UNIVERSITI TEKNOLOGI MARA
Cawangan Kedah
Kampus Sungai Petani

2021 iCMS
INTERNATIONAL CONFERENCE ON COMPUTING, MATHEMATICS AND STATISTICS

e-PROCEEDINGS

of The 5th International Conference on Computing, Mathematics and Statistics (iCMS2021)

4-5 August 2021

Driving Research Towards Excellence

e-Proceedings of the 5th International Conference on Computing, Mathematics and Statistics (iCMS 2021)

Driving Research Towards Excellence

Editor-in-Chief: Norin Rahayu Shamsuddin

Editorial team:
Dr. Afida Ahamad
Dr. Norliana Mohd Najib
Dr. Nor Athirah Mohd Zin
Dr. Siti Nur Alwani Salleh
Kartini Kasim
Dr. Ida Normaya Mohd Nasir
Kamarul Ariffin Mansor

e-ISBN: 978-967-2948-12-4
DOI

Library of Congress Control Number:

Publication by
Department of Mathematical Sciences
Faculty of Computer & Mathematical Sciences
UiTM Kedah

APPLY MACHINE LEARNING TO PREDICT CARDIOVASCULAR RISK IN RURAL CLINICS FROM MEXICO

Misael Zambrano-de la Torre[1], Maximiliano Guzmán-Fernández[2], Claudia Sifuentes-Gallardo[3], Hamurabi Gamboa-Rosales[4], Huizilopoztli Luna-García[5], Ernesto Sandoval-García[6], Ramiro Esquivel-Felix[7] and Héctor Durán-Muñoz[8]

[1,2,3,4,5,6,8] Unidad Académica de Ingeniería Eléctrica. Universidad Autónoma de Zacatecas, México,
[7] Carrera de Mecatrónica. Universidad Tecnológica del Estado de Zacatecas
([1] misaelzambrano1997@gmail.com, [2] maxguzman1@hotmail.com, [3] clauger17@gmail.com,
[4] hamurabigr@uaz.edu.mx, [5] hlugar@uaz.edu.mx, [6] esandoval@uaz.edu.mx,
[7] resquivel@utzac.edu.mx, [8] hectorduranm@hotmail.com)

Abstract. Approximately 41 million people in the world die each year from cardiovascular diseases. In Mexico, it is one of the main causes of death per year. This problem is even more critical in rural areas of Mexico. Due to the limited number of specialized medical equipment available in these clinics. Therefore, the objective of this work is to propose a new stage in the methodology used in machine learning for the classification of cardiovascular risk in rural clinics in Mexico. The importance of this work is being able to classify patients based only on non-invasive attributes, avoiding the use of specialized clinical equipment. For this purpose, the Heart Disease Data Set repository is used to implement the new stage. The methodology to be implemented consists of 6 stages. The performance of the three algorithms is compared in terms of four parameters. The results obtained show that only 4 attributes are required for classification with an 80% acceptance rate.

***Keywords*:** Machine Learning, Cardiovascular risk and Specialized medical equipment.

1. Introduction

Currently, according to the World Health Organization (WHO), degenerative diseases cause the death of approximately 41 million people per year, representing 71% of deaths worldwide. From this total, cardiovascular diseases account for 17.9 million deaths worldwide. In the case of Mexico, according to the National Institute of Statistics and Geography (INEGI), in 2019, there are a total of 747, 784 deaths. Cardiovascular diseases occupy 23.5% (156, 041 people), from 88.8% of the total deaths in 2019. It is a fact that cardiovascular diseases have a significant impact on the population of the country. This is why artificial intelligence has been looking to develop solutions through the use of machine learning algorithms. Training algorithms with different attributes and instances (observations), in order to obtain the best performance in a classification or prediction (Diwakar et al., 2020).

Artificial intelligence has developed supervised prediction or classification algorithms in the healthcare area (Ramesh et al., 2004). These algorithms are trained, tested and implemented in specific contexts, which are defined by experts and developers. In the cardiology context, by using machine learning and data mining, patient attributes are studied and help classify the risk of developing heart disease (Uddin et al., 2019). Data mining and information processing allow the use of different methodologies to make predictions or classifications. In other words, it can be adapted to different areas of health.

Some examples of this are the use of machine learning to assess cardiovascular disease in dialysis patients, where different algorithms were tested (Mezzatesta et al., 2019). In their work, it was reported that using the Support Vector Machine algorithm, a performance in terms of accuracy of 95.25% was obtained in the Italian health sector database and 92.15% in the US health sector database, following a five-stage methodology: database description, preprocessing, attribute description, algorithm training and model validation. On the other hand, a system that automatically selects and fits machine learning models is found in the literature. The constructed model was tested using data from 423,604 participants. The results show that the model obtains

an accuracy rate of 95%. There is also a record of a comparative analysis of machine learning methods with the Hellenic Score cardiovascular disease risk tool, as the name given by the author (Dimopoulos et al., 2018).

Once understood, how artificial intelligence, machine learning and classification algorithms are involved in cardiovascular health. The problem affecting physicians in rural clinics in Mexico was identified. Commonly, rural clinics do not have the infrastructure, supplies and technology to detect the risk of cardiovascular disease. This puts the patient's cardiovascular health at risk due to the deficiency of a specialized study. By considering "non-invasive" attributes, it is possible to develop a process to follow to classify patients with risk of cardiovascular problems. This will have less impact on the patient's integrity and will take advantage of the benefits of artificial intelligence.

2. Materials and Methods

The methodology used in this work is based in the literature, with 5 stages for information processing. However, in most of the reported works, it is common to use the available attributes indiscriminately. This process allows to obtain the best performance, considering any attribute. But the problem lies in the ability to obtain them. I.e., the monetary cost, the inconvenience to the patient and the tools available in rural areas of Mexico. Therefore, in this work it is proposed to add a new stage to the methodology commonly used, and avoid the previous problems. The proposed new stage is shown in Figure 1, Stage 3. This new stage consists of analyzing and assigning "Invasive" and "Non-invasive" attributes in the database. Trying to use only "Non-invasive" attributes and obtain acceptable performance. The 6 stages used in this work are described in the following subsections.

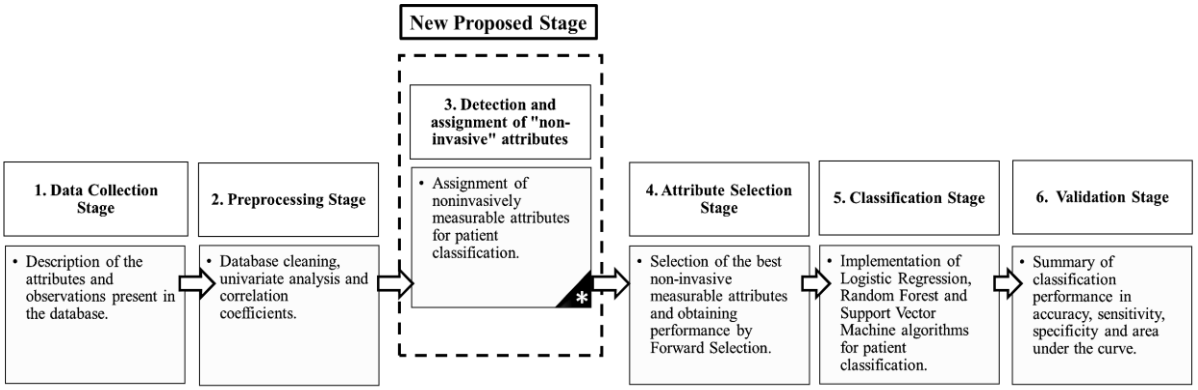

Figure 1: Methodology for classifying cardiovascular risk, incorporating the new stage 3: detection of "noninvasive" attributes

2.1 Stage 1: Data Collection

"The Heart Disease Data Set" from the CAD repository from the University of California was used. The main reason for using this database is because currently in Mexico there is no official registry of patients that can be freely used for research. In addition to not having the consent and privacy rights of patients. On the other hand, this database is previously validated data available for free use in research (Dua and Graff, 2017). This database has 75 attributes and 303 patients. In this work 14 attributes were used as shown in Table 1. Two of which are general patient information (age and sex). The remaining twelve attributes provide clinical information on the patient's cardiovascular health condition. More information can be found in the description of the repository. Where it is possible to find specialized terms that require general medical knowledge.

This database has 75 attributes and 303 patients. For practical purposes in this work, the file corresponding to the fourteen attributes was extracted directly from the repository. It is important to note that different versions of the database can be found on the web. It is recommended to visit the official page of the repository and extract the database.

Table 1: Attributes used in this work.

Attribute Name	Abbreviation	Attribute Type
Age	age	General information
Sex	sex	
Chest Pain	chest-pain	Cardiovascular Health
Resting Blood Pressure	rblood-pressure	
Serum cholesterol	cholesterol	
Fasting blood sugar	sugar	
Maximum heart rate	heart-rate	
Exercise-induced angina	induced-angina	
Resting electro results	electro-results	
Exercise-induced ST	Old-peak	
The slope ST segment	slope	
Number of main vessels	No-vessels	
Thalassemia	thalassemia	
Condition	condition	Classification

2.2 Stage 2: Database Preprocessing

In this stage, the database is preprocessed and any possible errors in the data are identified. Five null values were found, so the attributes of this patient were eliminated, reducing the number of patients to 298. Once the preprocessing is done, a database with 14 attributes is obtained, including general information and cardiovascular health attributes of the patients. Then, a univariate analysis was performed for each attribute. The Pearson's, Spearman's and Kendall's correlation coefficient was obtained. Finally, a correlation matrix was obtained to summarize the correlation coefficients between each of the attributes.

2.3 Stage 3: Detection of "Non-Invasive"

Commonly in the literature an indiscriminate selection of attributes is made, without considering the complications to obtain such attributes. For this reason, a new third stage is proposed. It consists of separating in two groups: "Invasive" and "Non-invasive". Based on how each of them can be measured. In this way, it is quicker and easier to make a classification. I.e., it does not require a specialized procedure. The assignment was made considering the opinion of an expert in the health area. Because in rural clinics in Mexico, there are only three tools for patient examination: an oximeter, a stethoscope and a mercury sphygmomanometer. Based on the above, the 14 attributes were classified into invasive and non-invasive. Subsequently, the summary of the univariate analysis was performed and a correlation matrix was made.

2.4 Stage 4: Attribute Selection

At this stage, eight attributes were selected. In order that the patient does not require clinical analysis or interventions. The attributes selected and assigned as "Non-invasive" to be used in the next stage for the training of the three algorithms. These attributes are: Age, Sex, Chest Pain, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Maximum Heart Rate and Exercise-Induced Angina. All these attributes can be measured quickly, easily and by low-cost electronic sensors, avoiding inconvenience and saving money. Subsequently, the Forward Selection technique was used to measure the performance of the algorithms, testing different combinations between the eight attributes. The parameter used to evaluate performance for each combination of attributes was the

area under the curve. If the value is close to 0, performance is not good. But if the value is close to 1, performance is acceptable. The operating parameters used for the Forward Selection technique were 70% of the data for the training stage of the model, and 30% of the remaining data for the testing. Finally, the number of attributes was reduced from eight to only four. Therefore, only four attributes are required to perform the classification and keep the similar performance when using eight attributes. I.e., it is necessary to measure only four attributes. This saves money, time, resources and computational time.

2.5 Stage 5: Classification

After attribute selection, patient classification is applied using three different classification algorithms: multivariable Logistic Regression (LR), Random Forest (RF) and Support Vector Machines (SVM). Patients were classified according to two conditions: high (1) or low (0) risk of cardiovascular problems. For each algorithm, the K-fold cross-validation technique was used. This validation technique consists of a random division of the data into k groups of equal size. Once the distribution of the data set was done, the training and testing of the three algorithms was performed. The operation configuration of the algorithms is shown below. As well as, the syntax to generate them.

2.5.1 Logistic Regression Algorithm

The four attributes obtained with the Forward Selection technique were used (sex, chest pain maximum heart rate, exercise-induced angina). The R language was used to implement the algorithm, based on the “glm” function, with the default configuration. The formula used is shown in Figure 2.

```
predictors<-c("sex","chest_pain","heart_rate","induced_angina" ) #attributes
output<-"condition" # output, attribute for classification

#Let's train a model with the logistic regression algorithm
model_glm<-train(train_data[,predictors], train_data[,output], method = "glm",
                 trControl = traintype,
                 tuneLength = 3) # Operating conditions for creating the model
                                 # from the logistic regression algorithm
```

Figure 2: Attributes and algorithms used for Logistic Regression.

This model generated the predictions for the 298 observations. For binarization, it was necessary to define a threshold. The threshold allows to classify 0 and 1 the probabilities obtained. Then, the threshold obtained from the ROC (Receiver Operating Characteristic) curve of the pROC library was used. Where the point of the best performance of the algorithm, as predicted, can be retained. This ensures that the maximum performance of the algorithm is obtained, i.e., as $p > 0.5 = 1$ or $p < 0.5 = 0$. The operating conditions of the model show an intercept of approximately 0.84 and the degree of contribution corresponding to each of the attributes involved in the model. Once the Logistic Regression model was built, the next algorithm to be trained and tested was Random Forest.

2.5.2 Random Forest Algorithm

This algorithm is used because it has good results in the health area. It has a high classification capability. For the implementation, the same 4 attributes were used as in Logistic Regression. The “randomforest” library was used for the implementation. The formula used is shown in Figure 3.

```
predictors<-c("sex","chest_pain","heart_rate","induced_angina" ) #attributes
output<-"condition" # output, attribute for classification

#Let's train a model with the random forest algorithm
model_RF<-train(train_data[,predictors], train_data[,output], method = "rf",
                trControl = traintype,
                tuneLength = 3) # Operating conditions for creating the model
                                # from the random forest algorithm
```

Figure 3: Attributes and algorithms used for Random Forest.

Binarization was carried out using the ROC curve implemented in logistic regression. The type of random forests used was regression type, commonly used in classification. A total of 500 decision trees was used and mean squared residual of 0.17. Finally, to conclude the classification stage, the third algorithm, Support Vector Machine, was trained and tested.

2.5.3 Support Vector Machine Algorithm

The Support Vector Machine, in its classification mode, is used to separate patients into two groups, based on the same 4 attributes used previously. I. e., to predict the problem of high or low risk of cardiovascular problems. The "svm" function was used, available in R library e1071 in version 4.0. As in the case of logistic regression, the binarization of the probabilities was achieved based on the best threshold obtained. The formula used is shown in Figure 4.

```
predictors<-c("sex","chest_pain","heart_rate","induced_angina" ) #attributes
output<-"condition" # output, attribute for classification

#Let's train a model with the support vector machine algorithm
model_svm<-train(train_data[,predictors], train_data[,output], method = "svmLinear",
                 trControl = traintype,
                 tuneLength = 3) # Operating conditions for creating the model
                                 # from the support vector machine algorithm
```

Figure 4: Attributes used for Support Vector Machines.

The number of support vectors used was 137 and uses the linear kernel function. The parameters were the following: $\epsilon = 0.1$ and for the cost C = 1. Finally, each algorithm was trained under the same number of observations and attributes. This was done in order to be able to compare their performance afterwards.

2.6 Stage 6: Validation

In the sixth stage, the performance of the algorithms is compared. The evaluation is carried out by performance obtained during the "test". Four parameters are used and obtained for the evaluation: accuracy, area under the curve, sensitivity and specificity. In order to evaluate the efficacy of the algorithms and their ability to correctly classify patients. This was carried out by statistical validation. Sensitivity was calculated from true positives (TP) and false negatives (FN), it represents the way that a subject who is sick is correctly detected. Equation 1 is shown below.

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

Specificity was calculated from true negatives (TN) and false positives (FP). It refers to the proportion of subjects without disease that were correctly classified. Equation 2 is shown below.

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

Accuracy was calculated from the TP, TN, FP and false negatives (FN). It refers to the total probability that a subject is correctly classified. Equation 3 is shown below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Finally, the area under the curve is commonly used in classification analysis to evaluate how well an algorithm performs in classification. It is calculated by evaluating the Receiver Operating Characteristic curve. From the relationship between sensitivity and 1 - specificity. It is calculated to

predict the probability that the algorithm will classify a randomly selected positive patient to a greater extent than a randomly selected negative patient. According as the algorithms show a higher probability of hitting the correct prediction, better is the performance of the algorithm. The range of the probability is between 0.5 and 1, where 0.5 means that the performance is not ideal for classification and a value close to 1 means that the performance is ideal for correctly classifying patients.

3. Results and Discussions

In the second stage, 5 observations with NA values were detected, leaving 298 remaining observations, 13 attributes of patient information and only one for condition. Therefore, the data set was fully adequate to proceed to the next stage. As a result of the new third stage, two groups of attributes were considered, "Invasive" and "Non-invasive". The assignment was made on the basis of the exploration method, according to the health care expert. Eight attributes were considered non-invasive and some of them can be obtained by simple questioning. The results of correlation coefficients by univariate analysis are shown in Figure 5 using a heat map of the correlation matrix.

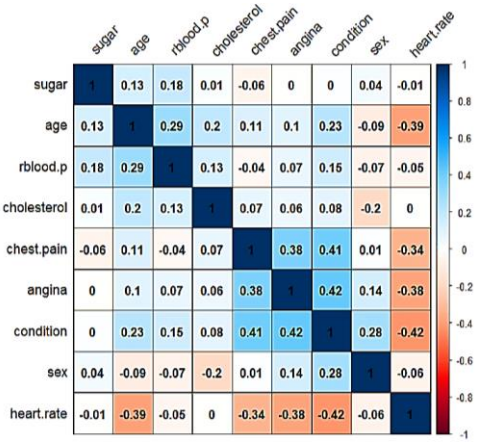

Figure 5: Heat map of the correlation matrix of eight non-invasively measurable attributes.

As shown in Figure 5, the attributes "sugar" and "cholesterol" have a low correlation coefficient with respect to the attribute "condition". This means that these two attributes will not contribute enough information to the correct classification. However, the attributes "angina", "heart rate" and "chest pain" are the attributes that will have the highest statistical correlation in the classification. Subsequently, in the attribute selection stage, the Forward Selection technique was used to determine the possible performance of the three algorithms. The parameter to measure performance was the area under the curve. Figure 6 shows graphically the performance obtained by Forward Selection.

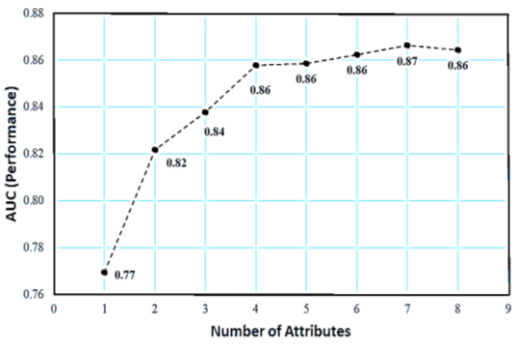

Figure 6: Forward selection in terms of the area under the curve of 8 attributes.

Based on Figure 6, it can be seen that only four attributes need to be considered. The performance obtained is similar when considering only eight attributes. In other words, the performance of the algorithms will not be affected by using only four attributes and not eight attributes. Therefore, it is observed that the computational cost and the exploration work by the doctor is reduced to 50%. Based on the Forward Selection analysis, the four attributes that allow classification with similar performance when using eight attributes are: Sex, Chest Pain, Maximum Heart Rate, Exercise-induced Angina and the Condition.

Analyzing the attributes found and selected, it can be seen that only a measurement instrument and some patient questions are required to obtain a prediction of high or low risk of cardiovascular problems. As a result, there will be certainty of correctly classifying 86% of the patients, with a margin of error of 14%. In other words, out of every 10 patients who submit to the study, 8 of them will be correctly classified and only 2 will not be correctly classified. From five to eight attributes, the difference in the area under the curve is approximately 0.01. This does not represent a profit in performance, but does represent a significant reduction in attributes. Once the attributes were selected, training and testing of the three classification algorithms were performed. Four parameters were obtained to measure their performance: accuracy, sensitivity, specificity and area under the curve. Performance was obtained for the 4 attributes obtained during the testing of the different algorithms. On the other hand, the same training process was carried out, but this time considering the total number of attributes in the database, i.e., 14 attributes. The main purpose is to make a direct comparison of the performances and finally, to check if the methodology proposed in this work has worked. Table 2 shows the results obtained.

Table 2: Performance comparison of Logistic Regression, Random Forest and Support Vector Machine algorithms with four and fourteen attributes

Attributes	Algorithm	Accuracy	Sensitivity	Specificity	Area Under the Curve	Attribute Type
4	Logistic Regression	0.6621	0.8181	0.5365	0.7888	Non-invasive
	Random Forest	0.7702	0.8085	0.7037	0.7934	
	Support Vector Machine	0.6891	0.8108	0.5675	0.7414	
14	Logistic Regression	0.6216	0.7368	0.5000	0.8944	Invasive / Non-invasive
	Random Forest	0.8648	0.8750	0.8461	0.9037	
	Support Vector Machine	0.8378	0.8269	0.8636	0.8711	

Considering the Random Forest algorithm with only 4 attributes and the performance obtained in the area under the curve with 0.79, it is the most ideal algorithm for classifying patients. On the other hand, it is important to highlight the difference between the performance obtained for 14 attributes (0.9) and for 4 attributes (0.79). I. e., the difference is approximately 0.1, which can be considered a big difference, but if it considers only 4 attributes and not 14, we have a reduction in the number of attributes of approximately 70%. This is reflected in lower computational cost and data collection. In practice, this will help doctors to generate patient classification with just a few tools. Obviously, measuring and classification will also be simpler and faster. It will be obtained in exchange for 0.1: reduction in monetary cost, measurement discomfort for patients, computational cost and avoidance of specialized and expensive instruments.

4. Conclusions

The new stage proposed in the methodology shown in Figure 1 proposes assigning the attributes as "Non-invasive". In this way, data collection is made easier and patients are classified in the best possible way. In this work, it is concluded that the new proposed stage proves to be useful, acceptable, easy to implement and reproduce, due to its correct performance. With this, it is possible to apply it to new health data sets in Mexico. The dataset used in this work allowed validating the proposed new

stage, thanks to the statistically significant results. It is important to highlight that this work allows the generation of an alternative methodology by considering a new stage. This makes it different from the methodologies already existing in the literature for the prediction and early detection of cardiovascular problems. The main difference is the way in which the attribute is measured or obtained and the complications it may produce in the patient. In this work, only non-invasive measurable attributes are considered. This makes it easier to obtain the attributes and therefore, the classification of the patient is also easy. This considering the conditions and tools available in rural clinics in Mexico.

From the results obtained, it is concluded that a reduction of approximately 70% of attributes was achieved in exchange for a loss in performance of 10%, in terms of the area under the curve. The main advantage is that only 4 attributes need to be measured for the classification. The computational cost, as well as the time to obtain attributes, will be much lower. These four attributes were shown to have a high statistical correlation coefficient in relation to the condition attribute (classification). Thanks to the methodology illustrated in Figure 1 and the proposed new step, it can be concluded that the process followed proved to be easy to implement and that it is possible to consider only attributes that are measurable in a non-invasive way for the classification of patients at risk of cardiovascular problems. This work does not seek to replace or avoid the intervention and judgment of the health expert. On the contrary, it is always advisable to consider the health expert during the process.

Acknowledgment

The authors gratefully acknowledge the scholarships for this work from CONACYT.

References

M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh, and N. Kumar, “Latest trends on heart disease prediction using machine learning and image fusion,” in *Materials Today: Proceedings*, Jan. 2020, vol. 37, no. Part 2, pp. 3213–3218, doi: 10.1016/j.matpr.2020.09.078.

A. N. Ramesh, C. Kambhampati, J. R. T. Monson, and P. J. Drew, “Artificial intelligence in medicine,” *Annals of the Royal College of Surgeons of England*, vol. 86, no. 5. Royal College of Surgeons of England, pp. 334–338, Sep. 2004, doi: 10.1308/147870804290.

S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 281, 2019, doi: 10.1186/s12911-019-1004-8.

S. Mezzatesta, C. Torino, P. De Meo, G. Fiumara, and A. Vilasi, “A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis,” *Comput. Methods Programs Biomed.*, vol. 177, pp. 9–15, Aug. 2019, doi: 10.1016/j.cmpb.2019.05.005.

A. C. Dimopoulos *et al.*, “Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk,” *BMC Med. Res. Methodol.*, vol. 18, no. 1, p. 179, Dec. 2018, doi: 10.1186/s12874-018-0644-1.

Dua, D. and Graff, C, “{UCI} Machine Learning Repository,” 2017. http://archive.ics.uci.edu/ml (accessed Dec. 04, 2020).

2021 ICMS

INTERNATIONAL CONFERENCE ON COMPUTING, MATHEMATICS AND STATISTICS

e ISBN 978-967-2948-12-4

9 7 8 9 6 7 2 9 4 8 1 2 4