TÉCNICO LISBOA

Exploration of Temporal Patterns in Classification Problems

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:	Professor Full Name
Supervisor:	Professor Full Name 1 (or 2)
Member of the Committee:	Professor Full Name 3

July 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: palavra-chave1, palavra-chave2,...

Abstract

The use of data mining techniques in the field of healthcare is still very much be-hind of where it needs to be, when a full advantage is taken of the available data with help from data mining and analytical tools.

An area where this can be used is in the process of diagnosis and prognosis, where it can help the physicians decide the correct path to take with a certain patient, based on his probable prognosis.

This is a process where a doctor takes into account an enormous amount of in-formation about the patients' state as well as his evolution, fact that is not being used in the current approaches. This is what we propose to develop, a technique that can, generally, help in the process of prognosis by using the evolution of the patient over time when making the predictions.

We will start this document by mentioning the state of affairs in the field of prognosis, by quickly explaining the most commonly used techniques followed by use cases in various diseases. We end by describing our approach to the solution and describe how we will validate our work.

Keywords: keyword1, keyword2,...

Contents

List of Tables

List of Figures

Chapter 1

Introduction

An enormous amount of data exist in the field of healthcare while the amount of knowledge that is being gathered from that data is still very limited. The data can be used to teach us new causal relations between symptoms and diseases as well a variety of other relevant information.

Another way where the use of this data can be very helpful is in the difficult task of prognosis, which helps patients and their carers decide and plan the best course of action to take, taking into account the most probable outcome in the patients dis-ease. Nowadays in data mining, prognosis is being done the exactly same way as diagnoses using the current state of the patient, but not taking into account his evolution.

To fix this problem of not using the evolution of the state of the patient we pro-pose an approach that includes this sequence of states by incorporating time in the construction of the prediction model.

We will start by describing what is diagnosis and prognosis and how data mining techniques have been used to help in these areas of healthcare. We show the state of affairs in terms of prognosis, how it is performed in a variety of diseases showing that the techniques all circle around the same. We finish by proposing a different approach for the prognosis problem and discuss how these approaches will be validated.

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 analyse 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 paper, 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 6). The paper concludes with a discussion of the improvements achieved, the issues constraining those improvements and proposing some guidelines for the next steps (chapter 7).

falta falar dos outros chapters

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 2.1.

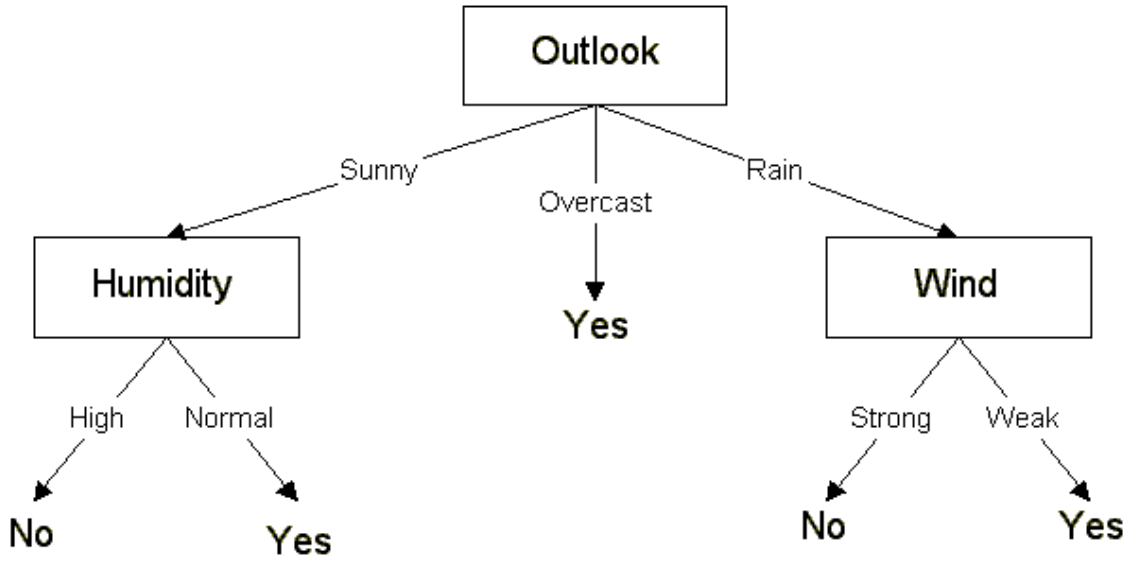


Figure 2.1: Example Decision Tree to predict if some students will play football.

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].

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 2.2.

Neural networks do not present an easily-understandable model. When looking at a decision tree, 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 analysed. 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.

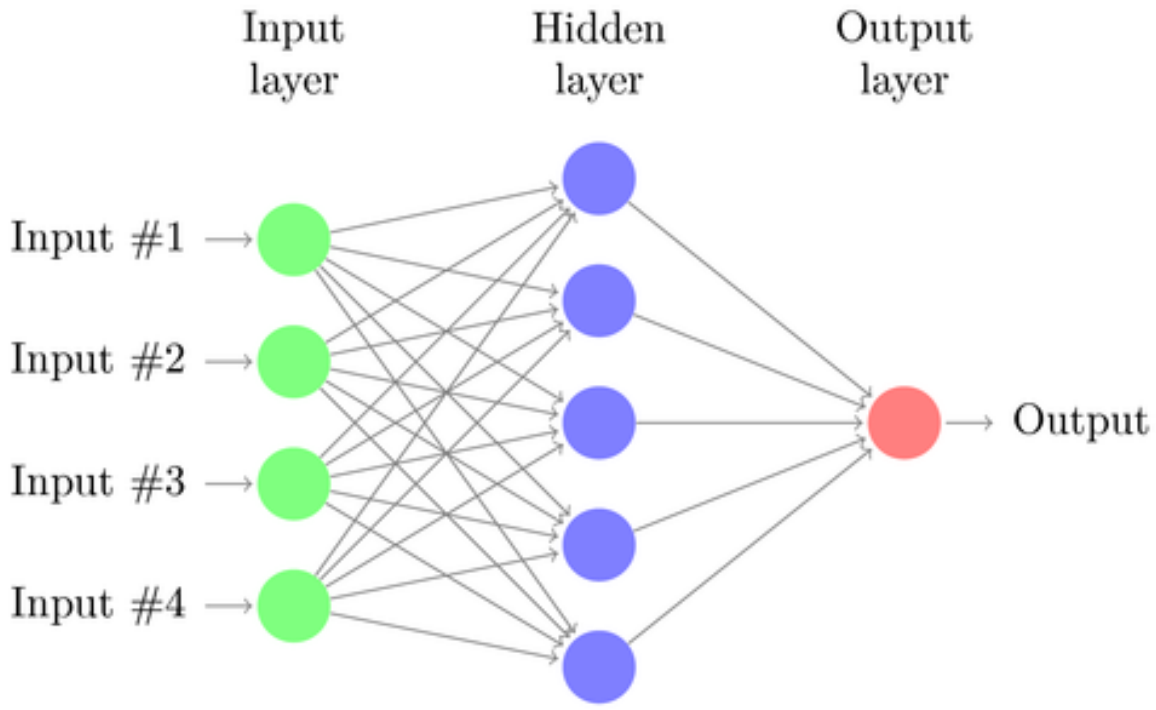


Figure 2.2: Artificial Neural Network structure.

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 2.3.

High accuracy, nice theoretical guarantees regarding overfitting, and with an appropriate kernel they can work well even if you’re data isn’t linearly separable in the base feature space. Especially popular in text classification problems where very high-dimensional spaces are the norm. Memory-intensive and kind of annoying to run and tune, though, so I think random forests are starting to steal the crown.

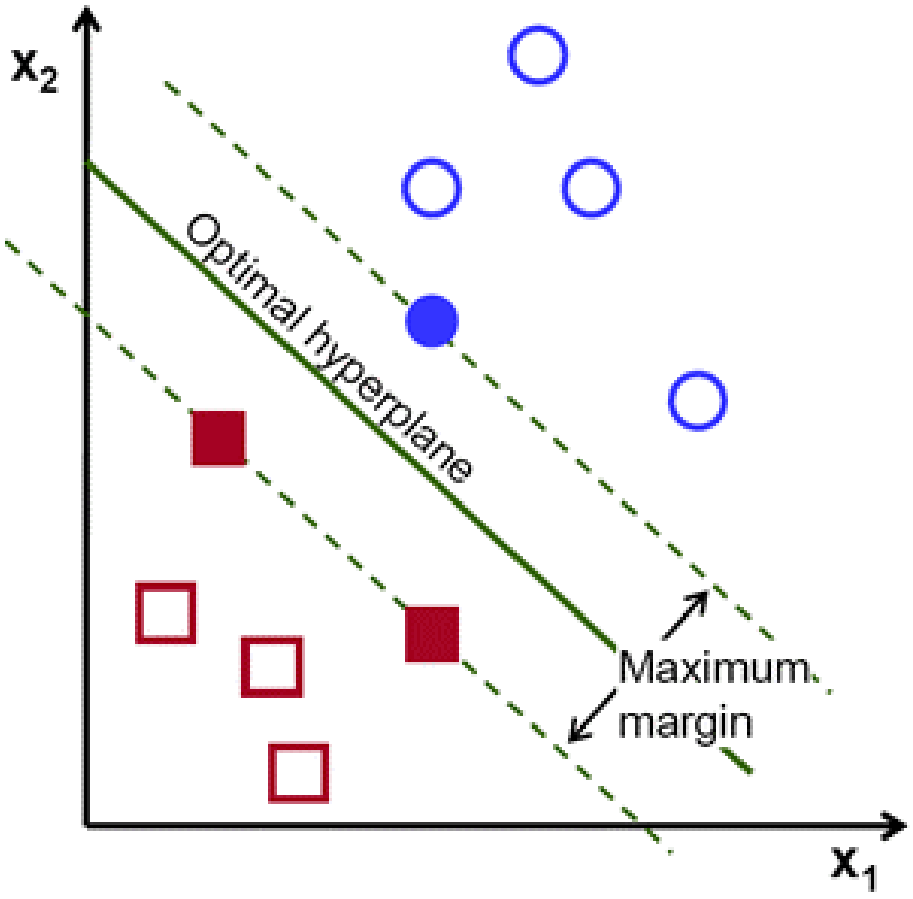


Figure 2.3: Example of 2D SVM optimal hyperplane

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 you need less training data

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 statistical methods.

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 can't

be applied. In those cases it is used logistic regression.

However, linear regression is appropriate only if the data can be modelled by a straight line function, which is often not the case. Also, linear regression cannot easily handle categorical variables nor is it easy to look for interactions between variables.

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.

As with general nonlinear regression, logistic regression cannot easily handle categorical variables nor is it good for detecting interactions between variables.

You also have a nice probabilistic interpretation, unlike decision trees or SVMs, and you can easily update your model to take in new data (using an online gradient descent method), again unlike decision trees or SVMs. Use it if you want a probabilistic framework (e.g., to easily adjust classification thresholds, to say when you're un-sure, or to get confidence intervals) or 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.

2.2.5 Dynamic Bayesian Networks/HMM

HMMs 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.

Chapter 3

Status of Affairs

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, it is said in [22] 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 behaviour. 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 modelling 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.

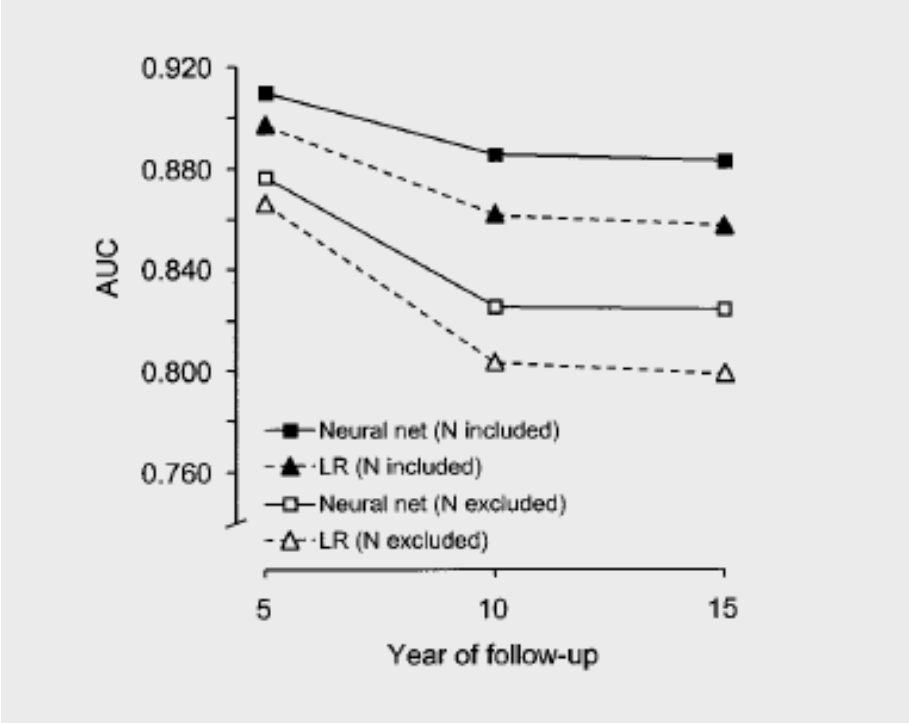


Figure 3.1: AUCs for the neural network and logistic regression (LR)

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 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.

[19] uses Logistic Regression model, Artificial Neural Network (ANN), Naive Bayes, Bayes Net, Decision trees with naïve Bayes, Decision Trees (ID3) and Decision Trees (J48) 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 cancer 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 Discriminant Analysis (PLS-LDA), k-means and Apriori Algorithm. In the end this study shows that PLS-DA

performs the best with lowest computation time and highest accuracy. [27]

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 trans-

plantation. 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 in the various diseases presented above, is that nowadays it is 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 very data dependent. [19] shows 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 that is notable in this review of prognosis work is that there is no evolution or search for improvement with only 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 Difficulties 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 of 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 can't 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 the data mining technique to be used.

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 in hand, difference between patients, using one predictor (or feature) is rarely descriptive enough to help. Doctor's use a various amount of features about patients to be able to give a prognosis and so also needs to happen when performing the prognosis computationally. A multivariable approach should be used in order to take into account the relations between features when developing the best prognostic model.

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

Classification is the problem of determining to which of a group of categories an observation belongs to. Formally you can represent the classification problem as finding the correct class y using the observation features $\bar{x}$ in the pair $(\bar{x}, y)$, where $\bar{x}$ is the feature vector and y the class.

$$\bar{x} =< x_1, x_2, ..., x_n > \tag{4.1}$$

Applied to the diagnosis/prognosis problem this is what is currently being done, where a model is created to discover y_n from $\bar{x}_n$. As already mentioned the difference between a prognostic model and a diagnostic model is just the time between the present and the supposed time of y_n.

That is why we are introducing time, the addition of time to the equation turns the pair $(\bar{x}, y)$ into the triplet $(\bar{x}_i, y_i, t_i)$ where t_i is a timestamp, $\bar{x}_i$ the feature vector in at time t_i and y_i the class also at time t_i.

$$\bar{x}_i =< x_{i1}, x_{i2}, ..., x_{im} > \tag{4.2}$$

In this case we have a sequence S of triplets, that is ordered and that can be used to predict a whole new triplet $(\bar{x}_{n+1}, y_{n+1}, t_{n+1})$.

$$S = (\bar{x}_1, y_1, t_1) \dots (\bar{x}_n, y_n, t_n) \Rightarrow (\bar{x}_{n+1}, y_{n+1}, t_{n+1}) \tag{4.3}$$

Supposing t_{n+1} is known our objective is to find y_{n+1}, the class in a future point in time.

In order to find y_{n+1} we propose the following approaches:

1. To only find y_{n+1}

   In this approach y_{n+1} is found just based on the values for the feature vector $\bar{x}$ at each point of time, ignoring the values of x_{n+1} and do not trying to estimate them.

   $$y_{n+1} = f(x_i : 1 \leq i \leq n) \tag{4.4}$$

2. To find $(\bar{x}_{n+1}, y_{n+1})$

Here y_{n+1} is found using diagnostic model on the feature vector $\bar{x}_{n+1}$ that, itself is found using various approaches:

(a) $x_{n+1,k} = f(\bar{x}_j : 1 \leq j \leq n)$

i. We can determine $x_{n+1,k}$ using the past values from the same feature, $x_{1,k}, ..., x_{n,k}$.

ii. We can determine $x_{n+1,k}$ using the past values from all the features $\bar{x}_{1,k}, ..., \bar{x}_{n,k}$ and $y_1, ..., y_n$.

In the approach $i.$, only the values of one feature are used to predict the value of that feature, i.e. in order to predict $x_{n+1,i}$ only the values $x_{1,i}, ..., x_{n,i}$ will be used.

This approach will be performed by using and/or adapting the techniques used in time series prediction. Here we need to find the technique that best deals with data that might not be numeric. This approach does not take into account the complex nature between features.

On the approach $ii.$, all the features and eventually also the class are used when predicting each feature value. This approach will be performed using core data mining techniques, like decision trees and SVM, which capture the dependence relations between features.

These two approaches, described above, will work as baseline for comparison with the rest of the work. Where one represents the use of time but lacks in capturing the intrinsic relation between features and another the opposite, while capturing the dependence relation between features misses in the explicit use of time.

Another approach is to develop a representation for $x_{1_k}, ..., x_{n_k}$ that captures its' evolution. Using that new representation of the data we can develop a new model that uses the relation between features as well as temporal patterns to correctly classify y_{n+1}.

4.1 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 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: Table 1

As seen in 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 2 approaches to doing that, in the first we only use the past values of a feature to predict the future value of that feature, that is, for example with the high blood pressure, we can use HBP in times $0, 1, 2, 3, 4$ in order to predict 5, and the same for every other feature.

In 4.2 we can see the data what would be used in order to predict HBP at time 5.

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

Table 4.2: Table 2

The second option we would 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 4.3 to predict HBP 5, as well as every other feature.

Patient ID	Gender	Age_0	HBP_0	LBP_0	Age_1	HBP_1	LBP_1	...	Age_4	HBP_4	LBP_4	HBP
25	M	65	160	88	65	140	90		65	141	88	?

Table 4.3: Table 3

At the end of this two approached the result is the same, which is a complete data row at time step 5.

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

Chapter 5

Approach2

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 5.1.

$$\bar{x}_i = (x_{i,1}, ..., x_{i,m}) \rightarrow y_i \tag{5.1}$$

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.

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 analyses 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.

Considering all of 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 5.2. Note that the different values for y_i^t may be observable (available) or

non-observable at time instant t for instance i.

$$(\bar{x}_i^1, y_i^1), ..., (\bar{x}_i^n, y_i^n) \rightarrow y_i^{n+1} \tag{5.2}$$

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], 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 5.3.

$$(\bar{x}_i^1, y_i^1), ..., (\bar{x}_i^n, y_i^n) \rightarrow y_i^{n+1} \tag{5.3}$$

The difference to the definition 5.2 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 5.1.

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 5.4, where n corresponds to the number of available observations and m the number of variables for describing each observation.

$$M_P((\bar{x}_i^1, y_i^1)...(\bar{x}_i^n, y_i^n)) = M_D(M_{E1}(\bar{x}_i^1...\bar{x}_i^n)...M_{Em}(\bar{x}_i^1...\bar{x}_i^n)) \tag{5.4}$$

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 5.6, 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.

$$M_{UvEk} : [Dom_{Xk}]^n \rightarrow Dom_{Xk} \qquad (5.5)$$
$$M_{UvEk}(x_k^1 ... x_k^n) = x_k^{n+1}$$

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 5.7.

$$M_{MvEk} : [Dom_{X1} \times ... \times Dom_{Xm}]^n \rightarrow Dom_{Xk} \qquad (5.6)$$
$$M_{MvEk}(\bar{x}^1 ... \bar{x}^n) = x_k^{n+1}$$

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.

5.1 Algorithm

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.

The difference between the models is on the creation of the estimation models (line 9), in particular on the creation of the training dataset for each variable. While for univariate model, it consists on the projection of D in relation to each X_k, the multivariate model uses the entire set of variables. 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 22).

Algorithm 1 Pseudocode for Univariate Estimation training

procedure UNIVARIATEESTIMATION(Dataset D, Function alg_{class}, int ρ)
  $//D$ – the training dataset with
  $//\quad D = \{(\bar{x}_i^t, y_i^t) : \forall i, t : 1 \leq i \leq |D| \wedge 1 \leq t \leq n\}$
  $//alg_{class}$ – 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**
    $D_k \leftarrow \pi_{Xk}(D) = \{(x_{ik}^{n-\rho}, ..., x_{ik}^n) : \forall \bar{x}_i \in D\}$
    $M_{Ek} \leftarrow alg_{class}(D_k)$
  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

Chapter 6

Validation and Experimental Results

6.1 Dataset Description

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

6.1.1 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 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% for the 1st class, 22.3% for the 2nd, 29.1% for the 3rd and 42.7% for the 4th, and 0.88% non-classified.

6.1.2 Hepatitis Dataset

The Hepatitis dataset was made available as part of the ECML/PKDD 2005 Discovery Challenge[2], it contains information 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).

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,

[1] https://nctu.partners.org/ProACT/
[2] http://lisp.vse.cz/challenge/CURRENT/

45.9% for 1, 21.35% for 2, 15.19% for 3 and 15.40% for 4.

6.1.3 SEER Dataset

The SEER data was requested from the Surveillance, Epidemiology, and End Results (SEER) web site. The data is composed by 9 files where each one contains data related to a specific cancer (breast, colon, urinary, etc.). It is composed by variables that give socio-demographic and cancer specific information concerning an incidence of cancer. Each record represents a particular patient-tumor pair within a registry. Each record is assigned a case number for each patient, and a unique record number for each specific tumor.

6.2 Validation Techniques

Both of the datasets have already been used successfully when performing the task of prognosis, the SEER data being use to predict patient survivability while the ALS data was used to predict disease's progression.

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

We will not be preprocessing any of the data, because that is not the objective of the proposed work.

Our objective is to focus on finding a way to use the time dimension when per-forming prognosis, use a patients' evolution over time, and with that build a generalizable technique whose results are not so dependent on the data.

For these reasons, and because there are other works where these datasets have been used and preprocessed, [15] [13] [27] , we can base our work on their results and use most of our time on the actual task in hand.

Having the two datasets and the model built, the usual classification metrics will be used, like accuracy, sensitivity and specificity.

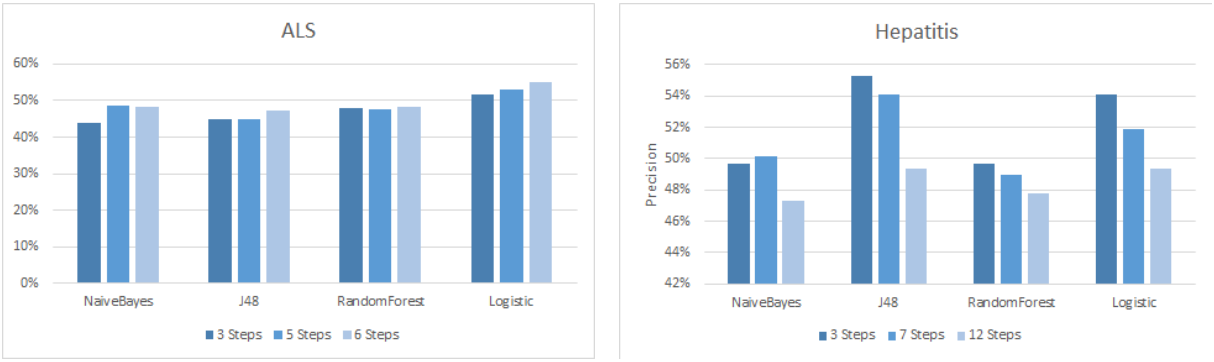
Accuracy is the ratio of correct classifications over all the cases,

$$Accuracy = \frac{TP + TN}{P + N} \tag{6.1}$$

with TP the number of true positives, TN the number of true negatives, and P and N the number of positive and negative cases, respectively. While sensitivity, also called true positive rate, is the ability of the model to identify positive cases, in other words this metric shows the overall percentage of positive classifications.

$$Sensitivity = \frac{TP}{TP + FN} \tag{6.2}$$

Because only measuring the ability to identify the positive cases is useless (a system that always



(a) ALS

(b) Hepatitis

Figure 6.1: BaselineSingleObs precision (several classifiers and number of observations).

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.

$$Specificity = \frac{TN}{FP + TN} \tag{6.3}$$

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 those showing the importance of time in this matter.

6.3 Experimental Results

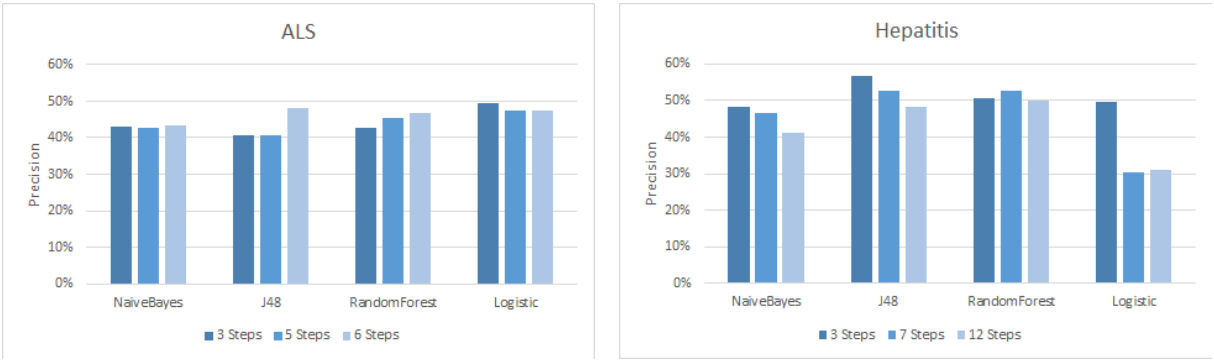
6.3.1 Diagnosis Model

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 precision ranged between 40% and 55%, depending on the dataset, technique and number of time points used, as seen in 6.1 and 6.2.

6.3.2 Regression Techniques

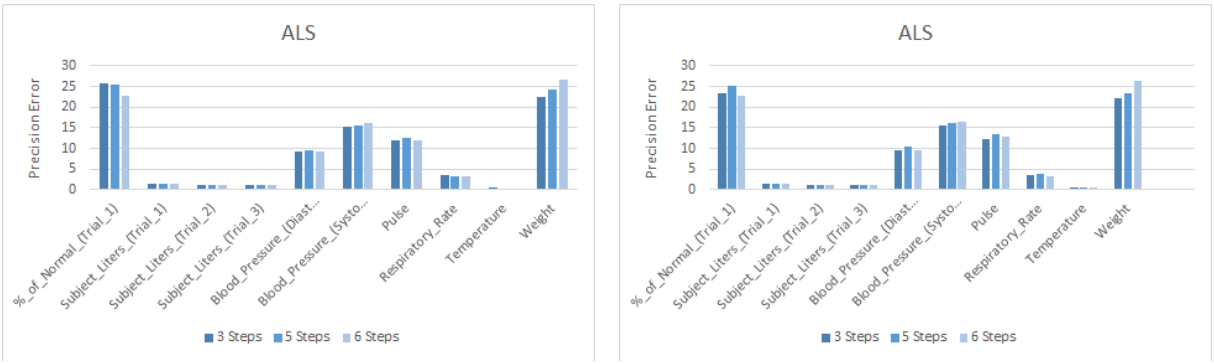
Because of the differences of the various datasets, some numeric and some nominal, different regression techniques have been applied in the estimation phase of this work, namely linear regression for the numeric datasets and logistic regression for the nominal.



(a) ALS

(b) Hepatitis

Figure 6.2: BaselineMultipleObs precision (several classifiers and number of observations).



(a) Univariate Estimation

(b) Multivariate Estimation

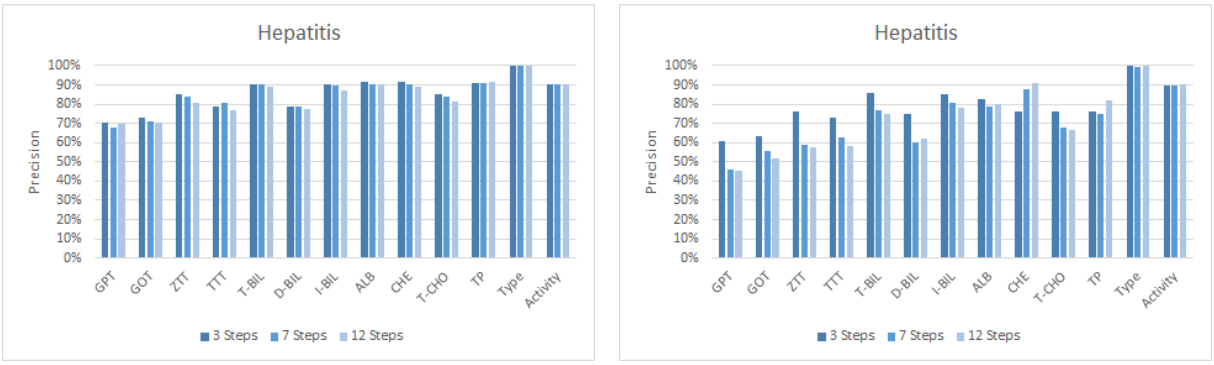
Figure 6.3: Impact of the number of observations on the precision of the linear regression estimation models for each variable, in the ALS dataset.

Estimation Models

The results with the univariate and multivariate estimation models for the ALS dataset, numeric, can be seen in 6.3. This models were built using linear regression as previously said. Both estimation models were applied using a different number of observations, and the previous. Because the dataset is numeric we evaluated our estimation by the error, 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 $9,4.$, with features having errors as high as 25 and as low as 0.35.

We can also see in 6.4 the results of using Logistic Regression on the Hepatitis dataset. In both cases the average precision of estimation rounded the 80% range, with the multivariate model being consistently a bit worse the the model that uses a single variable.

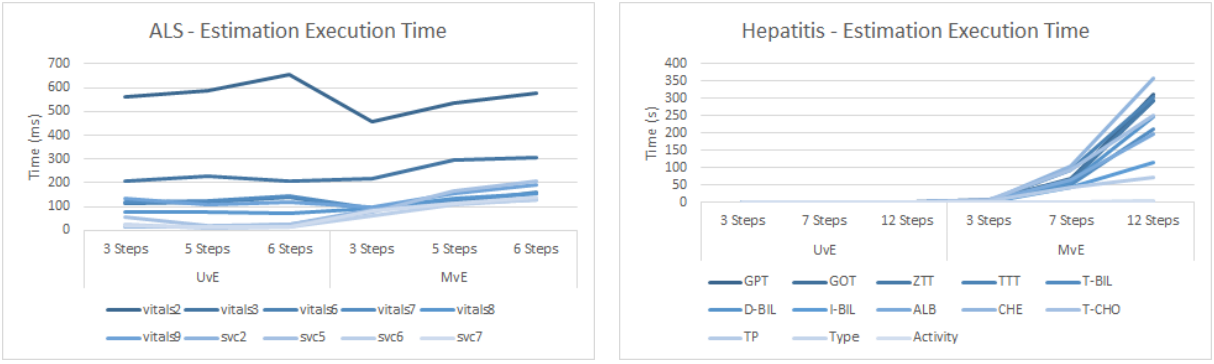
In 6.5, we can see the performance analysis, in milliseconds, of the estimation phase, divided by execution time per feature per feature. Its important to note that no significant difference was noticed between the univariate and multivariate estimations when using linear regression. The same cannot 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) Univariate Estimation

(b) Multivariate Estimation

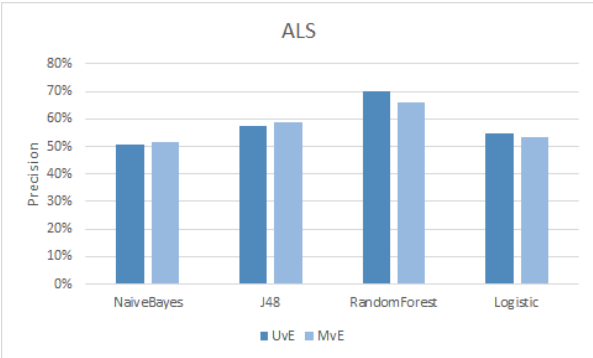
Figure 6.4: Impact of the number of observations on the precision of the logistic regression estimation models for each variable, in the hepatitis dataset.



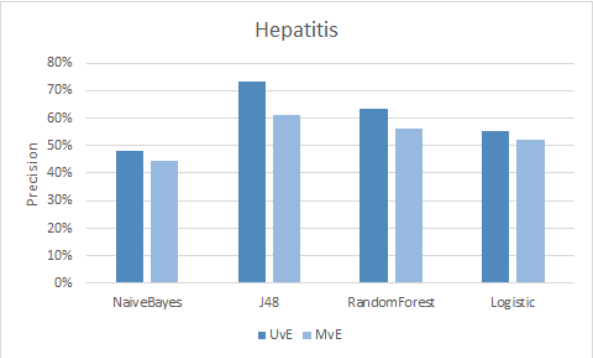
(a) ALS - Linear Regression

(b) Hepatitis - Logistic Regression

Figure 6.5: Execution time of feature estimation in both datasets using regression techniques.

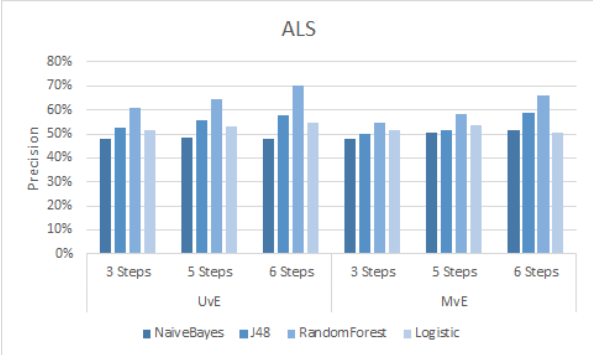


(a) ALS - linear regression estimations

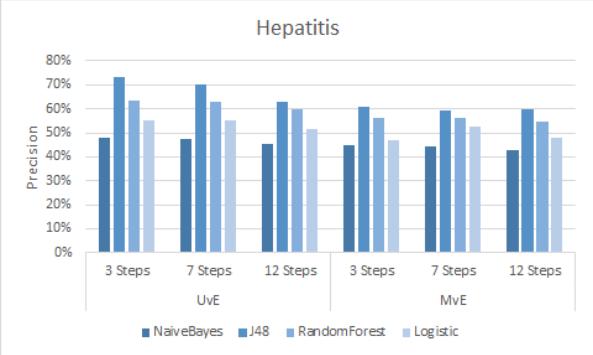


(b) Hepatitis - logistic regression estimations

Figure 6.6: Precision of different models.



(a) ALS - linear regression estimations



(b) Hepatitis - logistic regression estimations

Figure 6.7: Impact of the number of observation on prognosis models.

Prognosis Results

The overall prognosis precision achieved by using different techniques on our approaches, with the predictions achieved by using regression techniques, can be seen in 6.6.

6.7 shows the relation between the number of observations and the final precision of the prognosis, using both, UvE and MvE estimation models, and a variety of techniques.

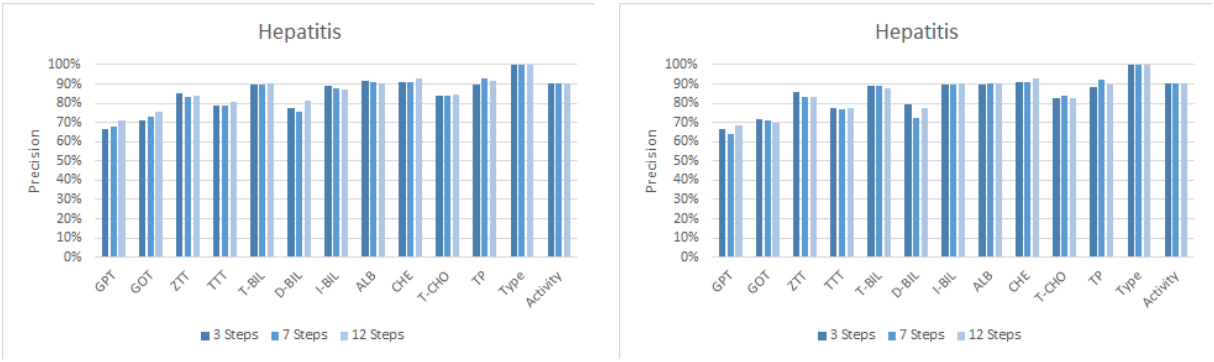
6.3.3 Decision Tree

In this section, J48 was used as the estimation technique. Because J48 cannot handle numeric classification this technique was only used on the hepatitis dataset and the results were as follows.

Estimation Models

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.

Since ALS observations are described by numerical variables and Hepatitis by nominal variables, we measured the prediction error (the distance from the predicted to the target value) and the number of correct predictions, respectively.



(a) Univariate Estimation

(b) Multivariate Estimation

Figure 6.8: Impact of the number of observations on the precision of the decision tree estimation models for each variable using both univariate and multivariate models.

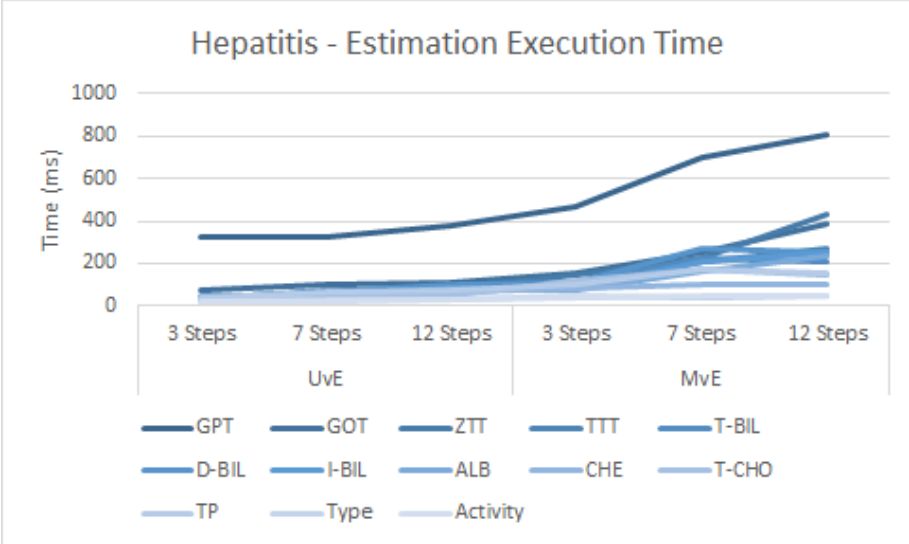


Figure 6.9: Execution time of feature estimation in the hepatitis dataset using decision trees.

6.8 shows the results with univariate and multivariate estimation models, respectively. Both estimation models were applied using a different number of observations, and the previous.

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 6.9, we can see the performance analysis, in milliseconds, of the estimation phase using decision trees, divided by execution time per feature.

It is important to note that even though the estimation using logistic regression had similar results (6.4), when looking into to the precision of the estimation, it took much much longer to estimate those results ($3\times$ more in the the fastest case and $800\times$ more in the slowest, 6.5).

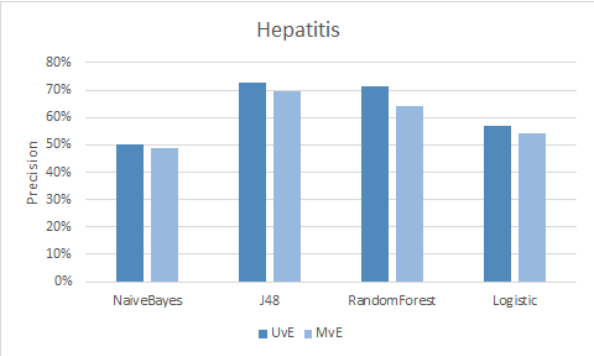


Figure 6.10: Precision of different models using the decision trees estimations.

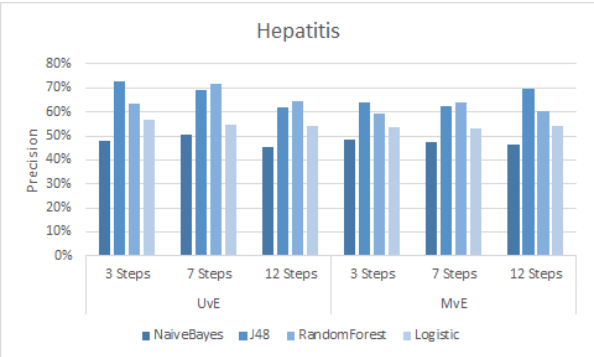


Figure 6.11: Impact of the number of observation on prognosis models using the decision trees estimations.

Prognosis Results

The overall prognosis precision achieved by using different techniques on our approaches can be seen in 6.10. The improvements on the precision of our approach are always present when compared to the ones achieved by baseline models (see 6.1 and 6.2). In Hepatitis dataset the improvements round about 20%.

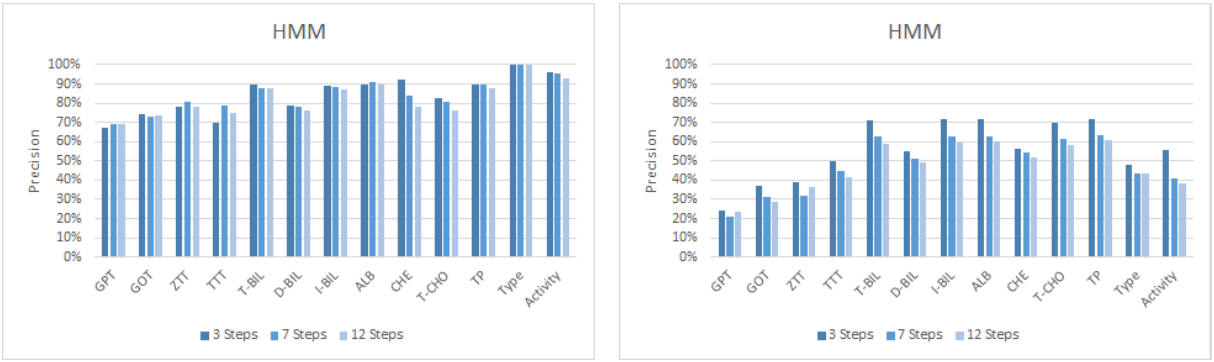
6.11 shows the relation between the number of observations and the final precision of the prognosis, using both, UvE and MvE estimation models, and a variety of techniques. It is interesting to note that the higher number of observations become prejudicial to the UvE model, which means that the values from the long past do not help to estimate future values.

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

6.3.4 HMM

In this final section, HMM were used in the estimation phase. Again like the C4.5 algorithm it doesn't handle numeric classes so only the hepatitis dataset was used.

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



(a) Univariate Estimation

(b) Multivariate Estimation

Figure 6.12: Impact of the number of observations on the precision of the HMM estimation models for each variable using both univariate and multivariate models.

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

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.

Since ALS observations are described by numerical variables and Hepatitis by nominal variables, we measured the prediction error (the distance from the predicted to the target value) and the number of correct predictions, respectively.

6.12 shows the results with univariate and multivariate estimation models, respectively. Both estimation models were applied using a different number of observations, and the previous.

In 6.13, we can see the performance analysis, in milliseconds, of the estimation phase using decision trees, divided by execution time per feature.

It is important to note that even though the estimation using logistic regression had similar results (6.4 and 6.8), when looking into to the precision of the estimation, it took much much longer to estimate those results (6.5 and 6.9).

Prognosis Results

6.3.5 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

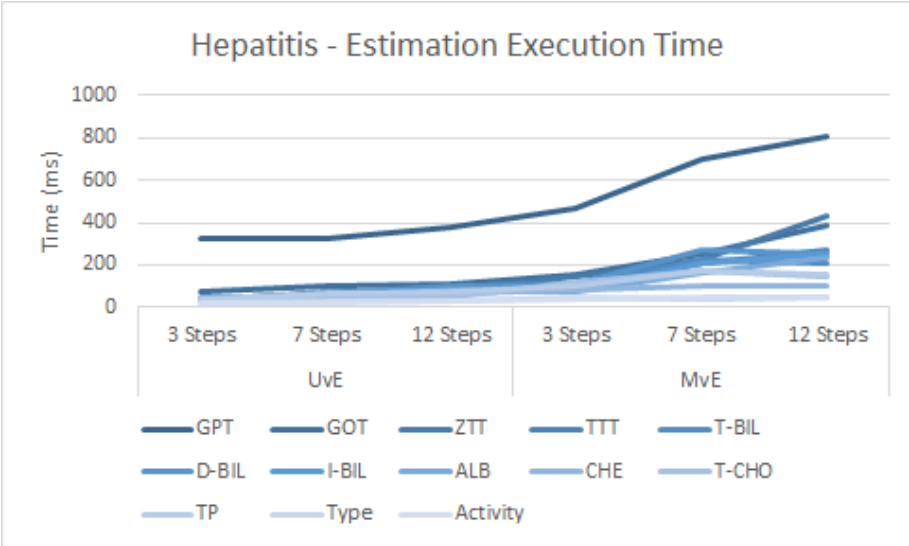


Figure 6.13: Execution time of feature estimation in the hepatitis dataset using HMMs.

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' evolution, considering the different observations made along time. In this paper, 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.

Chapter 7

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 paper 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 no evolution, being the 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 three possible solutions to the use of time in improving the results of a prognostic model.

We concluded by showing that this work will be validated in two different datasets in order to show its' generalizability, and that in each dataset the use of the usual classification metrics will be applied, along with a temporal pattern relevance analysis to show the relevance of time in the prognosis problem.

7.1 Achievements

The major achievements of the present work...

7.2 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 that, instead of using the values that result from the estimation phase, 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.