UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Classification methods applied to the diagnosis of Obstructive Sleep Apnea - a comparative study

Ana Margarida Rodrigues dos Santos Martins Pereira

Mestrado em Bioestatística

Dissertação orientada por:
Marília Cristina de Sousa Antunes
Vukosava Milic Torres

2021

Dedication

Dedicated to me and those who believed and helped me throughout this journey.
Determination, strength and integrity.

Also,

In loving memory of my grandpa

> ”We’ve got years ahead of us
>
> We’ve got people who care for us
>
> Yeah, we’ve got Sunday morning coffees in the sun”

Nothing Can Stop Us Now, Tim Minchin

Acknowledgements

To patients that voluntarily collaborated in this study. This Master thesis is partially supported by Harvard Medical School-Portugal Program Project HMSP-ICJ/0022/201: “Obstructive sleep apnea and associated metabolic/cardiovascular disorders: understanding mechanisms towards early diagnosis and prognosis prediction’ ’.

Abstract

Humans are innate pattern recognizers and we make use of such for important daily and future decisions affecting the society and surrounding environment. Recent technological development dated since the middle of the 20^{th} century made available booms of information that can better be acknowledge and useful for decision making if wisely used.

The purpose is to study patterns aimed for the classification of a specific outcome variable. The two main types of methodologies differentiated are the unsupervised and supervised, differing on the acknowledgement of the true classification, regarding the outcome, that is absent in the first type (pattern search) but present in the second type (pattern prediction). The scope of fields making use of pattern recognition is increasing, and it raises the necessity for a basic knowledge of the diverse methods for a conscious selection of those that are most adequate to the data. Being out of the scope of the MSc course, the interest and necessity to explore this subject comes along with its increasing importance in medical sciences research.

Hereupon, the dissertation is defined on two aims: present the most popular methods in classification and, posteriorly, apply those methods on the real case study for developing pre-screening techniques for Obstructive Sleep Apnea (OSA). OSA is a common sleep related disorder and a serious public health concern estimated to affect more than 100 million adults, characterized by the involuntary and intermittent obstruction of the upper airway cannal during sleep. This disturbance is associated to several health morbidities that in turn can aggravate the severity of the condition if untreated for a long-term. The present concern about OSA relates to the continued high underdiagnosis and difficulty for a premature diagnosis. The pre-diagnosis based on pathophysiological evaluation and symptoms recording has shown to be limiting for detecting most patients that should be tested for OSA. Field researchers highlight the importance of proteomics research, recently in expansion, for understanding the biochemical complexity of the disturbance and it is proposed the consideration of such on the pre-diagnosis since the demanded costs, resources and low capacity do not allow for the general population to undergo the gold standard Polysomnography test.

The data available for the case study exemplifies the conditions that are not desirable for a classification study: small sample size, imbalanced distribution, high dimensionality. In this way, a set of steps were taken before classification to prepare the data (*e.g.* data acquaintance and variable selection). For unsupervised classification, hierarchical and partitioning methods are applied. Decision Tree, Naïve Bayes and Logistic Regression are considered for supervised classification. Results show that the pattern search with unsupervised methods was not capable to adjust so well to the classification of OSA. Supervised methods, trained with the complete data and tested with Leave One Out Cross Validation, showed in turn a better performance for classification of the true outcome classes of both binary and multiclass outcome variables. From those, decision trees showed to be the best performing supervised method applied to the case study mainly due to the capability of better interpreting the results in comparison to the other methods since the performance was similar in all supervised methods. In terms of variable importance, a mixture of variables representing proteoform expression and clinical parameters associated to OSA may return the best set of variables for a possible pre-screening of the disturbance.

It is concluded that the most fruitful, consistent and generalized results of a pattern analysis can be provided by the good quality of the data, the concrete definition of the study purpose (whether to find patterns or set rules of decision for classification), and a wise selection of the mechanism behind the construction of the classifier.

Key-words: patterns, classification, pre-diagnosis, proteoform, OSA

Resumo

O ser Humano é um especialista nato na deteção de padrões naturais no meio envolvente. A consciência desta capacidade determinou o uso dos padrões detetados como pilar importante na tomada de decisões, não só do dia-a-dia, mas também de desenvolvimento futuro da sociedade e compreensão da complexidade do meio em que vivemos. Os recentes avanços tecnológicos manifestados desde meados do século 20 têm disponibilizado *booms* de informação albergada de natural complexidade e crucial para o estudo de padrões que, por sua vez, têm permitido interpretar e dar fundamento às conclusões necessárias. Para uma análise mais completa dos dados, tem-se tornado cada vez mais necessário o uso de métodos de fundamento matemático e estatístico que permitam a melhor análise da informação. Não só se verifica necessário o uso destas metodologias, mas também de uma maior capacidade de processamento das mesmas recorrendo ao seu desenvolvimento em computadores (*Machine Learning*) para, com maior eficácia e precisão, proceder à exploração de grandes conjuntos de informação e posterior tomada de decisão. Dois tipos de métodos se destacam quando se recorre ao uso de máquinas: (métodos não supervisionados) aplicados na procura de padrões recorrendo apenas às características observadas nos elementos em estudo e sem o conhecimento *à priori* de uma real classificação dos mesmos; (métodos supervisionados) aplicados na criação de modelos classificadores com base em padrões observados nos elementos previamente caracterizados e classificados para a variável de interesse, para a posterior aplicação das regras construídas para a discriminação de padrões em novos elementos com classificação desconhecida. Os métodos de classificação supervisionados são levados em maior consideração quando está pendente a construção de regras de classificação para a tomada de decisões nas mais diversas áreas da ciência, economia, métodos de reconhecimento inteligente, entre outras. Não abrangido no conteúdo do Mestrado em Bioestatística, o presente interesse no estudo de padrões é justificado pela crescente necessidade do seu uso na área da saúde que requer, consequentemente, um conhecimento prévio dos métodos existentes e do seu correto modo de aplicação.

A presente dissertação está definida em dois objetivos: (1) estudar o conteúdo teórico de um conjunto de métodos de classificação não supervisionada e supervisionada, tendo por consideração os mais popularmente usados; (2) aplicar os métodos apresentados no caso de estudo da Apneia Obstrutiva do Sono (AOS). AOS é um distúrbio comum do sono e um problema sério de saúde pública que se estima afetar mais de 100 milhões de adultos ao nível global. O distúrbio caracteriza-se por afetar os indivíduos durante o sono com a involuntária e intermitente obstrução (parcial e/ou completa) do canal de respiração superior. Os sintomas são múltiplos e manifestam-se em ambos os períodos de sono e acordado: comportamento de roncador, disrupção do sono, disrupção do ciclo de sono, dores de cabeça, problemas de concentração, *etc.* Para além destes, o distúrbio está associado a comorbidades do foro metabólico, respiratório, cardio e cérebro-vascular. Consoante a sua severidade, são identificados três estádios de classificação dos pacientes: estádios ligeiro, moderado e severo. Quando adequado à severidade e ao paciente, o processo de terapia para um diagnóstico positivo é geralmente bem sucedido com um tratamento de Pressão Positiva Contínua nas vias respiratórias (CPAP para *"Continuous Positive Airway Pressure"*), sobre a qual está relatada a sua eficácia para a redução da gravidade do distúrbio e melhoria da qualidade de vida. Presentemente, a maior preocupação relativa ao distúrbio destaca-se pela grandeza estimada de casos por diagnosticar devida à dificuldade de pré-diagnóstico e baixa acessibilidade da sociedade para se submeter ao diagnóstico com o teste da Polissonografia. Este *gold standard* de diagnóstico é maioritariamente recomendado pelo médico quando são relatadas queixas, por parte do próprio ou de um familiar, de sintomas relacionados e/ou quando o profissional de saúde identifica uma maior propensão para o distúrbio após avaliação física do doente para outras comorbidades de saúde com uma estudada associação com o distúrbio, como a obesidade. Não só os critérios de recomendação de diagnóstico poderão estar fundamentados em avaliações pouco consistentes, a própria acessibilidade da PSG é reduzida para a sociedade geral dado o seu preço, disponibilidade diária de camas nas facilidades de diagnóstico e a necessidade de equipas médicas altamente especializadas. Todos os pontos referidos justificam a crescente necessidade de analisar opções de um pré-diagnóstico do distúrbio mais facilitada, tanto em termos financeiros como em disponibilidade de testes, e que permita uma maior deteção de casos em indivíduos que não se integrem nos critérios de avaliação atualmente considerados no pré-diagnóstico. Até ao momento presente, o estudo da AOS tem dado destaque ao conhecimento dos sintomas e doenças associadas em comparação ao reduzido conhecimento dos complexos mecanismos bioquímicos do distúrbio que afetam o organismo humano, sendo essencial dar foco aos mesmos. Os desenvolvimentos tecnológicos das últimas décadas, previamente mencionados neste resumo, vieram beneficiar os estudos de investigação em proteómica, *i.e.* da análise integral dos componentes de proteínas ao nível celular, de tecidos e do organismo, permitindo a acessibilidade a uma maior diversidade de e abundância de variantes de proteínas. As proteoformas são naturalmente derivadas pelos mecanismos naturais de síntese e modificação de proteínas no organismo, sendo apontadas para a análise de candidatas a biomarcadores de associação e meios de diagnóstico do distúrbio.

Neste trabalho, são consideradas variantes do tipo Apolipoproteína A2 e Apolipoproteína C, observadas em pacientes de estudo aos quais foi realizado o teste de PSG e deste modo são conhecidos os seus diagnósticos para o distúrbio. Os dados disponibilizados serão considerados para a construção de modelos afinados para a análise de capacidade de classificação das proteoformas presentemente consideradas para o diagnóstico de novos pacientes nos quais sejam observadas as expressões destas mesmas proteoformas mas estes não tenham sido sujeitos ao teste de diagnóstico. Tratando-se de um caso de saúde pública, é importante o conhecimento total do processo de construção do modelo, afinação e interpretação direta dos passos de diagnóstico sobre novos pacientes. Deste modo, foram escolhidos os seguintes métodos de caixa visível (*white-box*): (métodos não supervisionados) partição com K-medoids e aglomeração hierárquica; (métodos supervisionados) Árvores de decisão, Naïve Bayes e Regressão Logística binária e ordinal. Adicionalmente, é apresentado um método supervisionado de caixa não visível (método de *ensembling* baseado em árvores de decisão) mas não aplicado ao dados de estudo.

O caso considerado para a aplicação dos métodos de classificação revela a importância de um bom conjunto de dados em termos de tamanho da amostra e qualidade das variáveis que são observadas nos elementos de estudo. Para o caso concreto, caracterizado por um reduzido tamanho amostral, uma grande dimensionalidade de variáveis, e uma distribuição desequilibrada dos elementos na variável de interesse (diagnóstico do distúrbio), foi necessário um bom conhecimento e preparação prévia dos dados que incluiu imputação de dados omissos e seleção de variáveis. A aplicação dos métodos não supervisionados revelou ser pouco útil para análise de padrões naturais nos dados e possível ajustamento dos clusters formados às verdadeiras classes de diagnóstico. Para a aplicação de métodos supervisionados, a criação das amostras de teste teve por base o uso do método *Leave One Out Cross Validation*. Destes, o método de Árvores de Decisão mostrou ser o melhor aplicado, principalmente devido à simplicidade de interpretação dos resultados já que em termos de desempenho foi semelhante aos restantes métodos supervisionados. Dos resultados obtidos, destacam-se a variável de representação de variação da expressão da proteoforma `dfA2DQ` relacionada com os níveis de colesterol, e da variável `insul` de medida dos níveis de insulina no organismo. Conclui-se para o caso de estudo que para a construção de técnicas de triagem seja considerada a análise de expressão de proteoformas em conjunto com variáveis clínicas cuja associação com a Apneia Obstructiva do Sono tenha sido anteriormente estudada.

Palavras-chave: classificação, qualidade de dados, saúde Humana, Apneia Obstructiva do Sono

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

This dissertation represents the will to deepen the knowledge on Pattern Recognition (PR), which subject is not introduced in the Master course but it is of great importance in the field of biostatistics applied to Human healthcare such as for the development of pre-diagnosis rules for the detection of medical conditions. The document is hereby introduced with a personal note from the student.

> *What is learnt in the classroom may represent the backbone contents of what we seek to understand, but progression is accomplished by constantly researching the abundance and diversity of information that is growing importance on our daily life. I understand this dissertation work as an intent to impel the students to become more independent on their knowledge seeking and to strengthen their capability to fulfill interest skills with experience and mattered knowledge. As a close and growing interest of mine, I have decided to explore the subject of Pattern Recognition. This subject is characterized by field multidisciplinarity and is designed for the study of patterns in the data. Humans are innate pattern recognizers, and because we have been able to learn and use them for important decision-making in the natural and artificial-made surrounding environments, the empirical importance of Pattern Recognition role is unquestionable.*

The importance of Pattern Recognition (PR) is well recognized and has an increasing history of success already dated in the 16^{th} century (Jain & Duin, 2004).
As the definition implies, PR is a system of assembled techniques (statistical, mathematical, heuristic, etc.) aimed to analyze patterns in the data; the integrated set of multidisciplinary techniques provides methodologies capable of performing several analysis to deepen the understanding of object description and data structures, therefore being the essence of PR (Liu, Sun, & Wang, 2006; Ross, 1998; Srihari, 1993).
Two aspects may be considered for a better use of the methodologies. First, a data sample always carries the natural complexity and noise existing in the source environments from where it is observed (the population). Secondly, better results on pattern search or better support for decision-making can be expected, respectively, from clustering or classification analysis if the most relevant patterns existing in that population are available in the data sample (Ross, 1998). Patterns are defined by the characteristics observed in the elements. The abundance and diversity that is necessary in the data to study them was made possible and available with recent technological developments (Fan & Li, 2006). Subsequently, this increase in resource availability since the middle of the 20^{th} century became a trigger for the constant development of computational systems with increased capacity to apply PR methodologies capable of transforming the incoming data into useful information (*Machine Learning*) (Ge et al., 2017; Liu, Sun, & Wang, 2006).
The set of methodologies is adapted to the type of pattern problem being studied; two situations are most commonly differentiated:

- Pattern discovery (unsupervised learning): defines the necessary analysis to find natural patterns in the data capable of clustering the elements without prior knowledge of class existence. In this scenario the unsupervised (clustering) methodologies are considered.
- Pattern classification (supervised learning): in the event of an outcome event for which the classes are well defined on a set of elements, the methodologies are intended to perform a group pattern characterization based on the elements of each class and create a set of rules intended for the future classification of newly observed elements for which the same characteristics are known but the class they belong to is not. The

reliance on characterized elements well defined in their class suggests a supervision possible to implement with these methodologies.

Pattern discovery and classification are visions defining possible approaches for the appliance of PR methodologies (Figure 1.1). Having defined the appropriate approach to analyze the data, the user may need further to decide the methodology based on the desired degree of transparency and interpretability of the step-by-step rationale justifying the results to be returned. The types white/glass box *vs* black box characterize the methodologies respetively into those enabling the know-how and interpretability of the process *vs* those having a complex structure not allowing the human interpretability. The user must measure the necessity for a transparent analysis of the particular data as it possibly may face a trade-off decision between the prediciton accuracy encountered with black box (trust in prediction) and the interpretability plus knowledge of the process strenghts and weaknesses met with white box (trust in the model).

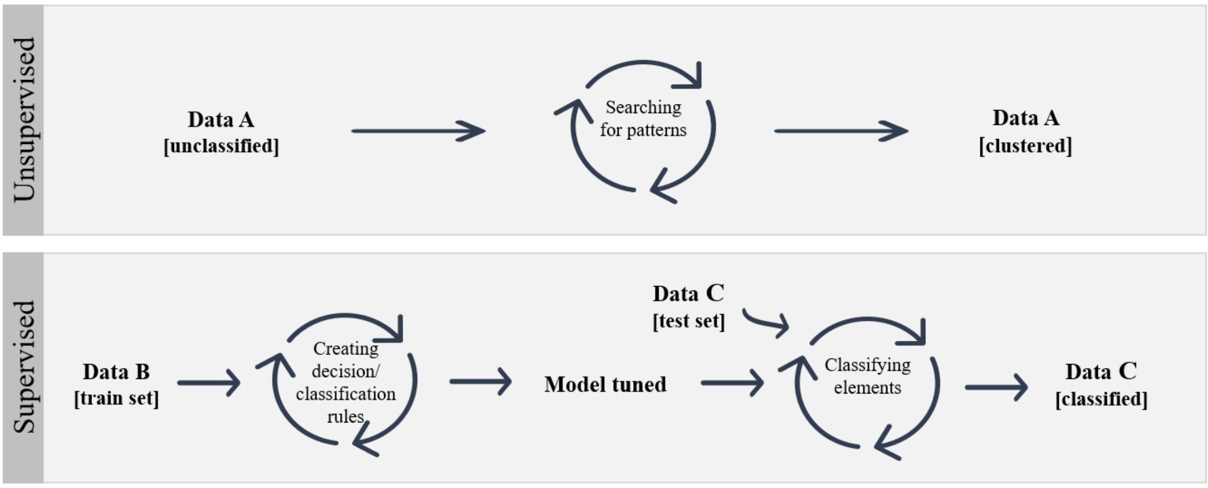


Figure 1.1: The mechanism of pattern recognition behind supervised and unsupervised methods.

In the context of the study of pattern recognition methodologies, a set of objectives was defined for this dissertation.

- Explore the most popular unsupervised and supervised algorithms in pattern recognition, preferably defined on the type of white-box methodologies (specific for the application on the case study as described by the 3^{rd} objective).
- Present the methods focused on their type of nature, theory behind the framework, data assumptions and application criteria.
- Apply the algorithms on a real case study to assess a set of comparable results: implementation complexity, results interpretability, accuracy of correct/consistent prediction of patterns, advantages and disadvantages to the case study.

Regarded in diverse scopes (*e.g.* finances, business, science, security, person identification, airplane piloting) (Jain & Duin, 2004 ; Ross, 1998), PR is hereby oriented to an urging need in the scope of public health sciences, for which the system use has been improving medical decision related to therapeutics and prognosis of patient's specific condition (Katsios & Roukos, 2010). Thus, the adequateness of this case is assured by its strong connection with statistics applied to health sciences, as pursued in the context of this Master degree, and the growing popularity, among field researchers, of PR methodologies to tackle the current issue at hands.

Case study - Diagnosis of Obstructive Sleep Apnea

Obstructive Sleep Apnea (OSA) is a common sleep related disorder and a serious public health concern, estimated to affect more than 100 million adults (Lurie, 2011; "Population ages 15-64, total," n.d.; Watson, 2016), two to three times more prevalent in men than women (Young, 1993), and between 1-4% of the pediatric population worldwide (Lumeng & Chervin, 2008).

If untreated for a long-term, the disorder may be responsible for health burdens of which most may represent a risk for increased disturbance severity: propensity for metabolic dysfunction, hypertension, diabetes, obesity, increasing mortality, cardio and brain vascular morbidity, among others (Coughlin et al., 2004; Feliciano et al., 2015; Partinen, Jamieson, & Guilleminault, 1988; Young et al., 2008; Young, Peppard, & Gottlieb, 2002).

For a conclusive diagnosis, patients must undergo the Polysomnography test (PSG). PSG is an overnight laboratory-based test and the gold standard for OSA diagnosis, on which the individual is monitored for sleep related parameters, *e.g.* respiratory events, brain activity, oxygen saturation and other physiological parameters (O'Connor, Thornley, & Hanly, 2000; White et al., 1995). The therapy process on a positive diagnosis is usually well succeeded. Most patients are indicated for a treatment with nasal Continuous Positive Airway Pressure (CPAP) and the method shows reported efficiency for reducing the disturbance's severity and improvement of the quality of life (Feliciano et al., 2015; Flemons, 2002; Gay et al., 2006; Weaver et al., 2007).

The present concern about OSA relates to the continued high underdiagnosis - non-detection of affected individuals with unnoticed health problems/symptoms - and difficulty for a premature diagnosis (Feliciano et al., 2015) as a preventing measure for futurely increased health problems.

Limitations of the actual pre-diagnosis are discussed. Most cases of medically indicated PSG are reasoned by patient's and relative's reported symptoms (Torres et al., 2017) or clinician's suspicion based on the interpretation of known risk factors associated to increased predisposition to OSA disturbance (further described in the preface of chapter 3).

The ready availability of the test may furthermore be limited to the general society due to its elevated cost, low daily diagnosis capacity, time consumption, requirement of highly specialized medical and technical teams to perform it, *etc* (Ferrie et al., 2011; Flemons et al., 2004).

Societal needs for a better management of the disturbance must be met. The extensive pathophysiological[1] study of OSA may not be sufficient to tackle the problem stated, as the knowledge gap on the complex biochemical mechanisms of the disturbance still persists. The recent technological developments from the past decades on *proteomics research*[2] became crucial for the assessment, detection and abundance quantification of a massive (unquantifiable) number and variety of protein forms. Field researchers enlarge the importance of proteomics research to improve the rather defective set of rules for OSA pre-diagnosis, as investigations have concluded in their data the association of OSA presence and severity with the abundance of protein forms.

This study is motivated by previous documented investigations on the analysis of OSA presence and severity associated to the abundance (expression) of proteoforms. First, in 2016, studies observed an association between the presence of proteoforms of modified transthyretin (TTR) and the diagnostic of OSA (Bodez et al., 2016). A year later, further investigation allowed for the construction of a supervised regression model based on the expression of TTR proteoforms that successfully discriminated individuals affected from those not affected, although the observed expressions of those predictors were ineffective for the discrimination of OSA severity (Torres et al., 2017). These recent studies initiate and reinforce the application of classification methods for the analysis of protoeforms' potential as candidate biomarkers for disturbance detection in the prospect of developing a simple and inexpensive point-of-care screening test (Bodez et al., 2016; Montesi, Bajwa, & Malhotra, 2012; Torres et al., 2017).

[1] *The changes of the body normal functioning due to a specified condition.*

[2] *"The analysis of the entire protein complement of a cell, tissue, or organism under a specific, defined set of conditions..."* (L.-R. Yu, Stewart, & Veenstra, 2010)

1.1 Chapter roadmap

This dissertation is composed in five chapters. Objectives and motivations have been presented in this introductive chapter, and the chapters ahead are briefed for a better comprehension of its structure.

The next chapter 2 addresses a framework methodology necessary for clustering and classification analysis. It will be commenced with theoretic presentation of the unsupervised and supervised methodologies that were selected, followed by a final subchapter focused on guidelines for data preparation particularized to the case study.

Having the necessary background knowledge of the steps for familiarizing and perform pattern analysis in data, the next step focus on the actual application on the case study and presentation of the corresponding results (chapter 3).

The discussion of the methods' application to the case study is centered in chapter 4, followed by the concluding chapter 5.

Chapter 2: Methodological framework

The present chapter attempts to fulfill the guidelines for the exploration, preparation and analysis of patterns in the data.

A first section provides a set of tools that are utilized in the methods presented next (2.1). This section is complementary but necessary and particularly intended for the reader that may not be familiarized with those tools that are essential for the operation of the methods.

The next section 2.2 encompasses the theoric presentation of the chosen set of PR methods, focused on multidimensional samples, in the prospects to (i) introduce the step-by-step mechanism for pattern recognition inherent to each particular method and (ii) guide for their proper use. Most methods proposed here have a white box type characterization and are amongst the most popularly used for the study of patterns. For unsupervised classification, the partitioning and hierarchical clustering methods are explored; as for supervised classification, four white-box and one black-box methods are respectively presented: Decision tree, Naïve Bayes, Binomial Logistic Regression, Ordinal Logistic regression and the Ensemble of decision trees (not applied to the case study).

The measures to evaluate performance and adequacy of each type of method (unsupervised and supervised learning) are summarized in section 2.3, followed by section 2.4 that concludes this chapter with important guidelines for data familiarizing and preparation for PR analysis.

2.1 Tools for classification

2.1.1 Data proximity

Relations are commanded by the characteristics observed in the elements and these can alert for the existence of natural patterns in the data. Proximity has a strong meaning for the measure of these relations as it focus on understanding why a certain element in the data may be more close to a second, but more apart from a third. The measure of proximity between elements i and j is based respectively on the vectors of p observed characteristics $\mathbf{x_i} = (x_{i1}, ..., x_{ip})$ and $\mathbf{x_j} = (x_{j1}, ..., x_{jp})$.

Considering the properties (i-iv) described below, the proximity may be inferred by the measure of similarity s_{ij}, or contrarily, dissimilarity d_{ij} between the two elements. Dissimilarity measures can be also referred to as distance measures but the equivalence only applies if the triangular inequality represented by property (v) is verified for all the three given pairs of elements (ij), (ih), and (jh).

(i) $s_{ij} = 1 - d_{ij}$

(ii) $d_{ii} = 0$ and $s_{ii} = 1 \; \forall \; i : 1 \leq i \leq n$

(iii) $d_{ij} \geq 0$ and $s_{ij} \geq 0$

(iv) $d_{ij} = d_{ji}$ and $s_{ij} = s_{ji}$

(v) $d_{ij} + d_{ih} \geq d_{jh}$

In table 2.1 is represented a summary of proximity measures compiled from Everitt *et al.* (2011b). The type of proximity measure used is generally based on the nature of the p set of variables that are characterizing the elements, simplified in three possible scenarios:

[a] All p variables are qualitative: measures of similarity;

[b] All p variables are quantitative: measures of dissimilarity (distance);

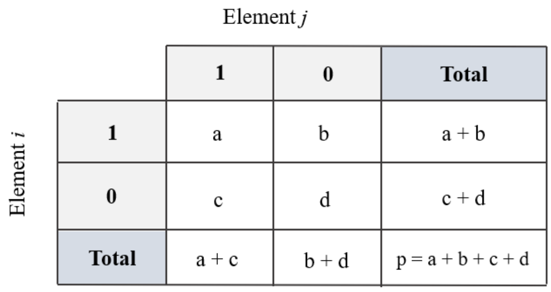
[c] Specific measures apply for a mixture set of variable natures, *i.e.* for f quantitative and q qualitative variables ($q = p - f \; \forall \; f$ and q variable: $1 \leq f \leq p$ and $1 \leq q \leq p$).

Table 2.1: Proximity measures applied to qualitattive, quantitative and mixed sets of variable natures.

Measure	Formula	Observations
S1: Matching coefficient ($w = 1$)	$s_{ij} = \frac{a+d}{a+w(b+c)+d}$	Rogers and Tanimoto ($w = 2$); Gower and Legendre (1986) ($w = 0.5$).
S2: Jaccard coefficient (1908) ($w = 1$)	$s_{ij} = \frac{a}{a+w(b+c)}$	Sneath and Sokal (1973) ($w = 2$); Gower and Legendre (1986) ($w = 0.5$).
S3: Score based coefficient	$s_{ij} = \frac{\sum_{m=1}^{p} s_{ijm}}{p}$	-
D1: Minkowski distance	$d_{ij} = \left(\sum_{m=1}^{p} w_m^r \lvert x_{im} - x_{jm} \rvert^r\right)^{\frac{1}{r}} \quad (r \geq 1)$	$w_m > 0$ ($1 \leq m \leq w$) are the weights given to variable m. Euclidean distance ($w_m = 1$, $r = 2$); Manhattan distance ($r = 1$).
D2: Mahalanobis distance	$d_{ij} = \sqrt{(\mathbf{x_i} - \mathbf{x_j})^T S^{-1} (\mathbf{x_i} - \mathbf{x_j})}$	-
S4: Gower's similarity distance (1971)	$s_{ij} = \frac{\sum_{m=1}^{p} w_{ijm} s_{ijm}}{\sum_{m=1}^{p} w_{ijm}}$	-

Similarity measures

The similarity measures S1 and S2 described in table 2.1 apply to binary variables. The parameters a, b, c and d represented in the formulas are originated from a 2×2 count table of observed correspondence for all p variables between two given elements i and j from the sample space (Figure 2.1). The sum of the totals corresponds to the total p number of variables characterizing the individuals as each count represents the match between the two elements for a given variable X_m in any of the cells represented in the figure.



Element j

Element i		1	0	Total
	1	a	b	a + b
	0	c	d	c + d
	Total	a + c	b + d	p = a + b + c + d

Figure 2.1: 2×2 count table of characteristic matching between two elements i and j.

The measures of similarity for binary variable types offer flexibility regarding the count of negative matches d (counts of attribute absence in both elements). Evaluated case wise, if the negative matches are as important as the positive matches a (counts of attribute presence in both elements) then measure S1 and measures respectively derived are to be used. Otherwise, measure S2 and corresponding derivations offer the alternative approach by excluding the counts of d.

In case the qualitative nature specifies more than two classes (multiclass), the score based coefficient calculates the similarity as the proportion of shared characteristics weighted equally by all the p variables (S3 from table 2.1). For each variable X_m ($1 \leq m \leq p$) is calculated the score s_{ijm} of binary nature returning one if the elements i and j share the presence of the same level within the *m-th* variable or zero if not.

Dissimilarity measures

In the same manner as for the measure of similarities, the elements are measured in pairs. The dissimilarity of the elements in the data is usually represented by a symmetric dissimilarity matrix ($d_{ij} = d_{ji}$) with a null diagonal ($d_{ii} = 0$). Each component of this matrix represents the estimated dissimilarity in a pair (figure 2.2).

$$D = \begin{bmatrix} 0 & d_{1,2} & \cdots & d_{1,j} & \cdots & \cdots & d_{1,n} \\ d_{2,1} & 0 & & & & & \vdots \\ \vdots & \vdots & & & & & \vdots \\ \vdots & \vdots & & \ddots & & & \vdots \\ d_{i,1} & \vdots & & & & & \vdots \\ \vdots & \vdots & & & & & d_{n-1,n} \\ d_{n,1} & \cdots & \cdots & \cdots & \cdots & d_{n,n-1} & 0 \end{bmatrix}_{n \times n}$$

Figure 2.2: Example of an $n \times n$ dissimilarity matrix.

The Euclidean distance is the most popularly known measure of dissimilarity (derived from distance D1 from table 2.1). The value of this distance between i and j can be interpreted as the real distance between the vector of characterizing variables of each element ($\mathbf{x_i}$ and $\mathbf{x_j}$) in the Euclidean space. The interpretation of variables' importance based on the Euclidean distance can become erroneous when the unit of measure is not the same for all the variables involved in the calculation of the distance, or the unit is the same but the order of magnitude is not. For these situations it is often considered the standardization of the values by subtracting the mean and dividing by the standard deviation of the respective variable, so that an increase g in the observed value of a variable may have the same meaning as for the same increase g in any other variable. Additionally, the relation between variables is not considered by the Euclidean distance, increasing the risk of existing redundancy between variable.

Mahalanobis distance (generalized squared) tackles the gaps referred for the use of Euclidean distance. Performing the standardization of the euclidean distance by the variance-covariance matrix of the data and correction of the existing variable correlation, this distance is scale-invariant, unitless and better prepared for analyzing relationships in multivariate data. The matricial formula for this distance accounts for the inverted variance-covariance matrix S^{-1} (distance D2 from table 2.1) (De Maesschalck, Jouan-Rimbaud, & Massart, 2000; Yoo et al., 2014).

Proximity in mixed variable natures

For these type of variable sets, the Gower's similarity distance (Gower, 1971) is usually considered (S4 from table 2.1). In the underlying formula, w_{ijm} is a flag indicator of the validity of the measure of similarity between i and j for the specific *m-th* characteristic; it can be set to zero if the observation is missing in at least one of the elements or when it is intended the exclusion of certain observations as for the removal of negative matches in binary variables. The overall similarity s_{ij} is obtained by weighting the sum of similarities calculated in each *m-th* characteristic observed (s_{ijm}) by the number of variable observations that were validated for the measure. If m is a qualitative variable, the value of s_{ijm} is one if the characteristics are matching between the two elements. For quantitative variables it is applied the measure $s_{ij} = 1 - \frac{|x_{im} - x_{jm}|}{R_m}$, where R_m is the range of observed values for the *m-th* variable.

2.1.2 Information value of the data

The value of information measures the importance and utility of the structure and consistency of the data. In the context here addressed, the measure of information is useful when it is necessary to evaluate the quality of an alteration made (*e.g.* a partition of the data) on the current state of the data. The higher the amount of information gained, the higher is the quality of the decision.

The information is studied in gain and indirectly by means of quantifying the augmented or diminished disorder between two states. Considering D_0 the representation of a measure of misinformation in the initial stage (prior to change) and D_1 after the change, the gain in information is the quantified amount of information of diminished disorder from a state to another (2.1):

$$Gain_{\,Information} = D_0 \; - \; D_1 \tag{2.1}$$

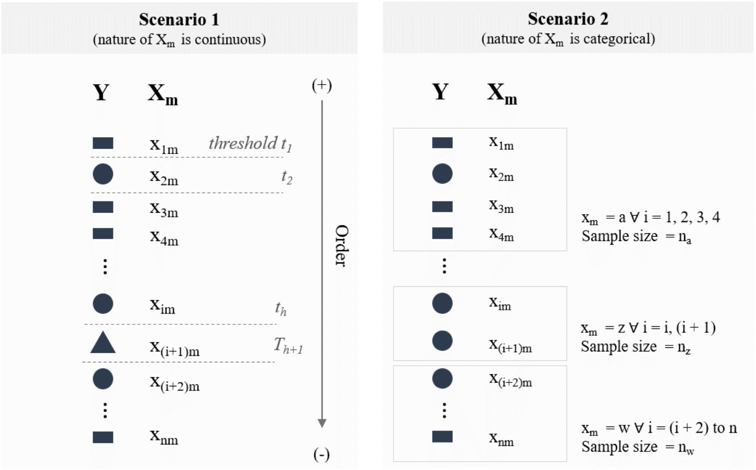
The measure of information is essential in the process of supervised classification undertaken by various methods such as the Decision tree and the derived Ensemble methods. In the light of these methodologies, the value of information is measured by assessing the disorder existing in the space of elements before and after the inherent step of partition is made. In this context, the disorder is represented by the amount of impurity that is defined according to the classes of the outcome variable present in the evaluated set of elements. The purity of a set is defined when all elements pertain to the same outcome class.

The two measures of disorder described in table 2.2 are particularly important: Shannon's entropy (Shannon, 1948) and Gini Index (Gini, 1912). For each measure, p_k represents the probability of the *k-th* outcome class ($1 \leq k \leq c$) in the set of elements analyzed. The formulas presented in the table are further demonstrated in Appendix A.

Table 2.2: Measures of disorder used in Decision tree classifiers.

Measure of disorder	**Formula**	**Maximum value (outcome classes: c)**
Shannon's Entropy	$-\sum_{k=1}^{c}[p_k \times \log_2(p_k)]$	$\log_2(c)$
Gini Impurity	$\sum_{k=1}^{c}[p_k(1-p_k)] = 1 - \sum_{k=1}^{c} p_k^2$	$\frac{c-1}{c}$

For each characterizing variable X_m, the information of that variable is estimated in the observed value x_m for which the alteration of the state produces more information gain or lowest entropy increase (threshold point). According to the nature of the variable (quantitative *vs* qualitative), the choice for the better threshold point can be accompanied with figure 2.3 and the steps described below.



Sample size: n
Y - random variable representing the outcome (categorical nature)
X_m - random variable representing the characteristic m, $1 \leq m \leq p$

Figure 2.3: Procedure for the analysis of information value in the data.

Steps:

For a set of elements of size n, proceed as follows:

[a] Identify the nature of the characterizing variable X_m.

- If X_m is of qualitative nature with l classes, calculate l times the entropy for every scenario of *"characteristic l is present vs absent"*.
- If X_m is continuous:
  - Order in a descending direction the values observed in the n elements regarding the *m-th* characteristic, not forgetting the outcome class to which these elements belong to.
  - For the element i with observed outcome $y_i = s$ and element $i+1$ with observed outcome $y_{i+1} = b$ ($y_i \neq y_{(i+1)}$), consider a candidate threshold the midpoint between the observed value of x_{im} and $x_{(i+1)m}$. In case of tied values, i.e., x_{im} = $x_{(i+1)m}$ for $y_i \neq y_{(i+1)}$, the midpoint between x_{im} and $x_{(i+1)m}$ is no longer a candidate value and the process continues to analyze further candidate thresholds.
  - Calculate the entropy for all the candidate thresholds and select the threshold for which the value of information gain is higher.

2.1.3 Estimator of probability distribution

Another important analysis capturing the structure of the data is the possibility to estimate the respective probability distribution. Particularly focusing on quantitative data, the probability distribution determined by the probability density function is of considerable importance in classification methods of probabilistic origin such as the supervised Naïve Bayes. In many cases, however, the behavior of the respective data does not allow for the adjustment of a specific standard parametric distribution.

This section presents an univariate non-parametric alternative for estimating the distribution of the data: univariate Kernel Density Estimator (KDE). The kernel is extensively applied for the density estimation constructed:

$$f(x|h) = \hat{f}_K(x;h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x^* - X_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x^* - X_i}{h}\right) = \frac{1}{nh}\sum_{i=1}^{n} K(u), \tag{2.2}$$

where $f(x|h)$ is the unkown true density distribution, $\hat{f}_K(x;h)$ the density distribution estimated with the kernel, $\mathbf{x_i} = (x_1, ..., x_i, ..., x_n)$ the univariate vector of elements, x^* the observed value of that variable for which is being estimated the probability, and h the smoothing parameter (bandwidth).

As for all probability density functions associated to a distribution family, the KDE must meet the following conditions:

$$\int_{\mathbb{R}} K(u)d(u) = 1 \quad \text{and} \quad K(u) \geq 0 \tag{2.3}$$

In this non-parametric function, h represents a continuous and strictly positive bandwidth ($h \in \mathbb{R}^+$). The bandwidth is an important piece of the method defining the smoothness of the kernel fit to the data (zooming

property). Consider the kernel estimation of the distribution density for the variable m. The kernel smooths the fit to the data by analyzing windowed pieces of the unknown distribution of observed values for that variable and adjust each piece with a known distribution. In another words, it controls the size of the neighborhood around x^* and consequently adjusts too well for smaller values of bandwidth and adjusts too little for larger values. Extreme values of bandwidth are not recommended. An underfitted curve for the set training the classifier may not capture the essential details for training the algorithm that is used afterwards for the class prediction of a new test set with unknown classification. On the other side, having an overfitted curve will overly capture the intrinsic details of the train sample and cannot be used as a representation of the general population.
The best kernel estimation of the train set is the one which may be represented between the underfitted and overfitted regions of the data density curve. An optimal bandwidth value results on better classification of the test sample with the trained algorithm. Table 2.3 regards some examples for univariate bandwidth estimation.

Based on the Gaussian probability density function described in equation (2.4), the gaussian kernel density estimator function $K(u)$ analyzes the fit of a gaussian distribution in each of the windows set by the bandwidth with the derived $K(u)$ (2.5). Other types of $K(u)$ functions can also be considered (*e.g.* epanechnikov, cosine, triangular) (Węglarczyk, 2018).

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{2.4}$$

where $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma > 0$

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-X_i}{h}\right)^2}, \tag{2.5}$$

where σ representing the standard deviation is replaced by h.

Table 2.3: Short summary of measures for bandwidth estimation. † From Silverman (1986).

Measure	**Formula**	**Observations**
"Gaussian" alike	$\hat{h}_{x_m} = \hat{\sigma}_{x_m}$	Unknown disitribution may be very similar to the gaussian
Rule of thumb (based on the asymptotic mean integrated square error) †	$\hat{h}_{x_m} = 1.06\hat{\sigma}_{x_m} n^{(-\frac{1}{5})}$	$\hat{\sigma}_{x_m}$: estimated standard deviation of variable x_m; n: sample size
Rule of thumb with reduced smoothing †	$\hat{h}_{x_m} = 0.9 \min\left(\hat{\sigma}_{x_m}, \frac{IQR}{1.34}\right) n^{-\frac{1}{5}}$	More fitting to non-unimodal distributions
Based on the interquartile range †	$\hat{h}_{x_m} = 0.79 IQR n^{-\frac{1}{5}}$	Estimator more robust to outliers

2.1.4 Data sets for supervised classification

The application of supervised classification implies the definition of a train set to tune the model and a test set to validate the classifier's performance. Both sets are defined within the available sample, and the train set is by rule the largest in order to allow more abundance and diversity in the information that is given to create a classifier with rules more generalized for new elements. Here are presented two possible techniques of facilitated implementation and understanding: Bootstrap sampling and Cross Validation.

2.1.4.1 Bootstrap sampling

The method of bootstrap proposes the random sampling, with replacement, of the elements in the data.
For a sample **S** consisting of n_S elements characterized y p variables ($n_S > 0, p > 0$), a bootstrap sample **B** of size n_B ($1 \leq n_B \leq n$) is generated by randomly resampling n_B times the elements from **S** with replacement. If two elements i and j are selected for **B** ($i, j : 1 \leq i \leq n$ and $1 \leq j \leq n$) the replacement rule expresses the condition that $\forall\ i, j \in$ **B**, $i = j \vee i \neq j$ (Davison & Kuonen, 2002).

The application of the method is found in the mechanism of some supervised methods such as on Ensemble based classification (see section 2.2.2.2). For the proper use of bootstrap some knowledge is necessary. Each time the bootstrap method is randomly resampling with replacement from a sample **S**, the probability of an individual to be resampled is $P(\text{sampled} = i) = w \times \left(\frac{1}{n_S}\right)$, where w may represent the importance of the element to be sampled in the specific iteration. Having no criteria to set the weights for the elements, the decision becomes naive as the same weight is usually given to all elements from **S**. If the sample size n_S is small, the chance of originating a bootstrap sample **B** too attached to **S** rises the problem of originating a sample **B** of low quality to be a representation of the population and mask the more realistic performance evaluation of the classifier.

2.1.4.2 Cross Validation

Another known method is the Cross Validation (CV) also named as K-fold Cross Validation. This technique creates f definitions of train and test sets based on the f number of folds the user wishes to apply.
Given the same sample **S** of size n_S presented in the Boostrap method, CV proceeds as follows for a f number of folds:

[a] The size of each fold is approximately calculated to $(\frac{n_S}{k})$. To each fold is given an unique identification id between one and f.

[b] For each iteration t ($1 \leq t \leq f$), select the fold with id = t as the test set and the remaining folds as the train set.

A known derivation of CV is the Leave One Out Cross Validation (LOOVC) in which the number of folds f equals the number of elements in the data n, thus leaving one element for test at each iteration.
Cross validation may present an alternative to the Boostrap method by reducing the bias that is produced when the bootstrap method is used. Moreover, this method ensures that every element is included in both the training and test sets, thus allowing for variety in the process.

2.2 Classification methodologies

2.2.1 Unsupervised methods

Clustering methods, as increasingly being known, are the specialists on revealing natural patterns. They indicate the origin of clusters to where the data points may be aggregated. But what defines a cluster?
Being a word often referred to qualitatively, clustering considers the aggregation of objects belonging to a data set. Quantitatively though, it is quite difficult to make a clear definition. Although it has been considered by authors that a definition depends on the user's judgement (Bonner, 1964), a more consistent definition, considering the vision of Pattern Recognition, may be the evaluation of two important intra and inter-cluster properties: the internal cohesion of the cluster (homogeneity) and external isolation from the remaining clusters (separation).
The diversity of methods is considerable and required a scope shortening to encompass here those fitting the objectives proposed. Here are presented two classic and most utilized types of unsupervised learning relying on the between objects proximity to command the clustering decisions made: partitioning and hierarchical clustering methods.

2.2.1.1 Partitioning

Partitioning is a non-sequential mechanism relying on the measure of proximity to perform one step clustering trials, *i.e.*, the creation of a partitioning scenario of the n elements from data D in which the clustering of an element i does not depend on the the clustering performed in the element j (j and $i \in D$, $i \neq j$) since clustering is performed simultaneously for all elements. Respecting the focus on the formation of exclusive clusters (each element belongs to a single cluster), the partitioning methods here presented are K-means and its improved version K-medoids.

K-means

First proposed by Macqueen (MacQueen, 1967), the designation of the method self-introduces its main properties: **(i)** K-means performs partition under the restricted number of k clusters previously decided and **(ii)** the method evaluates an item to be the representative of each k cluster centroid based on the mean of the vectors of observed characteristics in the elements assigned to the particular cluster.
K-means has a self-improving mechanism with successive iterated trials intended to make the best data partition, which is met when the clusters reach stability in terms of the constituting elements and, consequently, the position of their centroid. Self-improvement requires that the process of partition must be performed as many times as the number of iterations that are needed, each iteration being divided in two steps:

- Grouping of the elements based on proximity measurements to the centroid of the cluster (representative item) *(estimation)*

- Re-estimation of the representative items for a new evaluation of the clusters' composition *(maximization)*

The constitution of the iterations described resembles the techniques of estimation and maximization of the EM algorithm, for which the user can further be acquainted with in Dempster et al. (1977).

UNSUPERVISED CLASSIFICATION: K-MEANS

Iteration t^0 **-** Preparation of prerequisites (1) define the k number of clusters to consider and (2) the proximity measure to apply. (these may be further considered in the process of model tuning)

Iteration t^1 **-** The algorithm randomly selects a k number of objects from the sample space to be readily selected as the initial representative items for the pretended clusters.

From iterations t^2 **to** t^{z-1}**, repeat steps 1 and 2:**

Step 1 - Assign each element i ($1 \leq i \leq n$) to the cluster for which the distance to the centroid is smaller than for the centroids of the remaining clusters.

Step 2 - Re-estimate the centroid of each cluster determined by the average of the points in the multidimensional space that are representing each element.

Iteration t^z **-** Stop the process when the cluster constitution is not altered and the clusters' centroids remain unchanged from previous iteration.
(Note: z is an iteration resulting from convergence of the process and therefore not fixed from the start.)

K-medoids

Robustness of the K-means method is attempted with the developed K-medoids. The difference between the two lies in the method for estimating the centroid of the clusters. The representative item, now defined as medoid, is recalculated at each "maximization" step as the data object i belonging to the cluster C_l that minimizes the sum of all distances between that object i and every other object g belonging to the same cluster ($g \neq i$ and $\forall \quad g, i : 1 \leqslant g \leqslant n$ and $1 \leqslant i \leqslant n$):

$$\sum_{g \in C_l} d(i,g), \qquad i \neq g \tag{2.6}$$

The improvement of k-means results in the weight reduction of the outlying objects within a cluster since these may have a considerable influence on the estimation of the representative item (Figure 2.4). Each iteration is meant for improvement of cluster formation: the groups are continuously shaped by redefined medoids and the clusters are restructured based on the dissimilarity measures (Kaufman & Rousseeuw, 1990b).

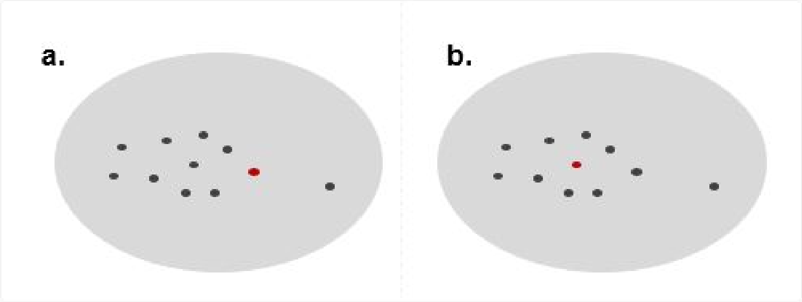


Figure 2.4: Representation of two partitioning methods. Representative item in (a) K-means is more influenced by potential outlying objects than (b) K-medoids.

UNSUPERVISED CLASSIFICATION: K-MEDOIDS

Iteration t^0 **-** Preparation of prerequisites (1) define the k number of clusters to consider and (2) the proximity measure to apply. (these may be further considered in the process of model tuning)

Iteration t^1 **-** The algorithm randomly selects a k number of objects from the sample space to be readily selected as the initial representative items for the pretended clusters.

From iterations t^2 **to** t^{z-1}**, repeat steps 1 and 2:**

Step 1 - Assign each element i ($1 \leq i \leq n$) to the cluster for which the distance to the medoid is shorter than for the medoids of the remaining clusters.

Step 2 - Re-estimate the clusters' medoids with equation (2.6).

Iteration t^z **-** Stop the process when the cluster constitution is not altered and the medoid elements do not change from the previous iteration.
(Note: z is an iteration resulting from convergence of the process and therefore not fixed from the start.)

2.2.1.2 Hierarchical clustering

The hierarchical method is very applicable and known for its process sequentiality and clustering in multiple stages. The method is defined by two main types of algorithms - divisive and agglomerative - that perform clustering in opposite directions of the hierarchical structure. The divisive method (top-down) performs the division of an initial cluster, constituted by the entire sample size n and respective characteristics, successively into several clusters whereas the bottom-up initially sets each element in a single cluster and performs successive cluster agglomeration. The number of clusters to form is based on a terminal condition defined by the user, otherwise the agglomerative algorithm fuses the entire sample into one cluster and the divisive will produce n number of singleton clusters (figure 2.5) (Everitt et al., 2011; Kaufman & Rousseeuw, 1990a).

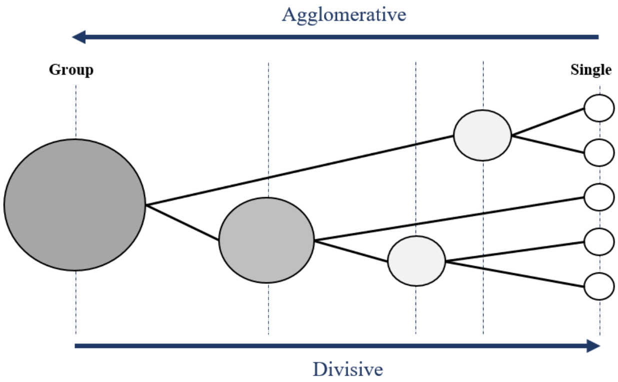


Figure 2.5: Top-down (divisive) and Bottom-up (agglomerative) types of hierarchical clustering.

The hierarchy is based on the construction of sequential nodes, for which the clustering step performed in a node depends on the step made by the previous one, and the formation of the cluster is finalized when all considered steps are made. Clustering relies on the estimation of elements proximity and these are evaluated in both within and between cluster environments, oftenly regarded with dissimilarity measures.

Agglomerative

Agglomeration based algorithms are the most widely used among the hierarchical methods (Everitt et al., 2011). For each method it is defined a linkage function allowing for the evaluation of the pair of closest clusters that are to be joined next. Considering two clusters, I and J: $I \neq J$, for which $i \in I \quad (1 \leq i \leq n_I)$ and $j \in J \quad (1 \leq j \leq n_J)$, table 2.4 compiles some of those functions, of which are described two among the most popular: complete-linkage (furthest-neighbor distance) and average-linkage (unweighted pair-group average method). Having decided the linkage function for the process, the mechanism will successively agglomerate two clusters at a time.

Table 2.4: Linkage functions used for agglomerative hierarchical clustering.

Function	Description	Formula
Single-linkage	Distance between the closest pair of observations in two clusters I and J.	$\min\limits_{i \in I, j \in J} \quad d(i,j)$
Complete-linkage	Distance between the furthest pair of observations in two clusters I and J.	$\max\limits_{i \in I, j \in J} \quad d(i,j)$
Average-linkage	Distance resultant from the sum of all pairwise distances between objects from clusters I and J , and averaged by the total number of distances measured.	$\frac{\sum_{i=1}^{n_I}\sum_{j=1}^{n_J} d(i,j)}{n_I . n_J}$

UNSUPERVISED CLASSIFICATION: AGLOMERATIVE HIERARCHICAL CLUSTERING

Step 0 - The sample space with size n is constituted by n clusters formed by a single element (singleton clusters). The following steps are considered for a desired number of $(n-s)$ final clusters.

Step 1 - Calculate the proximity between clusters and apply the agglomeration function. Agglomerate the pair of clusters for which the linkage dissimilarity is shorter. The number of clusters is now $(n-1)$.

Note: *The linkage function in step 1 equals to the simple measure of proximity between the elements composing the singleton clusters.*

Step 2 - Repeat step 1 to calculate the new between-cluster proximity matrix. The number of clusters after step 2 is $n-2$.

Steps 3 to s - Repeat step 1 until $(n-s)$ clusters have been reached on step s. Finalize the process.

Divisive

Processing in the opposite direction of the agglomerative method, divisive methods are more computationally demanding if all $2^{k-1}-1$ binary divisions are considered for a number of k elements at each stage of cluster division. These methods can be further detailed into monothetic and polythetic methods.

Monothetic methods consider the use of a single variable at each partition similarly like the supervised Decision Tree classifiers 2.2.2.1. The selection of the variable is based on a measure of information content similar to the measures presented in 2.1.2, of which the reader can further explore with Lance & Williams (1968) and Everitt et al. (2011).

The use of multiple variables to base a single division (polythetic) are more similar with the agglomerative method that can similarly use a proximity matrix. The steps are detailed for the method proposed by MacNaughton-Smith et al. (1964) that makes more appealing the use of polythetic methods. The particularity of this method consists in finding the element that is furthest away from the others clustered in the same group (splinter element).

UNSUPERVISED CLASSIFICATION: DIVISIVE HIERARCHICAL CLUSTERING

Step 0 - The sample space with size n is grouped in a single cluster **D**. The following steps are considered for a desired number of $(n - s)$ final clusters.

Step 1 - Calculate the proximity matrix between elements pertaining to the unique cluster. Select the element with the maximum average distance from all other elements to initiate the splinter group **P**. The number of clusters is now two.

Step 2 - For each element now belonging to D, calculate their average distance to **D** and to **P**. If the distance to **P** is shorter than to **D** for a r number of elements, select the one for which $avg.dist_{\mathbf{D}} - avg.dist_{\mathbf{P}}$ is smaller and below zero.

Repeat steps 1 and 2 until the elements are better clustered in either **D** or **P**. The number of clusters after step 2 is $n - 2$.

Steps 3 - When, for all elements in regard, $(avg.dist_{\mathbf{D}} - avg.dist_{\mathbf{P}}) > 0$, the steps 1 and 2 are continued separately in each of the two clusters, and subsequent clusters, until the s number of clusters is reached.

2.2.1.3 Summary

Table 2.5 summarizes the main differences found in the unsupervised methods approached.

Table 2.5: Summary of differences between partitioning and hierarchical unsupervised methods.

Type	**Number of clusters**	**Self-improvement**	**Methods presented**
Partitioning	Defined in the beginning	Reversible at each iteration	K-means, K-medoids
Hierarchical	Evaluated at the end of the clustering process	Sequentiality of the process does not allow reversion	Agglomerative and divisive hierarchical clustering

2.2.2 Supervised methodologies

Supervised methods assume the existence of a class variable and build method-specific computational rules for making such classification on non-classified elements. Among the presented are described probabilistic and non-probabilistic, white-box and black-box types regarded for the supervised classification considering a binary or multiclass outcome variable of interest.

2.2.2.1 Decision Tree

Application simplicity and interpretability are very appealing in a classification problem. By that, the decision tree is a popular method for the construction of a tree shaped classifier.

Given a complex problem, the method constructs a tree shaped classifier that recursively (or *branch* by *branch*) tries to decompose it into simpler problems that can be more easily solved. Decision trees are mostly biparted and always greedy as they evaluate the optimal decision at each step of the process so that the direct child nodes are the purest as possible (more discriminative for a particular outcome class). The term of purity comes with the use of splitting criteria (Gini Index or Entropy) to evaluate the amount of information in the stage previous de split and after the division step 2.1.2. The method constructs a model shaped as a tree of rules. Each tree **T** is constituted by two finite and existing sets of elements:

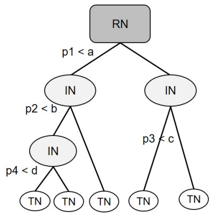


Figure 2.6: Exemplified structure of a Decision Tree.

i. Nodes **N** - each represents a decision made along the tree. There are three types of nodes:

- Root node: is the initial node of the tree constituted by all objects from the sample space; it is split once in the beginning and forms the first two descendent nodes.
- Internal nodes (when existing) - nodes located between the root and the terminal leaves; descendant from a node's division and precedent of two nodes formed by its partition.
- Terminal nodes (leaves) - not split, these are the last nodes formed by the splitting of an internal or root node and are responsible for the final classification of the belonging objects.

ii. Edges **E** - elements connecting a precedent decision (root or internal node) to one of the two following nodes.

These are directed trees, build up from a single node (root) and terminated at least by two leaves. Tree models are visually represented in a dendrogram (figure **??**) (Safavian & Landgrebe, 1991). The length of a tree path from the root to a particular leaf is dictated by the number of edges necessary for that path. In case the tree is extensively branched, reduction by pruning of the leaf nodes can be considered.

SUPERVISED CLASSIFICATION: DECISION TREE

I. Training Steps

Starting condition - Every train element, characterized by p features and classified for a target variable, are grouped in a single cluster (root node) regardless of their classification for the outcome.

Step 1 - The first split is made on the root node based on the selection of the best feature variable and corresponding decision rule evaluated by the splitting criterion.

Note: *In each node partition and for continuous deciding features the rule created for the two descendant nodes is given by two value range intervals that together cover the total value range observed for that deciding variable. For binary features the decision is based on the presence or absence of that characteristic in the elements. For feature variables constituted by more than two classes, the split decision is transformed binary and the decision is made upon the presence of a certain class and the absence of that class (presence of all other classes).*

Step 2 - The tree is biparted in two nodes for which 3 situations are possible: **(i)** the two nodes are terminal (process ends); **(ii)** one node is terminal and one is internal (the splitting continues on the second); **(iii)** the two nodes are internal and therefore both will be split during the process.

Note: *When situations **(ii)** or **(iii)** occur on the first split and on a certain* **d** *number of following splits of the internal nodes, repeat **Step 3**.*

Step 3 - Depending on the number of internal nodes, the splitting criterion estimates the best cut, based on a single feature variables, to take next in the tree. If there is only one internal node in the tree, the criterion estimates the threshold within that variable.

Step 4 - Training process terminates when all paths of the tree have ended in leaf nodes. The first model of classification is originated.

II. Model tuning

Model tuning can be performed by the alteration of specific method's parameters: splitting criteria do use, length of the tree, minimum number of elements in each terminal node, application of pruning steps. ***Step 1 -*** Evaluate the performance of the model and test parameter changing to tune the model in order to get the best performance for the classification of the test elements.

Classification and performance

Step 1 - Implement the classification rules created in the training process (after tuning).

Step 2 - Run the classifier on every test element and evaluate the performance of the decision tree.

2.2.2.2 Ensembling of decision trees

Rokach (2010) describes quite clearly the utilization of ensemble methodology as *"our second nature to seek several opinions before making a crucial decision."*. Ensemble methods perform classification based on the weighted decisions of numerous single methods in the aim to obtain the most tuned model.

Recalling figure 1.1, for the classification of elements a model is created and tuned with a training set of known outcome labeling so to be utilized on the classification of elements with unknown labeling (*”new elements”*). The ensemble methodology considers the application of W methods, each producing a model, that are contributors of the final decision derived from the weighting of the decisions taken by each model W model originated. In other words, ensemble methods ensemble a group of originated models. The ensemble d model requires also a process of tuning related to the number of contributors to consider and other parameters related to the methods that are utilized by the contributors alone.

According to the objectives that are set in the process of classification, the user must weight the gains and losses of considering this group of methods. On one hand, ensemble potentially leads to increased accuracy by creating a stronger model (learner) based on the single contributors, also defined as *”single weak learners”*. But on the other hand, the user is not capable of interpreting the steps justifying the decided classification as the method has lost its transparency (white-box decision tree turned into a black-box ensemble of trees).

Ensemble can be based on various types of single methods (weak learner) but, for being the most exemplified in the literature, here is focused the method that utilizes decision trees as the *single weak elements* (see previous method presented). The most commonly applied types of ensemble of trees are Bagging (**b**ootstrap **agg**regat**ing**) and Boosting (figure 2.7). The two types are summarized.

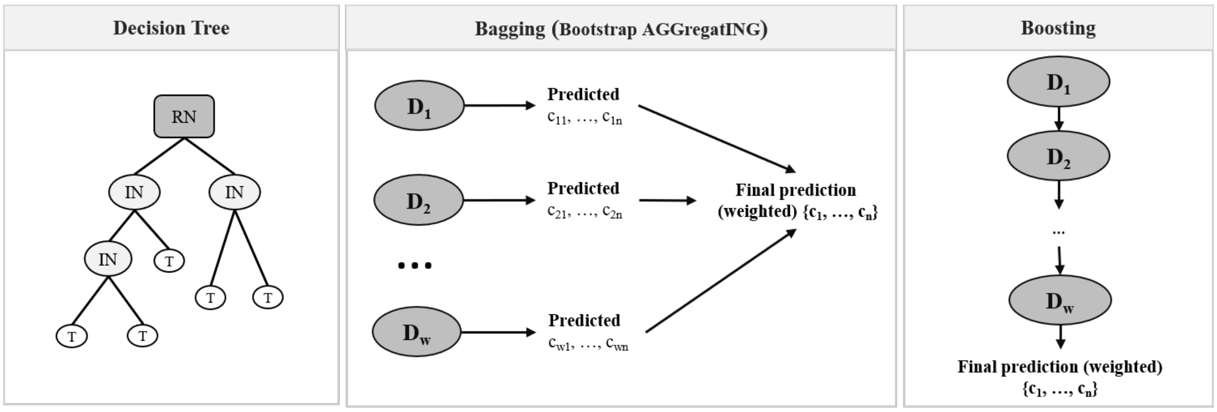


Figure 2.7: Type of ensembling method based on decision tree.

Bagging

Breiman introduced the Bagging method following the consideration that instability of the prediction method is the key element for a better accuracy in the classification process, by allowing the most variability. Being an ensemble, the method considers the use of W single learners and for each a set of training elements is chosen by bootstrapping the original set of data elements with replacement and having a set of test elements to be classified by the constructed model trees. Running in parallel, each single learner is allowed to grow to the most (user can decide on no or little pruning) and at the end produces an independent decision on the classification of the test elements. The weighted final decision on their classification is based on the majority vote of outcome label for each element.
From bagging evolved the method of Random Forest as an upgrade having the same characteristics inherent to the process except for the splitting moment, where not all feature variables from the sample space are considered for splitting evaluation but instead a subgroup of randomly selected variables is considered (increased variability between all the single trees that are grown in parallel).

Boosting

In boosting, the ensemble is structured so that all the single learners are dependent from each other by sequentially creating each following learner based on the accuracy of the decision results from the previous (learn from the previous). Boosting of the ensemble can be considered a stronger method than the bagging since it allows for self-improvement before the final decision is taken rather then running several trees in parallel and deciding by majority vote from those independently grown trees. Therefore, from tree to tree, the train set is decided by random selection with replacement from the elements in the data assigned with higher or lower weights (probability of being selected) if they were, respectively, incorrectly or correctly classified by the previous weak learner.

Table 2.6: Summary table of ensemble methods based on Decision tree classifiers.

Method	**Order of single learners**	**Train set selection**	**Advantages**	**Disadvantages**
Bagging	parallel (independent)	Elements have equal probability (weight) of being selected in boostrap	Variance reduction and variability increase	Bias reduction not taken into account, all single learners have the same importance for the final decision
Boosting	sequential (dependent)	Elements with probability of selection by bootstrap process weighted by the accuracy results from previous weak learners.	Variance reduction and variability increase	Over-fitting problems

2.2.2.3 Naïve Bayes

Naïve Bayes is a probabilistic method that constructs classifiers based on the fundament of the *Bayes Theorem of conditional probability* (Bayes, 1763). For two events A and B, the theorem describes the conditional probability of A given that B has occurred - $f(A = a \mid B = b)$:

$$\begin{aligned} f(A=a \mid B=b) &= \frac{f(A=a \cap B=b)}{f(B=b)} && \textbf{[a]} \\ &= \frac{f(B=b \mid A=a) \times f(A=a)}{f(B=b)} && \textbf{[b]} \\ &\propto \quad f(B=b \mid A=a) \times f(A=a) && \textbf{[c]} \end{aligned} \tag{2.7}$$

The equation is adjusted to a classification problem by expressing $f(A = a \mid B = b)$ as $f(C = c_l \mid X_i = x_i)$. This adaptation describes the conditional probability of the predicted class of the outcome, represented by estimator C, to be c_l given that the vector of values observed in the element i respective to the set of predictor variables, represented by the estimator X_i, equals to x_i. For each element i, the posterior probability of all k possible outcome classes is calculated based on equation (2.7). The evolution of the equation is explained step-by-step into peaces that are easily obtained in the data:

a. The Bayes theorem is applied for the conditional probability;

b. The theorem is again applied on the numerator in [a] as to facilitate the calculation of the joint probability with:

$$P(C = c_l) = \frac{n_{c_l}}{n} \qquad \textbf{\textit{a priori} probability} \tag{2.8}$$

n_{c_l} - number of objects belonging to class c_l
n - total number of objects from sample space

$$f(X_i = x_i \mid C = c_l) \qquad \textbf{density probability} \tag{2.9}$$

c. Only the numerator from [b] is important for the pretended probability because the denominator maintains the same for each object.

To the element i, with a specific vector $(\mathbf{x_i})$, is given a classification c_l for which the proportionality represented by [c] in (2.7) is highest among all the k classes' probabilities (2.10), which is also the one minimizing the classification error.

$$\underset{i \in A=\{1,\ldots,n\}, \quad c_l \in C=\{c_1,\ldots,c_k\}}{\text{classification:}} \qquad \max \quad f(C = c_l \mid X_i = x_i) \tag{2.10}$$

Naïve method

The term *Naïve* relates to the very strong assumption for multivariate data that the p candidate predictors are independent from each other when calculating $f(X_i = x_i \mid C = c_l)$, which in the reality is unlikely the case.

For an object i characterized by a single variable m, $f(X_{im} = x_{im} \mid C = c_l)$ is a mass probability (categorical type) or density probability function (continuous type) considered in the calculation of the probability of pertaining to one class:

$$f(C = c_k \mid X_{im} = x_{im}) \;\propto\; P(C = c_k) \times f(X_{im} = x_{im} \mid C = c_k) \tag{2.11}$$

When p predictor variables are included in the algorithm for classifying a sample of size n, the probability of i belonging to class c_l given $X_i = X = (x_1, \cdots, x_p)$ is:

$$\begin{aligned} &f(C = c_k \mid X_1 = x_1; \cdots; X_p = x_p) = \\ &\qquad = P(C = c_k) \times \frac{f(X_1 = x_1; \cdots; X_p = x_p \mid C = c_k)}{f(X_1 = x_1) \times \cdots \times f(X_p = x_p)} \\ &\qquad \propto\; P(C = c_k) \times f(X_1 = x_1 \mid C = c_k) \times \cdots \times f(X_p = x_p \mid C = c_k) \end{aligned} \tag{2.12}$$

SUPERVISED CLASSIFICATION: NAÏVE BAYES

Step 0 - Provide to the algorithm: the *a priori* probability of each class based on the train set, as in equation (2.8); the mass/density (if categorical/if continuous) probability function for each object $i \in \mathbf{A}$, with feature vector $\mathbf{x}_{i\in\mathbf{A}}$, given it belongs to the class c_l ($\in C = \{c_1, \ldots, c_k\}$).

For each object repeat:

Step 1 - Calculate the probability for each class with equation (2.7).

Step 2 - Decide the object's classification based on equation (2.10).

2.2.2.4 Logistic Regression

The use of logistic regression models is increasing in the medical research field. Logistic regression models are generalized linear models resulting from the extension of linear models. Considering the linear model, for a continuous outcome variable Y taking values in $\mathbb{R}$ it may be simply represented with an inclusion of a single explanatory variable X:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{2.13}$$

where β_0 is the expected value for Y when the observed value of X is $x = 0$, β_1 is the coefficient that determines the effect of that explanatory variable X on the value of Y, and ε a random variable representing the error.

Under the conditions of a classification problem, a logistic regression model must be considered rather than the linear model given the violation of the following assumptions:

[a] Continuous nature of Y

[b] Linear relationship between Y and the parameters β

[c] ε follows a normal distribution with mean $E(\varepsilon) = 0$ and variance $V(\varepsilon) = \sigma^2_\varepsilon$

Binary Logistic Regression

Still considering a single predictor, with a binary response Y the binary logistic regression usually considers the logit adaptation where the logarithm of the odds is modeled:

$$logit = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X, \tag{2.14}$$

where $\ln\left(\frac{p}{1-p}\right) \in \mathbb{R}$ and $p = P(Y = 1)$.

Deriving the direct calculation of the probability of Y=1 and Y=0 (reference level):

$$p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{2.15}$$

The probabilities $P(Y = y)$, $Y = \{0, 1\}$, are modeled. The base definition of logistic regression is for the study of a binary outcome variable. The simple representation of the binary logistic regression can be made with a binary outcome variable Y and a single explanatory variable X. The inclusion of more explanatory variables, say p, follows in a straightforward way, extending the systematic component of the model from $\beta_0 + \beta_1 X$ to $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.

For the study of an outcome variable with more than two classes it is required the use of more complex methods that are extensions of the binary logistic regression. Two types of extensions are identified based on the relationship between the classes of the outcome variable: if there is order among the classes the appropriate method is ordinal, otherwise multinomial. Multinomial logistic regression can be understood as the combination of $c - 1$ binomial logistic regressions (where c is the number of outcome classes). Likewise the binomial model, the coefficients are important parameters for the interpretation of the classification process.

Importance and interpretation of coefficients

Statistical significance is analyzed based on the p-value to understand the effect on the class a being analyzed compared to the baseline class being used on the logit function for class a.

For the comparisons with significant effect difference it is interesting to interpret the value for that difference, based on the Odds Ratio (OR).

The OR calculation below indicates which outcome is being compared to the reference class ($Y = 0$). To calculate the odds of being in class l with predictor value $x = a$ compared to the odds of being in that same class with $x = b$.

$$OR_l(a, b) = \frac{P(Y = l|x = a) \big/ P(Y = 0|x = a)}{P(Y = l|x = b) \big/ P(Y = 0|x = b)} \tag{2.16}$$

For a binary predictor, the values to compare are $x = 0$ (reference category) and $x = 1$.

Ordinal Logistic regression

The focus of this dissertation is on the extension of the binomial model regarded for ordered multiclass response, for which the ordinal logistic regression method (proportional odds model) is considered.
An ordered response variable implies the existence of an ordered structure and relationship between the response

classes. With ordinal logistic regression, this relationship and the estimation of the odds is calculated in terms of cumulative probabilities.

$$\text{logit}[P(Y \leq k)] = \ln\left(\frac{P(Y \leq k)}{P(Y > k)}\right)) = \alpha_k - \beta_1 x_1 - \cdots - \beta_p x_p \tag{2.17}$$

For a total of c response classes $c - 1$ cumulative logit equations are considered. Each cumulative logit has its own intercept α_k $(1 \leq k \leq c-1)$ that is orderly related with the other intercepts $(\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{c-1})$ according to the order of the k response class. As for the effect of each explanatory variable, the respective β coefficients are constant for every cumulative logit equation regardless of the category k considered. Subsequently, the relation between the vector of explanatory variables and the outcome response does not depend on the category of the response (Mccullagh, 1980; Abreu *et al.*, 2008).

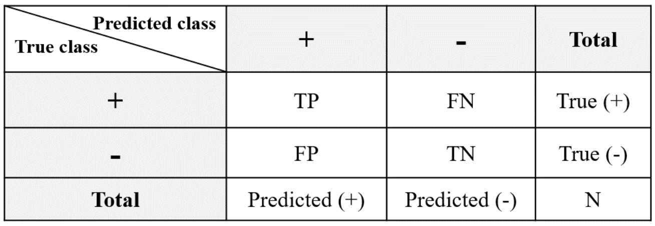
The $(c - 1)$ models are parallel linear equations that transform into a binomial logistic regression for the particular case of a binary response type. Given the characteristics of the model, it is required the validation of the proportional odds assumption to all explanatory variables included. In `R`, the functions `poTest` from `car` package and `brant` from `brant` package perform the test proposed by Brant (Brant, 1990) for the validation of the proportional odds assumption applied to an ordinal logistic regression model constructed with the `polr` function from `MASS` package.

If the assumption is not validated for at least one of the explanatory variables, less strict alternative regression methods, as the partial proportional odds models, are available for the use of regression methods on ordered response variables.

2.3 Performance evaluation

The performance of a classifier is evaluated on how well it behaved in the task of element classification. An ideal classifier would make no error on the classification of elements.
After training a learning algorithm is important to analyze its capability to use the characterizing variables for predicting the classification for a desired target variable. The evaluation of prediction performance takes into consideration important definitions. In the scenario of binary response, one class is defined “positive” and the other “negative” following the meaning these definitions have in medicine to assign, respectively, the class of presence and absence of a disease when performing this type of analysis. The primary interpretation relates to the count analysis of correct and incorrect classifications 2.8:

- True Positive - real positives correctly classified
- False Positive - real positives incorrectly classified
- True Negative - real negatives correctly classified
- False Negative - real negatives incorrectly classified



Predicted class True class	+	-	Total
+	TP	FN	True (+)
-	FP	TN	True (-)
Total	Predicted (+)	Predicted (-)	N

Figure 2.8: Confusion matrix (2x2) for binary response variables.

In table 2.7 are presented a set of performance measures for the classification with binary outcome. From those, the Geometric Mean (GM) and Matthews Correlation Coefficient (MCC) may give more adjusted interpretations for unbalanced response classes than the much utilized measure of Accuracy (ACC). Another measure for unbalanced data is the Balanced Accuracy, for which different weights are given to the sensitivity and specificity (w_{SE} and w_{SP}: $w_{SE} + w_{SP} = 1$. However, the decision must be thoroughly made since importance weights will be given to the outcome classes. Another measure highlighted is the Dice coefficient. This measure evaluates the matches between the true and predicted classifications represented in a confusion matrix (**CM**) and gives values of importance for each cell, originating a relevancy matrix **C**.

Table 2.7: Measures of performance for binary classification.

Measure	**Formula**
Sensitivity	$\frac{TP}{P}$, where $P = TP + FN$
Specificity	$\frac{TN}{N}$, where $N = TN + FP$
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{P+N}$
Balanced accuracy	$w_{SE}SE + w_{SP}SP$
Geometric Mean	$\sqrt{SP \times SE}$
Matthews Correlation Coefficient	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Dice coefficient	$\frac{\mathbf{C} \times \mathbf{CM}}{c}$
Index of Youden	(Sensitivity + Specificity) $- 1$

For multiclass outcome variables, a common practice consists on the averaged measures presented in table 2.7, were k scenarios of binary classification are created for the k number of outcome classes and those evaluated the presence of the *k-th* class *vs* its absence. In this way, the measures applied for binary classification can be calculated for every outcome class and the overall value of the measure is the average (Sokolova & Lapalme, 2009).

2.4 Data preparation

A prior preparation of the data is essential before stepping into deeper analysis as the case of the application of classification methodologies. In this step, the user gets acquainted with the data as he/she becomes able to identify the structural composition and the most important properties, which may indicate the necessity of preparation/adjustment, and ultimately allow for the selection of the most appropriate methods to apply.

2.4.1 Acquaint with the data

The first acquaintance of the data is the identification of its composition: number of elements in the sample, number of characterizing variables observed in those elements (dimensionality), which variable/s represent an outcome of interest (identification depends on the objective of the analysis, *e.g.* for classification this variable is identified) and which represent candidate predictors for the prediction of the identified outcome variable (classification) or for studying the relations naturally found between the elements (clustering).

To follow, the nature of the variables is defined (figure 2.9). This task is fundamental to have an insight on which type of methods should be decided for application based on their adequacy to the data.

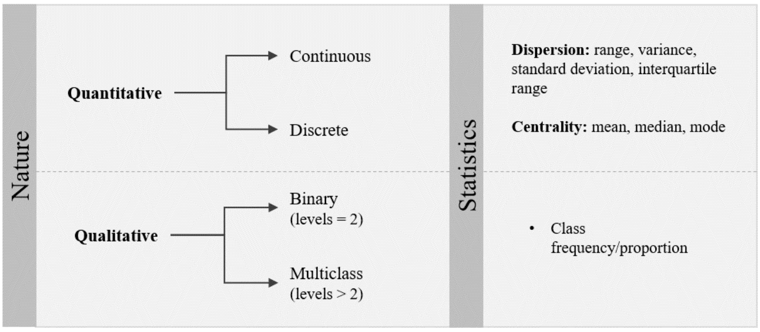

Nature
Quantitative
Continuous
Discrete
Qualitative
Binary (levels = 2)
Multiclass (levels > 2)
Statistics
Dispersion: range, variance, standard deviation, interquartile range
Centrality: mean, median, mode
• Class frequency/proportion

Figure 2.9: Initial acquaintance with the data.

The nature of the variables in the data data can be qualitative, if the characteristic is measured in terms of possible classes *e.g.* the color of the eyes, or qualitative if the characteristics is measured in a continuous space *e.g.* person's age or height. The calculation of summarizing statistics serves to understand important details of the distribution of the characteristics in the sample.

For an identified qualitative outcome variable further univariate hypothesis testing is recommendable to study the association found between the outcome and each of the characteristics measured.

2.4.1.1 Univariate discriminative potential

The analysis of univariate variable potential to discriminate the classes of outcome is also a interest for classification. The three following techniques exemplify possible methods for such analysis: hypothesis test p-value, Area Under the ROC Curve (empirical AUC) and Entropy.

i. Hypothesis test p-value

This value is obtained by the application of univariate statistical tests on grouped data to evaluate two proposed hypothesis generally described as follows:

- Null hypothesis (H_0): there is no statistically significant difference of the predictor distribution among the classes of the outcome variable.
- Alternative hypothesis (H_1): the distribution of the predictor variable is statistically different between the classes of the outcome variable.

The p-value returned is considered to analyze the discriminative potential of the specific predictor. Being an estimated probability defined in the interval [0,1], p-value represents the estimated probability that the difference described by H_1 would be observed assuming H_0 is true (thus ranging in the interval [0,1]). The lower the estimated value, the closer (or inside) it is from the region of H_0 rejection (marked by a threshold significance level represented by α), and therefore the variable has more potential to discriminate the response classes for the sample tested.

ii. Area Under the Curve

The Area Under the Curve (AUC) is a metric much utilized in the evaluation of the performance on binary classification (classes are described as presence/positive and absence/negative) and therefore a possible choice for analyzing discriminative potential. The AUC is calculated under the ROC curve (Receiver Operating Characteristic) obtained from plotting two probabilities:

- True positive rate (sensitivity): represents the probability that a true positive will be predicted as positive.
- True negative rate (1-specificity): represents the probability that a true negative will be predicted as negative.

The empirical calculation of the AUC for a single predictor variable uses the same approach as the calculation of the entropy described for quantitative variables in figure 2.3. The values for AUC range in the interval [0,1], in which for an $AUC = 0.5$ the capability to predict one class or another is the same (50%), for $0 \leq AUC < 0.5$ the classifier is being used in the opposite direction (predictions are opposite to the true labels), and for $0.5 > AUC \geq 1$ the polarity of prediction is according to the true labels and a closer value to 1 becomes closer to the perfect capacity to discriminate classes (Hand & Till, 2001).
In the scenario of a multiclass outcome variable it is suggested the calculation of an averaged AUC from all the estimated AUCs when dichothomizing the problem for each k outcome class (presence of k *vs* absence of k) so that a number of s AUC estimations for pair are made:

$$AUC_{multiclass} = \frac{2}{s(s-1)} \times \sum_{pair=1}^{s} AUC_{pair} \tag{2.18}$$

iii. Entropy

The entropy is explored in section 2.1.2. The evaluation of variables with an higher discrimination potential is determined by the lowest estimated entropy.

2.4.2 Imputation mechanisms

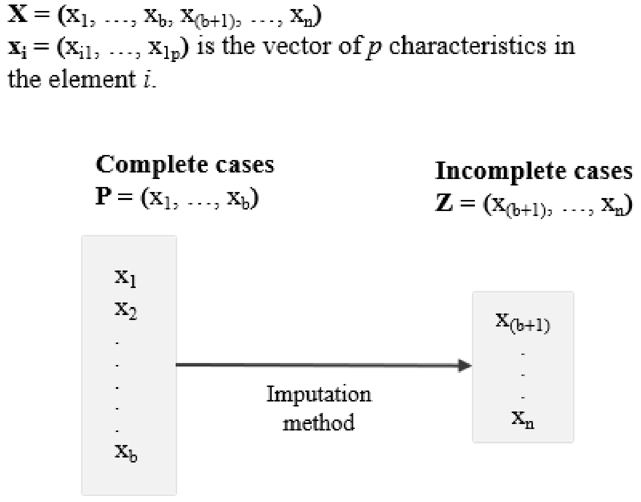
The assessment of omission in the data is crucial for studies in pattern recognition. Omission signifies that a given element in the data may not be completely observed and therefore there are missing values for the variables that are used to find patterns with unsupervised methodologies or create the structure for classification of new elements with supervised methodologies.
The types of omission are described below and interpreted in the context of the probability of a missing value "$P(missing)$" and for a given variable $X_{missing}$ for which a missing value occurs:

- **Missing completely at random (MCAR) –** $P(missing)$ is not related either with $X_{missing}$ or any other variable considered in the data. There is non-influence of the data for its occurrence.

- **Missing at random (MAR) –** $P(missing)$ is related to other variables but not with $X_{missing}$ itself.

- **Missing not at random (MNAR) –** $P(missing)$ is related to the variable with the missing values itself and the other variables in the data.

For a dataset with all complete cases, missing values can be dealt with by either dropping the incomplete ones or replace the missing with values estimated from statistical analysis based on the other elements of the data (Hegde et al., 2019). The first option may be more comfortable to use but it might mean a possible reduction of valuable information in the data. If considering the replacement of missing values (imputation), this mechanism requires some level of randomness in the event of omission in the data, thus being specifically utilized for MCAR and MAR types of omission. In table 2.8 follows a summary of imputation methods.



$\mathbf{X} = (x_1, \ldots, x_b, x_{(b+1)}, \ldots, x_n)$
$\mathbf{x_i} = (x_{i1}, \ldots, x_{1p})$ is the vector of p characteristics in the element i.

Figure 2.10: Mechanism of imputation.

Table 2.8: Table summary of imputation methods.

Type	Imputation method	Description	Advantages	Disadvantages
Simple imputation	Mean	Univariate analysis by replacement of the missing value with the mean of the values observed for that variable.	Simple to interpret and easy to implement.	The important relations that the variable may have with the other variables observed is disregarded for that element. Not suitable for large imputation procedures.
	Nearest-neighbor	Analyses the k nearest neighbors based on distance measures and calculates a value of centrality (mean or media) of the values of the k neighbors for the variable having the missing value.	Relation between variables is accounted for.	Not suitable for large imputation procedures.
Multiple imputation	Mice	Performs multivariate imputation by chained equations by running a series of regression models to predict the missing values conditional to the variables observed. The missing value is filled in multiple times creating multiple examples of complete datasets.	Adapted to large imputation procedures, the estimations from multiple imputation are less biased. Capable of handling diverse variable natures .	Higher complexity on the implementation of this technique (specially limiting for researchers related to field less applied to statistics).

2.4.3 Outlier detection

Awareness must be raised for the influence of outliers in the data as these can affect the application of methods for pattern recognition. An outlier may be defined as “an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism” (Hawkins, 1980) so that their behavior is very different from the majority of the elements present.

Outlying observations can be misleaders on the utilization of the methods as they may create sample-dependent results on a pattern search or on the creation of classification rules. Being much encountered in medical case studies, the presence of outliers is most frequent in the scenario of small sample sizes and, specifying the existence of an outcome variable (classification), weak differences in between the elements of different classes that prevent the broadening of the results to new elements (Ferrari et al., 2020; Osborne & Overbay, 2004).

Outliers can be from either two natures: (1) natural outliers existing in the environment (novelties in the data) or (2) artificial outliers originating from *e.g.* human errors during data entry or instrumental errors during measurement. Additionally, they can also be classified according to the dimensional space on which they are being detected: (1) univariate outliers when the detection of the same occurs for the analysis of a single variable or (2) multivariate outliers when the space in which the element is detected concerns the relation of multiple variables characterizing it. Given this, an observation may be considered a candidate outlier in the univariate space but not in the multivariate space (the other way around also applies).

When dealing with these deviating observations, the most common procedure considers their removal from the data or, alternatively, the specification of a weight (lower than for no candidate outliers) for the outlying observation. The diversity of methods for univariate and multivariate outlier detection includes statistical, clustering and distance-based mechanisms, those of which the user can have a deeper insight in Acuña & Rodriguez (2014) study.

2.4.4 Variable selection

The increasing number of candidate predictors (the curse of dimensionality) may pose a challenge for the study of patterns in the data. The most striking reasons for reducing dimensionality include (1) the exclusion of irrelevant data, (2) reduced time consumption, (3) development of the best model fit for the data and (4) reduced complexity of result interpretation (Andrews & McNicholas, 2014; L. Yu & Liu, 2003).

The measure of correlation has an important role for feature selection, well described by Yu & Liu (2003): $''$…if the correlation between a feature and the' ' outcome "class is high enough to make it relevant to … the class and the correlation between it and any other relevant features does not reach a level so that it can be predicted by any of the other relevant features, it will be regarded as a good feature for the classification task' '. A similar interpretation can be made for clustering tasks by considering the selection of the variables that are most active in cluster formation (Andrews & McNicholas, 2014).

In the scenario of classification or acknowledgement of the existence of an outcome variable, this selection can be made univariately upon the analysis of individual variable potential to discriminate the outcome and following selection of a top list of variables with higher estimated potential, such as the example techniques previously presented (p-value, empirical AUC and entropy). The foreseen disadvantage of this selection technique is the complete disregard of the relations between all the candidate predictor variables leading to the selection of variables that individually may have potential as predictors of the outcome, but together they may have not the same strength or may even be worst for prediction.

A much correct selection for multivariate data would be the one considering the relation between variables. For that, a recent method of variable selection introduced by Andrews & McNicholas (2014), with a very efficient and broader application for both clustering and classification methods, is presented.

The method of variable selection for clustering and classification is based in two important concepts: the within-group variance W and between-variable correlation ρ. The relation between these two concepts determines the variables that may be selected and is based on the following equation:

$$|\rho_{jr}| = 1 - W_j^d, \tag{2.19}$$

where j represents the variable that is being analyzed for entering or not in the group of variables selected V, r is a variable already belonging to those selected ($r \in V$), and d is the degree of the relationship ($1 \leq d \leq 5$).

The relation is defined so that more importance and approval is given to smaller values of W. Additionally, higher degree levels are more accepting for larger absolute values of ρ as represented by the following figure 2.11 extracted from Andrews & McNicholas (2014).

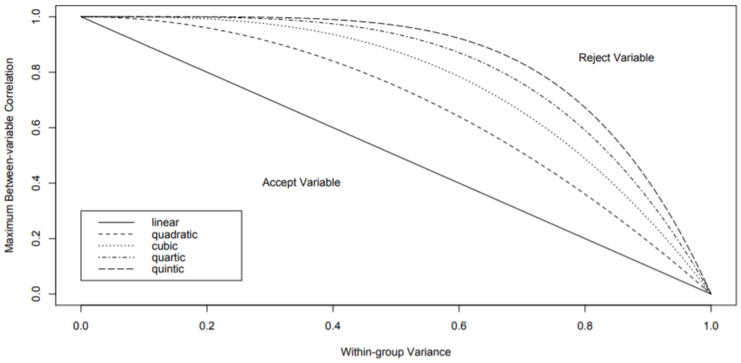


Figure 2.11: Values of acceptance according to the variance-correlation relationship degree (Andrews and McNicholas, 2014).

The method can be applied by the `vscc` (package `vscc`) in `R`, with the steps described below:

- V – space of currently selected variables
- r – represents a variable in space V
- C – space of variables candidate for selection
- j – represents a variable in space C

Step 0 - all variables are candidate for selection (V space is empty).

Step 1 - calculate the within-group variance for all j variables (W_j).

Step 2 - sort in ascending order the list of W_j and select the first variable l: W_l is the minimum of W_j $(1 \leq j \leq p)$. $V = \{W_l\}, C = \{W_1, \dots, W_p\} \quad \{W_l\}$

Step 3 - To evaluate the remaining j candidate variables, consider the evaluation of equation (2.19) between the j candidate being evaluated down the ordered list and all the r variables already selected. For the second selection, the first variable evaluated is the one positioned below the variable l that was just selected.

The application of this method implies an initial grouping of the elements so to evaluate the within-group variance for each candidate variables. If the selection is meant for clustering, the `vscc` function allows clustering of the data regarding the use of `mclust` or `teigen` methods that respectively perform Gaussian finite mixture models fitted via EM algorithm for model-based clustering, and model-based clustering and classification with the multivariate t distribution. For classification, the initial data groups can be specified as the true classes for the outcome variable being studied.

2.5 Summary

Every content aborded in this document regards important considerations when studying pattern recognition. Serving as guideline for the user, the contents here presented and their sequentiality in the process are summarized in scheme 2.12.

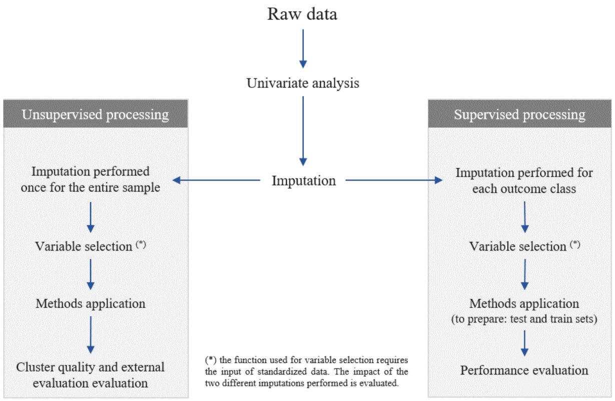

Raw data
Univariate analysis
Unsupervised processing
Supervised processing
Imputation performed once for the entire sample
Imputation
Imputation performed for each outcome class
Variable selection (*)
Variable selection (*)
Methods application
Methods application (to prepare: test and train sets)
Cluster quality and external evaluation evaluation
Performance evaluation
(*) the function used for variable selection requires the input of standardized data. The impact of the two different imputations performed is evaluated.

Figure 2.12: Schematization of the process of analysis for the aplication of pattern recognition methods.

Chapter 3: Data Analysis

The application of the classification methods to the case study of Obstructive Sleep Apnea is presented along this chapter.

Methodological note: Sample Collection & Biochemical Analysis

Patients with suspected OSA are hospitalized for an overnight polysomnography (PSG) study. Blood samples were collected into EDTA-coated polypropylene tubes before and after PSG, *i.e.*, between 8:00 pm and 09:30 pm (referred as ’evening’) and between 7:30 am and 09:00 am (referred as ’morning’). For proteomics study, samples were kept no longer than 4 hours at 4ºC until fractionation by centrifugation. Study was approved by Ethical committee and all participants were signed informed consent.

For the purpose of studying the diagnosis of OSA, a total of 69 male individuals with suspected presence of the disturbance were measured. Recommended to undergo the Polysomnnography, the individuals were monitored along the test for multiple parameters related to sleep quality, corporal movements and oxygen levels to assist the diagnosis of OSA primarily based on the “Respiratory Disturbance Index” (RDI). The RDI index measures the per hour sleep rate number of Apneas and Hypopneas (“Apnea-Hypopnea Index”) combined with the count of Respiratory Effort- Related Arousals (RERAs) for determining the diagnosis (Berry et al., 2012). The guidelines for the definitive diagnosis are described in Table 3.1.

Table 3.1: Criteria for the diagnosis of Obstructive Sleep Apnea.

Outcome class	**Diagnosis description**	**Index criteria**
Control	Not affected	RDI < 5
OSA-I	Mild stage	$5 \leq$ RDI < 15
OSA-II	Moderate stage	$15 \leq$ RDI < 30
OSA-III	Severe stage	RDI ≥ 30

All the parameters evaluated during the test are highly associated to the response and therefore those are not the focus of this study. Along with these, other three differentiated sets of characteristics were measured:

- **(i)** Quantitative expression of 22 proteoforms (16 Apo of family C and 6 Apo of family A2) from blood samples collected in two time periods: evening before the test and morning after. For this study, the evening expression and net expression between morning and evening (variation $= variable_{morning} - variable_{evening}$) were considered (appendix table B.3).
- **(ii)** Anthropometry, clinical history and lifestyle habits having a documented association to OSA (one-time measurement). From those are highlighted the individual’s increasing age, male gender and the presence of

metabolic disorders, particularly obesity estimated to affect 60 to 70% of the population positively diagnosed with OSA (Lurie, 2011; Coughlin et al., 2004) (appendix tables B.1 and B.2).

(iii) One-time measurement of clinical variables analyzed from blood and urine samples collected in the morning after the PSG examination (appendix tables B.1 and B.2).

3.1 Data preparation

3.1.1 Acquaintance and univariate analysis

The outcome variable Y is defined ordinal and it represents the diagnosis of the disturbance analyzed in two different definitions of the constituting classes:

- Y_2 - binary variable representing the absence *vs* presence of OSA;
- Y_4 - multiclass variable representing the absence of OSA or presence differentiated into the respective severity stages (table 3.1).

The imbalanced distribution of the outcome classes is evidenced. The distribution disparity is particularly stronger in the scenario of the outcome variable Y_2, where the observed class OSA represents 76.8% of the total sample size (table 3.2).

Table 3.2: Observed distribution of the outcome classes (binary and multiclass defined).

	Frequency (%)	
Outcome class	**Binary - Y_2**	**Multiclass - Y_4**
C	16 (23.2)	16 (23.2)
OSA-I	53 (76.8)	27 (39.1)
OSA-II		10 (14.5)
OSA-III		16 (23.2)

A total of $p = 76$ characterizing variables were measured. The quantitative nature is the most dominant (68 quantitative variables; 8 qualitative variables).
Measures of centrality (mean and median), dispersion (minimum, maximum, variance, standard deviation and quantiles) and data omission were estimated for quantitative variables. By visually analyzing the descriptives, it is noted the great difference in the order of magnitude of the values observed for some of the variables representing the proteoform expression (*e.g.* disparity in the order of magnitude between the values observed for variable `EA2DQ` and variable `EA2M` observed in Table C.4 from the appendix C). Descriptive univariate analysis was followed by hypothesis testing. To test the null hypothesis H_0 of statistically no significant difference on the median and shape distribution of the values observed on the explanatory variables along the classes of the outcome was performed with Mann-Whitney U test for the outcome Y_2 and Kruskal-Wallis test for the outcome Y_4 followed by Dunn's test for multiple pairwise comparison (Mann-Whitney and Kruskal-Wallis tests: $\alpha = 0.1$; Dunn's test: $\alpha = 0.05$). The use of non-parametric hypothesis tests is justified by the small size of the overall sample and the imbalanced distribution of the outcome classes.

For the clinical variables `abdo.perim.cm`, `cerv.perim.cm` and `bmi` there is a significant statistical difference in their observed values among the classes of the outcome Y_2. The same was observed for variables `abdo.perim.cm`, `cerv.perim.cm`, `bmi`, `trigl`, `glyc` and `hdl` regarding the outcome variable Y_4 (Table 3.3).

Table 3.3: Summary table of hypothesis test results for clinical variables.

	Mann-Whitney U test			Kruskall-Wallis	
Variable	Test statistic	p-value	Variable	Test statistic	p-value
abdo.perim.cm	234	0.020 **	cerv.perim.cm	13.567436	0.004 **
bmi	272	0.031 **	abdo.perim.cm	13.073025	0.004 **
cerv.perim.cm	251.5	0.040 **	bmi	11.473788	0.009 **
			trigl	8.273978	0.041 **
			glyc	8.146113	0.043 **
			hdl	6.494978	0.090 *

As for the variables representing proteoform expression, only the variables representing expression variation for Apolipoproteins A2 (`dfA2DQ`, `dfA2D2Q` and `dfA2M`) presented a significant statistical difference in the distribution of the observed values among both the outcomes Y_2 and Y_4 (Table 3.4).

Table 3.4: Summary table of hypothesis test results for proteoform expression.

	Mann-Whitney U test			Kruskall-Wallis	
Variable	Test statistic	p-value	Variable	Test statistic	p-value
dfA2DQ	175	0.001 **	dfA2DQ	11.5	0.009 **
dfA2D2Q	262	0.055 *	dfA2M	7.7	0.053 *
dfA2M	506	0.082 *	dfA2D2Q	7.4	0.060 *

For qualitative variables, frequency tables were analyzed and the Chi-Square test was applied to test the independence between the outcome class and each of the characterizing variables. The Fisher exact-test was considered instead of Chi-Square when the following conditions were found: (i) presence of zero count cells and (ii) $>$ 20% of the cells with count below 5 in the $c \times l$ contingency table for c number of outcome classes and l number of classes of the characterizing variable. For a significance value $\alpha = 0.1$, only the distribution of `tft` (p-value = 0.05) is significantly different among the classes of the binary outcome Y_2. No missing values were observed on these variables.

Table 3.5: Summary table of hypothesis test results for qualitative variables.
† : A test statistic is not produced with the Fisher-exact test.

	Binary (Y2)			Multiclass (Y4)	
Variable	Test statistic	p-value		Test statistic	p-value
tft	†	0.06	*	†	0.32
smoking.habits	2.11	0.32		†	0.51
morn.head	0.07	1.00		0.23	1.00
awakenings	1.35	0.37		†	0.75
card.path	0.99	0.40		1.26	0.72
resp.path	†	0.42		†	0.65
metab.path	1.38	0.25		1.43	0.74
endoc.path	†	0.66		†	0.89

3.1.1.1 Discriminative potential

The discriminative potential of each characterizing variable was analyzed with three methods for the quantitative variables (p-value, empirical AUC and entropy) and two methods for the qualitative (p-value and entropy). Overall, when both natures apply, the quantitative variables present an higher discriminative potential than the qualitative for both Y_2 and Y_4 (Tables C.9 to C.14).

The top 10 rank of quantitative variables considering the p-value and entropy methods show the presence of the same variables with the same ranks (Table 3.6). Comparing the top 10 obtained with the empirical AUC and the other two, the first shows some differences in the position of the variables ranked. For all methods, the variable `dfA2DQ` assumes the first position as the variable with higher potential to discriminate the classes of Y_2.

Table 3.6: Top 10 rank of quantitative variables with higher potential to discriminate the outcome classes of Y_2. Methods used: p-value, empirical AUC and Entropy.

Outcome variable: Y_2					
Variable	p-value (rank)	Variable	Entropy (rank)	Variable	AUC (rank)
dfA2DQ	1	dfA2DQ	1	dfA2DQ	1
abdo.perim.cm	2	EA2DQ	2	abdo.perim.cm	2
bmi	3	dfA2D2Q	3	bmi	3
cerv.perim.cm	4	dfC3v5	4	cerv.perim.cm	4
dfA2D2Q	5	age	5	dfA2D2Q	5
dfA2M	6	dfA2M	6	dfA2M	6
EA2DQ	7	bmi	7	EA2DQ	7
adren.u	8	abdo.perim.cm	8	adren.u	8
EC3v7	9	dfC3v8	9	EC3v7	9
age	10	adren.u	10	age	10

For the top 10 obtained for the discrimination of Y_4, the importance of the variables is altered compared to the ranks obtained for Y_2. The majority of the variables are identified in the three top ranks but in different positions. Additionally, for all three methods the variable `insul` presented the highest discriminative potential, always followed by variable `homa.ir` (Table 3.7).

Table 3.7: Top 10 rank of quantitative variables with higher potential to discriminate the outcome classes of Y_4. Methods used: p-value, empirical AUC and Entropy.

Outcome variable: Y_4					
Variable	p-value (rank)	Variable	Entropy (rank)	Variable	AUC (rank)
insul	1	insul	1	insul	1
homa.ir	2	homa.ir	2	homa.ir	2
cerv.perim.cm	3	bmi	3	abdo.perim.cm	3
abdo.perim.cm	4	abdo.perim.cm	4	dfA2DQ	4
bmi	5	dfA2DQ	5	cerv.perim.cm	5
dfA2DQ	6	cerv.perim.cm	6	bmi	6
trigl	7	trigl	7	glyc	7
glyc	8	dfA2D2Q	8	dfA2D2Q	8
dfA2M	9	adren.u	9	dfA2M	9
dfA2D2Q	10	sis.bp	10	trigl	10

The discriminative potential of qualitative variables highlights the variables `metab.path` and `smoking.habits` in both top 3 rank of variables assessed with the p-value and entropy methods for the outcome Y_2. For the evaluation of the discrimination of Y_4, `smoking.habits` is the only variable appearing consistently among the top 3 assessed with the methods mentioned. For those variables with entropy value absent due to the available sample and distribution that did not allow for the estimation of the entropy.

Table 3.8: Rank of the qualitative variables potential to discriminate the outcome classes Y_2 and Y_4. Methods used: p-value and entropy.

Outcome variable: Y_2			Outcome variable: Y_4		
Variable	p-value (rank)	Entropy (rank)	Variable	p-value (rank)	Entropy (rank)
tft	1	-	tft	1	-
metab.path	2	2	smoking.habits	2	1
smoking.habits	3	1	resp.path	3	-
awakenings	4	3	card.path	4	4
card.path	5	4	metab.path	5	3
resp.path	6	-	awakenings	6	2
endoc.path	7	5	endoc.path	7	5
morn.head	8	6	morn.head	8	6

3.1.2 Data imputation

In regards to the quantitative variables, a previous process of imputation was necessary for replacing the missing values found in those variables. Considering the previously presented methods for imputation in section 2.4.2, the K nearest neighbors was considered since it may return the best compromise between simplicity and quality of the imputed data. The `kNN` function from `VIM` package was applied in `R`. The code below exemplifies the execution of such process in `R`.

```
#### Imputation ####

# weights: equal weights for all variables (weight = 1)
# dist_vars: variables used to find the neighbors (quantitative + qualitative)
# k: number of neighbors to base imputation
# numFun: function to find the k neighbors to replace a quantitative missing value
  # median replaces the missing by the median value of the neighbors for that variable
# catFUN: function to find the k neighbors to replace a qualitative missing value
  # maxCat replaces the missing with the majority class in the neighbors
  # line commented because there are no qualitative data missing

VIM::kNN(data ,
        weights = rep(1,length(dist_vars)),
        dist_var = dist_vars,
        k= 5,
        numFun = median
        #catFUN = maxCat
        )
```

Recalling Figure 2.10, the method of imputation was differentiated for unsupervised and supervised classifications. The unsupervised imputation made no distinction of outcome classes, while the supervised imputation considered the imputation per class of the outcome (Figure 3.1). From this process were derived three imputed datasets: one dataset for unsupervised classification applied to both binary and multiclass cluster formation (D_{iu}), one dataset for supervised classification applied to binary classification (D_{is2}) and one dataset for supervised classification applied to multiclass classification (D_{is4}).

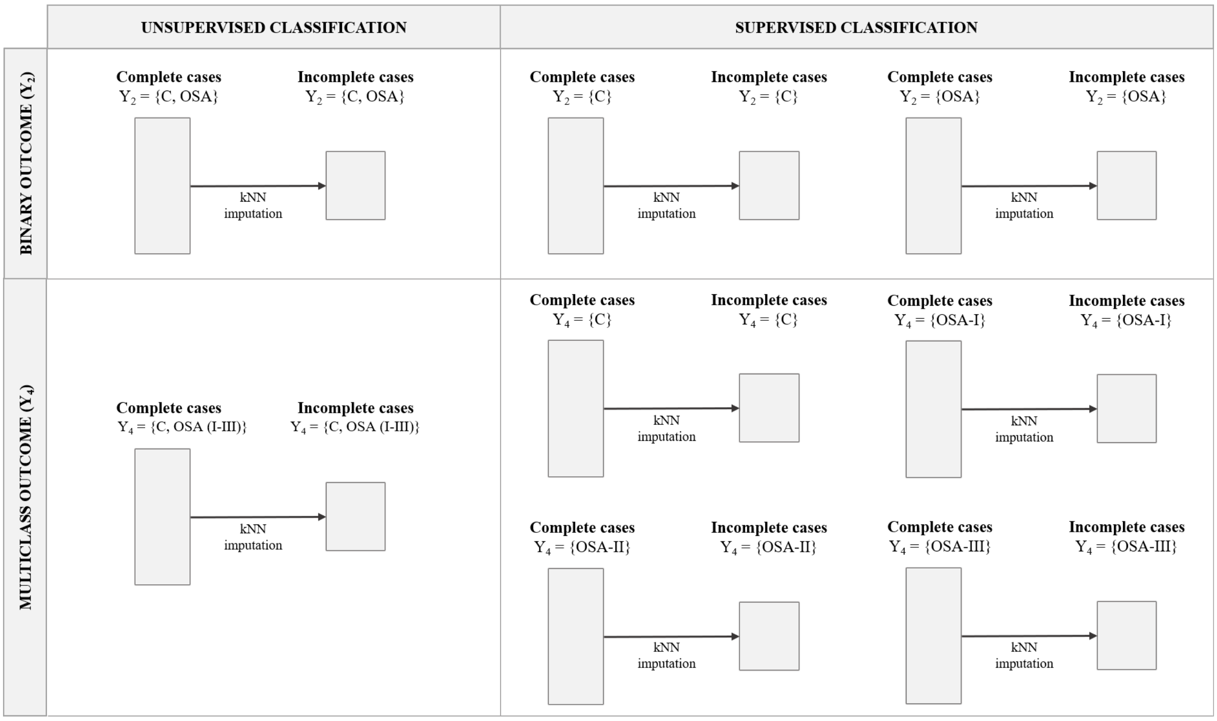

UNSUPERVISED CLASSIFICATION
SUPERVISED CLASSIFICATION
BINARY OUTCOME (Y_2)
MULTICLASS OUTCOME (Y_4)
Complete cases Y_2 = {C, OSA}
Incomplete cases Y_2 = {C, OSA}
kNN imputation
Complete cases Y_2 = {C}
Incomplete cases Y_2 = {C}
Complete cases Y_2 = {OSA}
Incomplete cases Y_2 = {OSA}
Complete cases Y_4 = {C, OSA (I-III)}
Incomplete cases Y_4 = {C, OSA (I-III)}
Complete cases Y_4 = {C}
Incomplete cases Y_4 = {C}
Complete cases Y_4 = {OSA-I}
Incomplete cases Y_4 = {OSA-I}
Complete cases Y_4 = {OSA-II}
Incomplete cases Y_4 = {OSA-II}
Complete cases Y_4 = {OSA-III}
Incomplete cases Y_4 = {OSA-III}

Figure 3.1: Application procedure of the k nearest neighbor for the imputation of missing data observed in quantitative variables of the case study.

3.1.3 Variable selection

The dimensionality of the data and its comparison to the sample size (76 variables observed on 69 individuals) required the use of variable selection techniques.

The selection of quantitative variables was considered by the use of the multivariate selection technique presented by Andrews & McNicholas (2014) and performed with function `vscc` available in `R` (recall section 2.4.4). For the purpose of clustering, the function `mclust` was selected to provide the data groups of the imputed dataset D_{iu}, which are required to start the process of selection (two or four groups according to the number of clusters that are being posteriorly searched during unsupervised classification). For supervised classification, the initial groups given for the process correspond to the true classes of the elements classification for OSA: Y_2 for two groups selection based on the imputed data D_{is2} and Y_4 for four groups selection based on the imputed data D_{is4}. The variables selected for clustering into both two and four groups are observed to be greatly different, in all degrees of variance-correlation relationship, from those selected for supervised classification (Tables 3.9 and 3.10). In terms of their position in the discriminative potential ranks, no variable selected for clustering is presented in any of the top 10 ranks respective to the quantitative variables (Tables 3.6 and 3.7).

Table 3.9: Set of variables selected for clustering per degree of variance-correlation relationship.

Variance-correlation relationship	**Two clusters**	**Four clusters**
linear	EC1dTP, homoc, ldl	EC1dTP, dop.u
quadratic	EC1dTP, EC3v3	EC1dTP, EC3v4, EC2dTQQPQQ, ECv4
cubic	EC1dTP, EC1n, ECv4, EC3v4	EC1dTP, EC3v4, EC2dTQQPQQ, ECv4, EC1n, dfC1dTP, EC2n, EC3v3
quartic	EC1dTP, EC1n, ECv4, EC3v4, EC3v3, dfC1dTP, EC2dTQQPQQ, EC2n	EC1dTP, EC3v4, EC2dTQQPQQ, ECv4, EC1n, dfC1dTP, EC2n, EC3v3, dfC3v4, EC3n, dfC2dTQQPQQ, EC3v2dA
quintic	EC1dTP, EC1n, ECv4, EC3v4, EC3v3, dfC1dTP, EC2dTQQPQQ, EC2n, EC3v2, dfC1n	EC1dTP, EC3v4, EC2dTQQPQQ, ECv4, EC1n, dfC1dTP, EC2n, EC3v3, EC3v2, dfC3v4, dfC1n, EC3n, dfC2dTQQPQQ

The same does not occur for variables selected for classification (Table 3.10), where the variables selected are among those with higher discriminative potential. Additionally, the first variable selected by this method is the same as the top variable in the univariate ranks (`dfA2DQ` for binary classification and `insul` for multiclass classification).

Table 3.10: Set of variables selected for classification per degree of variance-correlation relationship.

Variance-correlation relationship	**Binary (Y_2)**	**Multiclass (Y_4)**
linear	dfA2DQ, EC3v8	insul, cerv.perim.cm, dfA2DQ, dfA2D2Q, trigl, adren.u
quadratic	dfA2DQ, abdo.perim.cm, EC3v8, age	insul, homa.ir, abdo.perim.cm, bmi, dfA2DQ, trigl, noradren.u
cubic	dfA2DQ, abdo.perim.cm, EC3v8, dfA2D2Q	insul, homa.ir, cerv.perim.cm, abdo.perim.cm, bmi, dfA2DQ, trigl, noradren.u
quartic	dfA2DQ, abdo.perim.cm, cerv.perim.cm, EC3v8, dfA2D2Q, trigl	insul, homa.ir, cerv.perim.cm, abdo.perim.cm, bmi, dfA2DQ, dfA2D2Q, trigl, noradren.u, EA2MQ
quintic	dfA2DQ, abdo.perim.cm, cerv.perim.cm, EC3v8, dfA2D2Q, insul, trigl	insul, homa.ir, cerv.perim.cm, abdo.perim.cm, bmi, dfA2DQ, dfA2D2Q, trigl, noradren.u, noradren.u24, cigarettes, EA2MQ

As for qualitative variables the averaged rank of their potential to discriminate the outcome classes was considered (table 3.11). From the eight qualitative variables, tft and resp.path were excluded from selection as the available sample and distribution did not allow for the estimation of the entropy. From the remaining six, three variables were selected based on the average rank of qualitative variables between the ranks estimated with the p-value and the entropy (table 3.8).

Table 3.11: Selection of qualitative variables based on the rank average from the discriminative potential estimated with the p-value and entropy.

Binary (Y2)		**Multiclass (Y4)**	
Variable	Averaged rank	Variable	Averaged rank
metab.path	2	smoking.habits	1.5
smoking.habits	2	card.path	4
awakenings	3.5	metab.path	4
card.path	4.5	awakenings	4
endoc.path	6	endoc.path	6
morn.head	7	morn.head	7
tft	-	tft	-
resp.path	-	resp.path	-

3.2 Classification

After preparing the data, unsupervised and supervised classification were applied. The packages and functions from `R` considered for the production of the results are detailed in table D.1 of the appendix.

3.2.1 Unsupervised classification

For unsupervised classification, the K-medoids (partitioning) and the Agglomerative hierarchical clustering methods were applied. The use of such methods is aimed for comparing the results from natural pattern search and the true classification for the outcome (classification for OSA), considering the accuracy, specificity and sensitivity as measures for evaluation. For the evaluation of the cluster quality, the respective silhouette and separation of the clusters was extracted.
For each method, the parameters below were defined:

- Proximity measure: Mahalanobis distance
- Number of pretended clusters: two/four
- Variables selected: five sets of quantitative variables selected based on the variance-correlation relationship rule applied for clustering

3.2.1.1 K-medoids (partitioning)

i. Two clusters

From the partition results into two clusters, all clustering scenarios show a positive consistency (silhouette) and separation between the clusters formed (table 3.12). From those, the set of variables corresponding to the linear degree presents the best average silhouette, but the partition made for the data shows to be inadequate (67:2).

Table 3.12: Summary table of clustering quality per set of variables selected for unsupervised classification: two clusters formed on dataset D_{iu}.

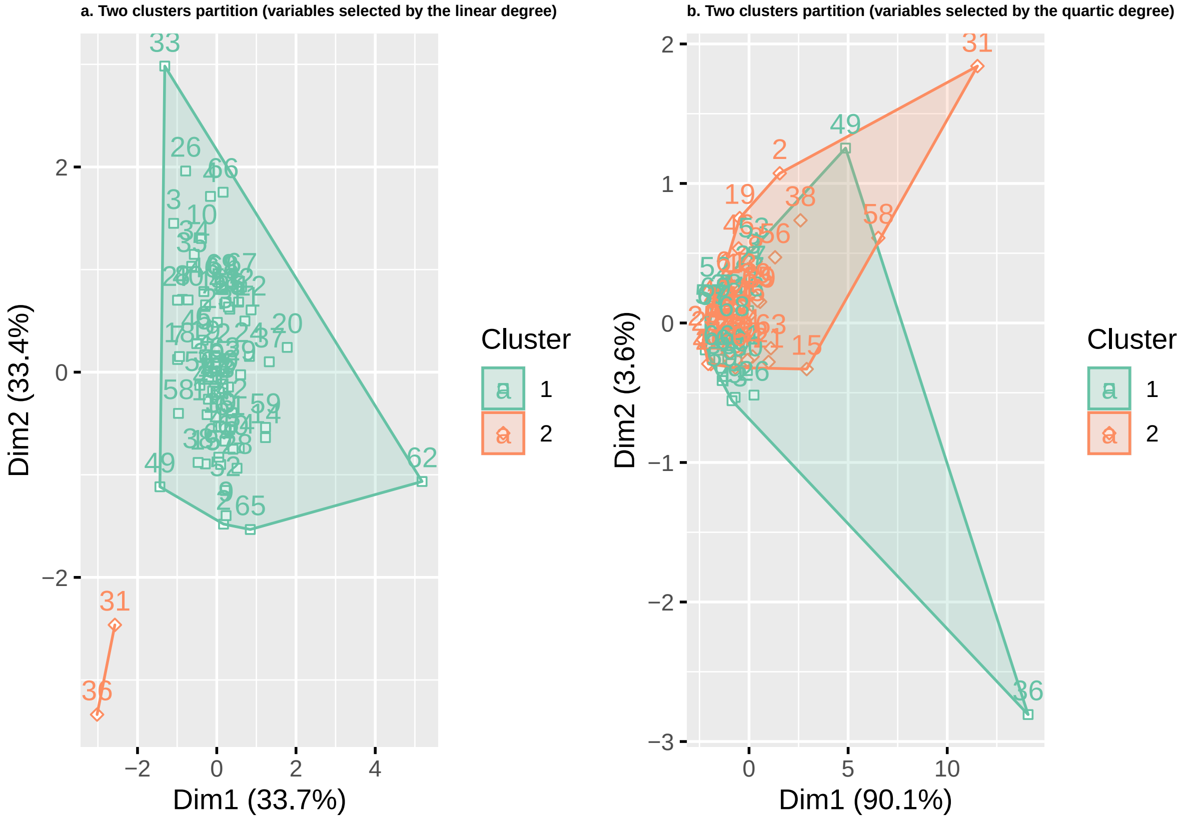
	Cluster 1		Cluster 2			
Variance-correlation relationship	Size	Silhouette	Size	Silhouette	Average sihlouette	Separation
linear	67	0.8	2	0.9	0.8	5
quadratic	68	0.9	1	0	0.5	31
cubic	68	0.9	1	0	0.4	37
quartic	25	0.2	44	0.2	0.2	0.9
quintic	48	0.5	21	-0.3	0.1	1

The fit of the data to the true outcome classes is considered, for two cluster formation, by the analysis of the two possible scenarios: (1) cluster one represents class C and cluster two represents class OSA; (2) cluster one represents class OSA and cluster two represents class C. Evaluating the fit of the data to the true outcome classes, the quartic degree of variance-correlation relationship shows one of the best results (with balanced values of accuracy,sensitivity and specificity) if the cluster one is considered to be class C and cluster two class OSA (table 3.13). The fitting results to the true class show to be rather close to the 50% proportion of correct classification.

Table 3.13: Summary table of clustering results adjusted to the real binary classification per set of variables selected for unsupervised classification: two clusters formed on dataset D_{iu}.

Variance-correlation relationship	Cluster identification	Sensitivity	Specificity	Accuracy
linear	cluster 1: C, cluster 2: OSA	0.04	1	0.26
	cluster 1: OSA, cluster 2: C	0.96	0	0.74
quadratic	cluster 1: C, cluster 2: OSA	0.02	1	0.25
	cluster 1: OSA, cluster 2: C	0.98	0	0.75
cubic	cluster 1: C, cluster 2: OSA	0.02	1	0.25
	cluster 1: OSA, cluster 2: C	0.98	0	0.75
quartic	cluster 1: C, cluster 2: OSA	0.66	0.44	0.61
	cluster 1: OSA, cluster 2: C	0.34	0.56	0.39
quintic	cluster 1: C, cluster 2: OSA	0.36	0.88	0.48
	cluster 1: OSA, cluster 2: C	0.64	0.12	0.52

The visualization of the clustering results is possible with the plot of the first two principal components of a principal component analysis. These first two components are those including the majority of the information present in the variables used. Viewing the plot result for the linear degree (plot a) it is verified that the separation of the clusters related with k-medoids is also observed for the two principal components calculated for the displaying of the plot (figure 3.2). For plot b (set of quartic relationship degree), the separation observed for k-medoids is not observed in the two principal components.



Four clusters

The partition into four clusters shows an overall positive consistency and separation for all sets of variables considered for clustering (table 3.14).

Table 3.14: Summary table of clustering quality per set of variables selected for unsupervised classification: four clusters formed on dataset D_{iu}.

	Cluster information				
Variance-correlation relationship	id	Size	Silhouette	Average sihlouette	Separation
linear	1	54	0.8	0.53	0.0
	2	12	0.4		0.0
	3	2	0.9		10.8
	4	1	0.0		12.7
quadratic	1	31	0.5	0.15	0.2
	2	27	0.2		0.2
	3	10	-0.1		0.4
	4	1	0.0		27.7
cubic	1	28	0.1	0.09	0.4
	2	39	0.3		0.4
	3	1	0.0		23.1
	4	1	0.0		51.9
quartic	1	32	0.0	0.06	2.0
	2	35	0.2		2.0
	3	1	0.0		57.4
	4	1	0.0		47.5
quintic	1	44	0.1	0.05	1.2
	2	23	0.1		1.2
	3	1	0.0		58.4
	4	1	0.0		39.2

The estimation of the measures of fit to the true classes of the outcome is of more complex evaluation since all possible combinations of true classes and predicted cluster must be accounted for. Visualizing the 4×4 confusion matrix between the true classes and predicted clusters helps for having an insight of any particular set of matching that reveals a good fitting scenario.

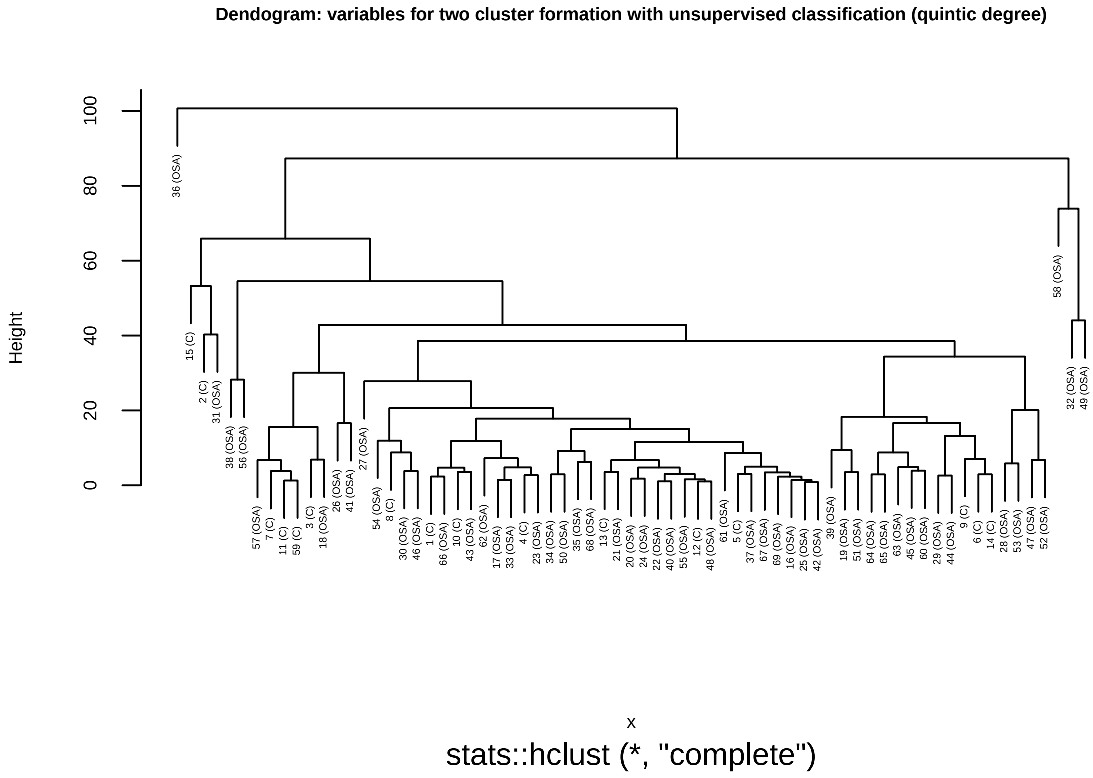
Exemplified with the set of variables of the quintic degree, the four clusters originated per set of variables showed a poor fitting pattern to the true outcome classes respective to the multiclass defined variable Y_4 (3.15).

Table 3.15: 4 x 4 confusion matrix between the four cluster, formed with the variables from the quintic variance-correlation relationship, and the true classes of the multiclass outcome Y_4

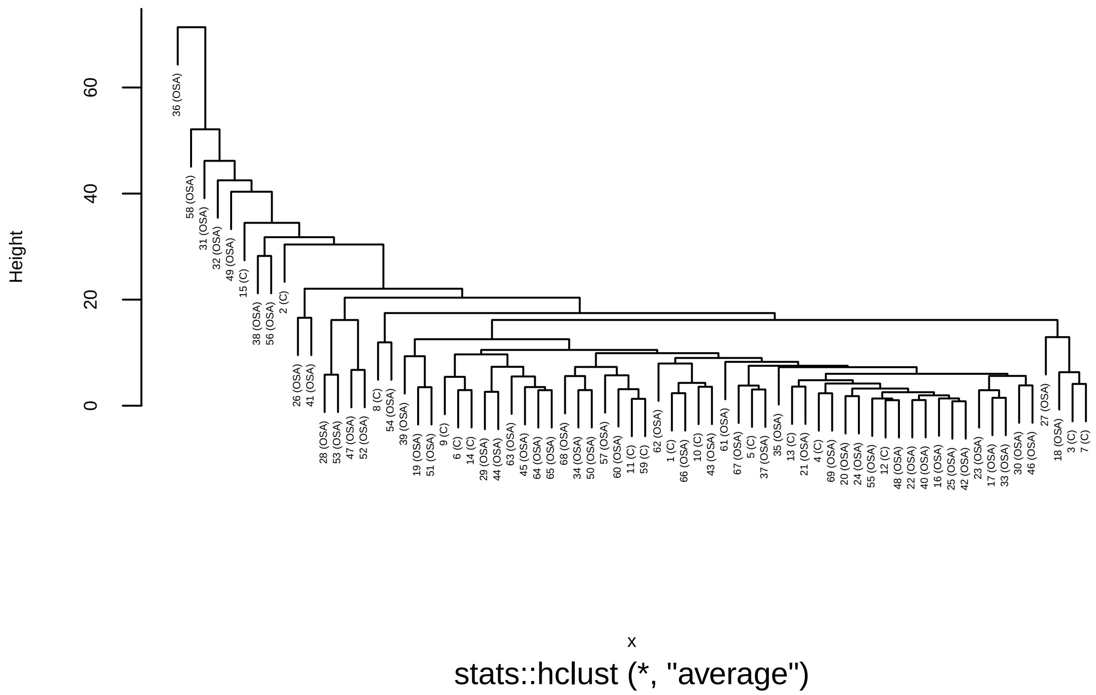
Quintic degree	cl1	cl2	cl3	cl4
C	10	6	0	0
OSA-I	18	8	1	0
OSA-II	6	4	0	0
OSA-III	10	5	0	1

3.2.1.2 Agglomerative Hierarchical clustering

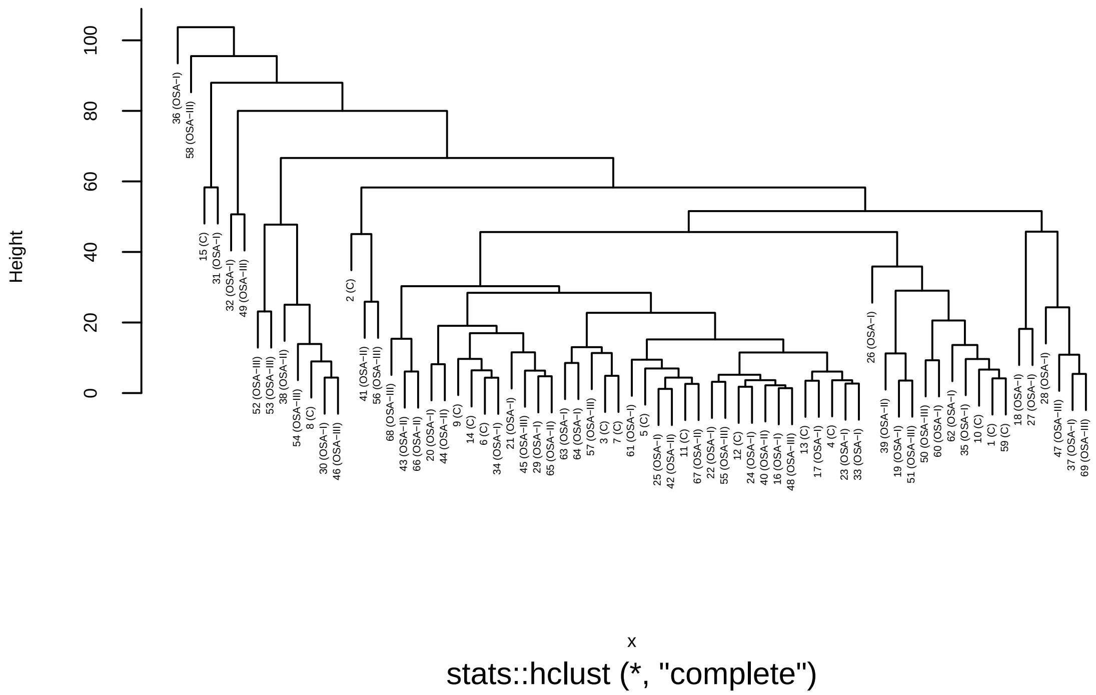
For the agglomerative hierarchical clustering, the complete and average linkage functions were considered. As this method does not require and initial decision on the number of cluster to form, the decision of the cut for two cluster (to compare with binary outcome variable Y_2) and four cluster (to compare with four outcome variable Y_4) was not possible for the results that were presented for all scenarios of variables selected for clustering provided by the five degrees of variance-correlation relationship. The set of variables of the quintic degree was used to exemplify the the poor results that are observed in all scenarios.



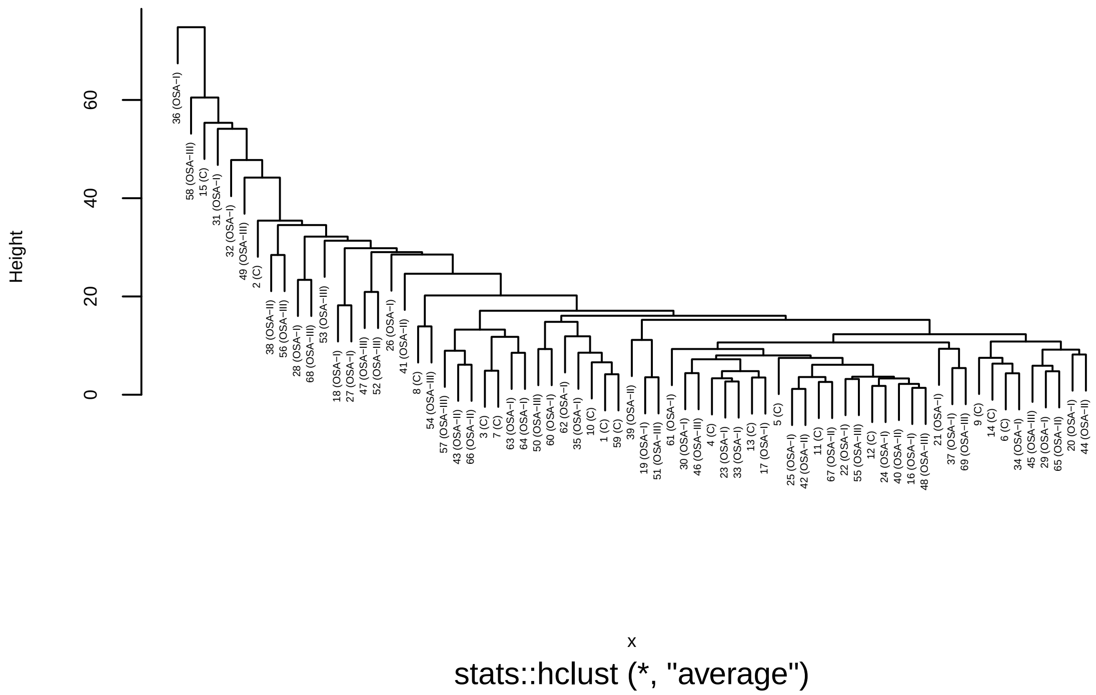
Dendogram: variables for two cluster formation with unsupervised classification (quintic degree)



Dendogram: variables for four cluster formation with unsupervised classification (quintic degree)



Dendogram: variables for two cluster formation with unsupervised classification (quintic degree)



3.2.2 Supervised classification

For supervised classification, the models presented for each method were tuned with the complete dataset. For testing those models, the Leave One Out Cross Validation (LOOCV) was considered. Measures of performance regarded: accuracy, sensitivity, specificity and dice coefficient.

3.2.2.1 Decision Tree

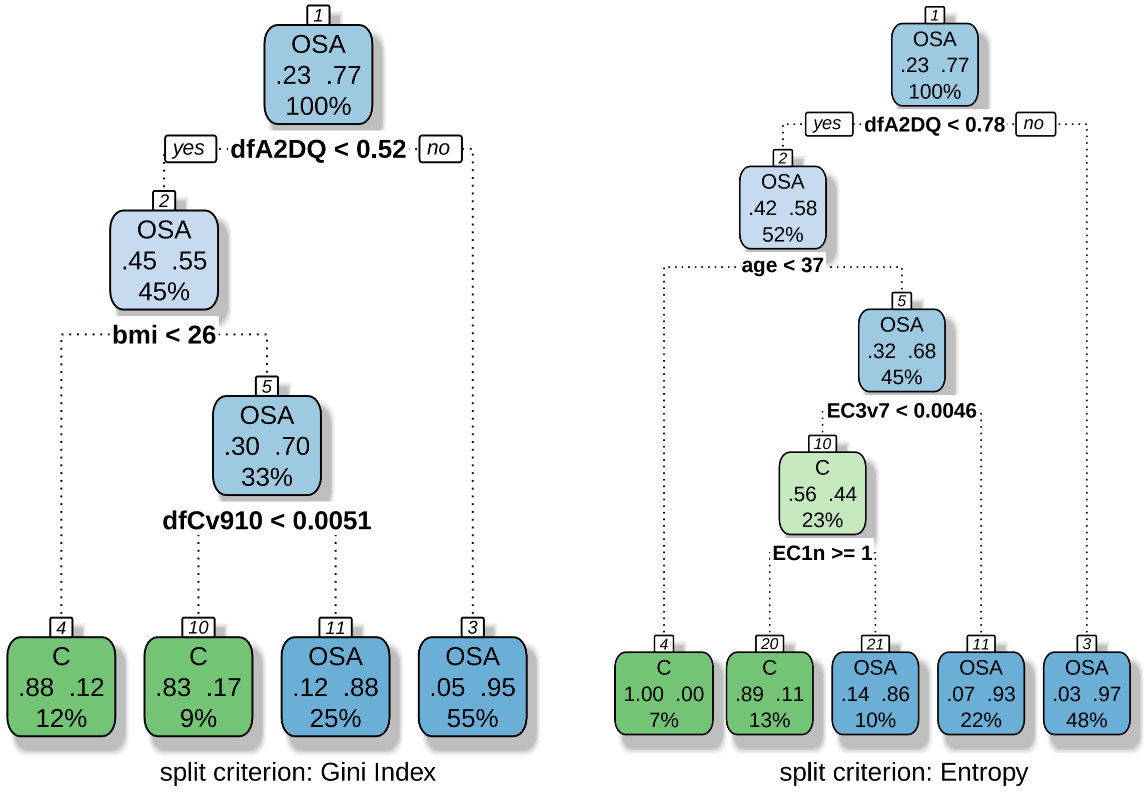
For this method all 76 variables (quantitative + qualitative) were made available for the tree to decide the best set for both binary and multiclass classification. Both Gini index and Entropy criteria were considered for the split, producing two models for binary and two models for multiclass outcomes.

Binary classification

The plots represent the model constructed by the decision tree for binary outcome. The variables `dfA2DQ`, `bmi` and `dfCv910` were selected by the Gini Index criterion to form a tree with four terminal nodes, and the variables `dfA2DQ`, `age`, `Ec3v7` and `EC1n` were selected by the Entropy criterion to form a tree with five terminal nodes. The reading of the dendrogram is exemplified below for the terminal node number four originated with Gini index criterion:

- Node information: `C` is the predominant class in the node in the proportion of 0.88 from the 12% of the total sample;

- Classification rule: if `dfA2DQ` < 0.52 and `bmi` < 26 then the elements that meet this criteria are classified as Controls (`C`).



split criterion: Gini Index

split criterion: Entropy

In table 3.16 are observed the performance results for the train and test sets considering the use of the models visualized in the dendrograms. For both splitting criteria used, the overall performance of the model for the train set is higher than for the respective evaluation of the test sets. Nevertheless, the performance results for the test sets remain close to those of the train sets, except for the measured value specificity for which is observed a greater decrease for a minimum of 0.5 respective to the use of the Gini index criterion. The results between the two splitting criteria are overall similar.

Table 3.16: Summary of performance of the Tree classifier trained for binary classification.

Classification type	**criterion**	**Set**	**Sensitivity**	**Specificity**	**Accuracy**	**Dice coefficient**
Binary	Gini Index	Train	0.96	0.75	0.91	0.96
		Test	0.89	0.50	0.80	0.90
	Entropy	Train	0.98	0.81	0.94	0.97
		Test	0.92	0.56	0.84	0.92

Multiclass classification

As for multiclass classification, compared to the binary classification, the trained classifiers observed in the following dendrograms show an increasing number of branches formed and, subsequently, terminal nodes formed. Additionally, the performance of the classification in both train and test sets is reduced in comparison to the performance results from the binary classification. The use of both splitting criteria show no great differences in the results observed (table 3.17).

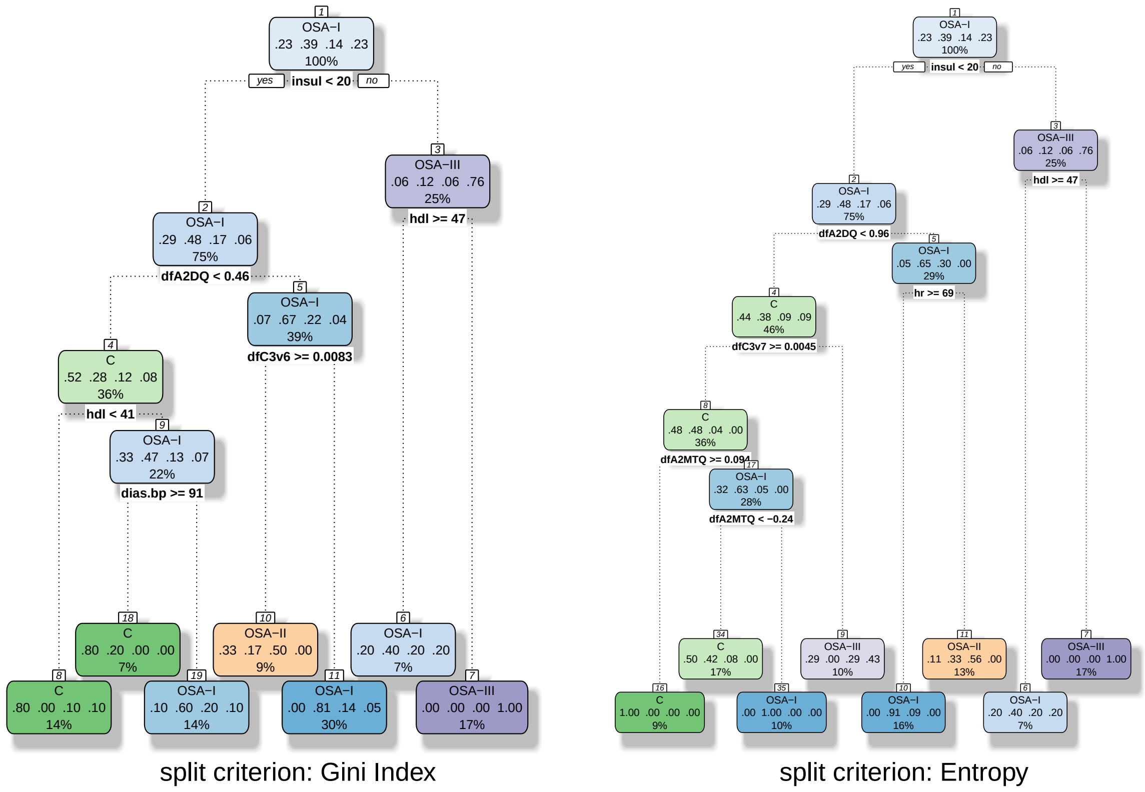


Table 3.17: Summary of performance of the Tree classifier trained for multiclass classification.

Classification type	criterion	Set	Sensitivity	Specificity	Accuracy	Dice coefficient
Multiclass	Gini Index	Train	0.68	0.90	0.75	0.91
		Test	0.48	0.79	0.57	0.83
	Entropy	Train	0.72	0.89	0.74	0.91
		Test	0.46	0.76	0.52	0.82

3.2.2.2 Naïve Bayes

The application of Naïve Bayes is facilitated by the function `naivebayes` (package `naive_bayes`), available in `R`, as it allows for the estimation of the density distribution of quantitative variables with an univariate kernel density estimator. Additionally, this function is capable of identifying count type of quantitative variables (use of poisson distribution) and qualitative variables (both binary and multiclass).

```r
classifier <- naivebayes::naive_bayes(y ~ variables,
                                      data = data,
                                      usekernel = T,
                                      usepoisson = T,
                                      kernel = "gaussian",
                                      bw = "nrd0") #silverman's rule of thumb
```

For both binary and multiclass classification, the five sets of variables selected for classification were considered. The density distribution of the quantitative variables regarded was estimated with an univariate kernel. Additionally, the top three qualitative variables (for binary and four class classification) were considered for the construction of the model (table 3.11).

Binary classification

In binary classification, the model trained with the variables selected with the quintic degree, plus the top three qualitative variables for binary classification, show performance results more balanced along the three measures. Therefore, the test set with LOOCV was performed for that model trained. Compared to the results of binary classification with decision trees, the performance in both train and test sets resultant from the set of considered quantitative and qualitative variables is similar. Difference of the results observed between the two methods are on the set of variables that compose the rules for classification, in which only the variable `dfA2DQ` is commonly considered.

Table 3.18: Summary of performance of the Naïve Bayes classifier trained for binary classification.

	Train set			Test set (LOOCV)			
Variance-correlation relationship	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Dice coefficient
linear	0.86	0.98	0.44				
quadratic	0.91	1.00	0.62				
cubic	0.93	1.00	0.69				
quartic	0.90	0.96	0.69				
quintic	0.90	0.94	0.75	0.80	0.87	0.56	0.9

Multiclass classification

In multiclass classification, the model trained with the variables selected with the quartic degree, plus the top three qualitative variables for binary classification, show performance results more balanced along the three measures. Therefore, the test set with LOOCV was performed for that model trained. The results of the test set decrease more substantially than for the results of binary classification with Naïve Bayes and also when compared to the multiclass classification with the decision tree.

Table 3.19: Summary of performance of the Naive Bayes classifier trained for multiclass classification.

	Train set			Test set (LOOCV)			
Variance-correlation relationship	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Dice coefficient
linear	0.77	0.73	0.92				
quadratic	0.80	0.78	0.93				
cubic	0.80	0.76	0.93				
quartic	0.81	0.78	0.93	0.45	0.43	0.81	0.81
quintic	0.78	0.78	0.93				

3.2.2.3 Logistic Regression

Binary logistic regression was applied for the classification regarding Y_2. For the multiclass outcome, considering that there is order in the classes of Y_4, the ordinal logistic regression was applied.

Although these methods can perform some type of variable selection, the total set of variables was not included due to the computational effort required. Instead, the set of variables select for classification respective to the quintic degree was considered given that it is the less strict relationship allowing for a greater number of variables to be included. The top qualitative variables selected was added to the set of variables for model training.

Binary Logistic Regression

The initial model considered the use of all variables described above. To get a parsimonious model, first the Variance inflation factor (VIF) was considered for evaluating variable multicollinearity. Posteriorly, a step of variable selection based on backward, forward and stepwise selection led to the derivation of the same altered model. The test comparing the initial model with the altered one revealed that they are not significantly different and, therefore, the model derived can be used (table 3.20).

Table 3.20: Selection of the model with the best parsimony agreement based on forward, backward and stepwise selection methods.

Full model	Model selected	ANOVA test (p-value)
y ~ dfA2DQ + abdo.perim.cm + EC3v8 + dfA2D2Q + insul + trigl + metab.path + smoking.habits + awakenings	y ~ dfA2DQ + abdo.perim.cm + EC3v8 + awakenings	0.527

ANOVA test for nested models evaluates de difference between two models that are nested with each other (in hierarchy). The shortened model is nested into the full model. Evaluates the difference of group covariates of one model and another. Being the null hypothesis of no statistically difference between them true, the shorter model can be selected over the full model as it does not have statistically significant difference from the full model and is more parsimonious because it does not need all the variables from the full model.

The model selected including the respective coefficients is written as follows:

$$logit(Y) = -6.121 + 0.808 \times dfA2DQ + 0.071 \times abdo.perim.cm - 833.042 \times EC3v8 + 1.387 \times awakenings1$$

Using the ROC curve, the best probability threshold was decided based on the highest Index of Youden. The best threshold probability is 0.661 and this results on the highest grouped specificity and sensitivity resulting from the classification.

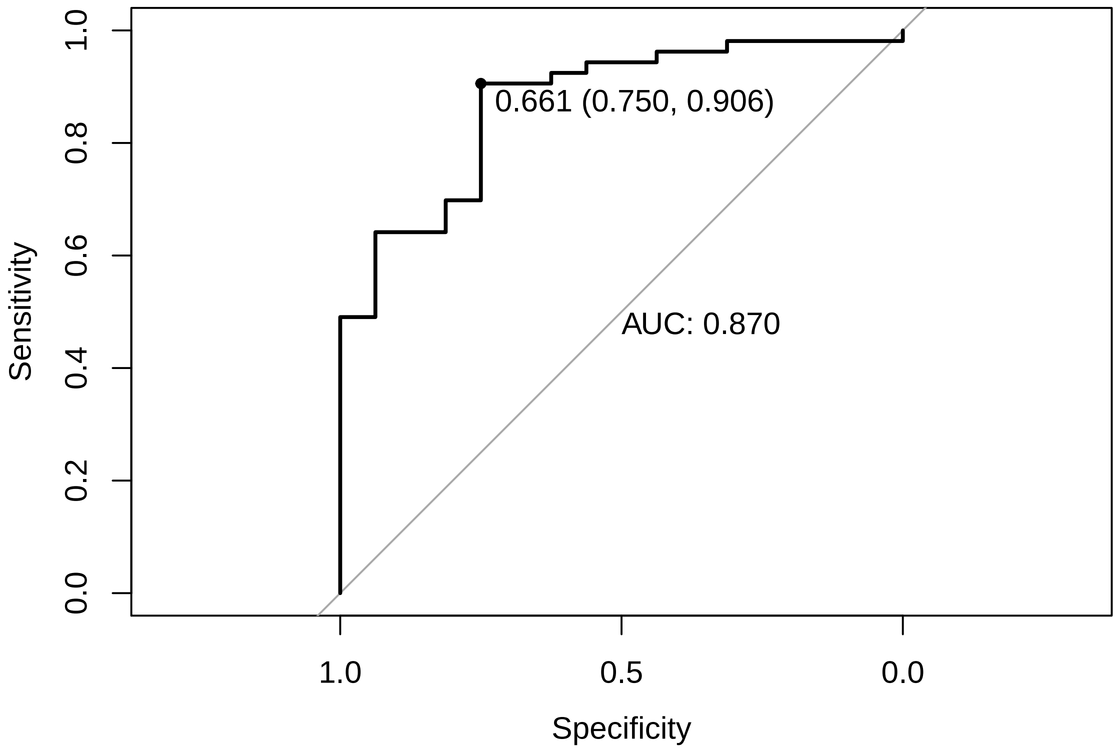


Figure 3.2: ROC Curve perfomed for finding the best probability threshold for classification with Binary logistic regression.

Table 3.21: Summary of performance of the Binomial logistic regression model trained for binary classification

Set	Best threshold (probability)	Index of Youden	Accuracy	Specificity	Sensitivity
Train	0.661	0.66	0.87	0.75	0.91
Test	(from train)	n/a	0.80	0.56	0.87

Evaluating the parameters (intercept and β coefficients) resultant from testing with each of the LOOCV samples, the mean value is similar to their value in the trained model (table 3.22).

Table 3.22: Comparison of parameters (intercept and β coefficients) estimated in the trained model and those estimated for each of the test models.

Parameter	Trained model	Summary statistics of Tested models					
		min	max	median	mean	quantil 25	quantil 75
(Intercept)	-6.121	-8.329	-5.028	-6.107	-6.145	-6.190	-5.997
dfA2DQ	0.808	0.721	1.037	0.805	0.811	0.796	0.808
abdo.perim.cm	0.071	0.060	0.093	0.070	0.071	0.069	0.071
EC3v8	-833.042	-1031.612	-635.429	-827.467	-834.902	-832.620	-820.664
awakenings1	1.387	1.109	1.797	1.380	1.391	1.345	1.411

Ordinal Logistic Regression

The set of variables for which the proportional odds assumption can be validated are those selected by the linear variance-correlation relationship together with the top three set of categorical variables selected for the multiclass response type (table 3.23).

Table 3.23: Summarized output result from Proportional Odds test using `polr` function from `MASS` package in `R`.

	b[polr]	b[>C]	b[>OSA-I]	b[>OSA-II]	Chi-square	df	Pr(>Chisq)
insul	0.09	0.03	0.09	0.15	3.94	2	0.14
cerv.perim.cm	0.13	0.10	0.19	0.24	0.51	2	0.77
dfA2DQ	0.23	0.58	0.09	-0.20	4.35	2	0.11
dfA2D2Q	0.03	0.49	-0.07	-0.30	1.58	2	0.45
trigl	0.00	0.00	0.00	0.00	1.05	2	0.59
adren.u	-0.02	-0.04	-0.01	0.06	2.10	2	0.35
smoking.habits1	0.05	-0.38	0.88	-0.05	2.55	2	0.28
smoking.habits2	0.60	0.55	1.05	0.58	0.71	2	0.70
card.path1	0.19	0.73	-0.01	0.27	0.94	2	0.62
metab.path1	0.16	0.58	-0.24	-0.18	1.21	2	0.55

The values of the parameters are shown in table 3.24. At a significance level $\alpha = 0.1$ the effect of the variable `insul` is statistically significant.

Table 3.24: Parameters of the ordinal logistic regression model. The effect of the explanatory variables is constant but the intercept varies according to the equation.

Parameter	Variable	Value	Std. Error	t value	p-value	
Coefficients	insul	0.09	0.04	2.39	0.017	*
	cerv.perim.cm	0.13	0.10	1.27	0.204	
	dfA2DQ	0.23	0.15	1.46	0.144	
	dfA2D2Q	0.03	0.32	0.10	0.920	
	trigl	0.00	0.00	0.45	0.653	
	adren.u	-0.02	0.04	-0.60	0.549	
	smoking.habits1	0.05	0.61	0.07	0.944	
	smoking.habits2	0.60	0.58	1.04	0.298	
	card.path1	0.19	0.50	0.38	0.704	
	metab.path1	0.16	0.50	0.33	0.741	
Intercepts	C|OSA-I	5.80	4.08	1.42	0.156	
	OSA-I|OSA-II	8.02	4.15	1.93	0.054	*
	OSA-II|OSA-III	8.97	4.17	2.15	0.032	*

The performance for classification shows an accuracy below 0.5 (less than half of the elements are correctly classified) in both trained and test models.

Table 3.25: Summary of performance of the Ordinal logistic regression model trained for multiclass classification

Set	Accuracy	Specificity	Sensitivity
Train	0.54	0.82	0.45
Test	0.38	0.76	0.30

Chapter 4: Discussion

The application of the methods is discussed.

The case study approached a scenario of small sample size and high dimensionality of the data (76 characteristics measured on 69 individuals). The diversity of the sample was well represented by both quantitative and qualitative types of nature. The task of classification for the diagnosis of OSA may already expect a bias in the selection of the observations and the availability of small sample sizes. Firstly, the patients that undergo the PSG test are usually recommended by a medical specialist due to suspicions of being affected with the disturbance. Although the use of suspicion is a poor criteria because it may mask individuals that are affected but no suspicions are raised, this criteria is fairly good for patients that meet the set of suspicions. Therefore, it may be associated to the extreme imbalance of the control class compared to the class of disturbance presence (OSA) (proportion approximated to 1 Control for 4 OSA). This imbalance is reduced when the multiclass outcome variable is considered.

Secondly, the sample size could already be expected to be small due to the costs, capacity limitation and the strict requirement for highly specialized medical teams to perform the test.

Data analysis proceeded initially by an univariate analysis of candidate predictors per class of the outcome. The application of the three different methods (p-value, entropy and empirical AUC) showed, separately per type of outcome variable, approximate results in the rank of variables with higher discrimination capacity. Variables already associated to the presence of the disturbance, such as the abdominal perimeter, BMI, cervical perimeter and age, stood out in the top 10 rank of quantitative variables with higher potential to discriminate the presence of OSA (binary classification). Regarding the discrimination of its severity types, variables of hormonal and metabolic type made presence in the rank. In both types of outcome, the proteoform expression stood out mainly from the Apo A2 proteoform types in the period of expression variation between morning and evening. To overcome the curse of dimensionality, a process of variable selection was necessary. This process considered the importance of each variable potential to discriminate the elements in groups and the existence of correlation between them. The variables selected for classification are among those presenting a better separation of the classes and reduced variable redundancy originated by correlated variables. The method for the selection of quantitative variables presents very clearly how discrepant can the variable selection be when the same is supervised or not.

Discussing the application of the classification methods, the poor results observed for the unsupervised classification show the importance of a good quality of the data, that may not be present in the data analyzed. Along the supervised classification process, the variables `dfA2DQ` and `insul` reveal a strict presence as the most important variables for the classification of patients for Obstructive Sleep Apnea. Observed for all supervised methods, the performance in binary classification is overall higher than the performance observed for multiclass classification. These results may occur because the classification complexity is increasing (from two classes to four classes) and the characteristics observed for patients classified in the same class have an associated variability aggravated by the smaller sample size observed in each of the classes of the multiclass response (figure 4.1).

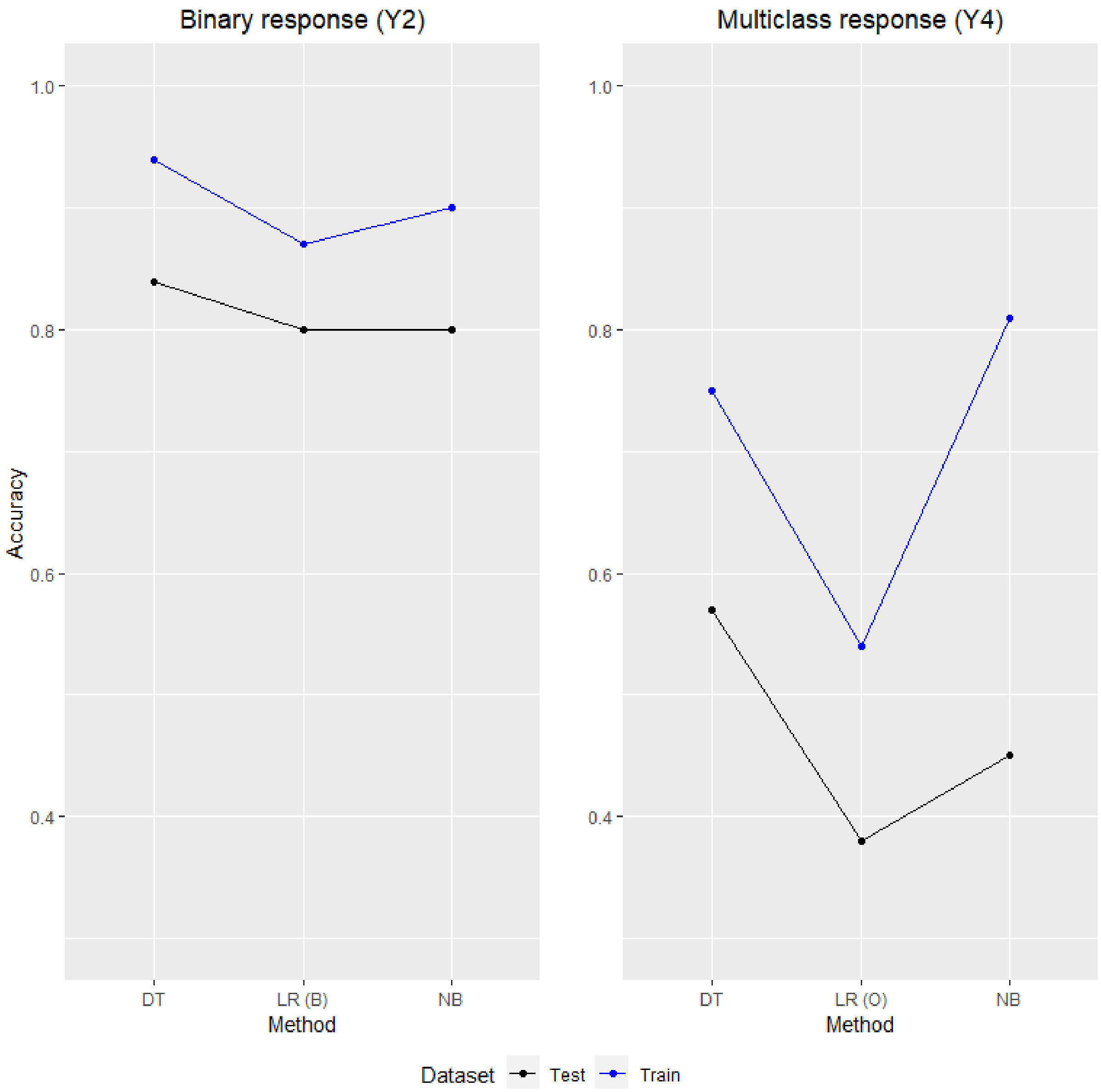


Figure 4.1: Accuracy of the models tuned with the train data and tested on the test data for each of the response variable types (binary and multiclass). DT: Decision Tree; LR(B): Binomial Logistic Regression; LR(O): Ordinal Logistic Regression; NB: Naïve Bayes.

The probabilistic methods revealed to be more sensitive to changes in the values observed for the covariates included in the model. Particularly for the Naïve Bayes method, the model trained with the set of variables obtained by the quartic and quintic degree are among the best performing models. These results can be expected as these variance-correlation relationship degrees require a less strict relationship between within-group variance and correlation that allows for variables more correlated to be in the set of selected variables. In this way, the greater number of variables selected with that relationship, compared to the other relationships, allows the models to use more information about the elements. This may also indicate that, although constantly being regarded among the variables with most potential for the discrimination of the outcome, the variables select in the linear and quadratic relationship degrees, for instance, are not capable alone to produce a better result than the addition of the remaining variables selected in the less strict relationships. As for the logistic regression models, the steps necessary for model training are quite complex and the rules create are not easily interpretable. Trees are particularly desirable for analysis and classification assignments in high dimensional data as they ease decision making on complex tasks by breaking down the problem, based on a large number of candidate predictors, into several simpler decision tasks and thus enabling a better human interpretability of the problem.

Guided by methods' performance, compared in figure 4.1, and the evaluation of result interpretation desired for the present case study, the applied Decision tree may be the best performing method. In terms of observed accuracy, no particular method differentiates from the remaining. Nevertheless, the model created by decision tree presents the highest accuracy for the classification according to a binary response type and considering the evaluation of performance separately per type of set being classified (train set *vs* test set). Evaluating the scenario of disturbance severity classification (multiclass response type), although the classification of the train set with the Decision tree is not the best among the methods applied, the decrease in performance observed on all methods when classifying the test sets is less steep with this method applied, which would possibly indicate model consistency if the data was representative of the population (which is not the case). Among all supervised methods applied, the power of simple interpretation of the results gained with the Decision tree may be more appealing to clinical specialists in the future prospects of applying such classification process for the pre-assessment of their patients' propensity for the definitive diagnosis test. Lastly regarding the importance of the characteristics, a pre-screening of OSA should consider the net expression of Apolipoproteins A2 proteorforms and the clinical variables with previously studied association to OSA.

Chapter 5: Conclusion

This dissertation contains the compilation of the most popular classification methods for Pattern recognition. Much more than the simple use of classification methodologies, for the application of pattern recognition it is essential a good acquaintance of the data and pre-preparation so that most diversity and abundance of the data information can be applied for classification.

As starting a classification analysis, it must be kept in mind that there is no method that can show primarily to be better than another. The extensive list of methods available, added to the high diversity of their underlying mechanisms, requires a thorough analysis of which method to select. This assessment must consider the purpose of the study - finding patterns or create supervised based rules for the classification of new data - and how transparent and interpretable it is desired the method to be at the risk of possibly harnessing the performance.

The case study of Obstructive Sleep Apnea is an example of pattern recognition applied to data in the less optimal condition desirable: imbalanced data, low sample size and high dimensionality allied to the poor quality of the characterizing variables collected. It is concluded that the use of proteoform expression combined with pathophysiological parameters previously associated to the disturbance revealed in this study for the classification of the individuals and may mark a commence point for the increased study of proteoform expression associated to the pre-diagnosis of one of the most urgent health issues in human medical care.

Suggestions are made for future application on the case study.

In regards to data preparation, an additional step can consider the application of sample balancing techniques to alleviate the limitations brought by the small sample size, such as over-sampling techniques (increase of small samples). For the process of classification with concrete model rules, black-box supervised methods such as Ensemble methods (*e.g.* Bagging and Adaboost) and Neural networks may be considered for evaluating the improvement of classification performance, although keeping in mind that result interpretability will be lost. If a white-box characterization is to be maintained, the neural network with a single layer (Single Layer Perceptron) is an additional candidate method to apply.

Bibliography

Andrews, J. L., & McNicholas, P. D. (2014). Variable selection for clustering and classification. *Journal of Classification*, *31*(2), 136–153.

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Phil. Trans. R. Soc.*, *53*, 370–418. http://doi.org/10.1093/biomet/45.3-4.296

Berry, R. B., Budhiraja, R., Gottlieb, D. J., Gozal, D., Iber, C., Kapur, V. K., … others. (2012). Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events: Deliberations of the sleep apnea definitions task force of the american academy of sleep medicine. *Journal of Clinical Sleep Medicine*, *8*(5), 597–619.

Bodez, D., Guellich, A., Kharoubi, M., Covali-Noroc, A., Tissot, C.-M., Guendouz, S., … others. (2016). Prevalence, severity, and prognostic value of sleep apnea syndromes in cardiac amyloidosis. *Sleep*, *39*(7), 1333–1341.

Bonner, R. E. (1964). On some clustering techniques. *IBM Journal of Research and Development*, *8*(1), 22–32. http://doi.org/10.1147/rd.81.0022

Coughlin, S., Mawdsley, L., Mugarza, J., Calverley, P., & Wilding, J. (2004). Obstructive sleep apnoea is independently associated with an increased prevalence of metabolic syndrome. *European Heart Journal*, *25*(9), 735–741. http://doi.org/10.1016/j.ehj.2004.02.021

Davison, A., & Kuonen, D. (2002). An introduction to the bootstrap with applications in r. *Statistical Computing & Statistical Graphics Newsletter*, *13*(1), 6–11.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*(1), 1–18.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Hierarchical clustering. In *Cluster analysis* (5th ed., pp. 43–69). John Wiley.

Fan, J., & Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv Preprint Math/0602133*.

Feliciano, A., Torres, V. M., Vaz, F., Carvalho, A. S., Matthiesen, R., Pinto, P., … Penque, D. (2015). Overview of proteomics studies in obstructive sleep apnea. *Sleep Medicine*, *16*(4), 437–445. http://doi.org/10.1016/j.sleep.2014.11.014

Ferrari, E., Bosco, P., Calderoni, S., Oliva, P., Palumbo, L., Spera, G., … Retico, A. (2020). Dealing with confounders and outliers in classification medical studies: The autism spectrum disorders case study. *Artificial Intelligence in Medicine*, *108*, 101926. http://doi.org/https://doi.org/10.1016/j.artmed.2020.101926

Ferrie, J. E., Kumari, M., Salo, P., Singh-Manoux, A., & Kivimaki, M. (2011). Sleep epidemiology–a rapidly growing field. *International Journal of Epidemiology*, *40*(6), 1431–1437. http://doi.org/10.1093/ije/dyr203

Flemons, W. W. (2002). Obstructive Sleep Apnea. *New England Journal of Medicine*, *347*(7), 498–504. http://doi.org/10.1056/NEJMcp012849

Flemons, W. W., Douglas, N. J., Kuna, S. T., Rodenstein, D. O., & Wheatley, J. (2004). Access to Diagnosis and Treatment of Patients with Suspected Sleep Apnea. *American Journal of Respiratory and Critical Care Medicine*, *169*(6), 668–672. http://doi.org/10.1164/rccm.200308-1124PP

Gay, P., Weaver, T., Loube, D., & Iber, C. (2006). Evaluation of Positive Airway Pressure Treatment for Sleep Related Breathing Disorders in Adults. *Sleep*, *29*(3), 381–401. http://doi.org/10.1093/sleep/29.3.381

Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data mining and analytics in the process industry: The role of machine learning. *IEEE Access*, *5*, 20590–20616. http://doi.org/10.1109/ACCESS.2017.2756872

Gini, C. (1912). Variabilità e mutabilità. *Reprinted in Memorie Di Metodologica Statistica (Ed. Pizetti E.*

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857–871.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, *45*(2), 171–186.

Hawkins, D. M. (1980). Introduction. In *Identification of outliers* (pp. 1–12). Springer Netherlands. http://doi.org/10.1007/978-94-015-3994-4_1

Hegde, H., Shimpi, N., Panny, A., Glurich, I., Christie, P., & Acharya, A. (2019). MICE vs PPCA: Missing data imputation in healthcare. *Informatics in Medicine Unlocked*, *17*, 100275. http://doi.org/https://doi.org/10.1016/j.imu.2019.100275

Jain, A. K., & Duin, R. P. W. (2004). Introduction to pattern recognition. *This Text Is Taken from a Contribution of the Authors in RL Gregory (Eds.), The Oxford Companion to the Mind, Second Edition, Oxford University Press, Oxford, UK*, 698–703.

Katsios, C., & Roukos, D. H. (2010). Individual genomes and personalized medicine: Life diversity and complexity. *Personalized Medicine*, *7*(4), 347–350.

Kaufman, L., & Rousseeuw, P. J. (1990a). *Introduction* (pp. 1–44). John Wiley & Sons.

Kaufman, L., & Rousseeuw, P. J. (1990b). *Partitioning Around Medoids (Program PAM)* (pp. 68–125). John Wiley & Sons.

Liu, J., Sun, J., & Wang, S. (2006). Pattern recognition: An overview. *IJCSNS International Journal of Computer Science and Network Security*, *6*(6), 57–61.

Lumeng, J. C., & Chervin, R. D. (2008). Epidemiology of Pediatric Obstructive Sleep Apnea. *Proceedings of the American Thoracic Society*, *5*(2), 242–252. http://doi.org/10.1513/pats.200708-135MG

Lurie, A. (2011). Obstructive Sleep Apnea in Adults: Epidemiology, Clinical Presentation, and Treatment Options. In *Obstructive sleep apnea in adults* (pp. 1–42). Karger. http://doi.org/10.1159/000327660

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). Oakland, CA, USA.

Montesi, S. B., Bajwa, E. K., & Malhotra, A. (2012). Biomarkers of sleep apnea. *Chest*, *142*(1), 239–245.

O'Connor, C., Thornley, K. S., & Hanly, P. J. (2000). Gender Differences in the Polysomnographic Features of Obstructive Sleep Apnea. *American Journal of Respiratory and Critical Care Medicine*, *161*(5), 1465–1472. http://doi.org/10.1164/ajrccm.161.5.9904121

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, *9*, Article 6. Retrieved from https://doi.org/10.7275/qf69-7k43

Partinen, M., Jamieson, A., & Guilleminault, C. (1988). Long-term Outcome for Obstructive Sleep Apnea Syndrome Patients. *Chest*, *94*(6), 1200–1204. http://doi.org/10.1378/chest.94.6.1200

Population ages 15-64, total. (n.d.). https://data.worldbank.org/indicator/SP.POP.1564.TO?view=chart.

Ross, P. E. (1998). Introduction to pattern recognition. *Forbes*, 98–104.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, *21*(3), 660–674.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437.

Srihari, S. N. (1993). Recognition of handwritten and machine-printed text for postal address interpretation. *Pattern Recognition Letters*, *14*(4), 291–302.

Torres, V. M., Feliciano, A., Antunes, M., Vaz, F., & Penque, D. (2017). Proteoforms of transthyretin-candidate biomarkers in diagnosis of obstructive sleep. In *HUPO2017-16th human proteome organization world congress 17-21 september 2017*.

Watson, N. F. (2016). Health Care Savings: The Economic Value of Diagnostic and Therapeutic Care for Obstructive Sleep Apnea. *Journal of Clinical Sleep Medicine*, *12*(08), 1075–1077. http://doi.org/10.5664/jcsm.6034

Weaver, T. E., Maislin, G., Dinges, D. F., Bloxham, T., George, C. F. P., Greenberg, H., … Pack, A. I. (2007). Relationship Between Hours of CPAP Use and Achieving Normal Levels of Sleepiness and Daily Functioning. *Sleep*, *30*(6), 711–719. http://doi.org/10.1093/sleep/30.6.711

Węglarczyk, S. (2018). Kernel density estimation and its application. In *ITM web of conferences* (Vol. 23). EDP Sciences.

White, D. P., Gibb, T. J., Wall, J. M., & Westbrook, P. R. (1995). Assessment of Accuracy and Analysis Time of a Novel Device to Monitor Sleep and Breathing in the Home. *Sleep*, *18*(2), 115–126. http://doi.org/10.1093/sleep/18.2.115

Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q. P., & Lillard Jr, J. W. (2014). A study of effects of multicollinearity in the multivariable analysis. *International Journal of Applied Science and Technology*, *4*(5), 9.

Young, T. (1993). Analytic Epidemiology Studies of Sleep Disordered Breathing—What Explains the Gender Difference in Sleep Disordered Breathing? *Sleep*, *16*(suppl_8), S1–S2. http://doi.org/10.1093/sleep/16.suppl_8.S1

Young, T., Finn, L., Peppard, P. E., Szklo-Coxe, M., Austin, D., Nieto, F. J., … Hla, K. M. (2008). Sleep disordered breathing and mortality: eighteen-year follow-up of the Wisconsin sleep cohort. *Sleep*, *31*(8), 1071–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18714778%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2542952

Young, T., Peppard, P. E., & Gottlieb, D. J. (2002). Epidemiology of obstructive sleep apnea: a population health perspective. *American Journal of Respiratory and Critical Care Medicine*, *165*(9), 1217–39. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11991871

Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 856–863).

Yu, L.-R., Stewart, N. A., & Veenstra, T. D. (2010). Proteomics: The deciphering of the functional genome. In *Essentials of genomic and personalized medicine* (pp. 89–96). Elsevier.

Appendix A: Information value: demonstrations

Demonstration: Gini Impurity calculation

$$\begin{aligned} Gini &= \sum_{k=1}^{c} [p_k(1-p_k)] = \sum_{k=1}^{c}(p_k - p_k^2) \\ &= \sum_{k=1}^{c} p_k - \sum_{k=1}^{c} p_k^2 = 1 - \sum_{k=1}^{c} p_k^2 \quad (q.e.d.) \end{aligned} \tag{A.1}$$

Demonstration: Maximum disorder

The disorder will be the highest if all outcome classes have equal probability in the data space considered. For a c total number of classes, the proportion of each for maximum disorder is estimated be $\left(\frac{1}{c}\right)$.

[1] Maximum Entropy

$$\begin{aligned} E_{max} &= -\sum_{k=1}^{c} [p_k \times \log_2(p_k)] = -\sum_{k=1}^{c} \left[\frac{1}{c} \times log_2\left(\frac{1}{c}\right)\right] \\ &= -\frac{1}{c} \times \sum_{k=1}^{c} \log_2\left(\frac{1}{c}\right) = -\frac{1}{c} \times c \times log_2\left(\frac{1}{c}\right) = -\log_2\left(\frac{1}{c}\right) \\ &= \log_2(c) \quad (q.e.d.) \end{aligned} \tag{A.2}$$

[2] Maximum Gini impurity

$$\begin{aligned} Gini_{max} &= 1 - \sum_{k=1}^{c} p_k^2 = 1 - \sum_{k=1}^{c} \left(\frac{1}{c}\right)^2 = 1 - \sum_{k=1}^{c} \frac{1}{c^2} \\ &= 1 - c \times \frac{1}{c^2} = 1 - \frac{1}{c} = \frac{c-1}{c} \quad (q.e.d) \end{aligned} \tag{A.3}$$

Appendix B: Case study: variable description

Table B.1: Description of clinical variables (continuous nature).

Period	Parameter type	Variable	Description	Measure unit
Cross-sectional measurement	clinical assessment	age	Age	
		cigarettes	Number of packs smoked per day for the years of smoking	pack-year unit
		abdo.perim.cm	Abdominal perimeter	cm
		cerv.perim.cm	Neck circumference	cm
		bmi	Body Mass Index	kg/m^2
Morning after the test	respiratory	oxi.morn	Daytime measurement of peripheral blood oxygen saturation (morning pulse oximetry)	%
	metabolic	glyc	Blood sugar level (glicemia)	mg/dl
		hbglyc	Glycated hemoglobin	%
		insul	Insulin (hormone)	mlU/L (IU international units)
		homa.ir	Homeostasis model assessment of insulin resistance (reference value < 2.15)	its score (o units) †
		cholest	Total cholesterol	mg/dl
		ldl	Low-density lipoproteins (”bad” cholesterol)	mg/dl
		hdl	High-density lipoproteins (”good” cholesterol)	mg/dl
		trigl	Triglyceride levels	mg/dl
	cardiac	sis.bp	Systolic blood pressure	mmHg
		dias.bp	Diastolic blood pressure	mmHg
		hr	Heart rate	BPM (beats per minute)
		homoc	Homocysteine levels (biomarker for cardiovascular disease)	μmol/L
	hormonal	adren.u	Adrenaline levels in the urine	pg/mL
		adren.u24	Adrenaline levels in the urine (24 hours)	pg/mL
		noradren.u	Noradrenaline levels in the urine	pg/mL
		noradren.u24	Noradrenaline levels in the urine (24 hours)	pg/mL
		dop.u	Dopamine levels in the urine	pg/L
		dop.u24	Dopamine levels in the urine (24 hours)	pg/L

Table B.2: Description of clinical variables (categorical nature).

Period	**Parameter type**	**Variable**	**Description**	**Categories**
Polisomnography test	response variable	Y_2	OSA diagnosis defined by presence vs absence	C, OSA
		Y_4	OSA diagnosis defined by severity	C, OSA-I, OSA-II, OSA-III
Cross-sectional measurement	clinical assessment	smoking.habits	Smoking habits	0 - non-smoker; 1 - smoker; 2 - former smoker
		morn.head	Morning headaches	0 - no; 1 - yes
		awakenings	Awakenings during sleep	0 - no; 1 - yes
		card.path	Cardiac pathology.	0 - no; 1 - yes
		resp.path	Respiratory pathology	0 - no; 1 - asthma; 2 - chronic obstructive pulmonary disease (COPD); 3 - pulmonary neoplasm; 4 - sarcoidosis
		metab.path	Metabolic pathology	0 - no; 1 - dyslipidemia
		endoc.path	Endocrine pathology	0 - no; 1 - diabetes
Morning after the test	hormonal	tft	Thyroid function tests	0 - normal; 1 - hiper; 2 - hipo

Table B.3: Description of variables representing proteoform expression (continuous nature).

	Variable		
Proteoform description	**Evening measure**	**Morning measure**	**Variation**
ApoC1 des TP	EC1dTP	MC1dTP	dfC1dTP
Apo C1 n	EC1n	MC1n	dfC1n
ApoC2 des TQQPQQ	EC2dTQQPQQ	MC2dTQQPQQ	dfC2dTQQPQQ
Apo C3 nat	EC3n	MC3n	dfC3n
Apo C2	EC2n	MC2n	dfC2n
Apo C3 var1	EC3v2	MC3v1	dfC3v12
Apo C3 var2 des A	EC3v2dA	MC3v2dA	dfC3v2dA
Apo C3 var 2	EC3v3	MC3v2	dfC3v23
Apo C3 var 3 des A	EC3v3dA	MC3v3dA	dfC3v3dA
Apo C3 var 3	ECv4	MCv3	dfCv34
Apo C3 var 4	EC3v4	MC3v4	dfC3v4
Apo C3 var5	EC3v5	MC3v5	dfC3v5
Apo C3 var 6	EC3v6	MC3v6	dfC3v6
Apo C3 var 7	EC3v7	MC3v7	dfC3v7
Apo C3 var 8	EC3v8	MC3v8	dfC3v8
Apo C var 9	ECv10	MCv9	dfCv910
Apo AII-M %	EA2M	MA2M	dfA2M
Apo AII-MQ %	EA2MQ	MA2MQ	dfA2MQ
Apo AII-MTQ %	EA2MTQ	MA2MTQ	dfA2MTQ
Apo AII-D %	EA2D	MA2D	dfA2D
Apo AII-DQ %	EA2DQ	MA2DQ	dfA2DQ
Apo AII-D2Q %	EA2D2Q	MA2D2Q	dfA2D2Q

Appendix C: Case study: result tables

Table C.1: Descriptive analysis of clinical variables (continuous nature) and hypothesis tests for the binary outcome variable Y_2. Significance level for Mann-Whitney hypothesis test: $\alpha = 0.1$; * $0.05 \leq$ p-value < 0.1; ** $0.001 \leq$ p-value < 0.05; *** p-value < 0.001.

	Min ; Max		Quantiles (25;50;75)		Mean (Std)		Missing values		Mann-Whitney U test	
Variables	C	OSA	C	OSA	C	OSA	C	OSA	Statistic	p-value
abdo.perim.cm	91.0 ; 122.0	84.0 ; 132.0	93.0 ; 98.5 ; 104.0	99.0 ; 107.5 ; 114.0	100.9 (9.7)	107.4 (10.5)	0	5	234.0	0.020 **
bmi	23.9 ; 37.7	21.7 ; 42.5	25.6 ; 27.1 ; 29.1	27.7 ; 30.5 ; 32.2	28.3 (4.0)	30.3 (4.0)	0	0	272.0	0.031 **
cerv.perim.cm	38.0 ; 46.0	37.0 ; 50.0	38.9 ; 41.0 ; 42.6	41.0 ; 43.0 ; 44.6	41.1 (2.7)	42.8 (2.9)	0	5	251.5	0.040 **
adren.u	0.0 ; 27.7	0.0 ; 23.1	8.7 ; 9.3 ; 12.5	0.0 ; 6.8 ; 13.1	10.0 (7.0)	7.5 (6.8)	0	2	502.5	0.161
age	27.0 ; 59.0	32.0 ; 60.0	34.2 ; 46.5 ; 51.0	43.0 ; 49.0 ; 54.0	43.6 (10.3)	47.8 (7.2)	0	0	329.5	0.181
insul	3.3 ; 25.4	1.9 ; 51.4	8.5 ; 10.8 ; 13.9	8.2 ; 13.3 ; 21.9	11.8 (5.8)	16.0 (10.1)	0	0	333.5	0.201
adren.u24	0.0 ; 58.2	0.0 ; 34.8	11.4 ; 14.6 ; 17.9	0.0 ; 10.7 ; 18.6	15.6 (13.4)	10.8 (9.1)	0	2	482.0	0.273
homa.ir	0.7 ; 6.7	0.3 ; 13.7	2.0 ; 2.7 ; 3.6	1.8 ; 3.1 ; 6.8	3.0 (1.6)	4.2 (3.1)	0	0	355.5	0.334
oxi.morn	0.9 ; 1.0	0.9 ; 1.0	1.0 ; 1.0 ; 1.0	1.0 ; 1.0 ; 1.0	1.0 (0.0)	1.0 (0.0)	0	1	478.0	0.355
dop.u24	88.0 ; 884.2	33.5 ; 1247.4	254.7 ; 358.2 ; 415.7	212.5 ; 294.1 ; 424.1	373.2 (193.8)	350.2 (223.6)	0	1	480.0	0.359
hdl	30.0 ; 66.0	30.0 ; 73.0	37.8 ; 39.0 ; 48.0	38.0 ; 44.0 ; 49.0	43.0 (9.9)	44.9 (9.5)	0	0	362.0	0.381
trigl	50.0 ; 255.0	34.0 ; 428.0	74.5 ; 104.5 ; 169.0	94.0 ; 126.0 ; 184.0	124.6 (65.3)	152.2 (95.0)	0	0	362.0	0.382
dias.bp	63.0 ; 111.0	59.0 ; 116.7	78.0 ; 85.5 ; 94.2	75.0 ; 82.0 ; 90.0	86.0 (13.2)	83.4 (12.5)	0	0	476.5	0.459
cigarettes	0.0 ; 88.0	0.0 ; 60.0	0.0 ; 16.5 ; 30.0	0.0 ; 5.5 ; 23.5	19.9 (22.6)	14.3 (16.9)	0	5	429.0	0.481
dop.u	99.6 ; 556.1	67.0 ; 1386.0	152.0 ; 212.5 ; 289.3	124.4 ; 211.5 ; 296.3	253.3 (136.6)	259.7 (214.5)	0	1	450.5	0.623
homoc	10.4 ; 22.5	10.2 ; 51.7	11.8 ; 14.4 ; 17.3	12.8 ; 15.3 ; 17.5	15.4 (4.2)	16.1 (6.1)	0	1	384.0	0.649
cholest	100.0 ; 268.0	128.0 ; 303.0	156.0 ; 177.0 ; 214.8	166.0 ; 190.0 ; 208.0	186.0 (48.6)	191.6 (35.0)	0	0	395.0	0.685
sis.bp	110.0 ; 175.0	88.0 ; 181.0	125.2 ; 131.5 ; 144.2	120.0 ; 134.0 ; 143.0	137.4 (19.8)	133.1 (16.5)	0	0	448.5	0.733
noradren.u	17.0 ; 97.7	10.3 ; 263.9	30.5 ; 42.7 ; 61.9	25.8 ; 41.1 ; 62.5	47.6 (22.7)	52.7 (44.5)	0	1	439.0	0.745
ldl	57.0 ; 178.0	64.0 ; 222.0	92.5 ; 120.5 ; 139.2	96.5 ; 112.0 ; 135.2	118.2 (35.5)	116.8 (30.4)	0	1	435.5	0.784
hr	54.0 ; 90.0	48.0 ; 110.0	62.8 ; 67.5 ; 74.8	62.0 ; 70.0 ; 74.5	69.6 (10.6)	70.6 (11.9)	2	6	315.5	0.823
glyc	84.0 ; 218.0	67.0 ; 186.0	89.5 ; 93.5 ; 102.0	88.0 ; 97.0 ; 111.0	102.6 (31.8)	101.6 (22.2)	0	0	409.0	0.837
noradren.u24	26.4 ; 149.0	30.8 ; 237.5	46.2 ; 58.1 ; 86.8	41.6 ; 56.6 ; 84.8	69.3 (33.8)	71.5 (43.5)	0	1	430.0	0.845
hbglyc	4.9 ; 9.3	4.9 ; 8.5	5.4 ; 5.8 ; 5.8	5.3 ; 5.6 ; 5.9	5.8 (1.0)	5.8 (0.7)	0	0	433.5	0.898

Table C.2: Descriptive analysis of clinical variables (continuous nature).

	Min ; Max				Quantiles (25;50;75)				Mean (Std)				Missing values			
Variables	C	I	II	III	C	I	II	III	C	I	II	III	C	I	II	III
insul	3.3 ; 25.4	1.9 ; 42.0	6.7 ; 26.1	7.0 ; 51.4	8.5 ; 10.8 ; 13.9	7.6 ; 9.5 ; 13.1	9.9 ; 13.8 ; 15.9	21.0 ; 24.6 ; 28.7	11.8 (5.8)	11.7 (8.0)	13.7 (5.6)	24.6 (10.4)	0	0	0	0
homa.ir	0.7 ; 6.7	0.3 ; 13.7	1.4 ; 9.2	1.5 ; 11.6	2.0 ; 2.7 ; 3.6	1.5 ; 2.4 ; 3.2	2.1 ; 3.0 ; 3.4	6.2 ; 7.0 ; 8.3	3.0 (1.6)	3.0 (2.6)	3.4 (2.2)	6.9 (2.9)	0	0	0	0
cerv.perim.cm	38.0 ; 46.0	37.0 ; 48.0	39.0 ; 50.0	42.0 ; 48.0	38.9 ; 41.0 ; 42.6	39.8 ; 42.0 ; 43.2	40.0 ; 42.0 ; 44.0	43.0 ; 44.0 ; 45.5	41.1 (2.7)	41.7 (2.7)	42.9 (3.8)	44.4 (1.8)	0	3	1	1
abdo.perim.cm	91.0 ; 122.0	84.0 ; 123.0	90.0 ; 132.0	97.0 ; 128.0	93.0 ; 98.5 ; 104.0	97.0 ; 102.2 ; 112.2	99.0 ; 105.0 ; 108.0	109.0 ; 112.0 ; 117.0	100.9 (9.7)	104.0 (9.6)	106.7 (12.6)	113.4 (8.3)	0	3	1	1
bmi	23.9 ; 37.7	21.7 ; 37.5	24.7 ; 35.5	26.2 ; 42.5	25.6 ; 27.1 ; 29.1	26.6 ; 28.1 ; 31.7	27.8 ; 30.0 ; 31.8	30.2 ; 31.3 ; 34.7	28.3 (4.0)	29.1 (3.6)	29.9 (3.1)	32.7 (4.2)	0	0	0	0
trigl	50.0 ; 255.0	36.0 ; 311.0	34.0 ; 374.0	96.0 ; 428.0	74.5 ; 104.5 ; 169.0	76.5 ; 104.0 ; 142.5	97.2 ; 132.5 ; 305.8	128.0 ; 154.0 ; 197.0	124.6 (65.3)	120.9 (67.7)	182.5 (129.4)	186.2 (98.4)	0	0	0	0
glyc	84.0 ; 218.0	67.0 ; 151.0	77.0 ; 143.0	84.0 ; 186.0	89.5 ; 93.5 ; 102.0	88.0 ; 95.0 ; 105.0	83.2 ; 86.0 ; 100.5	94.0 ; 108.0 ; 121.0	102.6 (31.8)	97.7 (18.0)	93.9 (19.6)	113.1 (26.8)	0	0	0	0
hdl	30.0 ; 66.0	30.0 ; 73.0	33.0 ; 59.0	30.0 ; 54.0	37.8 ; 39.0 ; 48.0	40.5 ; 47.0 ; 50.5	40.5 ; 44.0 ; 49.0	35.2 ; 42.0 ; 44.0	43.0 (9.9)	47.6 (10.4)	45.2 (8.5)	40.1 (6.8)	0	0	0	0
sis.bp	110.0 ; 175.0	88.0 ; 157.0	120.0 ; 147.0	105.0 ; 181.0	125.2 ; 131.5 ; 144.2	117.0 ; 127.0 ; 142.0	127.2 ; 129.5 ; 135.5	130.5 ; 143.0 ; 148.0	137.4 (19.8)	129.5 (16.6)	131.1 (8.0)	140.5 (18.7)	0	0	0	0
noradren.u24	26.4 ; 149.0	31.2 ; 184.6	30.8 ; 87.6	37.3 ; 237.5	46.2 ; 58.1 ; 86.8	37.7 ; 50.2 ; 73.9	39.6 ; 48.2 ; 72.8	47.3 ; 80.2 ; 111.6	69.3 (33.8)	65.3 (38.5)	55.8 (20.6)	90.9 (54.9)	0	0	1	0
age	27.0 ; 59.0	33.0 ; 60.0	32.0 ; 58.0	38.0 ; 59.0	34.2 ; 46.5 ; 51.0	46.5 ; 49.0 ; 52.5	38.8 ; 43.0 ; 49.5	43.8 ; 49.0 ; 55.0	43.6 (10.3)	48.3 (6.5)	44.2 (8.6)	49.2 (7.2)	0	0	0	0
adren.u	0.0 ; 27.7	0.0 ; 23.1	0.0 ; 15.3	0.0 ; 20.4	8.7 ; 9.3 ; 12.5	0.0 ; 6.2 ; 12.8	0.0 ; 0.0 ; 12.9	4.3 ; 8.6 ; 12.7	10.0 (7.0)	7.5 (6.9)	5.2 (7.2)	8.7 (6.6)	0	0	2	0
dias.bp	63.0 ; 111.0	59.0 ; 116.7	62.0 ; 91.0	68.0 ; 112.0	78.0 ; 85.5 ; 94.2	72.0 ; 80.0 ; 89.5	73.8 ; 80.5 ; 86.0	81.0 ; 86.5 ; 96.0	86.0 (13.2)	82.2 (13.0)	79.3 (9.1)	88.1 (12.6)	0	0	0	0
adren.u24	0.0 ; 58.2	0.0 ; 21.7	0.0 ; 21.6	0.0 ; 34.8	11.4 ; 14.6 ; 17.9	0.0 ; 10.5 ; 18.4	0.0 ; 0.0 ; 19.4	7.2 ; 14.1 ; 18.8	15.6 (13.4)	10.1 (8.1)	7.7 (10.7)	13.4 (9.9)	0	0	2	0
cigarettes	0.0 ; 88.0	0.0 ; 40.0	0.0 ; 50.0	0.0 ; 60.0	0.0 ; 16.5 ; 30.0	0.0 ; 4.0 ; 22.8	10.2 ; 20.0 ; 25.0	0.0 ; 7.5 ; 27.5	19.9 (22.6)	10.7 (13.7)	20.9 (16.8)	17.2 (21.4)	0	1	2	2
hbglyc	4.9 ; 9.3	4.9 ; 7.7	5.1 ; 7.1	5.1 ; 8.5	5.4 ; 5.8 ; 5.8	5.3 ; 5.6 ; 5.8	5.3 ; 5.5 ; 5.8	5.5 ; 5.8 ; 6.3	5.8 (1.0)	5.7 (0.6)	5.7 (0.6)	6.1 (0.9)	0	0	0	0
noradren.u	17.0 ; 97.7	15.7 ; 111.9	23.8 ; 62.6	10.3 ; 263.9	30.5 ; 42.7 ; 61.9	24.9 ; 35.5 ; 61.3	29.3 ; 36.4 ; 45.0	30.4 ; 48.2 ; 76.1	47.6 (22.7)	45.8 (27.4)	39.7 (13.8)	71.8 (68.8)	0	0	1	0
dop.u24	88.0 ; 884.2	127.9 ; 876.3	97.7 ; 502.5	33.5 ; 1247.4	254.7 ; 358.2 ; 415.7	198.2 ; 280.1 ; 418.8	160.8 ; 266.7 ; 481.9	263.9 ; 297.9 ; 423.3	373.2 (193.8)	334.5 (189.2)	302.2 (158.3)	403.6 (299.4)	0	0	1	0
oxi.morn	0.9 ; 1.0	0.9 ; 1.0	0.9 ; 1.0	1.0 ; 1.0	1.0 ; 1.0 ; 1.0	1.0 ; 1.0 ; 1.0	1.0 ; 1.0 ; 1.0	1.0 ; 1.0 ; 1.0	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0	1	0	0
dop.u	99.6 ; 556.1	98.7 ; 793.8	69.8 ; 344.2	67.0 ; 1386.0	152.0 ; 212.5 ; 289.3	111.3 ; 206.3 ; 289.3	134.0 ; 248.6 ; 289.9	159.8 ; 220.5 ; 357.8	253.3 (136.6)	240.7 (166.8)	214.6 (97.8)	317.0 (312.6)	0	0	1	0
ldl	57.0 ; 178.0	78.0 ; 222.0	66.0 ; 176.0	64.0 ; 145.0	92.5 ; 120.5 ; 139.2	103.0 ; 114.0 ; 128.5	89.8 ; 100.0 ; 129.2	100.5 ; 126.0 ; 136.5	118.2 (35.5)	120.7 (32.0)	109.3 (32.2)	115.0 (26.8)	0	0	0	1
hr	54.0 ; 90.0	48.0 ; 110.0	61.0 ; 91.0	57.0 ; 95.0	62.8 ; 67.5 ; 74.8	59.0 ; 72.0 ; 75.5	62.0 ; 68.0 ; 71.0	63.5 ; 69.0 ; 77.5	69.6 (10.6)	71.0 (14.2)	68.7 (9.3)	71.1 (10.0)	2	4	1	1
homoc	10.4 ; 22.5	11.1 ; 51.7	11.1 ; 19.6	10.2 ; 19.8	11.8 ; 14.4 ; 17.3	13.1 ; 15.2 ; 18.5	12.9 ; 14.3 ; 19.4	12.3 ; 15.8 ; 16.8	15.4 (4.2)	17.1 (7.9)	15.4 (3.4)	14.8 (2.8)	0	0	1	0
cholest	100.0 ; 268.0	128.0 ; 303.0	138.0 ; 249.0	134.0 ; 231.0	156.0 ; 177.0 ; 214.8	168.5 ; 185.0 ; 205.5	163.2 ; 188.5 ; 209.8	176.8 ; 196.0 ; 208.0	186.0 (48.6)	192.4 (39.4)	190.9 (38.8)	190.7 (25.5)	0	0	0	0

Table C.3: Hypothesis tests for the multiclass outcome variable Y_4.
Significance level for Kruskal-Wallis hypothesis test: $\alpha_k = 0.1$; Degrees of freedom: 3;
Significance level for Dunn's hypothesis test: $\alpha = \frac{\alpha_k}{2} = 0.05$;
* $0.05 \leq$ p-value < 0.1; ** $0.001 \leq$ p-value < 0.05; *** p-value < 0.001.

	Kruskal-Wallis test		Dunn's Test (Multiple Pairwise Comparison)					
Variables	Statistic	p-value	C - I	C - II	I - II	C - III	I - III	II - III
insul	18.8	< 0.001 ***	1.000	1.000	0.755	0.002 **	< 0.001 ***	0.086 *
homa.ir	18.1	< 0.001 ***	1.000	1.000	1.000	0.004 **	< 0.001 ***	0.04 **
cerv.perim.cm	13.6	0.004 **	1.000	0.632	1.000	0.002 **	0.01 **	0.312
abdo.perim.cm	13.1	0.004 **	0.796	0.601	1.000	0.001 **	0.02 **	0.262
bmi	11.5	0.009 **	0.917	0.481	1.000	0.004 **	0.028 **	0.456
trigl	8.3	0.041 **	1.000	1.000	0.572	0.099 *	0.021 **	1.000
glyc	8.1	0.043 **	1.000	0.633	0.660	0.266	0.116	0.018 **
hdl	6.5	0.090 *	0.302	1.000	1.000	1.000	0.043 **	0.407
sis.bp	4.6	0.207	-	-	-	-	-	-
noradren.u24	4.4	0.225	-	-	-	-	-	-
age	4.3	0.232	-	-	-	-	-	-
adren.u	4.1	0.250	-	-	-	-	-	-
dias.bp	3.6	0.309	-	-	-	-	-	-
adren.u24	3.3	0.343	-	-	-	-	-	-
cigarettes	3.0	0.388	-	-	-	-	-	-
hbglyc	3.0	0.394	-	-	-	-	-	-
noradren.u	2.3	0.509	-	-	-	-	-	-
dop.u24	1.6	0.662	-	-	-	-	-	-
oxi.morn	1.1	0.787	-	-	-	-	-	-
dop.u	1.0	0.797	-	-	-	-	-	-
ldl	1.0	0.806	-	-	-	-	-	-
hr	0.7	0.882	-	-	-	-	-	-
homoc	0.5	0.929	-	-	-	-	-	-
cholest	0.3	0.965	-	-	-	-	-	-

Table C.4: Descriptive analysis of evening proteoform expression (continuous nature) and hypothesis tests for the binary outcome variable Y_2. Significance level for Mann-Whitney hypothesis test: $\alpha = 0.1$; * $0.05 \leq$ p-value < 0.1; ** $0.001 \leq$ p-value < 0.05; *** p-value < 0.001.

	Min ; Max		Quantiles (25;50;75)		Mean (Std)		Missing values		Mann-Whitney U test	
Variables	C	OSA	C	OSA	C	OSA	C	OSA	Statistic	p-value
EA2DQ	2.5 ; 9.2	1.6 ; 9.9	4.0 ; 6.2 ; 7.8	4.2 ; 5.3 ; 6.0	5.9 (2.2)	5.3 (1.7)	0	0	524	0.157
EC3v7	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0	0	329	0.179
EC3v8	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0	0	508	0.235
EA2D2Q	0.2 ; 5.2	0.0 ; 3.9	0.9 ; 1.4 ; 2.3	0.8 ; 1.1 ; 1.8	1.7 (1.2)	1.4 (1.0)	0	0	486	0.382
ECv4	0.1 ; 1.1	0.1 ; 2.2	0.2 ; 0.3 ; 0.4	0.2 ; 0.3 ; 0.4	0.4 (0.2)	0.4 (0.4)	0	0	481	0.422
EC3n	0.0 ; 1.1	0.0 ; 1.8	0.1 ; 0.1 ; 0.2	0.1 ; 0.1 ; 0.2	0.2 (0.2)	0.2 (0.3)	0	0	368	0.430
EC3v5	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0	0	378	0.518
EC3v4	0.0 ; 0.3	0.0 ; 0.7	0.0 ; 0.1 ; 0.1	0.0 ; 0.0 ; 0.1	0.1 (0.1)	0.1 (0.1)	0	0	469	0.527
EA2M	36.9 ; 66.9	33.1 ; 74.0	48.9 ; 53.5 ; 63.6	50.3 ; 56.3 ; 60.4	54.5 (8.6)	55.6 (7.7)	0	0	381	0.546
ECv10	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0	0	465	0.565
EC3v3dA	0.0 ; 0.8	0.0 ; 1.1	0.1 ; 0.2 ; 0.3	0.1 ; 0.2 ; 0.2	0.2 (0.2)	0.2 (0.2)	0	0	463	0.584
EC2dTQQPQQ	0.1 ; 0.4	0.0 ; 1.6	0.1 ; 0.2 ; 0.2	0.1 ; 0.2 ; 0.3	0.2 (0.1)	0.2 (0.2)	0	0	388	0.614
EC3v6	0.0 ; 0.1	0.0 ; 0.1	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0	0	456	0.654
EA2MTQ	0.0 ; 3.0	0.0 ; 3.4	0.3 ; 0.4 ; 1.0	0.2 ; 0.4 ; 1.0	0.7 (0.7)	0.7 (0.6)	0	0	450	0.717
EC2n	0.7 ; 6.0	0.4 ; 18.0	1.8 ; 2.6 ; 3.9	1.9 ; 2.8 ; 3.7	2.9 (1.5)	3.5 (3.2)	0	0	407	0.815
EC3v3	0.3 ; 3.3	0.3 ; 7.9	0.8 ; 1.2 ; 1.6	0.8 ; 1.0 ; 1.6	1.3 (0.8)	1.5 (1.4)	0	0	439	0.837
EA2MQ	17.7 ; 45.1	17.4 ; 52.7	24.8 ; 30.2 ; 33.5	26.4 ; 29.8 ; 35.2	29.9 (7.3)	30.5 (6.9)	0	0	413	0.881
EC1n	0.4 ; 2.7	0.4 ; 8.7	1.1 ; 1.3 ; 1.9	0.9 ; 1.3 ; 2.0	1.5 (0.6)	1.8 (1.7)	0	0	417	0.926
EC3v2dA	0.0 ; 0.2	0.0 ; 0.6	0.1 ; 0.1 ; 0.1	0.1 ; 0.1 ; 0.1	0.1 (0.1)	0.1 (0.1)	0	0	430	0.938
EC3v2	0.2 ; 2.1	0.1 ; 5.4	0.5 ; 0.8 ; 1.2	0.6 ; 0.8 ; 1.1	0.9 (0.5)	1.1 (1.0)	0	0	419	0.949
EC1dTP	0.3 ; 1.2	0.2 ; 5.5	0.4 ; 0.6 ; 1.0	0.5 ; 0.6 ; 0.9	0.7 (0.3)	0.8 (0.9)	0	0	426	0.983
EA2D	2.5 ; 12.3	2.1 ; 12.6	5.7 ; 7.0 ; 7.9	5.4 ; 7.4 ; 8.6	7.2 (2.6)	7.0 (2.3)	0	0	423	0.994

Table C.5: Descriptive analysis of evening proteoform expression (continuous nature) and hypothesis tests for the multiclass outcome variable Y_4.
Significance level for Kruskal-Wallis hypothesis test: $\alpha = 0.1$; Degrees of freedom:3; * $0.05 \leq$ p-value < 0.1; ** $0.001 \leq$ p-value < 0.05; *** p-value < 0.001.

	Min ; Max				Quantiles (25;50;75)				Mean (Std)				Missing values				Kruskal-Wallis test	
Variables	C	I	II	III	C	I	II	III	C	I	II	III	C	I	II	III	Test statistiic	p-value
EA2D	2.5 ; 12.3	2.1 ; 12.6	3.7 ; 9.2	4.3 ; 11.2	5.7 ; 7.0 ; 7.9	4.4 ; 6.0 ; 8.4	5.9 ; 6.8 ; 8.3	7.5 ; 8.0 ; 9.3	7.2 (2.6)	6.4 (2.6)	6.9 (1.7)	8.0 (1.9)	0	0	0	0	5.1	0.163
EA2MTQ	0.0 ; 3.0	0.0 ; 3.4	0.1 ; 1.2	0.0 ; 1.6	0.3 ; 0.4 ; 1.0	0.2 ; 0.7 ; 1.2	0.2 ; 0.6 ; 1.0	0.1 ; 0.3 ; 0.8	0.7 (0.7)	0.8 (0.8)	0.6 (0.4)	0.4 (0.4)	0	0	0	0	4.4	0.219
EC3v7	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0	0	0	0	3.8	0.289
EA2D2Q	0.2 ; 5.2	0.0 ; 3.9	0.4 ; 3.4	0.2 ; 2.2	0.9 ; 1.4 ; 2.3	1.0 ; 1.3 ; 2.1	0.8 ; 1.0 ; 1.8	0.7 ; 0.9 ; 1.6	1.7 (1.2)	1.6 (1.1)	1.4 (0.9)	1.1 (0.6)	0	0	0	0	3.3	0.341
EA2DQ	2.5 ; 9.2	2.0 ; 9.1	3.1 ; 9.9	1.6 ; 6.0	4.0 ; 6.2 ; 7.8	4.4 ; 5.4 ; 6.4	5.0 ; 5.6 ; 6.7	4.1 ; 5.2 ; 5.8	5.9 (2.2)	5.4 (1.7)	5.7 (2.0)	4.8 (1.4)	0	0	0	0	3.3	0.344
EA2MQ	17.7 ; 45.1	17.4 ; 52.7	20.5 ; 40.7	20.5 ; 38.1	24.8 ; 30.2 ; 33.5	26.8 ; 30.1 ; 38.0	26.8 ; 29.6 ; 33.2	23.8 ; 27.7 ; 31.1	29.9 (7.3)	32.2 (7.7)	29.8 (5.9)	28.1 (5.4)	0	0	0	0	3.1	0.375
EC3n	0.0 ; 1.1	0.0 ; 1.8	0.0 ; 0.8	0.0 ; 0.6	0.1 ; 0.1 ; 0.2	0.1 ; 0.1 ; 0.2	0.1 ; 0.1 ; 0.2	0.1 ; 0.2 ; 0.2	0.2 (0.2)	0.2 (0.3)	0.2 (0.2)	0.2 (0.2)	0	0	0	0	2.7	0.442
EA2M	36.9 ; 66.9	33.1 ; 74.0	45.2 ; 66.7	49.6 ; 68.3	48.9 ; 53.5 ; 63.6	47.9 ; 55.6 ; 59.9	51.2 ; 56.1 ; 58.2	52.7 ; 58.1 ; 61.1	54.5 (8.6)	54.5 (8.9)	55.6 (7.1)	57.6 (5.6)	0	0	0	0	2.1	0.548
EC3v3dA	0.0 ; 0.8	0.0 ; 1.1	0.0 ; 0.6	0.1 ; 0.9	0.1 ; 0.2 ; 0.3	0.1 ; 0.2 ; 0.2	0.1 ; 0.2 ; 0.2	0.1 ; 0.2 ; 0.3	0.2 (0.2)	0.2 (0.2)	0.2 (0.2)	0.3 (0.2)	0	0	0	0	2.1	0.551
EC3v4	0.0 ; 0.3	0.0 ; 0.7	0.0 ; 0.1	0.0 ; 0.3	0.0 ; 0.1 ; 0.1	0.0 ; 0.0 ; 0.1	0.0 ; 0.1 ; 0.1	0.0 ; 0.0 ; 0.1	0.1 (0.1)	0.1 (0.2)	0.1 (0.0)	0.1 (0.1)	0	0	0	0	2.1	0.554
EC3v3	0.3 ; 3.3	0.3 ; 7.9	0.3 ; 3.3	0.7 ; 5.6	0.8 ; 1.2 ; 1.6	0.7 ; 1.0 ; 1.6	0.6 ; 0.9 ; 1.4	0.8 ; 1.3 ; 1.7	1.3 (0.8)	1.4 (1.6)	1.2 (0.9)	1.7 (1.4)	0	0	0	0	1.7	0.641
EC3v8	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0	0	0	0	1.5	0.691
EC3v5	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0	0	0	0	1.4	0.707
EC3v6	0.0 ; 0.1	0.0 ; 0.1	0.0 ; 0.0	0.0 ; 0.1	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0	0	0	0	1.1	0.771
ECv10	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0	0	0	0	1.0	0.799
EC3v2dA	0.0 ; 0.2	0.0 ; 0.3	0.0 ; 0.2	0.0 ; 0.6	0.1 ; 0.1 ; 0.1	0.1 ; 0.1 ; 0.1	0.0 ; 0.1 ; 0.1	0.1 ; 0.1 ; 0.1	0.1 (0.1)	0.1 (0.1)	0.1 (0.0)	0.1 (0.1)	0	0	0	0	0.8	0.840
EC3v2	0.2 ; 2.1	0.1 ; 5.4	0.2 ; 2.4	0.4 ; 3.4	0.5 ; 0.8 ; 1.2	0.6 ; 0.8 ; 1.1	0.5 ; 0.7 ; 1.0	0.6 ; 1.0 ; 1.2	0.9 (0.5)	1.0 (1.1)	0.9 (0.6)	1.2 (0.9)	0	0	0	0	0.8	0.850
ECv4	0.1 ; 1.1	0.1 ; 2.2	0.1 ; 0.5	0.1 ; 1.3	0.2 ; 0.3 ; 0.4	0.2 ; 0.3 ; 0.4	0.2 ; 0.3 ; 0.4	0.2 ; 0.2 ; 0.4	0.4 (0.2)	0.4 (0.5)	0.3 (0.2)	0.4 (0.3)	0	0	0	0	0.8	0.857
EC2n	0.7 ; 6.0	0.4 ; 18.0	0.9 ; 8.4	1.5 ; 9.8	1.8 ; 2.6 ; 3.9	2.0 ; 2.8 ; 3.2	1.8 ; 2.3 ; 3.8	1.9 ; 3.1 ; 3.9	2.9 (1.5)	3.6 (4.0)	3.1 (2.1)	3.8 (2.6)	0	0	0	0	0.6	0.902
EC2dTQQPQQ	0.1 ; 0.4	0.0 ; 1.6	0.0 ; 0.4	0.1 ; 0.7	0.1 ; 0.2 ; 0.2	0.1 ; 0.2 ; 0.2	0.1 ; 0.2 ; 0.3	0.1 ; 0.2 ; 0.3	0.2 (0.1)	0.3 (0.3)	0.2 (0.1)	0.2 (0.2)	0	0	0	0	0.4	0.942
EC1dTP	0.3 ; 1.2	0.2 ; 5.5	0.2 ; 1.4	0.2 ; 2.5	0.4 ; 0.6 ; 1.0	0.5 ; 0.6 ; 1.0	0.5 ; 0.7 ; 0.9	0.4 ; 0.6 ; 0.8	0.7 (0.3)	1.0 (1.2)	0.7 (0.4)	0.8 (0.7)	0	0	0	0	0.2	0.974
EC1n	0.4 ; 2.7	0.4 ; 8.7	0.4 ; 2.7	0.7 ; 5.8	1.1 ; 1.3 ; 1.9	1.0 ; 1.3 ; 2.1	0.9 ; 1.3 ; 2.1	1.1 ; 1.4 ; 2.0	1.5 (0.6)	1.9 (2.0)	1.5 (0.8)	1.9 (1.5)	0	0	0	0	0.2	0.982

Table C.6: Descriptive and univariate analysis of difference proteoform expression (dpe) for the two-category diagnosis of OSA.

	Min ; Max		Quantiles (25;50;75)		Mean (Std)		Missing values		Mann-Whitney U test	
Variables	C	OSA	C	OSA	C	OSA	C	OSA	Statistic	p-value
dfA2DQ	-2.5 ; 3.6	-2.2 ; 5.2	-1.1 ; -0.5 ; 0.3	0.0 ; 1.1 ; 2.2	-0.4 (1.5)	1.1 (1.6)	1	1	175	0.001 **
dfA2D2Q	-1.6 ; 2.0	-1.6 ; 2.5	-0.9 ; -0.3 ; -0.1	-0.5 ; 0.0 ; 0.6	-0.3 (1.0)	0.2 (0.8)	1	1	262	0.055 *
dfA2M	-6.2 ; 8.6	-30.7 ; 6.1	-1.7 ; -0.5 ; 2.5	-5.4 ; -1.1 ; 0.4	0.3 (4.2)	-2.8 (5.7)	1	1	506	0.082 *
dfC3v7	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	1	1	467	0.250
dfC3v5	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	1	1	456	0.325
dfC3v8	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	1	1	328	0.355
dfCv910	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	1	1	345	0.503
dfA2D	-4.7 ; 5.6	-3.7 ; 5.4	-0.7 ; 0.1 ; 2.3	-0.4 ; 1.1 ; 2.7	0.6 (2.7)	1.1 (2.1)	1	1	350	0.552
dfC3v3dA	-0.7 ; 0.4	-1.0 ; 0.9	-0.1 ; -0.1 ; 0.1	-0.1 ; 0.0 ; 0.1	-0.1 (0.3)	0.0 (0.3)	1	1	355	0.604
dfCv34	-0.9 ; 0.8	-2.0 ; 0.9	-0.1 ; -0.1 ; 0.1	-0.2 ; 0.0 ; 0.2	0.0 (0.4)	0.0 (0.5)	1	1	363	0.690
dfA2MTQ	-1.9 ; 1.5	-1.9 ; 1.2	-0.6 ; -0.3 ; 0.2	-0.4 ; -0.1 ; 0.1	-0.2 (0.8)	-0.2 (0.5)	1	1	368	0.746
dfA2MQ	-11.6 ; 9.6	-7.8 ; 10.8	-0.4 ; 0.4 ; 1.1	-1.0 ; 0.3 ; 1.9	0.0 (4.4)	0.1 (3.1)	1	1	407	0.804
dfC3v6	-0.1 ; 0.0	-0.1 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	1	1	406	0.816
dfC2dTQQPQQ	-0.4 ; 0.2	-1.5 ; 0.7	-0.1 ; 0.0 ; 0.1	-0.1 ; 0.0 ; 0.1	0.0 (0.1)	0.0 (0.3)	1	1	405	0.827
dfC2n	-4.9 ; 4.9	-15.9 ; 11.3	-1.4 ; -0.1 ; 1.6	-1.9 ; -0.5 ; 2.2	0.0 (2.8)	-0.2 (4.3)	1	1	404	0.839
dfC3v2dA	-0.2 ; 0.2	-0.5 ; 0.7	-0.1 ; 0.0 ; 0.0	-0.1 ; 0.0 ; 0.0	0.0 (0.1)	0.0 (0.2)	1	1	403	0.851
dfC3v4	-0.2 ; 0.2	-0.6 ; 0.2	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.1)	0.0 (0.1)	1	1	399	0.898
dfC1dTP	-0.9 ; 0.6	-4.9 ; 1.1	-0.2 ; -0.1 ; 0.2	-0.4 ; -0.1 ; 0.3	-0.1 (0.4)	-0.2 (1.0)	1	1	397	0.922
dfC3n	-1.0 ; 0.5	-1.6 ; 0.5	-0.1 ; -0.1 ; 0.0	-0.1 ; 0.0 ; 0.0	-0.1 (0.3)	-0.1 (0.3)	1	1	397	0.922
dfC3v23	-2.9 ; 1.9	-7.2 ; 4.9	-0.5 ; -0.2 ; 1.1	-0.6 ; 0.0 ; 1.0	0.0 (1.3)	0.0 (1.9)	1	1	385	0.946
dfC1n	-1.8 ; 1.1	-7.5 ; 2.8	-0.8 ; -0.4 ; 0.4	-1.2 ; -0.2 ; 0.8	-0.2 (0.9)	-0.4 (2.0)	1	1	388	0.982
dfC3v12	-1.8 ; 1.5	-4.8 ; 2.8	-0.4 ; -0.1 ; 0.5	-0.5 ; 0.0 ; 0.6	0.0 (0.9)	0.0 (1.3)	1	1	388	0.982

Table C.7: Descriptive analysis of difference proteoform expression (dpe) for the four-category diagnosis of OSA.

	Min ; Max				Quantiles (25;50;75)				Mean (Std)				Missing values			
Variables	C	I	II	III	C	I	II	III	C	I	II	III	C	I	II	III
dfA2DQ	-2.5 ; 3.6	-2.2 ; 5.2	-1.0 ; 3.4	-1.0 ; 3.5	-1.1 ; -0.5 ; 0.3	0.1 ; 1.3 ; 2.5	0.4 ; 1.7 ; 2.2	0.0 ; 0.9 ; 1.2	-0.4 (1.5)	1.2 (1.9)	1.4 (1.5)	0.8 (1.2)	1	1	0	0
dfA2M	-6.2 ; 8.6	-30.7 ; 3.2	-11.6 ; 1.6	-6.4 ; 6.1	-1.7 ; -0.5 ; 2.5	-6.2 ; -1.5 ; 0.1	-5.9 ; -3.6 ; -0.2	-1.2 ; 0.0 ; 1.1	0.3 (4.2)	-3.9 (7.0)	-3.7 (4.2)	-0.4 (3.0)	1	1	0	0
dfA2D2Q	-1.6 ; 2.0	-0.9 ; 2.5	-0.6 ; 1.9	-1.6 ; 0.9	-0.9 ; -0.3 ; -0.1	-0.3 ; 0.2 ; 1.0	-0.4 ; 0.3 ; 0.8	-0.6 ; -0.2 ; 0.0	-0.3 (1.0)	0.3 (0.9)	0.3 (0.8)	-0.2 (0.6)	1	1	0	0
dfC3v7	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1	1	0	0
dfA2MQ	-11.6 ; 9.6	-7.6 ; 10.8	-2.0 ; 4.0	-7.8 ; 3.0	-0.4 ; 0.4 ; 1.1	-0.8 ; 0.6 ; 2.2	-0.6 ; 0.4 ; 0.9	-1.2 ; -0.7 ; 0.9	0.0 (4.4)	0.7 (3.4)	0.3 (1.7)	-0.9 (3.0)	1	1	0	0
dfC3v5	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1	1	0	0
dfC3v4	-0.2 ; 0.2	-0.6 ; 0.2	-0.1 ; 0.1	-0.2 ; 0.1	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.1	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.1)	0.0 (0.2)	0.0 (0.0)	0.0 (0.1)	1	1	0	0
dfC3v6	-0.1 ; 0.0	-0.1 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1	1	0	0
dfCv910	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1	1	0	0
dfC3v3dA	-0.7 ; 0.4	-1.0 ; 0.9	-0.6 ; 0.2	-0.7 ; 0.6	-0.1 ; -0.1 ; 0.1	-0.1 ; 0.0 ; 0.2	-0.1 ; 0.0 ; 0.0	-0.2 ; -0.1 ; 0.0	-0.1 (0.3)	0.0 (0.4)	-0.1 (0.2)	0.0 (0.3)	1	1	0	0
dfA2MTQ	-1.9 ; 1.5	-1.9 ; 1.2	-0.7 ; 1.0	-0.9 ; 0.4	-0.6 ; -0.3 ; 0.2	-0.4 ; -0.2 ; 0.0	-0.4 ; -0.1 ; 0.1	-0.2 ; 0.0 ; 0.0	-0.2 (0.8)	-0.3 (0.6)	-0.1 (0.5)	-0.1 (0.4)	1	1	0	0
dfC3v2dA	-0.2 ; 0.2	-0.3 ; 0.7	-0.1 ; 0.1	-0.5 ; 0.3	-0.1 ; 0.0 ; 0.0	-0.1 ; 0.0 ; 0.1	-0.1 ; 0.0 ; 0.0	-0.1 ; -0.1 ; 0.0	0.0 (0.1)	0.0 (0.2)	0.0 (0.1)	0.0 (0.2)	1	1	0	0
dfA2D	-4.7 ; 5.6	-3.5 ; 4.8	-1.9 ; 5.4	-3.7 ; 5.3	-0.7 ; 0.1 ; 2.3	-0.3 ; 1.0 ; 2.2	-0.1 ; 2.1 ; 2.9	-0.8 ; 1.1 ; 2.7	0.6 (2.7)	0.9 (1.9)	1.7 (2.2)	0.8 (2.6)	1	1	0	0
dfC3n	-1.0 ; 0.5	-1.6 ; 0.5	-0.7 ; 0.3	-0.4 ; 0.5	-0.1 ; -0.1 ; 0.0	-0.1 ; 0.0 ; 0.1	-0.1 ; 0.0 ; 0.0	-0.2 ; -0.1 ; 0.0	-0.1 (0.3)	-0.1 (0.4)	-0.1 (0.3)	-0.1 (0.2)	1	1	0	0
dfC3v8	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 ; 0.0 ; 0.0	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1	1	0	0
dfC2n	-4.9 ; 4.9	-15.9 ; 11.3	-7.4 ; 3.9	-6.2 ; 6.0	-1.4 ; -0.1 ; 1.6	-1.6 ; -0.5 ; 2.2	-1.8 ; -0.4 ; 1.2	-2.1 ; -0.5 ; 2.9	0.0 (2.8)	-0.3 (5.2)	-0.6 (3.0)	0.1 (3.7)	1	1	0	0
dfCv34	-0.9 ; 0.8	-2.0 ; 0.9	-0.4 ; 0.3	-1.0 ; 0.6	-0.1 ; -0.1 ; 0.1	-0.2 ; -0.1 ; 0.2	-0.2 ; 0.0 ; 0.2	-0.1 ; 0.0 ; 0.2	0.0 (0.4)	-0.1 (0.6)	0.0 (0.2)	0.0 (0.4)	1	1	0	0
dfC2dTQQPQQ	-0.4 ; 0.2	-1.5 ; 0.7	-0.4 ; 0.2	-0.5 ; 0.3	-0.1 ; 0.0 ; 0.1	-0.1 ; 0.0 ; 0.1	-0.1 ; 0.0 ; 0.1	-0.1 ; 0.0 ; 0.2	0.0 (0.1)	0.0 (0.4)	0.0 (0.2)	0.0 (0.2)	1	1	0	0
dfC1n	-1.8 ; 1.1	-7.5 ; 2.8	-2.3 ; 1.4	-4.6 ; 2.4	-0.8 ; -0.4 ; 0.4	-1.0 ; -0.2 ; 0.6	-1.4 ; 0.0 ; 0.7	-1.1 ; -0.2 ; 1.2	-0.2 (0.9)	-0.6 (2.2)	-0.3 (1.3)	-0.3 (1.9)	1	1	0	0
dfC1dTP	-0.9 ; 0.6	-4.9 ; 1.1	-1.2 ; 0.6	-2.0 ; 1.1	-0.2 ; -0.1 ; 0.2	-0.4 ; -0.1 ; 0.2	-0.6 ; 0.0 ; 0.3	-0.4 ; -0.1 ; 0.4	-0.1 (0.4)	-0.3 (1.2)	-0.1 (0.6)	-0.1 (0.8)	1	1	0	0
dfC3v23	-2.9 ; 1.9	-7.2 ; 4.9	-2.8 ; 1.2	-4.5 ; 3.6	-0.5 ; -0.2 ; 1.1	-0.6 ; 0.0 ; 1.0	-0.5 ; 0.0 ; 0.5	-0.8 ; -0.2 ; 1.2	0.0 (1.3)	0.0 (2.2)	-0.1 (1.1)	0.1 (2.0)	1	1	0	0
dfC3v12	-1.8 ; 1.5	-4.8 ; 2.8	-2.1 ; 1.0	-2.4 ; 2.2	-0.4 ; -0.1 ; 0.5	-0.4 ; 0.0 ; 0.7	-0.5 ; 0.0 ; 0.3	-0.6 ; -0.1 ; 0.8	0.0 (0.9)	0.0 (1.5)	-0.1 (0.8)	0.1 (1.2)	1	1	0	0

Table C.8: Univariate analysis of difference proteoform expression (dpe) for the four-category diagnosis of OSA. Degrees of freedom: 3;

	Kruskal-Wallis test		Dunn's Test (Multiple Pairwise Comparison)					
Variables	Statistic	p-value	C - I	C - II	I - II	C - III	I - III	II - III
dfA2DQ	11.5	0.009 **	0.007 **	0.015 **	1.000	0.105	1.000	1.000
dfA2M	7.7	0.053 *	0.119	0.140	1.000	1.000	0.180	0.196
dfA2D2Q	7.4	0.060 *	0.07 *	0.192	1.000	1.000	0.217	0.428
dfC3v7	3.4	0.331	-	-	-	-	-	-
dfA2MQ	2.5	0.481	-	-	-	-	-	-
dfC3v5	2.3	0.519	-	-	-	-	-	-
dfC3v4	2.1	0.558	-	-	-	-	-	-
dfC3v6	1.7	0.630	-	-	-	-	-	-
dfCv910	1.6	0.670	-	-	-	-	-	-
dfC3v3dA	1.5	0.693	-	-	-	-	-	-
dfA2MTQ	1.4	0.704	-	-	-	-	-	-
dfC3v2dA	1.4	0.716	-	-	-	-	-	-
dfA2D	1.1	0.778	-	-	-	-	-	-
dfC3n	0.9	0.814	-	-	-	-	-	-
dfC3v8	0.9	0.816	-	-	-	-	-	-
dfC2n	0.3	0.968	-	-	-	-	-	-
dfCv34	0.2	0.971	-	-	-	-	-	-
dfC2dTQQPQQ	0.2	0.976	-	-	-	-	-	-
dfC1n	0.2	0.978	-	-	-	-	-	-
dfC1dTP	0.2	0.984	-	-	-	-	-	-
dfC3v23	0.1	0.991	-	-	-	-	-	-
dfC3v12	0.1	0.993	-	-	-	-	-	-

Table C.9: Estimated univariate potential of variables, based on the hypothesis tests p-value, to discriminate the classes of the binary outcome variable Y_2. †: variable of qualitative nature.

Variable		p-value		Rank	Variable		p-value	Rank
dfA2DQ		0.001	**	1	EC3v3dA		0.584	39
abdo.perim.cm		0.020	**	2	dfC3v3dA		0.604	40
bmi		0.031	**	3	EC2dTQQPQQ		0.614	41
cerv.perim.cm		0.040	**	4	dop.u		0.623	42
dfA2D2Q		0.055	*	5	homoc		0.649	43
tft	†	0.060	*	6	EC3v6		0.654	44
dfA2M		0.082	*	7	endoc.path	†	0.660	45
EA2DQ		0.157		8	cholest		0.685	46
adren.u		0.161		9	dfCv34		0.690	47
EC3v7		0.179		10	EA2MTQ		0.717	48
age		0.181		11	sis.bp		0.733	49
insul		0.201		12	noradren.u		0.745	50
EC3v8		0.235		13	dfA2MTQ		0.746	51
dfC3v7		0.250		14	ldl		0.784	52
metab.path	†	0.250		15	dfA2MQ		0.804	53
adren.u24		0.273		16	EC2n		0.815	54
smoking.habits	†	0.320		17	dfC3v6		0.816	55
dfC3v5		0.325		18	hr		0.823	56
homa.ir		0.334		19	dfC2dTQQPQQ		0.827	57
dfC3v8		0.355		20	glyc		0.837	58
oxi.morn		0.355		21	EC3v3		0.837	59
dop.u24		0.359		22	dfC2n		0.839	60
awakenings	†	0.370		23	noradren.u24		0.845	61
hdl		0.381		24	dfC3v2dA		0.851	62
trigl		0.382		25	EA2MQ		0.881	63
EA2D2Q		0.382		26	hbglyc		0.898	64
card.path	†	0.400		27	dfC3v4		0.898	65
resp.path	†	0.420		28	dfC1dTP		0.922	66
ECv4		0.422		29	dfC3n		0.922	67
EC3n		0.430		30	EC1n		0.926	68
dias.bp		0.459		31	EC3v2dA		0.938	69
cigarettes		0.481		32	dfC3v23		0.946	70
dfCv910		0.503		33	EC3v2		0.949	71
EC3v5		0.518		34	dfC1n		0.982	72
EC3v4		0.527		35	dfC3v12		0.982	73
EA2M		0.546		36	EC1dTP		0.983	74
dfA2D		0.552		37	EA2D		0.994	75
ECv10		0.565		38	morn.head	†	1.000	76

Table C.10: Estimated univariate potential of variables, based on the estimated entropy, to discriminate the classes of the binary outcome variable Y_2. †: variable of qualitative nature.

Variable		Threshold	Information gain	Rank
dfA2DQ		0.8	0.18	1
EA2DQ		7.1	0.11	2
dfA2D2Q		-1.0	0.10	3
dfC3v5		0.0	0.08	4
age		32.5	0.08	5
dfA2M		2.9	0.08	6
bmi		25.7	0.08	7
abdo.perim.cm		95.5	0.07	8
dfC3v8		0.0	0.07	9
adren.u		8.8	0.07	10
dfC3v2dA		-0.1	0.06	11
EC3v5		0.0	0.06	12
sis.bp		164.5	0.05	13
EC3v8		0.0	0.05	14
hdl		40.5	0.05	15
insul		20.4	0.05	16
dfC3v4		0.1	0.04	17
dfC2dTQQPQQ		-0.1	0.04	18
homa.ir		6.1	0.04	19
EC2dTQQPQQ		0.3	0.04	20
dfA2MQ		-0.4	0.04	21
dop.u24		309.1	0.04	22
dfCv34		-0.1	0.04	23
dfCv910		0.0	0.04	24
dfA2D		0.4	0.04	25
dfC3v6		0.0	0.04	26
EC3v4		0.0	0.04	27
dfC3v7		0.0	0.04	28
dfC1n		-1.3	0.04	29
trigl		86.0	0.04	30
ECv4		0.3	0.04	31
dfA2MTQ		-0.2	0.04	32
EC3v7		0.0	0.04	33
adren.u24		11.7	0.04	34
dfC3v3dA		0.0	0.03	35
EA2M		55.1	0.03	36
dop.u		127.7	0.03	37
glyc		113.0	0.03	38
dfC2n		-1.7	0.03	39
EA2D		7.6	0.03	40
dfC3n		0.1	0.03	41
EA2MQ		20.4	0.03	42
cholest		164.5	0.03	43
EA2D2Q		2.0	0.03	44
ECv10		0.0	0.03	45
cerv.perim.cm		41.8	0.02	46
dfC1dTP		-0.2	0.02	47
EC3v6		0.0	0.02	48
dfC3v12		-0.1	0.02	49
dfC3v23		1.4	0.02	50
EA2MTQ		0.3	0.02	51
smoking.habits	†	NA	0.02	52
EC3n		0.1	0.02	53
EC1dTP		1.0	0.02	54
ldl		145.5	0.02	55
homoc		20.4	0.02	56
EC3v3dA		0.2	0.02	57
dias.bp		97.0	0.02	58
hr		68.5	0.02	59
EC1n		1.0	0.02	60
metab.path	†	NA	0.01	61
hbglyc		6.2	0.01	62
awakenings	†	NA	0.01	63
noradren.u		24.6	0.01	64
EC3v2		1.1	0.01	65
card.path	†	NA	0.01	66
EC2n		2.1	0.01	67
EC3v3		1.4	0.01	68
EC3v2dA		0.0	0.01	69
noradren.u24		54.7	0.01	70
oxi.morn		1.0	0.01	71
endoc.path	†	NA	0.00	72
morn.head	†	NA	0.00	73
cigarettes		11.5	-0.02	74
tft	†	NA	NaN	75
resp.path	†	NA	NaN	76

Table C.11: Estimated univariate potential of variables, based on the empirical Area Under the Curve, to discriminate the classes of the binary outcome variable Y_2. †: variable of qualitative nature.

Variable	AUC	Rank	Variable	AUC	Rank
dfA2DQ	0.776	1	dop.u	0.541	35
abdo.perim.cm	0.695	2	homoc	0.538	36
bmi	0.679	3	EC3v6	0.538	37
cerv.perim.cm	0.673	4	dfCv34	0.535	38
dfA2D2Q	0.664	5	cholest	0.534	39
dfA2M	0.649	6	dfA2MTQ	0.528	40
EA2DQ	0.618	7	noradren.u	0.528	41
adren.u	0.616	8	ldl	0.523	42
EC3v7	0.612	9	dfA2MQ	0.522	43
age	0.611	10	hr	0.521	44
insul	0.607	11	EC2n	0.520	45
EC3v8	0.599	12	dfC2dTQQPQQ	0.519	46
dfC3v7	0.599	13	dfC2n	0.518	47
adren.u24	0.591	14	glyc	0.518	48
dfC3v5	0.585	15	EC3v3	0.518	49
homa.ir	0.581	16	noradren.u24	0.517	50
dfC3v8	0.579	17	dfC3v4	0.512	51
dop.u24	0.577	18	hbglyc	0.511	52
hdl	0.573	19	dfC1dTP	0.509	53
trigl	0.573	20	EC1n	0.508	54
EA2D2Q	0.573	21	dfC3v23	0.506	55
ECv4	0.567	22	EC3v2	0.506	56
EC3n	0.566	23	dfC1n	0.503	57
dias.bp	0.562	24	dfC3v12	0.503	58
cigarettes	0.559	25	EA2D	0.501	59
dfCv910	0.558	26	EC1dTP	0.498	60
EC3v5	0.554	27	EC3v2dA	0.493	61
EC3v4	0.553	28	dfC3n	0.491	62
dfA2D	0.551	29	EA2MQ	0.487	63
EA2M	0.551	30	dfC3v2dA	0.483	64
ECv10	0.548	31	dfC3v6	0.479	65
EC3v3dA	0.546	32	sis.bp	0.471	66
dfC3v3dA	0.545	33	EA2MTQ	0.469	67
EC2dTQQPQQ	0.542	34	oxi.morn	0.425	68

Table C.12: Estimated univariate potential of variables, based on the hypothesis tests p-value, to discriminate the classes of the multiclass outcome variable Y_4. †: variable of qualitative nature.

Variable		p-value		Rank	Variable		p-value	Rank
insul		< 0.001	***	1	resp.path	†	0.650	39
homa.ir		< 0.001	***	2	dop.u24		0.662	40
cerv.perim.cm		0.004	**	3	dfCv910		0.670	41
abdo.perim.cm		0.004	**	4	EC3v8		0.691	42
bmi		0.009	**	5	dfC3v3dA		0.693	43
dfA2DQ		0.009	**	6	dfA2MTQ		0.704	44
trigl		0.041	**	7	EC3v5		0.707	45
glyc		0.043	**	8	dfC3v2dA		0.716	46
dfA2M		0.053	*	9	card.path	†	0.720	47
dfA2D2Q		0.060	*	10	metab.path	†	0.740	48
hdl		0.090	*	11	awakenings	†	0.750	49
EA2D		0.163		12	EC3v6		0.771	50
sis.bp		0.207		13	dfA2D		0.778	51
EA2MTQ		0.219		14	oxi.morn		0.787	52
noradren.u24		0.225		15	dop.u		0.797	53
age		0.232		16	ECv10		0.799	54
adren.u		0.250		17	ldl		0.806	55
EC3v7		0.289		18	dfC3n		0.814	56
dias.bp		0.309		19	dfC3v8		0.816	57
tft	†	0.320		20	EC3v2dA		0.840	58
dfC3v7		0.331		21	EC3v2		0.850	59
EA2D2Q		0.341		22	ECv4		0.857	60
adren.u24		0.343		23	hr		0.882	61
EA2DQ		0.344		24	endoc.path	†	0.890	62
EA2MQ		0.375		25	EC2n		0.902	63
cigarettes		0.388		26	homoc		0.929	64
hbglyc		0.394		27	EC2dTQQPQQ		0.942	65
EC3n		0.442		28	cholest		0.965	66
dfA2MQ		0.481		29	dfC2n		0.968	67
noradren.u		0.509		30	dfCv34		0.971	68
smoking.habits	†	0.510		31	EC1dTP		0.974	69
dfC3v5		0.519		32	dfC2dTQQPQQ		0.976	70
EA2M		0.548		33	dfC1n		0.978	71
EC3v3dA		0.551		34	EC1n		0.982	72
EC3v4		0.554		35	dfC1dTP		0.984	73
dfC3v4		0.558		36	dfC3v23		0.991	74
dfC3v6		0.630		37	dfC3v12		0.993	75
EC3v3		0.641		38	morn.head	†	1.000	76

Table C.13: Estimated univariate potential of variables, based on the estimated entropy, to discriminate the classes of the multiclass outcome variable Y_4. †: variable of qualitative nature.

Variable	Threshold	Information gain	Rank	Variable	Threshold	Information gain	Rank
insul	20.4	0.35	1	hbglyc	6.2	0.06	39
homa.ir	5.1	0.32	2	EA2M	49.8	0.06	40
bmi	28.9	0.21	3	EC2dTQQPQQ	0.3	0.05	41
abdo.perim.cm	103.2	0.20	4	cholest	164.5	0.05	42
dfA2DQ	1.2	0.17	5	dfC3v6	0.0	0.05	43
cerv.perim.cm	42.8	0.16	6	oxi.morn	1.0	0.05	44
trigl	127.0	0.13	7	dfA2MTQ	-0.3	0.05	45
dfA2D2Q	0.1	0.12	8	dfA2MQ	0.1	0.05	46
adren.u	8.8	0.11	9	EC3v8	0.0	0.05	47
sis.bp	138.5	0.11	10	EC1dTP	0.6	0.05	48
adren.u24	11.7	0.11	11	dfC3v8	0.0	0.05	49
cigarettes	8.0	0.11	12	dfCv910	0.0	0.05	50
dfA2M	-1.1	0.11	13	dfC3v3dA	0.0	0.05	51
glyc	88.5	0.10	14	ECv4	0.3	0.04	52
EA2D	7.6	0.10	15	dfC1dTP	0.2	0.04	53
dfC3v5	0.0	0.10	16	EC3v6	0.0	0.04	54
noradren.u24	77.3	0.09	17	EC3v3dA	0.2	0.04	55
dop.u	127.7	0.09	18	EC2n	3.3	0.04	56
EC3v4	0.0	0.08	19	dfC2dTQQPQQ	-0.1	0.03	57
dfC3v7	0.0	0.08	20	dfC2n	1.8	0.03	58
EC3v5	0.0	0.08	21	ECv10	0.0	0.03	59
age	43.5	0.08	22	EC3v3	1.3	0.03	60
hdl	44.5	0.08	23	hr	68.5	0.03	61
EA2MTQ	0.3	0.08	24	EC3v2	1.4	0.03	62
EC3n	0.1	0.07	25	EC1n	1.0	0.03	63
dfC1n	1.1	0.07	26	dfC3n	0.1	0.03	64
EA2MQ	34.5	0.07	27	dfCv34	-0.1	0.03	65
dfC3v2dA	0.0	0.07	28	dfA2D	1.6	0.03	66
ldl	106.5	0.07	29	dfC3v23	1.0	0.02	67
EA2D2Q	1.0	0.07	30	EC3v2dA	0.1	0.02	68
noradren.u	44.0	0.07	31	dfC3v12	-0.8	0.02	69
EC3v7	0.0	0.06	32	awakenings †	NA	0.01	70
dop.u24	309.1	0.06	33	metab.path †	NA	0.01	71
smoking.habits †	NA	0.06	34	card.path †	NA	0.01	72
homoc	18.1	0.06	35	endoc.path †	NA	0.00	73
dias.bp	80.5	0.06	36	morn.head †	NA	0.00	74
EA2DQ	6.0	0.06	37	tft †	NA	NaN	75
dfC3v4	0.0	0.06	38	resp.path †	NA	NaN	76

Table C.14: Estimated univariate potential of variables, based on the empirical Area Under the Curve, to discriminate the classes of the multiclass outcome variable Y_4. †: variable of qualitative nature.

Variable	AUC	Rank	Variable	AUC	Rank
insul	0.723	1	EC3v3	0.571	35
homa.ir	0.699	2	dfC3v4	0.569	36
abdo.perim.cm	0.696	3	EC3v5	0.565	37
dfA2DQ	0.685	4	dfC3v3dA	0.563	38
cerv.perim.cm	0.684	5	dfCv910	0.561	39
bmi	0.681	6	dfA2D	0.561	40
glyc	0.664	7	EA2MQ	0.558	41
dfA2D2Q	0.641	8	dfA2MTQ	0.558	42
dfA2M	0.639	9	ECv10	0.555	43
trigl	0.637	10	EC3v8	0.555	44
EA2D	0.616	11	EC3v2dA	0.553	45
noradren.u24	0.615	12	ldl	0.551	46
sis.bp	0.614	13	EC3v2	0.548	47
age	0.612	14	dfC3v2dA	0.547	48
adren.u	0.611	15	dfC3v6	0.546	49
hdl	0.608	16	dfC3v8	0.545	50
EC3v7	0.603	17	ECv4	0.543	51
EA2DQ	0.601	18	EC3v6	0.540	52
EC3n	0.599	19	hr	0.537	53
cigarettes	0.598	20	dfC3n	0.536	54
EA2MTQ	0.597	21	EC2n	0.535	55
EA2D2Q	0.596	22	EC1dTP	0.527	56
dfC3v7	0.594	23	cholest	0.525	57
dias.bp	0.593	24	dfC2dTQQPQQ	0.520	58
adren.u24	0.592	25	dop.u	0.520	59
hbglyc	0.588	26	EC1n	0.517	60
noradren.u	0.584	27	dfCv34	0.506	61
EC3v3dA	0.580	28	EC2dTQQPQQ	0.504	62
oxi.morn	0.576	29	homoc	0.498	63
EC3v4	0.574	30	dfC1dTP	0.497	64
dfA2MQ	0.574	31	dfC2n	0.497	65
dfC3v5	0.574	32	dfC3v12	0.489	66
EA2M	0.572	33	dfC1n	0.488	67
dop.u24	0.571	34	dfC3v23	0.485	68

Appendix D: Packages from R

Table D.1: Packages and functions from R exemplified for the application of the methods presented. Most packages were applied to the case study.

Technique set	Description	Package	Function	Details
Data preparing	Imputation with k nearest neighbors.	`VIM`	`kNN`	
Data familiarizing	Empirical AUC (outcome classes: 4)	`pROC`	`multiclass.roc`	Performs multiclass empirical AUC
	Empirical AUC (outcome classes: 2)	`pROC`	`roc`	
Variable selection	Variable selection for clustering and classififcation	`vscc`	`vscc`	
Unsupervised learning	Partitioning around medoids (PAM)	`cluster`	`pam`	
	Agglomerative hierarquical clustering	`stats`	`hclust`	
	K-medoids	`cluster`	`pam`	
	Silhouette analysis - K-medoids	`cluster`	`silhouette`	
	Hierarchical agglomerative clustering	`factoextra`	`hcut`	
	Silhouette analysis - hierarchical clustering	`factoextra`	`fviz_silhouette`	
Supervised	Decision tree	`rpart`	`rpart`	
	Adaboost	`adabag`	`boosting`	Performs multiclass Adaboost.M1
	Adaboost	`fastAdaboost`	`adaboost`	Performs a binary classification task
	Naive bayes	`naivebayes`	`naive_bayes`	
	Binomial logistic regression	`stats`	`glm`	
	Ordinal logistic regression	`MASS`	`polr`	
	Test for Proportional Odds assumption (ordinal logistic regression)	`car`	`poTest`	Applied to objects returned by `polr` function
	Multinomial logistic regression	`nnet`	`multinom`	
	VIF (variance inflation factor)	`car`	`vif`	Applied for evaluating variable multicollinearity (binomial logistic regression)