and median score as predictive performance measure was used to build the tree. This combination, in fact, resulted be the best one according to predictive performance evaluation in comparison to the other scenarios proposed in the previous paragraph (*Section 4.2*)

I used as the training set a subset composed by 70% of the observations, randomly chosen from the overall dataset, and as the validation set a subset of 30% of the overall observations. The training set was used to build the classification tree using 10-fold cross-validation and the 1-SE rule to prune the tree. Only individuals with no missing outcome were considered. Thus, 7744 subjects were considered in the analysis of red meat (5419 in the training set and 2325 in the validation set), and 7702 in the analysis of processed meat (5390 in the training set and 2312 in the validation set).

A small change in food groups was made as compared to those proposed in Table 3.1. Total fruit intake was computed as the sum of citrus fruit and other fruits intake, total vegetable as the sum of leafy vegetables, fruiting vegetables, root vegetables, cruciferous vegetables and other vegetables, and total sweet was computed as the sum of soft drinks and fruit juices, desserts, and sugar and candies.

4.3.2 Results

Figure 4.8 shows the ordinal classification tree built for red meat as the response variable.

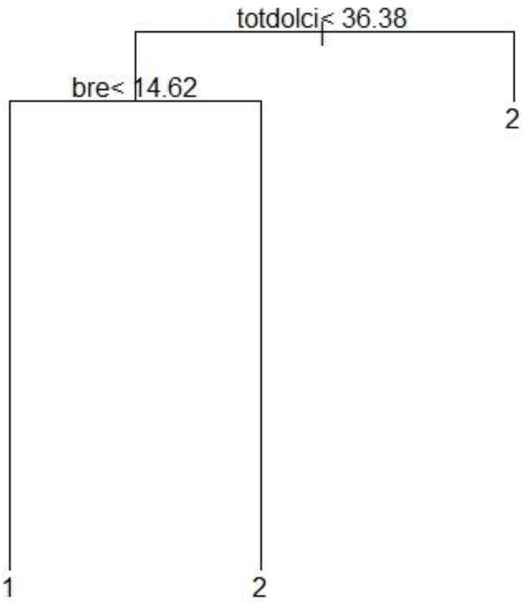

Figure 4.8*. Classification tree for red meat on the learning set.*

Important predictors for red meat consumption were total intake of sweets (first split) and bread consumption. Subjects consuming less than 36 portions/week of sweets and less than 2 portions of bread per day were classified as low (<50 g/day) consumption of red meat. The other categories were classified as intermediate (50-99 g/day) consumption. This tree does not allow to identify high (≥100 g/day) consumers of red meat.

Figure 4.9 shows the corresponding classification tree for the consumption of processed meat. This figure shows that important predictors of processed meat intake are the consumption of eggs, of bread and of sweets. According to this model, subjects eating less than 1 egg per week and less than 2 portions of bread per day have a small consumption (<25 g/day) of processed meat. On the other hand, individuals eating more than 1 egg per week and more than 30 portions of sweets per week have a great (≥50 g/day) consumption of processed meat. The other categories are predicted as intermediate (25-49 g/day) intake of processed meat.

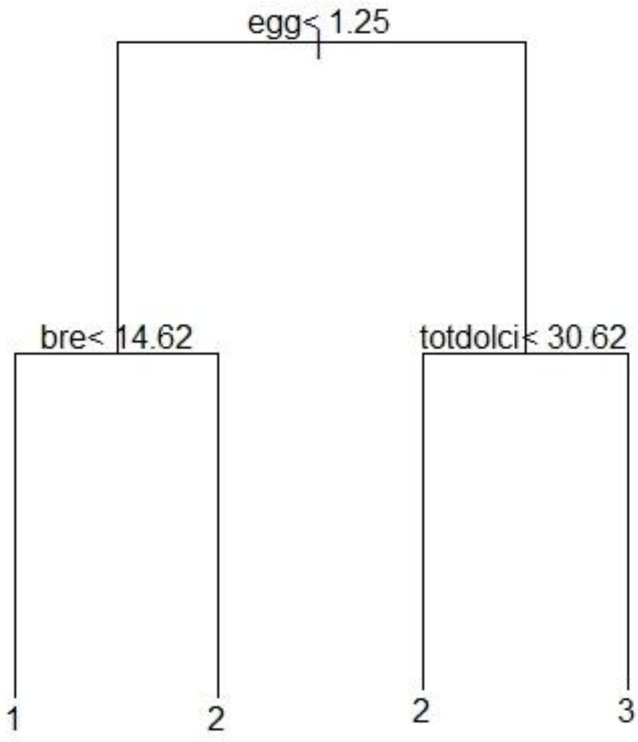

Figure 4.9. *Classification tree for processed meat on the learning set.*

We performed other two models in order to verify whether the intake of red meat and of processed meat were correlates, which means that individuals eating red great quantities of red meat eat also processed meat and vice-versa. Classification trees of these analyses are shown in **Figure 4.10** and **Figure 4.11**, respectively.

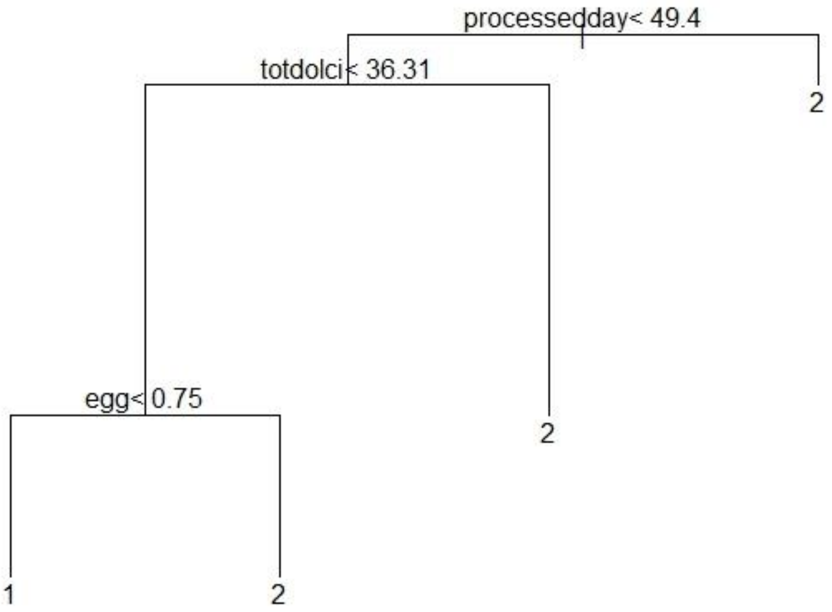

Figure 4.10. *Classification tree for red meat on the learning set, with processed meat as predictor.*

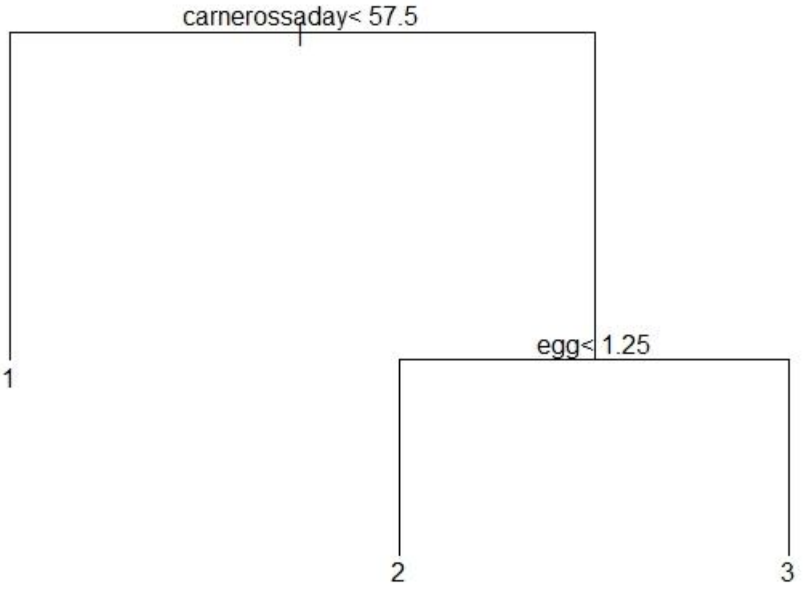

Figure 4.11. *Classification tree for processed meat on the learning set, with red meat as predictor.*

From these figures, it is evident that the intake of red meat and processed meat are related. In fact, processed meat intake is the first split when considering red meat as the response variable (**Figure 4.10**). Other important predictors for red meat are sweets and eggs. It emerged that individuals eating <50g/day of processed meat, <36 sweets/week, and <0.75 eggs/week are classified as low red meat consumers. Other categories predict intermediate consumers of red meat, while high consumers of red meat do not appear in this figure. Accordingly, also processed meat intake depends on red meat intake (**Figure 4.11**). Individuals eating <57 g/day of red meat are classified as eating <25 g/day of processed meat, those consuming more than 57 g/day of red meat and less than 1 egg per week are classified as intermediate consumers of processed meat, and those eating more than 1 egg per week are predicted eating more than 50 g/day of processed meat.

CONCLUSIONS

Ordinal classification trees were used to investigate the association between total energy intake, red meat and processed meat consumption and different food groups. The aim of the application on total energy intake was to compare the predictive ability of various classification tree methods to assess ordinal estimates on ordinal categorical outcome. Five different scenarios were compared, four ordinal classification tree and one nominal classification tree. The ordinal classification trees built with median value to predict within node class of the outcome had a better predictive performance, with that with quadratic misclassification cost slightly better than that with absolute cost. This findings were consistent both in the single-tree and in the resampling analysis. The use of nominal classification trees in case of ordinal categorical outcomes provides reasonably robust results according to misclassification of individuals. In fact, its predictive performance was similar to that obtained using ordinal classification trees with misclassification error rate as predictive performance measure. However, our findings also suggest that researchers should consider using classification tree models specifically designed for ordinal outcomes when feasible, particularly in situations where the predictive performance of nominal trees was less successful. These findings are in broad agreement with those provided by Wheeler and colleagues (Wheeler et al 2015). Moreover, also Galimberti, who proposed rpartScore package and to use median valued to predict the category of the outcome within a node, found higher predictive performance of classification trees built using misclassification cost rather than misclassification error rate as predictive performance measure (Galimberti et al 2012). However, in the application of classification trees in the prediction of total energy intake, no method had perfect predictive performance.

With the application on red meat and processed meat intake, we were able to investigate whether some dietary habits were common to those eating large quantities of red meat (more than 100 g per day) or processed meat (more than 50 g per day). The findings of these analyses showed that those eating eggs and large portions of sweets were classified by the ordinal classification tree as eaters of processed meat. Moreover, classification trees highlighted the evidence that the consumption of red meat and processed meat were strongly related.

Possible future researches should try to take advantage of findings obtained with classification tree methodologies in order to investigate the relationship between red meat and processed meat intake and the risk of colorectal cancer and the risk of other neoplasms. Moreover, the application of recursive partitioning techniques in predictive settings, including data on cancer screening, may be of interest for future researches.

REFERENCES

Agresti A (2002). *Categorical Data Analysis*. Wiley: Gainesville, Florida.

Archer KJ (2010). rpartOrdinal: An R Package for Deriving a Classification Tree for Predicting an Ordinal Response. *J Stat Softw* **34:** 7.

Biesbroek S, van der AD, Brosens MC, Beulens JW, Verschuren WM, van der Schouw YT *et al* (2015). Identifying cardiovascular risk factor-related dietary patterns with reduced rank regression and random forest in the EPIC-NL cohort. *Am J Clin Nutr* **102:** 146-154.

Breiman L, Friedman J, Stone CJ, Olshen RA (1984). *Classification And Regression Trees*.

Decarli A, Franceschi S, Ferraroni M, Gnagnarella P, Parpinel MT, La Vecchia C *et al* (1996). Validation of a food-frequency questionnaire to assess dietary intakes in cancer studies in Italy. Results for specific nutrients. *Ann Epidemiol* **6:** 110-118.

Franceschi S, Negri E, Salvini S, Decarli A, Ferraroni M, Filiberti R *et al* (1993). Reproducibility of an Italian food frequency questionnaire for cancer studies: results for specific food items. *Eur J Cancer* **29A:** 2298-2305.

Galimberti G, Soffritti G, Di Maso M (2012). Classification Trees for Ordinal Responses in R: The rpartScore Package. *Journal of Statistical Software* **47**.

Hollander M, Wolfe DA, Chicken E (2014). *Nonparametric Statistical Methods*. Wiley.

Horton T, Hornik K, van de Wiel MA, Zeileis A (2008). Implementing a Class of Permutation Tests: The coin Package. *Journal of Statistical Software* **28**.

Navarro Silvera SA, Mayne ST, Gammon MD, Vaughan TL, Chow WH, Dubin JA *et al* (2014). Diet and lifestyle factors and risk of subtypes of esophageal and gastric cancers: classification tree analysis. *Ann Epidemiol* **24:** 50-57.

Piccarreta R (2001). A new measure of nominal-ordinal association. *Journal of Applied Statistics* **28:** 107-120.

Piccarreta R (2008). Classification trees for ordinal variables. *Computational Statistics* **23:** 407-427.

Somers RH (1962). A New Asymmetric Measure of Association for Ordinal Variables. *American Sociological Review* **27:** 799-811.

Wheeler DC, Archer KJ, Burstyn I, Yu K, Stewart PA, Colt JS *et al* (2015). Comparison of ordinal and nominal classification trees to predict ordinal expert-based occupational exposure estimates in a case-control study. *Ann Occup Hyg* **59:** 324-335.