TÉCNICO LISBOA

Exploration of Temporal Patterns in Classification Problems

Versão Provisória

Daniel Sousa Veloso de Oliveira Cardoso

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisor(s): Cláudia Martins Antunes

Examination Committee

Chairperson:	Prof. ...
Supervisor:	Prof. Cláudia Martins Antunes
Member of the Committee:	Prof. Sara Alexandra Cordeiro Madeira

September 2014

Dedicated to someone special...

Acknowledgments

A few words about the university, financial support, research advisor, dissertation readers, faculty or other professors, lab mates, other friends and family...

Resumo

Inserir o resumo em Português aqui com o máximo de 250 palavras e acompanhado de 4 a 6 palavras-chave...

Palavras-chave: Prognóstico, Classificação, Dependencias Temporais

Abstract

The use of data mining techniques in healthcare has been noticing an increased relevance over the last few years, being applied with a variety of objectives, with the most common one being the automatic diagnostic process. In this process, data mining techniques have achieved interesting and successful results. However, when it comes to prognosis the same quality of results is not being achieved. We argue that this happens thanks to the inability of the used techniques to capture the inherent temporal dependencies present on the data. Specifically, the temporal evolution of a patient is not being taken into account when performing prognosis. In this paper, we propose a different approach, independent of the domain, to address this issue. We present our preliminary results on two different datasets that show an improvement in the overall precision of the prognosis.

Keywords: Prognosis, Classification, Temporal Dependencies

Contents

Acknowledgments v
Resumo vii
Abstract ix
List of Tables xiii
List of Figures xvi
Nomenclature 1
Glossary 1

1 Introduction **1**

2 Data Mining in HealthCare **3**
2.1 Medical Diagnosis versus Prognosis 3
2.2 Data Mining Techniques 4
2.2.1 Decision Tree 4
2.2.2 Artificial Neural Networks & Support Vector Machines 5
2.2.3 Bayesian Classifiers 6
2.2.4 Regression Analysis 7

3 Related Work **9**
3.1 Alzheimer 10
3.2 Cancer 10
3.3 Diabetes 14
3.4 Venous Thromboembolism 15
3.5 HIV/AIDS 16
3.6 Kidney Failure 17
3.6.1 Organ Failure 18
3.7 Critical Analysis 18
3.7.1 Challenges in using classification for prognosis 19

4 Approach **21**
4.1 Estimation Algorithm 23
4.2 Implementation Issues 25

4.3 Example . . . 25

5 Validation and Experimental Results **27**

5.1 Dataset Description . . . 27

5.1.1 ALS Dataset . . . 27

5.1.2 Discret ALS Dataset . . . 27

5.1.3 Hepatitis Dataset . . . 28

5.2 Validation Techniques . . . 29

5.3 Experimental Results . . . 31

5.3.1 Diagnosis Model . . . 32

5.3.2 Regression Techniques . . . 33

5.3.3 Decision Tree . . . 35

5.3.4 HMM . . . 37

5.3.5 Discussion . . . 40

6 Conclusions **45**

6.1 Achievements . . . 45

6.2 Future Work . . . 46

Bibliography **50**

List of Tables

4.1 Patient data. 26
4.2 Data used on the univariate estimation approach. 26
4.3 Data used on the multivariate estimation approach. 26

5.1 Discretization for both Blood Pressure features. 28
5.2 Discretization for the Pulse feature. 28
5.3 Discretization for the Respiratory Rate features. 29
5.4 Discretization for Percentage of Normal feature. 29
5.5 Discretization for the Temperature feature. 30
5.6 Notations in a binary contingency table. Color coding indicates correct (green) and incorrect (pink) rates or counts in the contingency table. 31

List of Figures

2.1 Example Decision Tree to predict if some students will play football. 4
2.2 Artificial Neural Network structure. 5
2.3 Example of 2D SVM optimal hyperplane. 6

3.1 AUCs for the neural network and logistic regression (LR). 11

5.1 Class distribution of the ALS dataset. 28
5.2 Discretization for the Weight feature. 30
5.3 Class distribution of the Hepatitis dataset. 31
5.4 BaselineSingleObs accuracy (several classifiers and number of observations). 32
5.5 BaselineMultipleObs accuracy (several classifiers and number of observations). 32
5.6 Impact of the number of observations on the accuracy of the linear regression estimation models for each variable, in the ALS dataset. 33
5.7 Impact of the number of observations on the accuracy of the logistic regression estimation models for each variable, in the hepatitis dataset. 34
5.8 Execution time of feature estimation in the ALS dataset using Linear Regression. 34
5.9 Execution time of feature estimation in the hepatitis dataset using Logistic Regression. . . 35
5.10 Accuracy of different models. 35
5.11 Impact of the number of observation on prognosis models. 36
5.12 Different metrics for the overall prognosis using logistic regression on the Hepatitis dataset. 37
5.13 Impact of the number of observations on the accuracy of the decision tree estimation models for each variable using both univariate and multivariate models on the discrete ALS dataset. 38
5.14 Impact of the number of observations on the accuracy of the decision tree estimation models for each variable using both univariate and multivariate models on the Hepatitis dataset. 38
5.15 Execution time of feature estimation in the hepatitis dataset using Decision Trees. 39
5.16 Execution time of feature estimation in the hepatitis dataset using Decision Trees. 39
5.17 Accuracy of different models using the decision trees estimations. 40
5.18 Impact of the number of observation on prognosis models using the decision trees estimations. 40

5.19 Impact of the number of observations on the accuracy of the HMM estimation models for each variable using both univariate and multivariate models in the ALS dataset. 41
5.20 Impact of the number of observations on the accuracy of the HMM estimation models for each variable using both univariate and multivariate models in the Hepatitis dataset. . . . 41
5.21 Execution time of feature estimation in the ALS dataset using HMMs. 42
5.22 Execution time of feature estimation in the hepatitis dataset using HMMs. 42
5.23 Accuracy of different models using the HMM estimations. 43
5.24 Impact of the number of observation on prognosis models using the HMM estimations. . . 43

Chapter 1

Introduction

The role of data analysis in healthcare has gained more attention, as available mining techniques have achieved higher levels of maturity. In particular, classification methods become to play a decisive role when applied to clinical trials, by providing high quality external evidence to support evidence-based medicine [40]. The rigorous metrics available to evaluate the confidence about the collected evidence on those trials, allied to the variety of techniques suited to different kinds of data, revealed to be fundamental to keep expertise up-to-date and available worldwide.

Despite the success of those techniques, they are mostly appropriate to analyze tabular data, described by a set of independent variables. Actually, we can see this kind of data as a static snapshot of the status of some entity, which is completely suited to represent patient records collected during their diagnosing process. On the other hand, prognosis may be seen as the prediction of an outcome in a future instant, considering all available data collected along time. In this manner, we may think of prognosis as the task of predicting an outcome, given a set of time-ordered snapshots. While in a single snapshot, methods may assume some level of independency among variables, this assumption is clearly unlikely in a set of snapshots, where the same variable is measured along different instants of time.

Actually, and despite this dependency among snapshots, a large number of classification-based approaches have been proposed for prognosis (see [19], [34], [48], for example). In our opinion, the results achieved through them have been impaired due to the dependency among the different values for the same variable along time.

In this dissertation, we argue that the simple prediction of the prognosis outcome by traditional classification methods, given a set of snapshots, can be significantly improved by exploring the temporal relations, or evolution verified in each variable that compose the snapshots. In order to validate our claim, we formalize the problem addressed, and present an approach to take those dependencies into account in the process of outcome prediction. We also perform a comparative analysis between two techniques used to estimate the future values of some features.

After the formalization of the prognosis problem, we review a set of case studies on several different diseases, with the most well-known classification techniques (chapter 3). In chapter 4 we describe our

approach, and propose two distinct implementations of it, followed by a description of some experiments that compare the accuracy of both traditional classifiers and our approach using two different techniques for the estimation phase (chapter 5). The dissertation concludes with a discussion of the improvements achieved, the issues constraining those improvements and proposing some guidelines for the next steps (chapter 6).

Chapter 2

Data Mining in HealthCare

Data Mining is the process of gathering knowledge from raw data. It is different from information retrieval because in that case what is retrieved is information that is present explicitly in the data and in the data mining case it discovers implicit patterns using analytical tools. There are two types of data mining, descriptive and predictive. The former, like the name says, describes characteristics and relations of the existing data and the latter use the existing data to predict some future value.

Data mining has been applied in a collection of fields like CRM, finance, social networks and health care.

One area that is becoming increasingly important is health, with the amount of data available and even the increase of the digitalization, to take full advantage of all this data, data mining tools need to be used. Data mining can help physicians to identify the most effective treatments, find adverse drug reactions, fraud detection, performing diagnostics and prognostics.

2.1 Medical Diagnosis versus Prognosis

Diagnosis is the use of patient's data, demographic and clinical, in order to under-stand and classify the current health condition of a patient.

Prognosis is the foreseeing or prediction of the risk or probability of a certain health event happening, in the future, using the clinical and non-clinical data. It is the medical prediction of how the pair patient disease is going to evolve in a specified period of time.

To do this prognosis, a physician will use data that relates the patient to a certain part of the population, i.e. demographic data, as well as the patient's and patient's family clinical history. This means that the evolution of the patient is important in the prediction of his next state. Simply putting, if a patient is showing improvement in a certain factor that is responsible for some disease, it is more probable that his prognosis related to that disease is better than if it the patient had the same value but that factor was deteriorating.

As previously stated, in the process of making a prognosis a physician uses the medical history of a patient. This includes the different states a patient has been in the form of various clinical analysis he

had done in different points over time. The need to use this sequential information shows the utmost importance that time has when predicting someone's survivability, risk of recurrence.

2.2 Data Mining Techniques

Different techniques have been used to perform all of those predictions, as we will show in the following chapter. We will start by describing the most common classification techniques used for prognosis, following with the cases where they have been applied to perform prognosis in different diseases.

2.2.1 Decision Tree

Decision trees are one of the most common classification techniques. They are a supervised learning technique that, based on the data features and a metric, that can be the Gini index, information gain, and Chi-squared test, tries to find the feature that best splits the data into more homogeneous sets in terms of the target variable.

By the end of the algorithm we have a tree where in each interior node there is one of the features and an edge per value of that feature. In the leaf nodes of this tree structure the class label is represented. An example of a decision tree is represented in Figure 2.1, where the outcome is whether some students will play football outside or not. If the outlook of the weather is overcast then the students will play, if it is sunny we need to look into the humidity, and if it is rainy then the wind is the deciding factor.

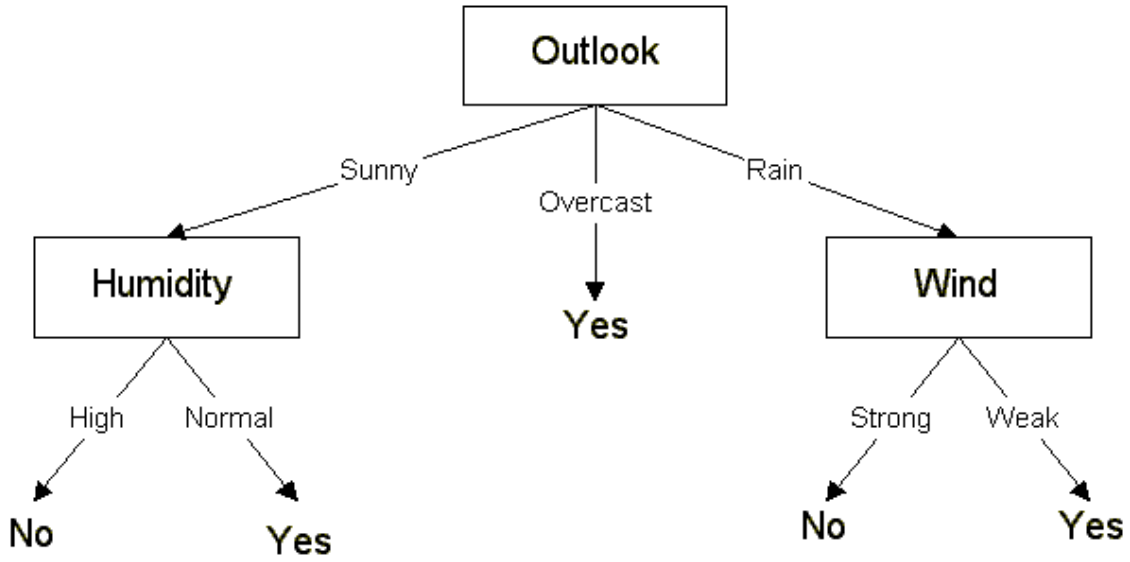


Figure 2.1: Example Decision Tree to predict if some students will play football.

The most common algorithms to build decision trees are Quinlan's ID3 [36], C4.5 [37] that came improve on ID3, and Breiman & et al.'s Classification And Regression Trees (CART) [10].

In the current work on prognosis,as seen in 3, the use of C5.0 is also found. C5.0 is an extension of C4.5 that, among several issues, presents a considerable performance optimization.

Decision Trees result in a very easily understandable, like Figure 2.1, it is easy to see that some initial variable divides the data into two categories and then other variables split the resulting child groups. This information is very useful to the researcher who is trying to understand the underlying nature of the data being analyzed.

2.2.2 Artificial Neural Networks & Support Vector Machines

Artificial Neural Networks are computational models that approximate the functioning of the brain, in the sense that they are highly complex and non-linear. These networks are composed by a group of interconnected nodes, also called neurons, and are used for classification. They have an input layer with nodes that correspond to data features, a various number of hidden layers and an output layer where the outcome is represented as seen in Figure 2.2.

Contrarily to decision trees, neural networks do not present an easily-understandable model. A neural network is more of a "black box" that delivers results without an explanation of how the results were derived. Thus, it is difficult or impossible to explain how decisions were made based on the output of the network.

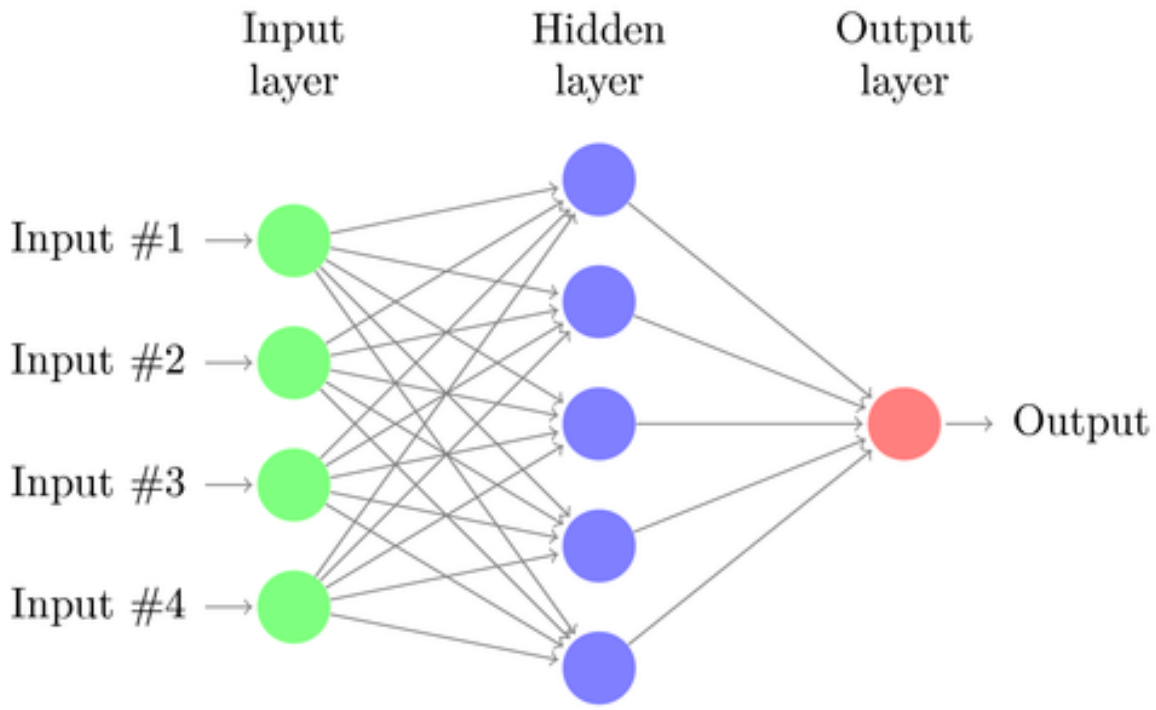


Figure 2.2: Artificial Neural Network structure.

SVMs are another supervised machine learning technique, where a hyperplane is found that correctly separates the spatial representation of the data into the various classes. For example, if the data is 2 dimensional, the hyperplane is a line that correctly divides the data and has the largest margin between itself and a data point, as seen in Figure 2.3.

SVMs have been used with high accuracy and can with the right kernel can have good results even

if the data is not linearly separable in the base feature space. They are memory intensive and require a lot of tunning and configuration.

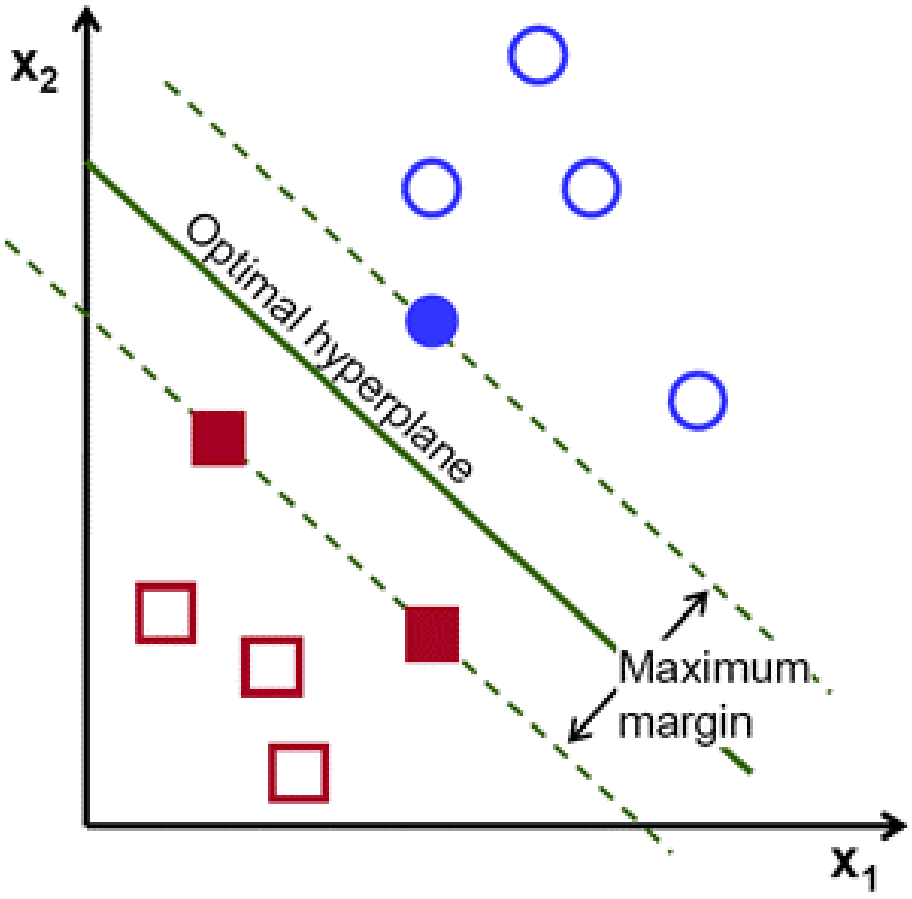


Figure 2.3: Example of 2D SVM optimal hyperplane.

2.2.3 Bayesian Classifiers

Bayesian classifiers are probabilistic classifiers that get their name by making use of the Bayes Rule of Inference.

Naïve Bayes Classifier calculates the probability of a certain outcome class by considering that all features are independent, in other words by seeing how their value alone influences the outcome class.

If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so less data would be necessary.

Bayesian Networks

Bayesian networks are probabilistic graphical models, represented as directed acyclic graphs, where nodes represent random variables and edges the probabilistic dependency between them. These dependencies between variables are found using the theory of information.

Exemplo

HMMs or Hidden Markov Models can be viewed as a specific case of the more general dynamic graphical models, where particular dependencies are assumed. Thus, HMMs and their variants can be interpreted as examples of DBNs. An HMM is a stochastic finite automaton, where each state generates (emits) an observation. An HMM is described by a quintuple, N, M, A, B, π where this symbols mean:

N = number of states in the model

M = number of distinct observation symbols per state (observation symbols correspond to the physical output of the system being modelled)

T = length of observation sequence

O = observation sequence, i.e. , $O_1, O_2, \ldots, O_T$

Q = state sequence $q_1, q_2, \ldots, q_T$ in the Markov model

A = a_{ij} transition matrix, where a_{ij} represents the transition probability from state i to state j

B = $b_j(O_t)$ observation emission matrix, where $b_j(O_t)$ represent the probability of observing O_t at state j

π = π_i the prior probability, where π_i represent the probability of being in state i at the beginning of the experiment, i.e., at time $t = 1$

λ = (A, B, π) the overall HMM model.

As mentioned above the HMM is characterized by N, M, A, B and π. The $a_{ij}, b_i(O_t)$, and π_i have the properties:

$\Sigma_j a_{ij} = 1, \Sigma_t b_i(O_t) = 1, \Sigma_i \pi_i = 1$ and $a_{ij}, b_i(O_t)$, and $\pi_i >= 0$ for all i, j, t.

2.2.4 Regression Analysis

Regression analysis is the use of a statistical analysis method used to measure the relation between variables. In other words, it helps to understand how a dependent variable varies with changes in one of the independent variables.

Linear Regression is an example of regression analysis where a linear function is used to model the data. When the outcome variable, the dependent variable is binary or categorical, linear regression cannot be applied. In those cases it is used logistic regression.

However, linear regression is appropriate only if the data can be modeled by a straight line function, which is often not the case.

Logistic Regression is a generalization of linear regression that, as just mentioned, is used to predict binary or categorical dependent variables. In this regression instead of predicting the estimate value of an event it predicts the probability of it occurring.

Regressions have comprehensible probabilistic interpretation and you can easily update your model to take in new data, unlike decision trees or SVMs. You can use this if you expect to receive more training data in the future that you want to be able to quickly incorporate into your model.

Another example of regression analysis that is also used in the healthcare domain is called Cox Proportional Hazard Models, which are a type of survival models, where the time to the occurrence of an event is related with one or more covariates that may be responsible. They show the influence of variables in the time to an event occurrence.

In medical studies Cox Proportional hazard models are the most common method used for survival outcomes.

It is an extension of the logistic model to the survival setting. Similar to conditional logistic regression with conditioning only at time of events. In the logistic method we us a linear predictor while in the COX mode a hazard function is used. The hazard function dictates the risk of the outcome during the follow up time.

$$\lambda(t|X) = \lambda(t)e^{\beta X} \tag{2.1}$$

Where $\lambda(t)$ is the hazard at time t, and is usually estimated at the mean values of the predictors and βX is the linear predictor, $\beta_1 \times x_1 + \beta_2 \times x_2 + ... + \beta_p \times x_p$

The linear predictor is usually centered at the mean value of the predictors, and $e^{\beta}X$ then indicates the hazard ratio compared to the average risk profile.

Chapter 3

Related Work

In this chapter it will be overviewed the work that has been done in the area of automatic prognostic and diagnostic. Diagnostic because, even though this thesis will be about prognosis, [22] states that the development, validation and impact assessment of both cases can be mutatis mutandis applied.

The prediction classification can be a diagnostic or a prognostic depending only on the amount of time until the outcome assessment. Being the options between the outcome assessment the present or the future, the former is a diagnostic and the latter a prognosis.

In the field of diagnosis the techniques used revolve around the same as in prognosis. Mainly it uses decision trees, artificial neural networks, association rules and Bayes classifiers as well as Support Vector Machines [25].

There are three types of prediction that can be done when talking about prognosis:

- We can try to predict the probability of developing a disease or a state of that disease, in other words we can perform a risk assessment or predict the disease susceptibility;

- We can predict if there will be recurrence of an event, for example if a cancer will recur after it was excised;

- We can predict if the patient will be alive at a certain time point, known as *survivability*.

We will separate the review on prognostic prediction by disease in order to allow the comparison between the work being done in the various diseases. Showing that even though the same techniques are used they require different preprocessing and the end results are very data dependent.

We can find work on prognostic prediction as far back as 1980 [31] where a regression analysis is used to find the predictive power of 17 features when predicting the survival of breast cancer patients. Also in the early 90s [21] where logistic regression is used to predict Survival of HIV infected patients and [30] where dynamic programming is used to predict the time to recurrence of an excised cancer.

3.1 Alzheimer

The Alzheimer's disease (AD) is the most common form of dementia. It causes problems with memory, thinking and behavior. Symptoms usually develop slowly and get worse over time, becoming severe enough to interfere with daily tasks and eventually leading to death. In order to predict the progress of the disease several techniques have been applied.

Alzheimer's disease is associated with variable but shortened life expectancy, even at relatively early stages. For that reason having a survivability expectancy might be important for the patients and their carers to understand and plan ahead.

In [34] they used Cox proportional hazards regression modeling for univariate and multivariate statistics.

On the multivariate analysis in order to find the most predictive features a forward stepwise approach was used followed by a backward stepwise linear regression in order to confirm if the results were robust.

The final model, SAM (Survival in Alzheimer's Model) is a 4 point risk scale according to whether a patient has or not the identified risk factors (increasing age, Constructional praxis, Gait apraxia). A patient with two risk factors will have an 80% chance of surviving 12 months, but less than 50% chance of surviving 3.5 years.

This study has some limitations like the fact that one of the features where it was built upon, was clinically obtained, by a standardized assessment by the same doctor. Also this model's generalizability may be limited because the cohort was a convenience sample and was not recruited to be representative of the larger population of people with AD.

In [48], Zhou *et al.* develop a new multi-task learning formulation based on the temporal group Lasso regularizer, in order to predict the Alzheimer's disease progression, based on the Mini Mental State Examination (MMSE) and Alzheimer's disease Assessment Scale cognitive subscale (ADAS-Cog) scores, that give the cognitive status of a patient. The multi-task regression approach captures the relation of the task, and the regularizer ensures that a small set of features is used for the regression and that a large deviation between successive time points is penalized.

3.2 Cancer

Cancer, known medically as a malignant neoplasm, is a class of diseases characterized by out-of-control cell growth. It becomes harmful when faulted cells grow into lumps of tissue that are called tumors. The cancer may also end up by spreading when cancerous cells move through the lymphatic system or bloodstream.

Cancer is seen as a deadly disease, as most people end up dying from the cancer or its treatment. The ones that actually survive have twice the probability of developing a second cancer than the people that were never diagnosed with cancer. [38]

Because of this the three types of prognosis are found in cancer research: there is the prediction of the probability of developing cancer, in other words we can perform a risk assessment or predict

the cancer susceptibility, the prediction if there will be recurrence in the cancer after if it excised or the prediction if the patient will be alive at a certain time point.

In these three areas of prognosis we have the following work.

To perform the prediction of survival at 5, 10 and 15 years after the diagnostic, [29] uses artificial neural networks and logistic regression, showing that neural networks are consistent with logistic regression as is represented in 3.1.

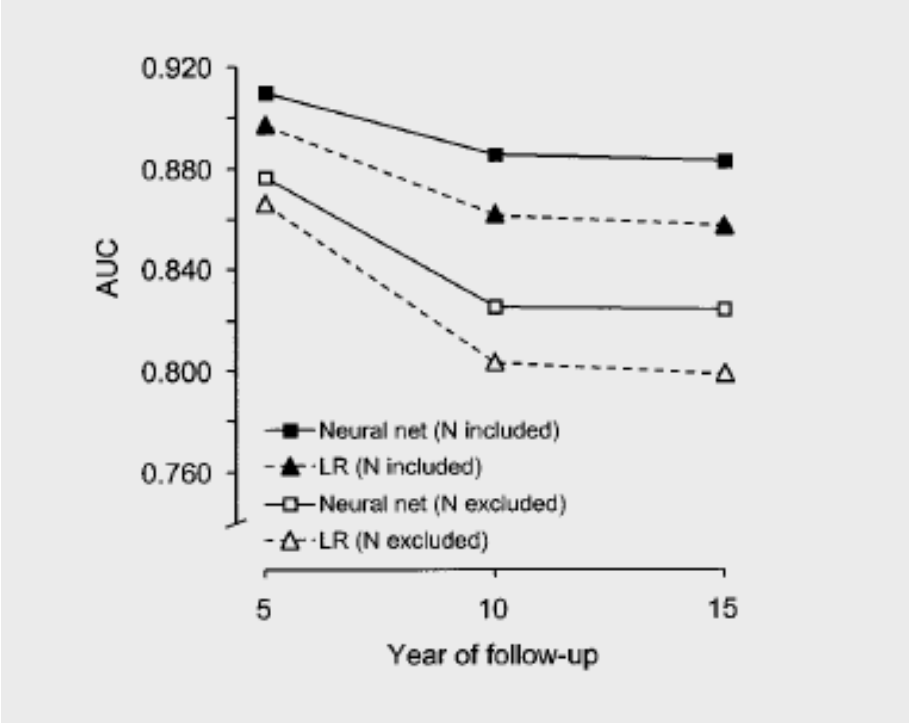


Figure 3.1: AUCs for the neural network and logistic regression (LR).

In [43], in order to decide which treatment is better for the patients' well-being, a Cox regression analysis is used to predict a score based on the regression coefficients, which classifies the patients in 3 different groups: the ones with good, intermediate or bad prognosis in terms of survivability. Using this knowledge of the degree of prognosis in addition with the short-term versus long-term benefits of each treatment, a better choice can be performed helping to improve the patient's quality of life.

To predict the overall survivability, at 1 year and 5 years mark, of patients with Acute Myeloid Leukemia, Breems *et al.* applied multivariate Cox regression analysis with stepwise backward selection on the patient's age at the time of the relapse, length of relapse free interval, previous stem cell transplant and cytogenetics.

Like in [43], they used the regression coefficients has a score function that is used to classify the patients [9]

The purpose of [15] is to develop predictive models and discover/explain relationships between certain independent variables and the survivability, 5 years after the diagnosis, in the context of breast cancer. Delen *et al.* per-form a comparative study with decision trees (C5.0), MLP neural network and logistic regression. Showing that with the SEER dataset and using a tenfold cross validation, the de-

cision tree performed the best out of the three with accuracy of 0.9362, closely followed by the neural network that achieved 0.9121 and the logistic regression that got 0.8920.

In the presence of microarray data, the clinical data is usually underused say Gevaert *et al.* that in [20] propose the usage of Bayesian networks to equally use both sources of data and that way get better results when performing the prognosis.

They evaluated three methods for integrating clinical and microarray data: decision integration, partial integration and full integration and used them to classify publicly available data on breast cancer patients into a poor and a good prognosis group. The partial integration method is most promising and has an independent test set area under the ROC curve of 0.845.

In the problem addressed in [3], a neural network calculates a time interval that corresponds to a possible right end-point of the patient's disease-free survival time, in other words it predicts the time to recur (TTR) by classifying the patient into 4 classes, TTR $\leq$ 1 year, TTR $\leq$ 3 years, TTR $\leq$ 6 years and TTR $>$ 6 years. The accuracy of the neural network was measured through a stratified tenfold cross validation approach. Sensitivity ranged between 80.5 and 91.8%, while specificity ranged between 91.9 and 97.9%, depending on the tested fold and the partition of the predicted period.

In [8] a comparison is made between different data mining techniques, Naïve Bayes, Neural Networks and Decision Trees when predicting survivability of breast cancer patients 5 years after the diagnose. For that comparison the Weka toolkit[1] and the SEER Dataset is used, which is composed by demographic data (age, race, etc.) and clinical data (Extension of tumor, stage of cancer, etc.). After the tests the conclusion was that both, decision trees and neural networks, had better and similar performance with accuracy around 86%, though in the computational time the approaches did differ where the neural networks model took 12 times more to be built.

Because of the neural networks' ability to consider variable relations and create non-linear predictions models they are a very used method for cancer survivability prediction, how long after surgery it is expected that the cancer will recur. Here in [12] it is shown that they can be used to predict the probability of survivability, and based on a threshold classify them as good or bad prognosis, with 2 different datasets.

In [19], Endo *et al.* use Logistic Regression model, Artificial Neural Network (ANN), Naive Bayes, Bayes Net, and a collection of Decision Trees (Decision trees with naïve Bayes, ID3 and J48 algorithms) to predict breast cancer survival at 5 years learning that Logistic regression has the highest accuracy along with J48. Decision trees tend to have high sensitivity. But is also shown that the best algorithm depends on the object and the dataset.

Because there is no use of fuzzy logic when performing cancer prognosis, most of the current work uses neural networks that yield difficult to understand models and that there is no use of hybridization of machine learning techniques, Muhammad Umer Khan *et al.* investigated a hybrid scheme based on fuzzy logic and decision trees on the SEER dataset. They performed experiments using different combinations of number of decision tree rules, types of fuzzy membership functions and inference techniques in order to predict the patient survivability. They end up by comparing the performance of each for can-

[1] http://www.cs.waikato.ac.nz/ml/weka/

cer prognosis and found hybrid fuzzy decision tree classification is more robust and balanced than the independently applied crisp classification. [24]

In [14], Delen uses a handful of data mining techniques, decision trees, artificial neural networks and sup-port vector machines along with the most common statistical analysis tool, logistic regression, to build a prediction model for prostate cancer survivability and comparing their performance. The results indicated that SVMs are the best predictor with a test data set accuracy of 92.85%, followed by ANNs with an accuracy of 91.07%, followed by decision trees with an accuracy of 90.00% and logistic regression with an accuracy of 89.61%.

Jong Pill Choi *et al.* compared the performance of an Artificial Neural Network, a Bayesian Network and a Hybrid Network used to predict breast cancer prognosis. The hybrid Network was a combination of ANN and Bayesian Network. All the techniques were used on nine variables of the SEER data that were clinically accepted. In this research the accuracy of ANN (88.8both performed much better than the Bayesian Network. [13]

In [44] improve the L1-L2 norm SVM that has automatic feature selection for prognostic prediction to use regression, and developed the algorithm to utilize the information of censored data. The proposed method is compared with other seven prognostic prediction methods, namely CART, MARS, RSA, RRLC, L1-norm SVM, L2-norm SVM, Elastic Net, penalized Buckley-James, on three real world data sets. The experimental results show that the proposed method performs consistently better than the medium performance and that it is more efficient than other algorithms that achieved similar performance.

Kharya performs a review of use cases where data mining has been used to per-form prognosis of cancer disease. It shows that the most common cases, while they may need to be tested on larger set of examples in order to find rules with higher level of statistical confidence, they do find statistically significant associations that can help predict a patients' future. In this study they show examples using decision trees, neural networks, logistic regression as well as Bayesian networks. [25]

In [45], the prediction of survivability on the 5 year mark after diagnose were performed using decision trees and logistic regression. Using the SEER dataset Wang *et al.* show that logistic regression, even though the accuracy is similar, outperforms decision trees by having a higher g-mean and by comparing the ROC curve and AUC.

In [41] instead of using the complete Wisconsin Prognostic Breast Cancer data set, a pre-processing technique is used in order to reduce the number of features and improve the accuracy of polynomial neural network that was later used. The pre-processing technique is called principal component analysis (PCA) and it is a statistical procedure that returns a set of principal components. These principal components are less than or equal to the number of original features and they are ordered by their importance in the variability of the outcome. It is shown that the use of PCA is preferred to normalization, having the former more accurate results.

Using the SEER database Lakshmi *et al.* perform a comparison of a number of techniques when diagnosing and predicting 5-year survivability of patients diagnosed with breast cancer. The techniques that were compared were: C4.5, SVM, PNN, k-NN, Binary Logistic Regression as well as Multinomial Logistic Regression, Partial Least Squares Regression (PLS-DA), Partial Least Squares Linear Discrimi-

nant Analysis (PLS-LDA), k-means and Apriori Algorithm. In the end, this study [27], shows that PLS-DA performs the best with lowest computation time and highest accuracy.

3.3 Diabetes

Diabetes, often referred to by doctors as diabetes mellitus, describes a group of metabolic diseases in which the person has high blood glucose (blood sugar), either because insulin production by the pancreas is inadequate, or because the body's cells do not respond properly to insulin, or both.

There are three types of diabetes: type 1 is when the body does not produce insulin, type 2 is when the body does not produce enough for normal function or the cells in the body do not react to insulin, insulin resistance and the third type affects females when pregnant. They develop high levels of blood sugar and don't have enough insulin to transport it.

In [**?**] a risk score to predict the incidence of diabetes was developed. The multivariate logistic regression model coefficients were used to assign each variable category a score. The Diabetes Risk Score was composed as the sum of these individual scores. In the final predictive model there were 7 features selected, Age, BMI, waist circumference, history of antihypertensive drug treatment and high blood glucose, physical activity, and daily consumption of fruits, berries, or vegetables.

The model was developed using a cohort study from 1987 and another from 1992 where the subjects received by mail a questionnaire on medical history and health behavior and an invitation to a clinical examination.

The score that was derived from the regression coefficients ranged from 0 to 20 and the value ≥ 9 was able to predict diabetes with a sensitivity of 0.78 and 0.81, specificity of 0.77 and 0.76 in the 1987 and 1992 cohorts, respectively.

In order to improve the work of Lindström *et al.* the author of [7] aims to describe sex-specific lifestyle and clinical diabetes risk factors in a French population followed over 9 years in order to aid in identifying those at risk for incident diabetes. The data is composed by clinical along with biological data that was gathered every 3 years over a period of 9 years. In this study patients with al-ready incident diabetes in the beginning were excluded as well as the patients with unknown status of diabetes at the end. The author performed a statistical analysis over the data in order to find the most predictive features. Balkau *et al.* used logistic model to test for interactions with sex, Parsimonious logistic regression models were selected using forwards and backwards as well best model selection criteria using all parameters; the Hosmer-Lemeshow goodness-of-fit test was the principal criteria for selection of a model.

The resulting models, clinical and clinical + biological, were able to predict the incidence of diabetes over the 9 year period. They studied the influence of gender in the model, learning that the predictive functions were different for each sex.

Because the currently available screening tools for identifying individuals at high risk of type 2 diabetes can be invasive, costly and time consuming Xie *et al.* devel-oped a tool to identify individuals in the Chinese general population with high risk of developing type 2 diabetes (Xie, *et al.*, 2010). Using data from 994 persons with type 2 diabetes and 13 129 persons with normal fasting glucose, test performed

to find diabetic patients, aged 35-74 years. After a Classification and regression tree (CART) analysis, performed separately in men and women, two risk trees were obtained: one with 5 risk levels for men, and another with 8 for women. Being that women with a diabetes risk level (DRL) of 8 and men with a DRL of 5 are at the highest risk of type 2 diabetes. The CART results were compared with multivariable logistic regression model including the same predictors achieving both the same AUC of 0.71 vs. 0.73 in women and 0.65 vs. 0.69 in men, in the training and testing samples, indicating a good prediction above chance.

In [11] a risk score is built for the prediction of type 2 diabetes in a 5 year follow up study between 1999 and 2004, using demographic data, like age, sex and ethnicity, some feature that represent the history of the patient and clinical tests. The score was built using a logistic regression analysis where the features' coefficients were rounded up and used as a score if that feature was present. It was found that this diabetes risk score was a useful non-evasive method to identify Australian adults at high risk of type 2 diabetes who might benefit from interventions to prevent or delay its onset.

3.4 Venous Thromboembolism

Venous Thrombosis is a blood clot that forms within a vein. A common cause of venous thrombosis is the deep vein thrombosis that can turn into a pulmonary embolism, which can be lethal. Venous thromboembolism is a disease that includes both deep vein thrombosis (DVT) and pulmonary embolism (PE).

In order to predict the outcome in a 30-day period of patients that had a pulmonary embolism, Aujesky *et al.* used clinical variables that were shown to be related with the death of patients with PE. These variables included demographics, comorbid conditions, physical examination findings, and laboratory and chest x-ray findings. On that data a stepwise logistic regression analysis was performed to create the pre-diction rules that classify within 5 levels of mortality risk. [6]

[18] Based on a cohort study of 929 patients that had a first unprovoked deep vein thrombosis, Eichinger *et al.* perform a Cox hazard proportional analysis to learn the relevance of, previously selected, clinical and laboratorial data in the recurrence of the thrombosis. Using those values a nomogram was created that can give risk probability of recurrence and correctly classify patients in risk categories.

The risk of recurrence in a patient that had an unprovoked thromboembolism is between 5 and 7% in the first year, that risk can be significantly reduced by the ad-ministration of oral anticoagulation therapy. On the other hand, the risk of major bleeding with ongoing oral anticoagulation therapy among venous thromboembolism patients is 0.9–3.0% per year with an estimated case-fatality rate of 13%. Given that the long-term risk of fatal hemorrhage appears to balance the risk of fatal recurrent pulmonary embolism among patients with an unprovoked venous thromboembolism, clinicians are unsure if continuing oral anticoagulation therapy beyond 6 months is necessary. In [39], Rodgers *et al.* used conditional logistic regression with forward variable selection, they conducted multivariable analysis with recurrent venous thromboembolism as the dependent variable in order to develop a risk score that may help clinicians decide whether to stop the anticoagulation therapy or not.

They concluded that it may be safe for women who have taken oral anticoagulants for 5–7 months after an unprovoked venous thromboembolism to discontinue therapy if they have 0 or 1 of the following signs or symptoms: hyperpigmentation, edema or redness of either leg; a D-dimer level of 250 μ/L or more while taking warfarin; BMI 30 $kg/m2$ or more; and age 65 years or more. A decision rule for mean was not able to be found.

In citeTosetto2012 another risk prediction score is develop for the same task as [39], to help clinicians know if the anticoagulant therapy may stop, in this case after an initial period of at least 3 months.

The score (DASH, D-dimer, Age, Sex, Hormonal therapy) was developed firstly by identifying variables highly correlated with the recurrence by using COX regression. In the initial full model there were 7 features: D-dimer; age; patient sex; hormone use at time of VTE (in women); mode of initial presentation (DVT alone or DVT and PE); and previous history of cancer, not active at the time of initial event. At first, the model was reduced using backward selection of features, but because this may lead to an overly optimistic model they evaluated the degree of over-optimism both by a heuristic formula and by linear shrinkage with bootstrapping, this means that they adjust the regression coefficient based on the calculated optimism.

In the end, by multiplying the corrected coefficient by a common value and rounding to the nearest integer the score was found. The annualized recurrence risk was 3.1% for a score ≤ 1, 6.4% for a score $= 2$ and 12.3% for a score ≥ 3. By considering at low recurrence risk those patients with a score ≤ 1, life-long anticoagulation might be avoided in about half of patients with unprovoked VTE.

3.5 HIV/AIDS

Human immunodeficiency virus/ acquired immunodeficiency syndrome (HIV/AIDS) is a disease that affects the human immune system when infected with HIV. Acquired Immunodeficiency Syndrome is the final stage of HIV infection. People at this stage of HIV disease have badly damaged immune systems, which put them at risk for opportunistic infections that may lead to death.

In terms of prognosis of HIV/AIDS, it usually refers to the likely outcome of HIV/AIDS. It may also include the duration of HIV/AIDS, chances of complications of HIV/AIDS, probable outcomes, prospects for recovery, recovery period for HIV/AIDS, survival rates, death rates, and other outcome possibilities in the overall prognosis of HIV/AIDS.

The ART Cohort Collaboration is an association between 13 cohort studies from Europe and North America, it gathers data from patients who are infected with HIV-1 and started highly active antiretroviral therapy (HAART). In [17], Egger *et al.* build a prognostic model to predict the development into AIDS or death and to death alone. The prognostic models were parametric survival models based on the Weibull, loglogistic, and lognormal distributions showing that the Weibull was the one that generalized best stratified by baseline CD4 cell count and transmission group (sexual contact, drug injection, etc.).

Using an adaptive fuzzy regression technique, Don *et al.* predicted the length of survival of AIDS patients based on their CD4, CD8 and viral load counts. A comparison was made with fuzzy neural networks getting both the techniques similar results. The accuracy of the prognosis ranged between 60

and 100% depending on what year was being predicted. [16]

With data from patients diagnosed with AIDS between 1987 and 2007 from the University Hospital of Kuala Lumpur. Abdul-Kareem *et al.* developed a Classification And Regression Tree (CART), based on clinical and demographic data, to predict survival of patients during that interval. The author managed to get an accuracy between 60-93% depending on the year that's being predicted. [1]

3.6 Kidney Failure

Kidney failure, also called renal failure or renal insufficiency, is a medical condition in which the kidneys fail to adequately filter waste products from the blood. There are 5 stages of kidney failure, being the first mildly diminished renal function, stage 2 and 3 need more level of care from the physician in order to deal with the dysfunction, stages 4 and 5 require the patients to endure in active treatment in order to survive. This active treatment may come in the form of dialysis or kidney transplant.

Due to the enormous amount of people in kidney transplantation waiting list Ahn *et al.* try to predict, in [2], the one year survival of patients with kidney transplantation in order to make a more informed decision when choosing a patient for transplant. For that they built a Bayesian Network on 35,366 kidney transplants performed in the United States between 1987 and 1991.

For the same task, in [35], are reported the results of training an ANN, that was able to correctly predict 84.95% of successful transplants and 71.7% of unsuccessful transplants.

Later, in [42], Shadabi *et al.* try to improve on Petrovsky's work. For this Shadabi *et al.* used artificial neural networks instead of the more usually used statistical techniques that don't provide enough information for complex problems. They tried to improve on Petrovsky and *et al.*'s work by using a radial basis function network and prediction of the outcome at the 2 years mark. The accuracy of this approach was very similar, when used on the same data set, to the one proposed in [35] and despite the use of a range of pre-processing and ANN solutions for prediction of outcomes of kidney transplants, they found that the resultant accuracy of approximately 62% was probably too low to be of any clinical use.

Like the previous papers [42], etc. in [32] artificial neural networks are used, on data monitored when providing kidney dialysis treatment, to determine the features are related with patients' life expectancy as well as detect the existence of renal failure. It provides a model that can help and support a better understanding of a patient's evaluation results.

Kusiak *et al.* [26] used rough sets and decision trees to predict the survival time of patients undergoing kidney dialysis. Although they had a limited dataset and the lack of many important variables they show the potential for making accurate decisions for individual patients is enormous and the classification accuracy is high enough (above 75–85%) to warrant the use of additional resources and further research.

Wolfe *et al.* [47] use Cox regression analysis to calculate the LYFT score (life years from transplant), in order to develop a novel kidney allocation system based on this prediction of lifespan. The LYFT score was higher for younger patients and smaller for diabetic patients.

Li *et al.* [28] present the development of a Bayesian belief network classifier for prediction of graft

status and survival period in renal transplantation using the patient profile information prior to the transplantation. They developed two classifiers one to predict the status of the graft and another to predict its survival period. While the first one achieved a prediction accuracy of 97.8% and true positive values of 0.967 and 0.988 for the living and failed classes the second model showed only 68% accuracy.

3.6.1 Organ Failure

Prognosis work that is not about kidney failure can also be found. In [33] a study is performed where the authors used neural networks, decision trees as well as logistic regression, when trying to predict a patients' survival after a combined heart-lung transplant. The predictive models' performance in terms of 10-fold cross-validation accuracy rates for two multi-imputed datasets ranged from 79% to 86% for neural networks, from 78% to 86% for logistic regression, and from 71% to 79% for decision trees.

Also, survival analysis of liver transplant patients in Canada was done by Hong *et al.* [23] here they apply Cox proportional hazards analysis to evaluate many clinical and physical parameters' relation to the survival of the patient. A drawback of that study is that they use a very limited set of variables.

Again in liver transplant there is this study [5], where the Kaplan–Meier method and Cox regression are used to evaluate the relevance of the up-to-seven criteria, with 7 being the sum of the size and number of tumors for any given hepatocellular carcinoma (HCC), when predicting the survival of patients with hepatocellular carcinoma that perform liver transplant.

3.7 Critical Analysis

What we can see from the status of automated prognosis for the various diseases presented above, is that it is usually made without contemplating any temporal information. And, it is our opinion, that using it may considerably improve the results achieved, since it will mimic physicians' procedures.

None of the previously presented approaches takes advantage of the evolution of a patient in order to increase its prognostic accuracy, and when it is used, it is in some sort of feature that represents this evolution. The time, as a dimension, is being over-looked when building a prognostic model and it should be included in the process.

Another disadvantage of the work that has been done is that its results are data dependent, and even domain dependent [19]. states that there is no best technique to perform overall prognosis and that the result of a technique depends highly on the data being used. In other words there is no general solution that can be used in more than one dataset maintaining their performance.

Another problem identified in this review of prognosis work is that there is no evolution or search for improvement, with just a few of the papers being based and working on improving some earlier work. There is a worry to develop new prediction models before validating the already existing ones.

3.7.1 Challenges in using classification for prognosis

One of the major setbacks when trying to perform prognosis, is the fact that the data is, what is called, *censored*. This means that the value of a feature in the data is only partially known. In our case this feature is the outcome, where, for example, when predicting cancer recurrence, we know the value if the cancer has recurred while on the other case, we cannot say with certainty that it won't recur, just the amount of time that has passed since the cancer was removed. This introduces a level of uncertainty in data that needs to be handled by data mining techniques.

Other difficulty when performing prognosis using classification is finding the correct dataset to train the model. The data should be from a cohort study, what enables better measurements of the features and helps to keep track on the outcome.

Also given the characteristics of the task at hand, difference between patients, using one predictor (or feature) is rarely descriptive enough to help. Doctors use several/a set of features about patients to be able to give a prognosis, and so also needs to happen when performing the prognosis computationally. As in medical prognosis, a multivariate approach should be used, by computer-bases systems, in order to take into account the relations between features. Features, also called predictors, can be data from the patient's demographic (age, gender, etc.), clinical history, physical tests, and disease characteristics. They should be well defined and, so they could be used in real clinical situations.

Chapter 4

Approach

In the medical context, *diagnosis* is the use of patient's data, demographic and clinical, in order to understand and classify the current health condition of a patient [43]. From a formal point of view, and in the computer-based context, let A be a set of *variables* (either known as *attributes*) and C a set of possible *classes*. Given an instance $\bar{x}_i$ described by a set of m variables from A, say $\bar{x}_i = (x_{i,1}, ..., x_{i,m})$, the goal is to discover the most probable value y_i, which corresponds to its class or status, with $y_i \in C$, as in 4.1.

$$\bar{x}_i = (x_{i,1}, ..., x_{i,m}) \rightarrow y_i \tag{4.1}$$

In a classification context, this is done in two steps: first by producing a classification model M_D, based on a set of known pairs (x_i, y_i) – the *training dataset*, and second, by applying the discovered model to each instance to classify.

On the other hand, *prognosis* is the foreseeing or prediction of the risk or probability of a certain health event happening in the future, using the clinical and non-clinical data. It is the medical prediction of how the pair patient/disease is going to evolve in a specified period of time.

Considering this, then the prognosis task can be formalized as follows:

Let a patient be represented by a sequence of pairs, $(\bar{x}_i^1, y_i^1), ..., (\bar{x}_i^n, y_i^n)$, then the goal is to predict his y_i^{n+1} value – equation 4.2. Note that the different values for y_i^t may be observable (available) or non-observable at time instant t for instance i.

$$(\bar{x}_i^1, y_i^1), ..., (\bar{x}_i^n, y_i^n) \rightarrow y_i^{n+1} \tag{4.2}$$

In this context each $\bar{x}_i^k$ is known as a *snapshot*. In other words, a *snapshot* is all the data that characterizes a patient in one, single, time point.

The traditional classification approach has been applied to prognosis with modest success, as seen above. In all described cases, the evolution of single variables was not explored, and actually, the different time instances of their values were addressed separately, ignoring any possible hidden structure, in the majority of approaches. On the other hand, the analysis of time series is applied to predict the next

outcome of a single variable.

By recognizing that estimation may be used to fill unseen variable outcomes, which in turn may be used to improve classifiers accuracy, as in asap classifiers [4]. With this purpose, we propose to transform the prognosis into a diagnosis task, by estimating the values of the variables that constitute the snapshot in the future point in time.

Formally, let A be a set of attributes, C be a set of possible classes and n be the number of observations. Let the t^{th} observation, described by m variables from A, be the pair given by $(\bar{x}_i^t, y_i^t) = (x_{i1}^t, ..., x_{im}^t, y_i^t)$ that says that at observation t the instance is described by x_i^t (the observable values) and classified as $y_i^t \in C$(the predicted value). Given an instance described by an ordered set of n observations, the goal is to predict the $n + 1^{th}$ observation, as in equation 4.3.

$$(\bar{x}_i^1, y_i^1), ..., (\bar{x}_i^n, y_i^n) \rightarrow (\bar{x}_i^{n+1}, y_i^{n+1}) \tag{4.3}$$

The difference to the definition 4.2 is the need to predict the entire $n + 1^{th}$ observation, not only the predicted value y_i^{n+1}. Indeed, if there is a model M_D, that from observable values is able to determine the predicted value, it is enough to estimate the observable values in the $n + 1^{th}$ observation, and from them to predict the predicted value. This model M_D is just a simple diagnosis model as in equation 4.1.

According to this formulation, a prognosis model, M_P, is then the composition of several models: one estimation model M_{Ek} per each observable variable X_k and a diagnosis model M_D able to predict the class given an observation, as in equation 4.4, where n corresponds to the number of available observations and m the number of variables for describing each observation.

$$M_P((\bar{x}_i^1, y_i^1)...(\bar{x}_i^n, y_i^n)) = M_D(M_{E1}(\bar{x}_i^1...\bar{x}_i^n)...M_{Em}(\bar{x}_i^1...\bar{x}_i^n)) \tag{4.4}$$

By transforming the prognosis problem into a diagnosis task, the challenge becomes to be able to estimate the observation in the time point to predict, which translates into the definition of the estimation models per each observable variable.

As stated above, the art of prognosis is based on the analysis of the evolution of the different variables along time. Therefore, estimation models should be able to recognize verified evolution trends in the estimation of future values.

In this manner, we propose that an estimation model for a single variable X_k, say M_{Ek} should be a function from a sequence of the observed values to an X_k value. In particular, we propose two different approaches: the *univariate-based* and the *multivariate-based estimations*.

A *univariate-based model* for variable $X_k(UvE)$ is a function from a sequence of n values of X_k to its next value, x_k^{n+1}, as in equation 4.5, where Dom_{Xk} represents the domain of variable X_k. These models only explore the individual values of a variable, ignoring any influence from other variables.

$$\begin{aligned} M_{UvEk} &: [Dom_{Xk}]^n \rightarrow Dom_{Xk} \\ M_{UvEk}(x_k^1...x_k^n) &= x_k^{n+1} \end{aligned} \tag{4.5}$$

On the counterpart, a *multivariate-based model* for variable $X_k(MvE)$ is a function from a sequence of n vectors of m variables, including X_k, to its next value, x_k^{n+1} – see equation 4.6.

$$M_{MvEk} : [Dom_{X1} \times ... \times Dom_{Xm}]^n \rightarrow Dom_{Xk}$$
$$M_{MvEk}(\bar{x}^1 ... \bar{x}^n) = x_k^{n+1} \tag{4.6}$$

By receiving a sequence of multi-values, recorded along n observations, multivariate estimator is able to contemplate the interdependencies among the different values, and having more informed inputs, is expected to output better estimations.

4.1 Estimation Algorithm

From the previous formulation, the algorithm required to train the new classifier is simple, and is similar for both estimation models.

Algorithm 1 Pseudocode for Univariate Estimation training

procedure UNIVARIATEESTIMATION(Dataset D, Function alg_{estim}, Function alg_{class}, int ρ)
  $//D$ – the training dataset with
  $//\quad D = \{(\bar{x}_i^t, y_i^t) : \forall i, t : 1 \leq i \leq |D| \wedge 1 \leq t \leq n\}$
  $//alg_{estim}$ – the estimation algorithm
  $//alg_{class}$ – the training algorithm
  $//\rho$ – the number of observations to use
  $A \leftarrow$ {the set of attributes describing D}

  // Training each estimation model
  for each variable X_k in A **do**
    $D_k \leftarrow \pi_{Xk}(D) = \{(x_{ik}^{n-\rho}, ..., x_{ik}^n) : \forall \bar{x}_i \in D\}$
    $M_{Ek} \leftarrow alg_{estim}(D_k)$
  end for

  // Estimating n+1 snapshot
  for each variable X_i in D **do**
    for each variable X_k in A **do**
      $x_{ik}^{n+1} \leftarrow M_{Ek}(x_{ik}^{n-\rho}, ..., x_{ik}^n)$
      $D^{n+1} \leftarrow D^{n+1} \cup \{(x_{i1}^{n+1}, ..., x_{im}^{n+1}, y_i^{n+1})\}$
    end for
  end for

  // Train the diagnosis model
  $M_D \leftarrow alg_{class}(D^{n+1})$

  // Output the composition of models
  Return $M_D \circ (M_{E1}, ..., M_{Ek})$
end procedure

Note, that the dataset has to be composed of records containing n snapshots, as described before, and ρ has to be less or equal to n. In terms of the classification training algorithm, it should be any tabular one, like a decision tree learner, an algorithm for training neural networks or just naïve Bayes.

Algorithm 2 Pseudocode for Multivariate Estimation training

procedure MULTIVARIATEESTIMATION(Dataset D, Function alg_{estim}, Function alg_{class}, int ρ)
  $//D$ – the training dataset with
  $//\qquad D = \{(\bar{x}_i^t, y_i^t) : \forall i, t : 1 \leq i \leq |D| \wedge 1 \leq t \leq n\}$
  $//alg_{estim}$ – the estimation algorithm
  $//alg_{class}$ – the training algorithm
  $//\rho$ – the number of observations to use
  $A \leftarrow$ {the set of attributes describing D}

  // Training each estimation model
  for each variable X_k in A **do**
    $D_k \leftarrow \pi_{Xk}(D) = \{(\bar{x}_i^{n-\rho}, ..., \bar{x}_i^n) : \forall \bar{x}_i \in D\}$
    $M_{Ek} \leftarrow alg_{estim}(D_k)$
  end for

  // Estimating n+1 snapshot
  for each variable X_i in D **do**
    for each variable X_k in A **do**
      $x_{ik}^{n+1} \leftarrow M_{Ek}(\bar{x}_i^{n-\rho}, ..., \bar{x}_i^n)$
      $D^{n+1} \leftarrow D^{n+1} \cup \{(x_{i1}^{n+1}, ..., x_{im}^{n+1}, y_i^{n+1})\}$
    end for
  end for

  // Train the diagnosis model
  $M_D \leftarrow alg_{class}(D^{n+1})$

  // Output the composition of models
  Return $M_D \circ (M_{E1}, ..., M_{Ek})$
end procedure

The difference between the *UvE* and *MvE* models is on the creation of the estimation models (line 10), in particular on the creation of the training dataset for each variable. While for the univariate model , Algorithm 1, it consists on the projection of D in relation to each X_k, the multivariate model, Algorithm 2, uses the entire set of variables. In other words, instead of just using the k^{th} variable values, the entire instances are used. In both cases, ρ corresponds to the number of snapshots to keep in the dataset. Since, it is usual that the instants more significant for determining the next value are the previous ones, only the last ρ snapshots are used.

After training the estimation model for each variable, the diagnosis model is learnt from the estimated snapshot for instant $n + 1$ and the known class label. Then, the algorithm outputs the model resulting from the composition of the different estimators and the diagnosis model learnt from the estimated values (line 23).

4.2 Implementation Issues

From the solution presented in the previous section there are few steps that deserve some special attention.

As it is possible to see in the pseudocode, of both approaches, a few parameters are configurable as arguments of our method. Namely the method to use in the estimation step,alg_{estim}, the method to use in the classification step,alg_{class} and the amount of time steps to be used, ρ.

ρ is the amount of time steps to be used, this parameter allows us to easily change the amount to information to be used on each run of the method. In this work, when a smaller amount than the number of steps available is passed to the algorithms the more recent steps are used. For example, if we have 10 time steps, and $\rho = 5$, the data from time steps 5 through 9 are taken into account.

alg_{class} is the method that is used to create the Diagnostic model. This parameter can be any classification technique, like Naive Bayes, Decision Trees or Support Vector Machines. In this dissertations we decided to use the techniques whose output is an easily understandable model. This gave us techniques like Decision Trees, Naive Bayes and Logistic Regression.

alg_{estim} is the method that is used in the estimation step of our solution. This technique is used to estimate the future value of every feature and it can be a various number of techniques. The technique can be chosen based on the characteristics and amount of the data available. For example, in this dissertation we use linear regression for the estimation when the data is numeric and logistic regression when it is nominal. We also try decision trees for the estimation and Hidden Markov Models.

This freedom in the configuration allows for a more generic and adaptive method that can be used on the most various data.

4.3 Example

Let's assume we have clinical and laboratorial data for a set of patients with Alzheimer's disease. The data itself could be represented like in Table 4.1.

Patient ID	Gender	Age	HBP	LBP	Degree of Progression	Time Step
25	M	65	160	88	1	0
25	M	65	140	90	1	1
25	M	65	138	85	1	2
25	M	65	134	81	1	3
25	M	65	141	88	2	4

Table 4.1: Patient data.

As seen in Table 4.1, we have N time steps of data, in this case 5, we are going to use those in order to perform prognosis. As described in the last section we have two approaches for doing that: in the first, we only use the past values of a feature to predict its future, that is, for example with the high blood pressure, we can use HBP at times $0, 1, 2, 3, 4$ in order to predict it at instant 5, and the same for every other feature.

In Table 4.2 we can see the data to be used in order to predict HBP at instant 5.

HBP_0	HBP_1	HBP_2	HBP_3	HBP_4	HBP_5
160	140	138	134	141	?

Table 4.2: Data used on the univariate estimation approach.

In the second option, we use every feature value to predict each one. In this case we would use the time steps 0 to 4 of every feature, as can be seen in Table 4.3 to predict HBP 5. And the same for every other feature.

Patient ID	Gender	Age_0	HBP_0	LBP_0	...	Age_4	HBP_4	LBP_4	HBP_5
25	M	65	160	88		65	141	88	?

Table 4.3: Data used on the multivariate estimation approach.

At the end of both approaches the result is the same, which is a complete snapshot at time step 5.

Using that predicted data on a diagnostic model that was previously trained on data (like in Table 4.1). We would get the final prognosis.

Chapter 5

Validation and Experimental Results

5.1 Dataset Description

In order to validate our proposal, we used two different real datasets from the healthcare field: the ALS and the Hepatitis datasets.

5.1.1 ALS Dataset

The ALS dataset[1] includes information from over 8500 ALS patients who participated in industry clinical trials. The data include demographic, family and medical history, the patient's history in terms of ALS symptoms, clinical and some laboratorial data. From these, we used a subset composed by the patients that had demographic data, had performed Slow Vital Capacity exams, as well as measurements of their vitals, counting 13 variables: gender, age, height, percentage of normal, subject liters (trial 1, 2 and 3), blood pressure (systolic and diastolic), pulse, respiratory rate, temperature and Weight.

The dataset is mostly composed by numeric attributes that were normalized into the range [0,1] using the *Feature Scaling* method, equation 5.1. Where X' is the new value, X the current value, X_{min} and X_{max} the minimum and maximum value of that feature, respectively, and a and b are the new range minimum and maximum, or in other words $a = 0$ and $b = 1$.

$$X' = a + \frac{(X - X_{min})(b - a)}{(X_{max} - X_{min})} \tag{5.1}$$

The outcome is a score that evaluates the state of the disease between 0 (severe) and 48 (normal), discretized into 4 classes (aggregations of 12 points). The subset contains 578 patients, with 5.9% for the 1st class, 22.3% for the 2nd, 29.1% for the 3rd and 42.7% for the 4th, as seen in Figure 5.1.

5.1.2 Discret ALS Dataset

Because some of the techniques used cannot be directly applied to numeric data the following discretization was applied on the ALS dataset.

[1] https://nctu.partners.org/ProACT/

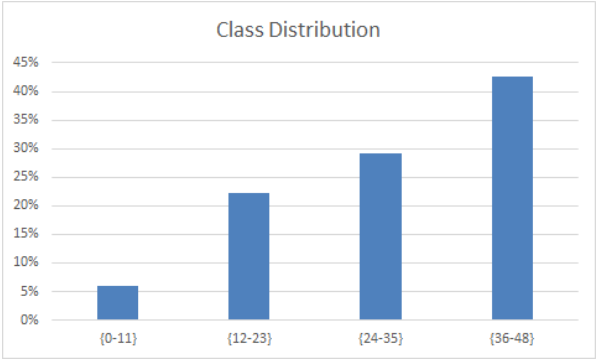


Figure 5.1: Class distribution of the ALS dataset.

For this discretization some domain knowledge was used to find a discretization that makes more sense than just dividing the feature into n bins. In Tables 5.1, 5.2, 5.3, 5.4, 5.5and Figure 5.2 the discretization used for the Blood Pressure, Pulse, Respiratory Rate, Percentage of Normal, Temperature and Weight features can be seen, respectively.

For the features where no decent class could be found, their values were discretized into n equally sized bins, with $n = 6$.

	Blood Pressure	
	Systolic	Diastolic
Normal	<120	<80
Pre-Hypertension	120-139	80-89
High Blood Pressure Stage 1	140-159	90-99
High Blood Pressure Stage 2	169-179	100-109
Hypertensive Crysis	>= 180	>= 110

Table 5.1: Discretization for both Blood Pressure features.

	Pulse
Slow (Bradycardia)	<60
Resting	60-100
Fast (Tachycardia)	>100

Table 5.2: Discretization for the Pulse feature.

5.1.3 Hepatitis Dataset

The Hepatitis dataset was made available as part of the ECML/PKDD 2005 Discovery Challenge[2], it contains data about 771 patients, and more than 2 million examinations between 1982 and 2001. Based on the work of [46] the data was reduced to the most significant exams. In the end 17 variables were used: gender, age, birthdate, birth decade, 11 of the most significant exams (GOT, GPT, ZTT, TTT, T-BIL, D-BIL, I-BIL, ALB, CHE, T-CHO and TP) and the results from the active biopsies at the time of the exams (type, activity and fibrosis).

[2]http://lisp.vse.cz/challenge/CURRENT/

Respiratory Rate	
Slow	$<$12
Normal	12-20
Fast	20-24
Very Fast	$>$24

Table 5.3: Discretization for the Respiratory Rate features.

% of Normal	
Very Low Breathing Capacity	$<$50
Deteriorating	50-80
Normal	80-100
Athlete	$>$100

Table 5.4: Discretization for Percentage of Normal feature.

Fibrosis is the objective class and it is described by integer values between 0 (no-fibrosis) and 4 (most severe). The subset contains 488 patients and the following distribution of classes: 2.05% of 0, 45.9% for 1, 21.35% for 2, 15.19% for 3 and 15.40% for 4, as seen in Figure 5.3.

5.2 Validation Techniques

Both of the datasets used describe a disease's progression, in this case Hepatitis and ALS.

The reason to use 2 datasets describing different diseases, instead of just one, is to show the generalizability of our approach which hopefully will show similar results in both.

As previously stated our objective was to focus on finding a way to use the time dimension when performing prognosis, use a patients' evolution over time, and with that build a generalizable technique whose results would not be so dependent on the data.

For these reasons, and because there are other works where these datasets have been used and preprocessed, [46], we can base our work on their results, not exploring other pre-processing techniques, and use most of our time on the actual task at hand.

In Table 5.6 we can see notation used. The table is composed by the positive and negative predictions,+P and -P respectively, and the positive and negative real values, +R and -R also respectively. Then TP means the True Positives, TN the number of True Negatives and similarly FP and FN the number of False Positives and False Negatives Respective, respectively. The sum by rows result in PP and PN which are the number of predicted positives and negatives while the sum by columns results in RP and RN, the number of Real Positives and Real Negatives, respectively. The sum of all the real and predicted values gives the size of the population, Pop.

Having the two datasets, the model built and the notation defined the usual evaluation metrics will be used, like *accuracy*, *precision*, *F-measure*, sensitivity and specificity.

Accuracy is the ratio of correct classifications over all the cases,

$$Accuracy = \frac{TP + TN}{Pop} \tag{5.2}$$

Temperature	
Hypothermia	<36.5
Normal	36.5-37.2
Fever	>37.2

Table 5.5: Discretization for the Temperature feature.

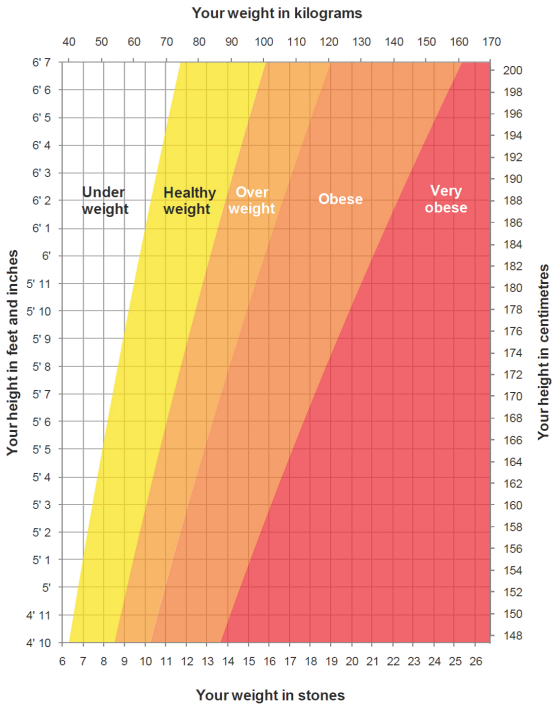


Figure 5.2: Discretization for the Weight feature.

Precision, also called positive predictive value, is the degree to which several measurements provide answers very close to each other. It is an indicator of the scatter in the data.The lesser the scatter, higher the accuracy.

$$Precision = \frac{TP}{PP} \tag{5.3}$$

Sensitivity, also called *true positive rate* or *recall*, is the ability of the model to identify positive cases, in other words this metric shows the overall percentage of correctly identified classifications.

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.4}$$

Because only measuring the ability to identify the positive cases is useless (a system that always classified something as positive would have a sensitivity of 1), we also use *specificity*. Similarly, *specificity* measures the ability of the system to identify the negative cases.

$$Specificity = \frac{TN}{FP + TN} \tag{5.5}$$

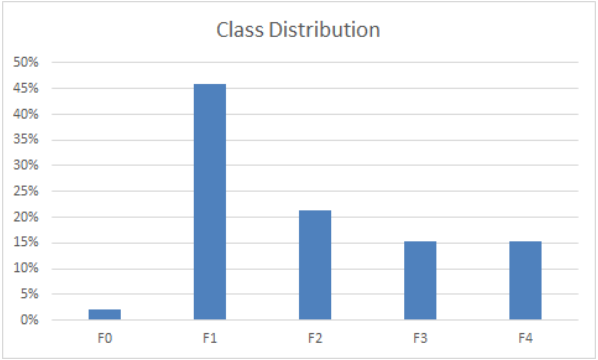


Figure 5.3: Class distribution of the Hepatitis dataset.

	+R	**-R**	
+P	TP	FP	PP
-P	FN	TN	PN
	RP	RN	Pop

Table 5.6: Notations in a binary contingency table. Color coding indicates correct (green) and incorrect (pink) rates or counts in the contingency table.

F-measure, also called *F1 Score*,is ..... Note that the F-measure effectively references the True Positives to the Arithmetic Mean of Predicted Positives and Real Positives, being a constructed rate normalized to an idealized value.

$$F\text{-}measure = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{5.6}$$

Because the use of time is the cornerstone of this work it is also needed to see if it was actually relevant. To do this we will look for the use of temporal patterns in the model created and their relevance in the decision process. I.e. if the model is a decision tree the closer to the root this temporal pattern rules, are the more relevant they are in the decision, and that shows the importance of time in this matter.

5.3 Experimental Results

In this section we will present the experimental results of this case study. We begin by describing the baseline for comparison and then present the performance of our approaches using different techniques.

This sections begins by presenting the baselines, the diagnosis models, the models that perform prognosis similarly to diagnosis. // Then, separating by technique used in the estimation phase (regression, decision tree or HMM), we present the performance of the different approaches compared to the baseline. It is important to note that, per estimation method, we separate the results by phase. This means that firstly we introduce the estimation performance and, after that, the overall prognosis performance using those same estimations.

All The results shown use the *accuracy* metric because it is a universal metric and the *F-measure* metric showed no significant change to the presented results. // We will also show a more in depth analysis of some cases where all the metrics will be taken into account. We will show only some, due to

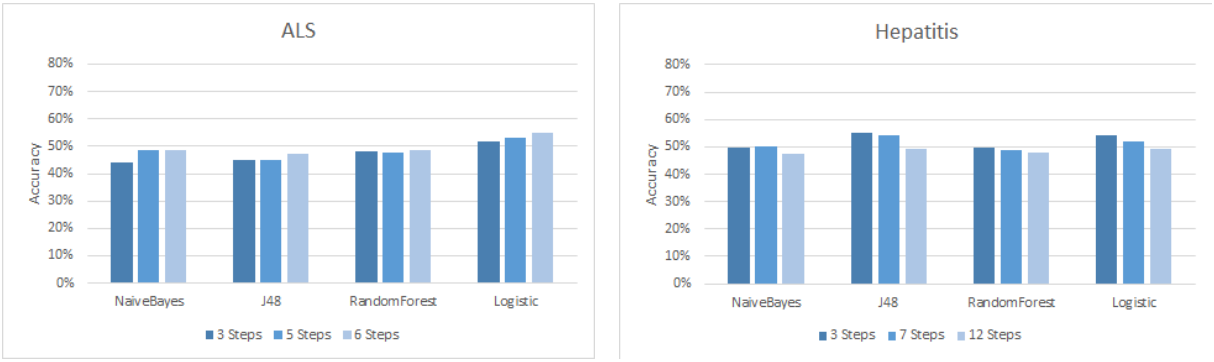
the large number of data and graphs generated by this analysis.

All tests were ran on an Asus U36SD with an Intel® Core™ i5 2430M/2410M Processor, clocked at 2.40 GHz, 8Gb of Ram and running 64 bit Windows 8.1 Pro. The Weka toolkit[3], version 3.7.10, was used for the regression and decision tree estimations and classifications. While to perform the HMM estimations, the package HMM [4] for the programming language R was used.

5.3.1 Diagnosis Model

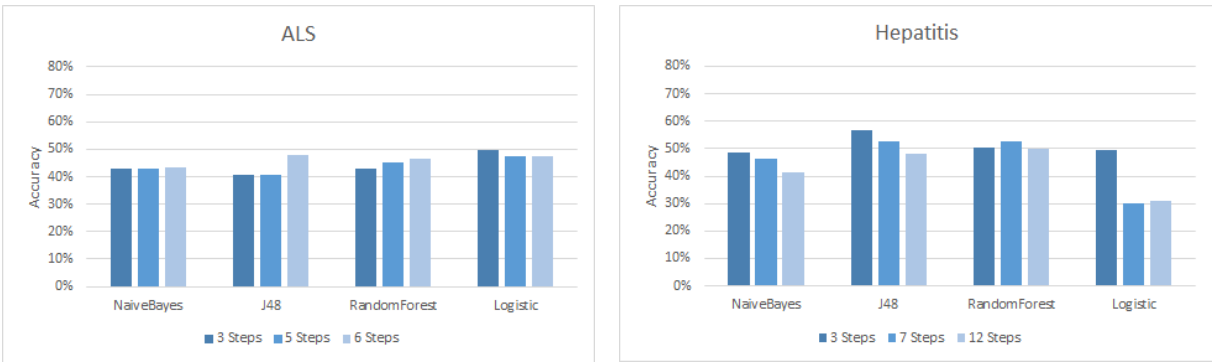
As a baseline for comparison with the proposed approaches we used two models. *BaselineSingleObservation* is a diagnostic model where a single observation in time is used to perform the prognosis. In other words, the state of a patient at instant n is used to predict his class at instant $n+1$. On the other hand, *BaselineMultipleObservation* instead of using a single observation, uses multiple observations: all information is used here to predict the class at instant $n+1$.

A collection of techniques were used with these models, with both achieving similar results: the accuracy ranged between 40% and 55%, depending on the dataset, technique and number of time points used, as seen in Figure 5.4 and Figure 5.5.



(a) ALS (b) Hepatitis

Figure 5.4: BaselineSingleObs accuracy (several classifiers and number of observations).



(a) ALS (b) Hepatitis

Figure 5.5: BaselineMultipleObs accuracy (several classifiers and number of observations).

[3] http://www.cs.waikato.ac.nz/ml/weka/
[4] http://cran.r-project.org/web/packages/HMM/index.html

It is interesting to note that the accuracy is almost constant for ALS, for different techniques and number of time steps. However, it is clear that for Hepatitis those variables are determinant for reaching higher accuracy. The best results tend to be achieved using 3 time steps and J48 and Logistic Regression.

It is also interesting to note that those differences are smoother in the presence of the multiple observations.

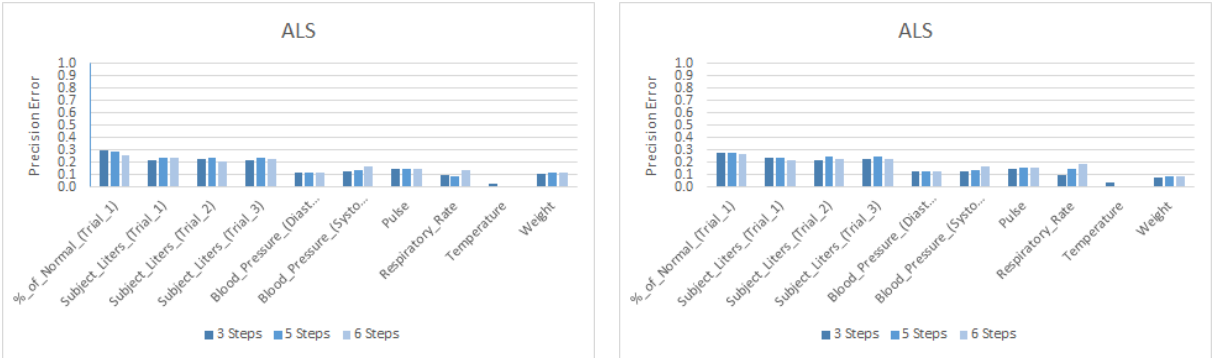
5.3.2 Regression Techniques

In this section the results of applying regression techniques with out approaches is shown. Because of the different characteristic of the datasets (numeric versus nominal attributes), different regression techniques have been applied in the estimation phase of this work, namely linear regression for the numeric datasets and logistic regression for the nominal.

Estimation Models

As previously mentioned we begin the presentation of our results by analyzing the estimation phase performance.

The results with the univariate and multivariate estimation models for the ALS dataset (numeric) can be seen in Figure 5.6. These models were built using linear regression as previously said. Both estimation models were applied using a different number of observations, $3, 5$ and 6 time steps. Because the dataset is numeric, we evaluated our estimation by the error, the distance, to the actual value at time t_{n+1}. Both the univariate and the multivariate estimation model presented an average estimation error of around 0.165, with features having errors as high as 0.30 and as low as 0.01.



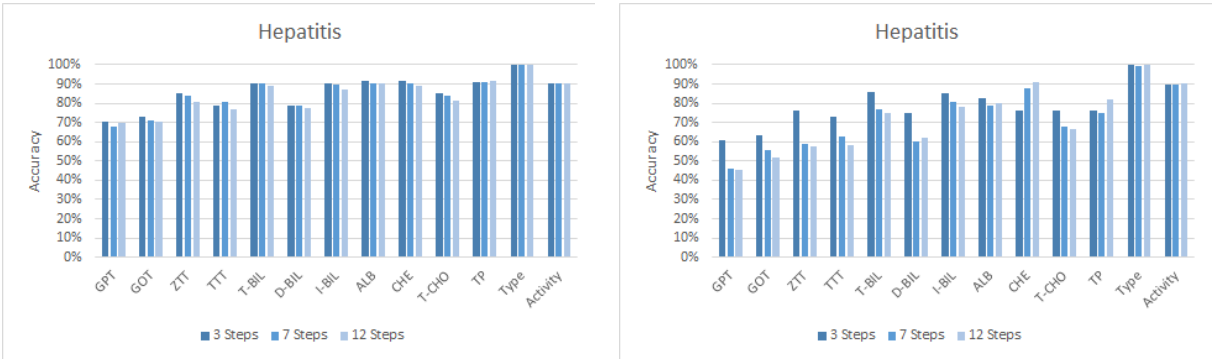
(a) Univariate Estimation

(b) Multivariate Estimation

Figure 5.6: Impact of the number of observations on the accuracy of the linear regression estimation models for each variable, in the ALS dataset.

We can also see in Figure 5.7 the results of using Logistic Regression on the Hepatitis dataset. In both cases the average accuracy of estimation rounded the 80% range, with the multivariate model being consistently a bit worse than the model that uses a single variable.

While in the ALS case we cannot descriminate any clear difference between the two estimation models in the Hepatitis dataset, as previously stated, the multivariate approach performed a bit worse



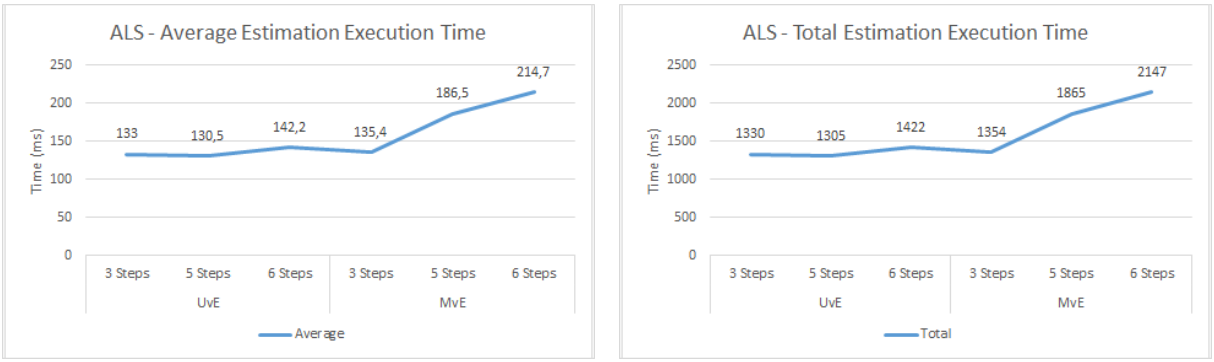
(a) Univariate Estimation

(b) Multivariate Estimation

Figure 5.7: Impact of the number of observations on the accuracy of the logistic regression estimation models for each variable, in the hepatitis dataset.

than the univariate model. Also in the ALS case we can see a slight trend of improvement in the estimations with the increase of snapshots used, while in the other case the oposite is noteable.

In Figure 5.8 and Figure 5.9, we can see the performance analysis, in milliseconds, of the estimation phase. It is important to note that no significant difference was noticed between the univariate and multivariate estimations when using linear regression. The same cannot be said about logistic regression where the overall estimation of the features on the multivariate approach took about $(Nsteps \times 3)$ times more than the univariate.



(a) Average Estimation Time

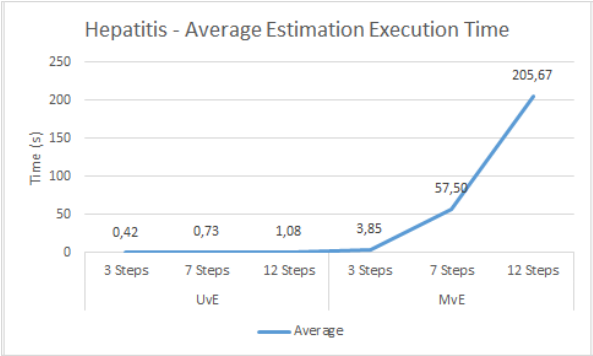
(b) Total Estimation Time

Figure 5.8: Execution time of feature estimation in the ALS dataset using Linear Regression.

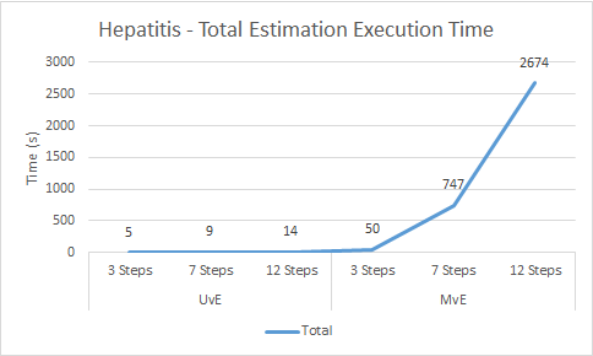
Prognosis Results

Finally we will evaluate the performance of the prognosis model, using the estimations presented in the previous section.

The overall prognosis accuracy achieved by using various techniques with our approaches, on the predictions achieved by using regression techniques, can be seen in Figure 5.10. It is observable that decision tree classifiers outperform the other techniques, with both of them, J48 and RandomForest, achieving better accuracies than the other techniques as well as the corresponding baselines. In the ALS dataset, J48 and RandomForest were able to improve the results in more than 15%, with both estimation models. In the Hepatitis dataset the UvE model clearly improved the final accuracy of the



(a) Average Estimation Time

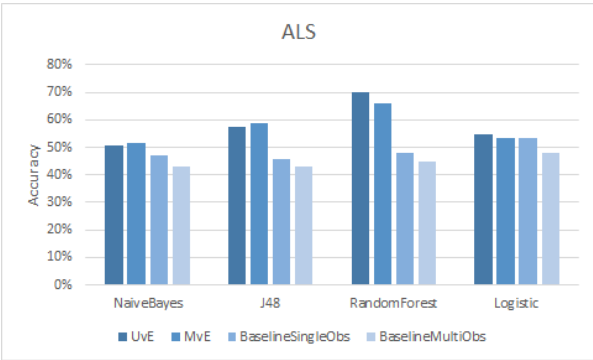


(b) Total Estimation Time

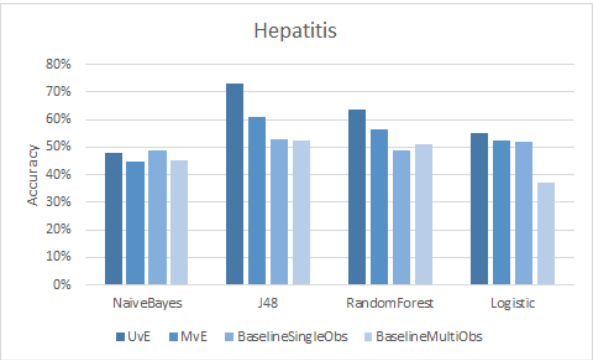
Figure 5.9: Execution time of feature estimation in the hepatitis dataset using Logistic Regression.

prognosis, achieving an improvement of about 20%. The MvE model didn't do so well only improving the final prognosis by 5%.

Figure 5.11 shows the relation between the number of observations and the final accuracy of the prognosis, using both, UvE and MvE estimation models, and a variety of techniques. It interesting to note that in the ALS dataset and using linear regression we can see distinctly that the number of steps used and the overall accuracy are directly proportional. In the Hepatitis datasets the opposite relation is notable, as the number of time steps used increases the accuracy decreases or maintains. This leads us to think that, in this dataset, the furthest points in time are not as relevant to perform the prognosis as the ones closer to t_{n+1}. If this happens because of the nature of the disease or because of the characteristics of the data we are not certain.



(a) ALS - linear regression estimations



(b) Hepatitis - logistic regression estimations

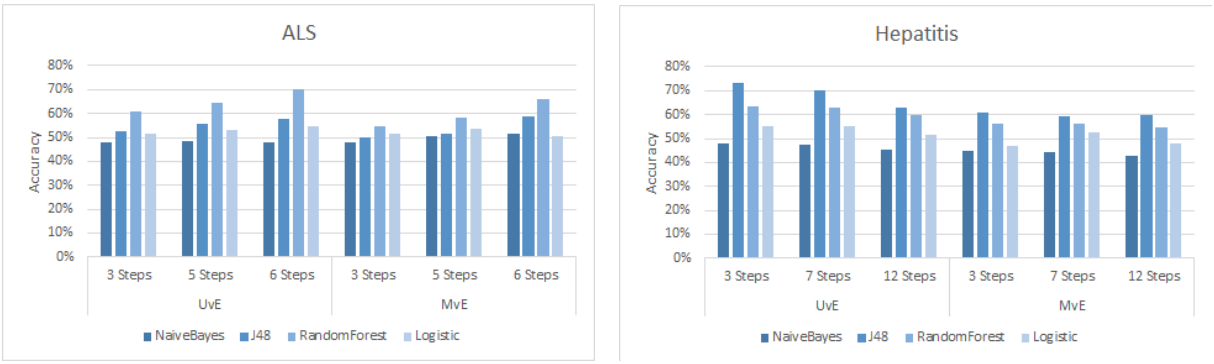
Figure 5.10: Accuracy of different models.

In Figure 5.12 the *F-measure* is show for the different approaches on both datasets.

5.3.3 Decision Tree

In this section, J48 [5] was used as the estimation technique. J48 is an implementation of Quinlan's C4.5 algorithm [37]. Because J48 cannot handle numeric classification this technique was used on the hepatitis dataset and on the discretized version of the ALS dataset. The results were as follows.

[5] http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html



(a) ALS - linear regression estimations

(b) Hepatitis - logistic regression estimations

Figure 5.11: Impact of the number of observation on prognosis models.

Estimation Models

Again, before assessing the results of our prognosis approach, we evaluate the impact of the number of observations used, on the quality of the estimations made through the two estimation models proposed.

Figure 5.13 and Figure 5.14 show the results with univariate and multivariate estimation models. Both estimation models were applied using a different number of observations.

The ALS dataset estimations perform very poorly with an average of 21% and 32% on the UvE and MvE models, respectively. While the results are poor, there can be noticed a slight improvement by using the multivariate model. This result may be caused by the discretization that was used.

On the Hepatitis dataset both models reach similar levels of accuracy, with quite good results for the majority of the Hepatitis variables (above 80%). It is interesting that there is a slight trend to increase the accuracy as the number of observations get higher.

Despite our expectations, it seems that there is no improvement on using multivariate-based estimation.

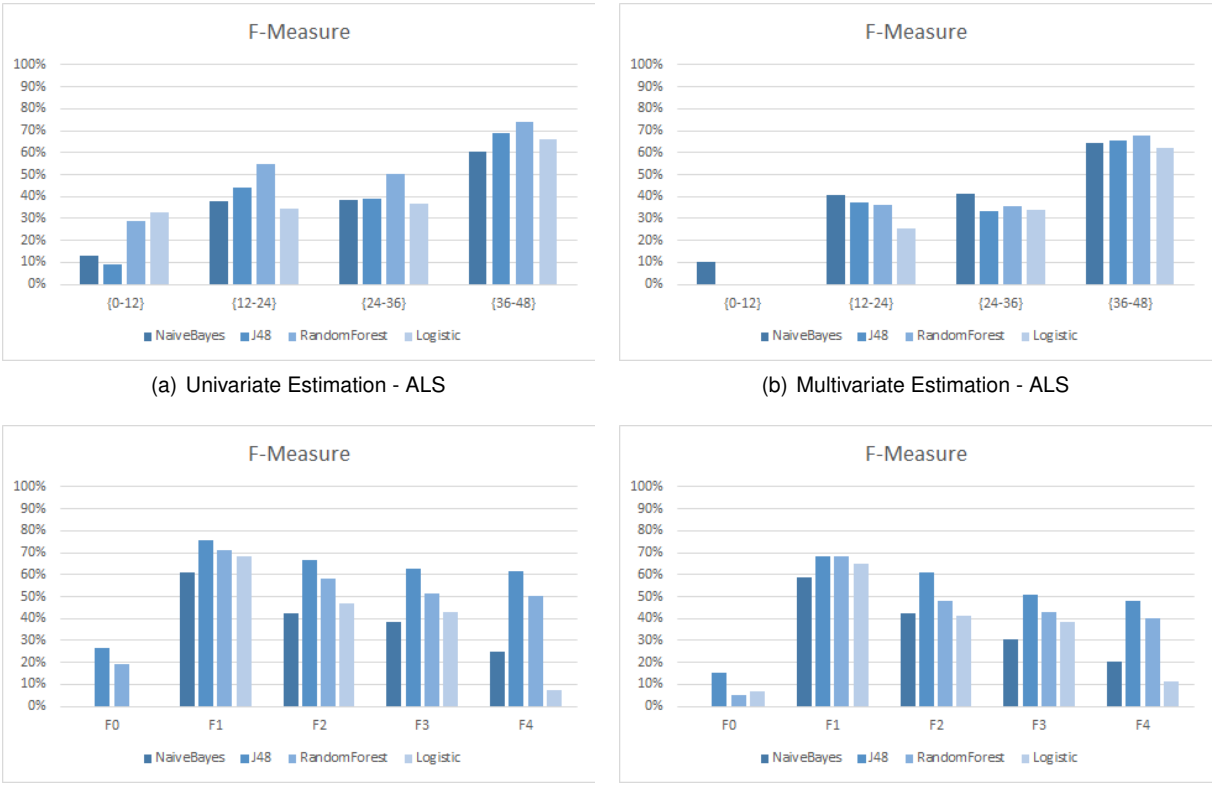
In Figures 5.15 and 5.16, the average and total time of execution, in milliseconds, for the feature estimation phase, using J48, is presented.

It is important to note that, on the Hepatitis dataset, even though the estimation using logistic regression had similar results (Figure 5.7), when looking into to the accuracy of the estimation, it took much longer to estimate the results ($3\times$ more in the fastest case and $800\times$ more in the slowest, (Figure 5.9)).

On the ALS case, the time performance was very similar to the linear regression estimation, (see Figure 5.8), with no significant difference worth mentioning.

Prognosis Results

The overall prognosis accuracy achieved by using different techniques on the decision tree estimations can be seen in Figure 5.17. The improvements on the accuracy of our approach are always present when compared to the ones achieved by baseline models also shown in the same figure. In the Hepatitis dataset the improvements round about 20%, while in the ALS dataset the improvements are more modest with the UvE model improving around 5%, with most classification techniques, and achieving



(a) Univariate Estimation - ALS

(b) Multivariate Estimation - ALS

(c) Univariate Estimation - Hepatitis

(d) Multivariate Estimation - Hepatitis

Figure 5.12: Different metrics for the overall prognosis using logistic regression on the Hepatitis dataset.

similar results as the baseline with the MvE model.

It is also curious to note that even though the MvE estimations were a little better, the overall prognosis using this estimations was consistently worse its univariate counterpart.

Figure 5.18 shows the relation between the number of observations and the final accuracy of the prognosis, using both, UvE and MvE estimation models, and a variety of techniques. Again, and similarly to the regression estimations, it is interesting to note that on the hepatitis case the higher number of observations become prejudicial to the UvE model, which means that the values from the long past do not help to estimate future values. And on the ALS case you can see the inverse relation while much less noticeable than when using linear regression to perform the estimations.

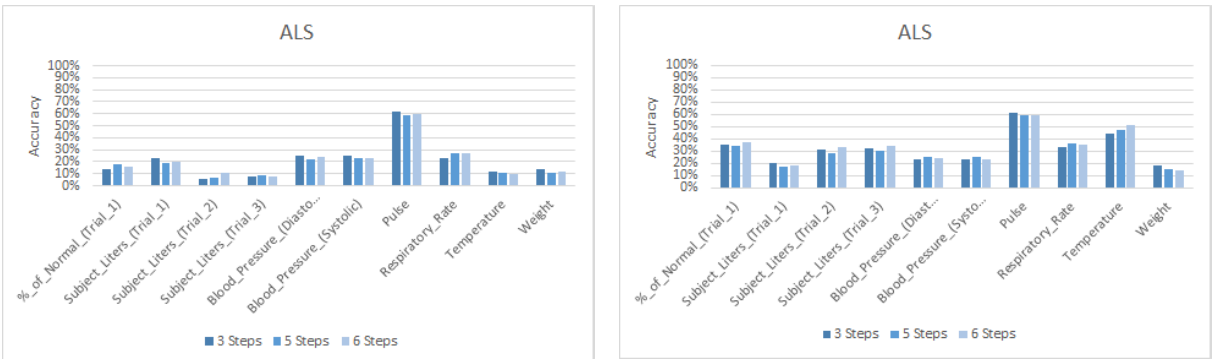
Again there is no clear difference between both estimation models, but decision trees (through C4.5 algorithm – J48 and the RandomForest ensemble) always perform better than the other models.

5.3.4 HMM

In this final section, HMMs were used in the estimation phase. Because HMMs cannot handle directly numeric classification the same discretization of the ALS dataset was used.

The HMMs we used had one state per time step, so if we had a sequence with data from 7 time instances our HMM would have 7 states. All the probability distributions, λ, would then be initialized randomly and normalized so that the probability distribution equals 1.

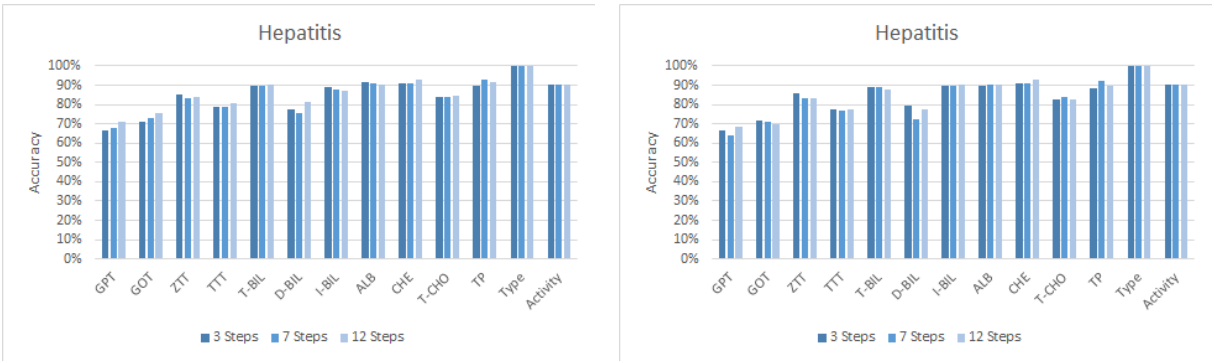
We would then train one HMM per class, using the Baum-Welch algorithm, which is used to adjust



(a) Univariate Estimation

(b) Multivariate Estimation

Figure 5.13: Impact of the number of observations on the accuracy of the decision tree estimation models for each variable using both univariate and multivariate models on the discrete ALS dataset.



(a) Univariate Estimation

(b) Multivariate Estimation

Figure 5.14: Impact of the number of observations on the accuracy of the decision tree estimation models for each variable using both univariate and multivariate models on the Hepatitis dataset.

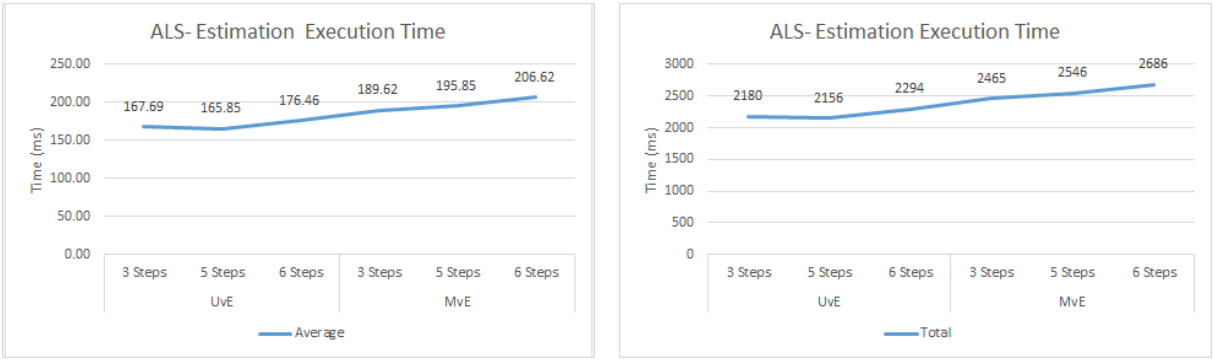
λ to maximize the likelihood of the training set. The training set was composed by a subset of the data that had the specific class.

The prediction phase was done by concatenating all the possible classes to the observed sequence and applying the forward algorithm with that sequence and the matching class HMM. The forward algorithm calculates the likelihood that the HMM generated the sequence. The sequence with the highest likelihood was chosen and so the concatenated class was the estimation.

Estimation Models

Figure 5.19 shows the results with univariate and multivariate estimation models in the ALS dataset, respectively, and Figure 5.20 shows the same results when applied on the Hepatitis dataset. A different number of time steps were used with each estimation model.

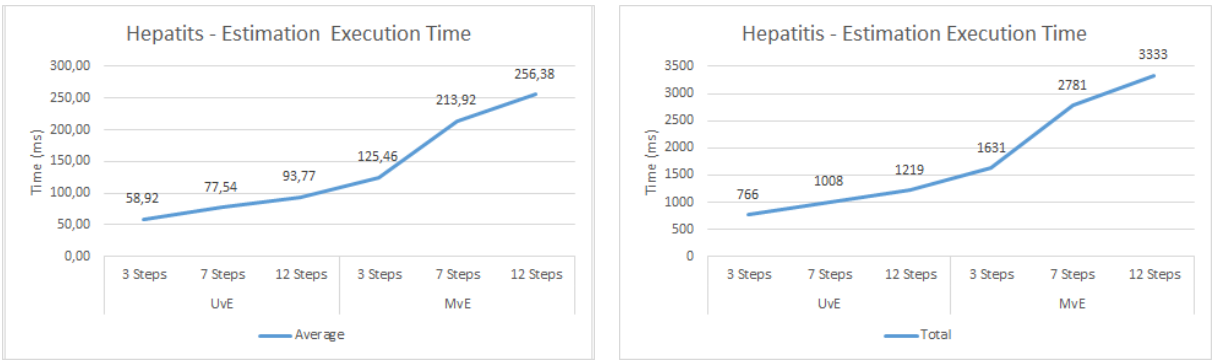
On the ALS case, the average estimation accuracy was of 13% and 4%, in the UvE and MvE approaches respectively. This poor performance might be caused by the discretization that was applied to the data. It can also be seen that the number of snapshots used has and inverse relation with the accuracy of the estimations, with the accuracy decreasing with the increase of the number of time steps used.



(a) Average Estimation Time

(b) Total Estimation Time

Figure 5.15: Execution time of feature estimation in the hepatitis dataset using Decision Trees.



(a) Average Estimation Time

(b) Total Estimation Time

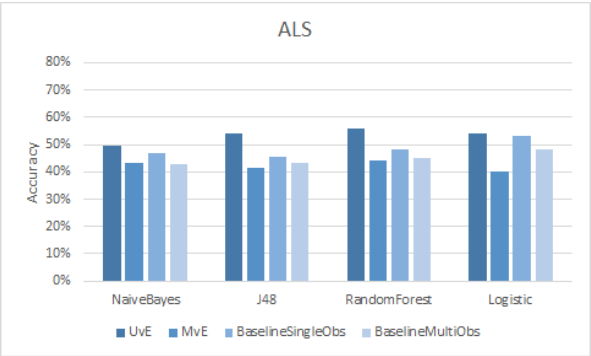
Figure 5.16: Execution time of feature estimation in the hepatitis dataset using Decision Trees.

On the other hand, on the hepatitis dataset the average estimation accuracy was of 83% and 49%, in the UvE and MvE approaches respectively. This is a very similar results to when using regression of decision trees on the UvE model of this dataset. The MvE performed considerably worse that any other technique used.

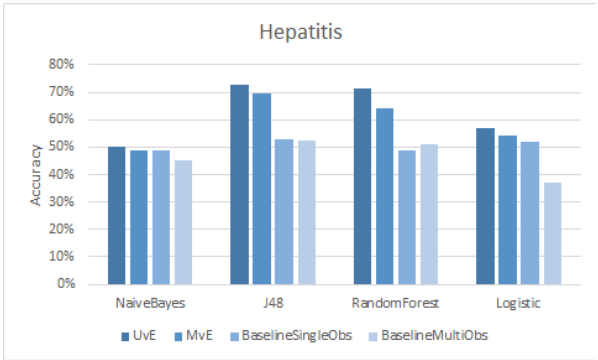
In Figure 5.21 and 5.22, we can see the performance analysis, in seconds, of the estimation phase using HMMs. As previously said the Baum-Welch algorithm was used in this step. This algorithm tries to maximize the likelihood of the training set and its result, the model's configuration, is a local maximum. Because of this fact this algorithm is ran X times, X iterations, to try to find the optimum solution. The execution times presented here represent the time taken to run 1 iteration of the algorithm, the estimation results shown, were achieved by performing 50 iterations.

While in the ALS case, this technique presented a very bad performance. In the Hepatitis case, it achieved similar results in the estimation accuracy, in the UvE approach while the MvE performed significantly worse than any other, (see Figure 5.7 and Figure 5.14).

But, in both cases, the time to make those estimations was much longer than any other technique. Even longer than the Logistic regression alternative on the hepatitis dataset, being between 8 and $200\times$ slower than it and between 100 and $6880\times$ slower than decision trees (Figure 5.9 and Figure 5.16).

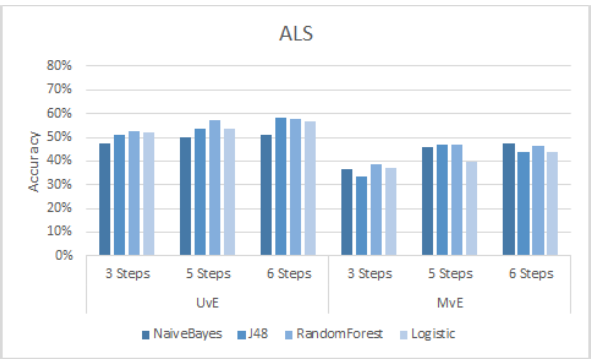


(a) ALS - accuracy using decision tree estimations

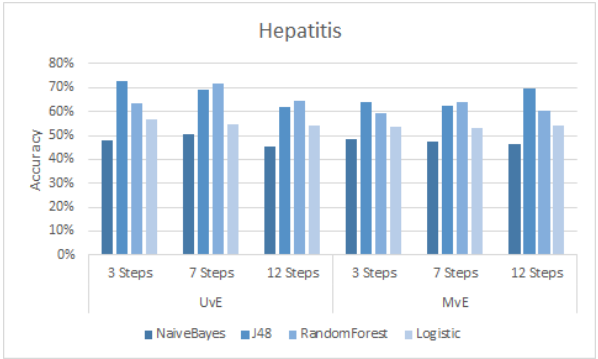


(b) Hepatitis - accuracy using decision tree estimations

Figure 5.17: Accuracy of different models using the decision trees estimations.



(a) ALS - accuracy using decision tree estimations



(b) Hepatitis - accuracy using decision tree estimations

Figure 5.18: Impact of the number of observation on prognosis models using the decision trees estimations.

Prognosis Results

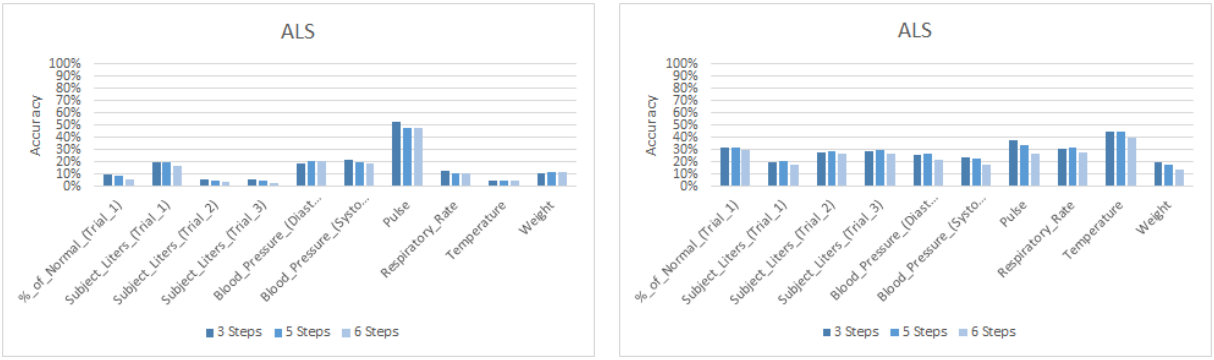
In Figure 5.23 the accuracy of the various models is shown using different techniques and the HMM estimations.In the ALS case, as is was to expect the overall accuracy was the worst found so far, being close to the baseline, but in most cases a bit worse. In the Hepatitis case it is curious to note that, even though the accuracy of the estimations in the univariate model is very similar, the overall prognosis accuracy is considerably worse. This might derive from the fact that the correct estimations made with this technique are not as relevant to the final prognosis as the ones made by using decision trees.

The impact of the amount of time steps, number of observations, used can be seen in Figure 5.24. Here no clear relation can be extracted.

5.3.5 Discussion

Currently, medical practice is helped by a variety of computer-aided tools, dedicated to help physicians taking the most appropriate decisions. However, despite the importance of prognosis, it did not deserved dedicated tools, and in the majority of situations, it has been addressed as a simple diagnosis problem, without exploring the temporality involved.

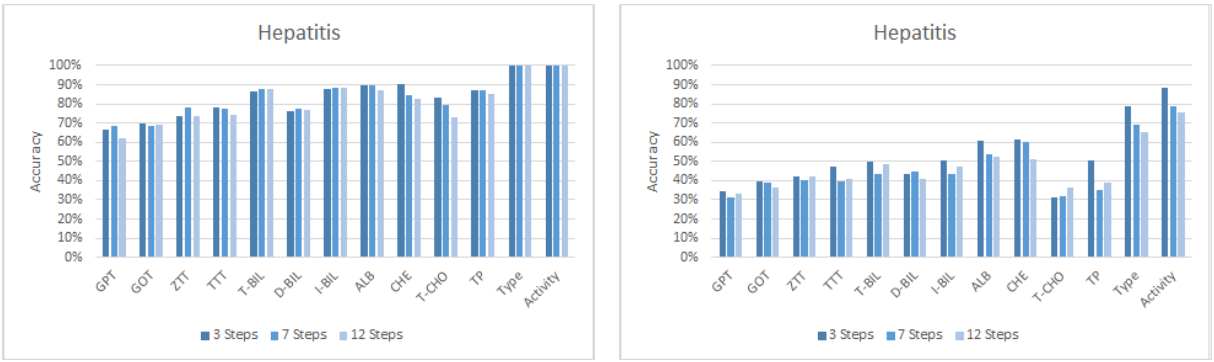
In order to mimic physicians practice, computer-aided prognosis should take into attention patients'



(a) Univariate Estimation

(b) Multivariate Estimation

Figure 5.19: Impact of the number of observations on the accuracy of the HMM estimation models for each variable using both univariate and multivariate models in the ALS dataset.



(a) Univariate Estimation

(b) Multivariate Estimation

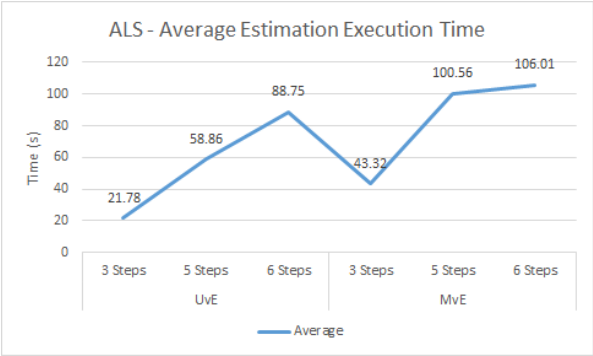
Figure 5.20: Impact of the number of observations on the accuracy of the HMM estimation models for each variable using both univariate and multivariate models in the Hepatitis dataset.

evolution, considering the different observations made along time. In this dissertation, we formalize both diagnosis and prognosis problems, making clear the differences between them, and propose a method to transform the prognosis into a diagnosis task, based on the composition of classification over the estimation of observation values. As described above, what distinguishes this approach, from what is found in the literature, is the use of temporal dependencies of the data in order to estimate the future values of every feature and with those values perform a diagnostic in the future.

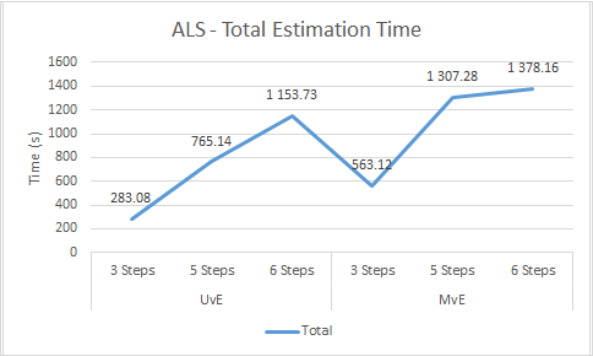
Taking into account the presented results of the techniques used, we can say that HMM were clearly the used method that performed the worse. Not only they took a lot longer to perform the estimations but, their results still palled when compared to the use of regressions or decision trees. The other two methods achieved really similar results, we can only say that when dealing with nominal datasets, using decision tree is a better approach when it comes to execution time, while the overall prognosis results are fairly close.

If dealing with numerical dataset, linear regression was the only tested technique in this work, it managed to achieve an improvement over the baseline.

Even though this work focus mostly on the estimation step of the proposed approach, the diagnosis phase still has room for improvement which if done can help improve the final results. One example is the use of more complex techniques and ensembles in the classification, that are known to have better

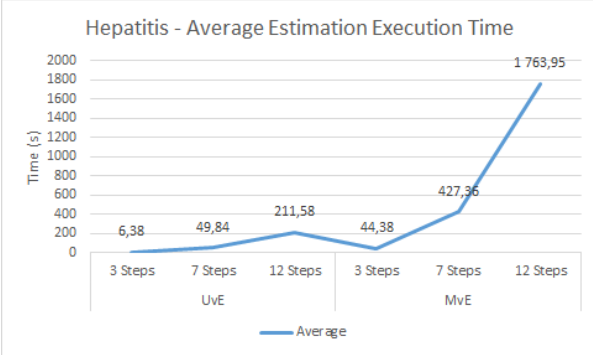


(a) Average Estimation Time

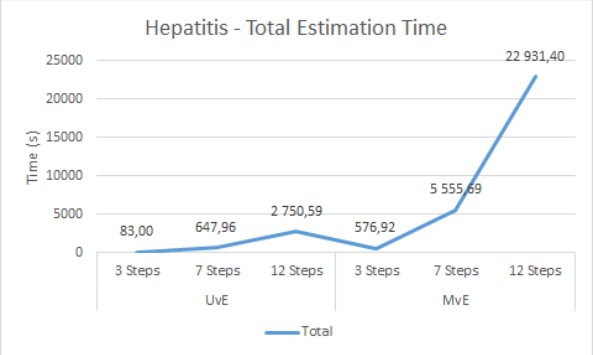


(b) Total Estimation Time

Figure 5.21: Execution time of feature estimation in the ALS dataset using HMMs.



(a) Average Estimation Time

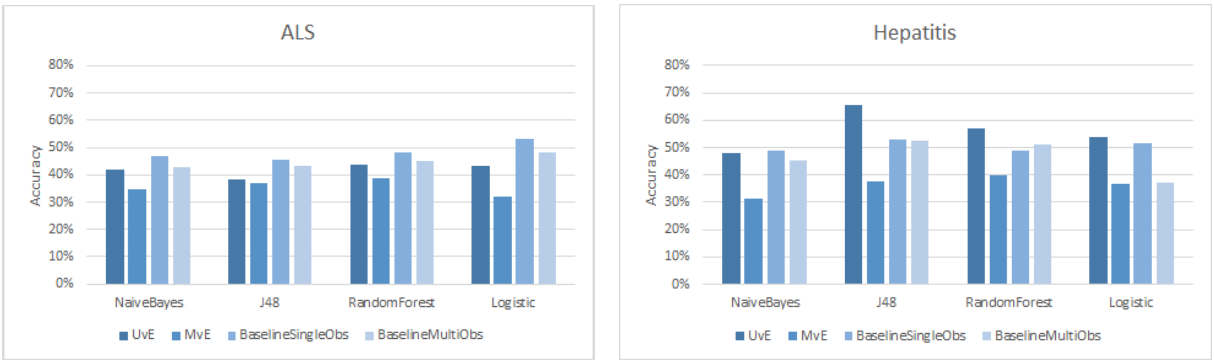


(b) Total Estimation Time

Figure 5.22: Execution time of feature estimation in the hepatitis dataset using HMMs.

performance than simpler decision trees or regressions.

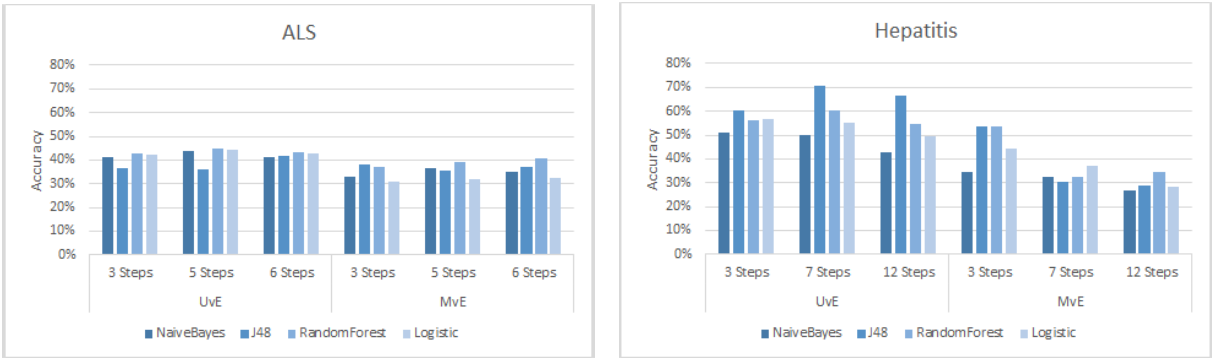
A curious result that counters what initially was thought when this approaches were planned is the performance of the multivariate approach. This approach was initially thought to be better than the univariate, because of its ability to find and use dependency relations between variables. That was in fact not the case in the datasets used. This might be because the data has too much noise or simply the relations between the used variables does not exist. Either way the univariate approach was consistently better.



(a) ALS - HMM estimations

(b) Hepatitis - HMM estimations

Figure 5.23: Accuracy of different models using the HMM estimations.



(a) ALS - HMM estimations

(b) Hepatitis - HMM estimations

Figure 5.24: Impact of the number of observation on prognosis models using the HMM estimations.

Chapter 6

Conclusions

There is a mismatch in the amount of data available in the field of healthcare and the data that is being used in order to gain knowledge. As it was shown in this dissertation, diagnosis and prognosis is a very relevant subject in the area of healthcare and that it has been subject to some work in the past years. This work shows that no novel techniques are being introduced, being the same techniques used consistently throughout the years, and a visible lack of work improving on previous research with predicting models being developed independently.

We also showed that the problem of prognosis is being tackled in the same way of diagnosis, not using the patients' evolution over time in order to improve the results.

In order to address this issue, we describe a novel approach that transforms prognosis into a diagnosis problem. This solution has two possible variants for the use of time on improving the results of a prognostic model. An univariate and a multivariate one where dependency relationships can be used to improve the final result. The method was then evaluated and discussed using two different datasets in order to show its' generalizability.

In this conclusion, we first highlight the main contributions of this work to the temporal pattern mining field and then discuss some directions for future work.

6.1 Achievements

From the survey presented the lack of use of temporality in the prognosis problem can clearly be identified. Our contribution was the proposal and definition of an extensible method that uses the temporality of the data to improve the prognosis result. It is extensible because a lot different techniques or methods can be used in each step of the approach accordingly to the characteristics of the problem and the data.

Generalidade e independencia do dominio

6.2 Future Work

From the experimental comparison of the different approaches, over two distinct datasets (with different data characteristics, either from the medical and the data points of view), it is clear an improvement trend when using the temporal informed methods proposed. The shallow differences between the results of the estimation models, need to be deeply studied and other techniques (like Dynamic Bayesian networks) should be explored to enrich the estimation process. In either cases, the temporality of this kind of data should be considered as a core aspect of the prognosis.

Another possible variation to tackle the prognosis problem presented in this thesis would be to, instead of using the values that result from the estimation phase, like in the current approach, the model that represents the evolutionary trend of that feature would be used. Then the final classification would be performed on these models.

Bibliography

[1] S. Abdul-Kareem, S. Raviraja, N. Awadh, A. Kamaruzaman, and A. Kajindran. Classification and regression tree in prediction of survival of aids patients. *Malaysian Journal of Computer Science*, 23(3):153–165, 2010.

[2] J. Ahn, J. Kwon, and Y. Lee. Prediction of 1-year graft survival rates in kidney transplantation: A bayesian network model. In *INFORMS & KORMS*, pages 505–513, 2000.

[3] I. Anagnostopoulos and I. Maglogiannis. Neural network-based diagnostic and prognostic estimations in breast cancer microscopic instances. *Medical and Biological Engineering and Computing*, 44(9):773–784, 2006.

[4] C. Antunes. Handbook for educational data mining. pages 353–363, New York, 2010. CRC Press.

[5] E. Ataide, M. Garcia, T. Mattosinho, J. Almeida, C. Escanhoela, and I. Boin. Predicting survival after liver transplantation using up-to-seven criteria in patients with hepatocellular carcinoma. *Transplantation Proceedings*, 44(8):2438–2440, 2012.

[6] D. Aujesky, D. Obrosky, R. Stone, T. Auble, A. Perrier, J. Cornuz, P.-M. Roy, and M. Fine. Derivation and validation of a prognostic model for pulmonary embolism. *American Journal of Respiratory and Critical Care Medicine*, 172(8):1041–1046, 2005.

[7] B. Balkau, C. Lange, L. Fezeu, J. Tichet, B. Lauzon-Guillain, S. Czernichow, F. Fumeron, P. Froguel, M. Vaxillaire, S. Cauchi, P. Ducimetière, and E. Eschwège. Predicting diabetes: clinical, biological, and genetic approaches: data from the epidemiological study on the insulin resistance syndrome (desir). *Diabetes Care*, 31(10):2056–2061, 2008.

[8] A. Bellaachia and E. Guven. Predicting breast cancer survivability using data mining techniques, 2006.

[9] D. Breems, W. V. Putten, P. Huijgens, G. Ossenkoppele, G. Verhoef, L. Verdonck, E. Vellenga, G. D. Greef, E. Jacky, J. der Lelie, M. Boogaerts, and B. Löwenberg. Prognostic index for adult patients with acute myeloid leukemia in first relapse. *Journal of Clinical Oncology*, 23(9):1969–1978., 2005.

[10] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software., Monterey, CA, 1984.

[11] L. Chen, D. Magliano, B. Balkau, S. Colagiuri, P. Zimmet, A. Tonkin, P. Mitchell, P. Phillips, and J. Shaw. Ausdrisk: an australian type 2 diabetes risk assessment tool based on demographic, lifestyle and simple anthropometric measures. *The Medical journal of Australia.*, 192(4):197–202, 2010.

[12] C.-L. Chi, W. Street, and W. Wolbergc. Application of artificial neural network-based survival analysis on two breast cancer datasets. In *American Medical Informatics Association Annual Symposium*, pages 130–134, 2007.

[13] J. Choi, T. Han, and R. Park. A hybrid bayesian network model for predicting breast cancer prognosis. *Journal of Korean Society of Medical*, 15(1):49–57, 2009.

[14] D. Delen. Analysis of cancer data: a data mining approach. *Expert Systems*, 26(1):100–112, 2009.

[15] D. Delen, G. Walker, and A. Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*, 34(2):113–127, 2005.

[16] R. Dom, S. Kareem, B. Abidin, A. Kamaruzaman, and A. Kajindran. The prediction of aids survival: A data mining approach. In *WSEAS Int'l Conf Multivariate Analysis and Its Application in Science and Engineering*, pages 48–53, 2009.

[17] M. Egger, M. May, G. Chêne, A. Phillips, B. Ledergerber, F. Dabis, D. Costagliola, A. Monforte, F. Wolf, P. Reiss, J. Lundgren, A. Justice, S. Staszewski, and et al. Prognosis of hiv-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies. *The Lancet*, 360(9327):119–129, 2002.

[18] S. Eichinger, G. Heinze, L. Jandeck, and P. Kyrle. Risk assessment of recurrence in patients with unprovoked deep vein thrombosis or pulmonary embolism. *Circulation*, 121(14):1630–1636, 2010.

[19] A. Endo, T. Shibata, and H. Tanaka. Comparison of seven algorithms to predict breast cancer survival. *Biomedical Soft Computing and Human Sciences*, 13(2):11–16, 2008.

[20] O. Gevaert, F. Smet, D. Timmerman, Y. Moreau, and B. Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):184–190, 2006.

[21] D. Hanson, C. Horsburgh, S. Fann, J. Havlik, and S. Thompson. Survival prognosis of hiv-infected patients. *Journal of acquired immune deficiency syndromes*, 6(6):624–629, 1993.

[22] J. T. Hendriksen, G. Geersing, K. M. Moons, and J. H. de Groot. Diagnostic and prognostic prediction models. *Journal of Thrombosis and Haemostasis*, 11(Supplement s1):129–141, 2013.

[23] Z. Hong, J. Wu, G. Smart, K. Kaita, S. Wen, S. Paton, and M. Dawood. Survival analysis of liver transplant patients in canada 1997–2002. *Transplantation Proceedings*, 38(9):2951–2956, 2006.

[24] M. Khan, J. Choi, H. Shin, and M. Kim. Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. In *Engineering in Medicine and Biology Society, 30th Annual International Conference of the IEEE*, pages 5148–5151, Vancouver, BC, 2008.

[25] S. Kharya. Using data mining techniques for diagnosis and prognosis of cancer disease. *Int'l Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, 2(2):55–66, 2012.

[26] A. Kusiak, B. Dixon, and S. Shaha. Predicting survival time for kidney dialysis patients: a data mining approach. *Computers in Biology and Medicine*, 35(4):311–327, 2005.

[27] K. Lakshmi, M. Krishna, and S. Kumar. Performance comparison of data mining techniques for prediction and diagnosis of breast cancer disease survivability. *Asian Journal of Computer Science And Information Technology*, 3(5):81–87, 2013.

[28] J. Li, G. Serpen, S. Selman, M. Franchetti, M. Riesen, and C. Schneider. Bayes net classifiers for prediction of renal graft status and survival period. *Int'l Journal of Medicine and Medical Sciences*, 1(4):215–221, 2010.

[29] M. Lundin, J. Lundin, H. Burke, S. Toikkanen, L. Pylkkänen, and H. Joensuu. Artificial neural networks applied to survival prediction in breast cancer. *Oncology*, 57:281–286, 1999.

[30] O. Mangasarian, W. Street, and W. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, 1995.

[31] C. Nash, S. Jones, T. Moon, S. Davis, and S. Salmon. Prediction of outcome in metastatic breast cancer treated with adriamycin combination chemotherapy. *Cancer*, 46(11):2380–2388, 1980.

[32] A. Osofisan, O. Adeyemo, B. Sawyerr, and O. Eweje. Prediction of kidney failure using artificial neural networks. *European Journal of Scientific Research*, 61(4):487, 2011.

[33] A. Oztekin, D. Delen, and Z. Kong. Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology. *Int'l Journal of Medical Informatics*, 78(12):84–96, 2009.

[34] M. Paradise, Z. Walker, C. Cooper, R. Blizard, and C. Regan. Prediction of survival in alzheimer's disease – the laser-ad longitudinal study. *Int'l Journal of Geriatic Psychiatry*, 24(7):739–747, 2009.

[35] N. Petrovsky, S. Tam, V. Brusic, G. Russ, L. Socha, and V. Bajic. Use of artificial neural networks in improving renal transplantation outcomes. *Graft*, 5(1):6–13, 2002.

[36] J. Quinlan. Induction of decision trees. *Machine Learning*, pages 81–106, 1986.

[37] J. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann, San Mateo, 1993.

[38] S. Rheingold, A. Neugut, and A. Meadows. Holland-frei cancer medicine. 5th edition. page Chapter 156, Hamilton (ON), 2000. BC Decker.

[39] M. Rodger, S. Kahn, P. Wells, D. Anderson, I. Chagnon, G. Gal, S. Solymoss, M. Crowther, A. Perrier, R. White, L. Vickars, T. Ramsay, M. Betancourt, and M. Kovacs. Identifying unprovoked thromboembolism patients at low risk for recurrence who can discontinue anticoagulant therapy. *Canadian Medical Association Journal*, 179(5):417–426, 2008.

[40] D. Sackett, W. Rosenberg, J. M. Gray, R. Haynes, and W. Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–2, 1996.

[41] S. Saxena, V. S. Kirar, and K. Burse. A polynomial neural network model for prognostic breast cancer prediction. *International Journal of Advanced Trends in Computer Science and Engineering*, 2(1):103–106, 2013.

[42] F. Shadabi, R. Cox, D. Sharma, and N. Petrovsky. Use of artificial neural networks in the prediction of kidney transplant outcomes. In *Knowledge-Based Intelligent Information and Engineering Systems*, volume 3215, pages 566–572, Welligton, 2004.

[43] E. Steyerberg, M. Homs, A. Stokvis, M. Essink-Bot, P. Siersema, and G. the SIREC Study. Stent placement or brachytherapy for palliation of dysphagia from esophageal cancer: a prognostic model to guide treatment selection. *Gastrointestinal Endoscopy*, 62(3):333–340, 2005.

[44] B.-Y. Sun, Z.-H. Zhu, J. Li, and B. Linghu. Combined feature selection and cancer prognosis using support vector machine. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 8(6):1671–1677, 2011.

[45] K.-M. Wang, B. Makond, W.-L. Wu, K.-J. Wang, and Y. Lin. Optimal data mining method for predicting breast cancer survivability. *Int'l Journal of Innovative Management, Information & Production*, 3(2):28–33, 2012.

[46] T. Watanabe, E. Suzuki, H. Yokoi, and K. Takabayashi. Application of prototypelines to chronic hepatitis data. In *ECML/PKDD Discovery Challenge*, Cavtat, Croatia, 2003.

[47] R. Wolfe, K. McCullough, D. Schaubel, J. Kalbfleisch, S. Murray, M. Stegall, and A. Leichtman. Calculating life years from transplant (lyft): Methods for kidney and kidney-pancreas candidates. *American Journal of Transplantation*, 8(4p2):997–1011, 2008.

[48] J. Zhou, L. Yuan, J. Liu, and J. Ye. A multi-task learning formulation for predicting disease progression. In *ACM SIGKDD Int'l Conf Knowledge Discovery and Data Mining*, pages 814–822. ACM, 2011.