Predicting the Conversion from Mild Cognitive Impairment to Alzheimer´s Disease using Evolution Patterns

Andreia Liliana Duarte Fernandes Ferreira

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Professor Sara Alexandra Cordeiro Madeira

Doutor Alexandre Valério de Mendonça

Examination Committee

Chairperson: Professor Raúl Daniel Lavado Carneiro Martins

Supervisor: Doutor Alexandre Valério de Mendonça

Member of the Committee: Professor Alexandra Sofia Martins de Carvalho

December 2014

Acknowledgments

My first acknowledgement goes to my supervisor, Sara Madeira, for guiding me throughout this work and for the opportunity to be part of this project. I also would like to thank to Alexandre Mendonça for the clinical feedback given and for the commitment in this project.

I express my gratitude to Luís Lemos for the constant availability to clarify my doubts and for the constructive critics and positive words. I would like to thank to Ricardo for the help and the constant encouraging words. I would also like to thank to the NEUROCLINOMICS group for being always promptly available to any question, especially to Telma for the unconditional support, for never let me doubt of myself, for the caring friendship and for the sharing moments along the last months.

I would like to acknowledgement my family, especially my mother, the strongest person I ever known, for the patience, for the guidance in all phases of my life and for teaching me to never give up.

I would like to thank my friends, especially to Simone, Serineu and Carolina. To Simone, I am grateful for our friendship, for the support in all decisions I made, for believing in me and for the countless hours we share learning from each other and growing up together. To Serineu and Carolina, thank you for the beautiful companionship, for being right next to me to celebrate every accomplishment and for the caress and the strength in the worst moments.

Finally, I would like to thank to my biomedical colleagues, especially to Teresa, my dear friend and companion along these five years, for the affection and the comfort words, for the unforgettable moments and the loudly laughers we share.

This work was supported by FCT through the NEUROCLINOMICS (PTDC/EIA-EIA/111239/2009) project.

Abstract

Declines in cognitive functions, together with other evidences of neurological degeneration, become increasingly likely as healthy people age [1,3]. Alzheimer´s Disease (AD) is a neurodegenerative disease characterized by progressive deterioration of cognitive function and is the most common cause of dementia in elderly people. Mild Cognitive Impairment is considered a prodromal state that represents the transitional period between normal ageing and dementia. As such is regarded with special attention since it represents higher risk to evolve to dementia. Thus, the definition of this clinical entity is fundamental to the timely administration of pharmaceutics and therapeutic interventions, improving patient's quality of life.

This thesis intends to predict the evolution of MCI patients to AD considering two approaches: the first where all patients are assumed to evolve similarly and the second where patient profiles are considered. Time windows for two to five years are used for prediction. Initially, we used supervised learning methods, using feature selection to effectively decrease the dimensionality of the problem. Then, standard clustering algorithms were applied with the purpose of studying the potential existence of MCI subtypes. The patients were also divided according to their state of depression, based on clinical information.

The results demonstrated the importance of considering longer time interval to predict conversion of MCI patients to AD and that the grouping of patients according to their depressive symptoms influences positively the prognosis results. The clustering analyses validated the importance to study MCI subgroups considering the different characteristics of this clinical entity in the prediction models.

Keywords: Alzheimer´s Disease, MCI, Temporal Window, Prognosis, Classification, Clustering.

Resumo

As evidências de degeneração neurológica tornam-se cada vez mais prováveis de acontecer com o envelhecimento [1,3]. Doença de Alzheimer (DA) é uma doença neuro-degenerativa caracterizada pela deterioração cognitiva progressiva e é a causa mais comum de demência nos idosos. A deficiência cognitiva ligeira (DCL) é um estado que representa o período de transição entre o envelhecimento normal e demência. Assim, é preciso conceder-lhe uma atenção especial por constituir maior risco de evoluir para demência. Assim, a definição desta entidade clínica é fundamental para a administração oportuna de produtos farmacêuticos e intervenções terapêuticas, melhorando a qualidade de vida do paciente.

Esta tese pretende prever a evolução de pacientes com DCL para DA considerando duas abordagens: a primeira que assume a evolução semelhante dos pacientes e a segunda que considera os seus perfis. Janelas temporais de dois a cinco anos são usadas para a previsão. Inicialmente foram utilizados métodos de aprendizagem supervisionada e técnicas de selecção de atributos para eficazmente diminuir a dimensionalidade do problema. Posteriormente, algoritmos *standard* de *clustering* foram aplicados com o objetivo de estudar potenciais subgrupos de DCL. Os pacientes também foram divididos de acordo com o seu estado de depressão, através de informações clínicas.

Os resultados demonstraram a importância de aumentar as janelas temporais na previsão da conversão de pacientes com DCL para DA e que a separação de pacientes de acordo com estados depressivos influencia positivamente os resultados do prognóstico. As análises de *clustering* validaram a importância do estudo dos subgrupos de DCL em modelos preditivos.

Palavras-Chave: Doença de Alzheimer, DCL, Janelas Temporais, Prognóstico, Classificação, *Clustering*

Contents

List of figures

Appendix B

List of Tables

Appendix A

Appendix B

Appendix C

List of Acronyms

AD	Alzheimer's Disease
MCI	Mild Cognitive Impairment
MRI	Magnetic Resonance Imaging
CV	Cross-Validation
EM	Expectation-Maximization algorithm
SVMs	Support Vector Machines
DT	Decision Tree
***k*NN**	*k*-Nearest Neighbour
GDS	Geriatric Depressive Scale
CDR	Clinical Dementia Rating scale

1. Introduction

Alzheimer´s Disease (AD) is the most common cause of dementia in elderly people, accounting for 60% to 80% of dementia cases, and its specific cause is still unknown [31]. This disease progresses gradually, beginning with early signs and subtle behavioural changes, followed by memory loss, impaired judgment and lower ability to participate in daily activities. In a last phase, the disease evolves to the incapacity of the patient to understand language or even to speak, besides the disability to control his body [24]. Even though the research on genetic, imaging and biomarkers correlated to AD has amplified in the past decades, the accurate assessment of AD is still a challenge, due to the several symptoms shared with other diseases, related with cognitive decline, and the little existence of available longitudinal studies to draw conclusions about [27, 30].

The number of advances to understand the biochemical nature of AD increased in the past ten years [5, 6, 8, 9, 10, 11, 12, 24]. The strongest epidemiological risk factors for AD are age and family history and other putative risk factors are head trauma, education level, hypothyroidism, apolipoprotein ε4 genotype, abnormal processing and accumulation of β-amyloid protein and tau protein.

Several reviews tried to identify the prominent neurodegenerative patterns during the prodromal stages of the disease, in which mild symptoms are evident, in order to prevent or postpone it by intervening early [3, 1, 6, 8, 55]. The risk of dementia in individuals with mild cognitive impairment (MCI) is higher when compared with cognitively normal (CN) subjects. The annual incidence rate of healthy subjects to develop AD is 1% to 2%, while the conversion rate from MCI to AD is reported to be approximately 10% to 15% per year [15]. As such, defining the MCI as a clinical stage will allow meaningful interventions to be taken and reliable predictions of progressing to AD to be made [21, 22]. Although questions have risen about MCI classification, some studies have applied MCI criteria with hindsight to acquired different datasets and have already provided important discernments to the clinical characterization [2, 16, 21, 22].

Many existing studies suggest that when functional impairment emerges and the patient is diagnosed, significant neurodegeneration has already happened many years before that [3, 20]. Cognitive functions are commonly measured through neuroimaging techniques and biomarkers, but neuropsychological tests have been acquiring value [7, 16, 29]. Moreover, these tests are not intrusive and less expensive. A set of recent papers showed that these tests are good to predict MCI progression to AD and that by exploiting state-of-the-art data mining methods is possible to go further in this direction [3, 8, 9, 15].

1.1 Problem Formulation

Previous studies intended to predict which individuals with MCI progress to AD [3, 49, 51, 55]. The disease has been studied through neuroimaging biomarkers [4, 29, 54], biomedical biomarkers [1, 29, 54] and neuropsychological assessments [16, 29]. These techniques provided insights into the disease biology, as well as staging and prognosis.

The aim of this work is to predict the evolution of MCI patients to AD considering two approaches: first where all patients are assumed to evolve similarly and, second where patient profiles are considered. Time

windows from two to five years are used for prediction. Demographic data and longitudinal data provided by the Cognitive Complaints Cohort (CCC) study in the Dementia Group at Instituto de Medicina Molecular (IMM) were used. The data correspond to a large number of neuropsychological assessments that each MCI patient was subjected during one or more evaluations. We first computed a baseline model to predict conversion of MCI individuals to AD through supervised learning techniques without using explicitly different MCI groups. Then, MCI subgroups were obtained, first by considering important characteristics of the patients and then through unsupervised learning methods. These groups were used to train models to predict the conversion to AD. The proposed models using patient profiles progression patterns should outperform the baseline model.

1.2 Goals

The goals of this thesis are:

1) Predict conversion from MCI to AD through supervised models, considering a time-window approach between 2 to 5 years (Baseline):

1.1) Train the prognosis approaches proposed by Lemos et al. [48] in a recent version of the Cognitive Complaints Cohort (CCC) study.

1.2) Train the prognosis model for the five years window, not done so far due to the lack of adequate data.

2) Predict conversion from MCI to AD through supervised models trained with different MCI groups, obtained using clinical knowledge and unsupervised methods, outperforming the baseline model in a ideal approach.

1.3 Dissertation Outline

This thesis is organized as follows. Chapter 2 presents the information relatively to Alzheimer´s Disease and MCI and reports previous studies to improve the prediction of conversion to the disease. Chapter 3 explains the experimental methodology to clean the data, the basic concepts on data mining techniques used and the model parameters used. Chapter 4 shows the application of prognostic prediction to the MCI patients, predicting if the patient will ever convert to AD and then if the patient will convert in a fixed period of time (time window). In chapter 5, a similar analysis is performed, considering putative differences in MCI subgroups. Chapter 6 is reserved for conclusions of the work and proposals for future work.

2. Background

In this chapter, it will be described the Alzheimer's disease and the neuropsychological tests used to characterize the different patients in the the used dataset.

The way that a great number of missing values and imbalance classes affect the results and the techniques to deal with these problems will be also described in this chapter. The different algorithms used to train the models will be succinctly explained as well. In order to understand what the previous works already achieved and improve in the same direction, few studies will be described emphasizing the evaluations similar to the ones done in the present thesis.

2.1 Alzheimer´s Disease and Mild Cognitive Impairment

Global population aging has been one of the most concerns of the twentieth century with massive economic, political and social consequences [31]. Dementia is prevalent among elderly and the complaints of a clinical phenotype, which includes episodic memory and other cognitive domain impairment, as language, visuospatial or visuoperceptual abilities, are common [30, 31].

Alzheimer´s Disease (AD) is a neurodegenerative disease clinically characterized by a progressive dementia. Although AD is still an irreversible disease, nowadays it is possible to identify traces or biomarkers that increase the likelihood that an AD pathological process is present, through Magnetic Resonance Imaging, Positron Emission Tomography scans, neuropsychological measures and others biomarkers [1, 3, 6, 7]. Numerous efforts have been made in the past decades in order to capture a full spectrum of the disease and apply it to research protocols and clinical trials, directed at early stages of the disease [6, 7, 15, 18].

Besides the fact that AD causes severe financial effects in society, its impact on social function and activities of daily living causes psychological and emotional changes in the patient, with the development of depression and other secondary features. Moreover, the appearance of cognitive, challenging behavioural and neuropsychiatric disturbances in the patient will also affect his family causing a psychological and a financial burden [24].

The term Mild Cognitive Impairment (MCI) was introduced by Reisberg et al [30], in the late 1980s, with the purpose of characterizing individuals in the intermediate stage, based on the Global Deterioration Scale [30]. In 1999, Petersen et al. [19] improved the concept, having in consideration an observational study of ageing. In the following years, the rising awareness of this grey zone of cognitive impairment had an important impact on public health, emerging clinical need to define it. The conception of this stage had been extensively used worldwide in the past ten years [19, 20, 30]. The MCI stage has been described also as dementia prodrome, incipient dementia or isolated memory impairment and its construct had the purpose of identifying individuals at an early stage in the cognitive decline, such that clinicians could intervene at this point [20].

Since MCI is a heterogeneous syndrome, where dementia symptoms are not fully expressed, the definition of MCI criteria represents a demanding task. MCI interposes between cognitive changes of healthy aging people, which interfere with the distinction from MCI individuals to those encountered by normal individuals as they

age, and very early dementia [19]. MCI is as well a preliminary stage of progressive cerebrovascular disease [28]. Moreover, in many studies, the MCI patients reverted to normal which lead us to believe in the wrong classification of such patients and validating the importance of defining this complex group [4, 21, 22, 19].

Although there is still no consensus concerning tests and precise definition of this concept, the factors related with the controversial results are becoming recognized. It is important to emphasize among them, the data mining algorithms versus clinical application of the criteria, the reliability of clinical judgment, the temporal changes in cognitive performances and the length of the follow-up of the subjects under study [28, 30]. Most of the studies performed have a maximal follow-up of the patients for three years, presenting unsatisfactory results [22, 19, 30, 1, 47]. Thus it would be of extreme importance to increase the temporal windows analysed.

Neuropsychiatric symptoms are common in subjects with MCI and the effects of these symptoms along with the separation of patients according to their state of depression is starting to assume a relevant role in the prognosis prediction of MCI patients, revealing important differences considering the neuropsychological state of the patient [16, 28, 19]. Generally, the test used to study this condition is the Geriatric Depression Scale (GDS) that is a self-report assessment used specially to identify depression [16].

In order to increase the accuracy of the MCI classification, many studies had in consideration alternate clinical subtypes of MCI. Therefore, with the combination of clinical subtypes and putative differences may become possible to reduce heterogeneity in the study groups and to match the timely administration of pharmacologic, tailored for specific targets and populations [19, 29, 16, 21, 22, 19, 28, 30]. Grouping into clusters proved to be useful among a massive amount of data resulted from neuropsychological tests to discover groups of similarity and to study interactions and relationships among them [28, 30, 40].

Therefore, MCI evolution has generated a lot of research from clinical, imaging, genetic, pathological and epidemiological perspectives, trying to enhance the uniformity of the outcomes, focusing the investigation on searching factors which make MCI patients more vulnerable to evolve to dementia [1, 2, 6, 8, 9, 30].

The assessment of tracers and biomarkers has been proposed to enhance the diagnosis of AD and other neurodegenerative diseases, at the dementia stage, the distinction in MCI, at a MCI stage, and the identification of healthy persons at risk of developing AD, at an early stage [6]. The most used techniques to identify the markers are Magnetic Resonance Imaging (MRI) volumetric studies [4, 6, 29, 51, 54], neurochemical analyses of the cerebrospinal fluid [1, 29, 54] and Positron Emission Tomography (PET) scans [7, 51]. In opposition to these methods, neuropsychological tests are relatively inexpensive and noninvasive to the patient, being widely used in the clinical assessment of AD [7, 16, 29]. Due to this, the upgrading of the value of neuropsychological tests to predict the progression of MCI patients to AD has great relevance. These tests were established by medical doctors and typically include orientation, new-learning/memory, intelligence, language, visuoperception, executive function, person's psychological, personal, interpersonal and wider contextual circumstances [25]. High values of accuracy in the prediction of dementia using neuropsychological tests were reported [1, 6, 9, 15].

A combination of features of different biomarkers proved to identify more accurately MCI patients at risk, outperforming sometimes a single modality of features, optimizing cost-benefit ratios in clinical trials [1, 3, 6, 7, 10]. This multivariate predictor will not be studied in this work.

2.2 Data Mining techniques

With the continuous expansion of data availability in large-scale, complex and networked systems, as surveillance, security, internet and finance, it becomes critical to extract useful knowledge of large datasets to support decision making. Currently, as stated by Singhal et al. [34] "we are drowning in data but starving for knowledge". With the purpose of extracting knowledge of these datasets, diverse activities have been referred such as data mining, knowledge discovery, pattern recognition and machine learning. In particular, Data Mining constitutes an iterative process of data gathering and analysis of large databases, with the purpose of finding regular patterns and extracting meaningful information to develop inductive learning models and adopt practical decisions based on the knowledge acquired [34, 37]. The methodologies of data mining are present in multiple fields of application, from marketing and manufacturing process control to medical diagnosis. In medical diagnosis, learning models are an irreplaceable tool within the medical field for the early detection of diseases using clinical test results [34, 37].

In order to extract useful information different steps have to be analysed carefully as data preprocessing; data mining, results evaluation and knowledge presentation [37].

2.2.1 Data Preprocessing

Real world data is imperfect and heterogeneous and the power to extract useful information and meaning results from it highly depends on the data quality. A raw data set is noisy, holding errors or outlier values that deviate from the expected, incomplete, and inconsistent, enclosing discrepancies in names and codes [34, 36].

In order to enhance the quality of the data, several preprocessing steps are taken [34, 36]:

(i) data cleaning involve the removal of incorrect values and the handling with outliers, missing values and inconsistencies;

(ii) data integration techniques combine the data from different databases and files (this step was not discussed in this thesis);

(iii) data reduction methods reduce the size of the data, preserving its important information, such as discretization and attribute selection and

(iv) data transformation converts a set of data values from the data format of a source data system into the data format of a destination data system.

2.2.2 Feature Selection

The success and accuracy of learning schemes, in their attempts to construct models from lots of data, is the reliable identification of a set of highly predictive attributes [32, 36]. Feature subset selectors are algorithms that remove the unhelpful and redundant information as possible, prior to knowledge discovery. The feature selection improves the classification accuracy substantially or equivalently, increases the speed and the accuracy of the learning process, and decreases the amount of training data and the storage requirements needed to obtain the desired level of performance [32, 37].

The feature selection algorithms are generally divided in the type of subset evaluator and the search methods and can be classified into three main categories: filter methods, wrapper methods and embedded methods [32, 37].

2.2.3 Unsupervised Learning Methods

In unsupervised learning algorithms, the class label attribute and the number of classes to learn in advance are unknown. Clustering methods represent a class of models of unsupervised learning, whose principle is to maximize the similarity of objects within one group, through appropriate metrics and notions of distance. The homogenous groups identified are called clusters, sharing a stronger resemblance between the instances that characterized them [40].

Clustering models may provide meaningful interpretation of regular patterns in the data, easily understood by experts in the application domain. For this reason, clustering has long been used in various disciplines, such as sciences, biology, astronomy, statistics, image recognition, processing of digital information, marketing and data mining. These techniques have proven to be effective when dealing with biomedical problems [40].

The determination a priori of the number of clusters is a difficult task, since it could be correct to separate one or more clusters by increasing the total number *k* of clusters identified or it may be appropriate to merge two or more clusters into a single one, reducing the number of clusters [40].

In what follows, the two clustering techniques used in this thesis are briefly described.

K-Means

K-means [40] is the simplest and most common used algorithm in unsupervised learning techniques. This algorithm partitions the data into k non-overlapping clusters $(C_1, C_2, \dots, C_k)$ represented by their centers or centroids $\mu_k = \frac{1}{N_k}\sum_{q=1}^{N_k} x_k$, where the x represent the observations and k the number of observations. The center of each cluster is calculated as the mean of all the instances belonging to that cluster.

The algorithm starts by arbitrarily selecting *k* centroids, one for each cluster. For every subsequent iteration, each instance is assigned to the cluster whose centroid is the closest, minimizing the total squared distance from the observation among all centroids. If no observation is reallocated to a cluster different from the one to which it belongs, determined during the previous iteration, the procedure stops. Otherwise the new centroid for each cluster is computed and a new assignment is made.

This method is computationally easy, fast and memory-efficient. Nonetheless, there are some variants that need to be taken in consideration. As the final clusters are quite sensitive to the initial cluster and different arrangements can arise from small changes in the initial random choice, it would be useful to generate different initial random assignments and then derive for each of them a different clustering, finally choosing the best one. Moreover, outliers can also affect the final result since they take large numerical values in the quadratic objective function. Hence, this method should be applied preferably only after outliers have been identified and removed.

Expectation-Maximization

Density-based methods [40] assume that the points that belong to each cluster can be represented by a probability distribution, where the entire data is assumed to be a mixture of several distributions. The aim of these methods is to identify the clusters and their distribution parameters. The problem of estimating the parameters of the probability distributions, in order to best fit the data, can be solved by the Expectation-Maximization (EM) algorithm [40].

This algorithm begins with an initial estimate of the parameter vector of the mixture model by selecting randomly *k* objects to better represent the cluster means. After that, the algorithm consists of two steps, an expectation step followed by a maximization step. In the expectation step it is computed the posterior probabilities for observations that belong to individual clusters, given the mixture and individual density functions. In maximization step the parameters that maximize the expected likelihood function are determined. These two steps are iterated until convergence, obtaining a locally optimal solution or a point of the underlying clustering criterion.

2.2.4 Supervised Learning Methods

Classification can be divided into two phases, the learning phase and the validation phase. In the learning phase, it is given a dataset containing predictive attributes and a categorical target attribute, called class or label. In our case, the instances are the several observations of each patient, the set of predictive attributes is extracted from the neuropsychological data and the class label is the patient state (evolution or not evolution to AD in a given temporal window). An appropriate algorithm is going to learn a model, identifying the optimally description of the relationship between the predictive values and the target class. In the validation phase, the learned classifier is applied to an independent set of instances, known as the validation set, to estimate its predictive performance [37, 38].

The classification problems can be categorized as binary or multicategory classification problems [37]. In the present thesis, we have a binary problem since only two classes are considered.

Currently, it is possible to distinguish four main categories of classification models with respect to the logic used for deriving the classifier [37]:

(i) Heuristics models: Use classification methods based on simple or intuitive algorithms, where Nearest Neighbour methods and Classification Trees are examples.

(ii) Separation models: Divide the attribute space in disjoint regions, separating the observations based on the target class. The most popular separation techniques include Perceptron methods, Neural Networks and Support Vector Machines.

(iii) Regression models: Identify a functional relationship, to predict the future value of the target attribute, based upon the functional relationship between the target variable and a subset of the remaining attributes contained in the dataset.

(iv) Probabilistic models: A hypothesis is formulated regarding the functional form of the conditional probabilities of the observations given the target class, where Naïve Bayes classifiers and Bayesian networks are examples.

It is also important to clarify the difference between eager learners and lazy learners. Eager learners build a generalization model during the training phase using a training set, independently of the validation set, whereas the lazy learners store the training instances and perform the generalization step when given the instance to classify, based on its similarity to the memorized training instances. Examples of lazy learners are the *k*-Nearest Neighbour classifier and the Lazy Bayesian Rules classifier. Examples of eager learners are the Decision Tree, Support Vector Machine, Multilayer Perceptron and Naïve Bayes [37].

The development of algorithms capable of learning from past experience has grown interest in neuroimaging and clinical based-diagnosis fields, since it represents a fundamental step in emulating the inductive capabilities of the human brain [3, 37]. A promising direction of enduring research is focused on increasing the accuracy and efficiency of machine learning techniques, to characterize prominent neurodegenerative patterns of diseases and the set of neuropsychological tests needed, in order to accelerate the diagnosis and prognosis process [7, 17].

Follows a short description of the supervised learning methods used in this thesis.

Support Vector Machines

Support Vector Machines (SVM) [15, 37, 38] are a family of separation methods for classification that represent the data points in a space mapped such that instances of each class are separated from the other.

a) Linearly separable case

Assume the training set $D = \{(X_1, y_1), .., (X_{D,} y_D)\}$ with instances $X_1, \dots, X_D$ and class labels $y_1, \dots, y_D \in \{+1,-1\}$. The data is linearly separable if it is possible to separate instances of both classes, in a 2-D graph, through a straight line. In this dimension, two hyperplanes can be chosen if they separate the data, with maximal distance between them, called maximal margin hyperplane satisfying $\boldsymbol{w}.\boldsymbol{X} + b = 0$.

The shortest distance from the hyperplane to one margin is equal to the shortest distance from the hyperplane to the other margin and the space between those margins is $\frac{2}{||w||}$. For that reason, $||w||$ must be minimized in order to maximize the separation between margins. To find the optimal hyperplane, a linear combination of training instances can be used, $W = \Sigma\, \alpha_i y_i \boldsymbol{X}_i$, where α_i are the Lagrangian multipliers coefficient indicating how difficult is to classify instances. The corresponding $\boldsymbol{X}_i$ are the support vectors, which lie on the margin satisfying $\boldsymbol{w}.\boldsymbol{X} + b = \pm 1$, which appear to be the most representative observations for each target class. Once the optimal hyperpace is determined, the classification between the two classes can be achieved by computing the decision boundary $d(\boldsymbol{X^T}) = sign(\sum_{i=1}^{l} \alpha_i y_i \boldsymbol{X}_i . \boldsymbol{X^T} + b)$, where $\boldsymbol{X}_i$ are the support vectors, $\boldsymbol{X^T}$ is the instance for which the class is unknown and *l* is the number of support vectors.

b) Linearly inseparable case

Simple linear classifiers are unable to correctly classify patterns which classes are not linearly separable. Support vector machines address this problem using linearly separable classes, not in the input space, but in the so-called feature space. Such modification is achieved by applying a non-linear kernel to the data, without resorting to mappings, carried out by means of kernel functions. The search for the hyperplane can be done using the transformed instances. The typically kernels used are:

(i) linear $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = X_i^T \boldsymbol{X}_j$;

(ii) polynomial $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = (X_i^T . \boldsymbol{X}_j + 1)^d$;

(iii) radial basic function (RBF) $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = e^{\frac{||X_i - X_j||^2}{2\sigma^2}}$; and

(iv) sigmoid $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \tanh(k\, X_i^T . \boldsymbol{X}_j - \delta)$.

Naïve Bayes

Bayesian methods belong to the family of probabilistic classification models and the Naïve Bayes (NB) is one of the simplest Bayesian network classifiers [37, 38]. What is being estimated is the conditional probability distribution of the values of the class attribute given the values of the other attributes. By means of Bayes theorem, the method calculates the posterior probability $P(C/X)$ expressing that a given observation belongs to a specific target class, once the prior probability $P(X)$ and the class conditional $P(X/C) = \frac{P(X/C).P(C)}{P(X)}$ known.

Let us consider a generic training set $X = \{x_1, \ldots x_n\}$, composed by n-dimensional attribute vectors and the respective class label, which can take *m* distinct values denoted by $C_1, \ldots, C_m$. Given an instance X, the classifier will predict to which class X belongs to, by choosing the class having the highest posterior probability conditioned, $P(C_i/X)$, called the *maximum posteori hyphothesis*, $P(C_i/X) > P(C_j/X)$ for $i \leq j \leq m; j \neq i$. Since $P(X)$ is constant for all classes, it is only necessary to maximize $P(X/C_i)P(C_i)$. If the class prior probabilities are unknown, then it is assumed that all classes are equally likely and only $P(X/C_i)$ must be maximized. Subsequently, as Naïve Bayes assume that attributes are conditionally independent, $P(X/C_i)$ is calculated by $P(X/C_i) = \prod_{k=1}^{n} P(x_k/C_i)$.

The estimation of $P(X/C_i)$ is performed differently having in account if the attributes are categorical or continuous. In the case of categorical attributes, $P(x_k/C_i)$ is evaluated through the ratio between the number of instances of class C_i in the training set, that have the value x_k for a certain attribute, and the total of instances of the class in the training set. In the case of continuous-valued attributes, the probability density function must be estimated and, to simplify, $P(x_k/C_i)$ is distributed with mean μ and standard deviation σ as $P(x_k/C_i) = \frac{1}{\sqrt{2\pi\sigma C_i}}$.

One advantage of NB classifier is that requires a small amount of training data to estimate these parameters. To predict the class label of the test instance X, $P(X/C_i)P(C_i)$ is evaluated for each class and then the most probable class C_i is obtained.

In spite of the simplistic assumption that attributes are independent and the easiest computational of conditional probabilities, the empirical evidences show that Bayesian network classifiers are a practical and efficient classifier, robust to noise and irrelevant attributes, whose performance is comparable to more complex approaches [15].

k-Nearest Neighbour Classifiers

The *k*-Nearest Neighbour (*k*NN) algorithm [37, 39] is a very simple and straightforward classifier whose classification is achieved by identifying the nearest neighbours to a query example and using those neighbours to determine the class of the query.

For each instance X_i in the validation set, the k parameter is defined empirically in the training set $(X_{i1}, X_{i2}, \ldots, X_{ik})$ stored previously. This algorithm stores all available cases and classifies new cases based on a

similarity measure. Each train instance has a vote for its class and the predicted class is the one with majority vote of its neighbours, measure by a distance function. In the present thesis, it is used the Euclidean distance $dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}$.

The simplicity and easy interpretability are the main advantages of *k*-NN algorithm that allow this classifier to usually outperformed algorithms of higher complexity. On the other hand, some significant disadvantages need to be referred. This algorithm is defined by poor run-time performance if the training set is too large, compromising the sensibility to irrelevant or redundant features. On very difficult classification tasks, *k*-NN algorithm may be outperformed by techniques as Support Vector Machines or Neural Networks. To overcome these limitations, the accuracy of this algorithm can be improved with a careful feature selection or feature weighting and the under-weighting of attributes with smaller range to the ones with larger range can be prevented with a normalization of the attributes using Euclidean distance [37, 39].

Decision Trees

Decision Trees (DT) are perhaps the most widely used learning methods in data mining applications [37]. This method has the advantage of providing clear explanation, understandable even by non-specialists, of which output values is produced for each given set of inputs, over other supervised systems such as Multilayer Perceptron or Support Vector Machine. In addition, DT are popular for their conceptual simplicity, computational speed and robustness to missing data and outliers [37, 41].

In a brief way, decision trees are a model structure that aim to obtain simple and explanatory rules for the relationship existing between the target variable and predictive attributes [38]. This model is composed by non-leafs containing a test on an attribute, branches representing outcomes of the tests and leafs containing class labels.

A basic algorithm for the construction of a decision tree receives as input a set of instances, an attribute list and an attribute selection method. The development of a classification tree corresponds to the training phase of the model and is regulated by the attribute selection method and a recursive procedure of heuristic nature based on a divide-and-conquer partitioning scheme, to select the attribute that better discriminates the instances according to the class. Since the algorithm is recursive, the set of instances at the beginning will change along the procedure. At each of the nodes determined, it is investigated if the attribute chosen presents no ambiguity, meaning that all of the instances that fall within the value of the chosen attribute have the same class, and then the value of the class, to which the majority of the observations in the node belong, is assigned to the attribute value. For the remaining values will be found the best way of partition the instances into individual classes, by means of the splitting rule determined by the attribute selection method. The process is repeated until no ambiguities still exist, or the algorithm runs out of attributes before attaining a perfect discrimination.

By varying the metrics used to identify the splitting rule, different versions of classification trees can be obtained. The most commonly are information gain, gini index and information gain ratio. At the end of the procedure, the set of splitting rules that can be found along the path, connecting the tree root to a leaf node, constitutes a classification rule, which is obeyed during the prediction phase, in order to assign the target class to a new observation [37, 38, 41].

The decision tree DT4.5 applies a measure called information gain ratio [42] as attribute selection measure, aiming to overcome one drawback of the information gain.

Firstly it is important to define these terms. The attribute selected to split is the one that provides the highest information gain. Consequently, this attribute maximizes the homogeneity of the label class and minimizes the expected information required to classify the instances of the resulting partitions.

The expected information needed to classify the instances in partition is defined as $Info(D) = -\sum_{i=1}^{k} p_i \, log_2 \, (p_i)$, where k is the number of classes and p_i is the probability that an instance belongs to class C_k, with 1 ≤ i ≤ k and it is estimated by $p_i = \left|C_{i,D}\right|/|D|$, where $\left|C_{i,D}\right|$ is the set of instances of class i in the partition D. The base 2 algorithm is used since the information is encoded in bits. To classify the expected information present in an attribute A, with possible outcomes n, the resulting partitions the definition is $Info_A(D) = -\sum_{r=1}^{n} \frac{|D_r|}{|D|} \times Info(D)$. The information gain is finally calculated through $Info(A) = Info(D) - Info_A(D)$. The information gain ratio attempts to normalize the information gain $SplitInfo_A(D) = -\sum_{r=1}^{n} \frac{|D_r|}{|D|} \times log_2 \left(\frac{|D_r|}{|D|}\right)$, which represents the potential information generated by splitting D into n partitions D_r (1 ≤ r ≤ n). The attribute with the highest information gain ratio $GainRatio(A) = \frac{Info(D) - Info_A(D)}{SplitInfo_A(D)}$ is chosen as a split criterion for the node.

2.2.5 Class Imbalance

The common understanding in the community is that imbalanced data corresponds to data where the number of instances of one class is significantly higher than the number of instances of the other classes. Is not that uncommon to testify imbalances in the order of 100:1, 1000:1, 10000:1, where in each case, one class severely represents another. This form is referred as the between-class imbalanced. Through singular assessment criteria the evaluation would not be accurate, since the error of classification would be 0.1%, in the second case referred above.

Although existing data engineering techniques have shown great success in the fundamental understanding of knowledge discovery in many real applications areas, the problem of learning from imbalanced data represents a recurring problem of high importance with wide-ranging implications, in both academia and industry [33]. In general, most classical machine learning algorithms perform poorly on imbalanced data, since they are biased toward the majority class, due to the assumption of balanced class distributions or equal misclassifications errors. As the minority class is underrepresented and outnumbered, the induction rules that describe these minority concepts are fewer and weaker comparatively to those of majority concepts [33, 36]. Owing to that, most informative assessment metrics should be used, as ROC Curves and precision-recall curves. This issue will be discussed latter.

Medical applications are a common source of imbalanced datasets, where the disease cases are much rarer than the healthy cases. It is important to understand that imbalanced datasets may be result from the minority class be rare in its own right, being composed by concepts with limited instances, which can hinder the classification process. This form is called the with-in-class imbalance.

Learning from such data requires new tools to transform vast amounts of raw data efficiently into information. Data-level solution has the purpose to resampling the datasets, through random sampling, oversampling or undersampling, in order to obtain balanced or optimal proportions. The data complexity comprises issues as overlapping and lack of representative data, besides the imbalanced representation of the

classes [33, 36]. Next, we briefly describe the Synthetic Minority Over-Sampling Technique (SMOTE) [33, 43] used in this work.

Synthetic Minority Over-Sampling Technique

The method Synthetic Minority Over-Sampling Technique (SMOTE) takes each minority class sample and introduces artificial minority instances, based on the feature space similarities. Depending upon the amount of over-sampling required, neighbours from the *k*-nearest neighbours are randomly chosen and the minority class is over-sampled along the line segments, joining any or all of the *k* minority class nearest neighbours [33, 43].

Synthetic samples are generated in the following way:

(i) One of the *k*-NNs of the X_i is selected to synthesize a new instance based on the instance X_i.

(ii) The difference vector is found between the selected neighbour and X_i.

(iii) This vector is multiplied by a random value between [0, 1].

(iv) The last result is added to X_i and the resulting instance is the new sample.

This approach effectively forces the decision region of the minority class to become more general.

Even though the synthetic examples cause the classifier to learned for the minority class samples, rather than those being subsumed by the majority class samples around them, the SMOTE algorithm has its drawbacks, including over generalization and variance [33, 36].

2.3 Result Evaluation

In the model development process, the available dataset is split into two subsets, the training and the validation set. The training set is used to identify a specific learning model. Use the training set to evaluate the model performance is not acceptable in data mining, generating overoptimistic and overfitted models, acquiring too much detail or noise from the train set. To avoid these problems, the validation set is used to assess the accuracy of the models generated during the training phase, in order to identify the best model for future predictions [38, 40].

One of the methods to evaluating models is the k-fold Cross-Validation, dividing the data into *k* subsets of equal size. This method guarantees that each observation of the dataset appears the same number of times in the training sets and exactly once in the validation sets, achieving an unbiased estimate of the model performance [59]. Usually it is ideal to choose higher values of k in order to obtain a more robust estimate of the accuracy. A popular choice in practice is 10-fold cross-validation, whereby the dataset is partitioned into 10 subsets. If k equals the sample size, this is called Leave-one-Out method [38, 40]. In the present thesis, the performance measures were estimated based on the 5-folds cross-validation method, due to the number of instances in the dataset.

Ideally, an assessment metric should be insensitive to the class proportion. However, the class distribution affects the metrics directly, if theirs definition involves instances of different classes, or indirectly, if denotes a differential error propagation which could lead to misleading information as well [33].

In this thesis, the results are analysed from multiple perspectives using accuracy, sensitivity, specificity and area under the curve (AUC). Although, these metrics provide a simple way of describing a classifier´s

performance, they can be deceiving and are highly sensitive to changes in the data. In order to obtained conclusive evaluations of performance, F-measure is also calculated [33, 43]. These mestrics will be described below.

Confusion Matrix

The Confusion matrices are useful for validating classification models since is not enough to consider the number of accurate predictions but it should also be accounted the type of error committed [33]. Table 1 presents the terms that compose a confusion matrix.

		Predicted Class	
Real Class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Table 1 Confusion Matrix.

True Positive (TP) refers to the number of positive instances that were correctly classified as positive, False Positives (FP) refers to the number of positive instances that were incorrectly classified as positives, False Negatives (FN) is the number of negative instances that were incorrectly classified as negatives and the True Negatives (TN) is the number of negative instances that were correctly classified as negatives.

Accuracy

Accuracy is the ratio between the number of correctly classified instances and the total number of instances defined as $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$. This metric is directly sensitive to class imbalance, not taking into account the imbalance of correct classifications. Therefore, as class distribution varies, the measures of the performance will change even though the underlying fundamental performance of the classifier does not. As stated in previous works [33, 36, 37], the accuracy metric does not provide adequate information on classifier´s functionality. For example, in a data set composed by 5% of minority class against the 95% of majority examples, a Naïve approach would provide an accuracy of 95%. Even though such result represents an outstanding accuracy, this value fails to reflect the fact that none of the minority examples are identified [33, 36, 37].

Sensitivity

Sensitivity measures how many examples of positive class were labelled correctly, represented as the ratio between the correctly classified positive instances over the number of real positive instances. Sensitivity is called

a measure of completeness and is defined as $Sensitivity = True\ Positive\ Rate = Recall = \frac{TP}{TP+FN}$. Sensitivity provides no insight to how many examples are incorrectly labelled as positive [33, 36, 37].

Specificity

Specificity is the proportion of negative instances that were correctly classified as $Specificity = True\ Negative\ Rate = \frac{TN}{TN+FP}$ [37].

F-measure

F-measure is a harmonic mean between precision and recall defined by $F - measure = 2 \times \frac{precision \times recall}{precision + recall}$. The F-measure provides more insight into the functionality of a classifier. This metric is motivated in this thesis since in many studies this is the ultimate measure of performance of a classifier, not depending on disease prevalence [32, 33, 36, 37]. However, we took in consideration the weights of the proportions of how many elements are in each class, considering a weighted F-measure defined as $Fmeasure\ weighted = \left(\frac{TP+FN}{TP+FN+FP+TN}\right)\left(\frac{2TP}{2TP+FP+FN}\right) + \left(\frac{FP+TN}{TP+FN+FP+TN}\right)\left(\frac{2TN}{2TN+FP+FN}\right)$.

Receiver Operating Characteristic (ROC) curves

ROC curves are useful because express the information content of a sequence of confusion matrices and provides a visual representation of the trade-offs between the benefits, reflected by the true positive rate (TPR), and costs, reflected by the false positive rate (FPR), of a given classifier model. These curves are an alternative to the assignment of misclassification costs. The ROC curve is drawn in a FPR (x-axis) versus TPR (y-axis) space, termed the ROC space.

Each classifier produces a (TPR, FPR) pair that corresponds to a single point in the ROC space. Figure 1 illustrates a typical ROC graph with points A, B, C, D, E, F and G. The point D (0,0) incomes that all test instances were labelled negative, the point B (1,0) is the case were all examples were classified incorrectly and the point A (0,1) represents a perfect classification. To compare two classifiers, the best is the one whose corresponding point in ROC space is closer to point A. Any classifier whose corresponding ROC point is in the line linking of (0,0) to (1,1), such as point E, will provide a random guess of the class labels, corresponding to a random classifier. Therefore, any classifier that appears in the lower or upper right triangle of ROC space performs worse or better than random guessing. Nonetheless, a point under the random line may represent an informative classifier whose information is incorrectly applied. Inverting the classification results of classifier F, the point G will be produced with the symmetric classification of point F.

In the case of probabilistic classifiers, a ROC curve is produced by a series of points varying the threshold of the continuous numeric values that represent the confidence of an instance belonging to the predicted class. In these cases, generally is used the area under the curve (AUC) to provide the average performance of the classifier. When working with non-probabilistic classifiers that return only the predicted class, the AUC is given by $\frac{TPR+TNR}{2}$, named as average accuracy [32, 33].

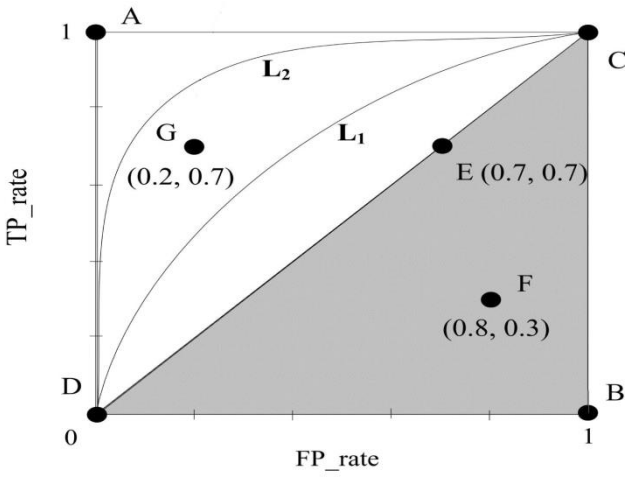

Figure 1 ROC curve representation [33].

In addition, visualization techniques also play an important role in the correct preprocessing and classification of the data. The visual tools as the boxplots and histograms used in the present thesis were adapted from Lemos et al. [48].

2.4 Related Work

Earlier works have been developed to predict MCI conversion to AD focused on neuropsychological tests, MRI and PET scan measures and biomarkers through diverse approaches.

Hinrichs *et al.* [3] aimed to introduce a new machine learning technique, based on the Multi-Kernel Learning (MKL), to predict markers for AD using different modalities. MKL allowed the incorporation of imaging modalities, as MR and PET scans, cognitive status and biological measures from CSF assays, Neuropsychological Status Exams (NPSE) and APOE genotype, to calculate various kernel matrices from each one of the modalities and to train an optimal combination of kernel and classifier. Firstly, the authors proposed a single predictive framework to classify AD, MCI and control subjects, obtaining 48 AD, 66 healthy controls and 119 MCI, in a total of 233 ADNI subjects. The results were used to project a Multi-Modality Disease Marker (MMDM) for MCI individuals and predict conversion to AD.

Using a 2-norm MKL, the imaging modality had 87.6% of accuracy, 78.9% of sensitivity and 93.8% of specificity; the biological measures had 70.4% of accuracy, 58.1% of sensitivity and 79.4% of specificity and the cognitive measures had 91.2% of accuracy, 89.2% of sensitivity and 92.6% of specificity. Although cognitive measures presented a good classification, combining all modalities, using a combination of kernels matrices while simultaneously training a classifier, t he best classifier was performed with a 92.4% of accuracy, 86.7% of sensitivity and 96.6% of specificity.

The results allowed the authors to understand that neurological biomarkers take increased importance after MCI diagnosis. The authors also assumed that NPSEs data can work better as a marker for individuals who have already had any AD cognitive declines and imaging modalities as future declines predictors.

Maroco *et al.* [17] calculated the sensitivity, specificity, global accuracy, AUC and Press'Q, to evaluate the performance of a classifier in prediction of evolution into dementia of MCI patients. Four hundred subjects were analysed, 257 labelled as MCI and 125 as AD patients, from the Laboratory of Language Studies, Santa Maria Hospital, and Memoclínica, from 1999 to 2007, with at least one follow-up neuropsychological assessment or clinical re-evaluation. The Neuropsychological predictors were a subset of tests from the Battery of Lisbon for the Assessment of Dementia (BLAD) such as Verbal Semantic Fluency, Interpretation of Proverbs and the Raven Progressive Matrices, Clock Draw, Digit Span Forward, Digit Span Backward, Word Recall, Verbal Paired-associate Learning and Logical Memory, the Mini-Mental State Examination and the Forgetting Index.

The classifiers used were the Linear Discriminant Analysis (LDA), the Quadratic Discriminant Analysis (QDA), the Binomial Logistic Regression (LR), the Neural Networks (NN), as Multilayer Perceptrons and Radial Basis Networks, the Support Vector Machines (SVM), the Classification Tress (CT), as Classification and Regression Tree (CART), Chi-squared Automatic Interaction Detector (CHAID), Quick Unbiased Efficient Statistical Tree (QUEST) and the Random Forests (RF).

Regarding the overall accuracy, SVM and RF had higher mean ranks, with 76% and 74%, respectively, from the other classifiers, which did not differ significantly in mean rank accuracy. On the topic of specificity, SVM had acquired 100%, followed by a second group composed by MLP, LR and RBF, with significant differences from a third group composed by LDA, QDA, classification trees and RF, with 64%. LDA, CART, QUEST and RF had the highest sensitivity values with 66% and SVM was the classifier with the significantly lowest sensitivity, with 30%. SVM showed the highest AUC and LDA, LR, MLP, RBF and RF were a homogenous group statistically different from the group composed by QDA, CHART and CHAID. QUEST had the significantly lowest AUC. The results of AUC related to the discriminant power of the classifiers, was appropriate for most classifiers, greater than 0.7 with the exception for classification trees with median AUC of 0.6. Relating to the Press'Q, SVM had the highest mean rank followed by RF, MLP, CHAID and LR.

In this analysis, some of the classifiers with the highest specificity, Neural Networks and SVM, were also the classifiers with the lowest sensitivity. This is due to the fact that small sample size and the chosen values for the tuning parameters affect the performance of some data mining methods assessed in this study. For binary classification problems, where sample size may compromise training and testing of popular data mining and machine learning methods, Random Forests and Linear Discriminant Analysis proved to have higher accuracy, sensitivity, specificity and discriminant power.

Lemos *et al.* [48] aimed to derive a classifier that predicts if a patient already suffers from AD or is already a MCI stage. The information used in this study was from Cognitive Complaints Cohort at the Institute of Molecular Medicine (IMM), using only patients labelled as MCI or AD and with at least two evaluations, considering only neuropsychological data. There were 583 patients labelled as MCI and 94 labelled as AD.

The authors had to deal with the high class imbalance problem between AD and MCI classes and with the high percentage of missing values, around 48%, in order to build the classifier that would maximize the prediction of AD subjects. To overcome the class imbalance problem, the authors used a Synthetic Minority Over-sampling Technique (SMOTE) on the minority class and perform a mean value imputation whenever the classification algorithm did not support the missing values. The authors also applied feature selection to the dataset, using filter-type techniques based on correlation, chi-squared statistic, gain ratio, information gain and

consistency subset evaluation. The classification techniques used to differentiate between MCI and AD were K-nearest techniques, C4.5 decision tree, Naïve Bayes, Artificial Neural Network and Support Vector Machine.

To assess the results, the authors applied a 10-fold cross validation. To compare the classifiers accuracy, sensitivity or true positive rate (TPR), specificity or true negative rate (TNR) were computed. The statistical significance of the different classification models was evaluated with 30 repetitions of a 10-fold cross-validation and a two sample paired t-test, to validate the null hypothesis that the classification results have similar distributions at 95% of confidence.

Two approaches were analysed, the first one considering the first and last evaluation of the patient and the other considering a time frame approach. By using a temporal window approach, they obtained better discriminative results. The authors concluded that the best models to use are the NB and SVMs algorithms, and that the mRMR feature set performed with better results than those using the original set or the correlated set. The temporal window with the highest discriminative power was the three years window. Using this window, the best model was obtained using radial SVM algorithm, with an accuracy of 82%, a sensitivity of 79%, a specificity of 86%, a ROC Area of 0.83 and a |TPR - FPR| of 0.64.

Since part of the work in the present thesis is related to differentiate MCI groups in order to improve the prediction of AD conversion, it was important to be aware of the work already done in this direction.

Espinosa *et al.* [2] intended to estimate conversion rates in different MCI subtypes and to determine neuropsychological test performances. For this purpose, the authors had access to a longitudinal follow-up between January 2006 and November 2011 of 550 MCI individuals with neurological examination, neuropsychological testing, social work evaluations, vascular risk factors and APOE genotyping. These subjects were classified according to Peterson´s criteria [19] and Lopez et al. [58] classification, in amnestic and non-amnestic single and multiple domain, considering Domain Pattern Impairment, in storage or retrieval amnestic MCI, considering Memory Pattern Impairment, and possible or probable MCI when there were comorbidities or there were none that could explain or contribute to cognitive deficits, respectively. The neuropsychological battery of Fundacío ACE (NBACE) included tests sensitive to processing speed, orientation, attention, verbal learning and memory, language, visuoperception, gnosis, praxis, and executive functions.

To analyse predictors of conversion, Cox proportional hazards analysis were used introducing different variables as NBACE variables, probable or possible MCI criteria, DPI, MPI, and presence or absence of at least one APOE ε4 allele, according to the different models considered.

The analysis time was in average 26.6 months, with 20.9% of Pr-aMCI, 42.5% of Pss- MCI, 6.7 % of Pr-naMCI, and 29.8% of Pss-naMCI. According to the DPI classification criteria, were classified 86.1% as Pr-aMCI-md, 88.5% as Pss-aMCI-md, 64.9% as Pr-naMCI-md and 73.8% as PssnaMCI- md, demonstrating that most of MCI patients displayed a multiple domain cognitive impairment and that multiple domain was more common in amnestic compared to non-amnestic MCI. In the amnestic MCI groups, in terms of the MPI, a memory storage deficit was more frequent, with 81.7%, among those subjects with Pr-aMCI than in Pss-aMCI patients, which corresponds to 57.7%. Probable amnestic-storage group had the highest risk of conversion to dementia, having 8.5 times more risk to convert to dementia than the possible non-amnestic MCI group, the group that resulted in the slowest conversion to dementia.

The analysis also demonstrated that orientation and verbal recognition are of great importance in the assessment of the amnestic and non-amnestic probable MCI subtypes and the authors support a therapeutic

intervention in probable amnestic-storage individuals, since almost all of them converted to dementia, especially to AD.

Kim *et al.* [21] proposed to evaluate if the performance of the mini-mental state examination (MMSE) could be used to identify groups with higher risk to convert to dementia. The patients analysed were recruited from the Clinical Research for Dementia of South Korea (CREDOS) study with a total of 519 individuals diagnosed with MCI, according to Peterson's criteria. Having in consideration the results of the Seoul Neuropsychological Screening Battery and the four cognitive domains, like memory, language, visuospatial and frontal functions, these patients were divided in 122 amnestic MCI-single domain (ASM), 303 amnestic MCI-multiple domains (AMM) and 94 non-amnestic MCI (NAM). Seven areas were analysed with the MMSE battery, such as orientation to time, orientation to place, registration of three words, attention and calculation, delayed recall of three words, language and copying interlocking pentagons.

In order to compare patient demographic findings, the authors used a Pearson's Chi-Square ($\varkappa^2$) test for categorical data and the ANOVA analysis for continuous variables. To compare the performance among the three subgroups, the univariate general linear model with adjustments to age and education was used. With the purpose to establish prediction models for each cluster, the authors performed a two-step cluster analysis, relating age, education and and implementing a discriminant analysis, which allowed obtaining two discriminant equations. Besides this, to understand the difference of conversion to dementia in the clusters, Kaplan-Meier survival analyses were used, with time to the event being defined as the time from study entry to the follow-up visit at which a first-time diagnosis of dementia was made. All the statistical analyses were developed in SPSS version 20.0.

The clusters obtained were divided in three groups as well, Cluster 1 consisted of 205 patients with AMM, corresponding to a 100 %, Cluster 2 was composed of 61 NAM, 33.3 %, and 122 ASM, 66.7 %, and Cluster 3 consisted of 33 NAM, 25.2 %, and 98 AMM, 74.8 %. Cluster 3 was significantly older, had a lower educational level and showed significantly lower ability for orientation to time and place, registration of three words, attention/calculation, language, and copying interlocking pentagons than clusters 1 and 2. At delayed recall, cluster 1 was significantly more impaired than cluster 2. For predict each cluster, the authors constructed a flowchart using the two equations obtained. From those, it was achieved a 64.4 % overall accuracy, with clusters 1, 2, and 3 acting out with 60.0, 44.8, and 98.5 % sensitivity and 72.0, 78.0, and 94.1 % specificity, respectively. The cluster 1 represented a more risk group for conversion to dementia, with 100% of AMM, whereas reversion into normal cognition occurred most commonly in the cluster 2. Cluster 3, which consisted of AMM and NAM, represented the oldest and lowest educational level patients, which seemed to be related with the risk for converting to dementia.

This study showed clustering the MCI group is a promising screening tool that could help define more risk groups for conversion to dementia.

3. Experimental Methodology

This chapter will describe the dataset under study as well as the preprocessing steps performed, including data transformation, the way of handling with missing values and the method of feature selection used. Besides this, the Classification Model used throughout this thesis will be also introduced.

3.1 Description of the Data

The Dementia Group at Instituto de Medicina Molecular (IMM) directed the Cognitive Complaints Cohort (CCC) study [15, 17] to investigate cognitive stability or dementia on subjects with cognitive complaints. The dataset, resulted from the application of neuropsychological tests, had in consideration the inclusion and exclusion criteria specified by Silva *et al.* [16] and medical doctors decision and was provided to the investigations occurring in the NEUROCLINOMICS project in INESC-ID Lisboa.

The dataset analysed in this work (Table 2) is composed of 1827 instances consisting of individual evaluations of 990 patients, categorized as Normal, Pre-MCI, MCI, AD and No Diagnosis, considering the patient's cognitive condition in their follow-up at IMM. The dataset has 192 attributes, both categorical and numerical, consisting in the results of the neuropsychological evaluation procedures, listed in Appendix A. Each patient has several evaluations (considered as different instances) where he/she was diagnosed with different stages of MCI or as AD, during the follow-up.

	Normal	**Pre-MCI**	**MCI**	**AD**	**No Diag**
Observations (%)	281 (15.4%)	79 (4.3%)	1248 (68.3%)	196 (10.7%)	23 (1.3%)
Age (Mean±SD)	64.8±10.2	65.8±9.1	70.2±8.6	73.8±8.3	71.7±7.3
Sex (Male/Female)	74/207	36/42	516/730	67/127	3/18
Schooling Years (M±SD)	10.1±4.7	10.8±4.8	13±4.9	8.8±5	6±3.3

Table 2 Dataset details.

3.2 Description of the Tools

The computational tools used in this work were MATLAB R2013a and Waikato Environment for Knowledge Analysis (WEKA) Version 3.7.11. MATLAB was used in the identification of missing values in the preprocessing steps. During all the work, Java programming language was used to unify different tasks from different sources.

3.3 Dataset Preprocessing

3.3.1 Outlier Detection

Outliers are objects with very extreme values in one or more attributes and can dramatically influence the results. The treatment depends on the nature of the outlier. Although the common method for identifying them relies on graphical techniques, nowadays the increase on the size of the databases forced the use of a variety of automated techniques [36]. In the present work, the outliers were identified by observing the distribution of each attribute. Since the number of outliers was small they were corrected manually.

3.3.2 Data Cleaning

Before explaining the steps made to clean the data, it is important to clarify the number of patients and their state in the first evaluation. Table 3 lists such values. Figure 2 shows the number of patients with one and more than one observations.

	Normal	**Pre-MCI**	**MCI**	**AD**	**No Diag**
Observations (%)	281 (15.4%)	79 (4.3%)	1248 (68.3%)	196 (10.7%)	23 (1.3%)
Patients (%)	179 (18.1%)	58 (5.9%)	738 (74.5%)	6 (0.6%)	9 (0.9%)

Table 3 Description of the patients in the first evaluation.

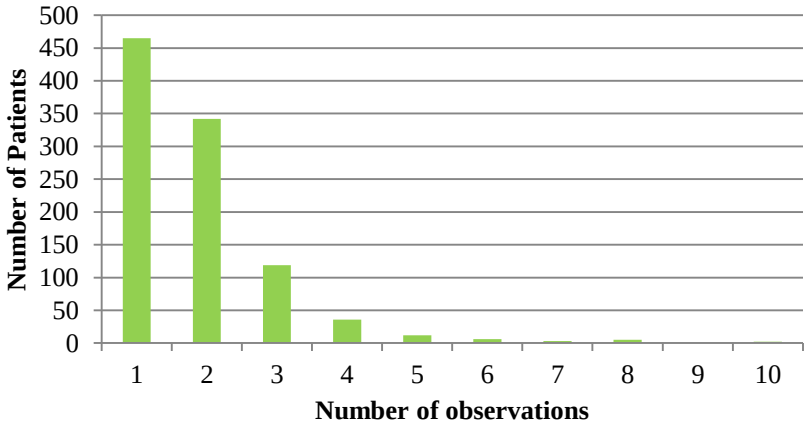

Figure 2 Histogram of the number of observations from all patients.

The first step was the exclusion of observations where the patients, in the different phases of the follow-up, are labelled as normal (281 observations), as pre-MCI (79 observations) and as without diagnosis (23 observations), resulting in the removal of 383 observations. Thus, the remaining dataset contains patients with evaluations labelled as MCI and AD. Later, the patients with only one evaluation corresponding to 384 (near 48.6% of the remaining patients) were eliminated leading to a significant reduction of number of patients analysed. This was performed since supervised learning needs, at least, two observations: one (or more) with the patient characteristics at a given time point and one with the prognosis outcome classes. Figure 3 shows the distributions of evaluations in the MCI and AD patients.

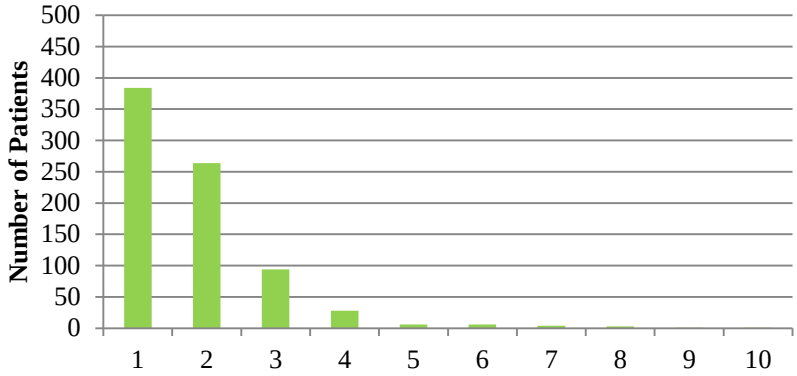

Figure 3 Histogram of the number of observations of MCI and AD patients.

The remaining patients constitute MCI and AD patients with two or more observations (Table 4). Since we want to predict conversion from MCI to AD, we only consider patients whose diagnosis in the first evaluation is MCI. In the following observations the patient can remain MCI or can evolve to AD.

	MCI	**AD**
Patients in the first evaluation (%)	407 (100%)	0 (0%)
Observations (%)	878 (82.8%)	182 (17.2%)

Table 4 Composition of the analysed patients.

3.3.3 Creating learning examples

The dataset is splitted in 75% for training and 25% for validating. This stratified partition had in consideration the distribution of attributes such the number of evaluations, the age, the sex, the schooling years and the class, keeping them constant in the training and validation sets.

In order to predict the conversion of MCI patients, two approaches were used. The first, which constitutes the baseline of this work and is normally used in similar problems [8, 9, 12], looks at the first and last evaluations of the patient in the dataset to see if the patient will ever convert from MCI to AD (Figure 4). A new set of learning examples is then created, as evolution (Evol) or no evolution (noEvol) instances. In this approach, called First Last Approach, each patient has only one entry in the post-processed dataset.

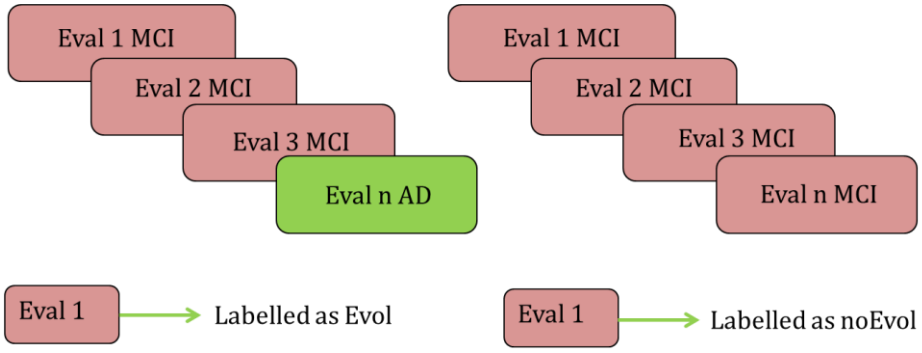

Figure 4 New class labels created for the First Last prognosis problem.

The second approach is related to the temporal window prognosis problem: predicting if a patients converts to AD in a given temporal window (Figure 5). The new set of learning examples is labelled as evolution (Evol) if

within the specified temporal window the MCI patient converts to AD and as no evolution (noEvol) if after the specified temporal window the MCI patient remains MCI. Moreover, there are instances where is not possible to define the class: instances with the second evaluation categorized as MCI still within the temporal window, without knowing the state of the patient in the next evaluation, and instances with only one observation in the specified temporal window. These instances are then removed, not creating learning examples. One patient that is not considered in one temporal window analysis can eventually be considered in another one. Each temporal window gives rise to different learning examples. The choice of the temporal window had in consideration the instances distribution between classes (Evol/noEvol) and the medical relevance of the problem, defined by consulting the medical partners of the NEUROCLINOMICS project. Lemos *et al* [48] considered the two, three and four years temporal window. In this thesis we go further on this topic and we analyse the five years temporal window, since data is now more complete and adequate for this task.

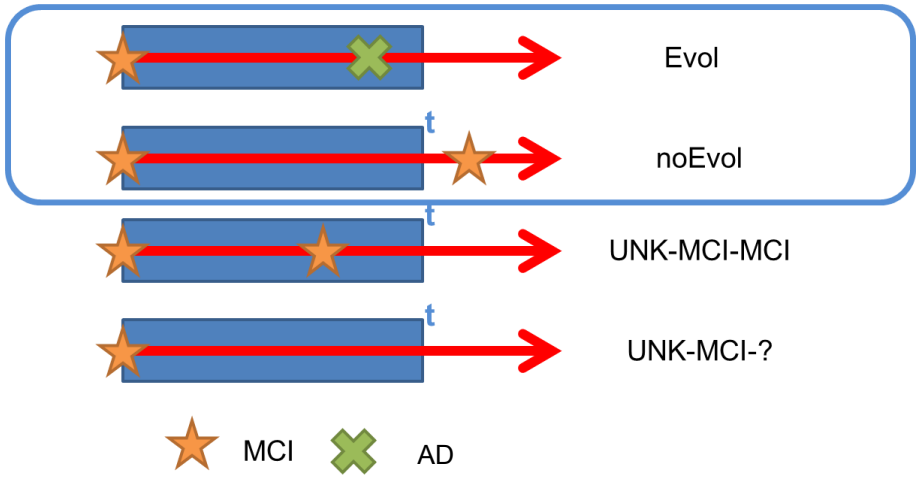

Figure 5 New class labels created for the Temporal Window prognosis problem [48].

In both approaches the Evol class is considered the positive class.

Instances with undefined class were excluded. Attributes such Case_number_for_this_dataset, Date, Observation_number_for_each_case and DiagNPS were removed since were too discriminative. Attributes as end_point, evolution_type, time_assessmet1_endpoint, TMTs_incomplete, time_begin_endpoint, Neuropsy_assessment_clinical, and Diagnosis-Code were not used in this analysis since were considered non-informative by the doctors, yielding a total of 182 attributes.

Tables 5 and 6 present the composition of the train and validation datasets used in the present thesis.

Total instances	noEvol	Evol
First Last approach	181 (59.7%)	122 (40.3%)
2Years Window	201 (70.5%)	84 (29.5%)
3Years Window	139 (54.7%)	115 (45.3%)
4Years Window	85 (37.3%)	143 (62.7%)
5Years Window	51 (23.7%)	164 (76.3%)

Table 5 Sizes of the train datasets after the above preprocessing steps.

Total instances	noEvol	Evol
First Last approach	66 (63.5%)	38 (36.5%)
2Years Window	61 (66.3%)	31 (33.7%)
3Years Window	40 (49.4%)	41 (50.6%)
4Years Window	24 (33.3%)	48 (66.7%)
5Years Window	16 (23.9%)	51 (76.1%)

Table 6 Sizes of the validation datasets after the above preprocessing steps.

3.3.4 Handling Missing values

Missing values and class imbalance may lead to misleading results and are thus important challenges [34].

Typically, databases include more than half of the entries as missing, including wrong manual data entry procedures, equipment errors and incorrect measurements. These empty entries make very difficult to mine the data, producing severe mistakes in calculation if not properly treated [35].

There are many strategies to deal with missing values, but it is important refer the existence of non-random missing values. Non-random missing values are produced intentionally and can thus have discriminative power. This kind of missing values are very common in clinical data. For example, a doctor may choose not to perform a specific test, if a patient is already too tired or achieving a low score in a related test or due to time restrictions of the patient [34, 35, 36]. The simplest way to deal with missing values is to discard the examples that have missing values. This was only reliable if the data contained a relatively small number of examples with missing values, which is not the case in this work. Lemos *et al.* [48] tested a set of strategies to deal with missing values and we followed the strategy they considered better: let the classifiers deal internally with the missing values. The Naïve Bayes classifier excludes the missing values from the calculations and the SVMs, the KNN and the C4.5 Decision Tree classifier use the internal way of median/ mode imputation, in case of numerical/categorical values, respectively. This strategy was applied to train and validation sets.

Most of the instances have between 10 a 70% of missing values. Initially, we thought to remove the instances with more than 70%, representing a minor portion of instances, or instances with 95% of missing, whose values will only contribute with noise to the classification. However, with the purpose of not introduce bias to the datasets removing important instances, no removal was made. After feature selection the percentage of missing values per instance and per feature were analysed as described in the next section.

3.3.5 Feature Selection

Attribute selection was performed for each dataset individually using a supervised filter. This kind of filters has two parameters: the type of subset evaluator and the search method. We used correlation-based feature subset selection and a greedy search, which evaluates the value of the attribute subsets considering their individual predictive ability and the redundancy among them [47].

When feature selection was performed in the different datasets, the percentage of missing values was also analysed. As shown in Figure 6, the features selected reached a maximum of 50-60% of missing values. Due to this, no exclusion of features was performed in the preprocessing steps. Figure 7 shows that after the feature

selection there were still instances with high percentages of missing values. Thus, the next step was to exclude instances with more than 50% missing values.

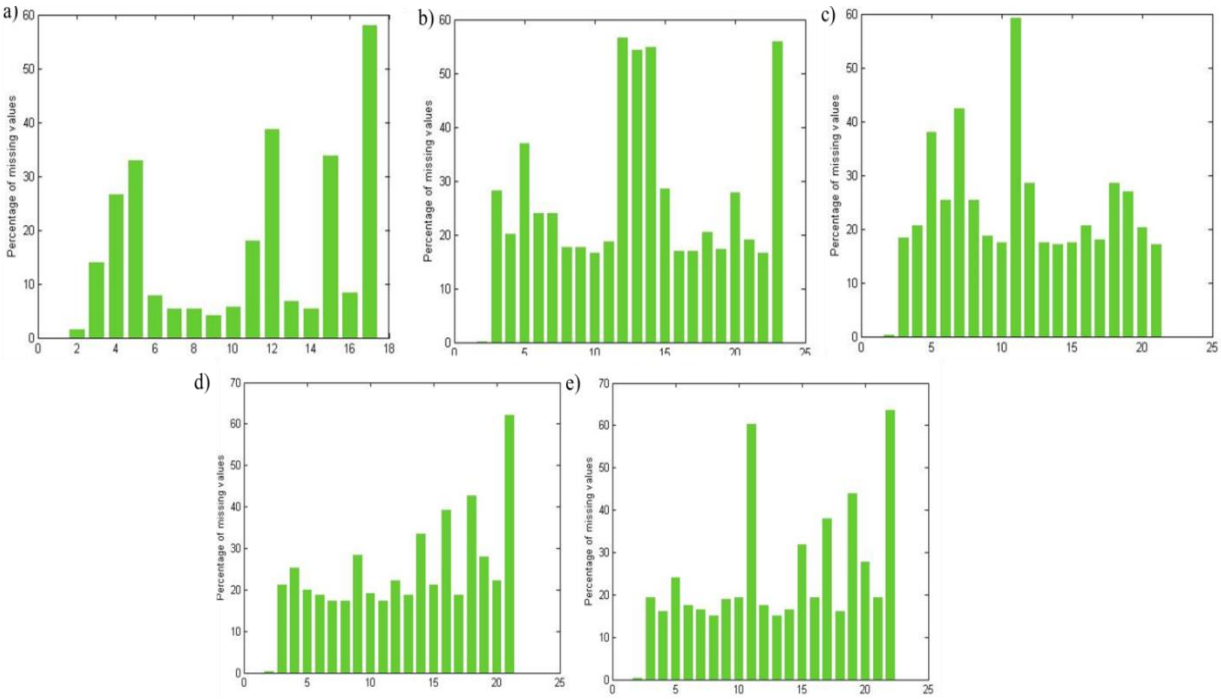

Figure 6 Percentage of missing values per feature selected by feature selection: a) in the First Last approach; b) in the two years temporal window; c) in the three years temporal window; d) in the four years temporal window; e) in the five years temporal window.

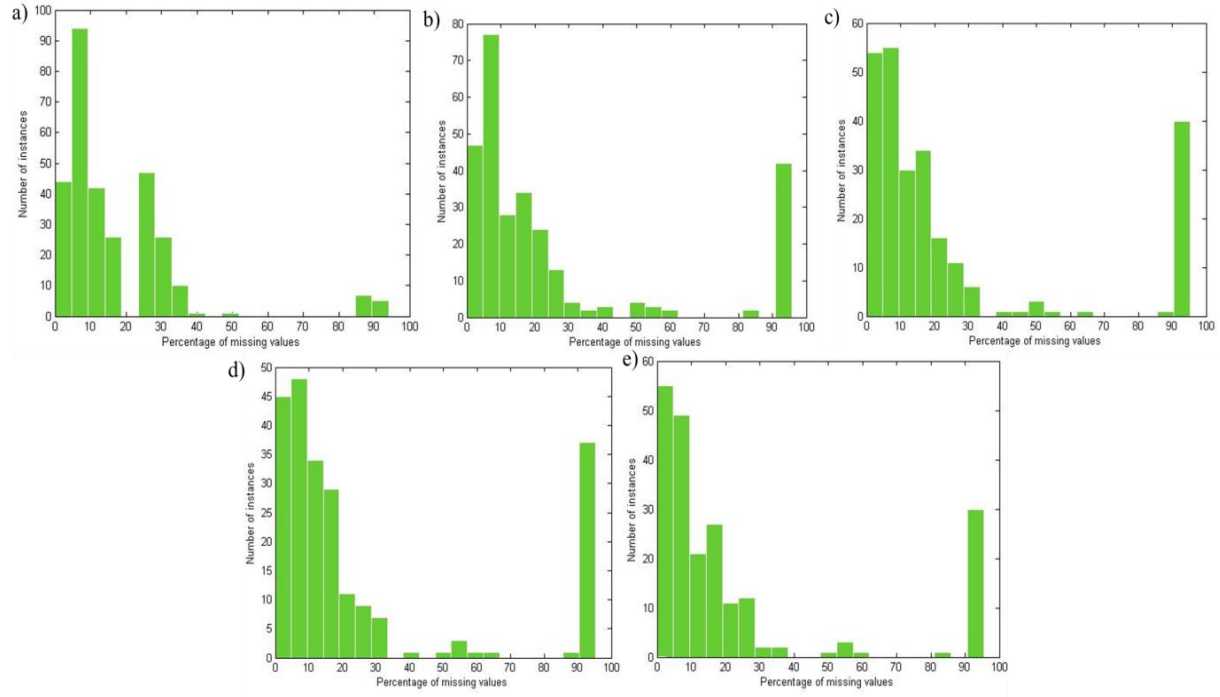

Figure 7 Numer of instances per percentage of missing values after feature selection: a) in the First Last approach; b) in the two years temporal window; c) in the three years temporal window; d) in the four years temporal window; e) in the five years temporal window.

In the present thesis, the attributes considered important for the medical doctors presented a normalization version to age and schooling years. The dataset is then composed by attributes with correction for age and school, represented by a z-score version, and attributes without correction. In this analysis we did not perform a discretization of the attributes since the better results were always obtained by feature selection. Table 7 shows the composition of the data after the preprocessing steps referred above.

Total instances	noEvol	Evol
First Last approach	171 (58.8%)	120 (41.2%)
2Years Window	161 (68.8%)	73 (31.2%)
3Years Window	112 (53.1%)	99 (46.9%)
4Years Window	66 (35.7%)	119 (64.3%)
5Years Window	42 (23.3%)	138 (76.7%)

Table 7 Sizes of the train datasets after Feature Selection.

Figure 8 presents a summary of the preprocessing steps.

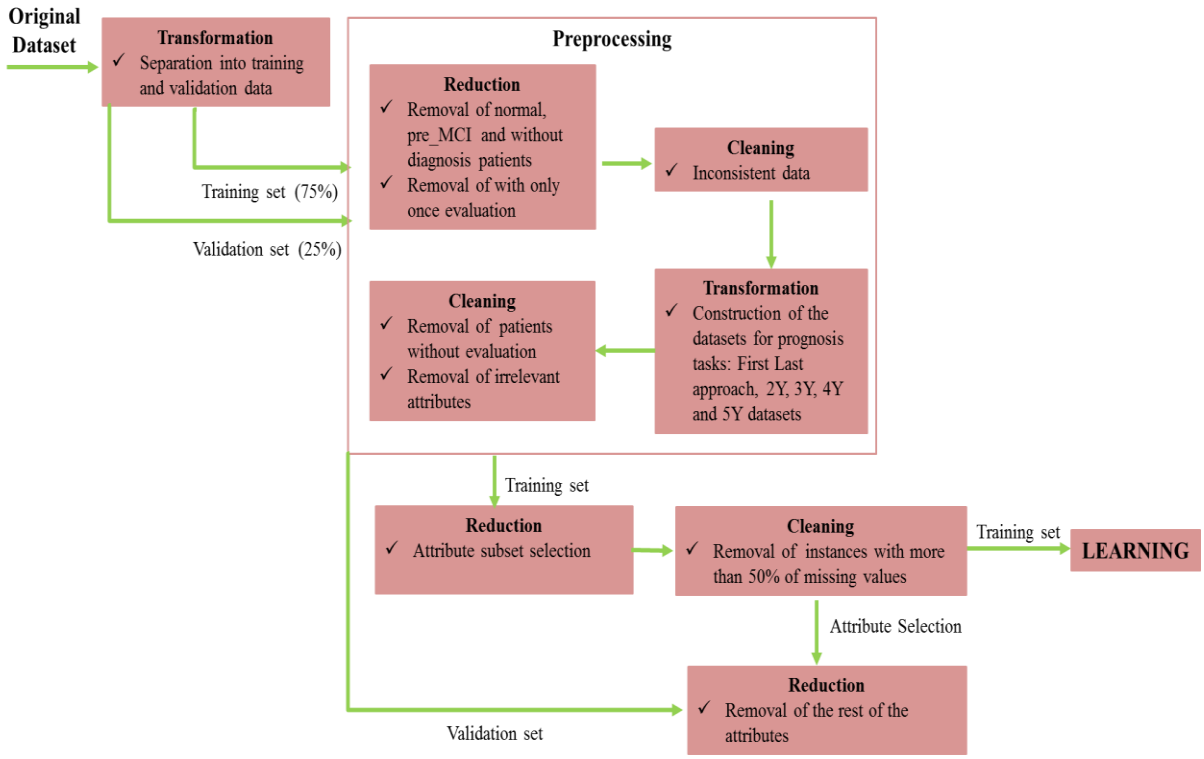

Figure 8 Preprocessing steps applied to the dataset.

3.4 Classification

In the methodology followed, five classifiers were used: NB, Radial Basic Functions SVM, Polynomial SVM, *k*NN and C4.5 DT, all implemented in WEKA.

Since, most classifiers are highly sensitive to the irregular distribution in class the data balancing technique SMOTE was used [43]. This algorithm is also implemented in WEKA. The classifier parameters and the percentage of oversampling to be used are determined using 5-fold cross validation on a greedy search approach.

As already mentioned, the classifiers have internal ways to deal with missing data. In both train and validation set, the NB classifier excludes the missing values from the calculations and the SVMs, the *k*NN and the C4.5 DT classifier use the internal way of median/ mode imputation, in case of numerical/categorical values, respectively.

3.4.1 Classification Model

Firstly, the classification model explained already was applied to the First and Last Evaluation Prognosis Problem, with the purpose of constructing a baseline comparison for further analyses. Afterwards the different temporal windows containing all patients was analysed regarding the same principles, expecting to outperform the baseline.

In a second phase of the work, two different studies are performed using different MCI groups. The first approach considers important characteristics of the patients as the Geriatric Depression Scale (GDS) of the patients. Depression is associated with MCI patients, however is still unknown the relation to Alzheimer's disease [22, 30]. Moreover, this state of the patients can be only temporary. As previously stated, an important approach would be to separate the depressed from the not depressed patients and to understand the relationship between these conditions and whether it would be useful to give different treatment to patients with a high level of depression for the prognosis of dementia. Clustering models is the second approach of this data mining project that will be followed by the application of the aforementioned methodologies within each cluster. Two clustering methods were applied, the K-Means and the Expectation Maximization (EM). Both of these algorithms are implemented in the software package WEKA. During the clustering study, respecting the definition of unsupervised learning, the target attribute (Progression classification) was not used when grouping the patients. Besides this attribute, the patient´s identification number was also excluded to avoid biasing the results. This analysis consisted on trying different number of clusters, in particular, 2, 3 and 4, with the purpose of identify the putative differences of MCI patients.

Training Model

To determine the best SMOTE percentage, it is important to have in consideration that such percentage changes for each problem, subset of features, classifier parameters and classifier itself. In order to determine the best SMOTE percentage and the best classifier parameters, it was necessary to cross all tested SMOTE percentages with all tested parameters sets in a grid search of a parameterized model. The classification model used in this thesis was the automated model created by the previous work of Lemos *et al.*[48] with slight changes, in order to optimize the grid search process.

The metric used to compare models for each SMOTE and parameter set was the F-Measure. This choice relates to the fact that the other metrics are highly sensitive to imbalanced data and the F-measure metric is a trade-off between the sensitivity and specificity, not depending on disease prevalence. Both the parameters sets and the SMOTE percentages are tested with 5-fold cross-validation. The SMOTE percentage was tested using 11 different values for each parameter combination.

The grid search was performed in all datasets and classifier models. The parameter intervals are presented in Table 8. After the search, five best triples are determined {*Classifier, Parameters, SMOTE percentage*}, one

for each classifier and were tested in 30 repetitions, using different seeds in the 5-fold cross validation for each repetition. A common way to test whether the difference between the results of two classifiers is non-random is to compute a paired t-test. In the present study, a paired t-test using the 30 repetitions was used and the t-test was only applied if the ANOVA test with 95% confidence level confirmed the existence of a significant difference [46].

Classifier	**Parameters**	**SMOTE (**11 steps)
Naïve Bayes	*Gaussian* or *Supervised Discrimination* or *Kernel*	0% to the inversion of the imbalanced
SVM RBF	*Complexity* ϵ [1,10] and $\gamma \epsilon [10^{-5}, 10^{-2}]$	0% to the inversion of the imbalanced
SVM Poly	*Complexity* ϵ [1,10] and Degree ϵ [0.5, 5.0]	0% to the inversion of the imbalanced
DT C4.5	*Confidence* ϵ [0.05, 0.5]	0% to the inversion of the imbalanced
*k*NN	$k \epsilon$ [1,10]	0% to the inversion of the imbalanced

Table 8 Grid Search parameter interval.

The feature selection is accomplished outside the cross-validation during the preprocessing step. Such decision was made to provide to the medical doctors the selected features instead of performing this evaluator inside the cross-validation.

Validation Model

The validation model was obtained by splitting the original dataset in 75% for training and 25% for validating. As aforementioned, this stratified partition consider the distribution of attributes as the number of evaluations, the age, the sex, the schooling years and the class, keeping them constant in the training and validation sets. The overfitting problem, which would put in cause the generalization of the method, is minimized with the split of the dataset, since the features and the parameters are only selected using 75% of the data.

The aim of the validation set is to evaluate the final models created during the training phase, by analysing the behaviour of the trained models in a real-world simulation, since the model has never been in contact with any instance of the validation patients.

The preprocessing steps were applied in an equivalent way to training and validation sets, except for the attribute subset selection. The attributes obtained in the training set after applying the filter were extrapolated directly to the validation sets.

Figure 9 presents the data stream in the grid search, including the diferent groups and analyses considered in this work.

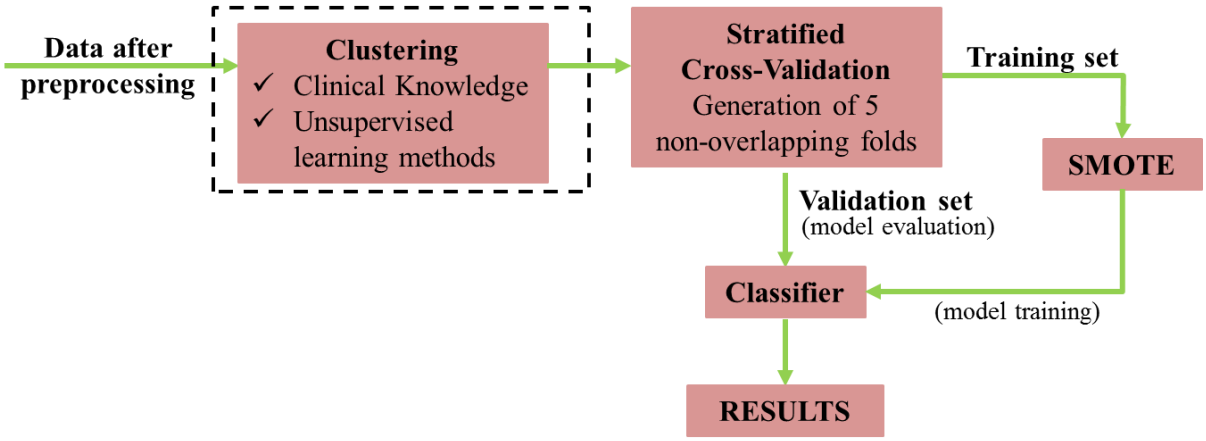

Figure 9 Data flow used in the parameter grid search for finding the classifiers parameters [48].

4. Predicting conversion from MCI to AD

In order to obtain the highest benefit from disease-modifying drugs once they became available, a prognosis of conversion to AD of patients in the prodromal stage is fundamental. This chapter describes a supervised learning analysis proposed to predict the conversion of MCI patients to AD. The data analysis was performed using the standard classifiers, NB, SVM RBF, SVM Poly, C4.5DT and *k*NN.

The demographic characteristics of the patients are important aspects to take in consideration. There are many studies relating the age, schooling years and the sex of the patients with the predisposition to progress to dementia [4, 7, 9, 16, 17, 30]. Additionally, it is unclear if the use of age and education corrections improves the prognostic value to dementia. In particular, Hessler *et al.* [45] investigated this problem and their analysis provided evidence that the omission of age norms in the diagnosis of MCI leads to a more accurate identification of individuals at risk to dementia and that the application of the education correction, which is often neglected, likely increases the predictive accuracy.

In order to corroborate such results and understand if the predictive value would be enhanced by taking these conditions in consideration, three different datasets were analysed separating the attributes corrected for age and school from the ones that were not normalized. In order to clearly understand the specifics of learning with the two distinct datasets, we applied feature selection. It was interesting to notice that the attributes chosen in the dataset comprising the attributes without correction were the ones corresponding to the corrected attributes selected in the dataset comprising the z-scores. The values of F-measure achieved in the studies do not differ significantly, which could also be understood by the fact that the attributes available present a simultaneously normalization both to age and schooling years. The dataset containing all features obtained better performances, creating a more generalized model. Hence, we had this in consideration during the following analyses.

As aforementioned, two approaches were considered to predict the conversion of MCI patients:
(i) predict if the patient will ever convert to AD corresponding to the First Last Approach (Baseline);
(ii) predict if the patient will evolve to AD in a given time window corresponding to the Time Window Approach.

4.1 First Last Approach

The dataset characteristics relative to the First Last approach is described in Table 9. It is interesting to notice the demographic characteristics of this set. The mean age of the evolution group is significantly higher than the no evolution group. The schooling years are relatively lower in the Evol group.

Groups First Last	noEvol	Evol
Total instances (%)	248 (60.8%)	160 (39.2%)
Age (Mean±SD)	68.8±8.5	71.7.0±7.9
Sex (Male/Female)	106/142 (42.7%/57.3%)	56/103 (35%/64.4%)
Schooling Years (Mean±SD)	9.1±4.9	8.6±4.7

Table 9 Dataset demographic characteristics after applying pre-processing for the First Last approach.

4.1.1 Cross-Validation results

The first problem to address is the class imbalanced found in the datasets used for prognosis. As hereinbefore, for simple binary classification problems, sample size may compromise training and validation [17] and the SMOTE algorithm was used to overcome this issue. Results are shown for the evaluation with and without the use of SMOTE in order to understand how the classifiers behave and how the dissimilar proportions of classes affect the evaluation. Table 10 presents the characteristics of the train dataset used.

First Last approach	noEvol	Evol
Total instances (%)	181 (59.7%)	122 (40.3%)

Table 10 Size of the train dataset after applying pre-processing for the First Last approach.

The proportion of classes is 59.7% of the instances classified as noEvolution (noEvol) and 40.3% classified as Evolution (Evol). Figure 10 presents the train results without the application of the SMOTE algorithm. The train results with the application of the SMOTE algorithm may be consulted in Appendix B.

After comparing the results with and without contemplating the SMOTE algorithm, we can conclude that the application is not so beneficial in this dataset and its use may lead to the overfitting of the data and bias the results. In the following analyses, the application of the SMOTE algorithm is only considered if the proportion of the majority class is above 70%.

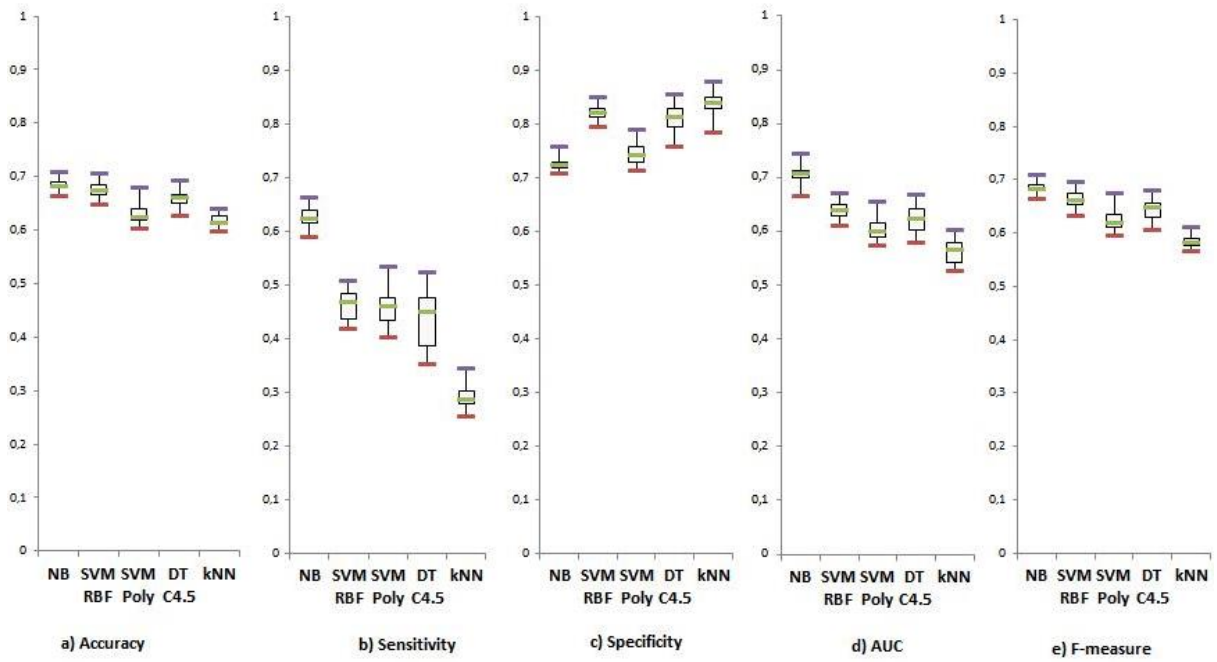

Figure 10 Train results of Prognosis without applying SMOTE using First Last approach.

Paired t-tests were used to compare the classifiers, using 30 repetitions. The t-test is only applied if the ANOVA test with 95% confidence level confirms the existence of a significant difference. The classifier NB is the best model and the *k*NN model got the worst results. Both present a significant difference among the other models.

Table 11 shows the confusion matrix of the NB classifier, showing that the model is learning both classes, with lower sensitivity and specificity, as seen in Figure 10. Table 12 presents the results for NB with the actual value of the assessed metrics.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	78 (63.9%)	44 (36.1%)
	noEvol	47 (26%)	134 (74%)

Table 11 Confusion Matrix of NB for the First Last approach, using 5-fold CV.

Classifier	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**F-measure**
NB	0.684±0.010	0.625±0.018	0.724±0.012	0.706±0.014	0.685±0.010

Table 12 Evaluation metrics of NB for the First Last approach, using 5-fold CV.

Feature selection was performed in order to determine if a reduction in the number of attributes used would increase the accuracy of the generated models. The above results were obtained from learning the dataset comprising all attributes (in a total of 182) and the following results were obtained through learning in the dataset composed with only attributes resulted from the correlation based-Feature Selection method, followed by cleaning of instances with more than 50% of missing values. Throughout the document, this last procedure is implicit whenever the feature selection was performed. The selected features for the prognosis are listed in the Appendix B as well as the classification model parameters for prognosis.

The train dataset size after feature selection, followed by the cleaning of instances with more than 50% of missing values, is described in Table 13. Figure 11 presents the train results of prognosis after applying feature selection without applying the SMOTE algorithm.

First Last approach	**noEvol**	**Evol**
Total instances (%)	171 (58.8%)	120 (41.2%)

Table 13 Size of the train dataset after applying Feature Selection using First Last approach.

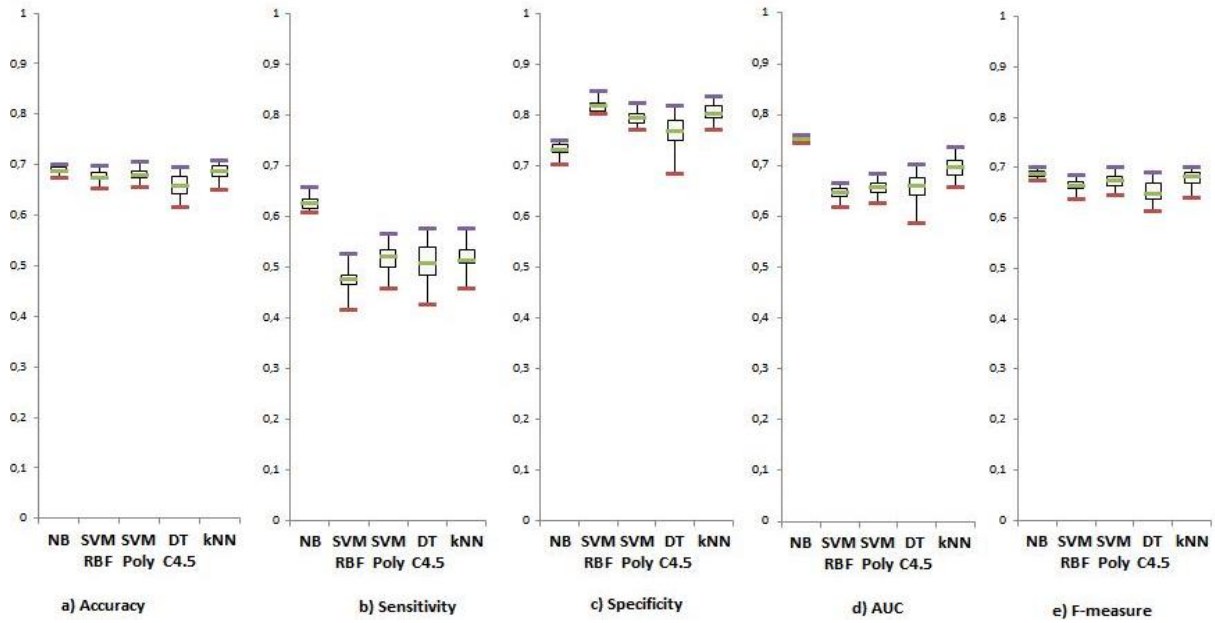

Figure 11 Train results of Prognosis after applying Feature Selection using First Last approach.

Paired t-tests were again used to compare the classifiers, using 30 repetitions. Using the correlated features, the NB is again the best model. The model C4.5DT has the worst performance. The dataset with all features is composed of 182 attributes, with 46.6% of missing values, while the reduced dataset obtained after feature selection contains 15 attributes, with only 13.1% of missing values.

Table 14 presents the confusion matrix for NB and Table 15 shows the results of this classifier.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	76 (63.9%)	44 (36.1%)
	noEvol	43 (26%)	128 (74%)

Table 14 Confusion Matrix of the NB with Feature Selection for the First Last approach, using 5-fold CV.

Classifier	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**F-measure**
NB	0.688±0.008	0.627±0.012	0.731±0.013	0.751±0.004	0.688±0.007

Table 15 Evaluation metrics of NB for the First Last approach, using 5-fold CV.

Regarding the results presented in Figure 12, it is possible to observe that the performance of learning using the reduced dataset through NB classifier is slightly lower comparing with the learning with all attributes. Since the variation of results was not statistically significant, the model with less attributes was chosen due to its simplicity, being less prone to overfitting.

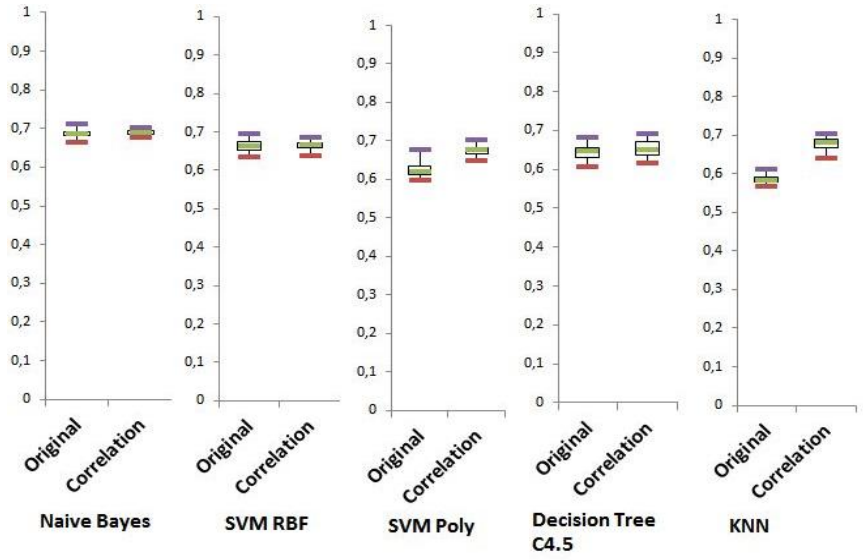

Figure 12 Train results of the F-measure metric in the two datasets using First Last approach.

This First Last approach was trained to define a baseline for the predictive power of the prognosis model. The temporal windows are expected to achieve better results [48].

4.1.2 Validation results

In the training phase, the best model was determined using CV and was then applied to the validation set, independent of the training set, whose description is found in Table 16.

First Last approach	noEvol	Evol
Total instances (%)	66 (63.5%)	38 (36.5%)

Table 16 Size of the validation dataset for First Last approach.

The Table 17 presents the confusion matrix of the NB and Table 18 shows the best results. The classifier NB has an accuracy of 0.74, a ROC Area above 0.75 and the compromise between the values of sensitivity and specificity leads to values of F-measure of 0.74.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	24 (63.2%)	14 (36.8%)
	noEvol	13 (19.7%)	53 (80.3%)

Table 17 Confusion Matrix of the NB for the First Last approach, using the validation set.

Classifier	FS	Accuracy	Sensitivity	Specificity	AUC	F-measure
NB	Correlation	0.74	0.632	0.803	0.753	0.740

Table 18 Best classifier for the First Last approach, using the validation set.

4.2 Two Years Temporal Window

The dataset characteristics relative to the two years temporal window is described in Table 19. The mean age of the evolution group is significantly higher than that of the no evolution group. The schooling years did not differ significantly between noEvol and Evol groups.

Groups 2 Years	noEvol	Evol
Total instances (%)	263 (69.5%)	115 (30.5%)
Age (Mean±SD)	69.8±7.6	72.5.0±7.8
Sex (Male/Female)	115/148 (43.7%/56.3%)	44/70 (38.3%/63.6%)
Schooling Years (Mean±SD)	9.1±5.2	8.7±5.2

Table 19 Dataset demographic after applying pre-processing for the two years temporal window.

4.2.1 Cross-Validation results

Table 20 presents the characteristics of the train dataset used in the following experiments. In the train dataset, the class distribution is 70.5% of the evaluations classified as noEvolution (noEvol) and 29.5% classified as Evolution (Evol) and thus the application of the SMOTE was used. Figure 13 shows the train results after applying SMOTE.

Two Years Temporal Window	noEvol	Evol
Total instances (%)	201 (70.5%)	84 (29.5%)

Table 20 Size of the train dataset after applying pre-processing for the two years temporal window.

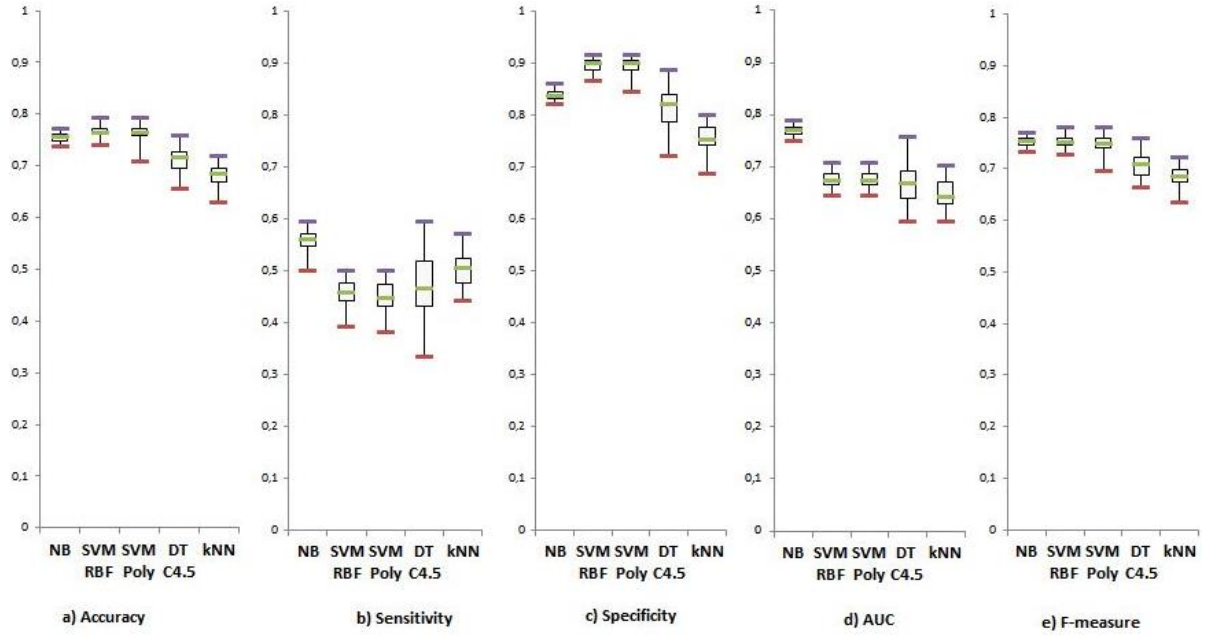

Figure 13 Train results of Prognosis after applying SMOTE using two years temporal window.

Paired t-tests were used to compare the classifiers, using 30 repetitions. The classifier NB is the best model. The classifier *k*NN presents the worst result, having no statistical significant difference with C4.5DT.

Table 21 shows the confusion matrix for NB and Table 22 displays the results of this classifier.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	46 (54.8%)	38 (45.2%)
	noEvol	34 (16.9%)	167 (83.1%)

Table 21 Confusion Matrix of the NB for the two years temporal window, using 5-fold CV.

Classifier	Accuracy	Sensitivity	Specificity	AUC	F-measure
NB	0.756±0.01	0.556±0.022	0.839±0.011	0.77±0.01	0.753±0.009

Table 22 Evaluation metrics of NB for the two years temporal window, using 5-fold CV.

The results demonstrate that the model is learning from both classes. Similarly to previous reports [1, 47, 7, 9, 4], the results in the two years temporal window are not great, probably due to the insufficient time for the patient to evolve. The predictive accuracy of the models is likely to improve with extended follow-up time.

The train dataset size relative to two years temporal window after the feature selection is described in Table 23. Figure 14 presents the train results of prognosis after applying feature selection. The dataset with all features is composed by 182 attributes, with 52.2% of missing values, whereas the reduced dataset obtained after feature selection contains 22 attributes, with only 13% of missings.

Two Years Temporal Window	noEvol	Evol
Total instances (%)	161 (68.8%)	73 (31.2%)

Table 23 Size of the train dataset after applying Feature Selection using two years temporal window.

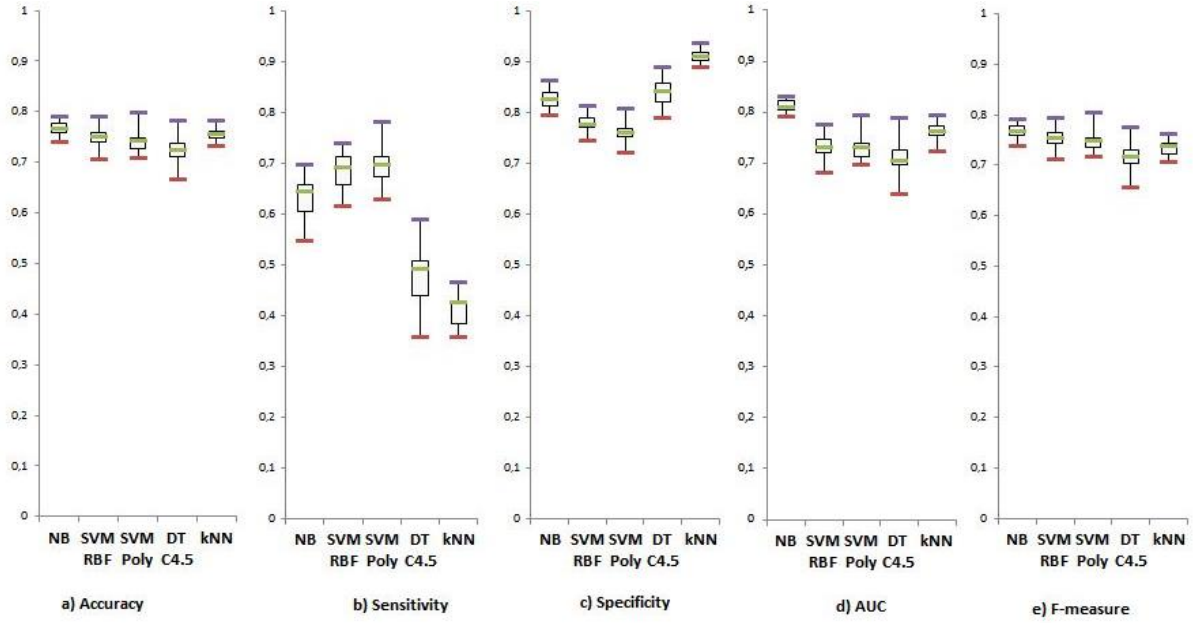

Figure 14 Train results of Prognosis after applying Feature Selection using two years temporal window.

Paired t-tests were used to compare the classifiers, using 30 repetitions. NB is the best model and the C4.5DT model got the worst results.

As we can notice in Figure 15, when the classifiers learn using the selected features have higher predictive capability.

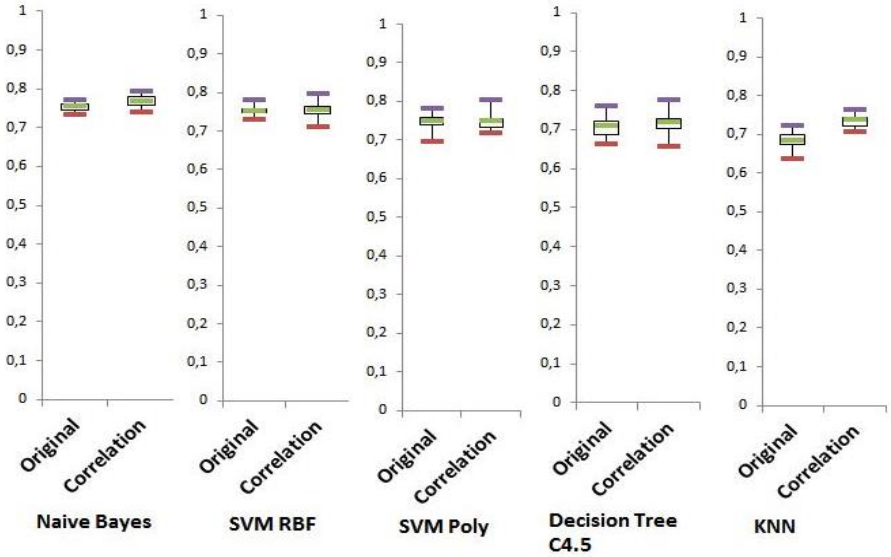

Figure 15 Train results of the F-measure metric in the two datasets using two years temporal window.

Table 24 shows the confusion matrix for NB and Table 25 displays the results of this classifier.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	45 (61.6%)	28 (38.4%)
	noEvol	27 (16.8%)	134 (83.2%)

Table 24 Confusion Matrix of NB with Feature Selection for two years temporal window, using 5-fold CV.

Classifier	Accuracy	Sensitivity	Specificity	AUC	F-measure
NB	0.767±0.014	0.635±0.035	0.827±0.018	0.811±0.012	0.768±0.009

Table 25 Evaluation metrics of NB for the two years temporal window, using 5-fold CV.

4.2.2 Validation results

The model NB trained in the reduced dataset was applied to the validation set described in Table 26.

Two Years Temporal Window	noEvol	Evol
Total instances (%)	61 (66.3%)	31 (33.7%)

Table 26 Size of the validation dataset using two years temporal window.

The Table 27 displays the confusion matrix of the NB classifier and Table 28 displays the best results considering other metrics. NB has an F-measure slightly above 0.65 and accuracy not higher than 70%.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	16 (51.6%)	15 (48.4%)
	noEvol	15 (24.6%)	46 (75.4%)

Table 27 Confusion Matrix of the NB for two years temporal window, using the validation set.

Classifier	FS	Accuracy	Sensitivity	Specificity	AUC	F-measure
NB	Correlation	0.674	0.516	0.754	0.744	0.674

Table 28 Best classifier for two years temporal window, using the validation set.

Previous works [1, 47] aimed to predict the conversion of a MCI patient to AD in a 2 or 2.5 years period, reporting disappointing results. To improve the results, they claim that it is necessary to increase the temporal window analysis, so an enhancement in the results is expected in the following temporal windows.

4.3 Three Years Temporal Window

The dataset characteristics relative to the three years temporal window is describe in Table 29. The mean age of the evolution group is significantly higher than the no evolution group and the schooling years are similar in both groups.

Groups 3 Years	noEvol	Evol
Total instances (%)	180 (53.7%)	156 (46.4%)
Age (Mean± SD)	69.7±7.5	72.6±7.9
Sex (Male/Female)	84/96 (46.7%/53.3%)	56/99 (35.9%/63.5%)
Schooling Years (Mean± SD)	9.3±5.3	8.5±4.7

Table 29 Dataset demographic after applying pre-processing for the three years temporal window.

4.3.1 Cross-Validation results

In the next sections only the results considering feature selection will be presented. The analyses without feature selection may be consulted in Appendix B.

The train dataset size relative to three years temporal window after the feature selection is described in Table 30. The SMOTE algorithm was not applied due the similar proportions of classes in this dataset.

Three Years Temporal Window	noEvol	Evol
Total instances (%)	112 (53.1%)	99 (46.9%)

Table 30 Size of the train dataset after applying Feature Selection using three years temporal window.

Figure 16 presents the train results after applying feature selection. All classifiers performed with high generalization with values of F-measure close to 0.8 using the correlated features. The three years temporal window dataset is the most balanced resulting in a significant improvement in all assessment metrics.

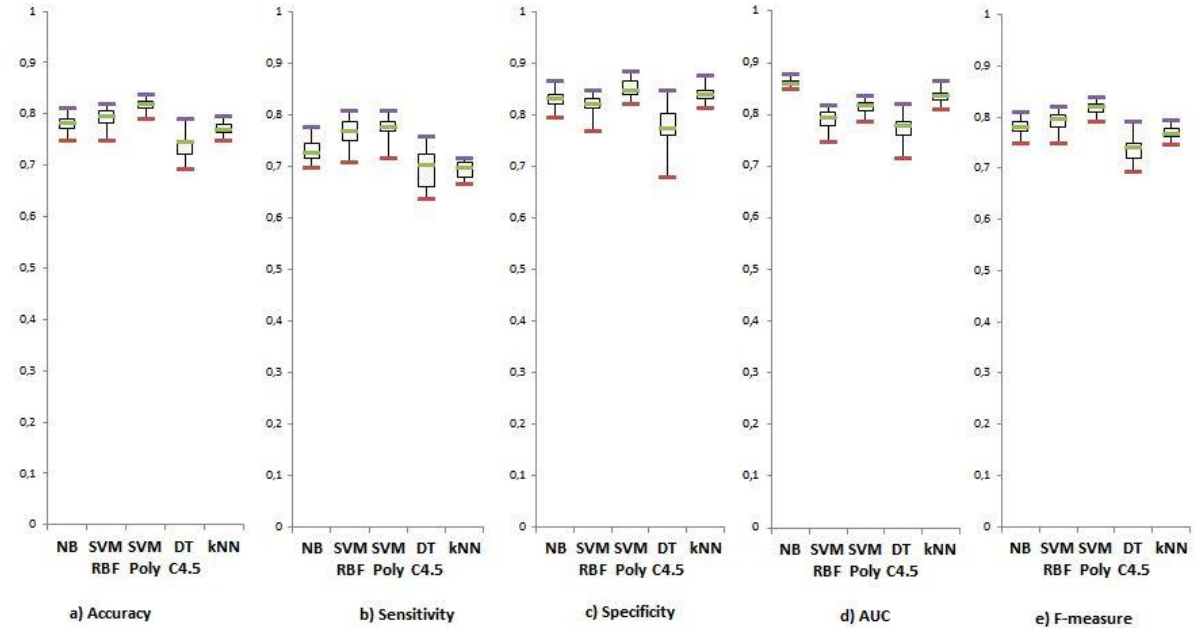

Figure 16 Train results of Prognosis after applying Feature Selection using three years temporal window.

The dataset with all features has 182 attributes, with 52.2% of missing values, whereas the reduced dataset obtained after feature selection contains 21 attributes, with only 9.7% of missings.

Paired t-tests were used to compare the classifiers, using 30 repetitions. SVM Poly is the best model in the reduced dataset and the C4.5DT got the worst results.

Table 31 shows the confusion matrix for SVM Poly and Table 32 displays the results of this classifier.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	77 (77.8%)	22 (22.2%)
	noEvol	19 (16.9%)	93 (83.1%)

Table 31 Confusion Matrix of the SVM Poly with Feature Selection for three years temporal window, using 5-fold CV.

Classifier	Accuracy	Sensitivity	Specificity	AUC	F-measure
SVM Poly	0.817±0.012	0.778±0.020	0.853±0.016	0.815±0.012	0.817±0.012

Table 32 Evaluation metrics of SVM Poly for the three years temporal window, using 5-fold CV.

Figure 17 displays the F-measure values of all classifiers in the two datasets analysed. Again, we note that the choice of feature selection has a great effect on the final outcome.

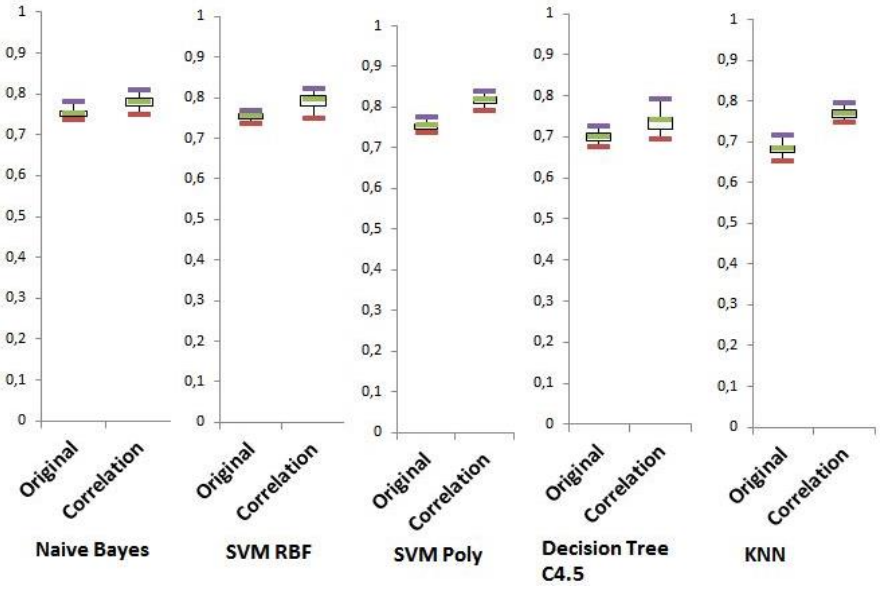

Figure 17 Train results of the F-measure metric in the two datasets using three years temporal window.

The previous results make us believe that most MCI patients that are going to evolve do so in this time frame. The trade-off between sensitivity and specificity is preferable than in the previous analyses, reflected by the higher values of F-measure.

4.3.2 Validation results

The performance of the best classifier applied to the validation set, described in Table 33, can be found in Table 35. Table 34 displays the confusion matrix of the SVM Poly.

The best classifier achieved an AUC near to 0.7 and performed with values of F-measure close to 0.7 and accuracy of 69.1%. Thus, in the three years temporal window, we obtained an overall good performance in the validation set.

Three Years Temporal Window	noEvol	Evol
Total instances (%)	40 (49.4%)	41 (50.6%)

Table 33 Size of the validation dataset using three years temporal window.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	27 (65.9%)	14 (34.1%)
	noEvol	11 (27.5%)	29 (72.5)

Table 34 Confusion Matrix of the SVM Poly for three years temporal window, using the validation set.

Classifier	FS	Accuracy	Sensitivity	Specificity	AUC	F-measure
SVM Poly	Correlation	0.691	0.659	0.725	0.692	0.691

Table 35 Best classifier for three years temporal window, using the validation set.

4.4 Four Years Temporal Window

The dataset characteristics relative to the four years temporal window is describe in Table 36. The mean age of the evolution group is significantly higher than the no evolution group. The schooling years are comparable in both groups.

Groups 4 Years	noEvol	Evol
Total instances (%)	109 (36.3%)	191 (63.7%)
Age (Mean± SD)	69±5.2	72±4.6
Sex (Male/Female)	55/54 (50.5%/49.5%)	71/119 (37.1%/62.3%)
Schooling Years (Mean± SD)	8.8±5.2	8.6±4.6

Table 36 Dataset demographic after applying pre-processing for the four years temporal window.

4.4.1 Cross-Validation results

The train dataset size relative to four years temporal window after the feature selection, followed with the cleaning of instances with more than 50% of missing values, is described in Table 37. In this dataset, the difference in the proportion of classes is again not that significant, so the SMOTE algorithm was not used. In Figure 18 is presented the train results after applying feature selection. The dataset with all features has 182 attributes with 54.1% of missing values, while the reduced dataset obtained after feature selection has 20, with 9.8% of missings.

Four Years Temporal Window	noEvol	Evol
Total instances (%)	66 (35.7%)	119 (64.3%)

Table 37 Size of the train dataset after applying Feature Selection using four years temporal window.

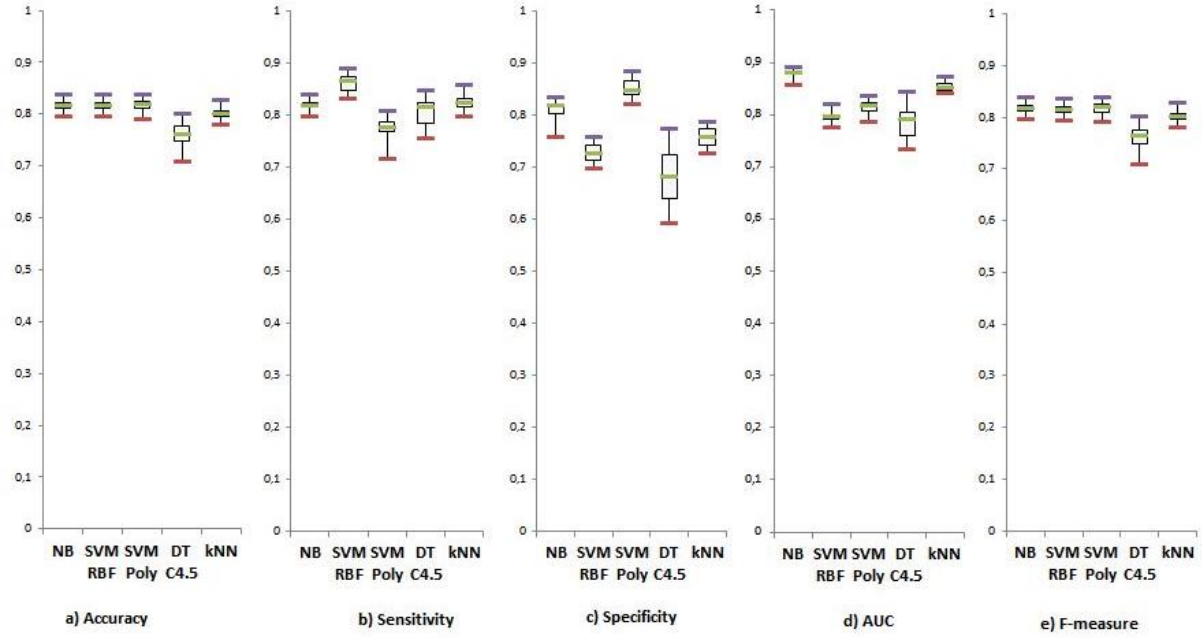

Figure 18 Train results of Prognosis after applying Feature Selection using four years temporal window.

Paired t-tests were used to compare the classifiers, using 30 repetitions. NB is the best model, with no statistical difference with SVM Poly and SVM RBF and the C4.5DT model got the worst results.

Table 38 shows the confusion matrix for NB and Table 39 displays the results of this classifier.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	98 (82.4%)	21 (17.6%)
	noEvol	12 (18.2%)	54 (81.8%)

Table 38 Confusion Matrix of the NB with Feature Selection for four years temporal window, using 5-fold CV.

Classifier	Accuracy	Sensitivity	Specificity	AUC	F-measure
NB	0.816±0.010	0.818±0.01	0.812±0.018	0.879±0.007	0.818±0.007

Table 39 Evaluation metrics of NB for the four years temporal window, using 5-fold CV.

Figure 19 displays the F-measure values of all classifiers in the two datasets analysed. As abovementioned, the increase of the temporal window leads to better results in the cross-validation without the application of feature selection. The use of correlated features even improves such good results.

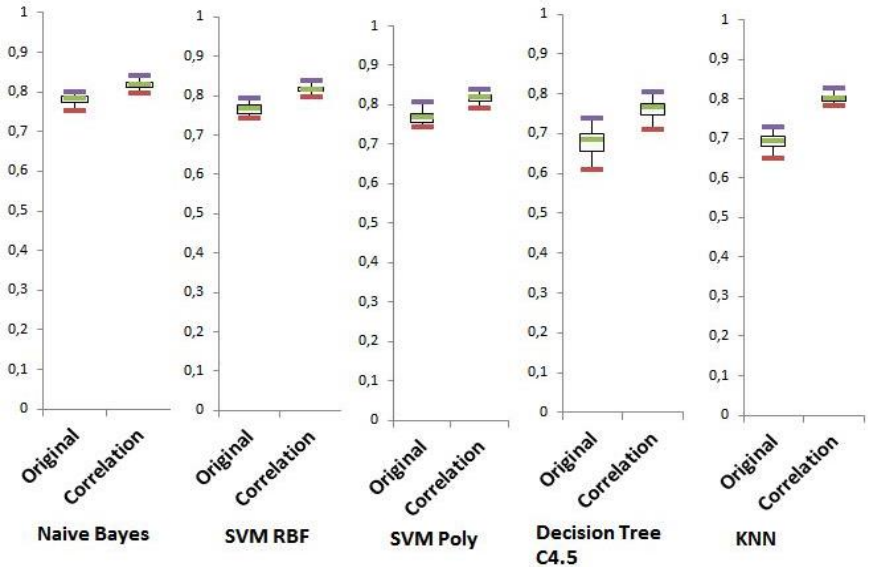

Figure 19 Train results of the F-measure metric in the two datasets using four years temporal window.

4.4.2 Validation results

Follows the second phase of the classification, in which the best classifier obtained during the training phase is applied to the validation set, described in Table 40. The Table 41 displays the confusion matrix of the NB classifier and Table 42 displays the best results considering other metrics. Using NB, the F-measure is higher than 0.7. The specificity has low values in contrast with the high values of sensitivity, probably due to the lower number of examples of noEvol.

Four Years Temporal Window	**noEvol**	**Evol**
Total instances (%)	24 (33.3%)	48 (66.7%)

Table 40 Size of the validation dataset using four years temporal.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	40 (83.3%)	8 (16.7%)
	noEvol	12 (50%)	12 (50%)

Table 41 Confusion Matrix of the NB for four years temporal window, using the validation set.

Classifier	**FS**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**F-measure**
NB	Correlation	0.722	0.833	0.5	0.825	0.715

Table 42 Best classifiers for four years temporal window, using the validation set.

4.5 Five Years Temporal Window

The dataset characteristics relative to the five years temporal window is describe in Table 43. The mean age of the evolution group is significantly higher than the no evolution group. The distributions of the schooling years are equivalent in both groups.

Groups 5 Years	noEvol	Evol
Total instances (%)	67 (23.8%)	215 (76.2%)
Age (Mean± SD)	68.1±6.7	72.3±7.7
Sex (Male/Female)	30/37 (44.8%/55.2%)	78/136 (36.3%/63.3%)
Schooling Years (Mean± SD)	8.8±5	8.5±4.6

Table 43 Dataset demographic after applying pre-processing for the five years temporal window.

4.5.1 Cross-Validation Results

The present study allowed the analysis of the five years temporal window, which is estimated to improve the previous results, since it was reported that MCI patients that will evolve, evolve in the five years time frame.

The train dataset size relative to the five years temporal window after the feature selection is described in Table 44. Figure 20 presents the train results after applying feature selection. The dataset with all features has 182 attributes, with 53.6% of missing values, while the reduced dataset obtained after feature selection has 21, with only 11.4 % of missings.

Five Years Temporal Window	noEvol	Evol
Size	42 (23.3%)	138 (76.7%)

Table 44 Size of the train dataset applying Feature Selection using five years temporal window.

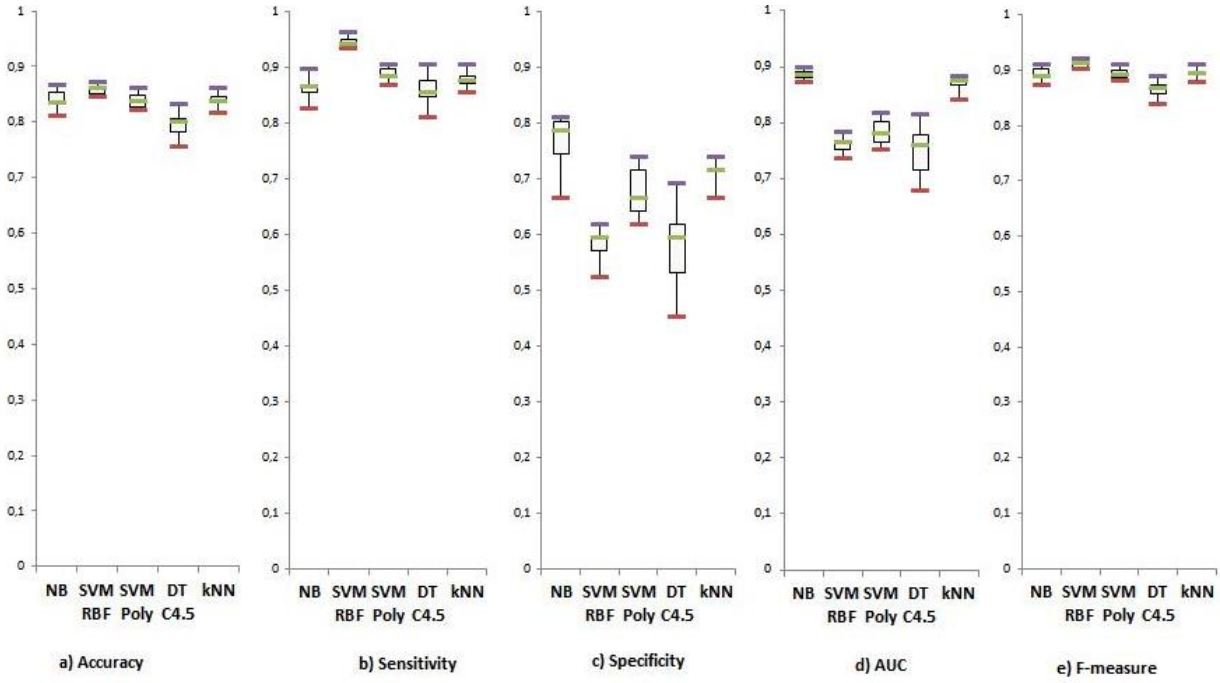

Figure 20 Train results of Prognosis after applying Feature Selection using five years temporal window.

Paired t-tests were used to compare the classifiers, using 30 repetitions. SVM RBF model is the best model with no statistical difference with NB and the C4.5DT model got the worst results.

Table 45 shows the confusion matrix for SVM RBF and Table 46 displays the results of this classifier. The values of sensitivity are very high in opposition to the values of specificity, justified by the dissimilar proportions of classes. Although these values are dissimilar, the model performed with high values of F-measure demonstrating the good predictive capability of the models. Conversion into dementia is the key prediction in this biomedical application requiring classifiers with high sensitivity as was obtained in the four and five temporal window.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	125 (90.6%)	13 (9.4%)
	noEvol	15 (35.7%)	27 (64.3%)

Table 45 Confusion Matrix of the SVM RBF with Feature Selection for five years temporal window, using 5-fold CV.

Classifier	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**F-measure**
SVM RBF	0.846±0.009	0.893±0.01	0.690±0.022	0.792±0.012	0.847±0.012

Table 46 Evaluation metrics of SVM RBF for the five years temporal window, using 5-fold CV.

Figure 21 displays the F-measure values of all classifiers in the two datasets analysed. The best result of F-measure using SVM RBF with the correlated features is near 0.9. The best values of F-measure in the cross-validation results were obtained in the five years temporal window.

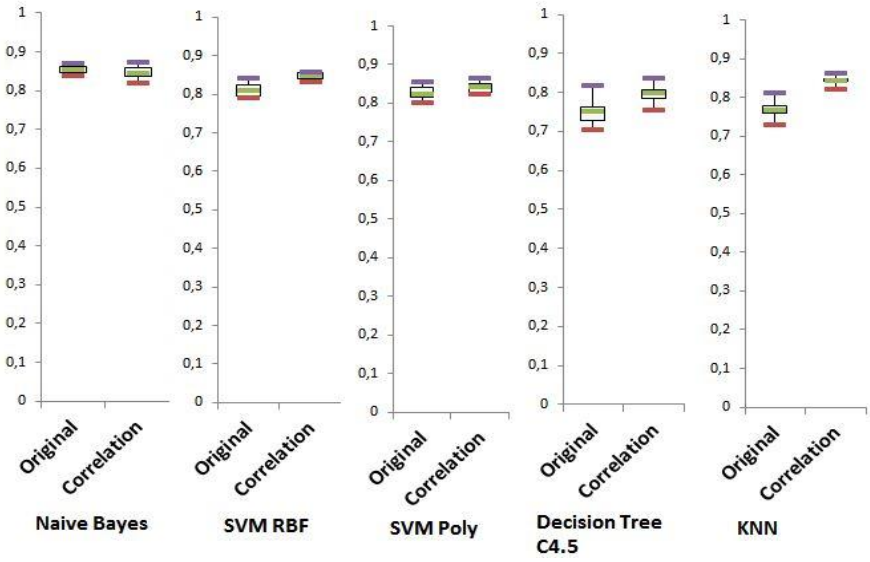

Figure 21 Train results of the F-measure metric in the two datasets using five years temporal window.

4.5.2 Validation Results

The best classifiers obtained during the training phase were applied to the validation set, described in Table 47. The Table 48 displays the confusion matrix of the SVM RBF classifier and Table 49 displays the best results considering other metrics.

Five Years Temporal Window	noEvol	Evol
Total instances (%)	16 (23.9%)	51 (76.1%)

Table 47 Size of the validation dataset using five years temporal.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	46 (95.8%)	5 (4.2%)
	noEvol	8 (50%)	8 (50%)

Table 48 Confusion Matrix of SVM RBF for five years temporal window, using the validation set.

Classifier	FS	Accuracy	Sensitivity	Specificity	ROC Area	F-measure
SVM RBF	Correlation	0.806	0.902	0.5	0.701	0.799

Table 49 Best classifiers for five years temporal window, using the validation set.

The SVM RBF was the best classifier in the cross-validation, performing with accuracy higher than 0.8, which was the best value comparing with the other temporal windows. Nonetheless, the total accuracy of the classifier is misleading since the model is good only at predicting the positive class (Evol class), with high sensitivity, but quite bad at predicting the noEvol class, with low specificity. The F-measure value is close to 0.8, having in account the values of sensitivity of 0.902 and the specificity of 0.5. As such, the results obtained in five temporal window should be interpreted with caution due to the relatively low number of patients used as examples and it would be relevant to replicate this study in a larger dataset with similar follow-ups.

4.6 Discussion

The analyses previously presented in each temporal window corroborated the use of SMOTE algorithm when the proportion of the classes is imbalanced. As such, in what follows, this algorithm will only be applied when the proportion of classes justified it.

In general, the original dataset containing the entire group of features produces lowest values in the evaluation metrics used. This can be explained by the fact that the original dataset includes a high number of features and some of them can even add noise to the learning process. The application of feature selection improved the accuracy of every classifier in each temporal window.

The results show that the temporal window approach mitigates the problem of different follow-up time associated with the First Last approach leading to better prognosis results. With the increase of the temporal window, it is noticed a higher prediction capability of the models. It is important to refer that the reduction of the

cardinality of instances along the temporal windows influence the way that the models learn, leading to opposite values of sensitivity and specificity, even though it was applied the SMOTE algorithm when necessary.

The 3, 4 and 5 years temporal window presented high values of F-measure. Such results corroborate the need to increase the time analysis of a patient. Regarding the results from these temporal windows, it would be important to determine if the increase in the F-measure values along these windows is statistically significant. Based on that, it would be essential to perform a statistical test, such as Friedman test [46], in order to compare the algorithms on the different datasets.

Several studies [1, 47] use a relatively short interval of two years follow-up to predict the conversion of MCI patient to AD obtaining comparable results to the ones we get using the same temporal window. The available data allowed the learning of models up to five years temporal window, not done in the work of Lemos *et al.* [48]. Between two and three years temporal window, there was a noticeable higher predictability of conversion to AD and in this regard we hypothesize that most patients that are going to evolve do so in the three years temporal window.

Lemos *et al* [48] reported that the three years temporal window is the one with higher discriminative power in concordance with the medical feedback. The results obtained in the three and four years temporal window are more reliable, being more suitable to predict conversion with higher predictive capability. Even though the good results achieved in the five temporal window with the highest values of accuracy and F-measure, the reduced number of examples lead to the lowest values of specificity. The prognosis in the five years temporal window allowed us to understand the importance of larger follow-up and that it would be important to replicate the study in a larger database to corroborate the results obtained. Wider time interval strategies seem to provide a good opportunity to monitor prognosis prediction of conversion of MCI patients to AD.

The integration of more patients to the database will allow a more accurate learning for both classes of patients (as in the case of the three years temporal window, presenting a comparable proportion of noEvol and Evol examples).

Table 50 presents a summary of the performance of the best models in each time window in the cross-validation.

	Classifier	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**F-measure**
First Last	NB	0.688±0.008	0.627±0.012	0.731±0.013	0.751±0.004	0.688±0.007
2 Years	NB	0.767±0.014	0.635±0.035	0.827±0.018	0.811±0.012	0.768±0.009
3 Years	SVM Poly	0.817±0.012	0.778±0.020	0.853±0.016	0.815±0.012	0.817±0.012
4 Years	NB	0.816±0.010	0.818±0.01	0.812±0.018	0.879±0.007	0.818±0.007
5 Years	SVM RBF	0.846±0.009	0.893±0.01	0.690±0.022	0.792±0.012	0.847±0.012

Table 50 Best models to each progression approach during the training phase.

As expected, the performance of all classifiers in validation set data (Table 51) decline in comparison with cross-validation.

	Classifier	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**F-measure**
First Last	NB	0.74	0.632	0.803	0.753	0.740
2 Years	NB	0.674	0.516	0.754	0.744	0.674
3 Years	SVM Poly	0.691	0.659	0.725	0.692	0.691
4 Years	NB	0.722	0.833	0.5	0.825	0.715
5 Years	SVM RBF	0.806	0.902	0.5	0.701	0.799

Table 51 Best models to each progression approach in the validation set.

Even though the cross-validation results confirmed the importance of the temporal window approach, in the validation set the results in the two, three and four temporal window did not outperform the First Last Approach. Such results are most likely due to the fact that the First Last Approach is trained with more examples than any temporal window approaches. However, the increase along the temporal window in the values of accuracy and F-measure is maintained and the proportions of sensitivity and specificity are also in agreement with what we observed in the cross-validation results.

It was expected that the three temporal window results outperformed the results when compared with the other temporal windows. Nevertheless, the validation results did not contradict the conclusions taken when analysing the cross-validation results. It is important to note the results of the five temporal window evaluation, which performed with the highest F-measure value with 90.2% of sensitivity, learning almost perfectly from the Evol class and with 50% of specificity, which substantiates the importance of the integration of more noEvol examples.

Another important issue to discuss is the features selected by the feature selection method that are common along the temporal window. The attributes age, Orient_T, or_Total, VerbalPaired_AssociateLearning_Z, LogicalMemory_A_Z, PA_Tot, LM_a_Interf, SemanticFluency_Z, Orientation_Z, Cube_Z and MPR_Z were chosen in all temporal windows having an important role in the classification of the patients independently of the temporal window under study. Such attributes are related to the orientation to time and place, memory evaluation, praxis and verbal fluency.

In order to understand the importance of such features, two different analyses were performed in all time windows. The first one considered only the features in common, displayed in Table 52, and the second one considered the features characteristic of each temporal window. Since the results did not outperform the ones obtained using the features selected in each temporal window, it is possible to report that neither the features in common nor the features aside the ones in common have sufficient discriminative power over all features, which assigns value to these features.

Features
Age
Orient_T
Or_Total
PA_Tot
LM_a_Interf
VerbalPaired_AssociateLearning_Z
LogicalMemory_A_Z
SemanticFluency_Z
Orientation_Z
Cube_Z
MPR_Z

Table 52 Features used in all temporal windows.

5. Predicting conversion from MCI to AD based on different MCI characteristics

Several studies [26, 28, 55] suggested that MCI is a complex clinical concept characterized by complex cognitive manifestations with a heterogeneous pattern of transitions, which can involve persistence of the symptoms, conversion to dementia or state improvement. The knowledge acquired so far that the separation of patients with different characteristics would improve the classification of the MCI either motivated us to develop a new approach predicting models trained with different MCI groups, based on clinical criteria or results of unsupervised learning methods [16, 13, 19, 28].

Despite the good performance of classifiers in the preceding analyses, the purpose of the following studies is thus to create an approach to classify patients by taking into account the putative differences between MCI patients. Ideally, this approach should outperform the classification approach where all MCI are assumed to evolve similarly.

5.1 Prognosis prediction based on clinical criteria: depressed/ not depressed

Previously in this work, we described the attribute that measures the state of depression of a patient named as GDS, Geriatric Depression Scale [16, 19, 28]. With the purpose of studying the influence of the depressive symptoms in classification, a new analysis was performed based on the GDS attribute. For this study, a short form of the self-report instrument was used, comprising score values between 0 and 14. It is expected that the separation of patients concerning their state of depression leads to a better learning of the models, since we believe that patients categorized as depressed have cognitive capabilities dissimilar from patients categorized as not depressed.

5.1.1 Clustering

According to clinical information, the patients are categorized as not depressed patients if their score in the GDS test is lower or equal to 4 and as depressed if is higher than 4. Table 53 shows the composition of the datasets regarding the GDS attribute in each temporal window. It is noticeable the occurrence of more than 50% of missing values. Because of that those instances were not taken in consideration.

Dataset	GDS score 0-4	GDS score 5-14	GDS value missing	# instances
First Last	72 (23.8%)	69 (22.8%)	162 (53.5%)	303
2Y	72 (25.3%)	63 (22.1%)	150 (52.6%)	285
3Y	58 (22.8%)	56 (22.1%)	140 (55.1%)	254
4Y	52 (22.8%)	48 (21.3%)	128 (56.1%)	228
5Y	46 (36.7%)	45 (5.6%)	124 (57.7%)	215

Table 53 Composition of instances considering the GDS attribute after preprocessing.

5.1.2 Classification

In the training phase two approaches were analysed in order to give answer to the question "Given that we know the state of depression of a patient, is it better to predict the conversion to AD using the general model containing all patients, or using a specific model considering their state?". In other words, the following analyses aim to determine if there is a higher capability to predict conversion based on the characteristics of the patients instead of assuming the same profile in the entire set of patients. The datasets used to create these models were separated in three, concerning the score of the attribute GDS. One considers only the instances with this attribute defined ($\mathfrak{D}_{all}$), other includes only instances with values from 0 to 4 corresponding to the not depressed patients (denominated as $\mathfrak{D}_{0-4}$) and another comprises instances with values from 5 to 14 corresponding to the depressed patients (denominated as $\mathfrak{D}_{5-14}$). The workflow of the two approaches is presented in Figure 22.

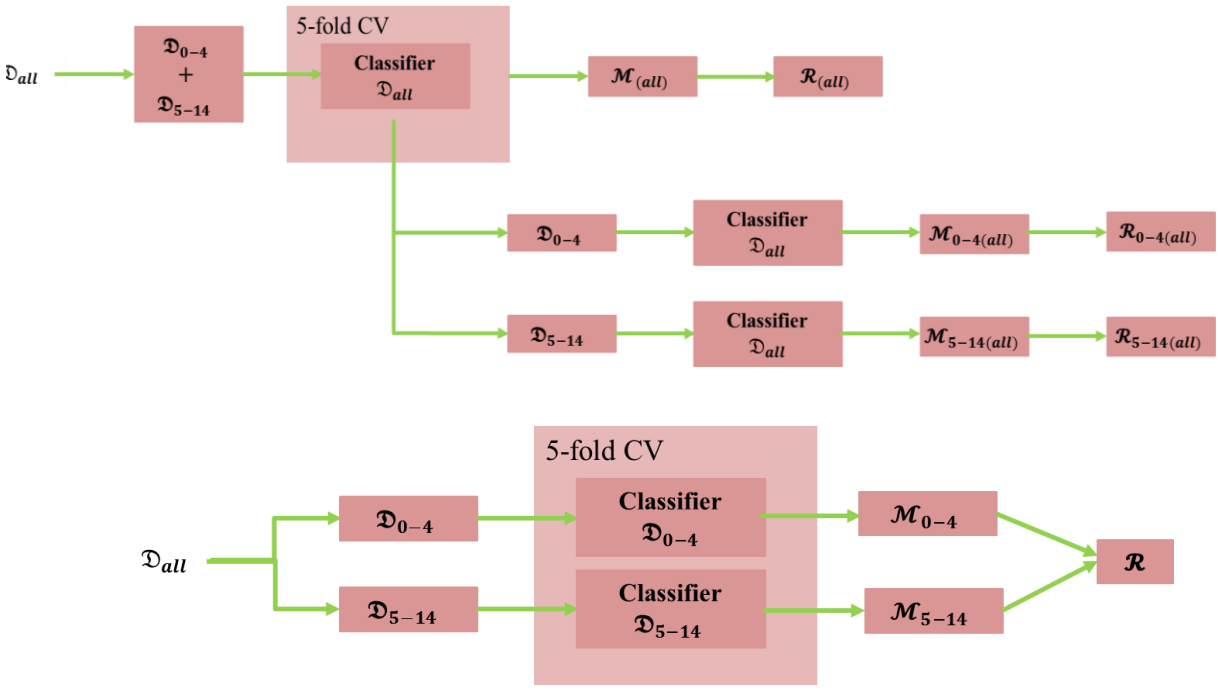

Figure 22 Workflow of the two approaches performed during the training phase.

As illustrated above, the first approach consists on training the models from the entire dataset $\mathfrak{D}_{all}$, to understand how these general models classify the patients according to their state of depression. The best classifier under the cross-validation evaluation is then applied to the restrictive datasets ($\mathfrak{D}_{0-4}$ and $\mathfrak{D}_{5-14}$) and

the respective result is generated for each group of patients. In the second approach, the dataset $\mathfrak{D}_{all}$ is divided in the two restrictive datasets and for each one of them is determined the best classifier under the cross-validation evaluation and from the particular models obtained from each group of patients is determined the output result.

The datasets were analysed after determining the set of attributes that allows the achievement of more accurate classifiers. The features selected can be consulted in Appendix C. In the succeeding analyses, the elimination of instances with more than 50% of missing values after the feature selection was not performed, since such analyses are based on comparisons between datasets, where it is important to keep the same number of instances, and the removals are minor.

5.1.2.1 Two Years Temporal Window

Cross-Validation results

The sizes of the train datasets for the two years temporal window are described in Table 54.

Two Years Temporal Window	noEvol	Evol
$\mathfrak{D}_{all}$	91 (67.4%)	44 (32.6%)
$\mathfrak{D}_{0-4}$	55 (76.4%)	17 (23.6%)
$\mathfrak{D}_{5-14}$	36 (57.1%)	27 (42.9%)

Table 54 Sizes of the train datasets which are differentiated based on GDS values, using two years temporal window.

In the dataset $\mathfrak{D}_{0-4}$, from the set of attributes elected in feature selection it is worth noting the attributes MPR, CVLT_rec_P, CancellationTask_Z, GeneralInformation_Z, MPR_Z in common with the attributes selected in the $\mathfrak{D}_{all}$ dataset. In the same way, the set of attributes LM_a_interf, Or_Total, Cube, Orientation_Z, A1_Z and Atot_Z were selected in the entire dataset and in the dataset $\mathfrak{D}_{5-14}$. Despite the attributes in common in the datasets, there are still attributes specific for each one of them. The list of attributes chosen may be consulted in Appendix C. Figures 23, 24 and 25 show the train results in the three distinct datasets.

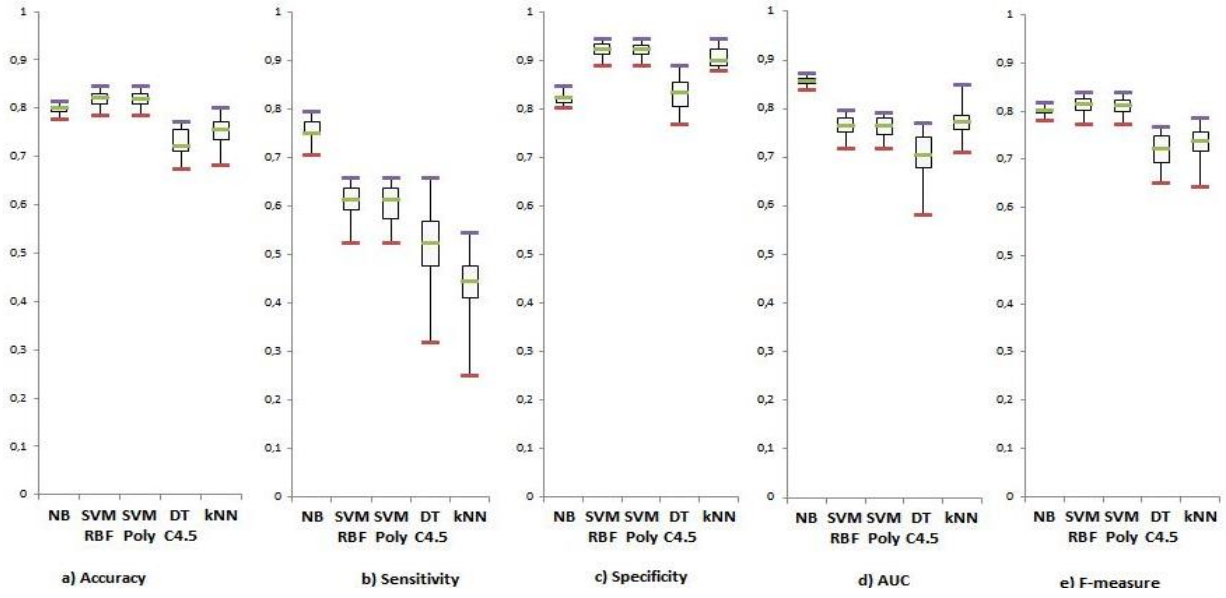

Figure 23 Train results of Prognosis of the dataset $\mathfrak{D}_{all}$, using two years temporal window.

All the evaluation metrics have higher values in the dataset with only patients with the GDS variable assigned, comparing with the dataset containing all patients analysed in the previous chapter. Thus, it would be important to update the database filling this attribute, in order to improve the prognosis evaluation.

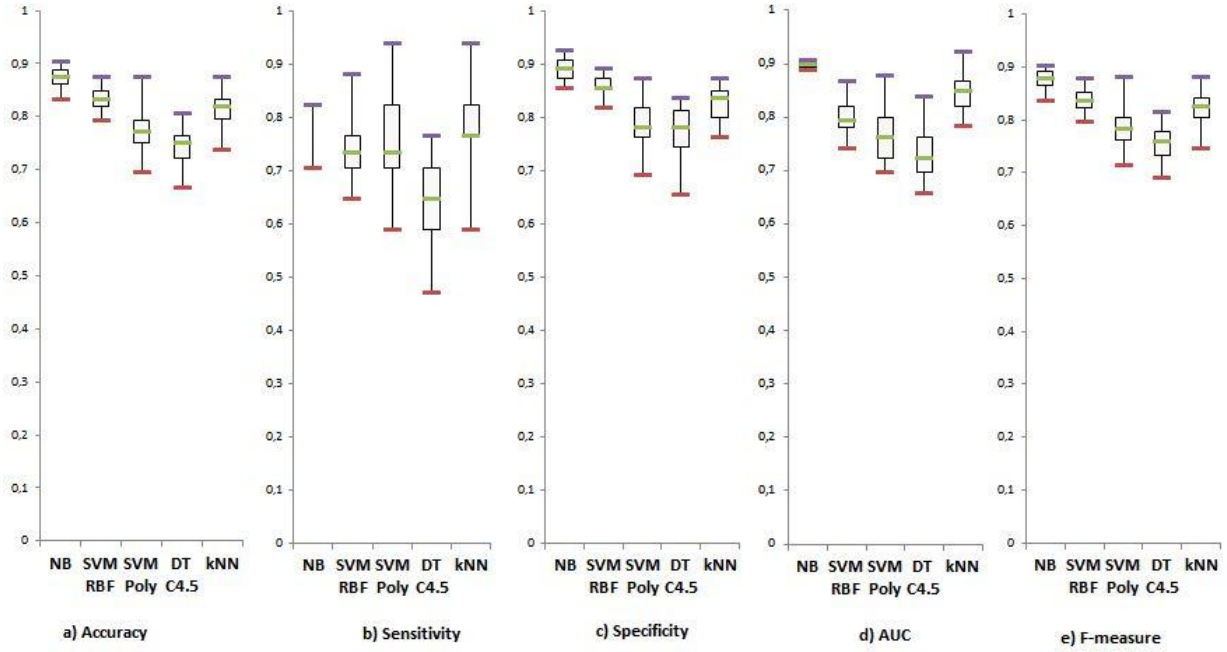

Figure 24 Train results of Prognosis of the dataset $\mathfrak{D}_{\mathbf{0-4}}$, using two years temporal window.

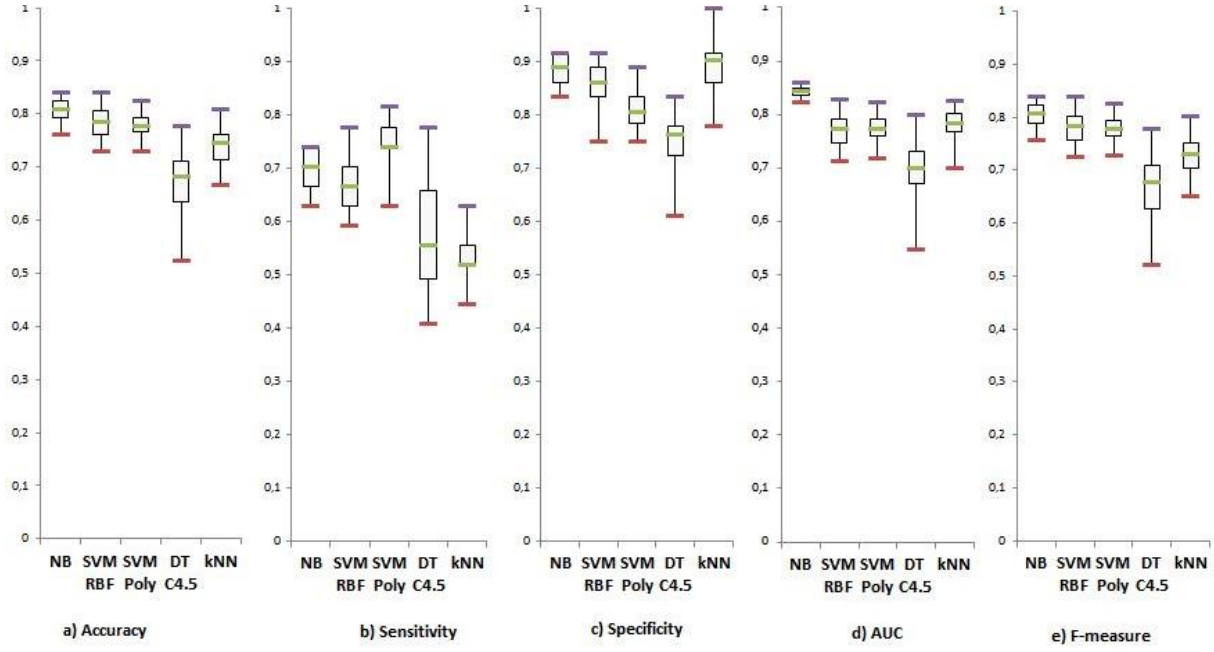

Figure 25 Train results of Prognosis of the dataset $\mathfrak{D}_{\mathbf{5-14}}$, using two years temporal window.

Paired t-tests were used to compare the classifiers, using 30 repetitions. The t-test is only applied if the ANOVA test with 95% confidence level confirms the existence of a significant difference. In $\mathfrak{D}_{all}$, SVM RBF is the best model, with no statistical difference with the SVM Poly, and the C4.5DT model got the worst results, with no statistically difference with *k*NN. In $\mathfrak{D}_{0-4}$, the classifier NB is the best model and the C4.5DT model got

the worst results, presenting both a significant difference among the other models. In $\mathfrak{D}_{5-14}$, the classifier NB the best model and the classifier C4.5DT presents the worst result.

Before determine how the specific models behave, it is important to understand how the general model classifies the patients according to their state of depression. It is important to consider the confusion matrices (Table 55) since it is relevant to know if the performance of the general model is related to the conditions of depression of the patients. We can observe that the model has a similar predictive capability when considers depressed or not depressed patients. Thus, its performance is class independent.

a)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	28 (63.6%)	16 (36.4%)
	noEvol	8 (8.8%)	83 (91.2%)

b)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	11(64.7%)	6 (35.3%)
	noEvol	5 (9.1%)	50(90.9%)

c)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	17 (63%)	10 (37%)
	noEvol	3 (8.3%)	33 (91.7%)

Table 55 Confusion Matrix of SVM RBF for two years temporal window, using 5-fold CV considering the $\mathfrak{D}_{all}$: a) considering all instances; b) considering only instances with GDS≤4; c) considering only instances with GDS>4.

Table 56 presents the evaluation metrics for the general model $\mathcal{M}_{(all)}$ and for each specific model $\mathcal{M}_{0-4(all)}$ and $\mathcal{M}_{5-14(all)}$, concerning the state of depression trained with this model. It is noticeable that the general model $\mathcal{M}_{(all)}$ has a good predictive capability independent of the score of the GDS of the patient. The values of sensitivity are low due to the reduced number of Evol examples trained, as reported in Table 54.

Model	**Classifier**	**Accuracy**	**Sensitivity**	**Specificity**	**F-measure**
$\mathcal{M}_{(all)}$	SVM RBF	0.822	0.636	0.912	0.817
$\mathcal{M}_{0-4(all)}$		0.847	0.647	0.909	0.846
$\mathcal{M}_{5-14(all)}$		0.794	0.63	0.917	0.787

Table 56 Evaluation metrics of SVM RBF model for two years temporal window, using 5-fold CV.

The second approach of the analysis, allows us to compare the performance of each specific model considering different characteristics of the patients with the performance of the general model relatively to these groups of patients. Table 57 shows the confusion matrices resulting from this approach. Table 58 presents the evaluation metrics for each specific model $\mathcal{M}_{0-4(all)}$ and $\mathcal{M}_{5-14(all)}$ and for the resulting model.

a)

Real Class	Predicted Class: Evol	Predicted Class: noEvol
Evol	14 (82.4%)	3 (17.6%)
noEvol	6 (10.9%)	49 (89.1%)

b)

Real Class	Predicted Class: Evol	Predicted Class: noEvol
Evol	18 (66.7%)	9 (33.3%)
noEvol	4 (11.1%)	32 (88.9%)

c)

Real Class	Predicted Class: Evol	Predicted Class: noEvol
Evol	32 (72.7%)	12 (27.3%)
noEvol	10 (11%)	81 (89%)

Table 57 Confusion Matrix for two years temporal window, using 5-fold CV considering a) NB in $\mathfrak{D}_{0-4}$; b) NB in $\mathfrak{D}_{5-14}$; c) sum of the confusion matrices of each model.

Model	Classifier	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{0-4}$	NB	0.875	0.824	0.891	0.878
$\mathcal{M}_{5-14}$	NB	0.793	0.667	0.889	0.789
$\mathcal{M}$		0.837	0.727	0.890	0.836

Table 58 Evaluation metrics of each model for two years temporal window, using 5-fold CV.

Looking to the values of F-measure and accuracy (Tables 56 and 58) there is no significant difference between the two approaches. The model $\mathcal{M}_{all}$ that is learning without considering the state of depression has similar behaviour predicting the different characteristics of the patients as the models $\mathcal{M}_{0-4}$ and $\mathcal{M}_{5-14}$ that are learning having in account this condition. However it is important to refer that the sensitivity correspondent to the $\mathcal{M}_{0-4}$ is higher than the one obtained by $\mathcal{M}_{0-4(all)}$, which represents the better capability of the specific model $\mathcal{M}_{0-4}$ to learn the positive class.

Validation results

Follows the second phase of the classification, in which the best classifier obtained during the training phase is applied to the validation set (Table 59).

Two Years Temporal Window	noEvol	Evol
$\mathfrak{D}_{all}$	27 (73%)	10 (21%)
$\mathfrak{D}_{0-4}$	12 (63.2%)	7 (36.8%)
$\mathfrak{D}_{5-14}$	15 (83.3%)	3 (16.7%)

Table 59 Sizes of the validation datasets, using two years temporal window.

The Table 60 displays the best results considering the two approaches used throughout this analysis. The results are identical both in the model trained with all patients and in the model considering the depressive symptoms of the patients.

Models	Classifier	Test set	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{0-4(all)}$	SVM RBF	$\mathfrak{D}_{all}$	0.842	0.571	1	0.829
$\mathcal{M}_{5-14(all)}$			0.667	0	0.8	0.667
$\mathcal{M}_{all}$			0.757	0.4	0.889	0.742
$\mathcal{M}_{0-4}$	NB	$\mathfrak{D}_{0-4}$	0.737	0.286	1	0.686
$\mathcal{M}_{5-14}$	NB	$\mathfrak{D}_{5-14}$	0.778	0.667	0.8	0.798
$\mathcal{M}_{0-4}$ + $\mathcal{M}_{5-14}$			0.757	0.4	0.889	0.742

Table 60 Best classifiers for two years temporal window, using the validation set.

5.1.2.2 Three Years Temporal Window

Cross-Validation results

The sizes of the train datasets for the three years temporal window are described in Table 61.

Three Years Temporal Window	noEvol	Evol
$\mathfrak{D}_{all}$	58 (50.9%)	56 (49.1%)
$\mathfrak{D}_{0-4}$	35 (60.3%)	23 (39.7%)
$\mathfrak{D}_{5-14}$	23 (41.1%)	33 (58.9%)

Table 61 Sizes of the train datasets which are differentiated based on GDS values, using three years temporal.

The attribute Or_Total was selected in the three datasets. In the dataset $\mathfrak{D}_{0-4}$, from the set of attributes selected it is worth noting As_tot, PA_tot, LM_a_Cued, MPR, CVLT_a2, CancellationTask_Z, VerbalPairedAssociateLearning_Z, LogicalMemory_A_Z, LM_DR_Z and MPR_Z which are common to the ones selected in the entire dataset. The set of attributes LM_a_Interf, LM_tot_Interf, Forgetting, Orient_T, and SemanticFluency_Z were selected in the entire dataset and in the dataset $\mathfrak{D}_{5-14}$. Again there are still attributes specific for each one of the datasets in study that may be consulted in Appendix C. Figures 26, 27 and 28 show the train results in the three distinct datasets.

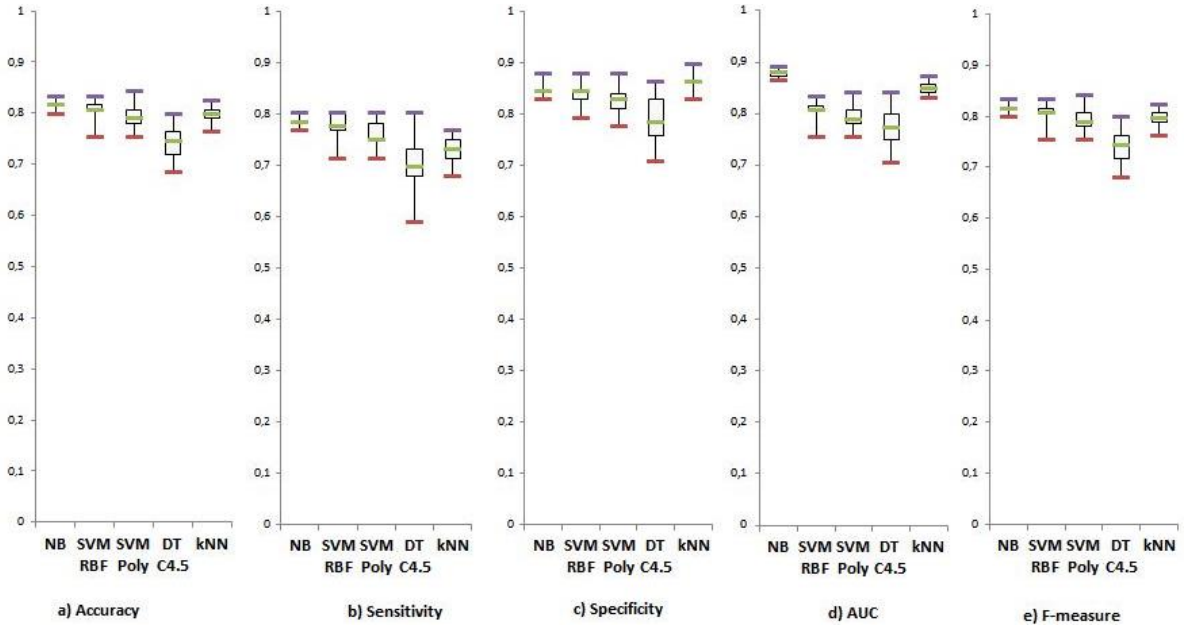

Figure 26 Train results of Prognosis of the dataset $\mathfrak{D}_{all}$, using three years temporal window.

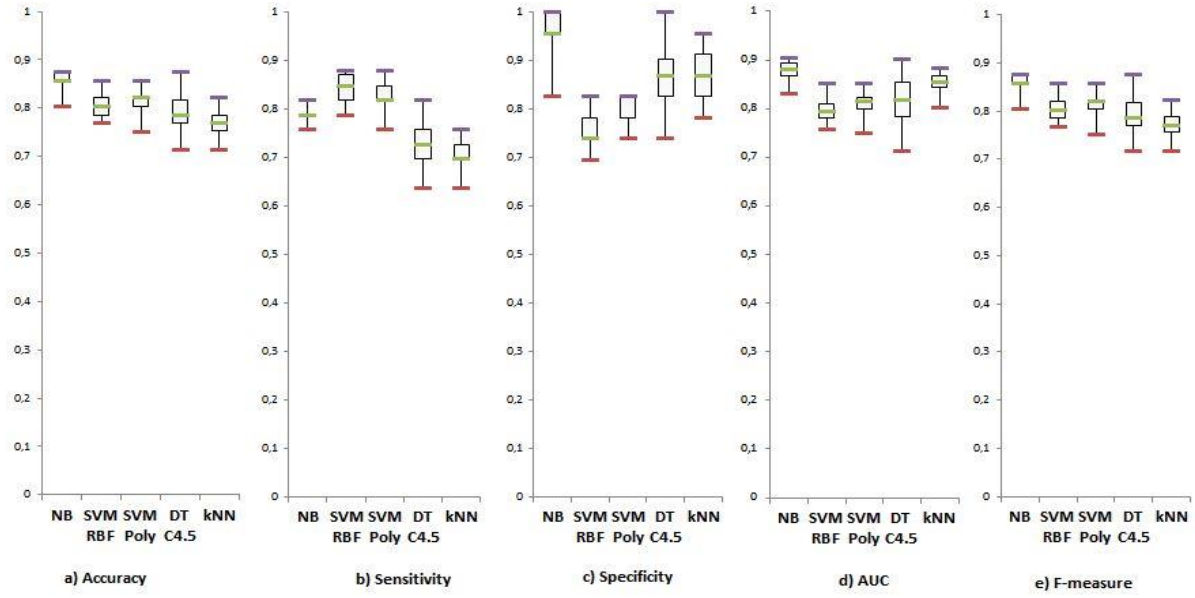

Figure 27 Train results of Prognosis of the dataset $\mathfrak{D}_{\mathbf{0-4}}$, using three years temporal window.

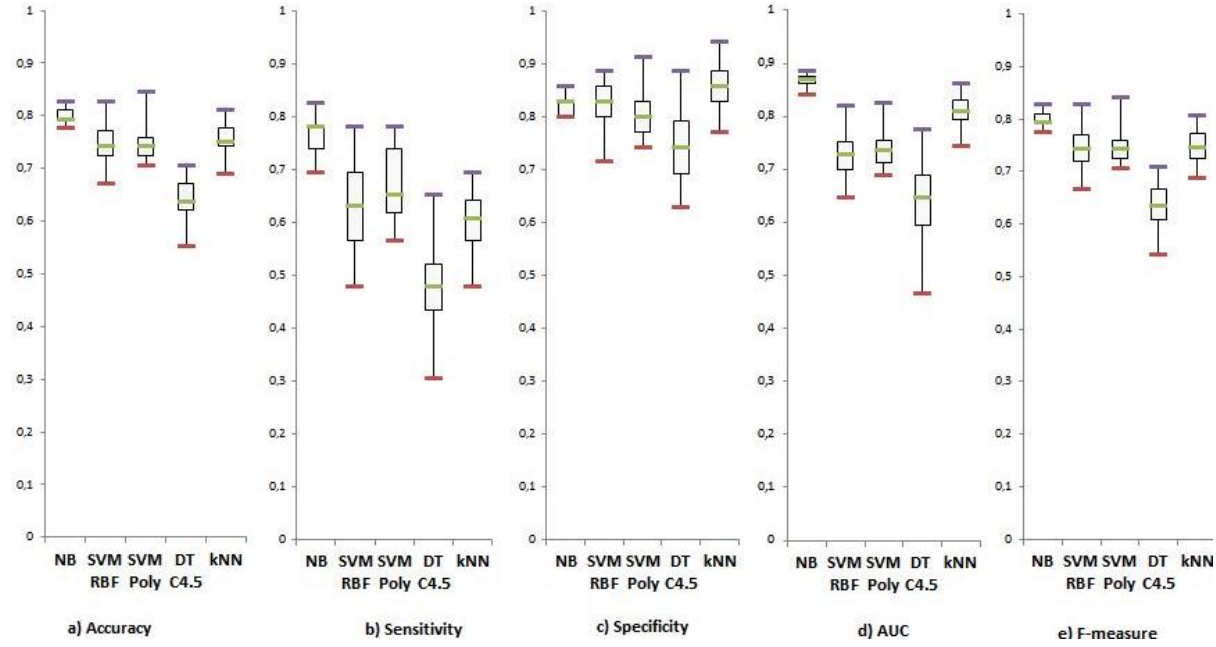

Figure 28 Train results of Prognosis of the dataset $\mathfrak{D}_{\mathbf{5-14}}$, using three years temporal window.

Paired t-tests were used to compare the classifiers, using 30 repetitions. In $\mathfrak{D}_{\boldsymbol{all}}$, the classifier NB is the best model and the C4.5DT got the worst results. In $\mathfrak{D}_{\mathbf{0-4}}$, the classifier NB is the best model and the C4.5DT model got the worst results. In $\mathfrak{D}_{\mathbf{5-14}}$, the classifier NB is the best model and the classifier *k*NN presents the lower results, with no statistical difference with C4.5DT model.

Concerning the confusion matrices presented in Table 62, once again it is noticeable that the behaviour of the general model is independent of the score of the GDS of the patient. In Table 63 it is presented the evaluation metrics for the general model containing all instances and the models for each restrictive dataset concerning the state of depression trained with the same classifier.

a)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	44 (78.6%)	12 (21.4%)
	noEvol	10 (17.2%)	48 (82.8%)

b)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	18 (78.3%)	5 (21.7%)
	noEvol	7 (17.2%)	28 (82.8%)

c)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	26 (78.8%)	7 (21.2%)
	noEvol	3 (13%)	20 (87%)

Table 62 Confusion Matrix of NB for three years temporal window, using 5-fold CV considering the $\mathfrak{D}_{all}$: a) considering all instances; b) considering instances with GDS≤4; c) considering instances with GDS>4.

Model	Classifier	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{all}$	NB	0.807	0.785	0.828	0.807
$\mathcal{M}_{\mathbf{0-4}(all)}$		0.793	0.783	0.8	0.794
$\mathcal{M}_{\mathbf{5-14}(all)}$		0.821	0.788	0.87	0.822

Table 63 Evaluation metrics of Naïve Bayes model for three years temporal window, using 5-fold CV.

Table 64 shows the confusion matrices obtained by the models obtained from each group of patients. Table 65 presents the evaluation metrics for each specific model $\mathcal{M}_{0-4}$ and $\mathcal{M}_{5-14}$, concerning the state of depression. In this temporal window, it is perceptible that the specific models, $\mathcal{M}_{\mathbf{0-4}}$ and $\mathcal{M}_{\mathbf{5-14}}$, considering different characteristics of the patients have higher predictive capability with F-measure of 0.81 and 0.876 respectively, when compared with the behaviour of the general model $\mathcal{M}_{\boldsymbol{all}}$ with 0.794 and 0.822. As regards to the results, the separation of patients concerning their state of depression seems to lead to a better learning of the models and produces valuable results.

a)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	17 (73.9%)	6 (26.1%)
	noEvol	5 (14.3%)	30 (85.7%)

b)

	Predicted Class	
	Evol	noEvol
Real Class Evol	26 (78.8%)	7 (21.2%)
noEvol	0 (0%)	23 (100%)

c)

	Predicted Class	
	Evol	noEvol
Real Class Evol	43 (76.7%)	13 (23.2%)
noEvol	5 (8.6%)	53 (91.4%)

Table 64 Confusion Matrix for three years temporal window, using 5-fold CV considering a) NB in $\mathfrak{D}_{0-4}$; b) NB in $\mathfrak{D}_{5-14}$; c) sum of the confusion matrices of each model.

Model	Classifier	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{\mathbf{0-4}}$	NB	0.81	0.739	0.857	0.81
$\boldsymbol{\mathcal{M}}_{\mathbf{5-14}}$	NB	0.875	0.788	1	0.876
$\boldsymbol{\mathcal{M}}$		0.842	0.768	0.914	0.841

Table 65 Evaluation metrics of each model for three years temporal window, using 5-fold CV.

Validation results

The best classifier obtained during the training phase is applied to the test set, described in Table 66. The Table 67 displays the best results considering the two approaches used throughout this analysis. In contradiction of what we obtained in the cross-validation results, the best results in this second phase of the classification are achieved with the model trained with all patients.

Three Years Temporal Window	noEvol	Evol
$\mathfrak{D}_{\boldsymbol{all}}$	19 (54.3%)	16 (45.7%)
$\mathfrak{D}_{\mathbf{0-4}}$	8 (47.1%)	9 (52.9%)
$\mathfrak{D}_{\mathbf{5-14}}$	11 (61.1%)	7 (38.9%)

Table 66 Sizes of the validation datasets, using three years temporal window.

Model	Classifier	Test set	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{\mathbf{0-4}(all)}$	NB	$\mathfrak{D}_{all}$	0.882	0.778	1	0.882
$\mathcal{M}_{\mathbf{5-14}(all)}$			0.696	0.571	0.75	0.701
$\mathcal{M}_{all}$			0.743	0.688	0.789	0.742
$\mathcal{M}_{\mathbf{0-4}}$	NB	$\mathfrak{D}_{\mathbf{0-4}}$	0.765	0.556	1	0.755
$\mathcal{M}_{\mathbf{5-14}}$	NB	$\mathfrak{D}_{\mathbf{5-14}}$	0.611	0.571	0.636	0.615
$\mathcal{M}_{\mathbf{0-4}}$ +$\mathcal{M}_{\mathbf{5-14}}$			0.686	0.563	0.789	0.681

Table 67 Best classifiers for three years temporal window, using the validation set.

5.1.2.3 Four Years Temporal Window

Cross-Validation results

The sizes of the train datasets size relative to the four years temporal window is described in Table 68.

Four Years Temporal Window	noEvol	Evol
$\mathfrak{D}_{all}$	32 (32%)	68 (68%)
$\mathfrak{D}_{\mathbf{0-4}}$	21 (40.4%)	31 (59.6%)
$\mathfrak{D}_{\mathbf{5-14}}$	11 (25%)	37 (75%)

Table 68 Sizes of the train datasets which are differentiated based on GDS values, using four years temporal.

The attributes SemanticFluency_Z, Or_Total and LogicalMemory_A_Z were selected in the three datasets. The attributes LM_tot_Interf, Orientation_Z, WordRecall_Z and Proverbs_Z were selected in the datasets $\mathfrak{D}_{all}$ and $\mathfrak{D}_{\mathbf{0-4}}$. The attribute MPR_Z was also selected in both datasets $\mathfrak{D}_{all}$ and $\mathfrak{D}_{\mathbf{5-14}}$. Figures 29, 30 and 31 show the train results in the three distinct datasets.

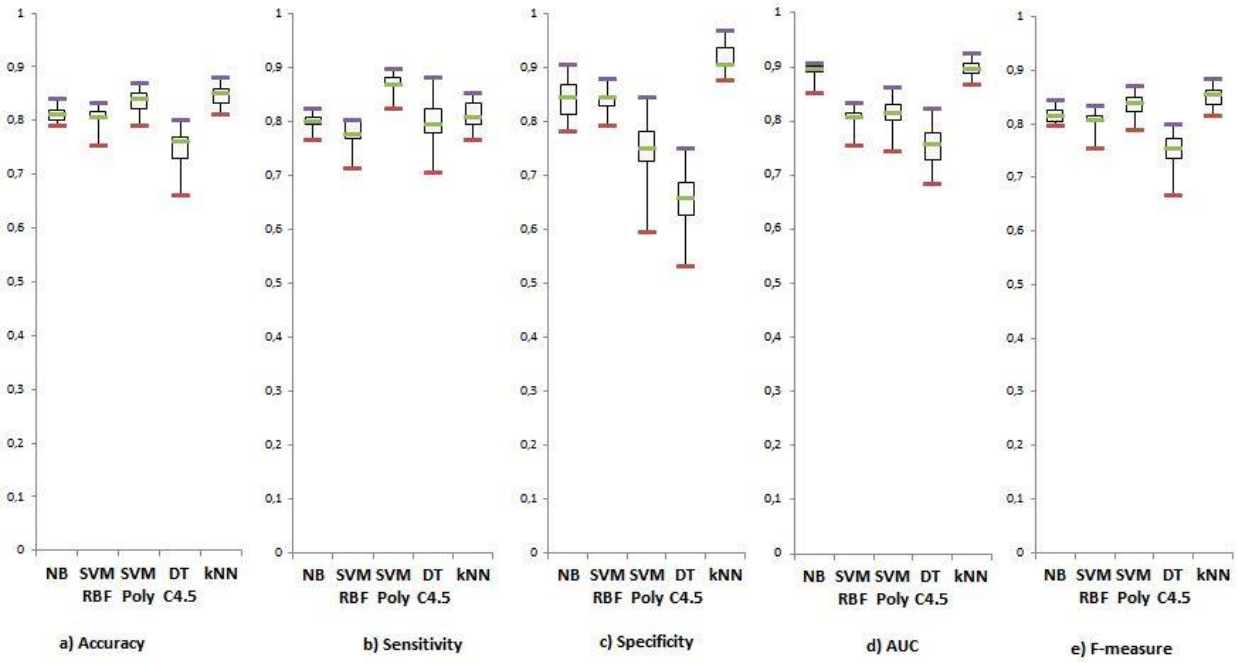

Figure 29 Train results of Prognosis of the dataset $\mathfrak{D}_{all}$, using four years temporal window.

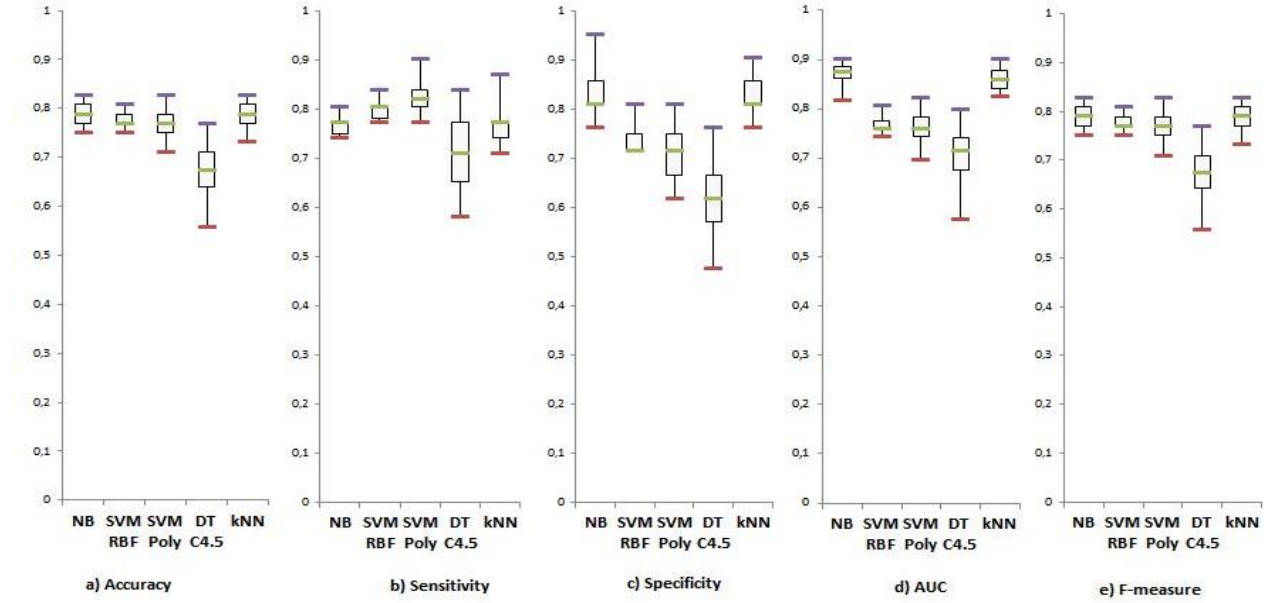

Figure 30 Train results of Prognosis of the dataset $\mathfrak{D}_{0-4}$, using four years temporal window.

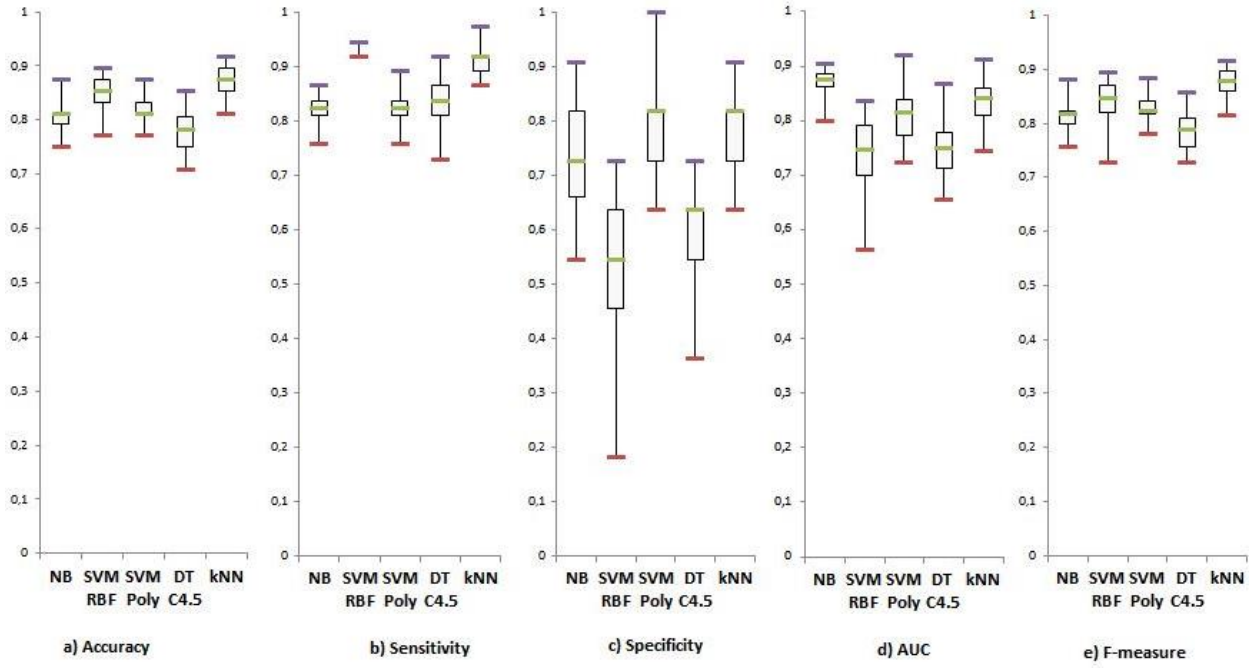

Figure 31 Train results of Prognosis of the dataset $\mathfrak{D}_{\mathbf{5-14}}$, using four years temporal window.

Paired t-tests were used to compare the classifiers, using 30 repetitions. In $\mathfrak{D}_{\mathbf{all}}$, kNN is the best model and the C4.5DT model got the worst results. In $\mathfrak{D}_{\mathbf{0-4}}$, the classifier NB is the best model, with no statistical difference with the kNN model and the C4.5DT model got the worst results. In $\mathfrak{D}_{\mathbf{5-14}}$, the classifier kNN is the best model and the classifier C4.5DT presents the lower results.

Table 69 shows the confusion matrices and Table 70 presents the evaluation metrics for the general model $\mathcal{M}_{(all)}$ and for each specific model $\mathcal{M}_{0-4(all)}$ and $\mathcal{M}_{5-14(all)}$, concerning the state of depression trained with this model.

a)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	49 (72.1%)	19 (27.9%)
	noEvol	3 (9.4%)	29 (90.6%)

b)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	20 (64.5%)	11 (35.5%)
	noEvol	2 (9.5%)	19 (90.5%)

c)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	29 (78.4%)	8 (21.6%)
	noEvol	1 (9.1%)	10 (90.9%)

Table 69 Confusion Matrix of *k*NN model for four years temporal window, using 5-fold CV considering the $\mathfrak{D}_{all}$: a) considering all instances; b) considering instances with GDS≤4; c) considering instances with GDS>4.

Model	Classifier	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{all}$	KNN	0.78	0.721	0.906	0.787
$\mathcal{M}_{\mathbf{0-4}(all)}$		0.75	0.645	0.905	0.751
$\mathcal{M}_{\mathbf{5-14}(all)}$		0.813	0.784	0.909	0.825

Table 70 : Evaluation metrics of *k*NN model for four years temporal window, using 5-fold CV.

In this temporal window, it is noticeable a lower cardinality of examples. Regarding the performance of the general model is worth mention the high percentage of false negatives, in particular when predicting instances with GDS≤4 with 35.5%. This justifies the low sensitivity of $\boldsymbol{\mathcal{M}_{\mathbf{0-4}(all)}}$ of 0.645.

Table 71 shows the confusion matrices obtained by the models obtained from each group of patients. Table 72 presents the evaluation metrics for each specific model, $\mathcal{M}_{0-4}$ and $\mathcal{M}_{5-14}$, concerning the state of depression. The datasets trained during this analysis have a reduce composition of both classes. In opposition to what happen in the previous temporal window, the models that learn considering the depressive symptoms of the patients, $\mathcal{M}_{0-4}$ and $\mathcal{M}_{5-14}$, and the general models, $\mathcal{M}_{\mathbf{0-4}(all)}$ and $\mathcal{M}_{\mathbf{5-14}(all)}$, do not differ in terms of accuracy and F-measure. This study should be repeated with larger groups of MCI patients since these results may be consequence of the smaller size of the dataset.

a)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	24 (77.4%)	7 (22.6%)
	noEvol	4 (19.1%)	17 (80.9%)

b)

	Predicted Class	
	Evol	noEvol
Real Class Evol	29 (78.4%)	8 (21.6%)
noEvol	3 (27.3%)	8 (72.7%)

c)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	53 (77.9%)	15 (22.1%)
	noEvol	7 (19%)	25 (81%)

Table 71 Confusion Matrix for four years temporal window, using 5-fold CV considering a) NB in $\mathfrak{D}_{0-4}$ b) *k*NN in $\mathfrak{D}_{5-14}$; c) sum of the confusion matrices of each model.

Model	Classifier	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{\mathbf{0-4}}$	NB	0.789	0.774	0.809	0.790
$\boldsymbol{\mathcal{M}_{\mathbf{5-14}}}$	KNN	0.771	0.784	0.727	0.784
$\boldsymbol{\mathcal{M}}$		0.78	0.779	0.781	0.785

Table 72 Evaluation metrics of each model for four years temporal window, using 5-fold CV.

Validation results

The second phase of the classification is presented below, in which the best classifiers obtained during the training phase were applied to the test set, described in Table 73. The Table 74 displays the best results. The best results in the validation phase are achieved with the specific models trained.

Four Years Temporal Window	noEvol	Evol
$\mathfrak{D}_{all}$	12 (40%)	18 (60%)
$\mathfrak{D}_{\mathbf{0-4}}$	6 (35.3%)	11 (64.7%)
$\mathfrak{D}_{\mathbf{5-14}}$	6 (46.2%)	7 (53.8%)

Table 73 : Sizes of the validation datasets, using four years temporal window.

Model	Classifier	Test set	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{\mathbf{0-4}(all)}$	*k*NN	$\mathfrak{D}_{all}$	0.882	0.818	0.818	0.885
$\mathcal{M}_{\mathbf{5-14}(all)}$			0.769	0.857	0.667	0.767
$\mathcal{M}_{all}$			0.833	0.833	0.833	0.834
$\mathcal{M}_{\mathbf{0-4}}$	NB	$\mathfrak{D}_{\mathbf{0-4}}$	0.882	1	0.667	0.875
$\mathcal{M}_{\mathbf{5-14}}$	*k*NN	$\mathfrak{D}_{\mathbf{5-14}}$	0.846	1	0.667	0.84
$\mathcal{M}_{\mathbf{0-4}}$ **+**$\mathcal{M}_{\mathbf{5-14}}$			0.867	1	0.667	0.86

Table 74 Best classifiers for four years temporal window, using the validation set.

5.1.2.4 Five Years Temporal Window

Cross-Validation results

The sizes of the train datasets size relative to the five years temporal window is described in Table 75.

Five Years Temporal Window	noEvol	Evol
$\mathfrak{D}_{all}$	16 (17.6%)	75 (82.4%)
$\mathfrak{D}_{\mathbf{0-4}}$	9 (19.6%)	37 (80.4%)
$\mathfrak{D}_{\mathbf{5-14}}$	7 (15.6%)	38 (84.4%)

Table 75 Sizes of the train datasets which are differentiated based on GDS values, using five years temporal.

In the dataset $\mathfrak{D}_{\mathbf{0-4}}$**,** from the set of attributes selected Or_Total, Orientation_Z, LogicalMemory_A_Z, and A5_Z are common to the attributes selected in the entire dataset. The set of attributes LM_a_Interf and SemanticFluency_Z were selected in the entire dataset and in the dataset $\mathfrak{D}_{\mathbf{5-14}}$. Figures 32, 33 and 34 show the train results in the three distinct datasets.

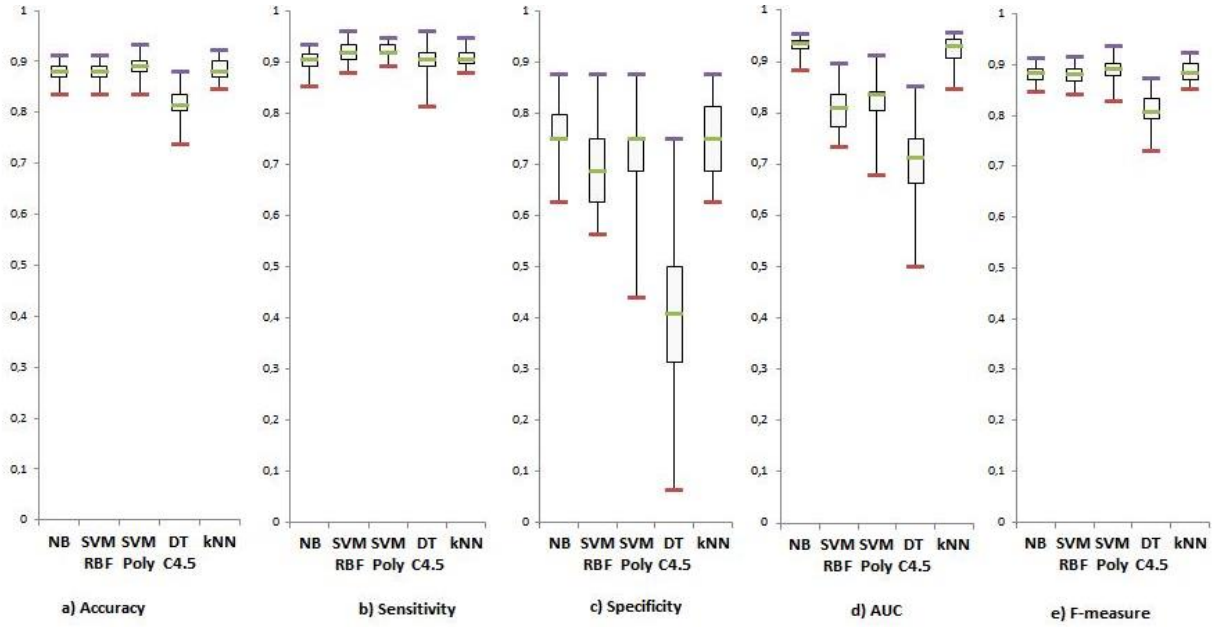

Figure 32 Train results of Prognosis of the dataset $\mathfrak{D}_{all}$, using five years temporal window.

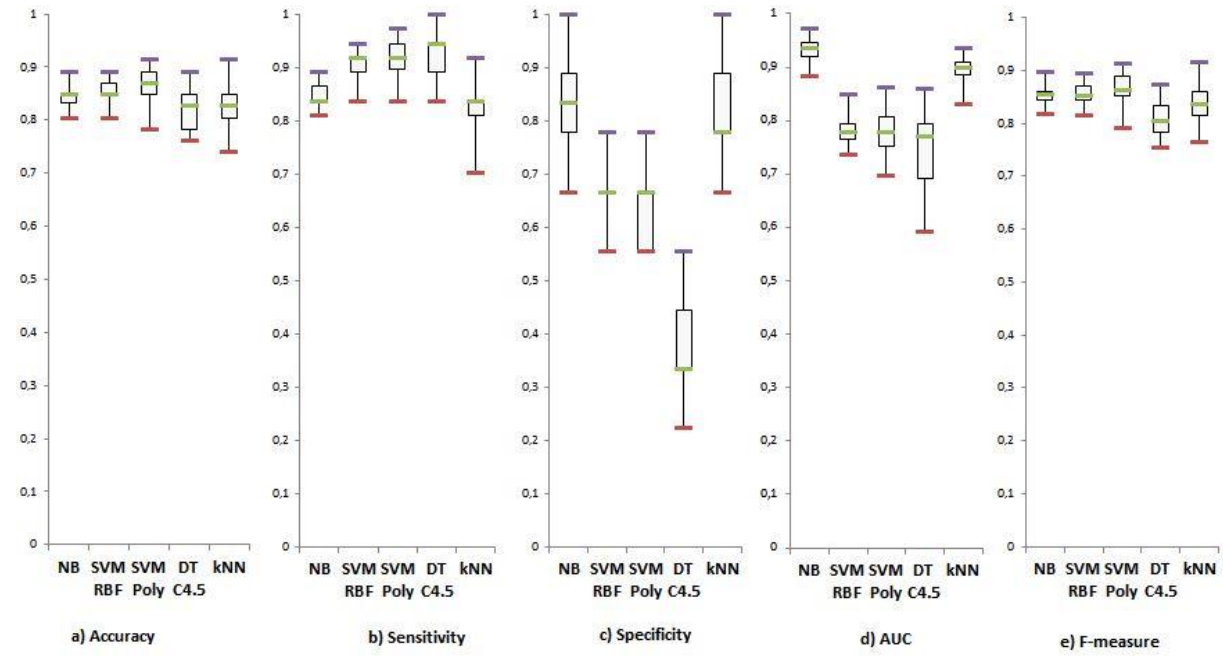

Figure 33 Train results of Prognosis of the dataset $\mathfrak{D}_{0-4}$, using five years temporal window.

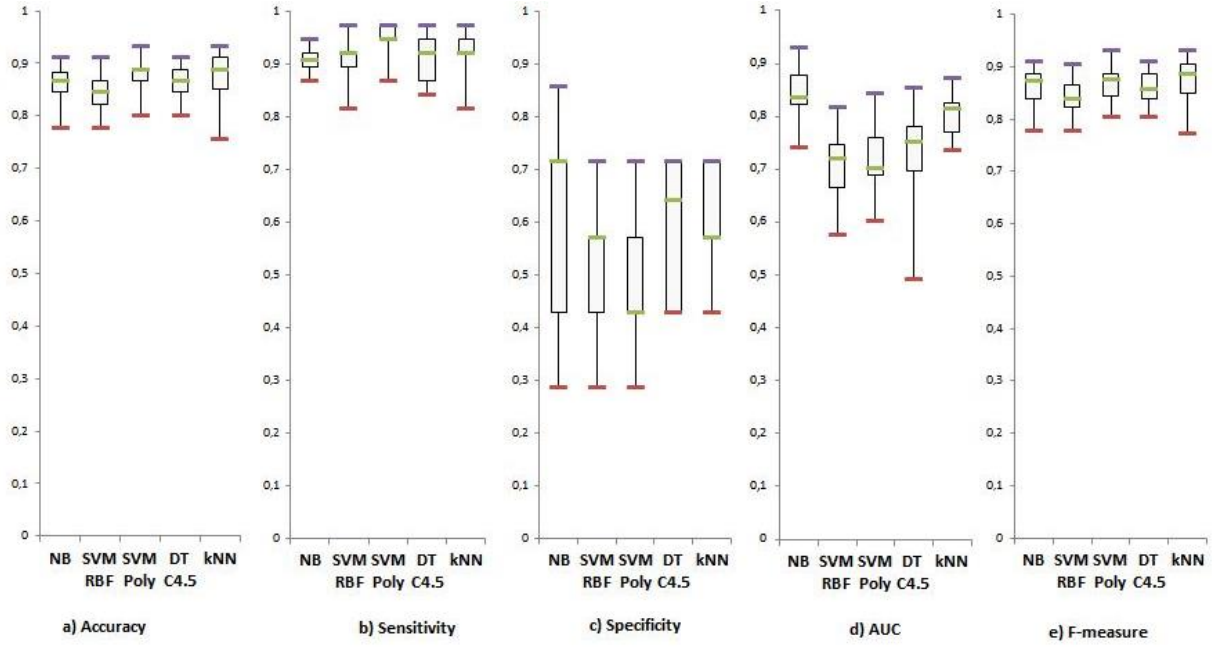

Figure 34 Train results of Prognosis of the dataset $\mathfrak{D}_{\mathbf{5-14}}$, using five years temporal window.

Paired t-tests were used to compare the classifiers, using 30 repetitions. In $\mathfrak{D}_{\boldsymbol{all}}$, SVM Poly classifier is the best model, with no statistical difference with the NB and *k*NN, and the C4.5DT model got the worst results. In $\mathfrak{D}_{\mathbf{0-4}}$, the SVM Poly is the best model, with no statistical difference with Naïve Bayes and SVM RBF, and the C4.5DT model got the worst results. In $\mathfrak{D}_{\mathbf{5-14}}$, the classifier *k*NN is the best model, with no statistical difference with Naïve Bayes, SVM Poly and C4.5DT. The classifier SVM RBF presents the lower results.

Table 76 shows the confusion matrices and Table 77 presents the evaluation metrics for the general model $\mathcal{M}_{(all)}$ and for each specific model $\mathcal{M}_{0-4(all)}$ and $\mathcal{M}_{5-14(all)}$, concerning the state of depression trained with this model.

a)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	70 (93.3%)	5 (6.7%)
	noEvol	5 (31.3%)	11 (68.7%)

b)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	34 (91.9%)	3 (8.1%)
	noEvol	3 (33.3%)	6 (66.7%)

c)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	36 (94.7%)	2 (5.3%)
	noEvol	2 (28.6%)	5 (71.4%)

Table 76 Confusion Matrix of SVM Poly for five years temporal window, using 5-fold CV considering the $\mathfrak{D}_{all}$: a) considering all instances; b) considering instances with GDS≤4; c) considering instances with GDS>4.

Model	Classifier	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}$	SVM Poly	0.824	0.907	0.438	0.845
$\mathcal{M}_{\mathbf{0-4}(all)}$		0.777	0.889	0.333	0.865
$\mathcal{M}_{\mathbf{5-14}(all)}$		0.87	0.923	0.571	0.923

Table 77 Evaluation metrics of SVM Poly model for five years temporal window, using 5-fold CV.

Looking to the high values of F-measure, consequence of the high values of sensitivity, we can conclude that the model is learning very well from the positive class, the Evol class. The low values of specificity is related with the reduce examples of noEvol.

Table 78 shows the confusion matrices obtained by the models obtained from each group of patients. Table 79 presents the evaluation metrics for each specific model $\mathcal{M}_{0-4}$ and $\mathcal{M}_{5-14}$, concerning the state of depression.

a)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	34 (91.9%)	3 (8.1%)
	noEvol	3 (33.3%)	6 (66.7%)

b)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	33 (86.8%)	5 (13.2%)
	noEvol	2 (28.6%)	5 (71.4%)

c)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	67 (89.3%)	8 (10.7%)
	noEvol	5 (31.3%)	11 (68.8%)

Table 78 Confusion Matrix for five years temporal window, using 5-fold CV considering a) SVM Poly in $\mathfrak{D}_{0-4}$; b) kNN in $\mathfrak{D}_{5-14}$ c) sum of the confusion matrices of each model.

Model	Classifier	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{\mathbf{0-4}}$	SVM Poly	0.87	0.919	0.667	0.87
$\mathcal{M}_{\mathbf{5-14}}$	*k*NN	0.844	0.868	0.714	0.855
$\mathcal{M}$		0.857	0.893	0.688	0.862

Table 79 Evaluation metrics of each model for five years temporal window, using 5-fold CV.

As presented in table 75, the datasets have a small number of examples, composed by low examples of the negative class, which influences the learning of the models. Though, we obtained a good overall performance in this temporal window separating the patients according to the depressive symptoms of the patient.

As seen in the previous temporal window, the separation according to the state of depression of the patient is an approach with really good potential and to take in consideration in future works.

Validation results

The best classifiers obtained during the training phase were applied to the test set, described in Table 80. The Table 81 displays the best results considering other metrics.

Five Years Temporal Window	noEvol	Evol
$\mathfrak{D}_{all}$	18 (69.2%)	8 (30.8%)
$\mathfrak{D}_{0-4}$	11 (78.6%)	3 (21.4%)
$\mathfrak{D}_{5-14}$	7 (58.3%)	5 (41.7%)

Table 80 Sizes of the validation datasets, using five years temporal window.

Model	Classifier	Test set	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{0-4(all)}$	SVM Poly	$\mathfrak{D}_{all}$	0.923	1	0.667	0.923
$\mathcal{M}_{5-14(all)}$			0.917	1	0.8	0.915
$\mathcal{M}_{all}$			0.923	1	0.75	0.92
$\mathcal{M}_{0-4}$	SVM Poly	$\mathfrak{D}_{0-4}$	0.857	1	0.333	0.827
$\mathcal{M}_{5-14}$	KNN	$\mathfrak{D}_{5-14}$	0.583	1	0	0.43
$\mathcal{M}_{0-4}$ +$\mathcal{M}_{5-14}$			0.731	1	0.125	0.648

Table 81 Best classifiers for five years temporal window, using the validation set.

In contradiction of what we obtained in the cross-validation results, the best results in this second phase of the classification are achieved with the model trained with all patients.

5.1.3 Discussion

The results presented in this study corroborate the hypothesis that the separation of the patients according to their state of depression influences positively the prognosis results. We expect to obtain even better results with a more complete database. The smaller size of the groups in this analysis in comparison with the total sample of the baseline study associated with the more than 50% of missing values regarding the GDS are limitations. It would be important the integration of more patients with the GDS information, which would allow for a higher prediction capability separating the patients with different characteristics. In this way, the repetition of this study would be recommended with a dataset with more follow-up assessments. It should be noted that even though the current study is based on a limited number of patients, this new data mining approach has potential as a possible strategy for the prediction of conversion of MCI patients to AD. In future works, it would be important to use parametric tests, as the Wilcoxon and the Friedman tests, to verify if the differences between the models with separation of the patients considering their state of depression have statistical differences.

One aspect to have in consideration is the susceptibility of clinical judgement to biases and faulty assumptions that can lead to diagnostic errors. Even though the results presented in the above sections demonstrated that to increase the ability to predict conversion to AD, future studies should differentiate MCI groups using clinical knowledge as the Clinical Dementia Rating scale (CDR). This test is starting to assume a relevant role in the conversion to dementia [29, 31, 13, 19]. CDR is a subjective assessment that characterizes six domains of cognitive and functional performance based on a structural clinical interview [19, 26, 53].

5.2 Prognosis prediction based on Patient Similarities

5.2.1 Clustering

Whenever carrying out unsupervised learning analyses, it is important to consider the relative scales of the attributes being measured. The cluster analysis algorithms depend on the concept of measuring the distance or other measure of similarity between the different patients we want to cluster. Therefore, this kind of analyses cannot be performed from all attributes together [1, 21]. Bondi *et al.* [26] transformed each of the neuropsychological measures to age-corrected z-scored based on the mean and standard deviations of the normal reference group and then the cluster analyses were conducted using the z-scores.

The analysis was performed with different set of attributes, resulting from the separation of the attributes that are not corrected for age and school (called as A attributes) and the ones with a z-score correspondent (called as B and B_Z attributes). One set of features analysed was labelled as A+B attributes and the other was labelled as A+B_Z attributes. In order to simplify the interpretation of the results, it only be shown the ones regarding the attributes corrected for age and school (A+B_Z attributes).

This section describes the results obtained by means of clustering algorithms without using clinical criteria to identify groups of patients, in order to investigate the potential existence of MCI groups (already defined in literature [26, 28, 55] or other). Since MCI is a state between the cognitive changes of normal ageing and what might constitute very early dementia, we first assume on the existence of three subgroups. The first one would be representative of a less severe cognitive decline profile within MCI groups, the other would be representative of the severe cognitive decline and the third group, with persistence of symptoms, would be representative of a mildly cognitive impairment. The next approach characterizes MCI using cluster analyses techniques that determine how individuals group together based on their patterns of performance, across the attributes selected from the baseline analysis in chapter 4. The features used throughout the clustering analysis can be consulted in Appendix C.

The choice of the optimal number of clusters *k* is a difficult task as well as the evaluation of the cluster´s quality. Follows a description of the noticeable changes in the groups of patients induced by the alteration of the predefined number of clusters, for each temporal window. In order to test the possible existence of three MCI groups, this study was performed for different number of clusters, in particular 2, 3 and 4.

Table 82 shows the confusion matrices resulted, after selecting the set of features of the two years temporal window in the baseline analysis.

a)

		Cluster	
		0	**1**
Real Class	**noEvol**	186	15
	Evol	51	33
# Instances		237	48

b)

		Cluster		
		0	**1**	**2**
Real Class	**noEvol**	82	94	25
	Evol	28	19	37
# Instances		110	113	62

c)

		Cluster			
		0	**1**	**2**	**3**
Real Class	**noEvol**	11	20	83	87
	Evol	0	34	19	31
# Instances		11	54	102	118

Table 82 Confusion matrix for the two years temporal window obtained by the EM algorithm with a) k=2; b) k=3; c) k=4.

A new row was added to each confusion matrix with the purpose of indicating the total number of the patients that belong to a given cluster.

In the generation of two clusters with EM algorithm, there is evident the existence of a larger group, mainly composed by noEvol examples and a smaller group mainly composed by Evol examples. When three groups are used (k=3), there is a variance between the proportions of the clusters, with a larger group, with 113 instances, a second group with 110 instances and smaller one with 62 instances. The smaller group is composed mainly by Evol patients and the other two groups include instances mostly from noEvol patients. When k=4, two larger groups are evident, containing the maximum amount of noEvol patients and it is possible to define a third group with 11 examples only composed by noEvol examples. The smaller group contains more Evol instances than noEvol. It was also noted that the formation of new clusters was supported by splitting the small group.

Table 83 shows the confusion matrices of the clustering resulted in the three years temporal window.

a)

		Cluster	
		0	**1**
Real Class	**noEvol**	126	13
	Evol	53	62
# Instances		179	75

b)

		Cluster		
		0	**1**	**2**
Real Class	**noEvol**	69	12	58
	Evol	23	57	35
# Instances		92	69	93

c)

		Cluster			
		0	**1**	**2**	**3**
Real Class	**noEvol**	7	12	57	63
	Evol	9	57	35	14
# Instances		16	69	92	77

Table 83 Confusion matrix for the three years temporal window obtained by the EM algorithm with a) k=2; b) k=3; c) k=4.

It is detected the same patterns. Using k=2, it is obtained a cluster mainly composed by Evol instances and another one mainly composed by noEvol instances. When three clusters are formed, there is a presence of two major groups mostly composed by noEvol patients and the formation of a smaller group with more Evol patients similarly to what happen in the two years temporal window. When k=4, two larger groups are evident, containing the maximum amount of noEvol patients and two other smaller groups mainly composed by Evol examples.

In general, the results achieved clusters formed by a disorganized blend of patients from both classes, demonstrating the difficult task of defining MCI groups.

Table 84 shows the confusion matrices of the clustering resulted in the four years temporal window.

a)

		Cluster	
		0	**1**
Real Class	**noEvol**	80	5
	Evol	71	72
# Instances		151	77

b)

		Cluster		
		0	**1**	**2**
Real Class	**noEvol**	43	37	5
	Evol	24	66	53
# Instances		67	103	58

c)

		Cluster			
		0	**1**	**2**	**3**
Real Class	**noEvol**	16	9	23	37
	Evol	23	85	8	27
# Instances		39	94	31	118

Table 84 Confusion matrix for the four years temporal window obtained by the EM algorithm with a) k=2; b) k=3; c) k=4.

In the four years temporal window is even more evident that dividing the dataset in three clusters will not be a choice since reduced number of examples of the smaller group will bias the results of the learning. It is expected to observe the same in the five temporal window. Even though we wanted to analyse the prognosis problem considering three automatic groups, due to the fact that the number of patients per cluster is reduced, invalidating the learning of the models, it was established to consider two clusters in the classification, in every temporal window. However, when the dataset is divided in three clusters, is it noticeable the existence of the two groups, with more Evol examples, and another one with more noEvol examples. When four groups are separated, there are three clusters mainly composed by Evol examples and one of them is mostly composed by noEvol instances.

Table 85 shows the confusion matrices of the clustering resulted in the five years temporal window.

a)

		Cluster	
		0	**1**
Real Class	**noEvol**	15	36
	Evol	128	26
# Instances		237	72

b)

		Cluster		
		0	**1**	**2**
Real Class	**noEvol**	32	3	16
	Evol	30	63	71
# Instances		62	66	87

c)

		Cluster			
		0	**1**	**2**	**3**
Real Class	**noEvol**	8	24	4	15
	Evol	24	14	63	63
# Instances		32	38	67	78

Table 85 Confusion matrix for the five years temporal window obtained by the EM algorithm with a) k=2; b) k=3; c) k=4.

Here, it is noticed once again a disordered blend of patients from both classes, without accomplishing the separation of the classes. When three clusters are formed, there is a presence of two major groups mostly composed by Evol patients and the formation of a smaller group with more noEvol patients. Using k=4, it is possible to observe the formation of the three groups mainly composed by Evol patients and another with mostly noEvol patients. As aforementioned, due to the low number of patients per cluster, it was established to consider two clusters in the classification.

5.2.2 Classification

In order to determine the differences between models, firstly the entire set of features from the dataset $\mathfrak{D}_{all}$ is reduced to the features from the baseline analysis. Then, the dataset comprising all the instances along with the set of attributes selected is divided in the two clusters and the clusters are analysed in a cross-validation evaluation. Figure 35 shows the workflow of the approach used.

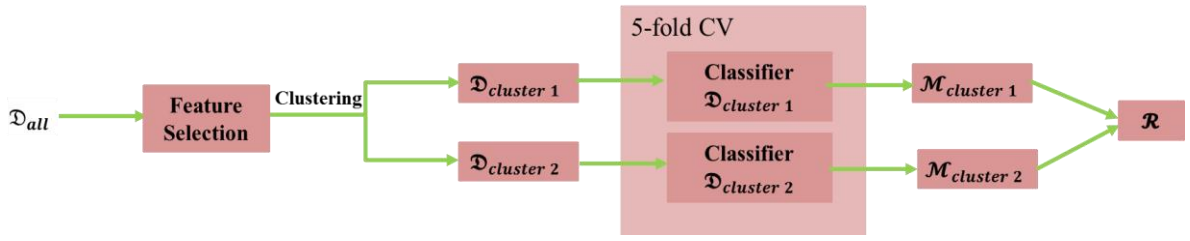

Figure 35 Workflow of the approach performed during the training phase.

5.2.2.1 Two Years Temporal Window

The sizes of the train datasets size relative to the two years temporal window is described in Table 86.

Two Years Temporal Window	noEvol	Evol
$\mathfrak{D}_{cluster\ 1}$	186 (78.5%)	51 (21.5%)
$\mathfrak{D}_{cluster\ 2}$	15 (31.3%)	33 (68.8%)

Table 86 Sizes of the train datasets formed by clustering using two years temporal window.

Paired t-tests were used to compare the classifiers, using 30 repetitions. The t-test is only applied if the ANOVA test with 95% confidence level confirms the existence of a significant difference. In $\mathfrak{D}_{cluster\ 1}$, SVM Poly, SVM RBF and NB are the best models, with no statistically difference, and the *k*NN model got the worst results. In $\mathfrak{D}_{cluster\ 2}$, SVM RBF is the best model, with no statistical difference with SVM Poly and the DT4.5 model got the worst results.

Table 87 shows the confusion matrices obtained by the clusters obtained. Table 88 presents the evaluation metrics for each model. It is noticeable the lower values of assessment metrics when the learning is performed considering the clusters obtained, comparing with the results from the baseline analysis.

a)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	15 (29.4%)	36 (70.6%)
	noEvol	40 (21.5%)	146 (78.5%)

b)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	26 (78.8%)	7 (70.6%)
	noEvol	5 (21.5%)	10 (78.5%)

c)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	41 (52.7%)	43 (39.3%)
	noEvol	45 (19.4%)	156 (80.6%)

Table 87 Confusion Matrix for two years temporal window, using 5-fold CV a) SVM Poly in $\mathfrak{D}_{cluster\ 1}$ b) SVM RBF in $\mathfrak{D}_{cluster\ 2}$; c) sum of the confusion matrices of each model.

Model	Classifier	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{cluster\ 1}$	SVM Poly	0.679	0.294	0.785	0.684
$\mathcal{M}_{cluster\ 2}$	SVM RBF	0.75	0.788	0.667	0.754
$\mathcal{M}$		0.691	0.488	0.776	0.692

Table 88 Evaluation metrics of each model for two years temporal window, using 5-fold CV.

5.2.2.2 Three Years Temporal Window

The sizes of the train datasets size relative to the three years temporal window is described in Table 89.

Three Years Temporal Window	noEvol	Evol
$\mathfrak{D}_{cluster\ 1}$	126 (70.4%)	53 (29.6%)
$\mathfrak{D}_{cluster\ 2}$	13 (17.3%)	62 (82.7%)

Table 89 Sizes of the train datasets formed by clustering using three years temporal window.

Paired t-tests were used to compare classifiers, using 30 repetitions. In $\mathfrak{D}_{cluster\ 1}$, NB presents the best results and the *k*NN model got the worst results. In $\mathfrak{D}_{cluster\ 2}$, the classifier SVM Poly is the best model, with no statistical difference with *k*NN, and the DT4.5 model got the worst results.

Table 90 shows the confusion matrices obtained by the clusters obtained. Table 91 presents the evaluation metrics for each model. Once again, the learning using the clusters obtained produce lower values of assessment metrics, comparing with the results from the baseline analysis.

a)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	21 (39.6%)	32 (60.4%)
	noEvol	12 (9.3%)	114 (90.5%)

b)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	57 (91.9%)	5 (8.1%)
	noEvol	9 (69.2%)	4 (30.8%)

c)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	78 (67.8%)	37 (32.2%)
	noEvol	21 (15.1%)	118 (84.9%)

Table 90 Confusion Matrix for three years temporal window, using 5-fold CV a) NB in $\mathfrak{D}_{cluster\ 1}$ b) SVM Poly in $\mathfrak{D}_{cluster\ 2}$; c) sum of the confusion matrices of each model.

Model	Classifier	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{cluster\ 1}$	NB	0.754	0.396	0.905	0.735
$\mathcal{M}_{cluster\ 2}$	SVM Poly	0.813	0.919	0.307	0.799
$\mathcal{M}$		0.772	0.678	0.849	0.769

Table 91 Evaluation metrics of each model for three years temporal window, using 5-fold CV.

5.2.2.3 Four Years Temporal Window

The sizes of the train datasets size relative to the four years temporal window is described in Table 92.

Four Years Temporal Window	noEvol	Evol
$\mathfrak{D}_{cluster\ 1}$	80 (53%)	71 (47%)
$\mathfrak{D}_{cluster\ 2}$	5 (6.5%)	72 (93.5%)

Table 92 Sizes of the train datasets formed by clustering using four years temporal window.

Paired t-tests were used to compare the classifiers, using 30 repetitions. In $\mathfrak{D}_{cluster\ 1}$, SVM RBF and NB are the best models, with no statistical difference, and the DT4.5, SVM Poly and *k*NN models got the worst results, with no statistical difference as well. In $\mathfrak{D}_{cluster\ 2}$, SVM RBF and NB are the best models, with no statistical difference and the DT4.5 model got the worst results.

Table 93 shows the confusion matrices obtained by the clusters obtained. Table 94 presents the evaluation metrics for each model. We can notice that that the values of assessment metrics are lower when the learning is performed considering the clusters obtained, similarly to what happen in the previous temporal window.

a)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	52 (73.2%)	19 (26.8%)
	noEvol	30 (37.5%)	50 (62.5%)

b)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	72 (100%)	0 (0%)
	noEvol	4 (80%)	1 (20%)

c)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	124 (86.7%)	19 (13.3%)
	noEvol	34 (40%)	51 (60%)

Table 93 Confusion Matrix for four years temporal window, using 5-fold CV a) SVM RBF in $\mathfrak{D}_{\mathbf{cluster\ 1}}$ b) SVM RBF in $\mathfrak{D}_{\mathbf{cluster\ 2}}$; c) sum of the confusion matrices of each model.

Model	Classifier	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{cluster\ 1}$	SVM RBF	0.675	0.732	0.625	0.675
$\mathcal{M}_{cluster\ 2}$	SVM RBF	0.948	1	0.2	0.931
$\mathcal{M}$		0.768	0.867	0.6	0.762

Table 94 Evaluation metrics of each model for four years temporal window, using 5-fold CV.

5.2.2.4 Five Years Temporal Window

The sizes of the train datasets size relative to the five years temporal window is described in Table 94.

Five Years Temporal Window	noEvol	Evol
$\mathfrak{D}_{cluster\ 1}$	15 (10.5%)	128 (89.5%)
$\mathfrak{D}_{cluster\ 2}$	51 (23.7%)	164 (76.3%)

Table 95 Sizes of the train datasets formed by clustering using five years temporal window.

Paired t-tests were used to compare the classifiers, using 30 repetitions. In $\mathfrak{D}_{cluster\ 1}$, SVM RBF, NB and DT4.5 are the best models, with no statistical difference, and *k*NN model got the worst results. In $\mathfrak{D}_{cluster\ 2}$, SVM RBF are the best model and the DT4.5 model got the worst results.

Table 96 shows the confusion matrices obtained by the clusters obtained. Table 97 presents the evaluation metrics for each model.

a)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	128 (100%)	0 (0%)
	noEvol	0 (0%)	15 (100%)

b)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	155 (94.5%)	9 (5.5%)
	noEvol	27 (52.9%)	24 (47.1%)

c)

		Predicted Class	
		Evol	noEvol
Real Class	Evol	283 (96.9%)	9 (3.1%)
	noEvol	27 (40.9%)	39 (59.1%)

Table 96 Confusion Matrix for five years temporal window, using 5-fold CV a) NB in $\mathfrak{D}_{\boldsymbol{cluster\ 1}}$ b) SVM RBF in $\mathfrak{D}_{\boldsymbol{cluster\ 2}}$; c) sum of the confusion matrices of each model.

Model	Classifier	Accuracy	Sensitivity	Specificity	F-measure
$\mathcal{M}_{\boldsymbol{cluster\ 1}}$	NB	1	1	1	1
$\mathcal{M}_{\boldsymbol{cluster\ 2}}$	SVM RBF	0.833	0.945	0.471	0.819
$\mathcal{M}$		0.899	0.969	0.591	0.893

Table 97 Evaluation metrics of each model for five years temporal window, using 5-fold CV.

5.2.3 Discussion

In general, the results achieved clusters constituted by a disorganized blend of patients from both classes, demonstrating the difficult task of defining MCI groups. Despite the fact that the results of the classification of patients based on unsupervised methods were not what we expected, this analysis validated the importance of the study of MCI subgroups considering the different characteristics of this clinical entity and recent research [26, 28, 53] using cluster analysis techniques has shown that MCI cohorts present heterogeneous cognitive profiles.

Thus, other techniques of clustering should be applied to corroborate such assumptions, such as similarity clustering algorithms. One of the approaches to try in future analyses is the feature selection for unsupervised learning [14], reducing the number of features and taken in consideration that in unsupervised learning we are not given the class labels.

Future analyse of our longitudinal study should include more follow-up assessments to provide new data supporting for a better understanding of the MCI subgroups. In order to validate the higher predictability of the specific models further application to larger groups of MCI patients monitored clinically are needed to confirm the current findings and we expect to obtain even better results with a more complete dataset.

6. Conclusions and Future Work

The work developed in this dissertation focused on the study of the conversion of Mild Cognitive Impairment (MCI) patients to Alzheimer' Disease (AD). MCI is a prodromal state that represents transitional period between normal changes in cognitive ageing and dementia. Furthermore, those patients represent higher risk of evolving to dementia. The construct of MCI proposes to identify individuals at an earlier point in the cognitive decline, such that if therapeutic interventions become available, clinicians can intervene at this juncture. Nevertheless, no consensus has been reached yet.

In this context, this work had two main challenges: (1) predict the conversion from MCI to AD using First and Last evaluations and temporal windows between two to five years and (2) predict conversion from MCI to AD using different MCI subgroups, obtained by unsupervised learning techniques and clinical knowledge. To perform this work we used a dataset consisting of neuropsychological tests and the corresponding diagnosis, obtained by the Cognitive Complaints Cohort (CCC) study that investigates cognitive stability or dementia on subjects with cognitive complaints. Those analyses were performed using standard classifiers as NB, SVM RBF, SVM Poly, C4.5DT and *k*NN. In both approaches, as in the case of most real clinical data, it was imperative to deal with the high percentage of missing values and class imbalance. To tackle these problems, the standard way that each classifier deals with the missing values was applied and the SMOTE algorithm was used whenever the proportion of classes on the evaluation was imbalanced. By means of a unique model, the data were evaluated using a grid search that combined oversampling and multidimensional parameters search in order to determine the best triples for each classifier {*Classifier, Parameters, SMOTE percentage*}. After this search, each one of the triples was tested in 30 repetitions, using different seeds in the 5-fold cross validation. Thus, it was possible to obtain the statistical difference between classifiers computing paired t-tests. The results were analysed from multiple goal perspectives as accuracy, sensitivity, specificity and AUC. Although, these metrics provide a simple way of describing a classifier´s performance, they can be deceiving and are highly sensitive to imbalanced data, so the F-measure metric is optimized during the grid search, being a trade-off between the sensitivity and specificity, not depending on disease prevalence.

In order to accomplish the first goal we used supervised learning methods to predict the conversion of MCI patients to AD, using different set of features both in the First Last approach and in the temporal windows approaches. First, it is important to mention that the reduced cardinality of the dataset impairs the classification analysis, in particular in the four and five temporal window, and consequently the results obtained. Our results showed that the temporal window approach leads to better discriminative results. Both in cross-validation and in the validation results, as long as the temporal window increases, the higher prediction capability of the models was noticed. Concerning the cross-validation results, in the five temporal window, even though we obtained the highest values of F-measure and accuracy, the values of specificity were the lowest, due to the reduced number of noEvol examples in this window. The results obtained in the three and four years temporal window are more reliable, being more suitable to predict conversion with higher predictive capability. These results are not analogous to the ones obtained with the validation set, where the only window that outperforms the results in the First Last approach was the five years temporal window. Such results are most likely due to an overfitting of the training set, which leads to a low generalization of the models, and the fact that the First Last dataset is

composed by more examples than any other temporal window. We concluded that the best models to use are the Naïve Bayes and SVMs classifiers. Our results highlighted the improvement of the accuracy and F-measure of every classifier after application of feature selection and the importance of the use of SMOTE algorithm when the proportion of the classes is imbalanced. For the three temporal window, the best model was obtained using polynomial SVM algorithm, which had an accuracy of 82%, a sensitivity of 78%, a specificity of 86%, an AUC of 0.82 and a F-measure of 0.82. For the four temporal window, the best model was Naïve Bayes, which had an accuracy of 82%, a sensitivity of 82%, a specificity of 81%, a AUC of 0.88 and a F-measure of 0.82. From the results, it is possible to understand the importance of wider time interval to prognosis of conversion of MCI patients to AD. Nevertheless, the study should be replicated in a larger database to corroborate these conclusions.

Our findings validate the growing evidence that MCI is an important clinical entity. In order to improve the predictive capability of conversion from MCI patients to AD, a more reliable approach would be to combine the output of different models. Several machine learning techniques do this by learning an ensemble of models and using them in combination, increasing the predictive performance over any of the constituent learning algorithms. Among these techniques the more prominent schemes are the bagging and the boosting. For instance, the Random Forest algorithm [57] combines random decisions trees with bagging to achieve higher classification accuracy and AdaBoost is the most common implementation of boosting [38].

Follows an overview of the conclusions achieved during the second phase of this work, concerning the development of a classification methodology for MCI subgroups based on clinical information and unsupervised methods, only in the temporal window approach. As already mentioned, the reduced cardinality of the dataset impairs the classification analysis and this is even more evident when separating the MCI group.

Considering the clinical information, the patients were divided according to their state of depression, specified as not depressed or depressed, dictated by the Geriatric Depressive State (GDS) attribute. The prevalence of more than 50% of missing values regarding this attribute was a drawback. The results presented corroborate the hypothesis that the separation of the patients according to their depressive symptoms influences positively the classification and that this data mining approach has potential as a possible strategy for the prognosis prediction of conversion of groups of MCI patients to AD. In future studies, it is expected that the integration of more patients with GDS information will allow for a higher prediction capability separating the patients with different characteristics. We concluded that the best models to use are the Naïve Bayes, *k*NN and polynomial SVM classifiers. The temporal windows with the highest discriminative power are the three and five temporal windows. For the three temporal window, when a cross-validation evaluation is applied to the restrictive datasets ($\mathfrak{D}_{\mathbf{0-4}}$ and $\mathfrak{D}_{\mathbf{5-14}}$) the best models allowed to obtain an accuracy of 84%, a sensitivity of 77%, a specificity of 91% and a F-measure of 0.84. For the five temporal window, the best models together allowed to obtain an accuracy of 86%, a sensitivity of 89%, a specificity of 67% and a F-measure of 0.86.

Regarding the formation of MCI subgroups through unsupervised leaning, EM and K-Means were the clustering algorithms applied, with distinct predefined number of clusters (k=2,3,4) to evaluate the stability of the clusters whenever the k-value was changed. As aforementioned, the identification of the ideal number of clusters is a demanding task. It is important to refer that the cluster analyses were performed using different set of features since in this kind of analyses is important to consider the relative scales of the variables being measured. Even though, we wanted to analyse the prognosis problem considering three automatic groups, due to the fact that the number of patients per cluster is reduced invalidating the learning of the models, it was decided

to consider two clusters in the prognosis problem, in every temporal window. However, the results did not outperform the classification approach where all MCI are assumed to evolve similarly. Notwithstanding, the diverse experiments performed concerning the evolutionary behaviour of the achieved subgroups seem to enhance the importance of the study of MCI subgroups considering the putative differences of this clinical entity. In the classification of the MCI subgroups is even more relevant to adopt a new approach by learning an ensemble of models and using them in combination.

Our results and recent research [26, 28, 53] using cluster analysis techniques have shown that MCI cohorts present heterogeneous cognitive profiles so other similarity clustering algorithms, such as hierarchical clustering and spectral clustering, should be applied to corroborate such assumptions. Other strategy to try in future analyses is the feature selection for unsupervised learning [14]. As already referred, some of the features may be redundant, others may be irrelevant. Since some unsupervised learning algorithms have higher difficulties dealing with high dimensional data misguiding clustering results, it is important to reduce the number of features which will enrich the problem. As we conclude, feature selection algorithms maximize the predictive accuracy being related to the class labels. However, in unsupervised learning we are not given the class labels. So in order to avoid the bias of the results, it would be interesting to perform a feature selection for unsupervised learning along with the application of more complex clustering techniques than the ones used in this study.

The results demonstrated that future studies should give priority to differentiate MCI subgroups using clinical knowledge. Besides the Geriatric Depression Scale (GDS), Clinical Dementia Rating scale (CDR) is starting to assume a relevant role in the conversion to dementia [29, 31, 13, 19] and thus it would be important to obtain a more complete database regarding this attribute to confirm such suppositions.

Besides the separation in MCI groups considering their putative differences, some MCI patients revert back to normal [4,22]. However, this reversion has not been fully characterized. In particular, the rates of reversion to normal may vary across the studies due to the differences in the sources of the study participants [4,22]. Even though reversion to normal occurs, many subjects subsequently progress back to MCI or Dementia. Therefore subjects who revert may have pathologies that lead to cognitive impairment. These subjects may be candidates for clinical trials to reduce their risk of future progression. It would be interesting to analyse the patients in this condition. The instances labelled as normal (removed in the preprocessing steps) should be analysed in detail in order to capture the characteristics behind the reversion. It is also important to have in consideration that reversion to normal may also occur because MCI is a clinical diagnosis, which is inherently characterized by the variability in the subject's performance, the caregiver's appraisal, the interaction between subject and clinician and the possible diagnostic error [4].

Statistical evaluation of experimental results has been considered an essential part of validation of machine learning methods [46]. In this present thesis, paired t-tests were used for the comparison of the classifiers in each dataset. However, as aforementioned, it would also be important to use a statistical test to compare the algorithms along the datasets in order to determine if the increase in the F-measure values along the windows is statistically significant. Paired t-tests are not appropriated to this kind of analysis, since these tests only make sense whenever the differences over the multiple data sets are commensurate and require that the differences between the two random variables compared are distributed normally [46]. As alternative to the paired t-tests, non-parametric tests as the Wilcoxon and the Friedman tests are suitable for these problems. These tests compensate the limitations of the paired t-tests [46].

7. References

[1] Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T. and Jin, J.S. (2011) Identification of Conversion from Mild Cognitive Impairment to Alzheimer's Disease Using Multivariate Predictors, PlosOne, 6(7), e21896.

[2] Espinosa, A., Alegret, M., Valero, S., Vinyes-Junqué, G., Hernández, I., Mauléon, A., Rosende-Roca, M., Ruiz, A., López, O., Tárraga, L. and Boada, M. (2013) A Longitudinal Follow-Up of 550 Mild Cognitive Impairment Patients: Evidence for Large Conversion to Dementia Rates and Detection of Major Risk Factors Involved, Journal of Alzheimer's Disease 34, 769–780.

[3] Hinrichs, C., Singh, V., Xu, G., and Johnson, S.C. (2011) Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population, NeuroImage 55, 574–589.

[4] Sachdev, P.S., Lipnicki, D. M., Crawford, J., Reppermund, S., Kochan, N. A., Trollor, J. N., Wen, W., Draper, B., Slavin, M. J., Kang, K., Lux, O., Mather, K.A. and Brodaty, H., (2013) Factors Predicting Reversion from Mild Cognitive Impairment to Normal Cognitive Functioning: A Population-Based Study, PlosOne, 8(3), e59649.

[5] Jonsson, T., Atwal, J. K., Steinberg, S.,Snaedal, J., Jonsson, P.V., Bjornsson, S.,Stefansson, H., Sulem, P., Gudbjartsson, D., Maloney, J., Hoyte, K., Gustafson, A., Liu, Y., Lu, Y., Bhangale, T.,Graham, R.R., Huttenlocher, J., Bjornsdottir, G., Andreassen. O.A., Jönsson, E.G., Palotie, A., Behrens, T.W., Magnusson, O.T., Kong, A., Thorsteinsdottir, U.,Watts, R.J. and Stefansson, K. (2012), A mutation in APP protects against Alzheimer's disease and age-related cognitive decline, Research Letter, 488, 96.

[6] Prestia, A., Carolia, A., Herholz, K., Reimand, E., Chend, K., Jaguste, W. J., Frisonia and G. B. Frisoni, (2013) Diagnostic accuracy of markers for prodromal Alzheimer's disease in independent clinical series, Alzheimer's & Dementia, 1-10.

[7] Liu, F., Wee, C., Chen, H. and Shen, D. (2014) Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's Disease and mild cognitive impairment identification, NeuroImage, 84, 466-475.

[8] Chapman, R. M., Mapstone, M., McCrary, J. W., Gardner, M. N., Porsteinsson, A., Sandoval, T.C., Guillily, M. D., DeGrush, E., and Reilly, L.A. Predicting conversion from mild cognitive impairment to Alzheimer's disease using neuropsychological tests and multivariate methods, (2011) Journal of Clinical Experimental Neuropsychology 33, 187-199.

[9] Ewers, M., Walsh, C., Trojanowski, J. Q., Shaw, L. M., Petersen, R. C., Jack Jr., C. R., Feldman, H. H., Bokde, A. L. W., Alexander, G. E., Scheltens, P., Vellas, B., Dubois, B., Weinera, M. and Hampel, H. (2012) Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance, Neurobiology of Aging, 33, 1203-1214.

[10] Li, S., Okonkwo, O., Albert, M. and Wang, M. (2013) Variation in Variables that Predict Progression from MCI to AD Dementia over Duration of Follow-up, Am J Alzheimer Dis (Columbia) 2, 12-28.

[11] Yu, L., Boyle, P., Wilson, R. S., Segawa, E., Leurgans, S., Jager, P. L. and Bennett, D. A. (2012), A Random Change Point Model for Cognitive Decline in Alzheimer's Disease and Mild Cognitive Impairment, Neuroepidemiology, 39, 73-83.

[12] Jack Jr, C. R., Wiste, H. J., Vemuri, P., Weigand, S. D., Senjem, M. L., Zeng, G., Bernstein, M. A., Gunter, J. L., Pankratz, V. S., Aisen, P. S., Weiner, M. W, Petersen, R. C., Shaw, L. M., Trojanowski, J. Q. and Knopman, D. S., (2010) Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to Alzheimer's disease, Brain-A Journal of Neurology, 133, 3336-3348.

[13] Kang, L., Xiong, C., Crane, P. and Tian, L. (2013) Linear combinations of biomarkers to improve diagnostic accuracy with three ordinal diagnostic categories, Statistics in Medicine, 32, 631-643.

[14] Dy, J.G. and Brodley, C.E., (2004) Feature Selection for Unsupervised Learning, Journal of Machine Learning Research, 5, 845–889.

[15] Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I. and Mendonça, A. (2011) Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests, BMC Research Notes, 4:299.

[16] Silva, D., Guerreiro, M., Maroco, J., Santana, I., Rodrigues, A., Marques, J. B. and Mendonça, A. (2012) Comparison of Four Verbal Memory Tests for the Diagnosis and Predictive Value of Mild Cognitive Impairment, Dementia and Geriatric Cognitive Disorders Extra, 2, 120-131.

[17] Maroco, J., Silva, D., Guerreiro, M., Mendonça, A. and Santana, I. (2011) Prediction of dementia patients: a comparative approach using parametric vs. non parametric classifiers, XIX Congresso Anual da Sociedade Portuguesa de Estatística.

[18] Dubois, B., (2010) Revising the definition of Alzheimer's disease: a new lexicon, Lancet Neurol, 9, 1118–27.

[19] Peterson, R. C., (2004) Mild Cognitive Impairment as a diagnostic entity, Journal of Internal Medicine, 256, 183-194.

[20] Peterson, R. C., (2009) Mild Cognitive Impairment, Ten years Later, Arch Neurol, 66, 1447-1455.

[21] Kim, S. Y., Lim, T. S., Lee, H. Y. and Moon, S. Y. (2014) Clustering mild cognitive impairment by mini-mental state examination, Neurol Sci.

[22] Roberts, R. O., Knopman, D. S., Mielke, M. M., Cha,R.H., Pankratz,V.S., Christianson, T.J.H., Geda,Y.E., Boeve,B.F., Ivnik, R.J.,Tangalos, E.G., Rocca,W.A. and Petersen, R.C. (2013) Higher risk of progression to dementia in MCI cases who convert to normal , Neurology, 82, 1-9.

[23] Sá. F., Pinto, P., Cunha, C., Lemos, R., Letra, L., Simões, M. and Santana, I., (2012) Differences between early and late-onset Alzheimer's disease in neuropsychological tests, Frontiers in neurology, volume 3, article 81.

[24] Sloane, P. D., Zimmerman, S., Suchindran, C., Reed, P., Wang, L., Boustani, M., and Sudha, S. (2002) The Public Health Impact of Alzheimer's Disease, 2000–2050: Potential Implication of Treatment Advances, Annu. Rev. Public Health, 23, 213–31.

[25] Mathuranath, P.S., Nestor, P. J., Berrios, G. E., Rakowicz, W., and Hodges, J. R. (2000) A brief cognitive test battery to differentiate Alzheimer's disease and frontotemporal dementia, Neurology, 55, 1613-1620.

[26] Bondi, M.W., Edmonds, E.C., Jak, A.J., Clark, L.R., Delano-Wood, L., McDonald, C.R., Nation, D.A., Libon, D.J, Au, R., Galasko, D. and Salmon, D.P., (2014) Neuropsychological Criteria for Mild Cognitive Impairment Improves Diagnostic Precision, Biomarker Associations, and Progression Rates, Journal of Alzheimer's Disease, 20, 1-15.

[27] McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D. and Stadlan, E.M. (1984) Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease, Neurology, 34, 939-944.

[28] Hessen, E., Reinvang, I., Eliassen, C.F., Nordlund, A., Gjerstad, L., Fladby, T. and Wallin, A., (2014), The Combination of Dysexecutive and Amnestic Deficits Strongly Predicts Conversion to Dementia in Young Mild Cognitive Impairment Patients: A Report from the Gothenburg-Oslo MCI Study, Dementia and Geriatric Cognitive Disorders Extra 4, 76-85.

[29] Young, A.L., Oxtoby, N.P, Daga, P. and Cash, D.M., on behalf of the Alzheimer's Disease Neuroimaging Initiative, Fox, N.C, Ourselin, S., Schott, J.M. and Alexander, D.C, (2014) A data-driven model of biomarker changes in sporadic Alzheimer's disease, Brain-A Journal of Neurology, 1-14.

[30] Petersen, R.C., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V. and Fratiglioni, L., (2014) Mild cognitive impairment: a concept in evolution, KeySymposium, Journal of International Medicine, 275, 214–228.

[31] Seixas, F.L., Zadrozny, B., Laks, J., Conci, A. and Saade, D.C.M, (2014) A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment, Elsevier, Computers in Biology and Medicine, 51, 140–158.

[32] Forman, G., An Extensive Empirical Study of Feature Selection Metrics for Text Classification, (2003), Journal of Machine Learning Research, 3, 1289-1305.

[33] He, H. and Garcia, E.A, (2009), Learning from Imbalanced Data, IEEE Transactions on knowledge and Data Engineering, 21(9), 1263-1284.

[34] Singhal, S. and Jena, M., A Study on WEKA Tool for Data Preprocessing, Classification and Clustering, (2013), International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN, Volume-2, Issue-6, 2278-3075.

[35] Farhangfar, A., Kurgan, L. and Dy, J., (2008), Impact of imputation of missing values on classification error for discrete data, Pattern Recognition, 41, 3692-3705.

[36] Gibert, K., Izquierdo, J., Holmes, G., Athanasiadis, J., Comas, J. and Sànchez-Marrè, M., On the role of pre and post-processing in environmental data mining, (2008), International Environmental Modelling and Software Society (iEMSs), 1937-1958.

[37] Vercellis, C., (2009), Business Intelligence: Data Mining and Optimization for Decision Making, John Wiley & Sons, Ltd.

[38] Witten, I. H., Frank, E. and Hall, M. A., (2001), Data Mining: Pratical Machine Learning Tools and Techniques, Elsevier Inc, 3rd Edition.

[39] Cunningham, P. and Delany, S.J., (2007), k-Nearest Neighbour Classifiers, Technical Report UCD-CSI

[40] Rokach, L., Maimon, O., Data Mining and Knowledge Discovery Handbook, (2010) Springer Science+Business Media, 2nd Edition.

[41] Quinlan, J.R., (1986), Induction of Decision Trees, Kluwer Academic Publishers, Boston, Machine Learning 1, 81-106.

[42] Raileanu, L.E. and Stoffel, K., (2004), Theoretical comparison between the Gini Index and Information Gain criteria, Kluwer Academic Publishers, Annals of Mathematics and Artificial.

[43] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., (2002), SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 16, 321–357.

[44] Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.J., Petersen, R.C., Snyder, P.J., Carrillo, M.C., Bill Thies and Phelps, C.H. (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimer's & Dementia, 7, 270–279.

[45] Hessler, J., Tucha, O., Förstl, H., Mösch, E. and Bickel, H., Age-Correction of Test Scores Reduces the Validity of Mild Cognitive Impairment in Predicting Progression to Dementia, PLOS ONE, 9(8), e106284.

[46] Demsar, J., (2006) Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research, 7, 1–30.

[47] Plant, C., Teipel, S. J., Oswald, A., Böhm, C., Meindl, T., Mourao-Miranda, J., Bokde, A. W., Hampeld, H. and Ewers, M., (2010) Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease, *Neuroimage*, 50-1, 162–174.

[48] Lemos, L. J. M., A data mining approach to predict conversion from mild cognitive impairment to Alzheimer's Disease (2012), Master degree, Instituto Superior Técnico.

[49] Molinuevo, J.L., Blennow, K., Dubois, B., Engelborghs, S., Lewczuk, P., Perret-Liaudet, A., Teunissen, C.E. and Parnetti, L., (2014) The clinical use of cerebrospinal fluid biomarker testing for Alzheimer's disease diagnosis: A consensus paper from the Alzheimer's Biomarkers Standardization Initiative, ELSEVIER, Alzheimer's & Dementia, 1-10.

[50] Visser, P.J., Wolf, H., Frisoni, G. and Gertz, H., Disclosure of Alzheimer's disease biomarker status in subjects with mild cognitive impairment (2012) Biomarkers Med, 6, 365-368.

[51] Bocchetta, M. et al, Galluzzi, S., Kehoe, P.G., Aguera, E.,Bernabei, R., Bullock, R., Ceccaldi, M., Dartigues, J.,Mendonça, A., Didic, M., Eriksdotter, M., Félician, O., Frölich, L., Gertz, H., Hallikainen, M., Hasselbalch, S.G., Hausner, L., Heuser, I., Jessen, F., Jones, R.W., Kurz, A., Lawlor, B., Lleo, A.,Martinez-Lage, P., Mecocci, P., Mehrabian, S., Monsch, A., Nobili, F.,Nordberg, A., Rikkert, M. O., Orgogozo, J., Pasquier, F.,Peters, O., Salmon, E., Sánchez-Castellano, C., Santana, I., Sarazin, M., Traykov, L, Tsolaki, M., Visser, P.J., Wallin, Å.K., Wilcock, G.,Wilkinson, D., Wolf, H., Yener, G., Zekry, D. and Frisoni, G.B. (2014) The use of biomarkers for the etiologic diagnosis of MCI in Europe: An EADC survey, ELSEVIER, Alzheimer's & Dementia, 1-12.

[52] A.J. Larner, (2013) Comparing Diagnostic Accuracy of Cognitive Screening Instruments: A Weighted Comparison Approach, Dementia Geriatric Cognitive Disorder Extra, 3, 60-65.

[53] Edmonds, E.C., Delano-Wood, L., Galasko, D.R., Salmon, D.P. and Bondi, M.W., (2014) Subjective Cognitive Complaints Contribute to Misdiagnosis of Mild Cognitive Impairment, Journal of International Neuropsychological Society, 20, 1-12.

[54] Eckerström, C., Olsson, E., Klasson, N., Berge, J., Nordlund, A., Bjerke, M. and Wallin, A., (2014) Multimodal Prediction of Dementia with up to 10 Years Follow Up: The Gothenburg MCI Study, Journal of Alzheimer's Disease, 1-10.

[55] Facal, D., Guàrdia-Olmos and Juncos-Rabadán, J., (2014) O. Diagnostic transitions in mild cognitive impairment by use of simple Markov models, International Journal of Geriatric Psychiatric.

[56] Moradi, E., Pepe, A., Gaser, C., Huttunen, H. and Tohka, J., (2014) Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects, NeuroImage, 1-34.

[57] Lebedev, A.V., Westman, E., VanWesten, G.J.P.,Kramberger, M.G, Lundervold, A., Aarsland, D., Soininen, H., Kłoszewska, I., Mecocci, P., Tsolaki, M., Vellas, B., Lovestone, S. and Simmons, A. (2014) Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness, NeuroImage, 6, 115-125.

[58] Lopez, O.L, Jagust, W.J, DeKosky, S.T, Becker, J.T, Fitzpatrick, A., Dulberg, C., Breitner, J., Lyketsos, C., Jones, B., Kawas, C., Carlson, M. and Kuller, L.H. (2003) Prevalence and classification of mild cognitive impairment in the Cardiovascular Health Study Cognition Study: Part , Arch Neurol, 60, 1385-1389.

[59] Kohavi, R., (1995), A Study of Cross-Validation and Boostrap for Accuracy Estimation and Model Selection, International Joint Conference on Artificial Inteligence (IJCAI).

A. Appendix Medical Exams

Feature	Type	Description
Cases_with_follow_up	Numeric	Type of cases
Observation_number_for_each_case	Numeric	Sequential count of the same case by observation
Case_number_for_this_database	Numeric	Rank of Cases
Age	Numeric	Age at evaluation
DiagNPS	String	Diagnosis from Psychologist
Diagnosis_code	Numeric	Neuropsychological and clinical Diagnosis
end_point	Numeric	Diagnosis on last assessment (or assessment of change)
evolution_type	Numeric	Type of evolution found between observations
time_assessmet1_endpoint	Numeric	time between 1st assessment and assessment of change of diagnosis (or last assessment)
time_begin_endpoint	Numeric	time between begin of symptoms and the assessment of diagnosis change (or last assessment)
Neuropsy_assessment_clinical	Numeric	Type of assessment (clinical or neuropsychological)
Date	Date	Date of the evaluation
School	Numeric	Years of formal education
Group	Numeric	Group in BLAD controls
Gender	Numeric	Gender
As_tot	Numeric	BLAD - Letter Cancellation (A) - TOTAL
DS_forw	Numeric	BLAD - Digit Span forward
DS_back	Numeric	BLAD - Digit Span backward
DS_tot	Numeric	BLAD - Digit Span TOTAL
PA_Tot	Numeric	BLAD - Verbal Paired-Associate Learning - TOTAL
LM_a	Numeric	BLAD - Logical Memory - A
LM_b	Numeric	BLAD - Logical Memory -B
LM_tot	Numeric	BLAD - Logical Memory - TOTAL
LM_a_Cued	Numeric	BLAD - Logical Memory - A - Cued
LM_b_Cued	Numeric	BLAD - Logical Memory - B - Cued
LM_a_Interf	Numeric	BLAD - Logical Memory with Interference - A
LM_b_Interf	Numeric	BLAD - Logical Memory with Interference - B
LM_tot_Interf	Numeric	BLAD - Logical Memory with Interference - TOTAL
LM_a_Interf_Cued	Numeric	BLAD - Logical Memory with Interference A - Cued
LM_b_Interf_Cued	Numeric	BLAD - Logical Memory with Interference B - Cued
Forgetting	Numeric	Forgetting Index= [(LM delayed recall – LM immediate) / LM immediate)] * 100
MVI_Tot	Numeric	BLAD - Word Recall with Interference - TOTAL
Infor	Numeric	BLAD - Test about General Information
VisualM_A	Numeric	BLAD - Visual Memory A
VisualM_B	Numeric	BLAD - Visual Memory B
VisualM_C1	Numeric	BLAD - Visual Memory C1
VisualM_C2	Numeric	BLAD - Visual Memory C2
VisualM_total	Numeric	BLAD - Visual Memory - TOTAL
Orient_P	Numeric	BLAD - Orientation - Personal
Orient_S	Numeric	BLAD - Orientation - Spatial
Orient_T	Numeric	BLAD - Orientation - Temporal
Or_Total	Numeric	BLAD - Orientation - TOTAL
Fluency_Sem	Numeric	BLAD - Verbal Fluency
Fluency_Phon	Numeric	BLAD - Phonologic Fluency
M_Initiative	Numeric	BLAD - Motor Initiative
Gm_Initiative	Numeric	BLAD - Grafomotor Initiative
Writing	Numeric	BLAD - Writing

Ident	Numeric	BLAD - Objects Identification
Token_T	Numeric	BLAD - Token Orders - TOTAL
Naming	Numeric	BLAD - Naming
Repetition	Numeric	BLAD - Repetition
Prxs	Numeric	BLAD - Motor Coordenation
Cube	Numeric	BLAD - draw of a cube
Clock	Numeric	BLAD - draw of a clock
Calc	Numeric	BLAD - Calculation
MPR	Numeric	BLAD - Raven Progressive Matrices
Proverb	Numeric	BLAD - Verbal Abstraction
Token_Complete	Numeric	Complete version of Token
Snodgrass_clinical	Numeric	Snodgrass and Vanderwart - clinical impression
Public_Faces_clinical	Numeric	Public Faces - clinical impression
MMSE	Numeric	Mini-Mental State Examination
TPRT	Numeric	Toulouse-Pierón (Rendimento de Trabalho)
TPID	Numeric	Toulouse-Pierón (Índice de Dispersão)
TMT_A_temp	Numeric	Trail Making Test (Part A) - Tempo
TMT_A_err	Numeric	Trail Making Test (Part A) - Erros
TMT_B_temp	Numeric	Trail Making Test (Part B) - Tempo
TMT_B_err	Numeric	Trail Making Test (Part B) - Erros
TMTs_incomplete	Numeric	Trail Making Test (Part B) - Incompletos
CVLT_a1	Numeric	California Verbal Learning Test - Lista A - 1 evocação
CVLT_a2	Numeric	California Verbal Learning Test - Lista A - 2 evocação
CVLT_a3	Numeric	California Verbal Learning Test - Lista A - 3 evocação
CVLT_a4	Numeric	California Verbal Learning Test - Lista A - 4 evocação
CVLT_a5	Numeric	California Verbal Learning Test - Lista A - 5 evocação
CVLT_a1a5	Numeric	California Verbal Learning Test - Lista A de 1 a 5 Total
CVLT_a_pers	Numeric	California Verbal Learning Test - Lista A Perseverações
CVLT_a_intr	Numeric	California Verbal Learning Test - Lista A Intrusões
CVLT_b_tot	Numeric	California Verbal Learning Test - Lista B Total
CVLT_b_pers	Numeric	California Verbal Learning Test - Lista B Perseverações
CVLT_b_intr	Numeric	California Verbal Learning Test - Lista B Intrusões
CVLT_b_cs	Numeric	California Verbal Learning Test - Lista B CS
CVLT_a_cr_int	Numeric	California Verbal Learning Test - Lista A - Evocação espontânea após curto intervalo
CVLT_a_crint_pers	Numeric	California Verbal Learning Test - Ev. esp. curto intervalo - perseverações
CVLT_a_crint_intr	Numeric	California Verbal Learning Test - Ev. esp. curto intervalo - intrusões
CVLT_a_crint_cs	Numeric	California Verbal Learning Test - Ev. esp. curto intervalo - CS
CVLT_a_crint_ajsem	Numeric	California Verbal Learning Test - Evocação após curto intervalo com ajuda semântica
CVLT_a_crint_ajsem_pers	Numeric	California Verbal Learning Test - Ev. curto intervalo com ajuda semântica - perseverações
CVLT_a_crint_ajsem_intr	Numeric	California Verbal Learning Test - Ev. curto intervalo com ajuda semântica - intrusões
CVLT_a_lg_int	Numeric	California Verbal Learning Test - Lista A - Evocação após longo intervalo
CVLT_a_lgint_pers	Numeric	California Verbal Learning Test - Ev. esp. longo intervalo

		- perseverações
CVLT_a_lgint_intr	Numeric	California Verbal Learning Test - Ev. esp. longo intervalo - intrusões
CVLT_a_lgint_cs	Numeric	California Verbal Learning Test - Ev. esp. longo intervalo - CS
CVLT_a_lgint_ajsem	Numeric	California Verbal Learning Test - Evocação após longo intervalo com ajuda semêntica
CVLT_a_lgint_ajsem_pers	Numeric	California Verbal Learning Test - Ev. longo intervalo com ajuda semântica - perseverações
CVLT_a_lgint_ajsem_intr	Numeric	California Verbal Learning Test - Ev. longo intervalo com ajuda semântica - intrusões
CVLT_rec_a	Numeric	California Verbal Learning Test - Reconhecimento após longo intervalo - Lista A
CVLT_rec_Bp	Numeric	California Verbal Learning Test - Reconhecimento - Lista B partilhados
CVLT_rec_Bn	Numeric	California Verbal Learning Test - Reconhecimento - Lista B não partilhados
CVLT_rec_P	Numeric	California Verbal Learning Test - Reconhecimento - Protótipo
CVLT_rec_sr	Numeric	California Verbal Learning Test - Reconhecimento - sem relação
GDS_1	Numeric	
GDS_2	Numeric	
GDS_3	Numeric	
GDS_4	Numeric	
GDS_5	Numeric	
GDS_6	Numeric	
GDS_7	Numeric	
GDS_8	Numeric	
GDS_9	Numeric	
GDS_10	Numeric	
GDS_11	Numeric	
GDS_12	Numeric	
GDS_13	Numeric	
GDS_14	Numeric	
GDS_15	Numeric	
GDS_total	Numeric	Geriatric Depression Scale
QSM_1	Numeric	Queixas Subjectivas de Memória
QSM_2	Numeric	
QSM_3	Numeric	
QSM_4	Numeric	
QSM_5	Numeric	
QSM_6	Numeric	
QSM_7	Numeric	
QSM_8	Numeric	
QSM_9	Numeric	
QSM_10	Numeric	
QSM_total	Numeric	Escala de Queixas Subjectivas de Memória
Blessed1	Numeric	Blessed (itens)
Blessed2	Numeric	
Blessed3	Numeric	

Blessed4	Numeric	
Blessed5	Numeric	
Blessed6	Numeric	
Blessed7	Numeric	
Blessed8	Numeric	
BlessedAVD	Numeric	Blessed (Total of Part 1 - Daily living activities)
Blessed9	Numeric	
Blessed10	Numeric	
Blessed11	Numeric	
BlessedHAB	Numeric	Blessed (Total of Part 2 - Habits)
Blessed12	Numeric	
Blessed13	Numeric	
Blessed14	Numeric	
Blessed15	Numeric	
Blessed16	Numeric	
Blessed17	Numeric	
Blessed18	Numeric	
Blessed19	Numeric	
Blessed20	Numeric	
Blessed21	Numeric	
Blessed22	Numeric	
BlessedPERS	Numeric	Blessed (Total of Part 3 - Personality)
BlessedTOT	Numeric	Blessed TOTAL
CancellationTask_Z	Numeric	Z - Scores
DigitSpan_Z	Numeric	Z - Scores
DigitSpan_forward_Z	Numeric	Z - Scores
DigitSpan_backward_Z	Numeric	Z - Scores
SemanticFluency_Z	Numeric	Z - Scores
MotorInitiative_Z	Numeric	Z - Scores
GraphomotorInitiative_Z	Numeric	Z - Scores
Identification_Z	Numeric	Z - Scores
Token_Z	Numeric	Z - Scores
Token_Complete_Z	Numeric	Z - Scores
Naming_Z	Numeric	Z - Scores
Repetition_Z	Numeric	Z - Scores
Writing_Z	Numeric	Z - Scores
Orientation_Z	Numeric	Z - Scores
WordRecall_Z	Numeric	Z - Scores
GeneralInformation_Z	Numeric	Z - Scores
VerbalPaired_AssociateLearning_Z	Numeric	Z - Scores
LogicalMemory_Z	Numeric	Z - Scores
LogicalMemory_A_Z	Numeric	Z - Scores
LM_DR_Z	Numeric	Z - Scores
Forgetting_Z	Numeric	Z - Scores
VisualMemory_Z	Numeric	Z - Scores
Cube_Z	Numeric	Z - Scores

Clock_Z	Numeric	Z - Scores
Calculation_Z	Numeric	Z - Scores
MPR_Z	Numeric	Z - Scores
Proverbs_Z	Numeric	Z - Scores
TP_RT_Z	Numeric	Z - Scores
TP_ID_Z	Numeric	Z - Scores
TMT_A_Z	Numeric	Z - Scores
TMT_B_Z	Numeric	Z - Scores
A1_Z	Numeric	Z - Scores
A5_Z	Numeric	Z - Scores
Atot_Z	Numeric	Z - Scores
B_Z	Numeric	Z - Scores
SDFR_Z	Numeric	Z - Scores
SDCR_Z	Numeric	Z - Scores
LDFR_Z	Numeric	Z - Scores
LDCR_Z	Numeric	Z - Scores
REC_Z	Numeric	Z - Scores

Table A 1 Feature List.

B. Complementary Results: Chapter 4

First Last Approach

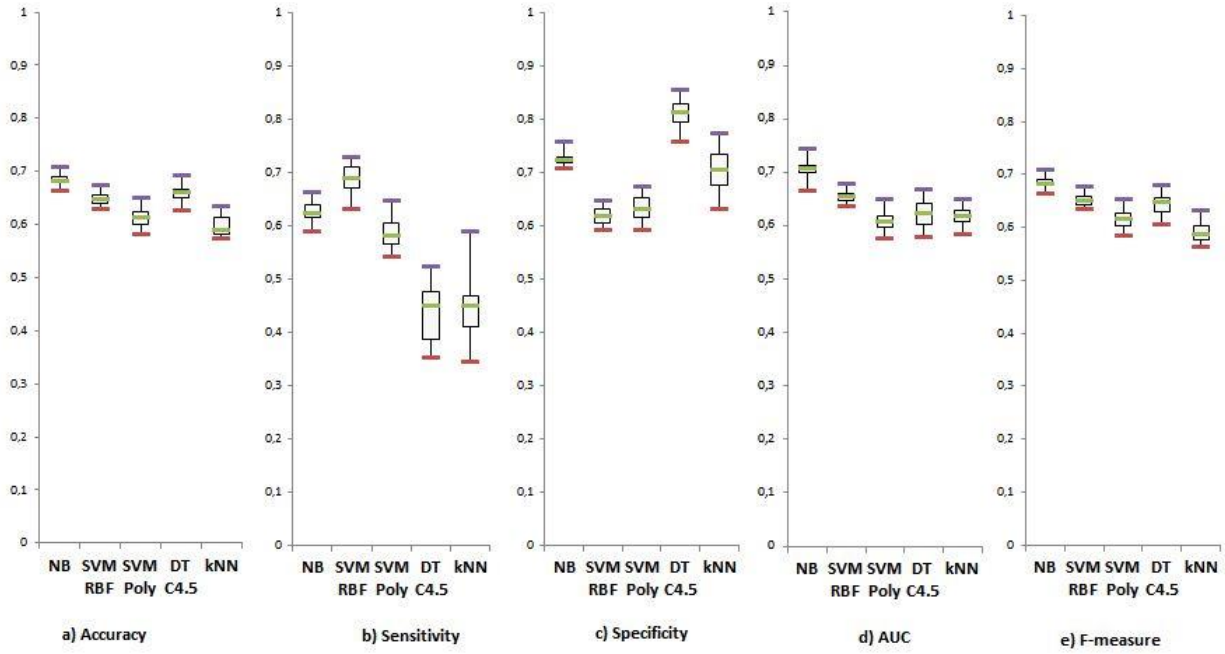

Figure B 1 Train results of Prognosis after applying SMOTE using First Last approach.

	Feature Selection	SMOTE (%)	Parameters
Naïve Bayes	Original	0	*Kernel*
	Correlation	0	*Gaussian*
SVM Poly	Original	0	*Compl*=0.5 and *Exp*=1
	Correlation	0	*Compl*=3.0 and *Exp*=1
SVM RBF	Original	0	*Compl*=4.0 and Υ=0.01
	Correlation	0	*Compl*=3.5 and Υ=0.1
Decision Tree C4.5	Original	0	*Conf*=0.05
	Correlation	0	*Conf*=0.1
kNN	Original	0	*k*=1
	Correlation	0	*k*=9

Table B 1 Classification model parameters for the prognosis using First Last approach.

Two Years Temporal Window

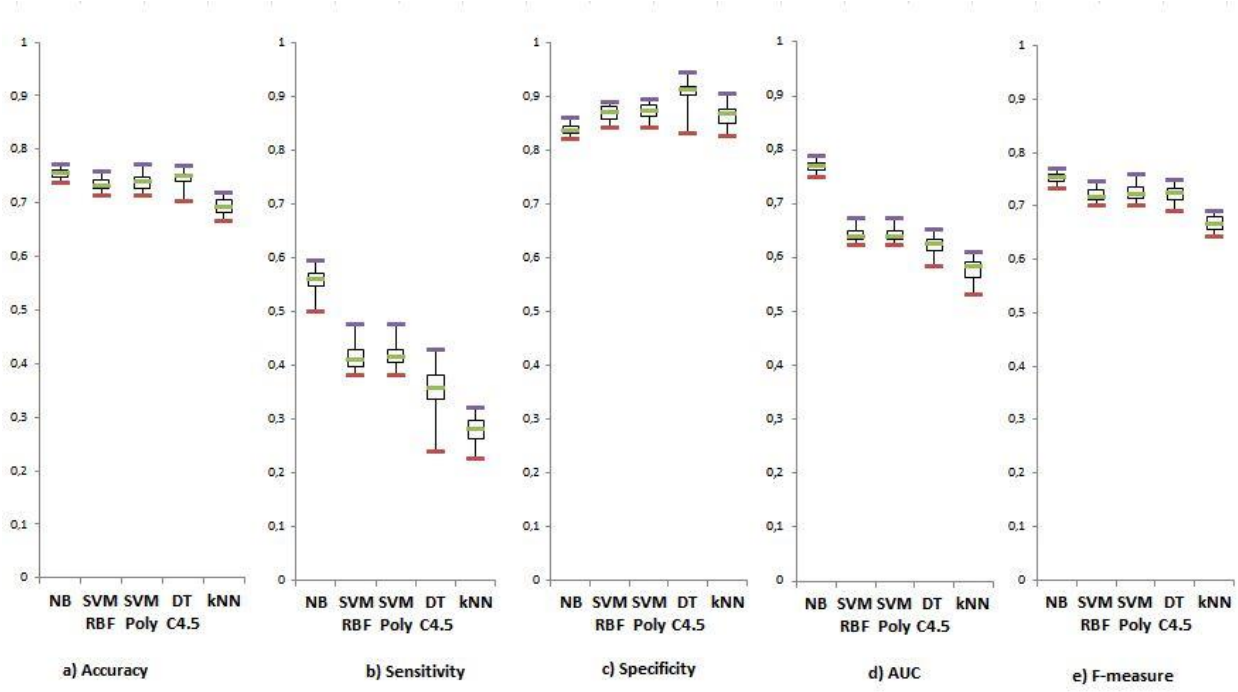

Figure B 2 Train results of Prognosis without applying SMOTE using two years temporal window.

	Feature Selection	SMOTE (%)	Parameters
Naïve Bayes	Original	0	*Kernel*
	Correlation	448	*Supervised Discretization*
SVM Poly	Original	0	*Compl*=0.5 and *Exp*=1
	Correlation	128	*Compl*=1.0 and *Exp*=2
SVM RBF	Original	67	*Compl*=2.0 and Υ=0.01
	Correlation	128	*Compl*=1.5 and Υ=1.0
Decision Tree C4.5	Original	67	*Conf*=0.3
	Correlation	0	*Conf*=0.45
kNN	Original	603	*k*=2
	Correlation	0	*k*=9

Table B 2 Classification model parameters for the prognosis using two years temporal window.

Three Years Temporal Window

3Years Window	noEvol	Evol
Total instances (%)	139 (54.7%)	115 (45.3%)

Table B 3 Train dataset after applying pre-processing for the three years temporal window.

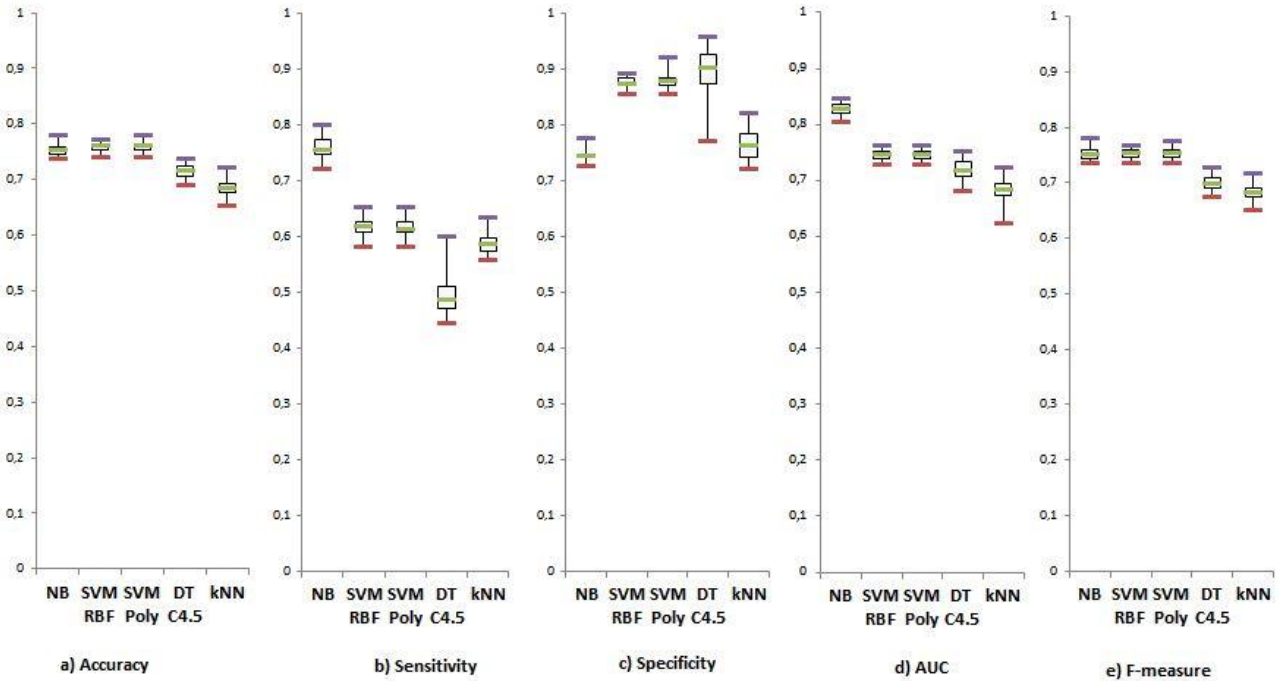

Figure B 3 Train results of Prognosis without applying SMOTE using three years temporal window.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	90(78.3%)	25(21.7%)
	noEvol	38(27.3%)	101(72.7%)

Table B 4 Confusion Matrix of the Naïve Bayes for the three years temporal window, in 5-fold CV.

Classifier	Accuracy	Sensitivity	Specificity	ROC Area	F-measure
NB	0.753±0.011	0.76±0.018	0.747±0.014	0.827±0.010	0.754±0.011

Table B 5 Evaluation metrics of SVM Poly for the three years temporal window, using 5-fold CV.

	Feature Selection	SMOTE (%)	Parameters
Naïve Bayes	Original	0	*Gaussian*
	Correlation	0	*Supervised Discretization*
SVM Poly	Original	0	*Compl*=2.5 and *Exp*=7
	Correlation	0	*Compl*=4.5 and *Exp*=1
SVM RBF	Original	0	*Compl*=2.0 and Υ=0.01
	Correlation	0	*Compl*=1.5 and Υ=1.0
Decision Tree C4.5	Original	0	*Conf*=0.05
	Correlation	0	*Conf*=0.45
kNN	Original	0	*k*=1
	Correlation	0	*k*=9

Table B 6 Classification model parameters for the prognosis using three years temporal window.

Four Years Temporal Window

4Years Temporal Window	noEvol	Evol
Size	85.0 (37.3%)	143.0 (62.7%)

Table B 7 Train dataset after applying pre-processing for the four years temporal window.

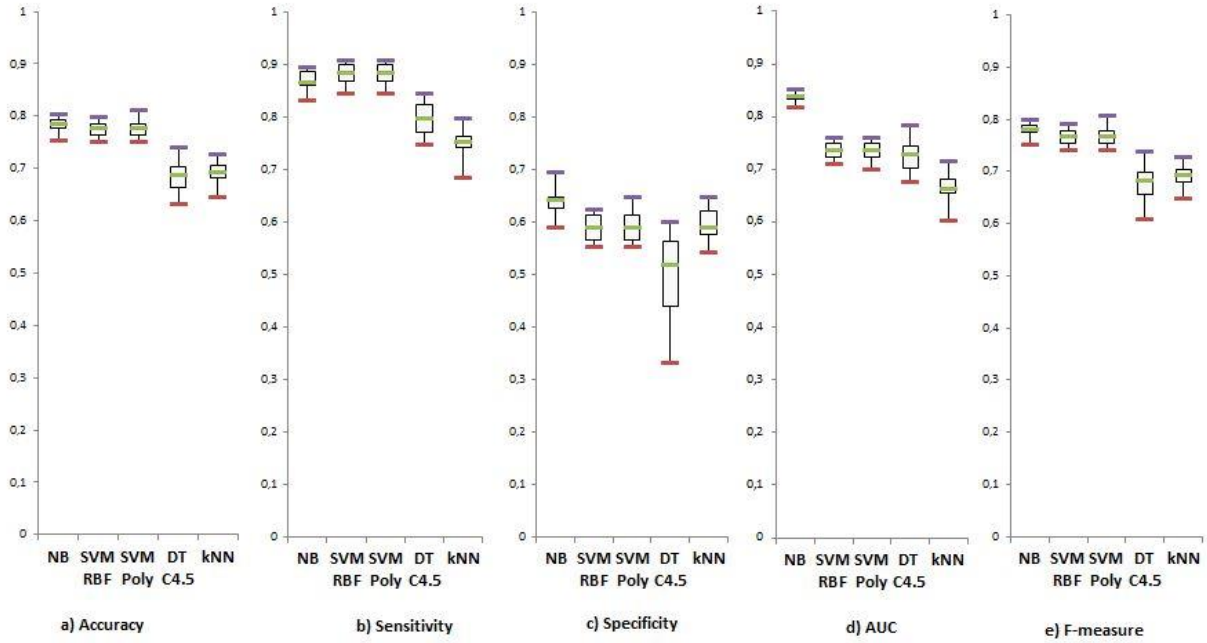

Figure B 4 Train results of Prognosis without applying SMOTE using four years temporal window.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	123(86.1%)	20(13.9%)
	noEvol	32(37.6%)	53(62.4%)

Table B 8 Confusion Matrix of the Naïve Bayes classifier for the four years temporal window, using 5-fold CV.

Classifier	Accuracy	Sensitivity	Specificity	ROC Area	F-measure
Naïve Bayes	0.785±0.012	0.871±0.017	0.64±0.02	0.836±0.007	0.781±0.012

Table B 9 Evaluation metrics of Naïve Bayes for the four years temporal window, using 5-fold CV.

	Feature Selection	SMOTE (%)	Parameters
Naïve Bayes	Original	0	*Kernel*
	Correlation	0	*Kernel*
SVM Poly	Original	0	*Compl*=0.5 and *Exp*=1
	Correlation	0	*Compl*=4.5 and *Exp*=1
SVM RBF	Original	0	*Compl*=2.5 and Υ=0.1
	Correlation	0	*Compl*=4.5 and Υ=0.1
Decision Tree C4.5	Original	0	*Conf*=0.05
	Correlation	0	*Conf*=0.4
kNN	Original	0	*k*=1
	Correlation	0	*k*=9

Table B 10 Classification model parameters for the prognosis using four years temporal window.

Five Years Temporal Window

5Years Temporal Window	noEvol	Evol
Size	51.0 (23.7%)	164.0 (76.3%)

Table B 11 Train dataset after applying pre-processing for the five years temporal window

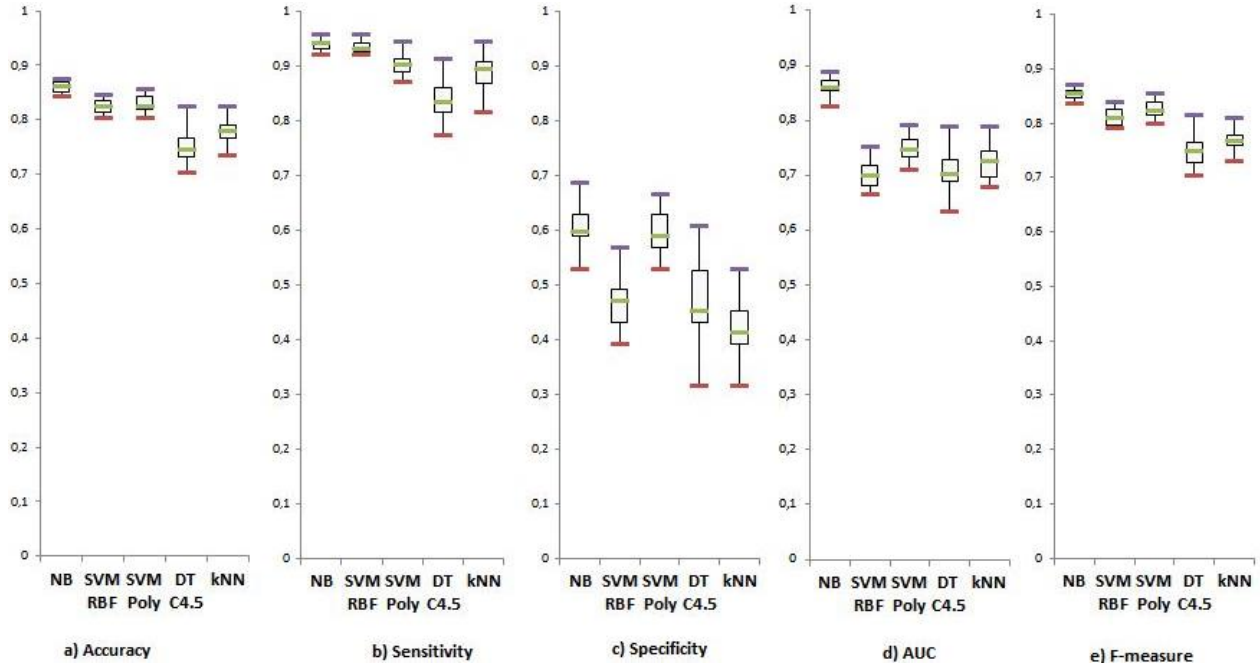

Figure B 5 Train results of Prognosis after applying SMOTE using five temporal window.

		Predicted Class	
		Evol	noEvol
Real Class	Evol	155(94.5%)	9(5.5%)
	noEvol	18(35.3%)	33(64.7%)

Table B 12 Confusion Matrix of the Naïve Bayes for the five years temporal window, using 5-fold CV.

Classifier	Accuracy	Sensitivity	Specificity	ROC Area	F-measure
Naïve Bayes	0.859±0.009	0.939±0.009	0.603±0.036	0.862±0.015	0.854±0.010

Table B 13 Evaluation metrics of Naïve Bayes for the five years temporal window, using 5-fold CV.

	Feature Selection	SMOTE (%)	Parameters
Naïve Bayes	Original	0	*Kernel*
	Correlation	182	*Supervised Discretization*
SVM Poly	Original	78	*Compl*=0.5 and *Exp*=1
	Correlation	52	*Compl*=0.5 and *Exp*=1
SVM RBF	Original	26	*Compl*=3.0 and Υ=0.1
	Correlation	78	*Compl*=5.0 and Υ=0.01
Decision Tree C4.5	Original	78	*Conf*=0.05
	Correlation	52	*Conf*=0.05
kNN	Original	26	*k*=9
	Correlation	0	*k*=10

Table B 14 Classification model parameters for the prognosis using five years temporal window.

FL	2Y	3Y	4Y	5Y
Age	Age	Age	Age	Age
As_tot	As_tot	DS_back	PA_Tot	PA_Tot
LM_tot	PA_Tot	PA_Tot	LM_a_Interf	LM_a
LM_b_Interf	LM_tot	LM_tot	Orient_T	LM_a_Interf
MVI_Tot	LM_a_Interf	LM_a_Interf	Or_Total	Orient_T
Orient_T	MVI_Tot	LM_tot_Interf	Fluency_Sem	Or_Total
Or_Total	Orient_S	MVI_Tot	Proverb	Fluency_Sem
Fluency_Sem	Orient_T	Orient_T	CancellationTask_Z	M_Initiative
M_Initiative	Or_Total	Or_Total	DigitSpan_backward_Z	MPR
Calc	MPR	CVLT_a1a5	SemanticFluency_Z	CVLT_a1
BlessedAVD	CVLT_a1a5	CancellationTask_Z	MotorInitiative_Z	DigitSpan_forward_Z
VerbalPaired_AssociateLearning_Z	BlessedAVD	DigitSpan_Z	Orientation_Z	SemanticFluency_Z
LogicalMemory_A_Z	BlessedTOT	SemanticFluency_Z	GeneralInformation_Z	Orientation_Z
LM_DR_Z	CancellationTask_Z	Orientation_Z	VerbalPaired_AssociateLearning_Z	GeneralInformation_Z
Clock_Z	SemanticFluency_Z	VerbalPaired_AssociateLearning_Z	LogicalMemory_Z	VerbalPaired_AssociateLearning_Z
Atot_Z	Orientation_Z	LogicalMemory_A_Z	LogicalMemory_A_Z	LogicalMemory_Z
	VerbalPaired_AssociateLearning_Z	Cube_Z	LM_DR_Z	LogicalMemory_A_Z
	LogicalMemory_A_Z	Calculation_Z	Cube_Z	LM_DR_Z
	Cube_Z	MPR_Z	MPR_Z	Cube_Z
	MPR_Z	Proverbs_Z	Atot_Z	MPR_Z
	Proverbs_Z			A5_Z
	A1_Z			

Table B 15 Features selected along the temporal windows.

C. Complementary Results: Chapter 5

C.1 Clinical criteria: depressed/ not depressed

Two Years Temporal Window

$\mathfrak{D}_{all}$	$\mathfrak{D}_{0-4}$	$\mathfrak{D}_{5-14}$
PA_Tot	As_tot	School
LM_a_Interf	MPR	LM_tot
Infor	CVLT_a5	LM_a_Interf
Or_Total	CVLT_rec_P	Orient_T
Fluency_Sem	QSM_total	Or_Total
Cube	Blessed7	Cube
MPR	CancellationTask_Z	CVLT_a1
Proverb	GeneralInformation_Z	CVLT_a3
CVLT_a2	MPR_Z	BlessedAVD
CVLT_rec_P	Proverbs_Z	Orientation_Z
BlessedTOT		LM_DR_Z
CancellationTask_Z		Forgetting_Z
Orientation_Z		Cube_Z
WordRecall_Z		A1_Z
GeneralInformation_Z		Atot_Z
VerbalPaired_AssociateLearning_Z		
LogicalMemory_A_Z		
LM_DR_Z		
MPR_Z		
A1_Z		
Atot_Z		

Table C1. 1 Features selected in the different datasets which are differentiated based on GDS values, using two years temporal window.

	Dataset	SMOTE (%)	Parameters
Naïve Bayes	$\mathfrak{D}_{all}$	0	*Gaussian*
	$\mathfrak{D}_{0-4}$	0	*Gaussian*
	$\mathfrak{D}_{5-14}$	0	*Gaussian*
SVM Poly	$\mathfrak{D}_{all}$	0	*Compl*=0.5 and *Exp*=1
	$\mathfrak{D}_{0-4}$	420	*Compl*=3.0 and *Exp*=5
	$\mathfrak{D}_{5-14}$	0	*Compl*=4.0 and *Exp*=1
SVM RBF	$\mathfrak{D}_{all}$	0	*Compl*=3.0 and γ=0.1
	$\mathfrak{D}_{0-4}$	588	*Compl*=4.0 and γ=1.0
	$\mathfrak{D}_{5-14}$	0	*Compl*=2.5 and γ=1.0
+Decision Tree C4.5	$\mathfrak{D}_{all}$	0	*Conf*=0.25
	$\mathfrak{D}_{0-4}$	168	*Conf*=0.2
	$\mathfrak{D}_{5-14}$	0	*Conf*=0.25
KNN	$\mathfrak{D}_{all}$	0	*k*=3
	$\mathfrak{D}_{0-4}$	672	*k*=2
	$\mathfrak{D}_{5-14}$	0	*k*=3

Table C1. 2 Classification model parameters for the prognosis using two years temporal window.

Three Years Temporal Window

$\mathfrak{D}_{all}$	$\mathfrak{D}_{1-4}$	$\mathfrak{D}_{5-14}$
As_tot	As_tot	LM_a
PA_Tot	PA_Tot	LM_a_Interf
LM_a_Cued	LM_a_Cued	LM_tot_Interf
LM_a_Interf	Or_Total	Forgetting
LM_tot_Interf	MPR	Orient_T
Forgetting	CVLT_a2	Or_Total
Orient_T	CVLT_a_crint_ajsem_intr	Cube
Or_Total	CancellationTask_Z	CVLT_a3
Fluency_Sem	WordRecall_Z	CVLT_b_tot
Cube	VerbalPaired_AssociateLearning_Z	SemanticFluency_Z
MPR	LogicalMemory_A_Z	LogicalMemory_Z
Proverb	LM_DR_Z	
CVLT_a2	MPR_Z	
CVLT_a5		
CancellationTask_Z		
SemanticFluency_Z		
Orientation_Z		
GeneralInformation_Z		
VerbalPaired_AssociateLearning_Z		
LogicalMemory_A_Z		
LM_DR_Z		
MPR_Z		
Proverbs_Z		
A1_Z		

Table C1. 3 Features selected in the different datasets which are differentiated based on GDS values, using three years temporal window.

	Dataset	SMOTE (%)	Parameters
Naïve Bayes	$\mathfrak{D}_{all}$	0	Gaussian
	$\mathfrak{D}_{0-4}$	0	*Kernel*
	$\mathfrak{D}_{5-14}$	0	*Supervised Discretization*
SVM Poly	$\mathfrak{D}_{all}$	0	*Compl*=1.5 and *Exp*=1
	$\mathfrak{D}_{0-4}$	0	*Compl*=1.5 and *Exp*=1
	$\mathfrak{D}_{5-14}$	0	*Compl*=0.5 and *Exp*=1
SVM RBF	$\mathfrak{D}_{all}$	0	*Compl*=1.0 and Υ=1.0
	$\mathfrak{D}_{0-4}$	0	*Compl*=3.0 and Υ=1
	$\mathfrak{D}_{5-14}$	0	*Compl*=1.5 and Υ=1.0
Decision Tree C4.5	$\mathfrak{D}_{all}$	0	*Conf*=0.3
	$\mathfrak{D}_{0-4}$	0	*Conf*=0.2
	$\mathfrak{D}_{5-14}$	0	*Conf*=0.5
KNN	$\mathfrak{D}_{all}$	0	*k*=9
	$\mathfrak{D}_{0-4}$	0	*k*=5
	$\mathfrak{D}_{5-14}$	0	*k*=7

Table C1. 4 Classification model parameters for the prognosis using three years temporal window.

Four Years Temporal Window

$\mathfrak{D}_{all}$	$\mathfrak{D}_{0-4}$	$\mathfrak{D}_{5-14}$
LM_tot_Interf	LM_a_Interf	LM_a_Interf
Forgetting	LM_tot_Interf	Or_Total
Orient_T	Or_Total	CVLT_a1a5
Or_Total	Proverb	SemanticFluency_Z
Fluency_Sem	CVLT_a_intr	LogicalMemory_A_Z
Cube	SemanticFluency_Z	MPR_Z
MPR	Orientation_Z	A5_Z
Proverb	WordRecall_Z	
CVLT_a3	GeneralInformation_Z	
SemanticFluency_Z	LogicalMemory_A_Z	
Orientation_Z	Proverbs_Z	
WordRecall_Z	TP_RT_Z	
VerbalPaired_AssociateLearning_Z	TMT_B_Z	
LogicalMemory_A_Z	LDFR_Z	
LM_DR_Z		
MPR_Z		
Proverbs_Z		
Atot_Z		

Table C1. 5 Features selected in the different datasets which are differentiated based on GDS values, using four years temporal window.

	Dataset	**SMOTE (%)**	**Parameters**
Naïve Bayes	$\mathfrak{D}_{all}$	0	*Gaussian*
	$\mathfrak{D}_{0-4}$	0	*Gaussian*
	$\mathfrak{D}_{5-14}$	208	*Supervised Discretization*
SVM Poly	$\mathfrak{D}_{all}$	0	*Compl*=2.5 and *Exp*=2
	$\mathfrak{D}_{0-4}$	0	*Compl*=4.0 and *Exp*=2
	$\mathfrak{D}_{5-14}$	104	*Compl*=2.0 and *Exp*=1
SVM RBF	$\mathfrak{D}_{all}$	0	*Compl*=1.0 and Υ=1.0
	$\mathfrak{D}_{0-4}$	0	*Compl*=1.0 and Υ=0.1
	$\mathfrak{D}_{5-14}$	156	*Compl*=1.0 and Υ=10.0
Decision Tree C4.5	$\mathfrak{D}_{all}$	0	*Conf*=0.1
	$\mathfrak{D}_{0-4}$	0	*Conf*=0.15
	$\mathfrak{D}_{5-14}$	130	*Conf*=0.2
KNN	$\mathfrak{D}_{all}$	0	*k*=4
	$\mathfrak{D}_{0-4}$	0	*k*=8
	$\mathfrak{D}_{5-14}$	0	*k*=1

Table C1. 6 Classification model parameters for the prognosis using four years temporal window.

Five Years Temporal Window

$\mathfrak{D}_{all}$	$\mathfrak{D}_{0-4}$	$\mathfrak{D}_{5-14}$
LM_a_Interf	PA_Tot	PA_Tot
Or_Total	Or_Total	LM_a_Interf
Fluency_Sem	Proverb	CVLT_a5
MPR	TPRT	SemanticFluency_Z
CVLT_a_crint_ajsem	CVLT_a_pers	LM_DR_Z
SemanticFluency_Z	Orientation_Z	
Orientation_Z	LogicalMemory_A_Z	
WordRecall_Z	LM_DR_Z	
VerbalPaired_AssociateLearning_Z	TMT_B_Z	
LogicalMemory_A_Z	A5_Z	
LM_DR_Z	SDCR_Z	
MPR_Z		
TP_RT_Z		
A5_Z		
LDFR_Z		

Table C1. 7 Features selected in the different datasets which are differentiated based on GDS values, using five years temporal window.

	Dataset	**SMOTE (%)**	**Parameters**
Naïve Bayes	$\mathfrak{D}_{all}$	168	*Supervised Discretization*
	$\mathfrak{D}_{0-4}$	48	*Supervised Discretization*
	$\mathfrak{D}_{5-14}$	48	*Supervised Discretization*
SVM Poly	$\mathfrak{D}_{all}$	48	*Compl*=2.5 and *Exp*=1
	$\mathfrak{D}_{0-4}$	72	*Compl*=3.0 and *Exp*=1
	$\mathfrak{D}_{5-14}$	0	*Compl*=2.0 and *Exp*=4
SVM RBF	$\mathfrak{D}_{all}$	240	*Compl*=4.0 and Υ=1.0
	$\mathfrak{D}_{0-4}$	120	*Compl*=5.0 and Υ=0.1
	$\mathfrak{D}_{5-14}$	92	*Compl*=2.0 and Υ=1.0
Decision Tree C4.5	$\mathfrak{D}_{all}$	24	*Conf*=0.15
	$\mathfrak{D}_{0-4}$	0	*Conf*=0.05
	$\mathfrak{D}_{5-14}$	69	*Conf*=0.05
KNN	$\mathfrak{D}_{all}$	48	*k*=6
	$\mathfrak{D}_{0-4}$	120	*k*=9
	$\mathfrak{D}_{5-14}$	69	*k*=7

Table C1. 8 Classification model parameters for the prognosis using five years temporal window.

C.2 Patient Similarities

Two Years Temporal Window

$\mathfrak{D}_{A+B}$	$\mathfrak{D}_{A+B\_Z}$
Age	LM_a_Interf
As_tot	Orient_S
PA_Tot	Orient_T
LM_tot	BlessedAVD
LM_a_Interf	BlessedTOT
MVI_tot	CancellationTask_Z
Orient_S	SemanticFluency _Z
Orient_T	Orientation_Z
Or_Total	VerbalPaired_AssociateLearning_Z
MPR	LogicalMemory_A_Z
CVLT_a1a5	Cube_Z
BlessedAVD	MPR_Z
BlessedTOT	Proverbs_Z
	A1_Z

Table C2. 1 Features selected in the different datasets in the Baseline analysis, using two years temporal window.

		Cluster	
		0	**1**
Real Class	**noEvol**	15	186
	Evol	35	49
# Instances		50	235

		Cluster		
		0	**1**	**2**
Real Class	**noEvol**	10	96	95
	Evol	28	19	37
# Instances		38	115	132

		Cluster			
		0	**1**	**2**	**3**
Real Class	**noEvol**	87	96	17	1
	Evol	26	19	34	5
# Instances		113	115	51	6

Table C2. 2 Confusion matrix for the two years temporal window obtained by the EM algorithm using $\mathfrak{D}_{A+B}$ with a) k=2; b) k=3; c) k=4

Three Years Temporal Window

$\mathfrak{D}_{A+B}$	$\mathfrak{D}_{A+B\_Z}$
Age	LM_a_Interf
DS_back	Orient_T
PA_Tot	CancellationTask_Z
LM_tot	DigitSpan_Z
LM_a_Interf	SemanticFluency _Z
LM_tot_Interf	Orientation_Z
MVI_tot	VerbalPaired_AssociateLearning_Z
Orient_T	LogicalMemory_A_Z
Or_Total	Cube_Z
CVLT_a1a5	Calculation_Z
	MPR_Z
	Proverbs_Z

Table C2. 3 Features selected in the different datasets in the Baseline analysis, using three years temporal window.

		Cluster	
		0	**1**
Real Class	**noEvol**	71	68
	Evol	19	96
# Instances		90	164

		Cluster		
		0	**1**	**2**
Real Class	**noEvol**	85	19	35
	Evol	21	73	21
# Instances		106	92	56

		Cluster			
		0	**1**	**2**	**3**
Real Class	**noEvol**	41	5	63	30
	Evol	29	53	16	17
# Instances		70	58	79	47

Table C2. 4 Confusion matrix for the three years temporal window obtained by the EM algorithm using $\mathfrak{D}_{A+B}$ with a) k=2; b) k=3; c) k=4

Four Years Temporal Window

$\mathfrak{D}_{A+B}$	$\mathfrak{D}_{A+B\_Z}$
Age	LM_a_Interf
PA_tot	Orient_T
LM_a_Interf	CancellationTask_Z
Orient_T	LogicalMemory_A_Z
Or_Total	SemanticFluency_Z
Fluency_Sem	MotorInitiative_Z
Proverb	Orientation_Z
	GeneralInformation_Z
	VerbalPaired_AssociateLearning_Z
	LogicalMemory_Z
	LogicalMemory_A_Z
	LM_DR_Z
	Cube_Z
	MPR_Z
	Atot_Z

Table C2. 5 Features selected in the different datasets in the Baseline analysis, using four years temporal window.

		Cluster	
		0	**1**
Real Class	**noEvol**	81	4
	Evol	83	60
# Instances		164	64

		Cluster		
		0	**1**	**2**
Real Class	**noEvol**	0	47	38
	Evol	52	30	61
# Instances		52	77	99

		Cluster			
		0	**1**	**2**	**3**
Real Class	**noEvol**	2	11	26	46
	Evol	57	23	39	24
# Instances		59	34	65	70

Table C2. 6 Confusion matrix for the four years temporal window obtained by the EM algorithm using $\mathfrak{D}_{A+B}$ with a) k=2; b) k=3; c) k=4

Five Years Temporal Window

$\mathfrak{D}_{A+B}$	$\mathfrak{D}_{A+B\_Z}$
Age	LM_a_Interf
PA_tot	Orient_T
LM_a	DigitSpan_forward_Z
LM_a_Interf	SemanticFluency_Z
Orient_T	Orientation_Z
Or_Total	GeneralInformation_Z
Fluency_Sem	VerbalPaired_AssociateLearning_Z
M_Initiative	LogicalMemory_Z
MPR	LogicalMemory_A_Z
CVLT_a1	LM_DR_Z
	Cube_Z
	MPR_Z
	A5_Z

Table C2. 7 Features selected in the different datasets in the Baseline analysis, using five years temporal window.

		Cluster	
		0	**1**
Real Class	**noEvol**	33	18
	Evol	45	119
# Instances		78	137

		Cluster		
		0	**1**	**2**
Real Class	**noEvol**	21	0	30
	Evol	58	71	35
# Instances		79	71	65

		Cluster			
		0	**1**	**2**	**3**
Real Class	**noEvol**	32	0	19	0
	Evol	30	28	60	46
# Instances		62	28	79	46

Table C2. 8 Confusion matrix for the five years temporal window obtained by the EM algorithm using $\mathfrak{D}_{A+B}$ with a) k=2; b) k=3; c) k=4