Data mining electronic health records of type 2 diabetes uncontrolled patients towards clustering LDL-cholesterol patterns

Universidade do Algarve

Mihai Daniel Petrovici

Dissertação

Mestrado Integrado em Engenharia Eletrónica e Telecomunicações

Trabalho efetuado sob a orientação de:

Prof$^{\underline{a}}$. Doutora Maria da Graça Cristo dos Santos Lopes Ruano

Prof. Doutor Rogério José Tavares Ribeiro

2018

UAlg FCT
UNIVERSIDADE DO ALGARVE
FACULDADE DE CIÊNCIAS E TECNOLOGIA

Data mining electronic health records of type 2 diabetes uncontrolled patients towards clustering LDL-cholesterol patterns

Universidade do Algarve

Mihai Daniel Petrovici

Dissertação

Mestrado Integrado em Engenharia Eletrónica e Telecomunicações

Trabalho efetuado sob a orientação de:

Profª. Doutora Maria da Graça Cristo dos Santos Lopes Ruano

Prof. Doutor Rogério José Tavares Ribeiro

2018

Declaração de autoria de trabalho

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

Mihai Daniel Petrovici

Acknowledgements

As dori sa le multumesc parintilor mei pentru tot ajutorul acordat pe durata intregului curs si pentru posibilitatea de a frecventa fara griji acest curs.

Queria agradecer a minha noiva por todo o apoio constante dedicado durante à minha vida académica e pelo amor e carinho que me foi dado.

Um obrigado a toda a familia MIEET que me ajudou e apoiou durante este percurso académico.

Um especial agradecimento também à Professora Dr.ª Maria Graça Ruano, pelo apoio recebido durante a realização da dissertação, como também pela enorme ajuda constante que me foi dada ao longo do percurso académico na Universidade do Algarve.

Um grande obrigado à Associação Protetora de Diabetes de Portugal e ao Professor Dr. Rogério Ribeiro pela oportunidade oferecida para trabalhar num tema de bastante interesse.

Não poderia faltar o meu agradecimento ao Professor Dr. António Ruano, pela sua disponibilidade em me ajudar sempre que foi necessário durante a realização deste projeto.

I would also like to acknowledge the support provided by AMGEN to the project "Prevalence of dyslipidemia in patients with diabetes mellitus at APDP - Diabetes Portugal", to which my present work is related.

Muito Obrigado a todos.

Va Multumesc.

Resumo

As doenças cardiovasculares (CVD) continuam a ser a maior causa de morte no mundo e constituem um fator de risco para diabéticos para além de os diabéticos terem maior propensão para desenvolver CVD. No entanto, apesar de as diretrizes recentes cobrirem o risco de CVD, o efetivo controlo lipídico está longe de ser conseguido. Além disso, a autogestão lipídica em conjunto com o gerenciamento de decisões terapêuticas, nem sempre assume a prioridade adequada quer pelos pacientes quer pelos profissionais de saúde.

Pretendendo compreender melhor a influência dos parâmetros clínicos no colesterol de lipoproteínas de baixa densidade (LDL) de doentes diabéticos tipo 2, doentes estes cujo gerenciamento dos valores lipídicos se suspeitam instáveis, recorreu-se a registos eletrónicos de saúde (EHR) providenciados pela APDP (Associação Protetora de Diabetes Portugal) para fazer um estudo baseado em técnicas de mineração de dados.

O banco de dados foi inicialmente analisado para compreender a integridade da base de dados, nomeadamente no que consta a variabilidade de informação associada a cada paciente e a identificação de valores corruptos ou incompreensíveis. Para cada um dos parâmetros clínicos com registo numérico foi estudada a sua distribuição estatística ao longo das consultas médicas (MA) com vista á identificação do seu comportamento individual e qual a dimensão da amostra da população que poderia ser usada para modelar o LDL. Considerou-se relevante assumir primeiramente uma abordagem linear para modelar o LDL. Utilizaram-se as abordagens mínimo quadrático ordinário e a sua variante passo a passo ('stepwise'), a qual permite ignorar os dados mais distantes da nuvem de dados. Depois, recorrendo aos mesmos conjuntos de dados usados nos modelos lineares testados foram testados modelos não-lineares, e as suas performances foram comparadas com as dos modelos lineares. A EHR disponibilizada incluía 32577 consultas médicas relativas a 1767 pacientes. Estas consultas foram registadas no período de janeiro de 2008 a fevereiro de 2018. Foram identificados como parâmetros clínicos registados na base de dados com elevado número de registos numéricos, os seguintes: hemoglobina glicada (HbA1c), colesterol LDL, colesterol de lipoproteínas de alta densidade (HDL), triglicéridos, gama-glutamil transferase sérica (GGT), plaquetas, microalbuminúria (MAU), proteinúria, creatinina, e, modificação da dieta na doença renal (MDRD).

Da análise estatística efetuada verificou-se, entre outros fatores, que nem todas as consultas de um paciente continham registos de parâmetros clínicos, e que nem todas as consultas com registo de parâmetros clínicos continha o mesmo tipo de parâmetros. Tal conduziu ao estabelecimento de uma restrição da população a utilizar nos estudos de modelação: apenas seriam utilizados os dados de pacientes que tivessem pelo menos 5 consultas médicas na década de registos nas quais houvesse registo de HbA1c. Do total de consultas médicas verificou-se que 32% continham registo de valores de LDL e que 63% descreviam os valores de

HbA1c. Contudo, nem todas as consultas com registo de HbA1c continham em simultâneo registo do valor de LDL do paciente.

Baseado nas ferramentas de análise estatística do Matlab, o colesterol LDL foi modelado por modelos lineares. Foram testados seis modelos lineares que diferiam pelo tipo de variáveis que os compunham. Consideraram-se modelos com seis variáveis, tal que, cinco delas representariam os parâmetros clínicos com maior frequência de registo na base de dados, ou seja, colesterol total, LDL, HDL, triglicéridos e HbA1c, e a sexta variável iria sendo um dos restantes parâmetros clínicos. De notar que dispúnhamos 4476 consultas médicas com registo das primeiras cinco variáveis, o que se considerava estatisticamente confiável para modelação do LDL. Analisando o desempenho dos modelos lineares verificou-se que o modelo linear mais simples, envolvendo LDL e Colesterol Total, HDL, Triglicerídeos, HbA1c e Proteinuria, apresentava um erro quadrático médio (RMSE) de 0,054. Contudo, este modelo utilizava uma quantidade de dados muito escassa (38 casos), motivo pelo qual foi desconsiderado. Sem este tipo de restrição pode apontar-se o modelo linear 3, resultante da combinação dos parâmetros LDL, Colesterol Total, HDL, Triglicerídeos, HbA1c e Plaquetas, o qual apresentou um erro quadrático de 0,07.

Foram depois testados modelos não-lineares baseados em modelação de redes neuronais. Recorreu-se ao algoritmo genético multi-objetivo (MOGA) disponível no laboratório de investigação. Procedeu-se a um pré-processamento de dados eliminando os dados que se encontravam acima de limiares estabelecidos experimentalmente. De seguida recorreu-se à normalização dos dados, conforme requerido pelos algoritmos a utilizar. Após estes pré-processamentos, o MOGA foi executado testando duas condições, uma sem limitações e a segunda impondo restrições no RMSE de treino com o intuito de obter modelos mais precisos. Considerando-se que cada execução do MOGA utilizaria estratégia semelhante à utilizada nos modelos lineares, foram testados seis modelos não-lineares, cujas variáveis correspondiam às cinco variáveis comuns entre os modelos lineares testados e as restantes variáveis iriam alternando entre os restantes parâmetros clínicos existentes na base de dados. De notar que nestes modelos foram considerados mais parâmetros que no caso dos modelos lineares. Cada uma das 6 populações consideradas recorreria às variáveis correspondentes e poderia utilizar de 2 a 25 neurónios. O algoritmo faria cinco testes de treino para encontrar o melhor compromisso que satisfizesse os objetivos. Para cada um destes modelos dividiu-se a população em três grupos: 60% da população foi usada para treino da rede neuronal, 20% da população para teste do modelo e os restantes 20% da população foi usada para a validação do modelo. As populações utilizadas em cada execução do MOGA, foram sujeitas ao algoritmo passo a passo para avaliação da relevância de cada variável no desempenho do modelo, criando dessa forma um novo modelo. O modelo MOGA com melhor desempenho na fase de

treino (RMSE=0.034) foi o modelo 4 envolvendo os parâmetros LDL, colesterol total, HDL, triglicéridos, HbA1c, GGT, plaquetas, MAU, creatinina, MDRD, sexo e idade. Contudo este modelo recorreu a apenas 830 consultas médicas o que se considerou pouco relevante estatisticamente. O modelo MOGA com menor RMSE na fase de validação foi o modelo 2, com RMSE=0.057. Este modelo, em vez dos parâmetros GGT e plaquetas introduzia o mês da consulta médica, sendo a população em estudo correspondente a 1410 consultas. Deve ressaltar-se, no entanto, que o modelo linear utilizando a população identificada pelo MOGA 5 conseguiu ainda uma performance melhor apresentando um RMSE=0.054. Neste modelo, LDL é função de colesterol total, HDL, triglicéridos, HbA1c, MAU, creatinina, MDRD, sexo e idade:

LDL = 1+1.05(Colesterol Total) -0.314(HDL)-0.124(triglicéridos)-0.005(HbA1c)-0.009(MAU) + 0.003(creatinina) + 0.017(MDRD) + 0.005(sexo) + 0.009(idade)

Como trabalho futuro recomenda-se a exploração da influência da medicação e das complicações no modelo do colesterol. Também é aconselhável que se mantenha um registo mais completo dos parâmetros clínicos para se poder ver a evolução temporal de cada parâmetro.

Palavras-Chave: Lipoproteína de Baixa Densidade, Diabetes, Data Mining, Modelo Linear, Multi-Objective Genetic Algorithm

Abstract

Cardiovascular Diseases (CVD) present the highest world health rate, constituting a risk factor to patients with diabetes and simultaneously a consequence of dyslipidemia. Effective lipid management of patients with diabetes is still largely unattained, requiring better perception of both patients and healthcare professionals. Aiming at better understanding the influence of clinical parameters on Low Density Lipoprotein (LDL)-cholesterol patterns of type 2 diabetes uncontrolled patients, the Electronic Health Records (EHR) provided by APDP (Associação Protetora de Diabetes Portugal) have been subject to data mining techniques.

The database content was primarily analyzed to understand data integrity and to avoid usage of EHR's corrupted values or misleading information. The statistical distribution of each clinical parameter reported in the data base took place to identify their individual behavior and to enable statistically coherent identification of the cohort to be used when modeling LDL.

As a first approach, LDL linear modeling was considered, using both ordinary least-squares and stepwise approaches. Then, LDL non-linear modeling was tested, using the same populations employed on linear modeling to assess the most accurate and practical LDL model. The provided EHR included 32577 medical appointments held by 1767 patients between January 2008 and February 2018. More than 10 clinical features were studied, leading to the decision of limiting the case-study population to those patients who had at least 5 Medical Appointments (MA) during the decade. From all MA's, 32% and 63% reported LDL and Glycated Hemoglobin (HbA1c) measurements, respectively, but some MA's did not report both simultaneously.

Six linear models, relating different sets of 6 clinical parameters were tested. The linear model 3, involving LDL, Total Cholesterol, HDL, Triglyceride, HbA1c and Platelet is the elected linear model with a Root Mean Square Error (RMSE) of 0.07. The model where Platelets are substituted by Proteinuria presents a RMSE of just 0.054 but employed solely 38 case-studies.

Neural network-based modeling strategies were tested as an alternative to linear models. In this sense, the Multi-Objective Genetic Algorithm (MOGA) was used. After data pre-processing, MOGA was performed twice using different threshold values. Six models were developed considering different combinations of clinical parameters. For each model, the population was divided into 3 groups: 60% of the population was used to train the network, 20% to test the model and the remaining 20% to validate the model.

Using the populations employed by each MOGA run, the stepwise algorithm was used to identify the relevance of each clinical parameter in the model and create another linear model using this parameter set. The MOGA model with the best training performance was Model 4, while model 2 was the one performing best in validation with RMSE of 0.057. However, linear model 5 created using the parameter selection identified by the MOGA presented a RMSE of 0.054 during validation when total cholesterol, HDL, triglyceride, HbA1c, microalbuminuria, creatinine, MDRD, sex and age are used in the composition of the LDL linear model.

Therefore, we can conclude that LDL can be modeled by a linear model using 6 or 10 clinical variables with very low mean square error.

Keywords: Low Density Lipoprotein, Diabetes Mellitus, Data Mining, Linear Model, Multi-Objective Genetic Algorithm

Contents

List of Figures

List of Tables

1 Introduction

According to the World Health Organization (WHO) cardiovascular diseases (CVD) are the major cause of human death. WHO also refers diabetes as one of the CVD risk factors. On the other hand, people with diabetes have a well-established relation to increased CVD risk, both as an independent risk factor and as a co-factor together with dyslipidemia. However, despite that recent guidelines cover the therapeutic goals, effective lipid management of patients with diabetes to reduce CVD risk is still largely unattainable. Furthermore, under the burden of self-management and shared decision making, lipid management seldom assumes the adequate priority in the perception of both patients and healthcare professionals.

The main motivation of this thesis is to understand the reasoning for the very low percentage of type 2 diabetes patients who, besides following medical monitoring, present low density lipoprotein (LDL) above acceptable thresholds. A preliminary analysis of the electronic health records (EHR) available at the Portuguese Diabetes Association (APDP – Diabetes Portugal) found that, within a sample of >5000 diabetes patients, only 12.3% had a last known LDL-cholesterol measurement below 100 mg/dl.

Regarding this truly impressive evidence, the major goal of this thesis is to identify the prevalence of uncontrolled LDL-cholesterol on type 2 diabetes patients within APDP's EHR aiming at posterior identification of lipid management clusters.

The patient's sample is provided by APDP. It consist in a group of almost 500 people diagnosed with diabetes mellitus type 2. Those individuals with an age greater than 18 have at least 5 appointments for 10 followed years. During those years the patients had several medical appointments and laboratory analysis. Some of the latter did not occur at the same date as registered medical appointments.

So the first step on the development of this thesis is to identify the pertinent information and values contained in the database, and to discern which medical records contain statistically usable data to perform the proposed study.

In this sense, after describing the fundamental concepts behind this work in Chapter 2, Chapter 3 describes the statistical study developed to understand the database content and the statistical behavior of the clinical parameters. Additionally, Chapter 3 includes the description of different linear models tested to model LDL using different combinations of clinical parameters.

Chapter 4 focus on the application of neural network based models to model LDL. Different models were tested, results presented, and the obtained results were compared with those obtained with linear model applied to the set of parameters identified as most relevant by the neural network approach.

Chapter 5 presents concluding remarks about the work developed in this thesis and suggests research topics to be addressed in the near future.

2 Reviewed Concepts

2.1 Diabetes Mellitus

Diabetes Mellitus (DM) is largely explained as "*a metabolic disorder of multiple set of causes characterized by chronic hyperglycemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both. The effects of DM include long-term damage, dysfunction and failure of various organs. DM present characteristic symptoms such as thirst, polyuria, blurring of vision, and weight loss. In its most severe forms, ketoacidosis or a non-ketotic hyperosmolar state may develop and lead to stupor, coma and, in absence of effective treatment, death*" [3].

As described in [4] DM has severals factors which lead to its development such as: autoimmune destruction of the pancreatic β-cells and resistance to insulin action causing hyperglycemia. [4]

Subjects with diabetes mellitus can be categorized according to clinical stage as Type 1 or Type 2, although exist various different form of sub-diabetes these two are the most used. [3]

Diabetes presents several comorbidities the most frequent being cardiovascular disease (CVD). [5]

Dyslipidemia contributes to the increased risk of cardiovascular disease whose causes are attributed to elevated levels of triglycerides and to low levels of high-density lipoprotein (HDL) cholesterol.[6]

2.1.1 Type 1

Type 1 diabetes affects a small amount of people. Only 5 -10% of the population with diabetes is Type 1. For this type of DM the patients require insulin for survival. [4]

Patients that present this form of diabetes are rarely obese, but that does not imply the inexistence of this type of DM. [3]

This type of DM occurs more frequently in childhood and adolescence, but can occur at any age. [4]

2.1.2 Type 2

Type 2 is one of the most common form of diabetes and is also known as non-insulin dependent diabetes (NIDDM).

The patients with this type of DM present disorders of the insulin action and secretion. The reasoning of developing Type 2 DM is still a subject of clinical research [3], but it can be said that is largely influenced by big quantity of nutrient ingestion and the person having a sedentary life style.

Type 2 accounts for 90-95% of the population with diabetes, and most of the patients are obese. [4]

2.1.3 Most common measurements in patients with diabetes

The most typical measurements that doctors ask for in order to diagnose or control the patient's disease are presented in this section.The list below represents the dyslipidemia measurements contained in the database under study, and therefore the parameters to be considered in this study.

Glycated Hemoglobin (HbA1c)

HbA1c is one of the most relevant parameters used in the state assessment and progression of DM. It allows evaluating the concentration of glucose in the blood over the preecending 8-12 weeks before the measurement. [7] [8]

In Table 2.1 we have the most common reference values that are used either to diagnose if the patients have DM, or to control those who already have the disease.

HbA1c	Diagnostic	Control
	4.5 to 5.6% Normal	4 to 6% Controlled
	5.7 to 6.4% Prediabetes	6 to 7% Partially Controlled
	>6.5% Possibly Diabetes	> 7% Not Controlled

Table 2.1: Reference values for HbA1c in diagnosis and control of DM
[9]

LDL, HDL and Triglycerides in NIDDM patients

The two most important types of lipoproteins that carry cholesterol to and away from the cell: one is low-density-lipoprotein (LDL), the other is high-density-lipoprotein (HDL).

The measurements of these types of cholesterol can be easily made through a blood test.

As we know LDL is considered the "bad" cholesterol because it contributes to the accumulation of fat in the arteries (atherosclerosis) leading to a narrowing of the arteries and increased risk of heart attack, stroke and peripheral arterial disease.

Also HDL is considered as the "good" cholesterol as it acts as a cleanser by taking LDL (bad cholesterol) away from the arteries and back to the liver, where the LDL is eliminated. A healthy HDL cholesterol level can protect against heart attacks and strokes. In the case of HDL cholesterol, higher levels are actually better.

Triglycerides are the most common types of fat in the body as they store the excess energy from our diet.

A high level of triglycerides combined with high LDL or low HDL cholesterol is linked to the accumulation of fat inside the artery walls, which increases the risk of myocardial infarction and stroke.

Patients with Type 2 diabetes are usally obese, they also have high values of tryglicerides and low HDL. According to [10], higher values than 400mg/dL a metabolism problem may be present. [11][12]

Table 2.2 presents the diagnostic values for the LDL, HDL and Triglycerides that we will use during this thesis.

HDL	>60 mg/dL - High (Great)
	41 a 60 mg/dL Normal
	<40 mg/dL - Low (Bad)
LDL	<100 mg/dL - Optimal
	101 to 130 mg/dL - Normal
	131 to 160 mg/dL - Normal/High
	161 to 190 mg/dL - High
	>190 mg/dL - Very High
Triglyceride	<150 mg/dL - Normal
	150 to 199 mg/dL - Borderline
	200 to 500 mg/dl - High
	>500 mg/dL -Very High

Table 2.2: Reference Diagnostic Values for HDL, LDL and Triglyceride [13]

Serum gamma-glutamyl transferase (GGT)

Serum gamma-glutamyl transferase is an enzyme present in the liver [14] , but can also be found in other organs according to [15] such as kidney, lung etc.

As stated in [16] GGT is used to test the hepatic inflammation of the liver, and also used as a risk marker for CVD.

According to [16], GGT was positively correlated with triglycerides, body mass index, LDL, age, sex and blood pressure. These parameters are to be used in our work too.

Table 2.3 presents the diagnostic values for the GGT parameter that we will use as reference during this thesis.

GGT	Men <45 U/L
	Woman <35 U/L

Table 2.3: Reference normal values for GGT [2]

Platelets

Platelets are very tiny blood cells which help our body to stop bleeding.

Reference [17] states that a high platelet agregation can lead to development of CVD.

The normal number of platelets in the blood is 150,000 to 400,000 platelets per microliter (mcL) or 150 to 400 × 109/L.

Microalbuminuria (MAU) , Proteinuria and Creatinine

MAU is an urinary secretion whose values are used to detect an early kidney disease.[18]

MAU is also used to predict CVD and some metabolic problems leading to insulin resistance. [19][20]

Proteinuria is an urinary protein resultant from a collection of 24 hour urine sample. [21]

Acording to [22] Creatinine is used to assess the Glomerular Filtration Rate (GFR), constituting a rough estimate of renal function. [23]

Table 2.4 presents the diagnostic values of microalbuminuria, proteinuria and creatinine parameters that we will use during this thesis.

Microalbuminuria	< 20 mg
Proteinuria	< 150 mg/day
Creatinine	Masculine: 50 - 100 mmol/L
	Feminine: 40 - 80 mmol/L

Table 2.4: Reference normal values for Microalbuminuria, Proteinuria and Creatinine [21][24][25]

Modification of Diet in Renal Disease (MDRD)

MDRD is a study equation used for detection of chronic kidney disease. [26]

The original equation shown in [27], was based on 6 variables: age; sex; ethnicity; and serum levels of creatinine, urea, and albumin. Later, one equation using only age, sex, ethnicity, and serum creatinine levels was proposed to simplify its use.

According to [26] the 6-variable and the 4-variable MDRD equations are the most accurate, being those the equations now widely accepted, and used by the most clinical laboratories to report Glomerular Filtration Rate estimates and assess kidney function.

2.2 Data mining

As the name suggests, Data mining is used to extract knowledge from large amounts of data. It develops from the natural evolution of information technology, involving data collection, data management and data analysis [28]. The same authors suggest the interpretation of Data mining as Knowledge Discovery from Data (KDD) or just as a step in the process of knowledge discovery. The process used for knowledge discovery is described in [28] as the folowing consecutive steps: Data cleaning; Data integration ; Data selection; Data transformation; Data mining; Pattern evaluation; Knowledge presentation. The first 4 steps are used for data pre-processing.

After pre-processing the data provided for the current study, a smaller data base was generated, containing the data selected for this particular study, this is, to identify LDL patterns. Different mining techniques exist. Since, as far as known, at the clinical environment LDL cholesterol is linearily modeled, the first attempt to analyse LDL behaviour was to apply linear modeling (section 2.3). Then a non-linear approach is considered, making use of artificial neural networks (section 2.4). Further studies towards knowledge discovery from the data base under study are required, as mentioned in chapter 5.

2.3 Linear Modeling

Linear regression is understood within the statistical field as a technique for estimating the expected value of a variable y, given the values of some other variables x. Linear regression is called "linear" because the output variable is represented as a linear function of input variables, whose weight parameters determine different linear models.

In this thesis the output variable is LDL. The simpler and understandable linear model makes use of ordinary least squares method. However, ordinary least squares modeling may produce model estimates with large variance, therefore a tradeoff between higher estimate accuracy and giving up use of some variables in the model becomes clarifying in terms of the population under study [29]. In this sense, the stepwise function is also considered.

2.3.1 Ordinary least squares

Ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being predicted) in the given dataset and those predicted by the linear function. [30]

In this thesis we use the function “*fitlm*”provided by *MATLAB* to create the initial linear models.

The functions "*fitlm*" according to the documentation of *MATLAB* has more than one form of being used. The two most used forms of the function in this thesis are:

"mdl = fitlm(X,y) returns a linear regression model of the responses y, fit to the data matrix X."

"mdl = fitlm(___,modelspec) defines the model specification using any of the input argument combinations in the previous syntaxes".[31]

For the "*modelspec*" we used a robust fitting with the weight function OLS.

The output of the function is presented as a graphical window where the Model is shown, and an output table with the next information:

- Formula of the linear model with the respective weights: y=1 + X1 + X2 + X3 + .., where y represents the output value of the parameter to be estimated, and X1, X2, X3, etc. represent the variables introduced in the model.

- For each variable we obtain the next information from the table:

"Estimate — Coefficient estimates for each corresponding term in the model.

SE — Standard error of the coefficients.

tStat — t-statistic for each coefficient to test the null hypothesis that the corresponding coefficient is zero against the alternative that it is different from zero, given the other predictors in the model. Note that tStat = Estimate/SE. .

pValue — p-value for the t-statistic of the hypothesis test that the corresponding coefficient is equal to zero or not."[32]

- A small summary statistic of the model:

"Number of observations — Number of rows without any NaN values.

Error degrees of freedom — n – p, where n is the number of observations, and p is the number of coefficients in the model, including the intercept.

Root mean squared error — Square root of the mean squared error, which estimates the standard deviation of the error distribution.

R-squared and Adjusted R-squared — Coefficient of determination and adjusted coefficient of determination, respectively.

F-statistic vs. constant model — Test statistic for the F-test on the regression model, which tests whether the model fits significantly better than a degenerate model consisting of only a constant term.

p-value — p-value for the F-test on the model." [32]

And finally, in the same graphical window a Figure is shown with the fitted model, presenting the adjusted data withing the Fit: y=d*X. and a linear 95% confidence bound.

2.3.2 Stepwise Function

The stepwise approach enables considering in the LDL model only the variables that most contribute for the model [29].

The Matlab function stepwise opens a window that show us the Coefficient of each parameters, t-stat and p-value. There are also presented the values of RMSE, R-squared, Adjusted R-Squared F-statistic and p-value of every previous model. This way we can choose the best set of inputs and see how the parameters interact with each other.

2.4 Artificial Neural Networks

Artificial Neural Networks (ANN) are especially used to perform non-linear mapping between an input space and an output space in order to obtain relationships between them or to detect templates within the input data. [1]

ANN were originally developed as an attempt to mimic human brain behavior. ANN also provides the ability to design algorithms that are applicable to many different domains by just setting some parameters based on the corresponding context. These algorithms can be used for statistical analysis and data modeling in several areas, such as medical diagnosis, financial market forecasting, energy consumption.[33]

2.4.1 Radial basis functions neural networks

Radial basis functions neural network (RBFNN) is a feed forward network used for pattern finding. It has the advantages of fast learning and a very high accuracy. [33]

A RBFNN is composed by three layers, as it can be seen in Figure 2.1 .[34]

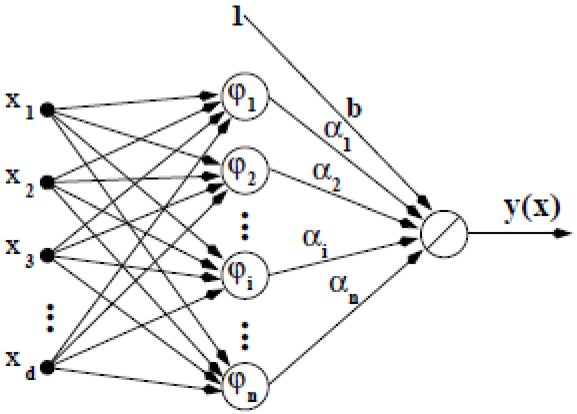


Figure 2.1: Radial basis function neural network structure.
[34]

First layer is a set of inputs connecting the network with its surroundings. The second, is a hidden layer where non-linear transformation of the input space is performed. The third layer unite the outputs of the second layer in order to obtain the overall output of the network. [34]

The hidden layer is formed by a set of neurons used for processing the information. Each neuron is expressed by a radial function defined in [34] such as:

$$\varphi_i\,(x) = \gamma(\mid c_i - x \mid) \tag{1}$$

where γ represents a transformation (usually non-linear), $c_i \in R^d$ (where d is the number of network inputs) is the function center and $x \in R^d$ is the point where the function is evaluated. To note that R^d represents the real space with dimension equal to the number of network inputs (d). " [34]:

The mostly applied radial function is the Gaussian function :

$$\varphi_i\,(x) = e^{-\frac{\|c_i - x\|^2}{2\sigma_i^2}} \tag{2}$$

where σ_i is the function spread. [34]

2.4.2 Learning Algorithms

Learning algorithms are classified by the type of learning mechanism they use or by the time interval between updating each parameter. In terms of the type of learning mechanism the algorithms can be classified as supervised, unsupervised, a combination of supervised and unsupervised, and, as reinforcement type of learning.

As a further classification we can say that learning algorithms can be categorized based on time that the parameters are updated:

- Offline: parameter update occurs after seeing all the data sample;
- Online: parameters update happens as each new data sample arrives. [33]

2.4.2.1 Supervised learning

Supervised learning can be explained as a type of learning where the network is provided with a correct answer (output) for every data sample that it is fed to the algorithm. [33][35]

2.4.2.2 Back propagation technique

Back propagation technique is a learning algorithm which looks for the minimum value of the error function, therefore improving the accuracy of predictions.

2.4.3 Genetic Algorithm

As explained in [33] the Genetic Algorithm (GA), introduced by John Holland and his students, is an algorithm created to find the best solution from a search space, through an imitation of the natural process of evolution. The two main principles on which the algorithm is based are:

"1- competition or survival of the fittest ;

2- child's inheritance of the parents genetic make-up." [33]

The GA creates an initial population acting as potential solutions. The population develops gradually over a number of generations, where each solution is evaluated and a measure of fitness is attributed. Only the best fitted ones are employed on the next generation, pursuing till the combination of input parameter settings solve the optimization problem. More details about the algorithm may be found in PhD thesis "Intelligent Support System for Cva Diagnosis By Cerebral Computerized Tomography" [33].

2.4.4 Multi-Objective Genetic Algorithm (MOGA)

MOGA [1] is an algorithm available at the laboratory where this work was developed that apply a blend between GA and state-of-the-art derivative based training algorithms to design neural network. [36]

Each solution is treated as a chromosome composed by a number of neurons and a vector for selecting features. This way the GA searches the best features. For the parameter estimation the algorithm uses a modified Levenberg-Marquardt (LM) algorithm in order to exploit the linear-nonlinear separability of the network parameters.[37].

The network training is a nonlinear problem, so for each potential solution is trained a number of times. As a form of terminating the training early-stopping is normally used.[36]

The MOGA cycle (Figure 2.2) consists of three actions: problem definition, solutions generation and analysis of results.

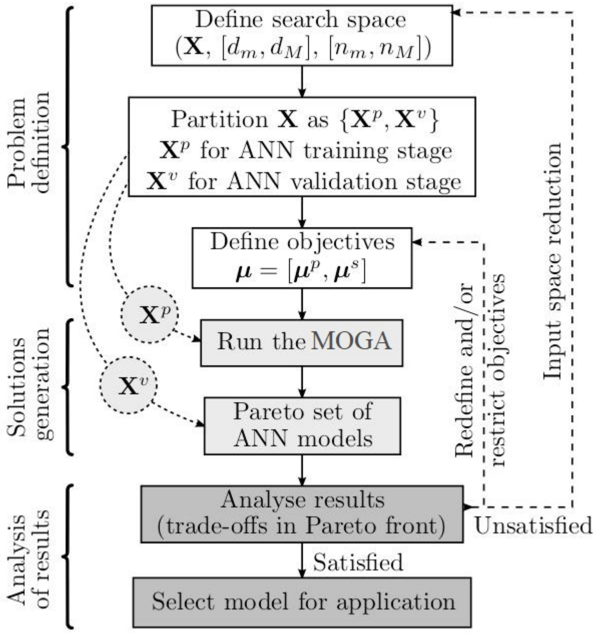

Problem definition
Solutions generation
Analysis of results
Define search space ($\mathbf{X}$, $[d_m, d_M]$, $[n_m, n_M]$)
Partition $\mathbf{X}$ as $\{\mathbf{X}^p, \mathbf{X}^v\}$
$\mathbf{X}^p$ for ANN training stage
$\mathbf{X}^v$ for ANN validation stage
Define objectives $\boldsymbol{\mu} = [\boldsymbol{\mu}^p, \boldsymbol{\mu}^s]$
$\mathbf{X}^p$
$\mathbf{X}^v$
Run the MOGA
Pareto set of ANN models
Analyse results (trade-offs in Pareto front)
Satisfied
Unsatisfied
Select model for application
Redefine and/or restrict objectives
Input space reduction

Figure 2.2: MOGA execution cycle [1]

Our data is supplied to MOGA in three different sets: for parameter estimation, early-stopping termination and for validation. We obtain them by using the ApproxHull algorithm. The validation set is used for evaluation of the performance and for selection of the final model. [36]

3 Statistical study of the database

In this section we start studying the database (DB) to identify the pertinent information and values contained in the database, and to discern which medical records contain statistically usable data to perform the proposed study. Then we test the clinical parameters to identify their correspondent statistical distribution, and how the parameters relate among them.

As previously mentioned it will be used a private clinical database provided by APDP. This database includes data from 32577 medical appointments held since January 2008 till February 2018, corresponding to 1767 patients. Although expected a monthly based update of clinical records, it is known that patients' appointments have a larger periodicity. It is also suspected that not all appointment records will provide values for all the medical parameters included in the database. A previous research on APDP database [38] reported that some fields of the data records were unfulfilled. In this sense, a primary analysis of the current database status and the integrity of the values and information is required to avoid usage of corrupted values or misleading information during present study.

3.1 Analysis of database content

The database consists of a total of 1767 patients, where 786 are female and 981 are male. Both present an average age of 60 years old on 2018. Women average height is 157cm while males present and average height of 170cm. The total number of medical appointments in the database is 32577 being held between January 2008 and February 2018.

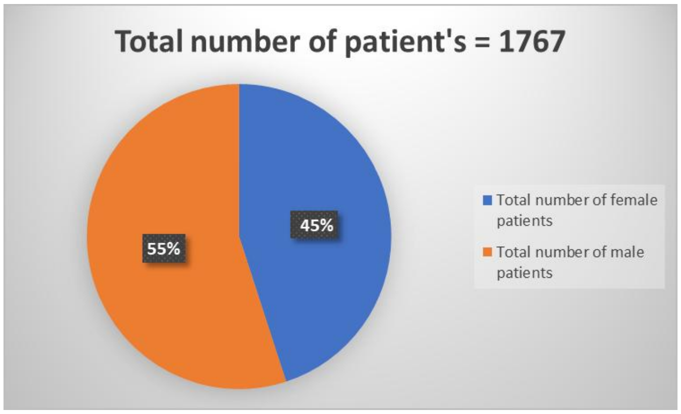
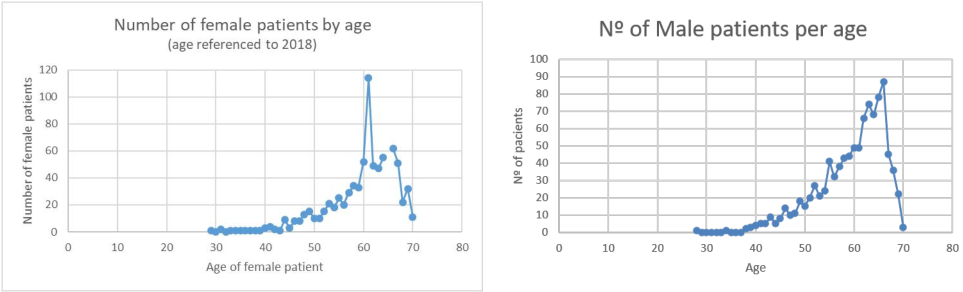


Figure 3.1: Distribution of Patients regarding sex

Among this population, in average, we can observe that both male and female present the same age of diabetes diagnose, the age of 43, they both joined the APDP at the age of 52 and their last medical appointment was performed at the age of 58. However, observing Figure 3.2, one can observe that the age of the majority of the patients in these classes is different: we observe that 114 women have 61 years old, while 87 men have 66 years old, and the distribution of number of patients per age varies as male or female is considered, as expected by the results obtained on other studies.



(a) (b)

Figure 3.2: Number of female (a) and male (b) patients per patient's age

The database includes, for each patients medical appointment several fields to be fulfilled by the clinician. Besides the patient's identification, age, dates of medical appointments, one should mention the clinical parameters of interest for this study as the columns labels provided in the database: Colesterol Total (total cholesterol), LDL, HDL, Triglicéridos (triglycerides), HbA1c, GGT, Plaquetas (platelets), Microalbuminuria, Proteinuria, Creatinina, MDRD, MDRD Estadio Peso (weight) , IMC (body mass index), TA systolic, TA diastolic, plus other fields were the proposed medication and complementary exams are fulfilled.

As a start, we considered the first 11 parameters above listed, and we analyzed the distribution of the total number of medical appointments for each patient during the 10 years period. From Figure 3.3 and Table 3.1 we observe that the most frequent occurrence is patients with 8 medical appointments (MA). Once we previously established as a patient's inclusive criteria that he/she should have at least 5 MA whose HbA1c was stored in the database, Table 3.1 results guarantee the accomplishment of the inclusive criteria established.

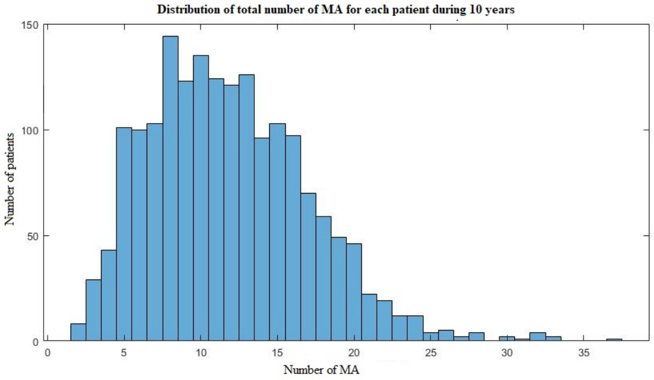


Figure 3.3: Distribution of total number of medical appointments for each patient during 10 years

MA	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Patients	0	8	29	43	101	100	103	144	123	135	124	121	126
MA	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
Patients	96	103	97	70	59	49	46	22	19	12	12	4	5
MA	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**		
Patients	2	4	0	2	1	4	2	0	0	0	1		

Table 3.1: Table values for Histogram in Figure 3.3

From these patients we need to choose the ones that have the most quantity of data available. In Figure 3.4(complemented by Table 3.2), we see that there is a big amount of MA, more precisely 6768 MA's, where only three parameters were fulfilled. Nevertheless, nearly 11% of the MA's have eleven parameters fulfilled.

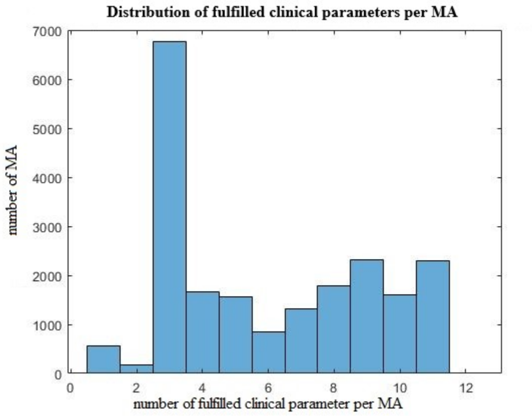


Figure 3.4: Distribution of number of fulfilled clinical parameter per MA

Number of MA	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Quantity Of Data	553	181	6768	1655	1552	852	1310	1776	2324	1605	2288	6
%	2.65	0.867	32.429	7.93	7.437	4.082	6.277	8.51	11.136	7.69	10.963	0.029

Table 3.2: Table values for Histogram in Figure 3.4

As we can see in Table 3.2, almost 43.8 ≈ 44% of the Medical Appointments have less than 5 parameters fulfilled. Only a really small population of the MA have all the clinical parameter values present, 6 (six) MA or 0.029%.

Concentrating on this subset of the database (Table 3.2), we have to know which parameters are the ones that are more frequently fulfilled. In order to do that we have to see for each of the selected parameter the respective histogram.

Referring to Figure 3.4, this histogram shows us promising values, but there is a need of analyzing the number of patients and for which of them, the database present at least 5 (five) medical appointments with the highest number of parameters.

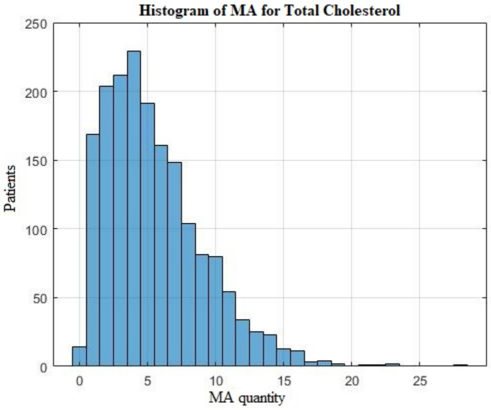


Figure 3.5: Total Cholesterol

As we can see in Figure 3.5 and from the Table 3.3 there are a lot of patients with five MA where Total Cholesterol is present, but we also have a significant amount of patients where the Total Cholesterol value is not present, or, that do not satisfy the five medical appointments limit. This subset corresponds to 828 patients, roughly 47% of the total population.

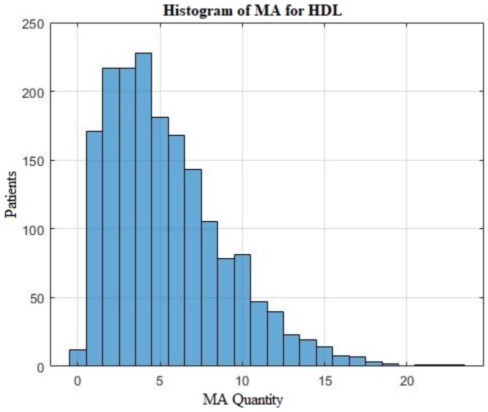


Figure 3.6: HDL

Looking at the Figure 3.6 we see the histogram of MA where HDL is present, unfortunately we observe that most of the patients have 4, 3 or 2 Medical Appointments. A total of 845 patients or 48% of the population is below the minimum proposed value of MA.

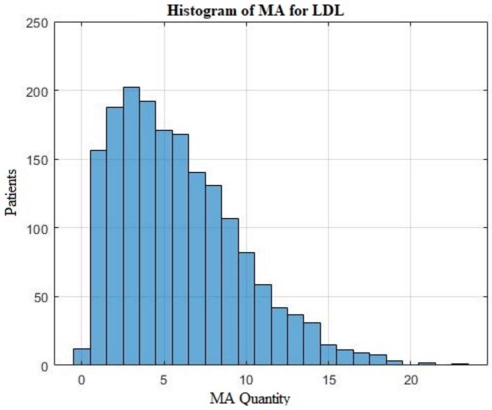


Figure 3.7: LDL

In Figure 3.7, the MA with LDL recordings, we see a more evenly distributed quantity of MA. Now we only have 750 patients or 42% below the threshold, it's a small improvement, but we also observe that a lot of patients present a number of MA between 5 and 10 which is very good for our research.

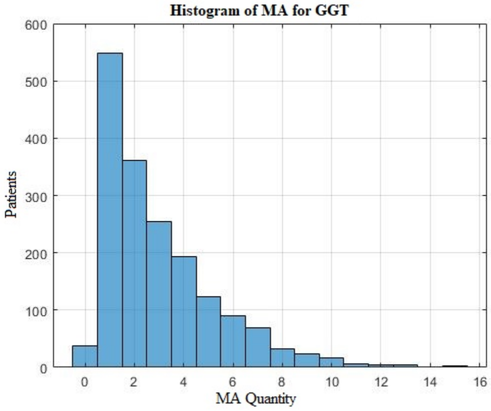


Figure 3.8: GGT

For the histogram of GGT (Figure 3.8) we observe a very high density of the population under the 5 (five) MA, more precisely 1396 patients have a number of MA under the imposed limit. Only 21% patients are above the threshold, which is a very bad result.

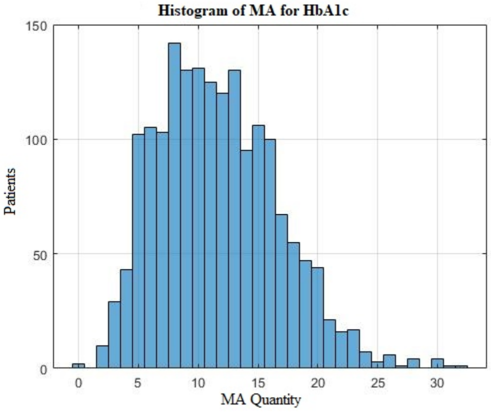


Figure 3.9: HbA1c

The HbA1c parameter is the most present parameter in our population. We have a very low number of patients under the threshold, only 84 as it can be seen in the Figure 3.9 and calculated from Table3.3. With only 4% of the population below the limit of 5 MA with HbA1c values registered we can consider that this parameter is one of the most important as it is the one that is the most measured.

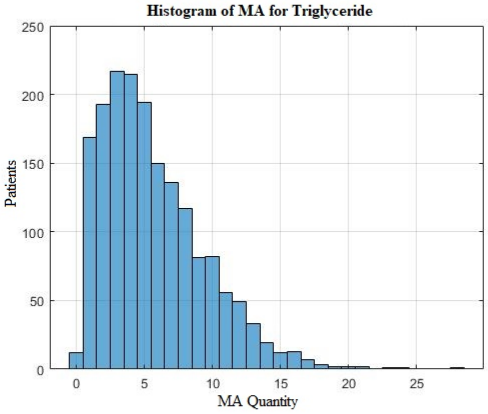


Figure 3.10: Triglyceride

For Triglyceride measurements, Figure 3.10 show a good amount of patients with less than five MA, 806 or 46%. We can see that most of them have 3 (three) MA with Triglyceride measurements. For our threshold we have a good amount of patients that present the parameter in at least 5 (five) observations.

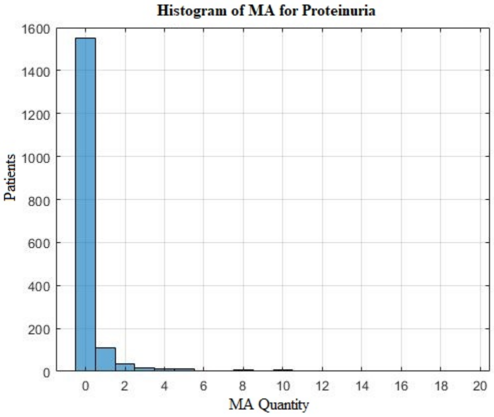


Figure 3.11: Proteinuria

Looking at Figure 3.11 we can see that Proteinuria is the parameter least registered. Very few patients have registered values and even less have over the established limit of five MA. Almost all of the population don't have this parameter present in their MA.

We think that one of the reasons behind this small amount of data is given by the fact that this clinical parameter is difficult to obtain with precision.

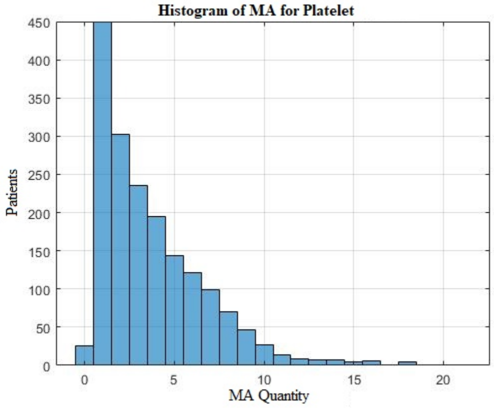


Figure 3.12: Platelets

The MA including Platelets values are also a small amount of data according to Figure 3.12, 1207 patients are below our limitation,which corresponds to almost 68% of total population.

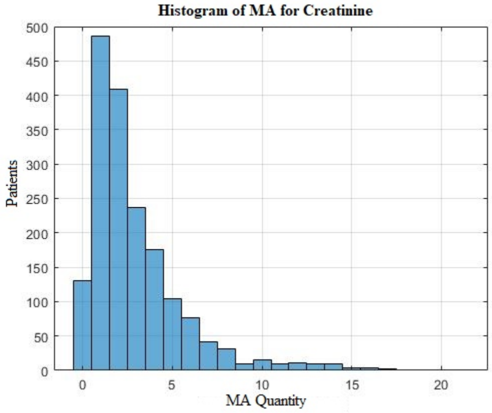


Figure 3.13: Creatinine

In Figure 3.13 we see that the majority of the patients have less than 5 (five) appointments, but the good thing is that we have a good amount of patients that have over 5 MA with registered values.

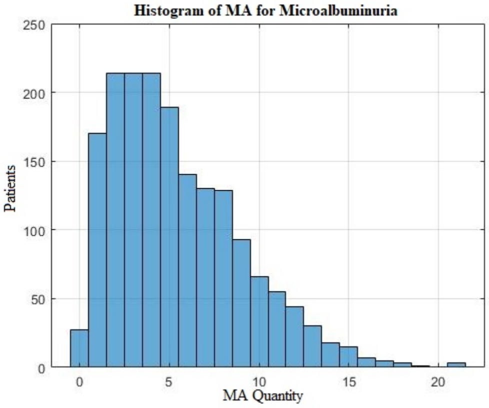


Figure 3.14: Microalbuminuria

In what concerns Microalbuminuria, Figure 3.14 shows a high amount of patients that have this clinical parameter present in their MA , and the majority of the patients also have at least 5 (five) MA. Therefore Microalbuminuria is worthwhile being included in our study.

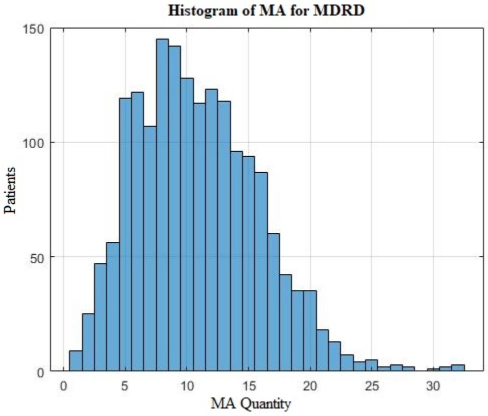


Figure 3.15: MDRD

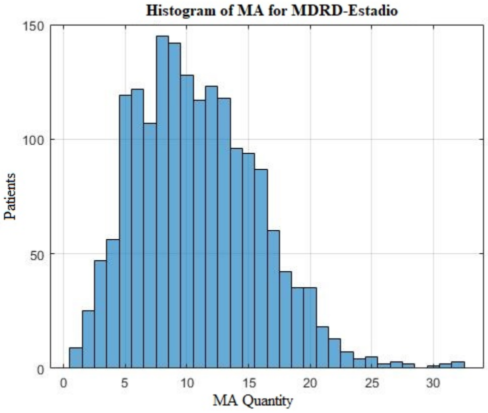


Figure 3.16: MDRD Estádio

Figure 3.15 and Figure 3.16 present the MDRD and MDRD-Estadio registration profile. Those parameters have a very good presence in our database. These two features are calculated the same way, MDRD-Estadio being just another annotation for MDRD.

MA	**Patients**											
	ColT	HDL	LDL	Trig	HbA1C	GGT	Crea	Micro	Plaq	Prot	MDRD	MDRD_E
0	14	12	12	12	2	38	130	27	25	1551	9	9
1	169	171	156	169	0	549	486	170	449	111	25	25
2	204	217	188	193	10	362	409	214	302	35	47	47
3	212	217	202	217	29	254	237	214	236	15	56	56
4	229	228	192	215	43	193	175	214	195	11	119	119
5	191	181	171	194	102	124	104	189	144	12	122	122
6	161	168	168	150	105	90	76	140	121	4	107	107
7	148	143	140	136	103	69	42	130	99	4	145	145
8	104	105	131	117	142	32	31	129	70	6	142	142
9	81	78	107	81	130	23	10	93	46	2	128	128
10	80	81	82	82	131	16	15	66	27	7	117	117
11	54	47	59	56	125	6	9	55	14	2	123	123
12	34	40	42	49	120	4	11	44	9	2	118	118
13	25	23	37	33	130	5	9	30	7	1	96	96
14	23	19	31	19	95	0	9	18	7	1	94	94
15	13	14	15	12	106	2	4	15	4	0	87	87
16	11	8	11	13	100		4	7	6	1	60	60
17	3	7	9	7	67		2	5	0	0	42	42
18	4	3	8	3	55		1	3	4	1	35	35
19	2	2	3	2	47		1	1	1	1	35	35
20	0	0	0	2	44		1	0	0		18	18
21	1	1	2	2	21		1	3	1		13	13
22	1	1	0	0	16						7	7
23	2	1	1	1	17						4	4
24	0			1	7						5	5
25	0			0	3						2	2
26	0			0	6						3	3
27	0			0	1						2	2
28	1			1	4						0	0
29					0						1	1
30					4						2	2
31					1						3	3
32					1							

Table 3.3: Histogram values of the parameters registered for 1767 patients

As we can see in Table 3.3, we have a number of patients where the MA=0, which means they do not have any registered value for that clinical parameter. We see that Protinuria (Prot) is the clinical parameter where the most patients do not have a value.

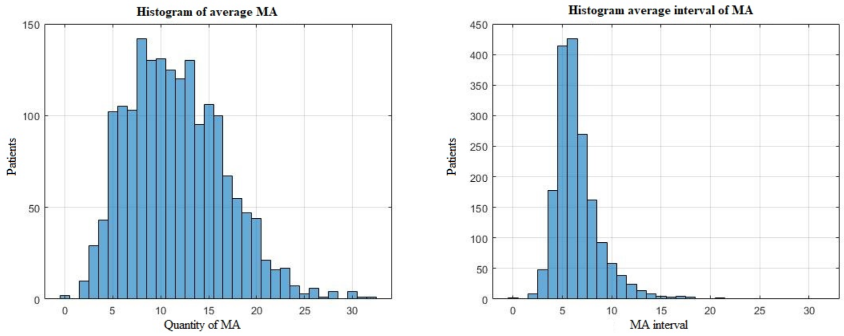
3.2 Analysis of clinical parameters statistical behavior

We are now interested in analyzing the temporal evolution of the clinical parameters. Since this would be a fastidious task due to the high number of patients, we are analyzing only some randomly chosen patients.

3.2.1 Analyzing medical appointment's distribution

After a close look at our database we extracted the MA that present obligatory the HbA1c parameter, and the results we obtained are as follows:

- Total MA under consideration: 20588
- Average interval between MA's: 6
- Mean value of MA held per patient: 12



(a)

(b)

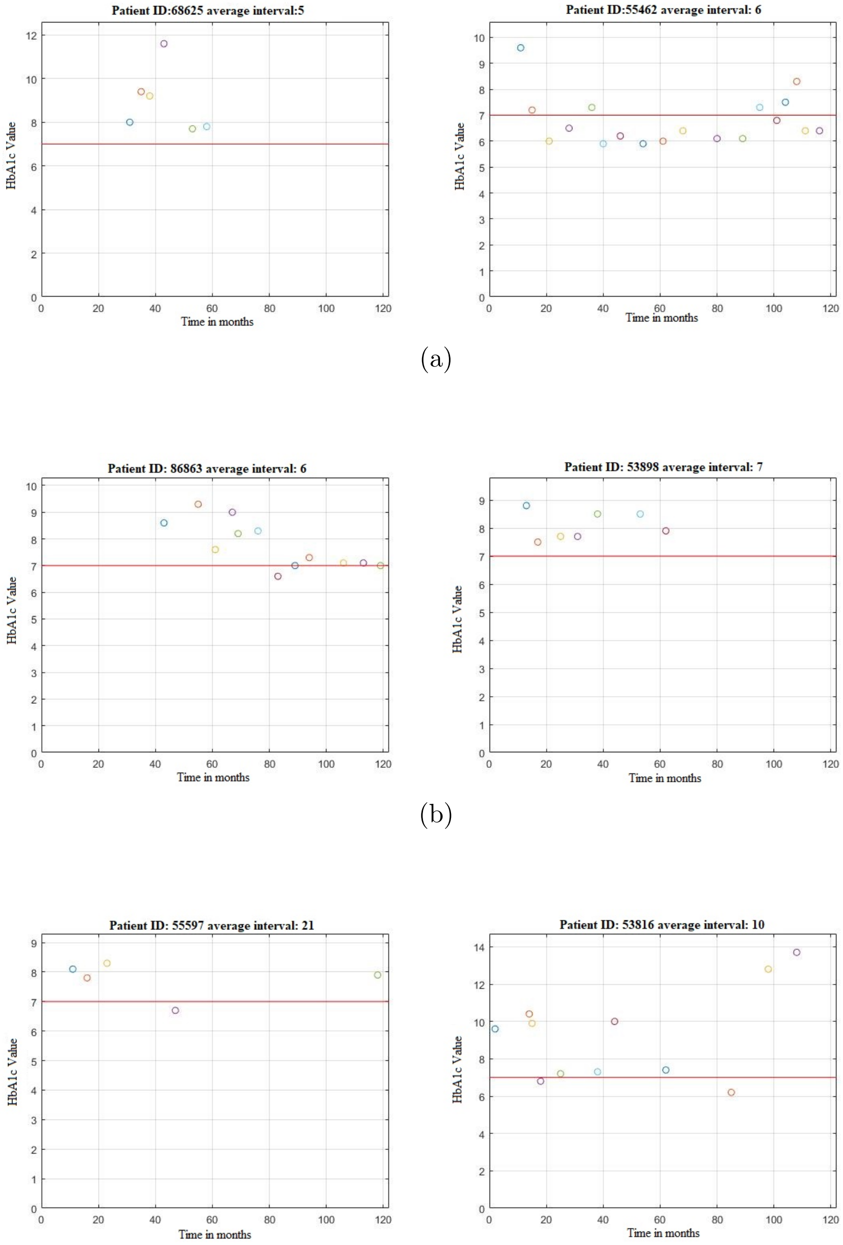
Figure 3.17: (a) number of medical appointments per patient (b) Average time elapsed between consecutive MA for each patient. All patients considered had HbA1c registrations

When plotting the histogram of the average number of medical appointments per patient (Figure 3.17a) the highest column corresponds to the number of patients which had MA with HbA1c registered. In Figure 3.17b, where the average time elapsed between consecutive MA for each patient is represented, we see that, in general, patients have 5 and 6 months between every MA where the HbA1c parameter is present.

3.2.2 HbA1c evolution according to time elapse between medical appointments

Considering only the MA's reported in the database which present values on the column of HbA1c parameter we obtain a matrix of every value of HbA1c of every patient, and we randomly choose 10 of them. Through the temporal evolution of HbA1c values we can see if the patients reacts to treatment by lowering his HbA1c value below the threshold established for diabetic patients, this is, below 7.

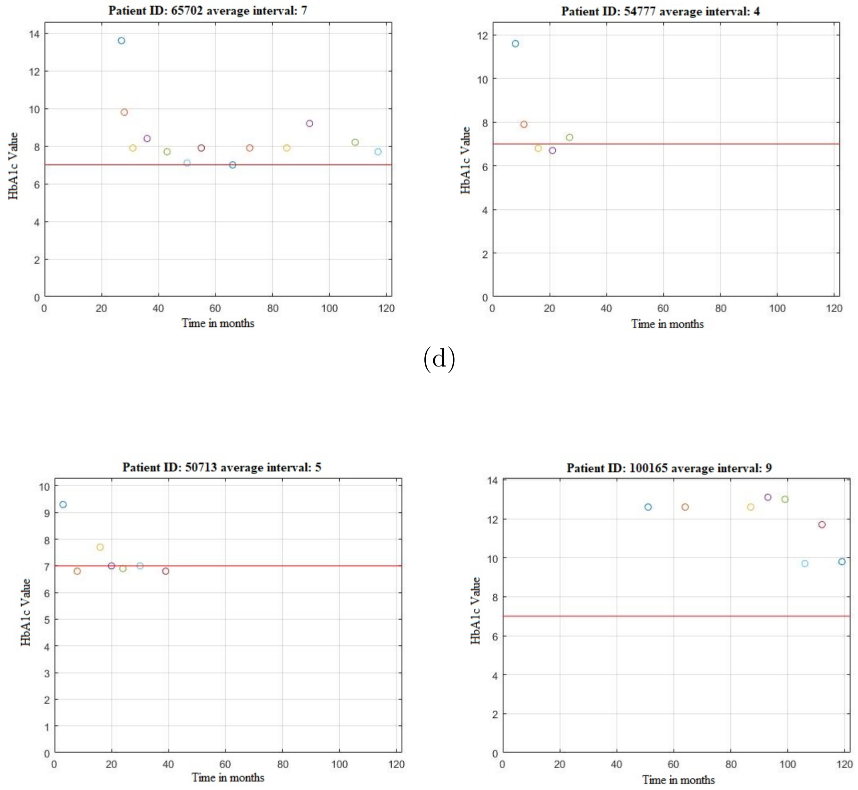
Figure 3.18 describe the temporal evolution of HbA1c.



(a)

(b)

(c)



(d)

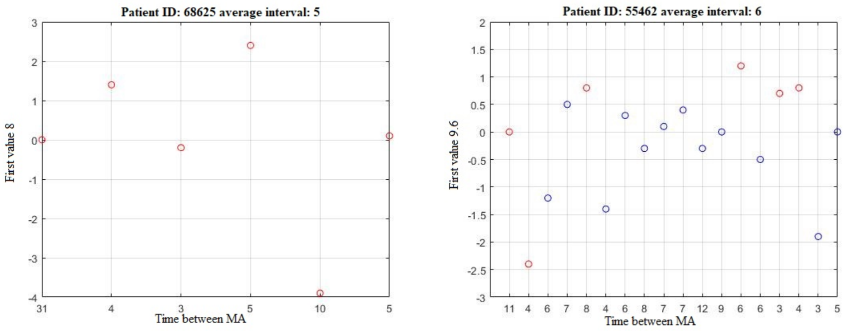
(e)

Figure 3.18: Temporal evolution of HbA1c for 10 patients (randomly selected) in months

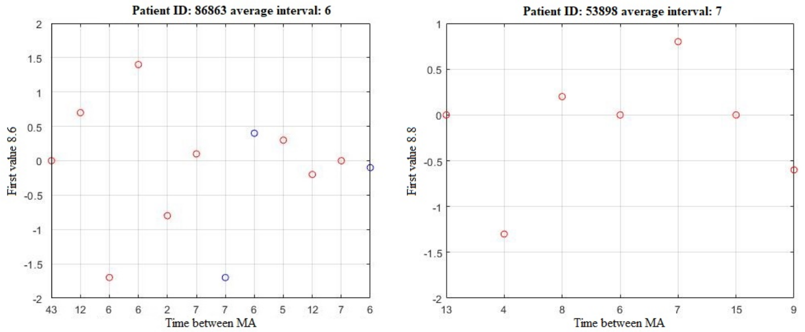
As we can see in Figure 3.18a we have two different patients, patient number 68625 has all of the values above the limit 7. The second patient, ID=55462, has a good control over the parameter, staying a good amount of time below the threshold.

In Figure 3.18b, 3.18c and 3.18d some patients reach the threshold but none of them can maintain it below the limit, which indicates that they may not follow as they should the recommended treatment. Figure 3.18e shows two individuals where one reaches the threshold and stays under the limit, and the other struggles to reach it.

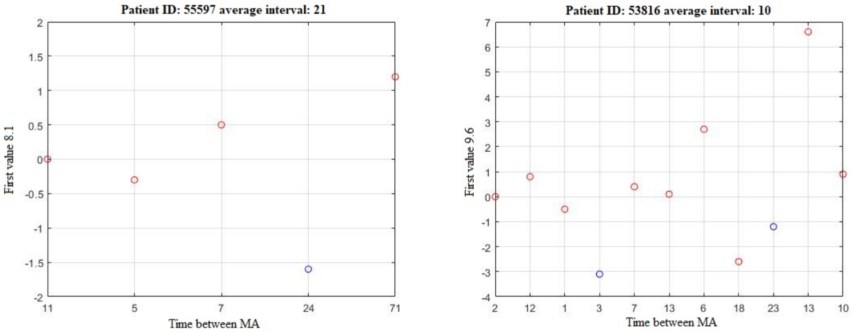
To have a better understanding as how the patients develop along time, we must also see the time interval between every MA, and observe the difference (positive or negative) of the "actual" parameter regarding the "past" one, this is the previously registered HbA1c value. Figure 3.19 the number of months between January 2008 and the first MA where HbA1c has been registered located at the y-axis label. The rest of the x-axis values represent the time elapse between consecutive MA's. The y-axis expresses the HBA1c difference from the initially obtained value (presented in the y-axis label). The blue and red circles indicate if the clinical parameter is below or above the the admissible HBA1c value, respectively.


Patient ID: 68625 average interval: 5
First value 8
3
2
1
0
-1
-2
-3
-4
31
4
3
5
10
5
Time between MA
Patient ID: 55462 average interval: 6
First value 9.6
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
11
4
6
7
8
4
6
8
7
7
12
9
6
6
3
4
3
5
Time between MA

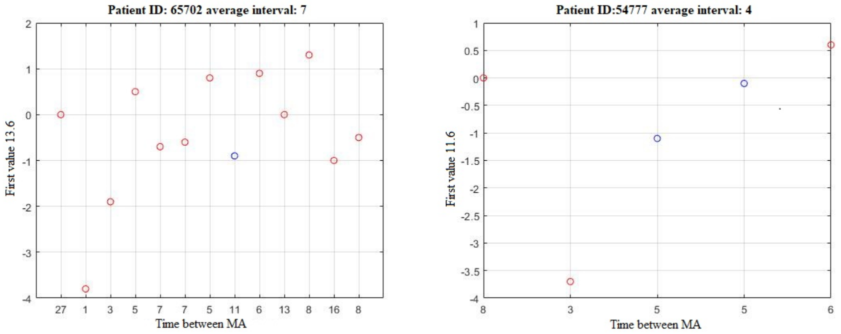
(a)


Patient ID: 86863 average interval: 6
First value 8.6
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
43
12
6
6
2
7
7
6
5
12
7
6
Time between MA
Patient ID: 53898 average interval: 7
First value 8.8
1
0.5
0
-0.5
-1
-1.5
-2
13
4
8
6
7
15
9
Time between MA

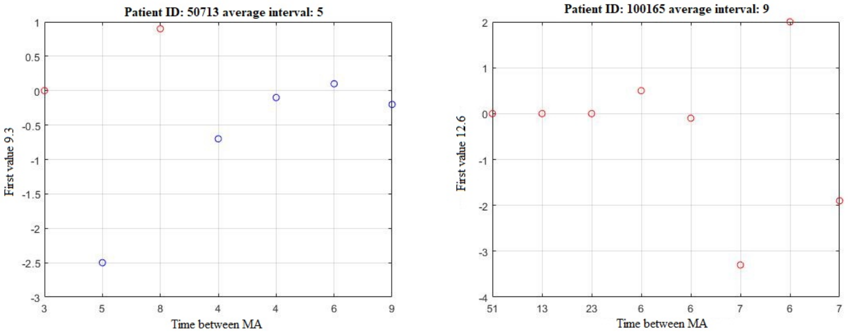
(b)


Patient ID: 55597 average interval: 21
First value 8.1
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
11
5
7
24
71
Time between MA
Patient ID: 53816 average interval: 10
First value 9.6
7
6
5
4
3
2
1
0
-1
-2
-3
-4
2
12
1
3
7
13
6
18
23
13
10
Time between MA

(c)


Patient ID: 65702 average interval: 7
First value 13.6
2
1
0
-1
-2
-3
-4
27
1
3
5
7
7
5
11
6
13
8
16
8
Time between MA
Patient ID:54777 average interval: 4
First value 11.6
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
8
3
5
5
6
Time between MA

(d)



(e)

Figure 3.19: Difference between HbA1c values of consecutive MA's for 10 patients (referred in Figure 3.18)

In Figure 3.19a we have two patients, on the left side we can see that patient 68625 has some fluctuations especially after the fourth MA where we see a big decrease of almost 4% in the value. On the right we have a patient with a lot more MA, presenting fluctuations of the HbA1c values and the majority of them are in the blue zone, indicating that therapy is improving his health status.

The left patient presented in Figure 3.19b starts with big fluctuations and presenting values over the limit, but managing to reduce the margins and finally reaching the threshold. The patient 53898 has a small try at reducing the value of HbA1c, but doesn't succeed, keeping all of his measurements above the limit.

Figure 3.19c, 3.19d and 3.19e show HbA1c values of patients with difficulties in reducing their values to the limit or below, only one of them managing to reduce it and to keep it under the limit (Figure 3.19e left).

From all these Figures we can conclude that the small sample of patients have at least five values for the parameter HbA1c. Some of them manage to reduce between every MA. Also one can see that the majority of the patients succeed to get at least once a value below the threshold (HbA1c=7).

3.2.3 Analyzing HbA1c values through MA's evolution

Our database presents a quantity of 3836 MA with HbA1c below threshold.

The next histogram, Figure 3.20, presents the number of patients whose HbA1c measurement reached the limit of value 7 or below this limit according to the months taken to achieve so, considering the beginning of treatment at APDP as the beginning. As may be seen many patients started the MA at APDP with values under the limit, explaining the histogram column height at 0 on the x-axis.

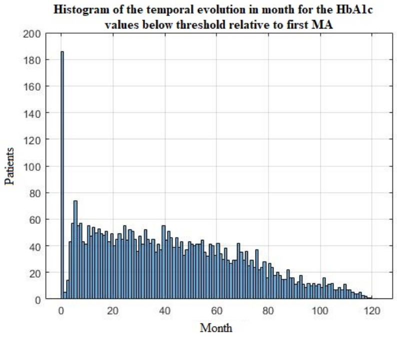


Figure 3.20: Number of patients whose HbA1c measurement reached the limit of value 7 or below according to the months taken to achieve so, considering the beginning of treatment at APDP as the temporal beginning

One can also see that the individuals that obtain a reduction of their HbA1c values do it in the first months of their treatment. There are also many patients that reduce HbA1c values just after several months of treatment , but it can be considered that some of these late MA's achievements (more to the left of the graphic) may be due to patients who had already reached the desired levels before. We saw in previous Figures that the patients presented fluctuations in HbA1c values, and if they reach the goal in one MA not necessarily maintain the value under the limits on the following MA's.

If we now analyze the same population but considering only males and only females (Figure 3.21 left and right respectively) we see that both gender patients present more MA above limit than below limit. Also we have a big population with only one value below the limit, for both sexes.

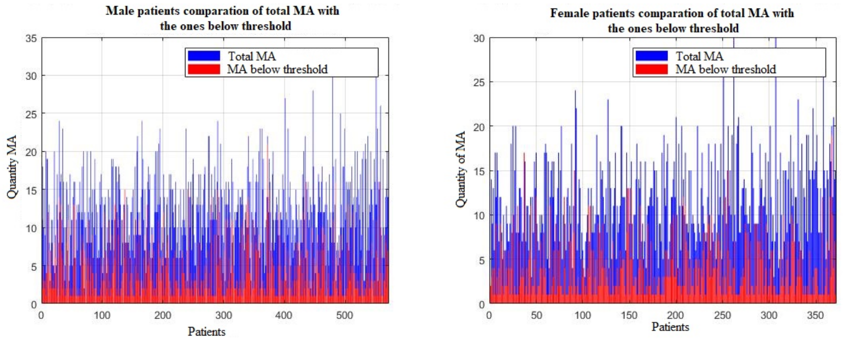


Figure 3.21: Quantity of MA every patient have below HbA1c threshold (7) by sex

There are several patients in both sexes that manage to reduce after just one MA, as show in Figure 3.22 (Left-Male, Right- Female).

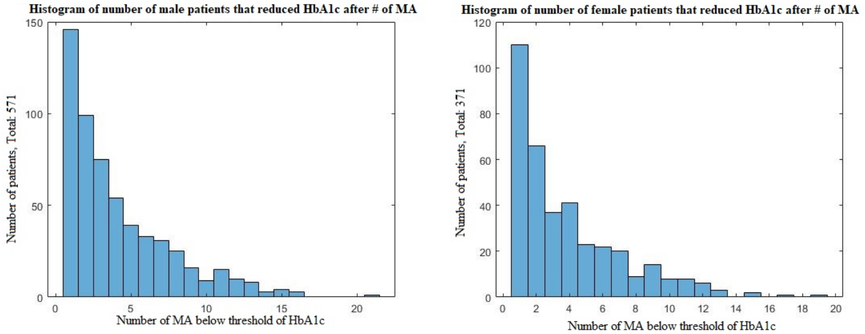


Figure 3.22: Patients who managed to reduce HbA1c below the threshold after number of # MA

As we can see there is a tendency in both histograms: along x-axis the values of y reduce in relation to the last value. We identify some cases where that does not happen, and tried to see if there is any relation with the season of the year.

We chose the patients that reduced their HbA1c values 9 times (of the both sexes), because there is a good amount of individuals in this situation, and Figure 3.23 and Figure 3.24 represents all the HbA1c values they had at the MA's for females and males respectively.

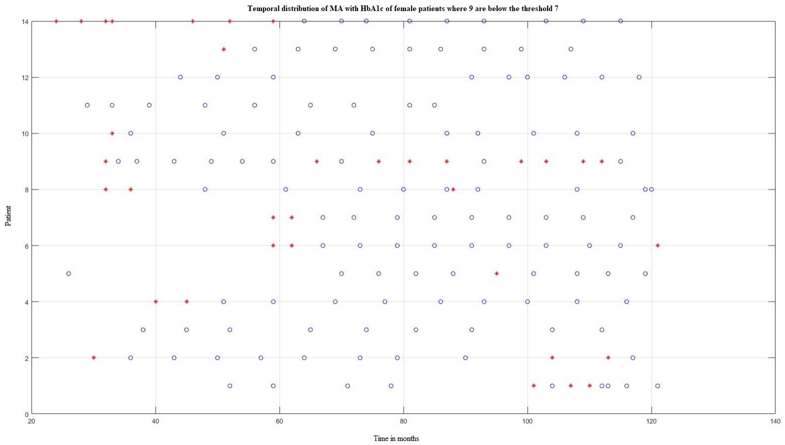


Figure 3.23: Female patients with 9 MA where HbA1c is below the limit along time

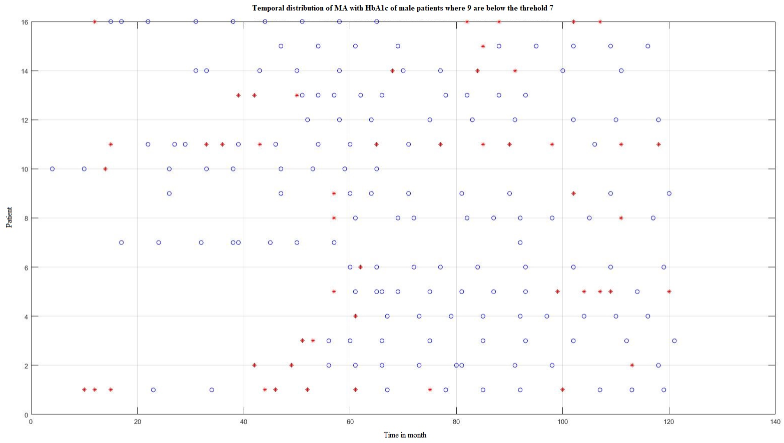
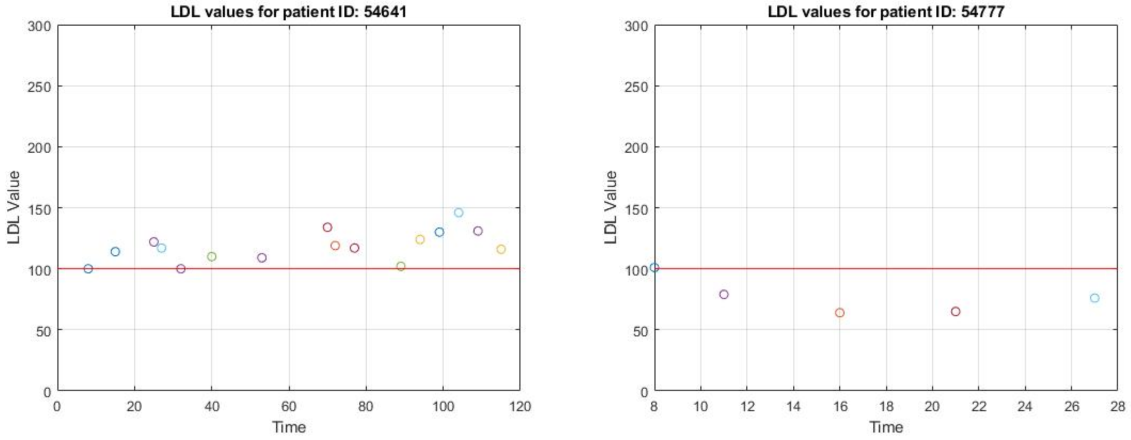


Figure 3.24: Male patients with 9 MA where HbA1c is below the limit along time

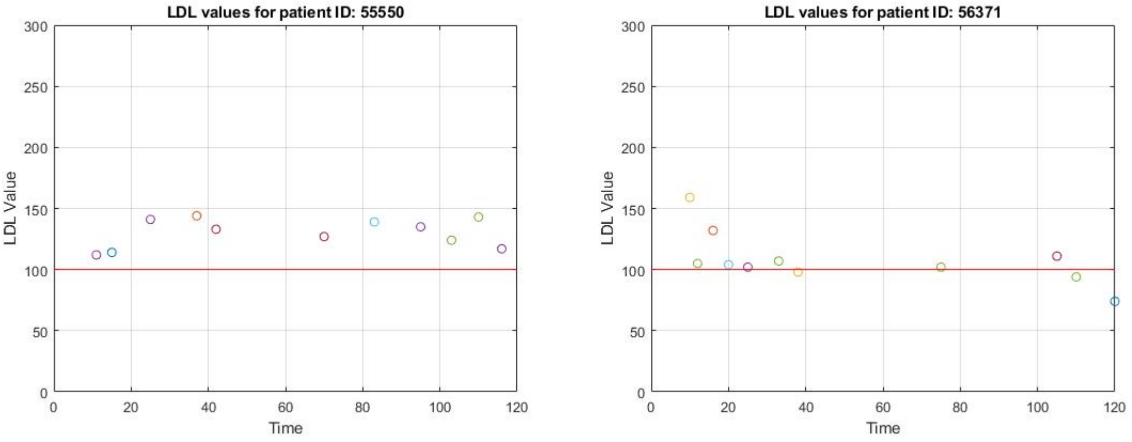
We conclude that, for both sexes, the periods of the year such as Christmas or other festive period, do not influence the raise in HbA1c. Therefore we may conclude that seasons where typically people eat more does not influence the diabetes evolution.

3.2.4 Analyzing LDL values through MA's evolution

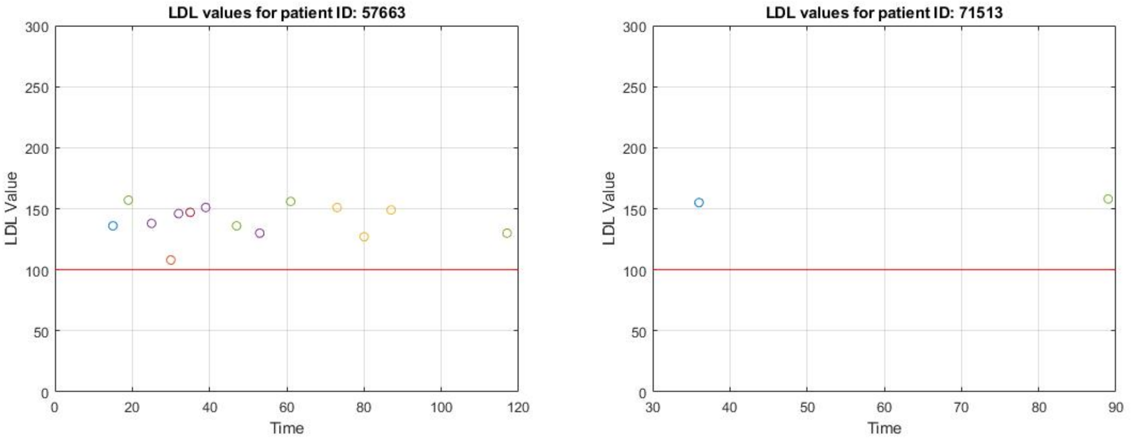
In this part we will consider only the MA presenting LDL values. We obtain a matrix of every value of LDL of every patient, and we randomly choose 10 patients to observe their LDL evolution along the sequence of MA. In this case we consider that a patient is improving his/her health condition if his/her LDL values are below 100 mg/dL.



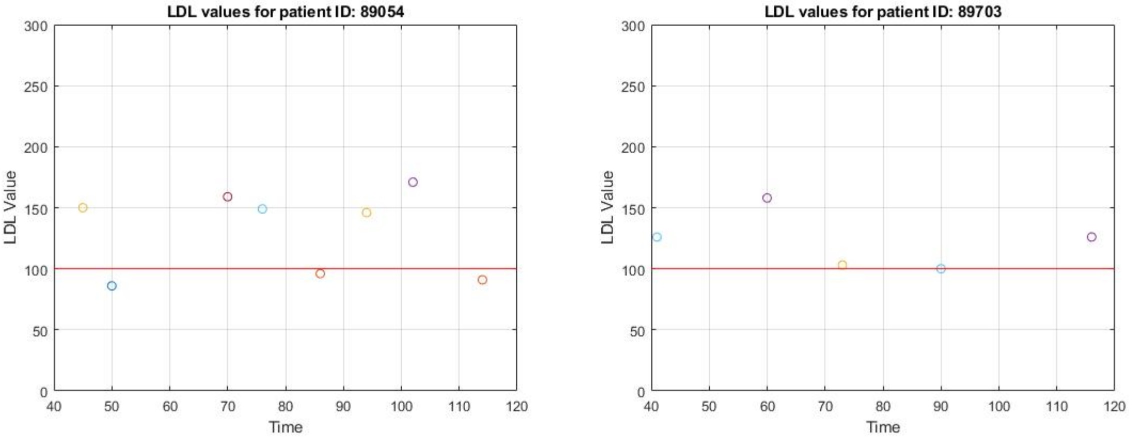
(a)


LDL values for patient ID: 55550
LDL values for patient ID: 56371
LDL Value
Time
0
50
100
150
200
250
300
0
20
40
60
80
100
120

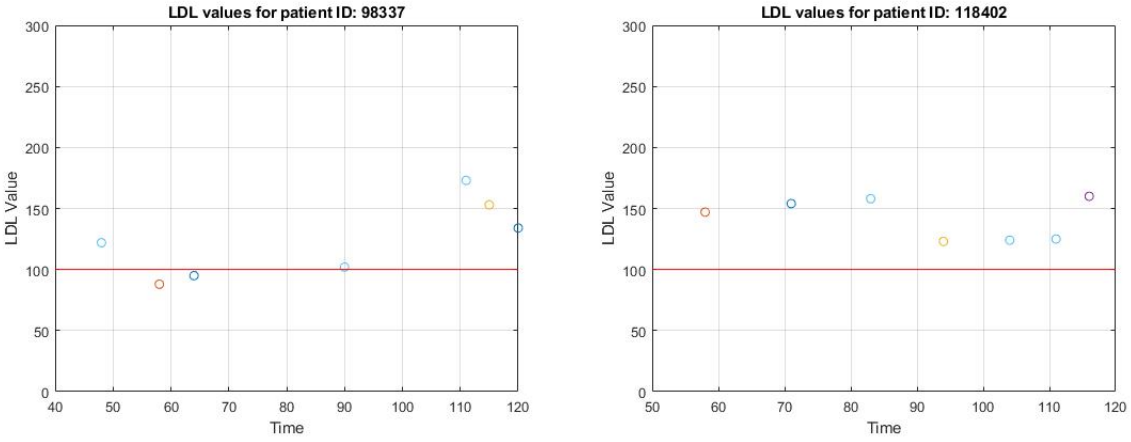
(b)


LDL values for patient ID: 57663
LDL values for patient ID: 71513
LDL Value
Time
0
50
100
150
200
250
300
0
20
40
60
80
100
120
30
40
50
60
70
80
90

(c)


LDL values for patient ID: 89054
LDL values for patient ID: 89703
LDL Value
Time
0
50
100
150
200
250
300
40
50
60
70
80
90
100
110
120

(d)



(e)

Figure 3.25: Temporal evolution of LDL for 10 patients (randomly selected) in months

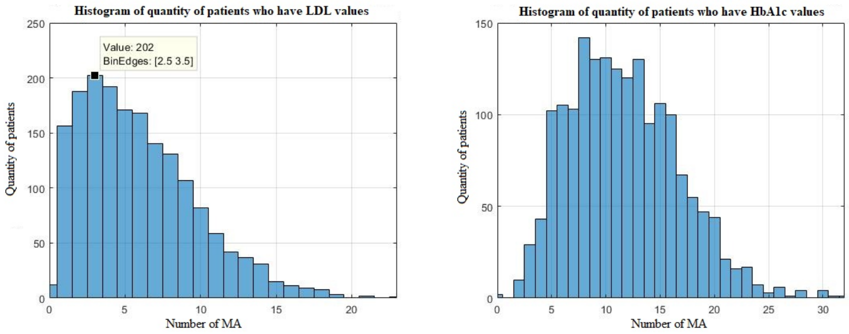
As we can see in the Figure 3.25 all patients have fluctuations in their LDL values. What we can see from those patients is that most of them reduce the values below the optimal value of 100 mg/dL, but they don't manage to keep LDL under the limit, confirming the need for deeper studies on LDL patterns.

We also have one patient ID: 71513 who only has two values for LDL and both are above the limit, with only these values we cannot conclude anything for this patient therefore he will be excluded from the study. Patient ID: 54777 has all the values below the proposed threshold, which is a strong indicator that the parameter is under control.

The database includes a total of 10468 MA where the parameter LDL is registered. In average, the time elapse between Medical Appointments of these patients is 7.9 months.

3.2.5 Joint analysis of HbA1c and LDL values registration through MA's

In this case we want to see how these two parameters interact which each other, and how they limit the amount of patients in regard to the MA provided in the database. We first analyze the histogram of each parameter in Figure 3.26a and 3.26b, and then we consider the patients who have registration of both parameters at the same MA.



(a) (b)

Figure 3.26: Histogram of quantity of patients who have (a) LDL values (b) HbA1c values registered at a MA record

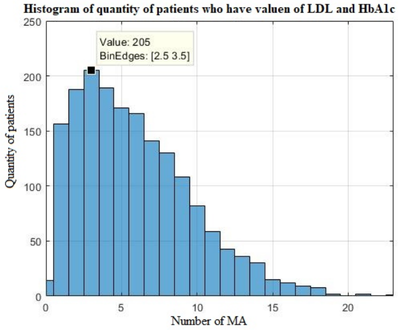


Figure 3.27: The number of patients who have on their MA values of LDL and HbA1c simultaneously registered

As we can see in Figure 3.26 the "dominant" parameter is LDL, since the quantity of patients with more than 5 (five) MA's is dramatically reduced if we consider the histogram (Figure 3.26b) related to HbA1c registrations. In fact, comparing, Figure 3.26a and 3.27, we see that they are similarly shaped.

3.3 Linear Modeling

In this section we present the results obtained with the MATLAB function fitlm, described in section 2.3. Those models are initially applied to all of the data just to give us an idea of how the parameters interfere with each other. The next linear models are established with fewer entries in order to see how each one of the features influence the expected results.

The aim of this research step is to find how we can prevent dyslipidemia by better controlling all the proteins associated with it. As explained in Section 2, controlling LDL is expected to lower CVD related problems, so we decided to create models where the output is the low density lipoprotein (LDL) and we considered it as a function of other linearly related clinical parameters. Doing this we hope to see how LDL is influenced by the parameters and in which amount relative to each other.

After having analyzed the database content and performing the statistical studies previously reported, we selected as sample population all patients who have at least five medical appointments where the reported clinical parameters are: Total Cholesterol, LDL, HDL, Triglyceride and HbA1c. Under these constraints we obtain a database with 4476 MA, which, in statistical terms is very good.

Every tested model assumes LDL as the output variable.

A pre-selection of data, to remove the values that are too far away from a rational range was performed. After removing the outliers, the parameter values to be considered in this study were all normalized between -1 and 1, by considering the maximum value reached by each parameter.

	1st Best Fit	2nd Best Fit	3rd Best Fit
ID	’logistic’	’normal’	’generalized extreme value’
Mês	’generalized pareto’	’generalized extreme value’	’normal’
ColT	’generalized extreme value’	’normal’	’tlocationscale’
LDL	’generalized extreme value’	’normal’	’tlocationscale’
HDL	’generalized extreme value’	’logistic’	’tlocationscale’
Trig	’generalized extreme value’	’tlocationscale’	’logistic’
HbA1c	’generalized extreme value’	’normal’	’tlocationscale’
GGT	’tlocationscale’	’generalized pareto’	’logistic’
Plaq	’generalized extreme value’	’tlocationscale’	’normal’
micro	’generalized pareto’	’generalized extreme value’	’tlocationscale’
Prot	’generalized pareto’	’generalized extreme value’	’logistic’
Crea	’generalized extreme value’	’normal’	’generalized pareto’
MDRD	’extreme value’	’generalized extreme value’	’normal’
MDRD_est	’tlocationscale’	’generalized pareto’	’logistic’
Sex	’extreme value’	’normal’	’tlocationscale’
Age	’extreme value’	’generalized extreme value’	’logistic’
Height	’generalized extreme value’	’normal’	’extreme value’
Weight	’extreme value’	’normal’	’tlocationscale’
BMI	’extreme value’	’generalized pareto’	’normal’
TA_sys	’extreme value’	’logistic’	’normal’
TA_dia	’extreme value’	’logistic’	’normal’

Table 3.4: Best fitted distribution for every parameter

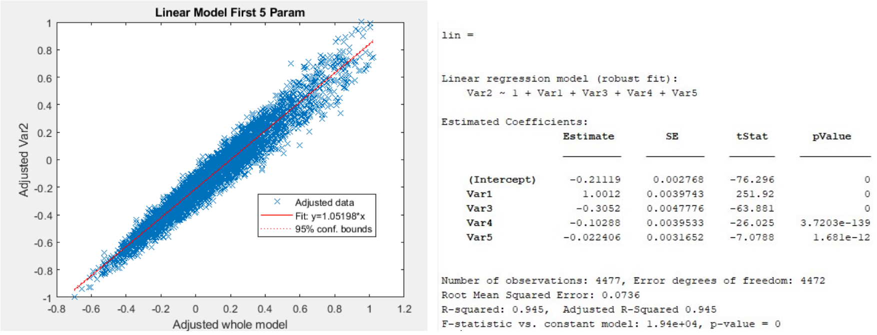
By using the algorithm ’allfitdist.m’, a Matlab function created by Michael Sheppard from MIT Lincoln Laboratory we managed to determine, for each clinical parameter, what fitting distribution would better fit the 1st best fitting model (1st column, in Table 3.4) or the 2nd and 3rd best fitting distributions (2nd and 3rd columns in Table 3.4).

From Table 3.4 we can see that normal distributions appear frequently and for almost all parameters, although the distributions tails differ giving rise to differently named distribution.

Following this stage, some linear models were created by using functions as *fitlm* available in *Matlab*.

3.3.1 Linear Model 1: Total Cholesterol, LDL, HDL, Triglyceride and HbA1c

As a primarily attempt we decided to determine the best fitted model when the most frequently registered parameters are considered, this is, Total Cholesterol, LDL, HDL, Triglyceride and HbA1c. The correspondence of these clinical parameters with the Matlab description is the: Var1 - Total Cholesterol, Var2 - LDL, Var3 - HDL, Var4 - Triglyceride and Var5 - HbA1c. .



```
lin =


Linear regression model (robust fit):
    Var2 ~ 1 + Var1 + Var3 + Var4 + Var5

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   _________   _________   ______   ___________

    (Intercept)     -0.21119    0.002768   -76.296              0
    Var1              1.0012   0.0039743    251.92              0
    Var3             -0.3052   0.0047776   -63.881              0
    Var4            -0.10288   0.0039533   -26.025    3.7203e-139
    Var5           -0.022406   0.0031652   -7.0788      1.681e-12


Number of observations: 4477, Error degrees of freedom: 4472
Root Mean Squared Error: 0.0736
R-squared: 0.945,  Adjusted R-Squared 0.945
F-statistic vs. constant model: 1.94e+04, p-value = 0
```

Figure 3.28: Linear Modelusing Total Cholesterol (Var1), LDL (Var2), HDL (Var3), Triglyceride (Var4), and HbA1c (Var5)

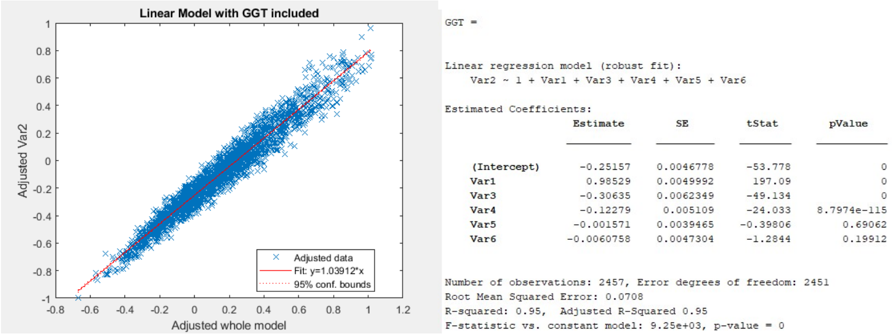
As we can see in Fig 3.28 the model presents a suitable distribution. As our data is already normalized, between -1 and 1, we can see that the RMSE value (0.0736) is very small denoting that the model has a good potential. An additional conclusion is that all these clinical variables influence the obtained LDL values since the resulting linear equation presented in Fig 3.28 involves all clinical parameters, with privilege among them given by the order Total Cholesterol (Var1), HDL (Var3), Triglyceride (Var4) and HbA1c (Var5) as described by the respective coefficients.

In terms of the correspondent data contained in the database we see that 4477 MA are included in the analysis.

As so, we may conclude that this model shows a good LDL prediction.

3.3.2 Linear Model 2 (Model 1 parameters + GGT)

In this model we tested the situation of adding up the clinical parameter GGT. The Matlab numbering of these parameters is the same as in the previous model and Var6 corresponds to GGT.



```
GGT =


Linear regression model (robust fit):
    Var2 ~ 1 + Var1 + Var3 + Var4 + Var5 + Var6

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                  __________    _________    ________    __________

    (Intercept)     -0.25157    0.0046778     -53.778             0
    Var1             0.98529    0.0049992      197.09             0
    Var3            -0.30635    0.0062349     -49.134             0
    Var4            -0.12279     0.005109     -24.033    8.7974e-115
    Var5           -0.001571    0.0039465    -0.39806       0.69062
    Var6          -0.0060758    0.0047304     -1.2844       0.19912


Number of observations: 2457, Error degrees of freedom: 2451
Root Mean Squared Error: 0.0708
R-squared: 0.95,  Adjusted R-Squared 0.95
F-statistic vs. constant model: 9.25e+03, p-value = 0
```

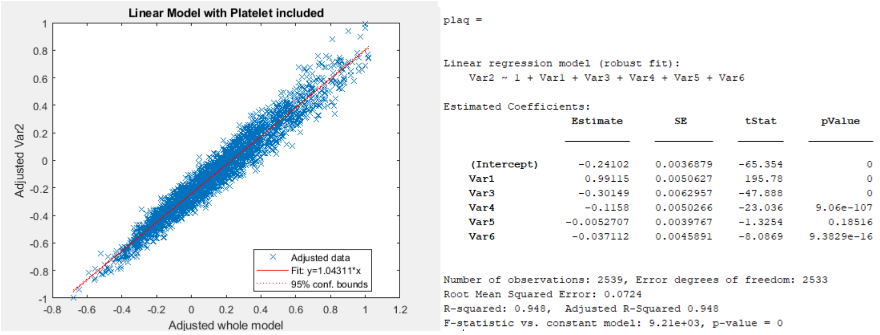
Figure 3.29: Linear Model using Total Cholesterol (Var1), LDL (Var2), HDL (Var3), Triglyceride (Var4), HbA1c (Var5) and GGT (Var6)

As we can see from Figure 3.29 the number of medical appointments diminished drastically, now the model just uses 2457 MA's, which is almost half from our previous model (Linear Model 1). In terms of database data such a short number of MA's limits our expectation of obtaining good fitting results, and it might implicate that we no longer guarantee the insertion criterion of this study of having at least five MA per patient. If this situation happens we are disabled of obtaining the temporal evolution of the parameters.

Analyzing the RMSE results we can see that the obtained value is smaller than the one obtained with model 1 (0.07), indicating that, besides the restriction above mentioned, this model can possibly be used as a good LDL predictor. In terms of the variables' influence on the model, we see that Var5 keeps its low impact on the model, as well as Var6, although Var6 has almost four times higher influence than the Var5. The other variables weight the whole model with coefficients almost 100 times greater than Var5 and Var6.

3.3.3 Linear Model 3 (Model 1 parameters + Platelet)

Now we considered the 5 clinical parameters used on model 2 and instead of using GGT the platelet parameter was considered. As previously, the Matlab's assigned variables are follows: Var1 - Total Cholesterol, Var2 - LDL, Var3 - HDL, Var4 - Triglyceride, Var5 - HbA1c and Var6 - Platelet



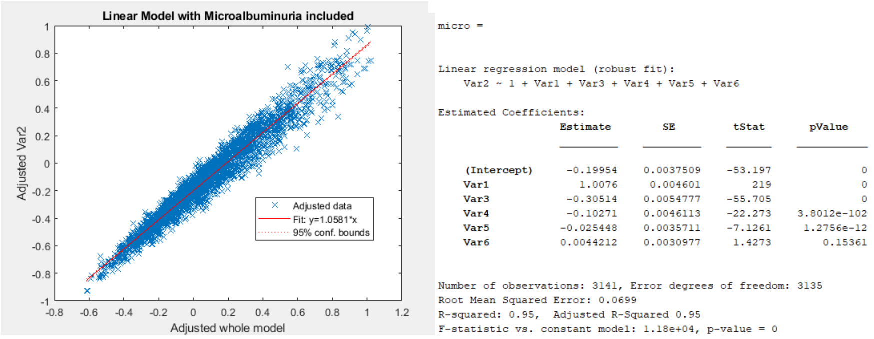
```
plaq =


Linear regression model (robust fit):
    Var2 ~ 1 + Var1 + Var3 + Var4 + Var5 + Var6

Estimated Coefficients:
                    Estimate        SE         tStat       pValue  
                   __________    _________    _______    __________

    (Intercept)      -0.24102    0.0036879    -65.354             0
    Var1              0.99115    0.0050627     195.78             0
    Var3             -0.30149    0.0062957    -47.888             0
    Var4              -0.1158    0.0050266    -23.036     9.06e-107
    Var5           -0.0052707    0.0039767    -1.3254       0.18516
    Var6            -0.037112    0.0045891    -8.0869    9.3829e-16


Number of observations: 2539, Error degrees of freedom: 2533
Root Mean Squared Error: 0.0724
R-squared: 0.948,  Adjusted R-Squared 0.948
F-statistic vs. constant model: 9.21e+03, p-value = 0
```

Figure 3.30: Linear Model using Total Cholesterol (Var1), LDL (Var2), HDL (Var3), Triglyceride (Var4), HbA1c (Var5) and Platelet (Var6)

Figure 3.30 represents the results obtained for this model. The number of Medical appointments involved is similar to those involved in model 2, leading to the same conclusion about the volume and type of data to create the model. The RMSE value was also kept on lower values, which again is a good indicator. In terms of variable influences, the Var6 (platelet) coefficient is higher than in the last model. Therefore we may say that platelet is more influent on the linear modelling of LDL, being more influent (at least ten times) than HbA1c.

3.3.4 Linear Model 4 (Model 1 parameters + Microalbuminuria)

In this model the effect of Microalbuminuria (MAU) in model 1's parameters will be tested. Matlab maintained the same labeling of the first 5 variables and attributed Var6 to Microalbuminuria.



```
micro =


Linear regression model (robust fit):
    Var2 ~ 1 + Var1 + Var3 + Var4 + Var5 + Var6

Estimated Coefficients:
                    Estimate        SE         tStat       pValue   
                   _________    _________    _______    ___________

    (Intercept)     -0.19954    0.0037509    -53.197              0
    Var1              1.0076     0.004601        219              0
    Var3            -0.30514    0.0054777    -55.705              0
    Var4            -0.10271    0.0046113    -22.273    3.8012e-102
    Var5           -0.025448    0.0035711    -7.1261     1.2756e-12
    Var6           0.0044212    0.0030977     1.4273        0.15361


Number of observations: 3141, Error degrees of freedom: 3135
Root Mean Squared Error: 0.0699
R-squared: 0.95,  Adjusted R-Squared 0.95
F-statistic vs. constant model: 1.18e+04, p-value = 0
```

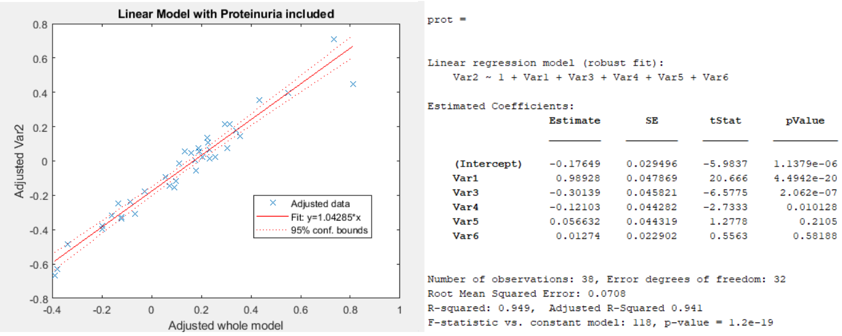
Figure 3.31: Linear Model using Total Cholesterol (Var1), LDL (Var2), HDL (Var3), Triglyceride (Var4), HbA1c (Var5) and Microalbuminuria (Var6)

From Figure 3.31 one may conclude that MAU has a very small influence on the LDL values. The number of medical appointments is higher than in the previous model and, again, the value of RMSE shows promising results for this model.

Regarding the weights of HbA1c and MAU on the whole model we may conclude that these two parameters have very few influence on the model. This conclusion is supported by Linear Model 1's results since when on both models we can see that HbA1c has not a strong supported by the results obtained on Linear Model 1, since on both models we see that both HBA1c and MAU present weak influence on the model.

3.3.5 Linear Model 5 (Model 1 parameters + Proteinuria)

Just like in model 4, this model is obtained by adding Proteinuria to the set of 5 initial parameters, and again, *Matlab* numbered this sixth variable as Var6.



```
prot =


Linear regression model (robust fit):
    Var2 ~ 1 + Var1 + Var3 + Var4 + Var5 + Var6

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)    -0.17649    0.029496    -5.9837    1.1379e-06
    Var1            0.98928    0.047869     20.666    4.4942e-20
    Var3           -0.30139    0.045821    -6.5775     2.062e-07
    Var4           -0.12103    0.044282    -2.7333      0.010128
    Var5           0.056632    0.044319     1.2778        0.2105
    Var6            0.01274    0.022902     0.5563       0.58188


Number of observations: 38, Error degrees of freedom: 32
Root Mean Squared Error: 0.0708
R-squared: 0.949,  Adjusted R-Squared 0.941
F-statistic vs. constant model: 118, p-value = 1.2e-19
```

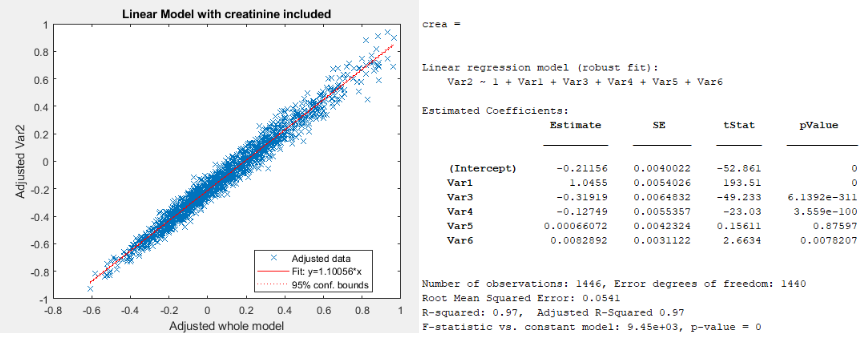
Figure 3.32: Linear Model using Total Cholesterol (Var1), LDL (Var2), HDL (Var3), Triglyceride (Var4), HbA1c (Var5) and Proteinuira (Var6)

The variable we added now, as we can see in the Figure 3.32, cannot be used to model LDL as a linear model because it is a feature with very little presence in the database, this is, with only 38 observations.

Looking at the results obtained, the values of RMSE and the weight of Proteinuria in the model is considerable comparing to the influence of HbA1c on model 2 (for instance). But the available number of studying-cases to be considered would have no statistical reliability.

3.3.6 Linear Model 6 (Model 1 parameters + Creatinine)

In this last model tested, the five parameters of model 1 are now combined with Creatinine. Matlab assigned Var6 to Creatinine



```
crea =


Linear regression model (robust fit):
    Var2 ~ 1 + Var1 + Var3 + Var4 + Var5 + Var6

Estimated Coefficients:
                     Estimate        SE         tStat       pValue
                    __________    _________    ________    ___________

    (Intercept)       -0.21156    0.0040022     -52.861              0
    Var1                1.0455    0.0054026      193.51              0
    Var3              -0.31919    0.0064832     -49.233    6.1392e-311
    Var4              -0.12749    0.0055357      -23.03    3.559e-100
    Var5            0.00066072    0.0042324     0.15611        0.87597
    Var6             0.0082892    0.0031122      2.6634      0.0078207


Number of observations: 1446, Error degrees of freedom: 1440
Root Mean Squared Error: 0.0541
R-squared: 0.97,  Adjusted R-Squared 0.97
F-statistic vs. constant model: 9.45e+03, p-value = 0
```

Figure 3.33: Linear Model using Total Cholesterol (Var1), LDL (Var2), HDL (Var3), Triglyceride (Var4), HbA1c (Var5) and Creatinine (Var6)

The results of this model are presented in Figure 3.33 demonstrating that parameters HbA1c (Var5) and Creatinine (Var6) do not have major influence in the LDL linear model. The model presents a very good RMSE value (0.0541), but what really matters is the influence of the features, and in this respect we conclude that the Creatinine do not change significantly the expected results of predicted LDL values.

3.3.7 Performance summary

Linear Models	RMSE	R^2
Model 1	0.074	0.945
Model 2	0.071	0.950
Model 3	0.070	0.950
Model 4	0.072	0.948
Model 5	0.054	0.970
Model 6	0.071	0.949

Table 3.5: Comparative Results of the Linear Models

To conclude, Table 3.5 enables a comparison of the six models' performance. RMSE indicates how the predicted data are around our linear model, while the R2 coefficient is a statistical measurement of how well the regression predictions approximate the real data points.

The minimum RMSE (0,054) and the maximum R2 (0,970) is obtained on model 5, this is to say, the best linear modeling of LDL should consider Total Cholesterol, HDL, Triglyceride, HbA1c and Proteinuria. But, since only 38 observations were considered, this model has to be discarded. So, Model 3, presenting an RMSE of 0.07 and the maximum R2 of 0.950 should be the linear model to consider. This model considers Total Cholesterol, LDL, HDL, Triglyceride, HbA1c and Platelet. Within 2539 observations, the model gives more than 10 times more relevance to Platelets than to HbA1c, both of them presenting much less influence on the model than the other clinical parameters.

Next chapter describes the tests implemented to model LDL with non-linear models, provided by a Multi-Objective Genetic Algorithm.

4 Multi-Objective Genetic Algorithm (MOGA)

Multi-objective formulations tend to present more realistic models for many complex optimization problems. Quite often, in real-life problems such as this particular case-study where clinical data is under analysis, the objectives under consideration conflict with each other. So, when optimizing a particular solution with respect to a single objective (as was the case of the linear models described in chapter 3) we may obtain unacceptable results with respect to the other objectives. A reasonable solution to a multi-objective problem is to investigate a set of solutions, each of which satisfies the objectives at an acceptable level without being dominated by any other solution.

In this chapter the experiments made to model LDL using the model proposed by the Multi-Objective Genetic Algorithm (MOGA) are described.

Making use of a Matlab interactive tool, the so called Stepwise algorithm, a complementary study was developed to understand at what extend each clinical parameter influences the LDL modeling. In this context, the training matrix of each MOGA model tested was therefore analyzed through the Stepwise algorithm. Results obtained are reported and comments on their comparison are included. Results are also compared with the results obtained with the linear models obtained when using the same populations identified by MOGA.

Using the same strategy as described in Chapter 3, particularly section 3.3, we decided to use most of the clinical parameters available in the database, except for those database fields concerned with complications and medication, due to the type of data of their content.

Care is also taken to avoid creating models based on statistically irrelevant populations and those patients who had less than 5 MAs during the 10 years period under analysis.

As a first processing stage, the population outliers are excluded and then we proceed to a data normalization, using the same strategy as when creating linear models (Chapter 3).

To remove the outliers we assumed ID>0 and month >0, and used the following restrictions, indicating for each variable the maximum values that can be considered within a normal range (and not an error):

Total Cholesterol<300;

LDL<250;

HDL<100;

Triglycerides <600;

HbA1c<13;

GGT<200;

Platelet<450;

Microalbuminuria<500;

Proteinuria<500;

Creatinine<500;

MDRD<500;

MDRD_Estadio<500;

Sex<500;

Age<500;

Height<500;

Weight <500;

Body Mass Index<500;

TA Sistólica<500;

TA Diastólica<500.

We assumed this maximum limitations to remove the most unreal values each parameter has. The 500 limit for some parameter are just to make sure that we don't have values that are unreal.

After scaling the values, the population was divided into three sets, such that 60% of the data is used for training the models, 20% for testing the models and the last 20% was used to validate the model.

Researching the performance of different models, several MOGA runs with different formulations were performed. To ease the description we will enumerate every test run as MOGA #, giving rise to the below sections.

Each case-study will use a number between 2 and 25 neurons and the input terms will vary according to the model. The algorithm will do five training tests for each candidate using the best compromise trial to compute the desired objectives. For the execution part we will use a number of 50 generations each with a population size of 100.

For each modeling case, MOGA was run twice. The first time without any restriction and the second time imposing a restriction on the value of RMSE for the training set aiming at obtaining better models. The RMSE maximum allowed value was imposed according to the particular model in study.

To select the restriction goals we used a Matlab script named "analisa_arx_v3" provided by Prof. Doutor António Ruano. This script used the obtained non-dominated solutions from MOGA whose Euclidean norm of the vector of linear weights (w) are below or equal to a user-specifed threshold (EN).

In all experiments LDL was considered to be the feature to be predicted. LDL was considered as output variable and the other input variables varied according to the specific experiment.

4.1 MOGA 1

In this experiment we modeled LDL as a function of 9 variables. For simplicity they will be numbered as follows: Var0 -LDL, Var1 - Total Cholesterol, Var2 - HDL, Var3 - Triglyceride, Var4 - HbA1c, Var5 - GGT, Var6 - Platelets,Var7 - MAU, Var8 - Creatinine, Var9 - Month of MA.

With all of these features only 830 MA from the database could be used, which is a very small population. The population was divided in three sets as follows, where the training set included 60% of the whole population:

- Training - 498 Observations
- Testing - 166 Observations
- Validation - 166 Observations.

After obtaining the matrices we introduced them in the MOGA software using the parameters of the algorithm as explained before.

4.1.1 Results

The given results comes to reinforce what we saw in section 3.4 using some of the features alone with the first five, and as we can see in the Figure 4.1 Var4 - HbA1c, Var5 - GGT, Var6 - Platelets presents the least presence in the obtained models.

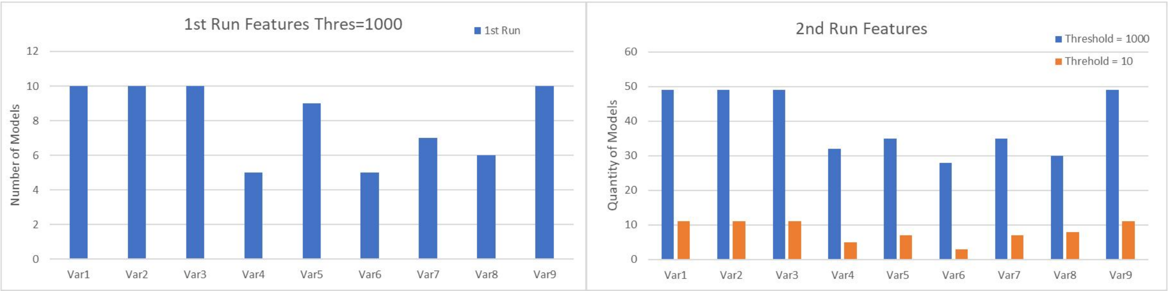


Figure 4.1: MOGA 1 - Histograms of number of models using each variable

Features	1stRun Thres:1000	2ndRun Thres:1000	2ndRun Thres:10
Var1	10	49	11
Var2	10	49	11
Var3	10	49	11
Var4	5	32	5
Var5	9	35	7
Var6	5	28	3
Var7	7	35	7
Var8	6	30	8
Var9	10	49	11

Table 4.1: MOGA 1 - Table with values referent to Figure 4.1

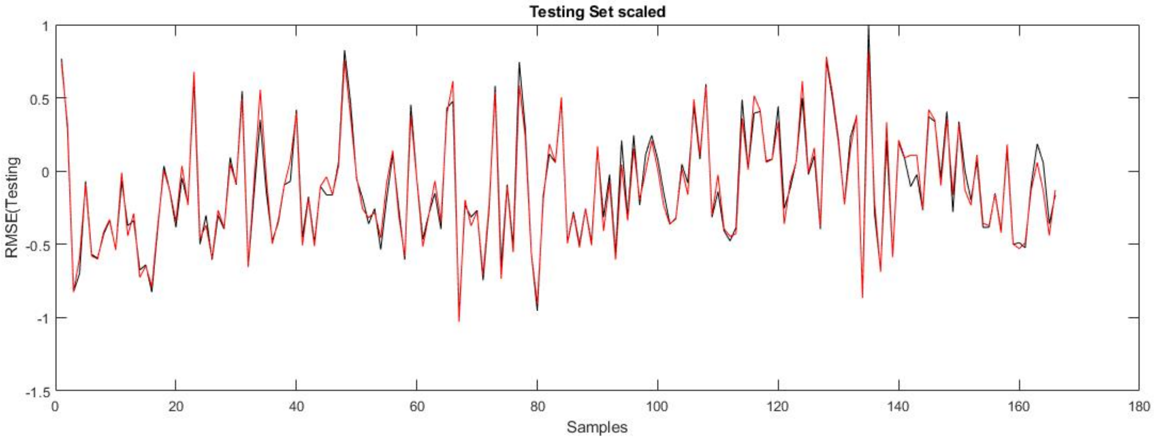
For the second run the selected RMSE training goal was: 0.037852.

We analyzed the second run with two EN thresholds in order to find the best model but also to see how the presence of the features evolves. Looking to both histograms their pattern is similar.

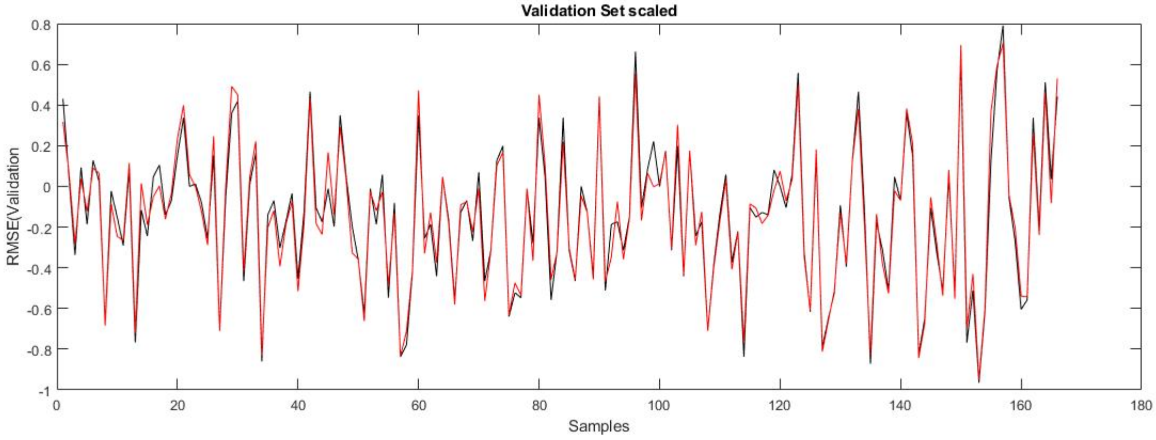
model	1819
y(k)=f(v1(k),v2(k),v3(k),v4(k),v5(k),v7(k),v8(k),)	
RMSE Training scaled	0.048
RMSE Testing scaled	0.065
RMSE Validation scaled	0.064

Table 4.2: MOGA 1 - Preferred Model

The model we selected was chosen by looking at the list we obtained using the lowest threshold and we chose the one that had the smallest RMSE training and validation values. In the Table 4.2 we have the values obtained by the model and we can also see that this model uses as parameters Var1, Var2, Var3, Var4, Var5, Var7 and Var8.



(a)



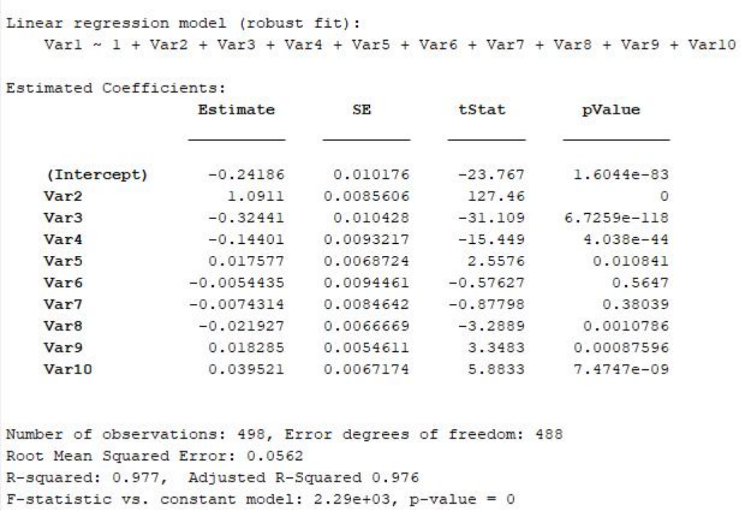
(b)

Figure 4.2: Predicted values (red) and Original Data (blue) of Model 1819 during Testing (a) and Validation (b) procedures

Figure 4.2 shows us how the model predicted the data versus the original values using the testing and validation observations. We can see the predicted data colored in red and the original values in blue, also if we look closely at both graphics presented in Figure 4.2 we see that the predicted data follows the original really close with some minor deviations. The results are pretty good besides the limitation of using a low number of observations.

4.1.2 Linear Models

In this sub-section we are going to use the same data used in section 4.1.1 and see how a linear model will behave and what result we obtain. Due to different software used Var1 to Var10 is to simplify the usage of available algorithms and will be assigned as follow : Var1 -LDL, Var2 - Total Cholesterol, Var3 - HDL, Var4 - Triglyceride, Var5 - HbA1c, Var6 - GGT, Var7 - Platelets,Var8 - MAU, Var9 - Creatinine, Var10 - Month of MA. In section 3, these clinical parameters were also numbered for facility of representation.



```
Linear regression model (robust fit):
    Var1 ~ 1 + Var2 + Var3 + Var4 + Var5 + Var6 + Var7 + Var8 + Var9 + Var10

Estimated Coefficients:
                   Estimate        SE          tStat        pValue
                  __________    _________    _________    __________

    (Intercept)     -0.24186     0.010176      -23.767    1.6044e-83
    Var2              1.0911    0.0085606       127.46             0
    Var3            -0.32441     0.010428      -31.109   6.7259e-118
    Var4            -0.14401    0.0093217      -15.449     4.038e-44
    Var5            0.017577    0.0068724       2.5576      0.010841
    Var6          -0.0054435    0.0094461     -0.57627        0.5647
    Var7          -0.0074314    0.0084642     -0.87798       0.38039
    Var8           -0.021927    0.0066669      -3.2889     0.0010786
    Var9            0.018285    0.0054611       3.3483    0.00087596
    Var10           0.039521    0.0067174       5.8833    7.4747e-09


Number of observations: 498, Error degrees of freedom: 488
Root Mean Squared Error: 0.0562
R-squared: 0.977,  Adjusted R-Squared 0.976
F-statistic vs. constant model: 2.29e+03, p-value = 0
```

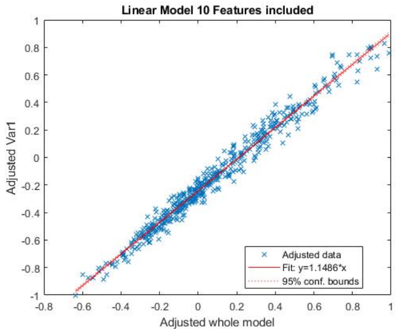
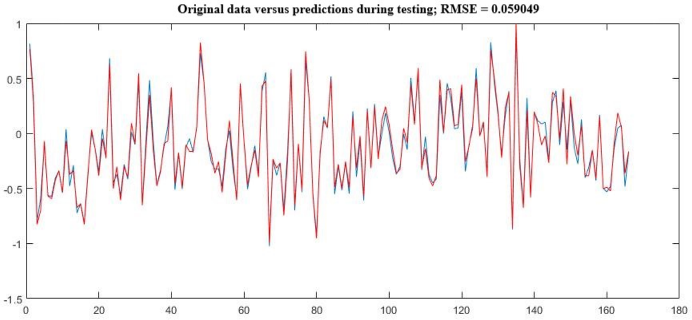
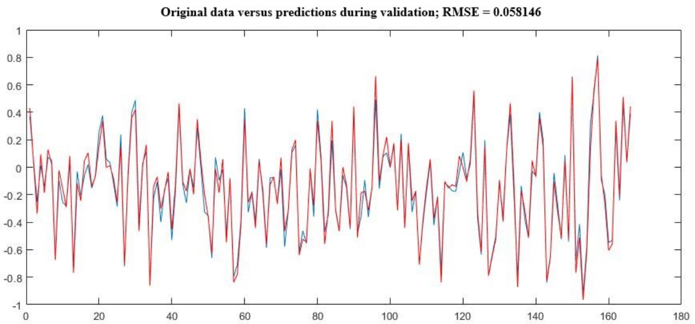


Figure 4.3: Linear model using MOGA 1 Training matrix

From Figure 4.3 we see the values of the weights attributed to each feature. The results confirm the weak contribution of Var6 and var7 (GGT and Platelets respectively) on the model just like MOGA-1 did, as represented in Figure 4.1. The selected MOGA-1 model, model 1819 (Table 4.2) presented a RMSE during training of 0.048, this is, less than the linear model hereby presented where RMSE = 0.0562.



(a)



(b)

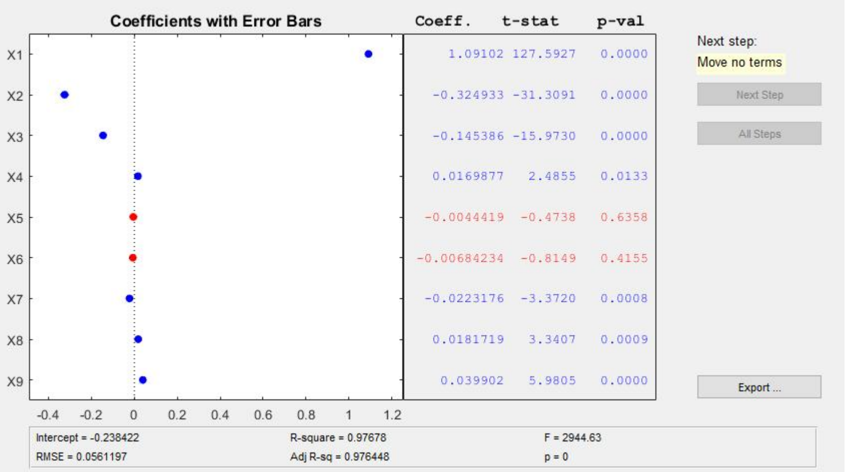
Figure 4.4: Predicted values (red) and Original Data (blue) of the Linear Model using Testing (a) and Validation (b) matrices provided by MOGA

Comparing Figure 4.4 with Figure 4.2 we can see that the result from MOGA are following better the expected values, by visual inspection. Which is confirmed by the RMSE correspondent values.

4.1.3 Stepwise

We use this interactive function to see how the linear models behaves by adding or removing features from the model. Labels of the parameters: X1 - Total Cholesterol, X2 - HDL, X3 - Triglyceride, X4 - HbA1c, X5 - GGT, X6 - Platelets,X7 - MAU, X8 - Creatinine, X9 - Month of MA are now employed.

To test the values with the stepwise function we used the Training matrix obtained and used in previously models.



	Coeff.	t-stat	p-val
X1	1.09102	127.5927	0.0000
X2	-0.324933	-31.3091	0.0000
X3	-0.145386	-15.9730	0.0000
X4	0.0169877	2.4855	0.0133
X5	-0.0044419	-0.4738	0.6358
X6	-0.00684234	-0.8149	0.4155
X7	-0.0223176	-3.3720	0.0008
X8	0.0181719	3.3407	0.0009
X9	0.039902	5.9805	0.0000

Next step:
Move no terms

Intercept = -0.238422 R-square = 0.97678 F = 2944.63
RMSE = 0.0561197 Adj R-sq = 0.976448 p = 0

Figure 4.5: Stepwise panel obtained when MOGA-1 training matrix is used

As we can see in Figure 4.5 we conclude that GGT and Platelet should be discarded, as they are not statistically significant, since the algorithm did not included the variables signaled with red in the model once they would increase the model's RMSE.

4.2 MOGA 2

From section 4.1 we saw that some features could be discarded. This way we decided to eliminate GGT and Platelet. In section 3.4 we also saw that those two factors did have a low influence in the linear models.

In this section labels are as follows: Var0 -LDL, Var1 - Total Cholesterol, Var2 - HDL, Var3 - Triglyceride, Var4 - HbA1c, Var5 - MAU, Var6 - Creatinine, Var7 - Month of MA, Var8 - Sex, Var9 - Age.

With all these features we obtained a total of 1410 MA, which is still a low value for a statistical population but better than the last one. As in section 4.1, the observation was divided in three matrices as follows:

- Training - 846 Observations
- Testing - 282 Observations
- Validation - 282 Observations.

4.2.1 Results

Again the first run had no restrictions while, for the second run the selected RMSE training goal was: 0.042774.

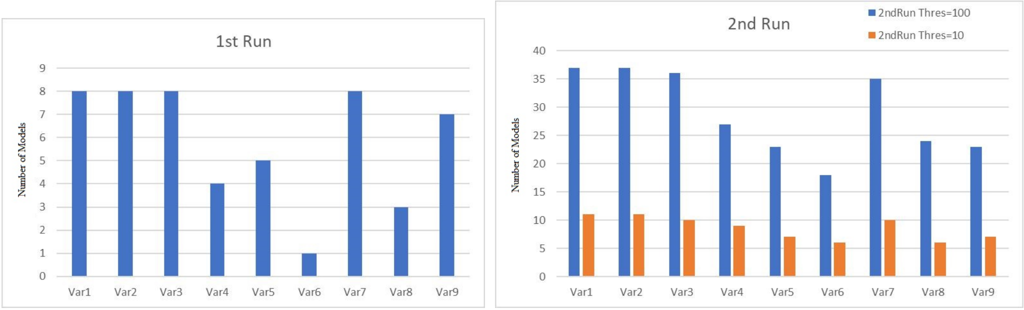


Figure 4.6: MOGA 2 - Histograms of number of models using each variable

Features	1stRun Thres:100	2ndRun Thres:10	2ndRun Thres:100
Var1	8	11	37
Var2	8	11	37
Var3	8	10	36
Var4	4	9	27
Var5	5	7	23
Var6	1	6	18
Var7	8	10	35
Var8	3	6	24
Var9	7	7	23

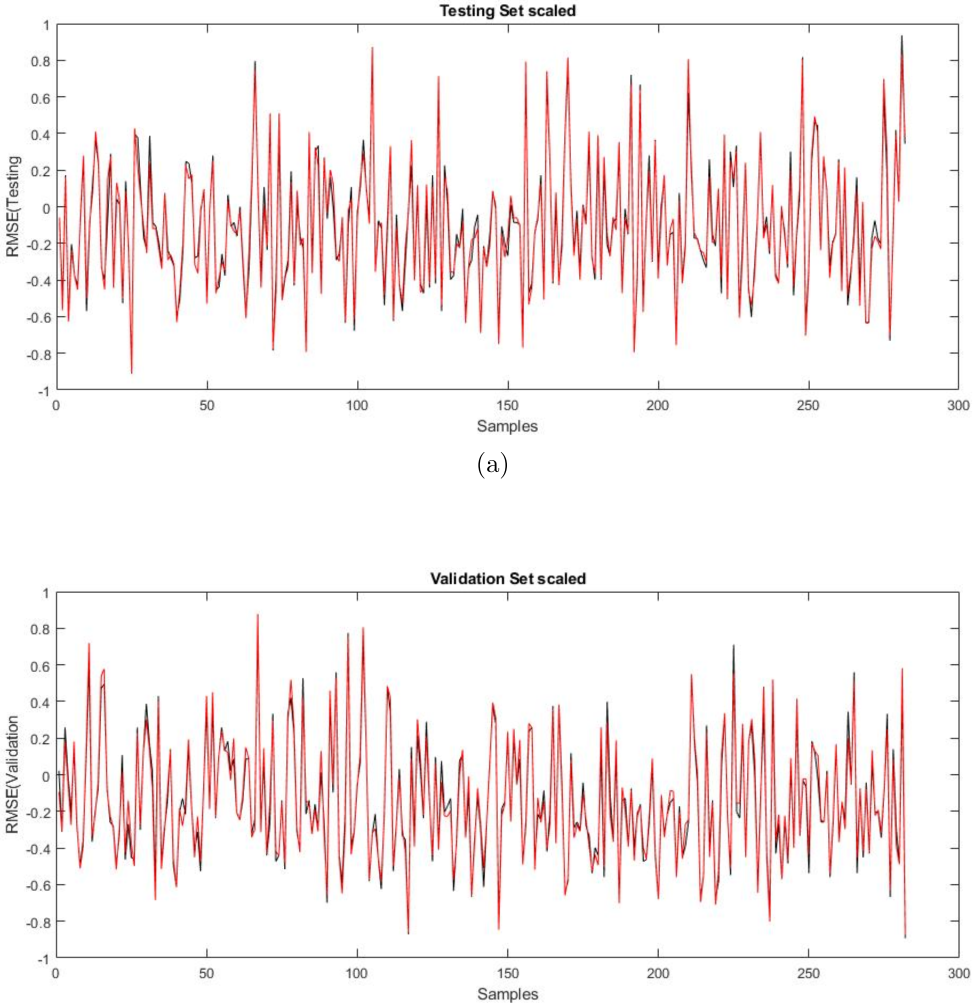
Table 4.3: MOGA 2 - Table with values referent to Figure 4.6

In terms of quantity of models where our clinical parameters are present we see that Creatinine (Var6) does not appear in many models, being the one that has the least occurrences. MAU (Var5) and age (Var8) also have little number of appearances so we can consider them as features with little impact on LDL as shown in Figure 4.6.

Model	599
y(k)=f(v1(k),v2(k),v3(k),v4(k),v5(k),v6(k),v7(k),v8(k),)	
RMSE Training scaled	0.044
RMSE Testing scaled	0.052
RMSE Validation scaled	0.057

Table 4.4: MOGA 2 - Preferred Model

Looking at the chosen model in the Table 4.4 we can see that it considers almost all the variables but not the age. As stated before age is a factor that has little presence. The selection of model 599 (Table 4.4) was somehow random, as we looked at various model within the least chosen threshold and selected the one that seemed to have the smallest RMSE values.



(a)

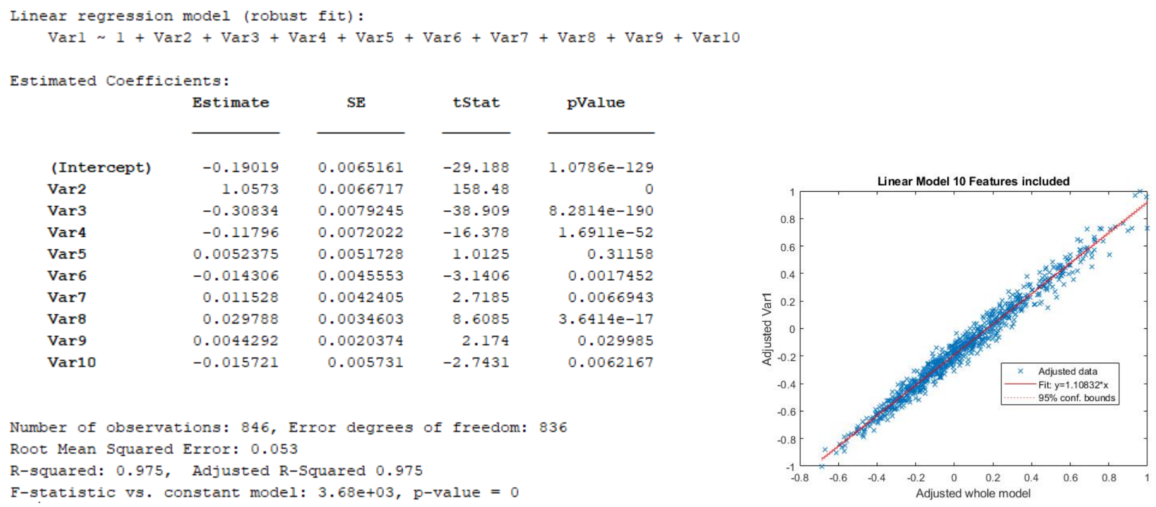
(b)

Figure 4.7: Predicted values (red) and Original Data (blue) of Model 599 during Testing (a) and Validation (b) procedures

Figure 4.7, shows how the model behaves during testing and validation stages, and we easily see that the predicted curves follow the original values (in blue) very close.

4.2.2 Linear Model

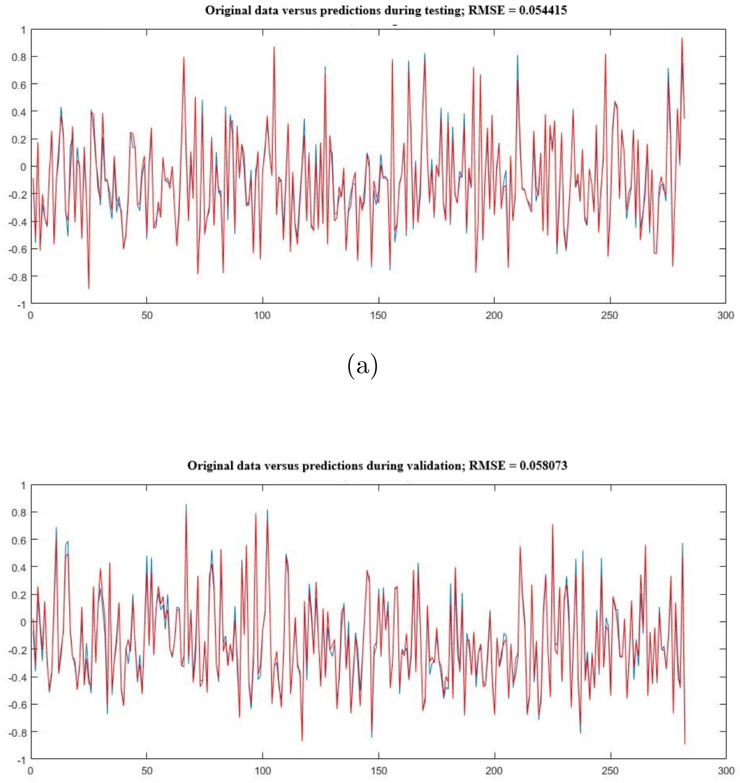
As we did before, MOGA-2 training, testing and validation matrices were used to build a linear model. Again, the linear model features were renamed as follows: Var1 - LDL, Var2 - Total Cholesterol, Var3 - HDL, Var4 - Triglyceride, Var5 - HbA1c, Var6 - MAU, Var7 - Creatinine, Var8 - Month of MA, Var9 - Sex, Var10 - Age.



```
Linear regression model (robust fit):
    Var1 ~ 1 + Var2 + Var3 + Var4 + Var5 + Var6 + Var7 + Var8 + Var9 + Var10

Estimated Coefficients:
                   Estimate        SE          tStat       pValue
                   _________    _________    _______    ___________

    (Intercept)     -0.19019    0.0065161    -29.188    1.0786e-129
    Var2              1.0573    0.0066717     158.48              0
    Var3            -0.30834    0.0079245    -38.909    8.2814e-190
    Var4            -0.11796    0.0072022    -16.378     1.6911e-52
    Var5           0.0052375    0.0051728     1.0125        0.31158
    Var6           -0.014306    0.0045553    -3.1406      0.0017452
    Var7            0.011528    0.0042405     2.7185      0.0066943
    Var8            0.029788    0.0034603     8.6085     3.6414e-17
    Var9           0.0044292    0.0020374      2.174       0.029985
    Var10          -0.015721     0.005731    -2.7431      0.0062167


Number of observations: 846, Error degrees of freedom: 836
Root Mean Squared Error: 0.053
R-squared: 0.975,  Adjusted R-Squared 0.975
F-statistic vs. constant model: 3.68e+03, p-value = 0
```

Figure 4.8: Linear model using the MOGA-2 Training matrix

The linear regression model we obtained is shown in Figure 4.8. According to the estimated values (left column of table) one can observe that Var5 - HbA1c and Var9 -Sex, have the lowest contribution to the model.



(a)

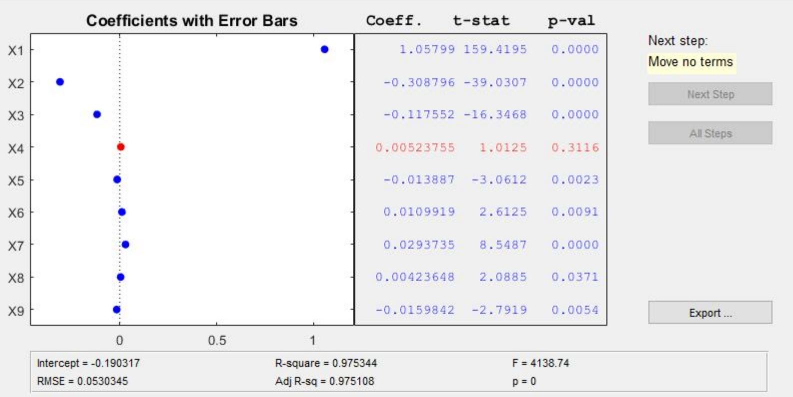
(b)

Figure 4.9: Predicted values (red) and Original Data (blue) of the Linear Model using the MOGA 2 Testing (a) and Validation(b) matrices

Figure 4.9 show the predicted values (in red) using the linear model follow the original values, both for testing and for validation, presenting similar behavior, with similar RMSE values, these being similar to the RMSE value obtained with MOGA 2 model 599, as may be seen by comparison with 4.7.

4.2.3 Stepwise

Labeling the clinical parameters as: X1 - Total Cholesterol, X2 - HDL, X3 - Triglyceride, X4 - HbA1c, X5 - MAU- , X6 - Creatinine,X7 - Month of MA, X8 - Sex, X9 - Age, and applying the Stepwise algorithm to the training matrix correspondent to MOGA-2 model,Figure 4.10, demonstrates that variable X4, this is HbA1c, presents a very weak contribution to the model, being recommended its exclusion from the model.



	Coeff.	t-stat	p-val
X1	1.05799	159.4195	0.0000
X2	-0.308796	-39.0307	0.0000
X3	-0.117552	-16.3468	0.0000
X4	0.00523755	1.0125	0.3116
X5	-0.013887	-3.0612	0.0023
X6	0.0109919	2.6125	0.0091
X7	0.0293735	8.5487	0.0000
X8	0.00423648	2.0885	0.0371
X9	-0.0159842	-2.7919	0.0054

Next step:
Move no terms

Intercept = -0.190317
RMSE = 0.0530345
R-square = 0.975344
Adj R-sq = 0.975108
F = 4138.74
p = 0

Figure 4.10: Stepwise panel obtained when MOGA 2 training matrix is employed

4.3 MOGA 3

After the results we obtained in the previous two models we considered that the variable month of the MAs should not be used as a MOGA feature. In fact, the month of the MA was referenced to the first month of 2008 corresponding to the time the patient joined APDP, but that did not correspond to the beginning of the disease. Also, we were testing the whole data available at the database, therefore we could not guarantee that all patients had more than 5 MA.

Labels of the features used in this model are: Var0 -LDL, Var1 - Total Cholesterol, Var2 - HDL, Var3 - Triglyceride, Var4 - HbA1c, Var5 - GGT, Var6 - Platelet, Var7 -MAU, Var8 - Creatinine.

With all this features we obtained a total of 830 MA, which is very low number since it was divided in three matrices as follows:

- Training - 498 Observations
- Testing - 166 Observations
- Validation - 166 Observations.

4.3.1 Results

Our result continue similar to the other models, Var5 - GGT and Var6 - Platelet mantain their low influence in the model. Now we can see that Var4 - HbA1c improved its influence and the other parameters keep their levels of occurrence.

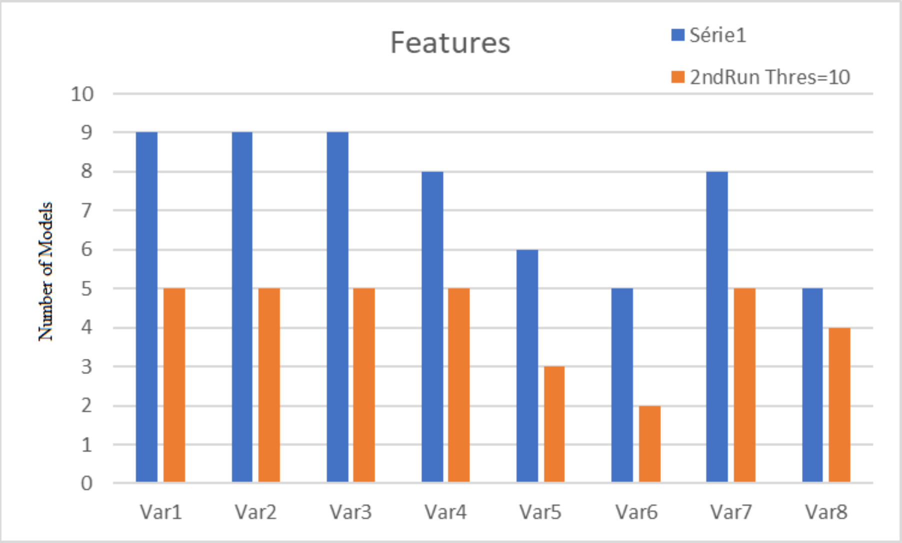


Figure 4.11: MOGA 3 - Histograms of number of models using each variable

Features	1stRun Thres:100	2ndRun Thres:10
Var1	9	5
Var2	9	5
Var3	9	5
Var4	8	5
Var5	6	3
Var6	5	2
Var7	8	5
Var8	5	4

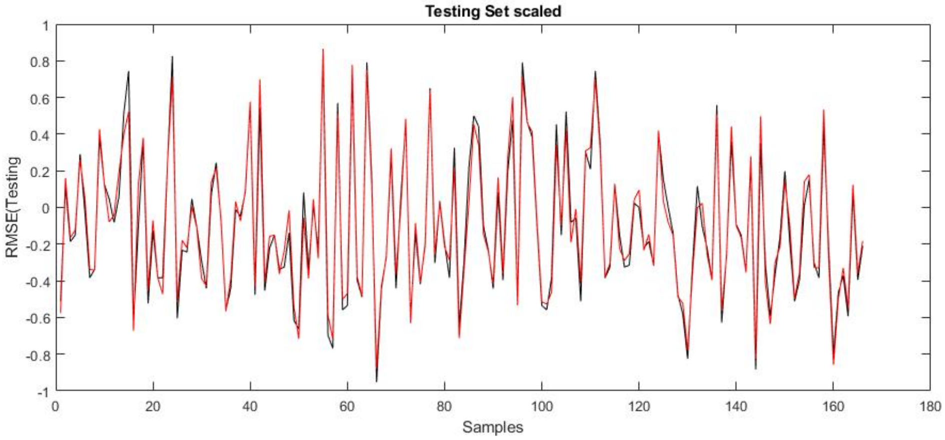
Table 4.5: MOGA 3 - Table with values referent to Figure 4.11

Model	31
y(k)=f(v1(k),v2(k),v3(k),v6(k),)	
RMSE Training scaled	0.053
RMSE Testing scaled	0.068
RMSE Validation scaled	0.064

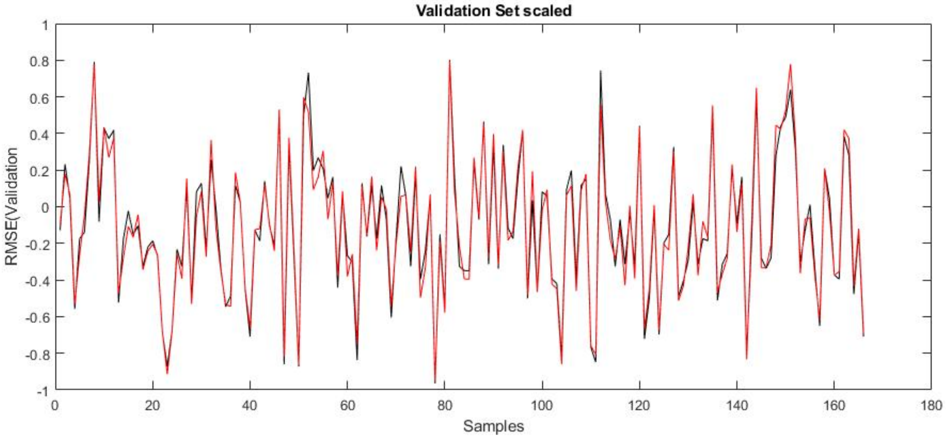
Table 4.6: MOGA 3 - Preferred Model

For the second run the selected RMSE training goal was: 0.048448.

As for the model we chosen we see that it has less features present and the value of RMSE for each matrix of data tests are low indicating a good model. Figure 4.12 also show us that the prediction of the model is within the range of the expected values.



(a)

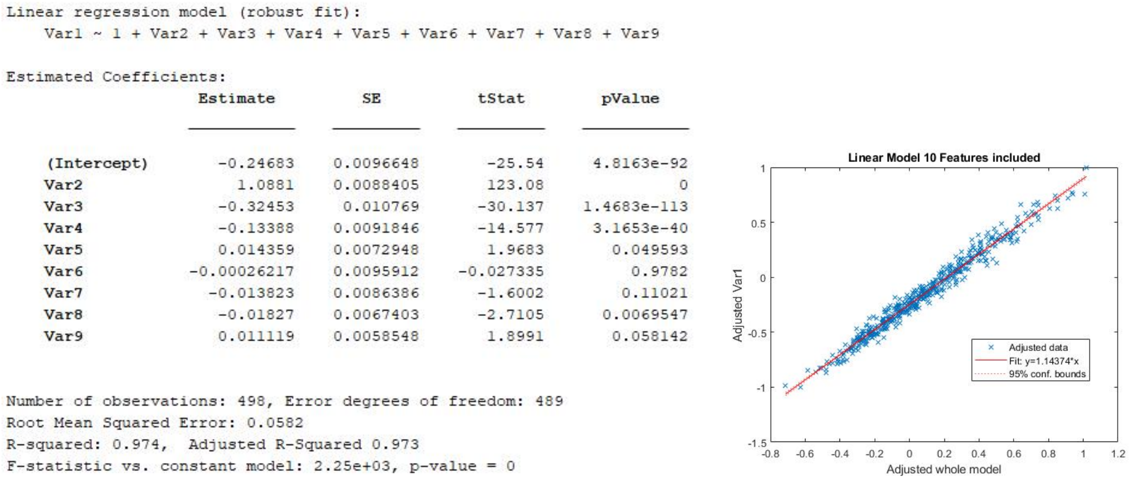


(b)

Figure 4.12: Predicted values (red) and Original Data (blue) of Model 31 during Testing (a) and Validation (b) procedures

4.3.2 Linear Model

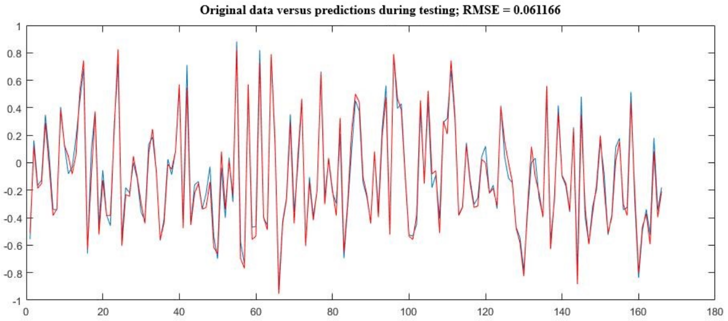
For the linear model to compare we use the following labels: Var1 -LDL, Var2 - Total Cholesterol, Var3 - HDL, Var4 - Triglyceride, Var5 - HbA1c, Var6 - GGT, Var7 - Platelet, Var8 - MAU, Var9 - Creatinine.



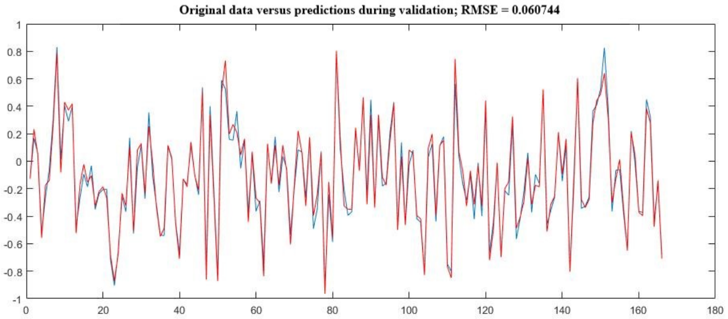
```
Linear regression model (robust fit):
    Var1 ~ 1 + Var2 + Var3 + Var4 + Var5 + Var6 + Var7 + Var8 + Var9

Estimated Coefficients:
                    Estimate        SE         tStat        pValue
                  __________    _________    _________    __________

    (Intercept)     -0.24683    0.0096648       -25.54     4.8163e-92
    Var2              1.0881    0.0088405       123.08              0
    Var3            -0.32453     0.010769      -30.137    1.4683e-113
    Var4            -0.13388    0.0091846      -14.577     3.1653e-40
    Var5            0.014359    0.0072948       1.9683       0.049593
    Var6         -0.00026217    0.0095912    -0.027335         0.9782
    Var7           -0.013823    0.0086386      -1.6002        0.11021
    Var8            -0.01827    0.0067403      -2.7105      0.0069547
    Var9            0.011119    0.0058548       1.8991       0.058142


Number of observations: 498, Error degrees of freedom: 489
Root Mean Squared Error: 0.0582
R-squared: 0.974,  Adjusted R-Squared 0.973
F-statistic vs. constant model: 2.25e+03, p-value = 0
```

Figure 4.13: MOGA 3 - Linear model using Training matrix used for MOGA

From the Figure 4.13 we see that the lowest influence on the model is presented by Var6 - GGT.



(a)



(b)

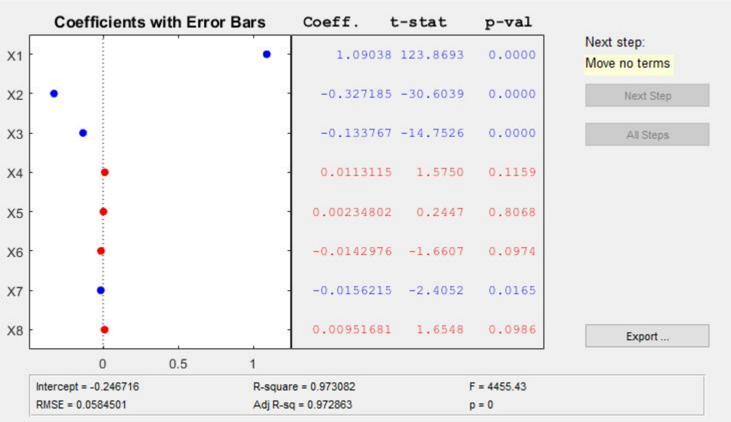
Figure 4.14: Predicted values (red) and Original Data (blue) of the Linear Model using the MOGA 3 Testing(a) and Validation(b) matrices

The model presents a very good RMSE value and good predictive results as seen in Figure 4.14.

Comparing MOGA 3 and the linear model we see that they obtain equivalent results, having both similar values of RMSE and good prediction values.

4.3.3 Stepwise

Once more the weight of each variable on the linear model is assessed using Stepwise algorithm. The correspondent labels of features are: X1 - Total Cholesterol, X2 - HDL, X3 - Triglyceride, X4 - HbA1c, X5 - GGT, X6 - Platelet, X7 - MAU, X8 - Creatinine.



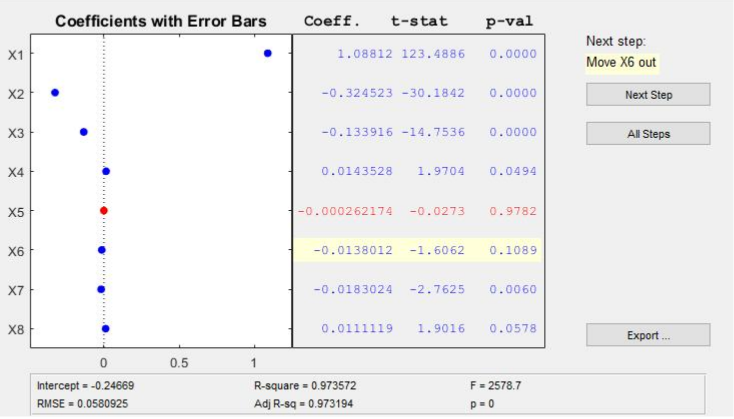
	Coeff.	t-stat	p-val
X1	1.09038	123.8693	0.0000
X2	-0.327185	-30.6039	0.0000
X3	-0.133767	-14.7526	0.0000
X4	0.0113115	1.5750	0.1159
X5	0.00234802	0.2447	0.8068
X6	-0.0142976	-1.6607	0.0974
X7	-0.0156215	-2.4052	0.0165
X8	0.00951681	1.6548	0.0986

Next step:
Move no terms

Intercept = -0.246716
RMSE = 0.0584501
R-square = 0.973082
Adj R-sq = 0.972863
F = 4455.43
p = 0

Figure 4.15: Stepwise panel obtained when MOGA 3 training matrix is employed

From Figure 4.15 we see that the function recommends discarding more variables than we previously did. We can see that it does not take into consideration HbA1c, GGT, Platelet and Creatinine. But looking at the values of RMSE and R squared, we observe that they are worse than in the MOGA 3 model, which can indicate us that our linear model could have better results.



Coefficients with Error Bars

	Coeff.	t-stat	p-val
X1	1.08812	123.4886	0.0000
X2	-0.324523	-30.1842	0.0000
X3	-0.133916	-14.7536	0.0000
X4	0.0143528	1.9704	0.0494
X5	-0.000262174	-0.0273	0.9782
X6	-0.0138012	-1.6062	0.1089
X7	-0.0183024	-2.7625	0.0060
X8	0.0111119	1.9016	0.0578

Next step:
Move X6 out

Intercept = -0.24669
RMSE = 0.0580925
R-square = 0.973572
Adj R-sq = 0.973194
F = 2578.7
p = 0

Figure 4.16: Stepwise panel obtained when MOGA 3 training matrix is employed with manually selecting the parameters used

In Figure 4.16 we see that manually adding some of the parameters discarded in the first run (4.15) lead to a slighter better RMSE value and R squared.

4.4 MOGA 4

As in the previous models we used only the features we considered most important due to their relevance on the identification and management of diseases, we now decided to introduce some additional features originated from the type of patient such as age, sex and one more clinical calculation the MDRD.

Features and correspondent labels are as follows: Var0 -LDL, Var1 - Total Cholesterol, Var2 - HDL, Var3 - Triglyceride, Var4 - HbA1c, Var5 - GGT, Var6 - Platelets,Var7 - MAU, Var8 - Creatinine, Var9 - MDRD, Var10 - Sex, Var11 - Age.

With all this features we just obtained a total of 830 MA divided in three matrices as follows:

- Training - 498 Observations
- Testing - 166 Observations
- Validation - 166 Observations.

4.4.1 Results

From Figure 4.17 we observe that now sex in the second run improved it's presence, Var5 - GGT still maintains its low occurrence in the models as it did previously. Although with the threshold we used we obtained a low number of models we can still conclude that GGT has a very low influence on LDL, not only based on this models but also taking into consideration the results we previously obtained.

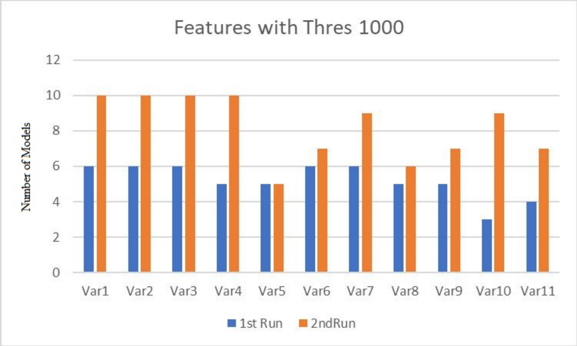


Figure 4.17: MOGA 4 -Histograms of number of models using each variable

Features	1stRun Thres:1000	2ndRun Thres:1000
Var1	6	10
Var2	6	10
Var3	6	10
Var4	5	10
Var5	5	5
Var6	6	7
Var7	6	9
Var8	5	6
Var9	5	7
Var10	3	9
Var11	4	7

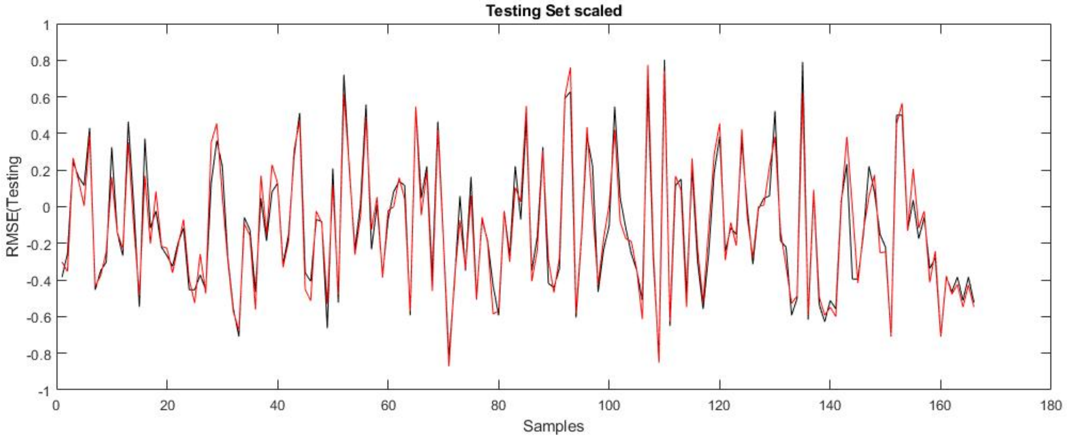
Table 4.7: MOGA 4 - Table with values referent to Figure 4.17

Model	2144
y(k)=f(v1(k),v2(k),v3(k),v4(k),v5(k),v6(k),)	
RMSE Training scaled	0.034
RMSE Testing scaled	0.084
RMSE Validation scaled	0.133

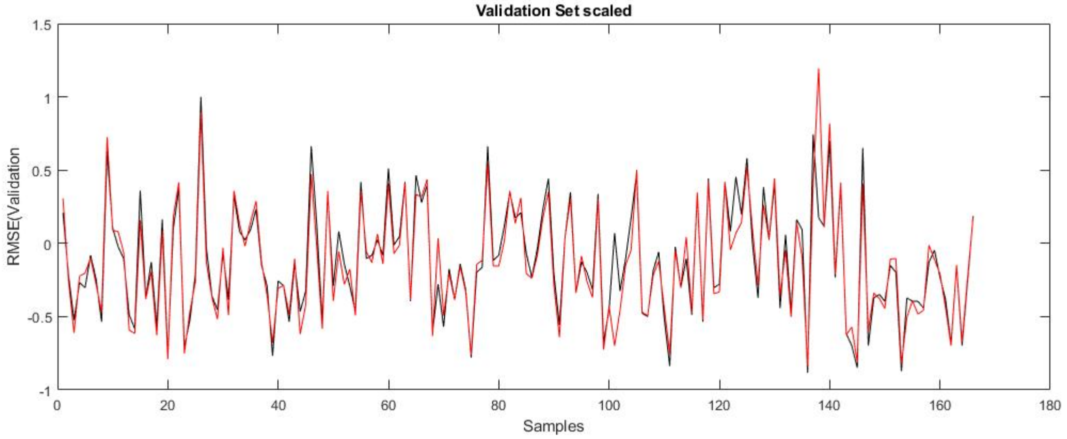
Table 4.8: MOGA 4 - Preferred Model

For the second run the selected RMSE training goal was: 0.040444.

The model we selected has the first 6 variables used, we chose it based on the RMSE training values and as we can see the results using Testing and Validation got worse than previous MOGA models.



(a)

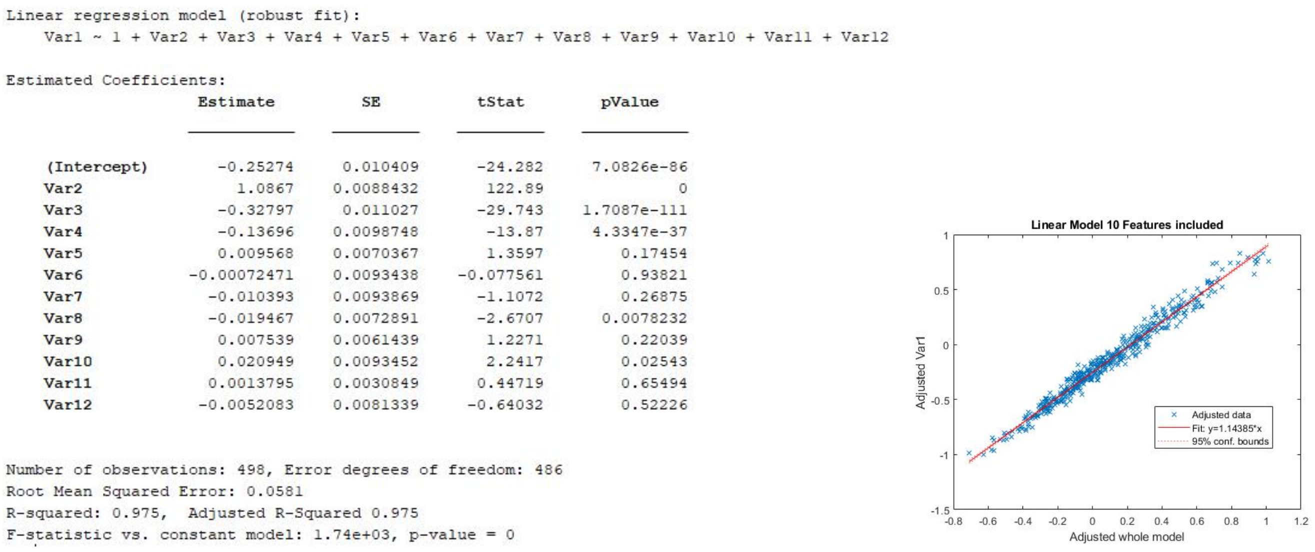


(b)

Figure 4.18: Predicted values (red) and Original Data (blue) of Model 2144 during Testing (a) and Validation (b) procedures

4.4.2 Linear Model

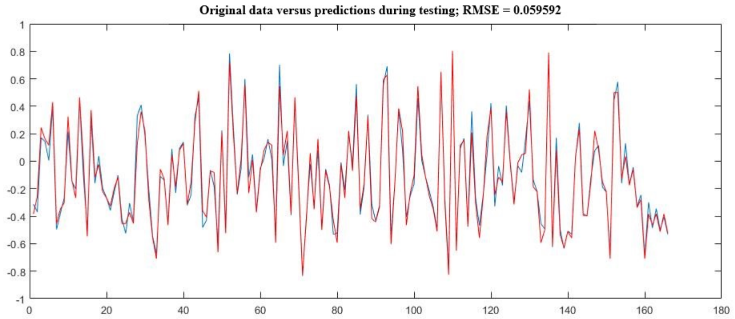
The labels we used in this linear model are as follows: Var1 -LDL, Var2 - Total Cholesterol, Var3 - HDL, Var4 - Triglyceride, Var5 - HbA1c, Var6 - GGT, Var7 - Platelets,Var8 - MAU, Var9 - Creatinine, Var10 - MDRD, Var11 - Sex, Var12 - Age



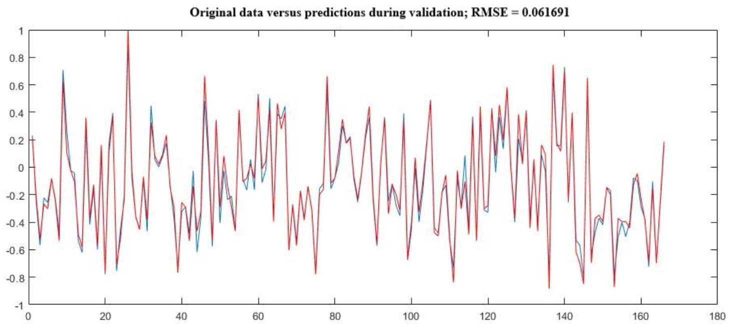
```
Linear regression model (robust fit):
    Var1 ~ 1 + Var2 + Var3 + Var4 + Var5 + Var6 + Var7 + Var8 + Var9 + Var10 + Var11 + Var12

Estimated Coefficients:
                    Estimate         SE         tStat        pValue
                   __________    _________    _________    __________

    (Intercept)      -0.25274     0.010409      -24.282    7.0826e-86
    Var2               1.0867    0.0088432       122.89             0
    Var3             -0.32797     0.011027      -29.743   1.7087e-111
    Var4             -0.13696    0.0098748       -13.87    4.3347e-37
    Var5             0.009568    0.0070367       1.3597       0.17454
    Var6          -0.00072471    0.0093438    -0.077561       0.93821
    Var7            -0.010393    0.0093869      -1.1072       0.26875
    Var8            -0.019467    0.0072891      -2.6707     0.0078232
    Var9             0.007539    0.0061439       1.2271       0.22039
    Var10            0.020949    0.0093452       2.2417       0.02543
    Var11           0.0013795    0.0030849      0.44719       0.65494
    Var12          -0.0052083    0.0081339     -0.64032       0.52226


Number of observations: 498, Error degrees of freedom: 486
Root Mean Squared Error: 0.0581
R-squared: 0.975,  Adjusted R-Squared 0.975
F-statistic vs. constant model: 1.74e+03, p-value = 0
```

Figure 4.19: Linear model using the MOGA 4Training matrix

Taking a closer look at Figure 4.19 we can see that this model presents very low values for Var6, Var11 and Var12 indicating that those features will have little impact in the prediction of our expected values.



(a)



(b)

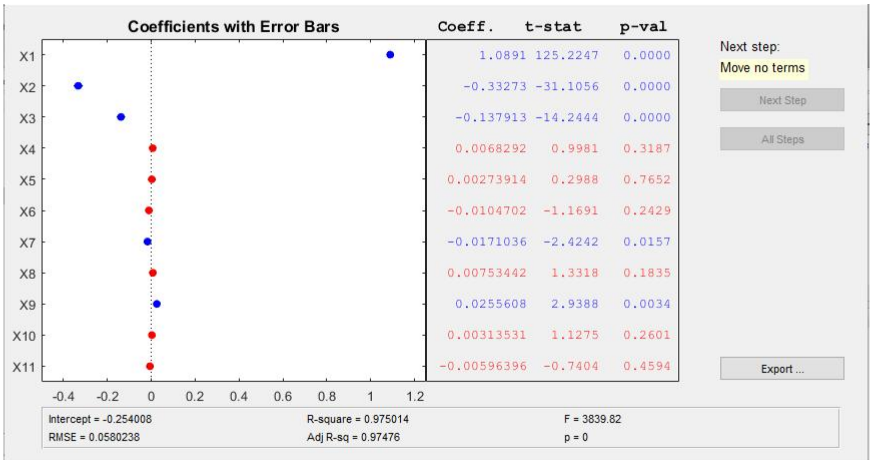
Figure 4.20: Predicted values (red) and Original Data (blue) of the Linear Model using the MOGA 4 Testing (a) and Validation (b) matrices

Figure 4.20 shows the result the model offers for testing and validation sets, and we can see small discrepancies in the predicted values (in red) and the original data (in blue) in red.

Comparing MOGA 4 and this linear model we see that both models are consistent in discarding the same variables.

4.4.3 Stepwise

To interpret 4.21 and 4.22, the following labels should be considered: X1 - Total Cholesterol, X2 - HDL, X3 - Triglyceride, X4 - HbA1c, X5 - GGT, X6 - Platelets, X7 - MAU, X8 - Creatinine, X9 - MDRD, X10 - Sex, X11 - Age.

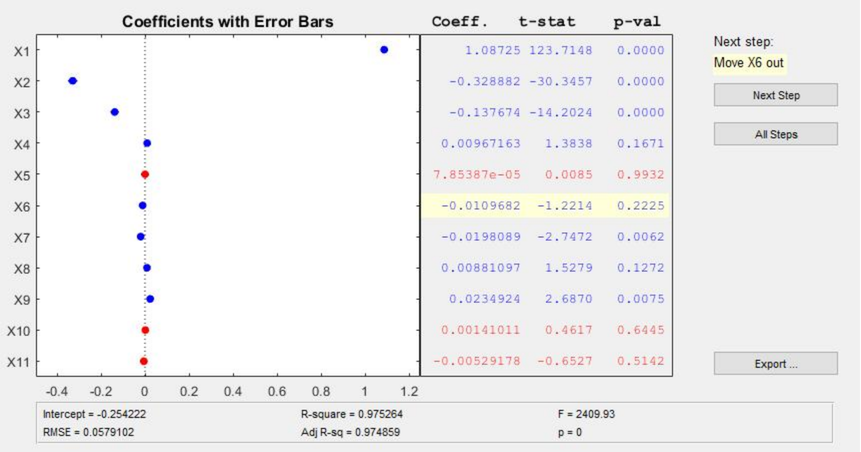


	Coeff.	t-stat	p-val
X1	1.0891	125.2247	0.0000
X2	-0.33273	-31.1056	0.0000
X3	-0.137913	-14.2444	0.0000
X4	0.0068292	0.9981	0.3187
X5	0.00273914	0.2988	0.7652
X6	-0.0104702	-1.1691	0.2429
X7	-0.0171036	-2.4242	0.0157
X8	0.00753442	1.3318	0.1835
X9	0.0255608	2.9388	0.0034
X10	0.00313531	1.1275	0.2601
X11	-0.00596396	-0.7404	0.4594

Intercept = -0.254008 R-square = 0.975014 F = 3839.82
RMSE = 0.0580238 Adj R-sq = 0.97476 p = 0

Figure 4.21: Stepwise panel obtained when MOGA 4 training matrix is employed

As observed in Figure 4.21 the function omits a few variables from the model, but after we add the parameters we consider relevant and only omit GGT, Sex and Age, we obtain a better model with lower RMSE and R-Square value a little higher as can be seen in Figure 4.22.



	Coeff.	t-stat	p-val
X1	1.08725	123.7148	0.0000
X2	-0.328882	-30.3457	0.0000
X3	-0.137674	-14.2024	0.0000
X4	0.00967163	1.3838	0.1671
X5	7.85387e-05	0.0085	0.9932
X6	-0.0109682	-1.2214	0.2225
X7	-0.0198089	-2.7472	0.0062
X8	0.00881097	1.5279	0.1272
X9	0.0234924	2.6870	0.0075
X10	0.00141011	0.4617	0.6445
X11	-0.00529178	-0.6527	0.5142

Intercept = -0.254222 R-square = 0.975264 F = 2409.93
RMSE = 0.0579102 Adj R-sq = 0.974859 p = 0

Figure 4.22: Stepwise panel with added features

4.5 MOGA 5

As we saw before GGT and Platelet have shown little influence on the models, so we did a new model without them and keeping MDRD, sex and age. Labels of the features are: Var0 -LDL, Var1 - Total Cholesterol, Var2 - HDL, Var3 - Triglyceride, Var4 - HbA1c, Var5 - MAU, Var6 - Creatinine, Var7 - MDRD, Var8 - Sex, Var9 - Age.

With all these features we obtained a total of 1409 MA divided in three matrices:

- Training - 845 Observations
- Testing - 281 Observations
- Validation - 283 Observations.

4.5.1 Results

From Figure 4.23 we see that the obtained results are somehow already the expected ones, Var5 - MAU and Var6 - Creatinine are showing the same behavior as in the previous models. In the second run of MOGA we see that Var6 diminished more than Var5 that keeps its level at four models, but in the second run we have less models within the threshold so we can say that the influence augmented a little.

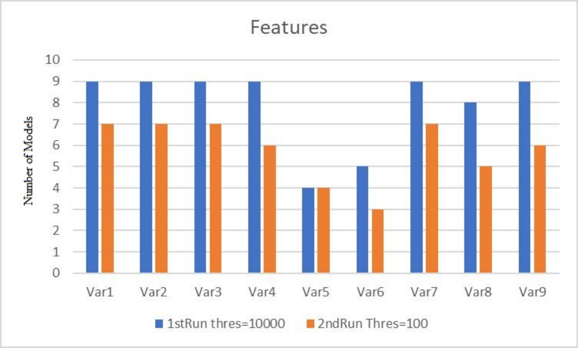


Figure 4.23: MOGA 5 -Histograms of number of models using each variable

Features	1stRun Thres:10000	2ndRun Thres:100
Var1	9	7
Var2	9	7
Var3	9	7
Var4	9	6
Var5	4	4
Var6	5	3
Var7	9	7
Var8	8	5
Var9	9	6

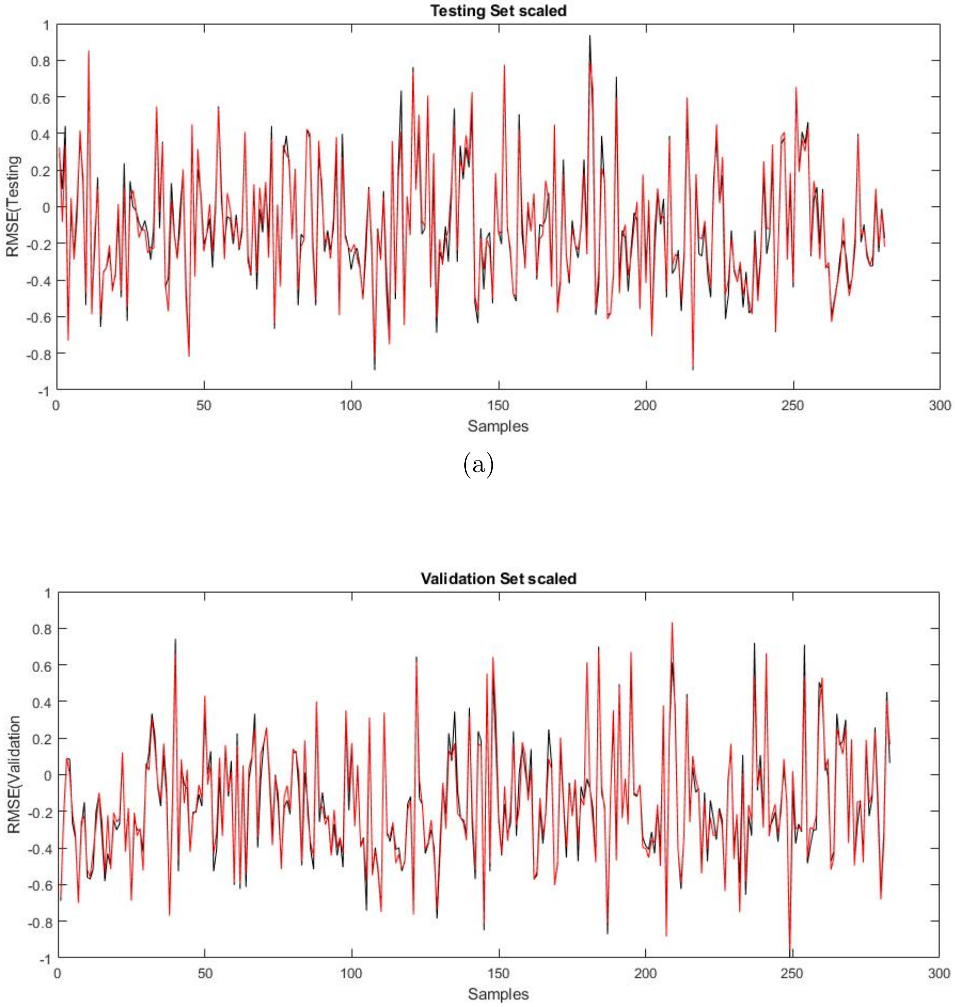
Table 4.9: MOGA 5 - Table with values referent to Figure 4.23

Model	48
y(k)=f(v1(k),v2(k),v3(k),v4(k),v5(k),v6(k),v7(k),v8(k),)	
RMSE Training scaled	0.046
RMSE Testing scaled	0.066
RMSE Validation scaled	0.074

Table 4.10: MOGA 5 - Preferred Model

For the second run the selected RMSE training goal was: 0.047118.

The chosen model has its results presented in Table 4.10. The prediction over testing and validation assume good results. This model uses almost all the features presented in the database and the evolution of the RMSE value using testing and validation matrices is good, showing us that the model respects the expected values.



(a)

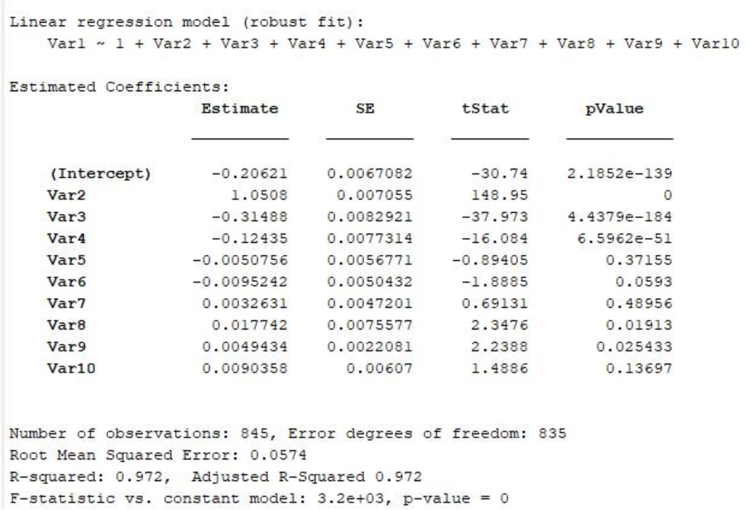
(b)

Figure 4.24: Predicted values (red) and Original Data (blue) of Model 48 during Testing (a) and Validation (b) procedures

In Figure 4.24 we see that the predicted values during testing and validation follow closely the original ones, giving us a good example that the model has a good fitting.

4.5.2 Linear Model

For this linear model we use the next labels: Var1 -LDL, Var2 - Total Cholesterol, Var3 - HDL, Var4 - Triglyceride, Var5 - HbA1c, Var6 - MAU, Var7 - Creatinine, Var8 - MDRD, Var9 - Sex, Var10 - Age.



```
Linear regression model (robust fit):
    Var1 ~ 1 + Var2 + Var3 + Var4 + Var5 + Var6 + Var7 + Var8 + Var9 + Var10

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                   __________    _________    ________    __________

    (Intercept)      -0.20621    0.0067082      -30.74    2.1852e-139
    Var2               1.0508     0.007055      148.95              0
    Var3             -0.31488    0.0082921     -37.973    4.4379e-184
    Var4             -0.12435    0.0077314     -16.084     6.5962e-51
    Var5           -0.0050756    0.0056771    -0.89405        0.37155
    Var6           -0.0095242    0.0050432     -1.8885         0.0593
    Var7            0.0032631    0.0047201     0.69131        0.48956
    Var8             0.017742    0.0075577      2.3476        0.01913
    Var9            0.0049434    0.0022081      2.2388       0.025433
    Var10           0.0090358      0.00607      1.4886        0.13697


Number of observations: 845, Error degrees of freedom: 835
Root Mean Squared Error: 0.0574
R-squared: 0.972,  Adjusted R-Squared 0.972
F-statistic vs. constant model: 3.2e+03, p-value = 0
```

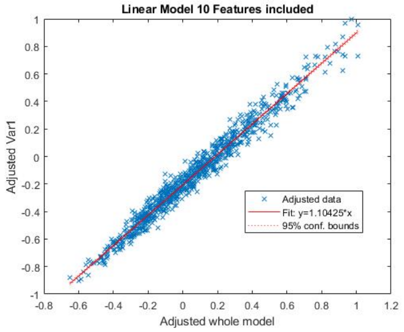
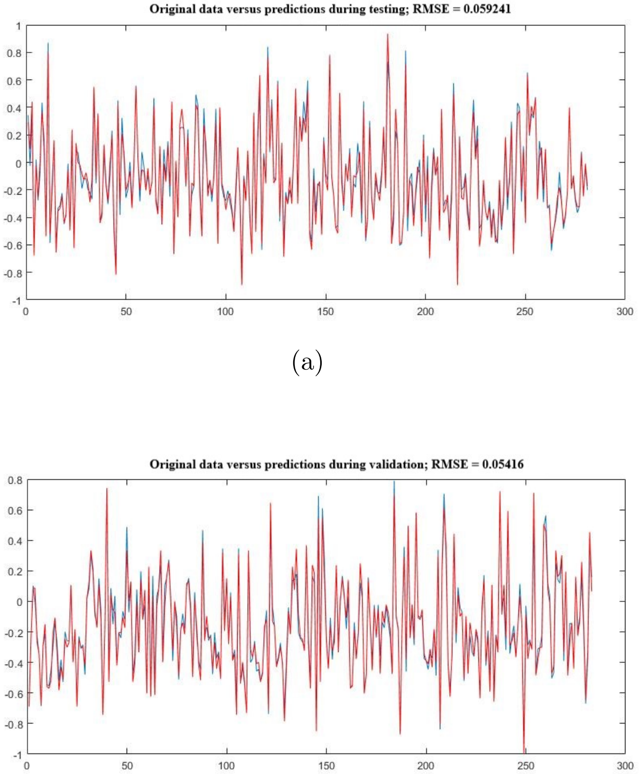


Figure 4.25: Linear model using the MOGA 5 Training matrix

Figure 4.25 shows us the weight of each parameter in obtaining the values of LDL using this model. We observe that Var5, Var6, Var7, Var9 and var10 have little impact in the model compared with the other ones.



(a)

(b)

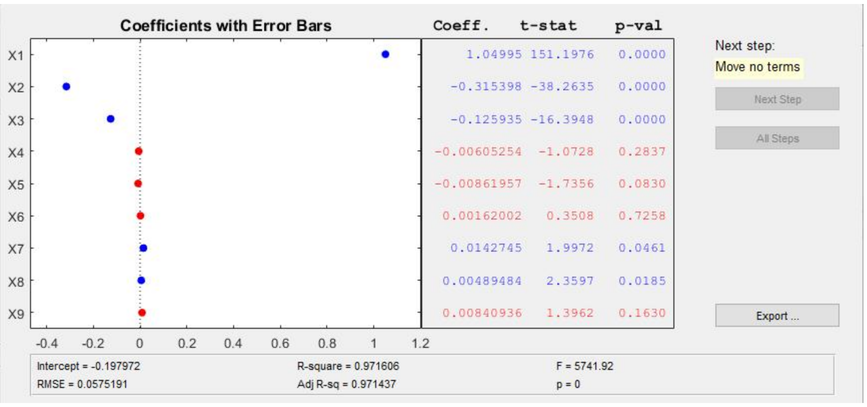
Figure 4.26: Predicted values (red) and Original Data (blue) of the Linear Model using MOGA 5 Testing (a) and Validation (b) matrices

Figure 4.26 shows us how the linear model predicts the values of LDL and we can see that it behaves well. As we can see the blue line which refers to the prediction values follows closely the original line in red.

Comparing the two models we see that both predict values of LDL really well. In terms of features we see that both consider the same variables(HbA1c, MAU, Creatinine, Sex and Age) having small influence on the models.

4.5.3 Stepwise

The correspondent labels for Stepwise algorithm are: X1 - Total Cholesterol, X2 - HDL, X3 - Triglyceride, X4 - HbA1c, X5 - MAU, X6 - Creatinine, X7 - MDRD, X8 - Sex, X9 - Age.

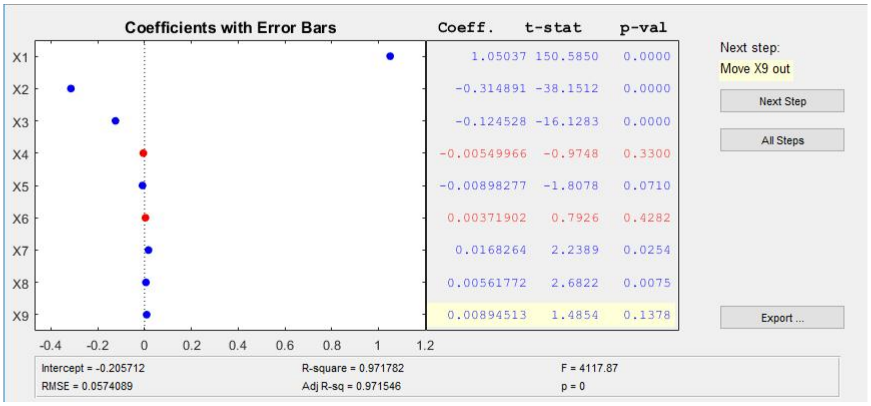


	Coeff.	t-stat	p-val
X1	1.04995	151.1976	0.0000
X2	-0.315398	-38.2635	0.0000
X3	-0.125935	-16.3948	0.0000
X4	-0.00605254	-1.0728	0.2837
X5	-0.00861957	-1.7356	0.0830
X6	0.00162002	0.3508	0.7258
X7	0.0142745	1.9972	0.0461
X8	0.00489484	2.3597	0.0185
X9	0.00840936	1.3962	0.1630

Next step:
Move no terms

Intercept = -0.197972
RMSE = 0.0575191
R-square = 0.971606
Adj R-sq = 0.971437
F = 5741.92
p = 0

Figure 4.27: Stepwise panel obtained when MOGA 5 training matrix is employed



	Coeff.	t-stat	p-val
X1	1.05037	150.5850	0.0000
X2	-0.314891	-38.1512	0.0000
X3	-0.124528	-16.1283	0.0000
X4	-0.00549966	-0.9748	0.3300
X5	-0.00898277	-1.8078	0.0710
X6	0.00371902	0.7926	0.4282
X7	0.0168264	2.2389	0.0254
X8	0.00561772	2.6822	0.0075
X9	0.00894513	1.4854	0.1378

Next step:
Move X9 out

Intercept = -0.205712
RMSE = 0.0574089
R-square = 0.971782
Adj R-sq = 0.971546
F = 4117.87
p = 0

Figure 4.28: MOGA 5 - Stepwise panel with added features

From Figure 4.27 and Figure 4.28 we can see that at first the tool excludes several variables but after we included some others (X5 - MAU and X9 - Age) the model presents better results. After the inclusion we conclude that X4- HbA1c and X6 - Creatinine influence the least our predictions.

4.6 MOGA 6

As a last model we wanted to see how LDL is influenced if we added the body mass index (BMI) in the features. BMI have a very close relationship with diabetes as is related to the weight of the patient, and as we know, the higher the weight implies higher BMI and simultaneously higher risk of DM.

Labels of the variables as follows: Var0 -LDL, Var1 - Total Cholesterol, Var2 - HDL, Var3 - Triglyceride, Var4 - HbA1c, Var5 - MAU, Var6 - Creatinine, Var7 - MDRD, Var8 - Sex, Var9 - Age, Var10 - BMI.

With all these features we obtained a total of 1113 MA. The total of the observations was divided in three matrices as follows:

- Training - 667 Observations
- Testing - 222 Observations
- Validation - 224 Observations.

4.6.1 Results

From Figure 4.29 we can see that almost all of the features have a good influence. The only ones that are less influent are Var5, Var6 and Var8 (MAU, Creatinine and Sex, respectively).

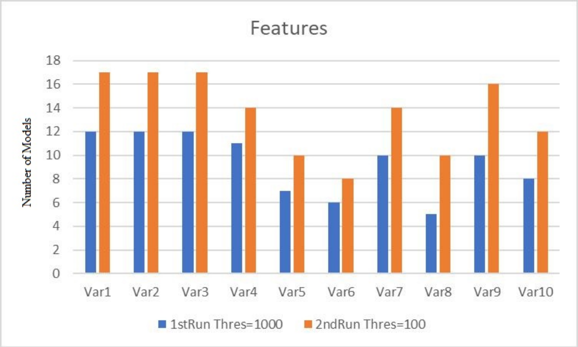


Figure 4.29: MOGA 6 - Histograms of number of models using each variable

Features	1stRun Thres:1000	2ndRun Thres:100
Var1	12	17
Var2	12	17
Var3	12	17
Var4	11	14
Var5	7	10
Var6	6	8
Var7	10	14
Var8	5	10
Var9	10	16
var10	8	12

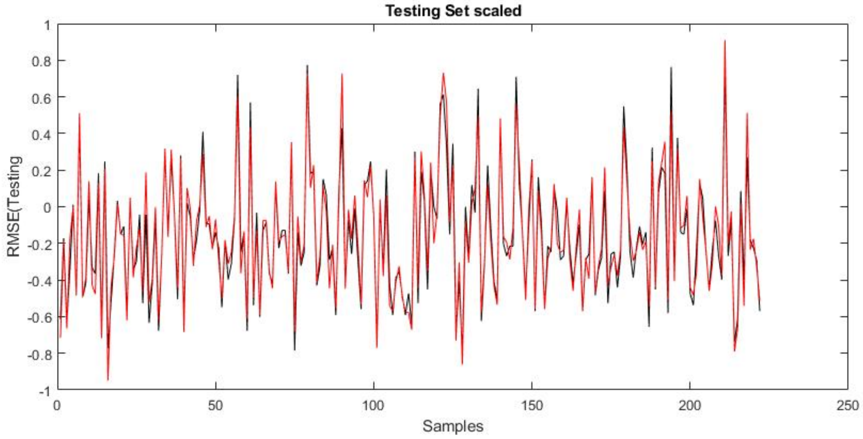
Table 4.11: MOGA 6 - Table with values referent to Figure 4.29

Model	2441
y(k)=f(v1(k),v2(k),v3(k),v4(k),v8(k),v9(k),)	
RMSE Training scaled	0.042
RMSE Testing scaled	0.079
RMSE Validation scaled	0.099

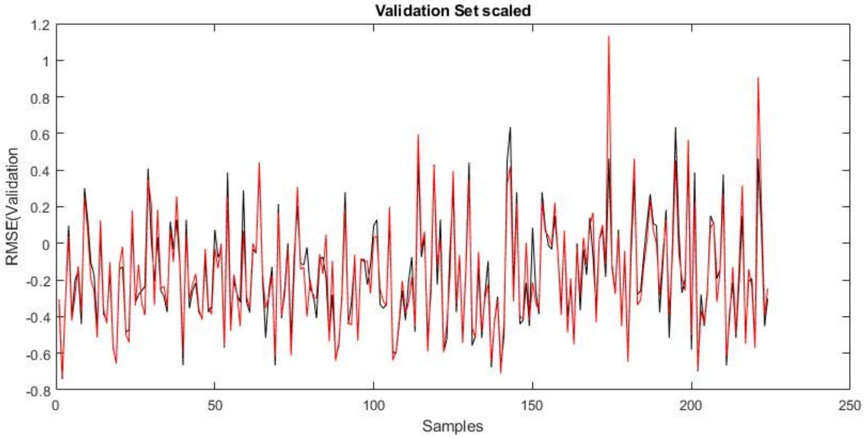
Table 4.12: MOGA 6 - Preferred Model

For the second run the selected RMSE training goal was: 0.045184.

The model obtained (shown in Table 4.12) uses the most influential features as shown before and the RMSE values are very good.



(a)



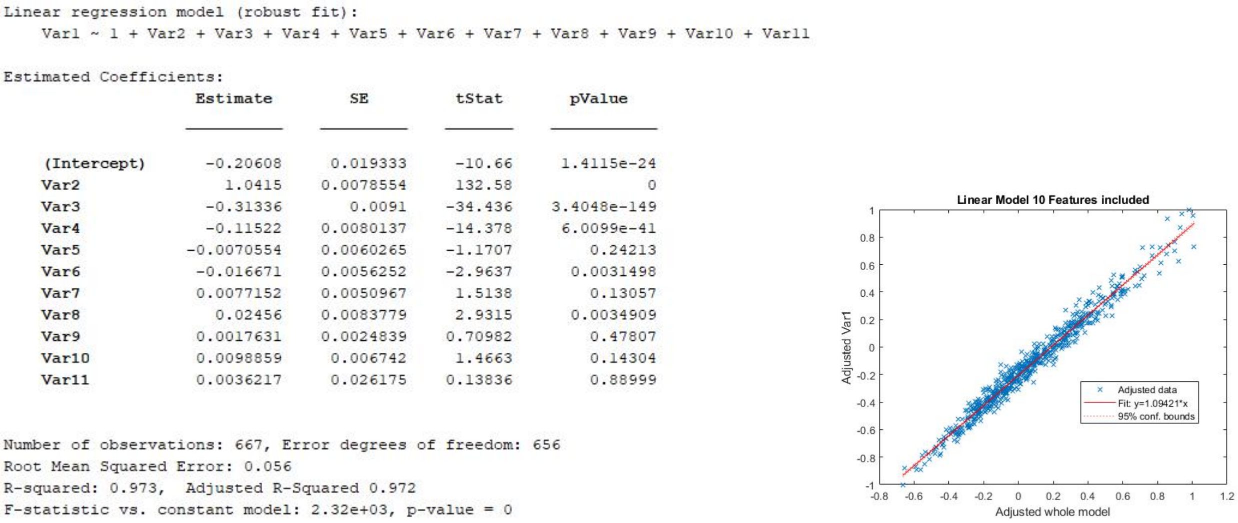
(b)

Figure 4.30: Predicted values (red) and Original Data (blue) of Model 2441 during Testing (a) and Validation (b) procedures

In Figure 4.30 we see the results of the model in testing and validation environment. The model behaves well as the curves follows each other closely, indicating very good levels of trust.

4.6.2 Linear Model

The labels used for the correspondent linear model are: Var1 -LDL, Var2 - Total Cholesterol, Var3 - HDL, Var4 - Triglyceride, Var5 - HbA1c, Var6 - MAU, Var7 - Creatinine, Var8 - MDRD, Var9 - Sex, Var10 - Age, Var11 - BMI.

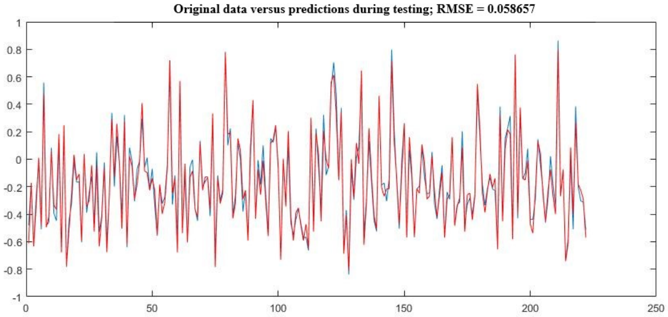


```
Linear regression model (robust fit):
    Var1 ~ 1 + Var2 + Var3 + Var4 + Var5 + Var6 + Var7 + Var8 + Var9 + Var10 + Var11

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                   __________    _________    _______    __________

    (Intercept)      -0.20608     0.019333     -10.66    1.4115e-24
    Var2               1.0415    0.0078554     132.58             0
    Var3             -0.31336       0.0091    -34.436    3.4048e-149
    Var4             -0.11522    0.0080137    -14.378    6.0099e-41
    Var5           -0.0070554    0.0060265    -1.1707       0.24213
    Var6            -0.016671    0.0056252    -2.9637     0.0031498
    Var7            0.0077152    0.0050967     1.5138       0.13057
    Var8              0.02456    0.0083779     2.9315     0.0034909
    Var9            0.0017631    0.0024839    0.70982       0.47807
    Var10           0.0098859     0.006742     1.4663       0.14304
    Var11           0.0036217     0.026175    0.13836       0.88999


Number of observations: 667, Error degrees of freedom: 656
Root Mean Squared Error: 0.056
R-squared: 0.973,  Adjusted R-Squared 0.972
F-statistic vs. constant model: 2.32e+03, p-value = 0
```

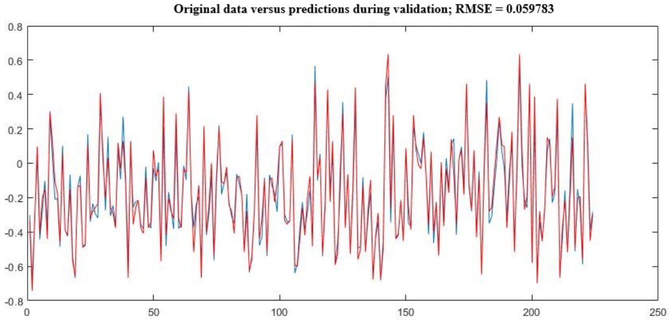
Figure 4.31: MOGA 6 - Linear model using Training matrix used for MOGA

As we see in the Figure 4.31 the linear model follows the same tendency as MOGA

6 showing good influence for almost all parameters, were almost all of them maintain the expected levels. The least influent factor is Var9 - Sex which should have a little more impact, but as we saw in the previous models it keeps influence low.



(a)



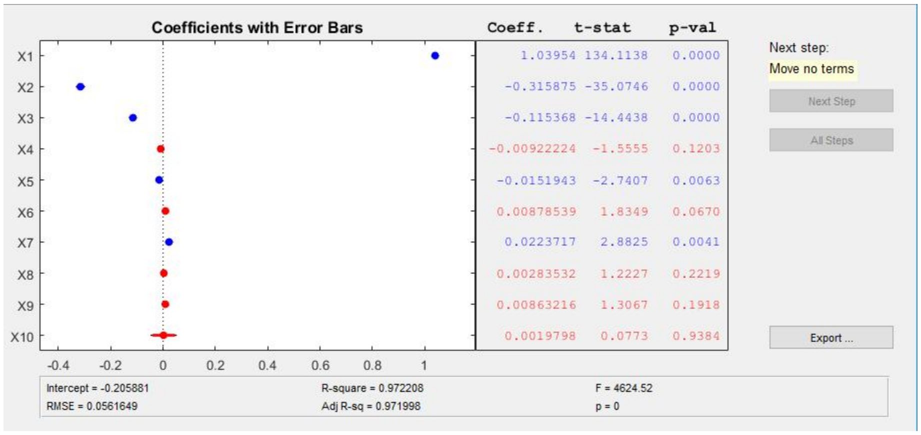
(b)

Figure 4.32: Predicted values (red) and Original Data (blue) of the Linear Model using the MOGA 6 Testing (a) and Validation (b) matrices

From Figure 4.32 we see that the model obtain relatively close values to the expected ones indicating very good behavior. The RMSE values for Training and Validation stay similar to the other models. .

4.6.3 Stepwise

Relabeling the variables as follows : X1 - Total Cholesterol, X2 - HDL, X3 - Triglyceride, X4 - HbA1c, X5 - MAU, X6 - Creatinine, X7 - MDRD, X8 - Sex, X9 - Age, X10 - BMI.

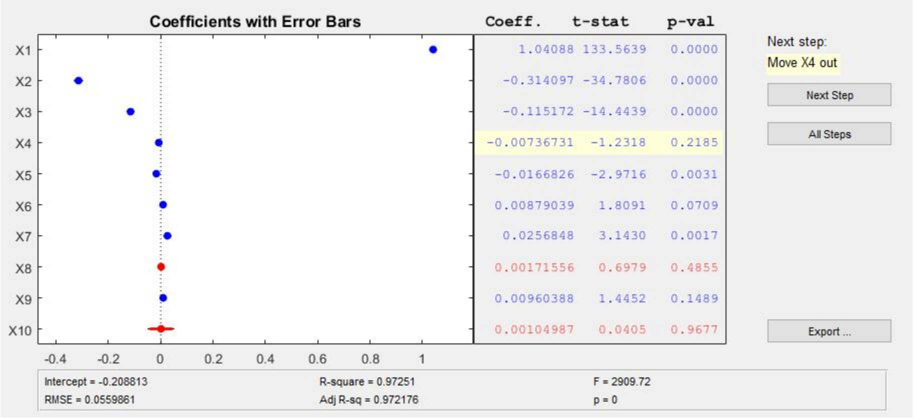


	Coeff.	t-stat	p-val
X1	1.03954	134.1138	0.0000
X2	-0.315875	-35.0746	0.0000
X3	-0.115368	-14.4438	0.0000
X4	-0.00922224	-1.5555	0.1203
X5	-0.0151943	-2.7407	0.0063
X6	0.00878539	1.8349	0.0670
X7	0.0223717	2.8825	0.0041
X8	0.00283532	1.2227	0.2219
X9	0.00863216	1.3067	0.1918
X10	0.0019798	0.0773	0.9384

Intercept = -0.205881 R-square = 0.972208 F = 4624.52
RMSE = 0.0561649 Adj R-sq = 0.971998 p = 0

Figure 4.33: Stepwise panel obtained when MOGA 6 training matrix is employed

The first obtained results using Stepwise are presented in Figure 4.33, where we see that there are a few variables excluded. After a few testing we managed to obtain a better result in terms of RMSE and R-square improving them as shown in Figure 4.34. The resultant model excludes X8 - Sex and X10-BMI as they do not improve the models performance values due to their small influence.



	Coeff.	t-stat	p-val
X1	1.04088	133.5639	0.0000
X2	-0.314097	-34.7806	0.0000
X3	-0.115172	-14.4439	0.0000
X4	-0.00736731	-1.2318	0.2185
X5	-0.0166826	-2.9716	0.0031
X6	0.00879039	1.8091	0.0709
X7	0.0256848	3.1430	0.0017
X8	0.00171556	0.6979	0.4855
X9	0.00960388	1.4452	0.1489
X10	0.00104987	0.0405	0.9677

Intercept = -0.208813 R-square = 0.97251 F = 2909.72
RMSE = 0.0559861 Adj R-sq = 0.972176 p = 0

Figure 4.34: Stepwise panels with added features

4.7 Conclusions

Model	Type of Error	MOGA 1	MOGA 2	MOGA 3	MOGA 4	MOGA 5	MOGA 6
MOGA	RMSE Training	0.048	0.044	0.053	0.034	0.046	0.042
	RMSE Testing	0.065	0.052	0.068	0.084	0.066	0.079
	RMSE Validation	0.064	0.057	0.064	0.133	0.074	0.099
Linear	RMSE Training	0.056	0.053	0.058	0.058	0.057	0.056
	RMSE Testing	0.059	0.054	0.061	0.059	0.059	0.059
	RMSE Validation	0.058	0.058	0.061	0.061	0.054	0.059
Stepwise	RMSE Training	0.056	0.053	0.058	0.058	0.0574	0.056

Table 4.13: Comparison of each RMSE values achieved by each model

As Table 4.13 shows we have some similar values for each model we trained. The best model using MOGA algorithm comparing the RMSE Validation values is obtained by MOGA 2. The best model using linear regression we obtained as the best model MOGA 5.

If we look at the Table 4.13 we see that the value of these two models are not very similar using the various approaches we have chosen (MOGA, linear regression or Stepwise). MOGA 5 only presents very good result in the linear regression, while MOGA 2 has the best values in MOGA and Stepwise. But, knowing that the linear regression model is much faster that an algorithm such as MOGA, we have to chose as the best MOGA 5.

5 Conclusions, remarks and future work

We did several models using the data at our disposable and we obtained satisfactory results, considering that we can conclude which of the parameters have the smallest and higher influence in the variations of LDL.

After the study of the different models we used we could say that the least influencing parameters on the LDL could be GGT and Platelet. Controlling these parameters will show little result in the objective of lowering the LDL values.

On the other side Total Cholesterol, HDL, Triglycerides, HbA1c, MDRD and BMI have promising results, each one of them show good presence and influence in the models, leading us to believe that if we can control them and keep them between the reference values we can achieve a lowering in the fluctuations of LDL leading us to a better control.

MAU and Creatinine showed us that they can have a little impact in the models and can have a role in the objective in controlling LDL.

Using all the models created we can say that one model that could interpret the relation of the LDL values with the other features should include obligatorially Total Cholesterol, HDL, Triglycerides, HbA1c, MDRD. As a second option we can include sex, age and BMI in order to introduce more information correspondent to the patients.

All these results are complemented by the stepwise results, as we saw from the tests for each linear model. In general, all the created models revealed the same features as being the lowest influencer factors.

Better controlling the parameters we choose as the ones with the greatest influence can lead to better control of the DM and the its subjacent comorbidities, in particular CVD.

Sadly we could not use medication and previous complication of the patients as input features for our models, given the small amount of observations we obtained from the database with the majority of features present. This fact leaded us to not manage to get a time line for the individuals so we could not tell how the medication would influence the control of LDL because we didn't have the next medical appointment present to see the evolution.

In subsection 3.3 the different RMSE and R-square values obtained, using the Linear Models presented, are shown in Table 3.5. From this Table we can easily conclude that the best model is Model 5 with an RMSE of 0.054 and an R-square of 0.97. Using this model the equation that better predicts LDL values is:

$LDL = 1 + 0.989(TotalCholesterol) - 0.301(HDL) - 0.121(Triglyceride) - 0.067(HbA1c) -$

$0.012(Proteinuria)$

But as previously said this model can't be used as there are so little observations, the next best model is Model 3 with an RMSE of 0.070 and an R-square of 0.950. The equation of the models is as follows:

$LDL = 1 + 0.991(TotalCholesterol) - 0.302(HDL) - 0.116(Triglyceride) - 0.005(HbA1c) - 0.037(Platelet)$

The results obtained with the selected population using the MOGA algorithm are presented in Table 4.13.

In order to compare all of the used methods firstly we have to compare RMSE Training. Looking at the Table 4.13 we see that for the Linear Models and Stepwise the best RMSE value (0.053) is obtained in MOGA 2. On the other side MOGA obtained the best RMSE value (0.034) using the population from MOGA 5. Comparing the three values we see that MOGA obtained far better results.

As we know in order to select the better model we have to compare the RMSE Validation values. This values were obtained only using MOGA and Linear Models. Table 4.13 shows us that using MOGA the best results for RMSE Validation (0.057) were obtained using population from MOGA 2. For the Linear Model the RMSE Validation (0.054) is the lowest using MOGA 5 and is better that the one obtained with the MOGA algorithm.

Using the best model to predict the LDL values we have the next formula:

$LDL = 1 + 1.05(TotalCholesterol) - 0.314(HDL) - 0.124(Triglyceride) - 0.005(HbA1c) - 0.009(MAU) + 0.003(Creatinine) + 0.017(MDRD) + 0.005(Sex) + 0.009(Age)$

From the Table 4.13 we see that the obtained values of each model are very similar between them leading us to conclude that the results obtained using linear models and prediction show us better results or similar results to the ones from the MOGA. As we know the neural network models take more time to be created and the linear ones are faster to obtain, and in our case as the result are similar we should think of using the linear ones.

One of the facts that lead to those results may be due to the small quantity of data we withdraw from the database provided. We could not increase the data selection as we initially hoped. As we previously showed, some of the values can be estimated using others and we thought that we could obtain some of them, but after we took a closer look we saw that when the factor to be estimated was missing also some of the elements needed for prediction also failed to be present.

As a future work I would suggest to keep a better track of the records introduced in the database, introducing all the values of the clinical parameters so when you look back in time you can extrapolate more easily all the parameters and use them to better predict the future.

Also in future the models should take into account the medication, which should be divided by classes. The complications should also be taken into account and introduced into the models. A more closely observations for each patient is required in order to better find the fluctuation in every clinical parameter.

References

[1] Pedro M Ferreira and António E Ruano. Evolutionary multiobjective neural network models identification: evolving task-optimised models. In *New Advances in Intelligent Signal Processing*, pages 21–53. Springer, 2011.

[2] O que é o GGT / Gamaglutamiltranspeptidase?, http://www.examedesangue.com/sanguineos/ggt/, Date Accessed: 29 Set. 2018.

[3] Kurt George Matthew Mayer Alberti and P Z ft Zimmet. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation. *Diabetic medicine*, 15(7):539–553, 1998.

[4] American Diabetes Association and Others. Diagnosis and classification of diabetes mellitus. *Diabetes care*, 37(Supplement 1):S81—-S90, 2014.

[5] Barbara V Howard, David C Robbins, Maurice L Sievers, Elisa T Lee, Dorothy Rhoades, Richard B Devereux, Linda D Cowan, R Stuart Gray, Thomas K Welty, Oscar T Go, and Others. LDL cholesterol as a strong predictor of coronary heart disease in diabetic individuals with insulin resistance and low LDL. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 20(3):830–835, 2000.

[6] James Shepherd, Philip Barter, Rafael Carmena, Prakash Deedwania, Jean-Charles Fruchart, Steven Haffner, Judith Hsia, Andrei Breazna, John LaRosa, Scott Grundy, and Others. Effect of lowering LDL cholesterol substantially below currently recommended levels in patients with coronary heart disease and diabetes. *Diabetes care*, 29(6):1220–1226, 2006.

[7] Randie R. Little and David B. Sacks. HbA1c: How do we measure it and what does it mean? *Current Opinion in Endocrinology, Diabetes and Obesity*, 16(2):113–118, 2009.

[8] Lorenz RA Goldstein DE, Little RR. Tests of glycemia in diabetes. *Diabetes Care*, 27:1761–1773, 2004.

[9] O que é Hemoglobina Glicada e valores de referência do exame, https://minutosaudavel.com.br/o-que-e-hemoglobina-glicada-e-valores-de-referencia-do-exame/, Date Accessed: 29 Set. 2018.

[10] John D Brunzell, William R Hazzard, Arno G Motulsky, and Edwin L Bierman. Evidence for diabetes mellitus and genetic forms of hypertriglyceridemia as independent entities. *Metabolism - Clinical and Experimental*, 24(10):1115–1121, oct 1975.

[11] W. V. Brown. Lipoprotein disorders in diabetes mellitus. *Medical Clinics of North America*, 78(1):143–161, 1994.

[12] Robert C Biesbroeck, John J Albers, Patricia W Wahl, Clarice R Weinberg, Martin L Bassett, and Edwin L Bierman. Abnormal Composition of High Density Lipoproteins in Non-insulin-dependent Diabetics. *Diabetes*, 31(2):126 LP – 131, feb 1982.

[13] Drª. Ana Luiza Lima. Quais os tipos de Colesterol, https://www.tuasaude.com/colesterol/, Date Accessed: 29 Set. 2018, 2018.

[14] Anoop Shankar and Jialiang Li. Association between serum gamma-glutamyltransferase level and prehypertension among US adults. *Circulation Journal*, 71(10):1567–1572, 2007.

[15] L Sacchetti, G Castaldo, G Fortunato, and F Salvatore. Improved procedure for measuring gamma-glutamyltransferase isoenzymes in serum. *Clinical chemistry*, 34(2):419–422, 1988.

[16] Jennifer E. Mason, Rodman D. Starke, and John E. Van Kirk. Gamma-glutamyl transferase: A novel cardiovascular risk biomarker. *Preventive Cardiology*, 13(1):36–41, 2010.

[17] P V Halushka, R Curtis Rogers, C Body Loadholt, and J A Colwell. Increased platelet thromboxane synthesis in diabetes mellitus. *The Journal of laboratory and clinical medicine*, 97(1):87–96, 1981.

[18] Graziella Bruno, Franco Merletti, Annibale Biggeri, Giuseppe Bargero, Stefania Ferrero, Gianfranco Pagano, and Paolo Cavallo Perin. Progression to overt nephropathy in type 2 diabetes: the Casale Monferrato Study. *Diabetes care*, 26(7):2150–2155, 2003.

[19] JohnS Yudkin, RichardD Forrest, and CarolineA Jackson. Microalbuminuria as predictor of vascular disease in non-diabetic subjects: Islington Diabetes Survey. *The Lancet*, 332(8610):530–533, 1988.

[20] L Groop, A Ekstrand, C Forsblom, E Widen, P-H Groop, A-M Teppo, and J Eriksson. Insulin resistance, hypertension and microalbuminuria in patients with type 2 (non-insulin-dependent) diabetes mellitus. *Diabetologia*, 36(7):642–647, 1993.

[21] S J Schwab, R L Christensen, K Dougherty, and S Klahr. Quantitation of proteinuria by the use of protein-creatinine ratios in single urine samples. *Arch.Intern.Med.*, 147:943–944, 1987.

[22] Andrew S Levey, Richard L Berg, Jennifer L Gassman, Phillip M Hall, and W Gordon Walker. Creatinine filtration, secretion and excretion during progressive renal disease. *Kidney international Supplement*, (27), 1989.

[23] Ronald D Perrone, Nicolaos E Madias, and Andrew S Levey. Serum creatinine as an index of renal function: new insights into old concepts. *Clinical chemistry*, 38(10):1933–1953, 1992.

[24] D. Kundu, A. Roy, T. Mandal, U. Bandyopadhyay, E. Ghosh, and D. Ray. Relation of microalbuminuria to glycosylated hemoglobin and duration of type 2 diabetes. *Nigerian Journal of Clinical Practice*, 16(2):216–220, 2013.

[25] Dra. Ângela Cassol. Quais são os valores de referência de creatinina?, https://medicoresponde.com.br/quais-sao-os-valores-de-referencia-de-creatinina/, Date Accessed: 29 Set. 2018.

[26] Andrew S Levey, Josef Coresh, Tom Greene, Lesley A Stevens, Yaping Lucy Zhang, Stephen Hendriksen, John W Kusek, and Frederick Van Lente. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. *Annals of internal medicine*, 145(4):247–254, 2006.

[27] A S Levey, J P Bosch, J Breyer Lewis, N Rogers, and D Roth. A simplified equation to predict glomerular filtration rate from serum creatinine. 11:A0828, 2000.

[28] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[29] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Second edi edition, 2008.

[30] Ordinary Least Squares. Wikipedia, Wikimedia Foundation, 16 Aug. 2018, en.wikipedia.org/w/index.php?title=Ordinary_least_squares&oldid=855225783. Date Accessed: 04 Set. 2018.

[31] Fit linear regression model, MathWorks, https://www.mathworks.com/help/stats/fitlm.html, Date accessed: 20 Jul. 2019.

[32] Linear regression model, MathWorks, https://www.mathworks.com/help/stats/linearmodel.html, Date accessed: 20 Jul. 2019.

[33] Elmira Hajimani. *Intelligent Support System for Cva Diagnosis By Cerebral Computerized Tomography*. PhD thesis, UNIVERSIDADE DO ALGARVE, 2016.

[34] César Alexandre Domingues Teixeira. *Soft-computing techniques applied to artificial tissue temperature estimation*. PhD thesis, Universidade do Algarve, 2008.

[35] Ani1 K Jain and Jianchang Mao. Artificial Neural Network: A Tutorial. *Communications*, 29:31–44, 1996.

[36] H. Harkat, A. Ruano, M. G. Ruano, and S. D. Bennani. Classifier Design by a Multi-Objective Genetic Algorithm Approach for GPR Automatic Target Detection. *IFAC-PapersOnLine*, 51(10):187–192, 2018.

[37] Pedro M Ferreira and António E Ruano. Exploiting the separability of linear and non-linear parameters in radial basis function networks. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, pages 321–326. IEEE, 2000.