Heart Failure Prediction Using Real Data Processed by Machine Learning Techniques

1st Daniel Badran
dept. of Mechanics and Industrial Management
Faculdade de Engenharia da Universidade do Porto
Daniel_badran@hotmail.com

***Abstract*—In the past decade, the health industry has been producing huge amounts of data that could be used to aid doctors diagnose or even predict future possible illnesses. The purpose of this paper is to describe a project that creates a linear regression model using the historical relationship between a dependent variable and multiple explanatory independent variables, to predict the future of the dependent variable for a given duration of time, for different patients. The model and the prediction are established using Gretl; precisely using time series model. Later on, we construct a comparison table between different algorithms used in the study, to deduce which one is the most accurate after testing each one in MATLAB.**

***Index Terms*—Heart failure, machine learning, Coronary Heart Disease, Linear Regression, prediction**

I. Introduction

Coronary Heart Disease (CHD), is characterized by a wax-like substance called plaque that cumulates up inside the coronary arteries; this disease can be caused by age, a bad diet or genetic factors [1]. These arteries supply oxygen-rich blood to your heart muscle. When plaque builds up in the arteries, the condition is called atherosclerosis. The buildup of plaque occurs over many years.CHD, specifically cardiovascular diseases (CVDs) are the number one cause of death globally; more people die annually from CVDs than any other cause. Over 17.6 million people died from CVDs by 2012, where that number was maintained at 17.7 million in 2015, representing 31.43% of all global mortality. [2] Out of these deaths, an estimated 7.4 million were due to Coronary Heart Disease (CHD) and 6.7 million were due to strokes. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies. Patient's cardiovascular disease or ones that are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management using counseling and medications. The paper is structured as follows, Section II will be dedicated to illustrating and explaining previous work to give a feel where this idea came from and to present the thesis. Section III will present our proposed method as well as our approach to resolving this problem. Section IV will feature all the experimentations and the obtained results.

II. Related Work

A. Prediction of Coronary Heart Disease Using Risk Factor Categories

The main idea behind the study was to use seven risk factors; Low-Density Lipoprotein cholesterol, High-Density Lipoprotein cholesterol, sex, age, diabetes, smoking, and blood pressure). Factors such as obesity left ventricular hypertrophy, family history of premature coronary heart disease and estrogen replacement therapy have been taken into consideration, as input in order to create prediction algorithms using regression models formed by linear and logistic regressions to forecast CHD risk, for a middle-aged population. The following table in figure 1 illustrates the obtained results.

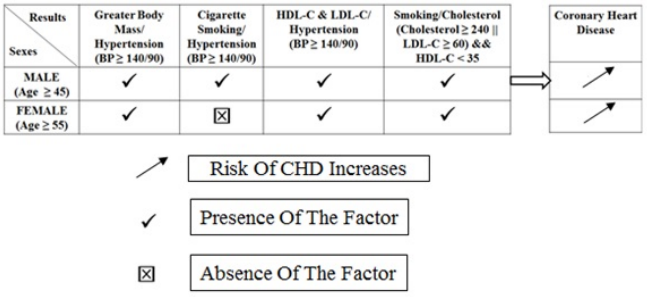

Fig. 1. Study's Obtained Results.

B. Reduced Number of Circulating Endothelial Progenitor Cells Predicts Future Cardiovascular Events

The objective of this project was to study over a period of 10 months using multiple cardiovascular events (Cardiovascular death, unstable angina, myocardial infarction, PTCA Percutaneous Transluminal Coronary Angioplasty: Procedure to open up blocked coronary arteries, CABG Coronary Artery Bypass Grafting: Surgery that improves blood flow to the heart, or ischemic stroke) to predict CHD risk. [3] To start the test EPCs were marked by CD34+KDR+; then the analysis was established using a normal distribution with the Kolmogorov-Smirnov fit test then compared by the Mann-Whitney U test using ANOVA. Comparison of categorical variables was generated by the Pearson X^2 test. The obtained results demonstrated the following:

- Documented Coronary Artery Disease (CAD) had higher number of risk factors.
- Reduced level of EPC is a surrogate marker of vascular function, and cumulative risk prediction and an identifier for future CHD risk.

C. Coronary Heart Disease Prediction From Lipoprotein Cholesterol Levels, Triglycerides, Lipoprotein (a), Apolipoproteins A-I and B, and HDL Density Subfractions

The main idea behind this project was to predict coronary heart disease from cholesterol levels and HDL density subfractions using Cox Proportional Hazards Regression Analysis, precisely using RR model. Analyses were performed separately; divided by sex. Mean lipid values were calculated for participants with and without incident CHD after age and race adjustment. The lowest CHD risk was found in the lowest LDL-C quintile, in women and men, and CHD risk accelerated with increasing values of LDL-C for both sexes. [4]

TABLE I
OBTAINED RESULTS

CHD Risk Level	LDL-C Level	Sex
High	High	M-F
High	High	M-F
Low	Low	M-F

D. Combination of Data Mining Methods With New Medical Data To Predict the Outcome Of Coronary Heart Disease

The goal of this study is to develop a data mining algorithm for predicting survival of the CHD patient. The analysis was performed during follow-ups using 3 data mining prediction models. The first being Support Vector Machine (SVM) with 10 folds cross-validation algorithm, which turned out to be the most accurate algorithm, the second model is, an artificial neural network using multi-layer perceptron with backpropagation, and the final model was decision trees. The results were divided into 3 major factors, accuracy, sensitivity and specificity. [5]

TABLE II
THIS STUDY'S RESULTS

Model	Accuracy	Sensitivity	Specificity
SVM	92.1% ;	92.87% ;	89.11%
ANN	91.0% ;	91.73% ;	88.12%
Decision Tree	89.6% ;	90.98% ;	84.16%

III. PROPOSED METHOD

A. Proposed Solution

Since the prediction will be based on 21 factors (Weight, Systolic Pressure, Diastolic Pressure, Pulse, Oxygen, Daily weight difference, Daily systolic pressure difference, Daily diastolic pressure difference, Daily pulse difference, Daily oxygen difference, Average variation in weight for the past three days, Average variation in systolic pressure for the past three days, Average variation in diastolic pressure for the past three days, Average variation in pulse for the past three days, Average variation in oxygen for the past three days, Average variation in weight for the past seven days, Average variation in systolic pressure for the past seven days, Average variation in diastolic pressure for the past seven days, Average variation in oxygen for the past seven days, Time). The time variable is added as a periodic variable, and the dummy variables are added to calculate the response (class variable).

B. Carried approaches

- Linear Regression:
  We propose to use the linear regression model to predict the heart failure risk level using the previously mentioned variables as input. Linear regression is the most basic type of regression and commonly used predictive analysis. The overall idea of linear regression is examining two things: does a set of predictor variables accurately predict an outcome variable? Which variables, in particular, are significant predictors of the dependent variable? [6] Linear regression uses the historical relationship (scatter plot) between an independent (x-axis) and a dependent variable (y-axis) to predict the future values of the dependent variables. The line with the smallest set of distances between the data points is the regression line. The trajectory of this line will best predict the future relationship between the two variables. In our case, the input is all the previously mentioned variables except for class variable, which is used as output. The following is the main equation of linear regression:

$$h_\theta(x) = \theta_0 + \theta_1(x) + \theta_2 \tag{1}$$

  h_θ (x) is the dependent value that the equation is trying to predict. θ_0 and θ_1 are selected so that the square of regression residuals is minimized. θ_2 is the linear residual
- Time Series Linear Regression
  Time series is a sequence of numerical data points in successive order. Commonly, time series is a sequence taken at successive equally spaced points in time, also called discrete time data. Time series are often used to examine how the changes associated with the chosen data point are compared to the shifts in other variables over the same time period. In forecasting, time series uses information regarding historical values and associated patterns to predict future activity. [7]F The main characteristic of time series is seasonality, in which, data experiences regular and predictable changes that recur within every defined time interval, in other words, periodic fluctuations. Seasonality may be caused by different factors, one of these factors is repetitive and generally is considered as regular predictable patterns in the level of time series. [7] Sometimes cyclic patterns may occur, its when data exhibits rises and falls that are not among the fixed period. In general, the average length of cycles is longer than the

length of a seasonal pattern, and the magnitude of cycles tends to be more variable than the magnitude of seasonal patterns. [8]

- Ordinary Least Square (OLS)
  The OLS or linear least square is a method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the sum of the squares of the differences between the observed responses in the given dataset and those predicted. To choose the most efficient OLS model, the following assumptions must be satisfied: [9]
  - All variables contained in a model are statistically significant. Meaning all their values must be greater than 0. Furthermore, no essential variable is omitted.
  - Residuals should not deviate significantly from 0 in any subset of the time series. Residuals represent the difference between the actual and the fitted value of each observation in the data series.
  - Residuals have a constant variance throughout the series. If the variance is not constant, then a number of remedies can be used, such as weighted regression power transformation, or generalized autoregressive conditional heteroscedasticity techniques may be applied to the data. Weighted regression is a technique that divides all series by the standard deviation of the errors term. A power transformation searches for the exponent of the series between -1 and +1 that results in a constant variance.
  - Residuals are free from autocorrelation for all lags.
  - Residuals are normally independently distributed. Failure of this assumption is linked to the failure of the previously mentioned condition.
  - Residuals are not a function of the lagged values of each of the independent variables.
  - X values in a series are not a function of the lagged residuals. If the combination of X values is a function of the lagged residuals, then a one-way causal model is the wrong functional term.
  - Residuals distribution is invariant over time, that is, one subset of the series data should have the same covariance structure [10] as another subset.

$$Y = \beta_0 + \sum_{p}^{j=1} \beta_j X_j + \tag{2}$$

Y is the dependant variable. β_0 is the intercept of the model. X_j is is the j^th explanatory variable of the model (j=1p. ϵ is the random error with variance σ^2.

- Classification Learner
  The classification learner helps us train models to classify data using supervised machine learning. These machine learning tasks are comprised by interactively exploring data, selecting features, specifying validation schemes, training models and asserting results. Few examples of this classification:
  - Decision Trees: Used to visually and explicitly represent decisions and decision making. It uses a tree-like model of decisions. Its a commonly used tool in data mining and machine learning for deriving a strategy to reach a particular goal. [11]
  - Support Vector Machines (SVM): Are a set of supervised learning methods, used for both classification and regression challenges. SVMs are the coordinates of the individual observation. SVM is a frontier which best segregates the two classes hyperplane/line. [12]

C. Used Softwares

- Gretl
  Gretl (GNU Regression, Econometrics and Time-Series Library) is an open-source cross-platform software package. It has a graphical user interface. Gretl offers features necessary for performing econometrics and time series analyses. [13]
- MATLAB
  Developed by MathWorks, MATLAB is a tool used to design and analyse systems or data using matrix-based MATLAB language to express computational mathematics.

IV. EXPERIMENTS AND RESULTS

A. Experimental Protocol

Our sample is a dataset of 1589 observations from all 3 patients; the dataset was split into 70% for training and 30% for testing. The models creation and training were conducted using the linear regression time series algorithm. Start date of our dataset was entered and the end date was calculated automatically by Gretl.

B. Obtained Results

We plot the selected variable (in our case we will plot the Class variable) using time series plot. This will display the variables progress through time.

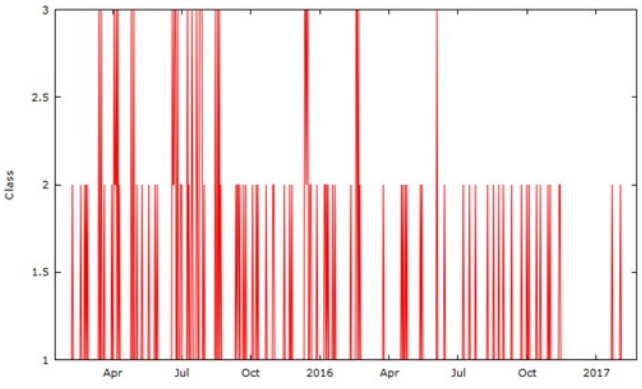

Fig. 2. Patient 1 Class Variable Graph.

The Class variable must be regressed against time by means of one-time variable and multiple dummy variables. Dummy

variable or indicator variable is an artificial variable created to represent an attribute with two or more distinct categories/levels. [14]
In order to compute the seasonality, one dummy variable is excluded. We obtained as results, R-squared = 0.708052 and Adjusted R-squared = 0.696440 = 69.6
Overall, we have an adjusted R-squared [15] of 69.6%. This is calculated based on independent variables.

$$R_A djusted^2 = 1 - (1 - R^2)(N - 1)/(N - p - 1) \quad (3)$$

Where: R^2=Sample R Squared. p=Number Of Predictors. N=Total Sample Size.
An in-sample forecast is run by using a constrained sample from our dataset. The obtained results for an in-sample forecast are Mean Absolute Error = 0.15601, Mean Percentage Error = 2.8008 and Mean Absolute Percentage Error = 12.935.

- Mean Absolute Error, [16] measures the difference between values (sample and population values) predicted by a model or an estimator and the values actually observed.
- If we take the mean absolute percentage error of 12.935% we note that the actual value and predicted value are 12.935% apart. The mean percentage error is calculated using the following formula: [16]

$$(1/N) \sum_{k-1}^{N} |A_k - F_k|/A_k \quad (4)$$

A_k is the actual value. F_k is the forecasted value. N is the total sample size.
The error of 12.9% can be considered as a good result.

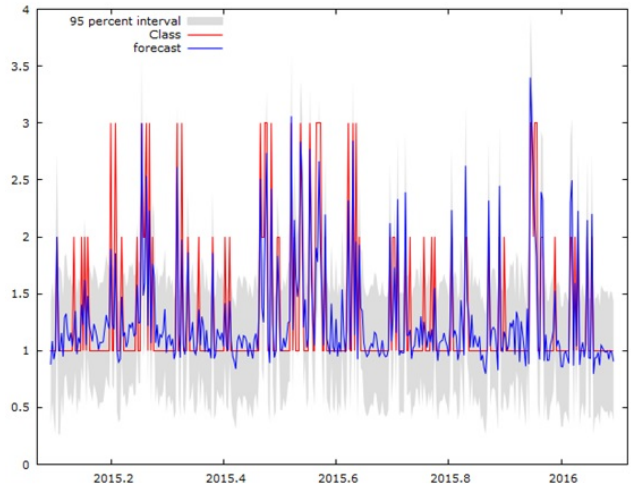

Fig. 3. Patient 1 Sample Prediction.

In this step, we will reset the sample to full range and rerun our linear regression model and add observations to the data, so we could predict the off-sample value of our targeted variable Class. After adding 10 observations to predict future unknown values, the above graph represents the results with 95% confidence interval. The blue line indicates the value of the classes each day; Mean Absolute Error = 0.24816, Mean Percentage Error = 5.4221 and Mean Absolute Percentage Error = 19.174. As the percentage error is 5.4221 this means, the accuracy is 94.5779%; with R-squared = 0.396388 and Adjusted R-squared = 0.365047 = 36.5%.

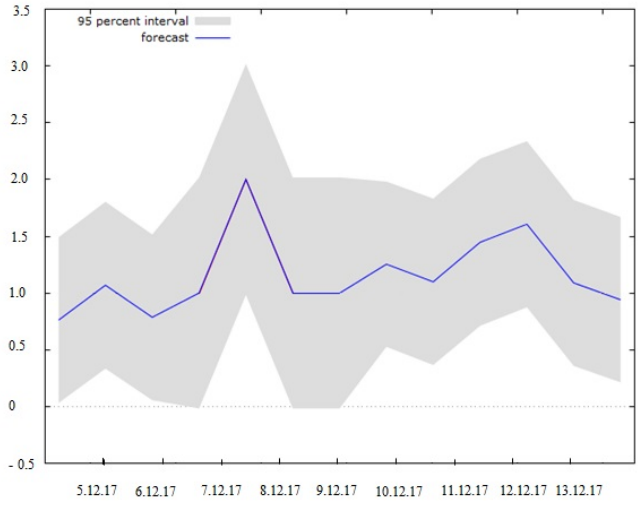

Fig. 4. 10 Day Off-Sample Prediction.

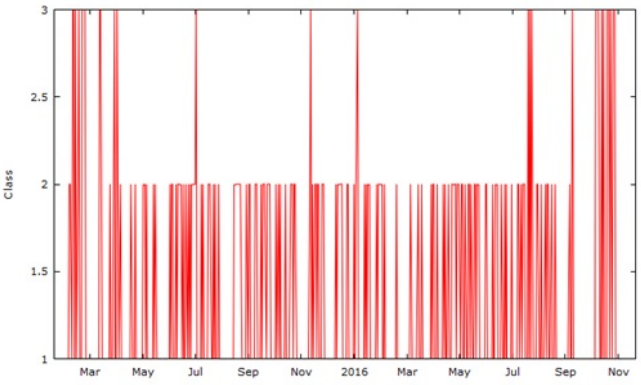

Fig. 5. Patient 2 Class Graph.

The in-sample forecast has the following results, Mean Absolute Error = 0.34503, Mean Percentage Error = -7.9653 and Mean Absolute Percentage Error = 25.823.

- If we take the Mean Percentage Error of 25.823% we note that the actual value and predicted value are 25.823% apart, which is considered unacceptable in clinical studies.

After adding 6 observations to predict future unknown values the above graph represents the results, with 95% confidence interval. The blue line indicates the value of the classes each day; showcasing a Mean Absolute Error = 0.2259, Mean

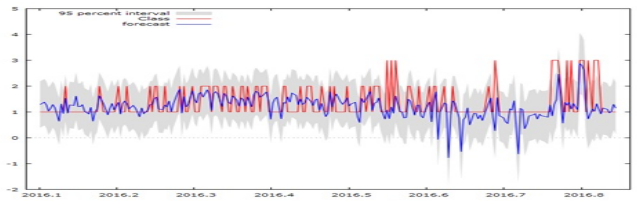

Fig. 6. Patient 2 Sample Prediction.

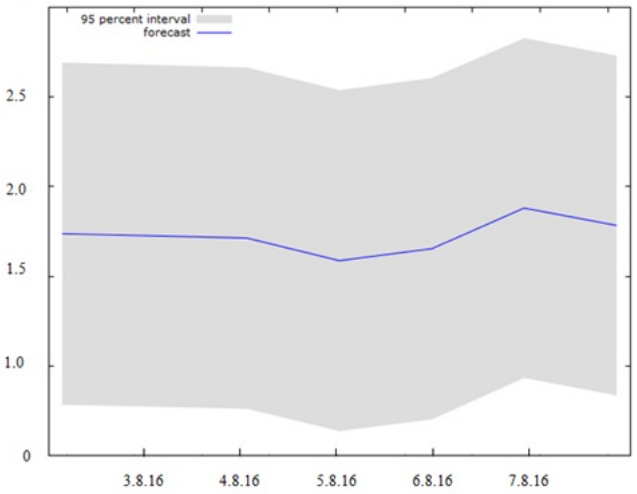

Fig. 7. Patient 2 Off-Sample Forecast Prediction.

Percentage Error = -22.59 and Mean Absolute Percentage Error = -22.59.
As the percentage error is 22.59 this means, the accuracy is 77.41%.

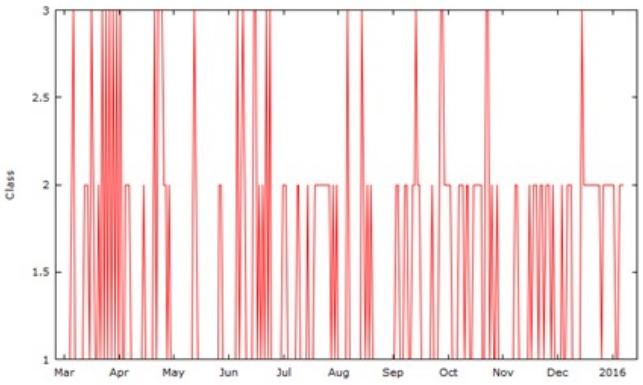

Fig. 8. Patient 3 Class Variable Graph.

The R-squared = 0.127255 and Adjusted R-squared = 0.039667 = 3.9667%, this in-sample forecast has the following results: Mean Absolute Error = 0.49991, Mean Percentage Error = -14.677 and Mean Absolute Percentage Error = 35.186.

- If we take the mean percentage error of 35.186% we note that the actual value and predicted value are 35.186% apart, which is considered unacceptable in clinical studies.

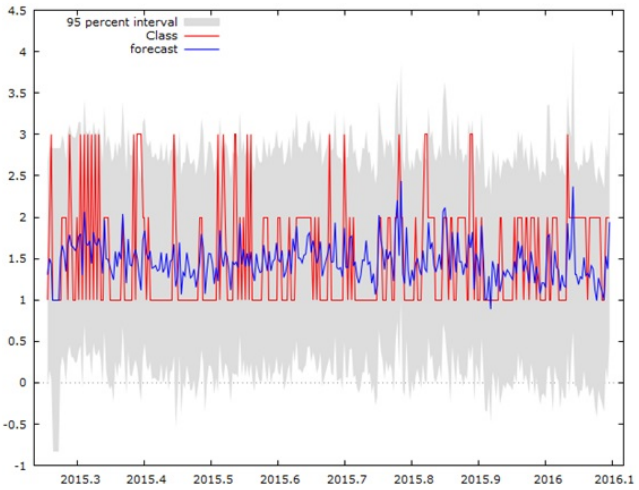

Fig. 9. Patient 3 In-Sample Prediction.

After resetting the sample to full range, and rerunning our linear regression model, we could now predict future observations with 95% confidence integral. We added 6 off-sample observations and the results are in fig. 10.

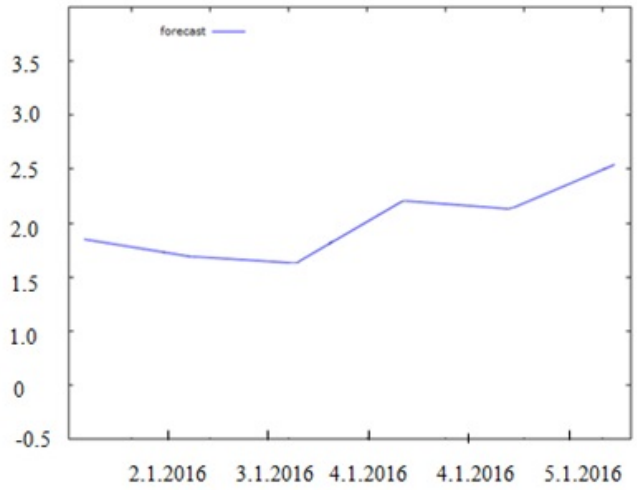

Fig. 10. Patient 3 Off-Sample Forecast.

For this patient the Mean Absolute Error = 0.49991, Mean Percentage Error = -14.677 and Mean Absolute Percentage Error = -22.59; as the Mean Percentage Error is 14.677 this means, the accuracy is 85.323%.

C. Results and Discussions

- Gretl Results:
  The following table illustrates all the previously obtained results using Gretl.

TABLE III
PATIENTS RESULTS SUMMARY

Patients	Accuracy
Patient 1	94.5779%
Patient 2	77.41%
Patient 3	94.8376%

- MATLAB Results:
  On MATLAB, we used all available classification algorithms (in classification learner) to study the data, and generate confusion matrices. 5 folds were used for testing the model. The following table presents the most accurate algorithm for all 3 patients.
  The best classification algorithm is Complex Tree: The average accuracy is equal to 92.8667%

TABLE IV
MATLAB RESULTS

Algorithm	Accuracy
Medium and Complex tree	95.6%
Medium and Complex tree	93.1%
Medium and Complex tree	89.9%

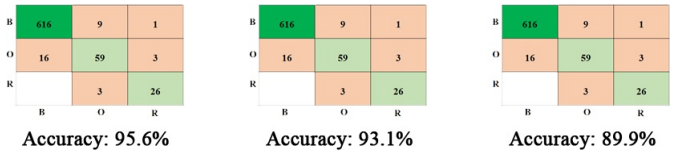

Fig. 11. Generate Confusion Matrixes.

V. CONCLUSION

This research highlighted the death risk of coronary heart disease and the main factors contributing to the patient's death, where we were able to implement a linear regression model capable of predicting which class can occur in the near future, based on multiple input samples. Also, when we look at this project, we find that it answers two main questions; the accuracy of the model and the prediction of a correct risk class, by having high efficiency and expandability; while including off-samples which are disregarded by other studies. Finally, recommendations for future studies should include larger forecast size and detecting what is/are the main factor(s) for risk evolvement. This will increase the relevance of the results and show even higher significance, thus we can cover more patients and reduce mortality rate.

REFERENCES

[1] B. G. Nordestgaard, M. J. Chapman, S. E. Humphries, H. N. Ginsberg, L. Masana, O. S. Descamps, O. Wiklund, R. A. Hegele, F. J. Raal, J. C. Defesche, A. Wiegman, R. D. Santos, G. F. Watts, K. G. Parhofer, G. K. Hovingh, P. T. Kovanen, C. Boileau, M. Averna, J. Born, E. Bruckert, A. L. Catapano, J. A. Kuivenhoven, P. Pajukanta, K. Ray, A. F. H. Stalenhoef, E. Stroes, M.-R. Taskinen, A. Tybjrg-Hansen, and for the European Atherosclerosis Society Consensus Panel, "Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease consensus statement of the european atherosclerosis society," *European Heart Journal*, vol. 34, no. 45, pp. 3478–3490, 2013.

[2] C. J. McAloon, L. M. Boylan, T. Hamborg, N. Stallard, F. Osman, P. B. Lim, and S. A. Hayat, "The changing face of cardiovascular disease 2000–2012: An analysis of the world health organisation global health estimates data," *International journal of cardiology*, vol. 224, pp. 256–264, 2016.

[3] S.-L. Caroline, R. Lothar, F. Stephan, V. Mariuca, B. Martina, K. Ulrike, D. Stefnaie, and Z. Amdreas, "educed number of circulating endothelial progenitor cells predicts future cardiovascular events. proof of concept for the clinical importance of endogenous vascular repair.," vol. 16, pp. 2981–2987, 2005.

[4] A. Sharrett, C. Ballantyne, S. Coady, G. Heiss, P. Sorlie, and W. Patsch, "Coronary heart disease prediction from lipoprotein cholesterol levels, triglycerides, lipoprotein(a), apolipoproteins a-i and b, and hdl density subfractions.," vol. 10, pp. 1108–1113, 2001.

[5] X. Yanwei, W. Jie, and Z. Zhihong, "Combination data mining methods with new medical data to predicting outcome of coronary heart disease, convergence information.," pp. 868–872, 2007.

[6] R. B. Darlington and A. F. Hayes, *Regression analysis and linear models: Concepts, applications, and implementation*. Guilford Publications, 2016.

[7] J. J. Faraway, *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, vol. 124. CRC press, 2016.

[8] C. Chatfield, *The analysis of time series: an introduction*. CRC press, 2016.

[9] J. C. Pickett, D. P. Reilly, and R. M. McIntyre, "How to select a most efficient ols model for a time series data," *THE JOURNAL OF BUSINESS*, vol. 11, 2005.

[10] R. Wolfinger, "Covariance structure selection in general mixed models," *Communications in statistics-Simulation and computation*, vol. 22, no. 4, pp. 1079–1106, 1993.

[11] https://www.mathworks.com/help/stats/classification-nearest-neighbors.html, n.d.

[12] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[13] R. Lucchetti *et al.*, "Who uses gretl? an analysis of the sourceforge download data," in *Econometrics with gretl. Proceedings of the gretl conference 2009, Bilbao, Spain*, pp. 45–55, 2009.

[14] S. Skrivanek, "The use of dummy variables in regression analysis," *More Steam, LLC*, 2009.

[15] J. Frost, "Multiple regression analysis: Use adjusted r-squared and predicted r-squared to include the correct number of variables," *Minitab Blog*, vol. 13, no. 06, 2013.

[16] A. De Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean absolute percentage error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, 2016.