Comparative study of machine learning algorithms (SVM, Logistic Regression and KNN) to predict cardiovascular diseases

Mohammed Marouane Saim [1], *Hassan Ammor* [2]

[1] ERSC, E3S Research Center, Mohammed V University of Rabat, Morocco; mohammedmarouanesaim@research.emi.ac.ma,
[2] ERSC, E3S Research Center, Mohammed V University of Rabat, Morocco; ammor@emi.ac.ma,

Abstract. Artificial intelligence has had an impact on a variety of fields, including medicine and, most importantly, cardiovascular diseases. Indeed, early diagnosis of many disorders is a serious medical issue. In this article, we will compare various machine learning algorithms in order to select the optimal one for diagnosing people who might suffer from heart disease based on a variety of clinical data from patients. The effort in this article is focused on studying the dataset using data mining algorithms, and also explaining the used machine learning algorithms in predicting heart disease, in order to assist future researchers in getting the most out of these skills.

Keywords: Data Mining Algorithms, Heart Disease, Risk Prediction, Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Boruta, Performance metrics.

1. Introduction

Artificial intelligence approaches were used in various areas of medical science, particularly cardiovascular illnesses, to produce innovative medical assistant systems. Disease diagnosis has progressed beyond internal treatment approaches; presently, doctors are focusing more on early identification or prediction of illnesses based on risk factors derived from the patient's lifestyle. The main cause of death in the globe is cardiovascular disease, with more people dying from it each year than from any other cause. Cardiovascular disease is predicted to account for 17.7 million deaths worldwide, or 31% of overall mortality . Coronary heart disease accounts for 7.4 million of these deaths, whereas stroke accounts for 6.7 million. (2015 data), both of which are expected to rise in the future [1].

According to a study from 2018, the top two causes of death in Morocco are: coronary heart disease with a percentage of 28.7% and stroke with a percentage of 9.42% [2]. Amidst all of these numbers, According to figures gathered from the evaluation of cardiovascular illnesses, 16.1 % have high blood pressure, 38.9 % have low physical activity, and 43.9 % are overweight, and 10.8 % use cigarettes, which is one of the leading causes of heart disease.

The diagnosis of heart problems is a challenging undertaking as well as a vital responsibility in the medical sector, and it necessitates the utmost care. However, there are some methods for data extraction and analysis, plus the presence of a large set of medical data leads to the proper diagnosis of the disease. It is possible to raise the chance of heart disease prediction by using medical data such as age, sex, blood sugar, and blood pressure. These data must be collected in an organized manner so that the preventative system can be developed. It's possible to prevent that outcome by predicting a person's likelihood of developing heart disease based on their clinical data. The good news is that heart attacks are highly preventable, and that early treatment with minor lifestyle adjustments improves the prognosis significantly.

This article is divided into six sections. The definition of the problem and the purpose of this essay are presented in the first two sections. The dataset and data mining methods utilized in this article are discussed in part three, while the three algorithms (k-nearest neighbours, logistic regression, and support vector machine) employed in the prediction of heart illnesses are discussed in section four. The fifth section offers a comparison of each algorithm's metrics. The conclusion and future propositions about the subject of this paper are presented in section six.

2. Heart diseases

The cardiovascular apparatus is the conductor of all the organs; it ensures blood circulation to provide nutrients and oxygen to the different cells, which helps maintain the general metabolism. Coronary heart disease, high blood pressure, and hypertensive heart disease are only a few of the illnesses that impact the heart and blood arteries.

Obesity, smoking, a poor diet, a lack of physical activity, problematic alcohol consumption, diabetes, hypertension, and other risk factors all have a contribution in the development of different heart diseases. These "intermediate risk factors" can be assessed in primary care settings and signal a high risk of heart attack, heart failure, stroke, and other complications.

Risk evaluation: the Framingham questionnaire [3]: This survey is used to forecast the risk of cardiovascular disease over the next ten years. It can be minimal (less than 10%), medium (10% to 19%), or strong (more than 20%). The findings assist doctors in making treatment decisions. The treatment will be more intensive if the risk is high. Age, blood pressure, cholesterol levels, and other risk variables are all considered in this survey. It is widely used by doctors in Canada and the United States. It was created in the United States of America. There are many different sorts of surveys because they must be adjusted to the people who utilize them. One of the most extensively used in Europe is SCORE (Systematic COronary Risk Evaluation).

3. Dataset

The dataset used in this article is publicly available on the Kaggle website; it's from a cardiovascular disease study. The classification's purpose is to determine whether the patient is at risk for heart disease in the next 10 years. The dataset offers patient information; it incorporates over 4000 records and features. The attributes can be divided on 4 categories: Demographic, behavioral, information on medical history and information on current medical conditions.

3.1. Data cleaning

As the amount of data we have has increased, so has the risk of error. As a result, we must rely on data cleaning to improve the efficiency of our data management procedures. By reducing inaccuracies, data cleaning strengthens the data's integrity. A range of activities aimed at detecting and fixing damaged data is known as data cleaning. This crucial phase in data processing increases the data's dependability and worth. Missing values and entries that do not display in the correct spot are the most typical sources of data inaccuracy. In the dataset used in this article, the major problem we found was missing data with a percentage of 12, 74 %.

3.2. Exploratory data analysis

Unlike the traditional hypothesis testing approach, which aims to verify preconceived hypotheses on the relationships between variables, exploratory data analysis makes it possible to define systematic relationships between variables, in the absence of these presumptions of the nature of these relationships. In a typical exploratory analysis, the researcher takes into account and compares, using various techniques, many variables to reveal systematic structures.

Exploratory data analysis (EDA) is used in this project to provide a better understanding for the dataset, it helps determine how best to use the data. To explore data and find the best way to use these features, we used different types of graphs for example: Histogram: to check the distribution of the different attributes.

3.3. Feature selection

Almost every dataset has hundreds of assembled features, but not all these attributes have the same importance in different projects. Which makes selecting the significant ones and integrating them is the model an important step in the project to reach a better accuracy. We call this step "Feature selection". We can find many different algorithms to use but we chose "Boruta Feature Selection"[4], when working with biomedical data, this algorithm is the most practical because in this type of data we collect a lot of measurements (AVC prevalent, HYP prevalent, Tot Cholesterol…) and that makes it hard to know which one is more important for each illness. Boruta [5] is based on a simple idea and it works in the following way:

- First, it adds randomness to the dataset by developing scrambled copies of all attributes, we call them shadow features.
- Then, it trains a random forest classifier on the new dataset and applies an attribute importance measure to rate the importance of each feature where an elevated score means it's an important feature.
- At every iterance, it verifies whether a real feature has more importance than shadow features and persistently removes attributes which are considered unimportant.
- Finally, the algorithm stops after checking all features.

4. Models: Machine Learning algorithms

After choosing the most important features in the dataset, we can start building our models based on different Artificial intelligence algorithms. In this article, we chose to compare three supervised learning algorithms: K-nearest neighbors, Support vector machine, and Logistic regression. The choice of these models is based on exciting systems [6].

4.1. **: K-nearest Neighbors**

The supervised machine learning technique K-nearest neighbors (KNN) [7] can be utilized for classification and regression tasks. The KNN algorithm will generate a prediction based on the full dataset. Indeed, the algorithm will look for the K dataset examples that are closest to our observation for the trait we wish to predict. The method will then calculate the value of the variable y of the observation that we wish to predict based on the output variables y of these K neighbors.

In this article, we used this algorithm because it's easy to implement and also doesn't require a long training period. The accuracy of KNN in our project is 82,45%, the F1 score is 82,25% and the AUC is 83,72%.

4.2. **Logistic Regression**

In the field of machine learning, logistic regression [8] has become a significant technique. This method allows the machine learning application to classify incoming data using an algorithm based on historical data. The system is better able to predict classifications inside datasets with more relevant data input. A logistic regression model examines the relationship between one or more independent factors to forecast a dependent data variable. In contrast to linear regression, logistic regression predicts the outcome as probability of the default class. Therefore, the result belongs to the interval [0,1]. It is between 0 and 1, and since this is a probability, the output value of y is generated by changing the value of x using the logistic function [9]:

$$\mathbf{h\ (x) = 1\ /\ (1 + e^{x})} \qquad (1)$$

The threshold is used to convert this probability into a binary classification.

In this article we used this algorithm because it's easy to implement and also doesn't require a long training period. The accuracy of Logistic Regression in our project is 65,5 %, the F1 score is 62,27% and the AUC is 70,8%.

4.3.**Support vector machine**

SVM [10] is a classification problem solution. It falls under the group of linear classifiers, which use a linear separation of data and have their own method for determining the category boundary. It is required to provide training data to the SVM in order for it to find this border. The SVM will determine the most likely location of the boundary based on this data: this is the training period, which is required for any machine learning method. After the training phase was done, the SVM used the training data to find the border's intended location. In other ways, the training data helped him figure out where the border was. Furthermore, the SVM can now forecast the category of an entry that it has never seen before.

One of SVM's advantages, and this is crucial to notice, is that they are particularly efficient when there is little training data available: we observe that SVMs are significantly more efficient than other algorithms when there is little training data available. When there is too much data, however, SVM's performance decreases.

In this article we used this algorithm because it's easy to implement and also doesn't require a long training period. The accuracy of Support Vector Machine in our project is 86,64%, the F1 score is 85,75% and the AUC is 93,34%.

5. **Comparison**

In order to compare the performance of machine learning models we use the following metrics:

The accuracy score: The accuracy score, which is effectively a ratio of perfectly anticipated observations to total observations, is the most intuitive performance measure. True, accuracy is an excellent statistic, but only when you have symmetric datasets with about equal values of false positives and false negatives. As a result, you'll have to rely on other metrics to evaluate the performance of your model. SVM outperforms the other two models in this project in terms of accuracy.
The accuracy results of the three algorithms are illustrated in figure 1.

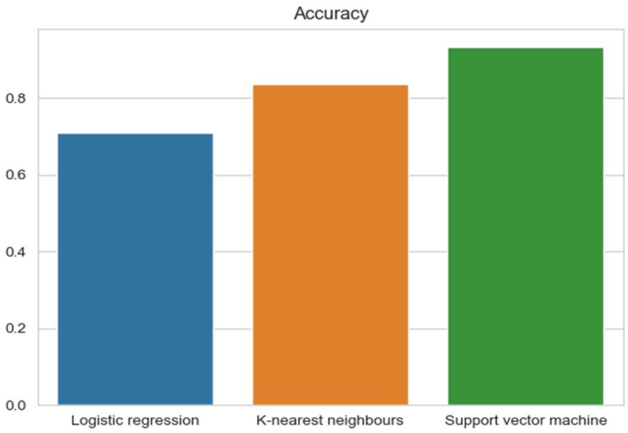

Fig. 1. The prediction accuracy of the three algorithms.

The F1 score: it's usually more useful than accuracy, particularly if you have an unequal class distribution; accuracy works better if false positives and false negatives have similar cost. If the cost of false positives and false negatives are highly different, it's better to look at both Precision and Recall because F1 Score is the weighted average of both of them. Therefore, this score takes both false positives and false negatives into account. The results of the F1 score showed the same info as the accuracy in this project: SVM has a higher score but the difference between it and K-nearest neighbors isn't so significant which is the opposite for logistic regression. The F1 score results of the three algorithms are illustrated in figure 2.

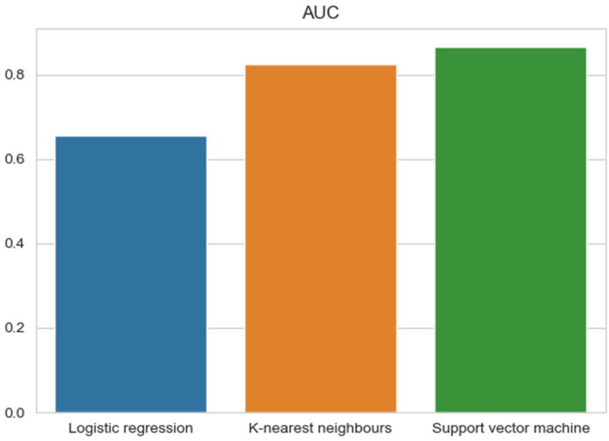

Fig. 2. The F1 score of the three algorithms.

The area under the ROC curve (AUC): A Receiver Operating Characteristic (ROC) curve is a graph representing the performance of a classification model for all classification thresholds. This curve plots the rate of true positives as a function of the rate of false positives. AUC measures the entire two-dimensional area under the entire ROC curve by integral calculations from (0,0) to (1,1). AUC provides an aggregate measure of performance for all possible classification thresholds. One can interpret the AUC as a measure of the probability that the model will rank a random positive example above a random negative example. The same as the first two metrics the difference between Support Vector Machine and k-nearest neighbor isn't as considerable as the one between logistic regression and the other two models. The area under the ROC curve (AUC) results of the three algorithms are illustrated in figure 3.

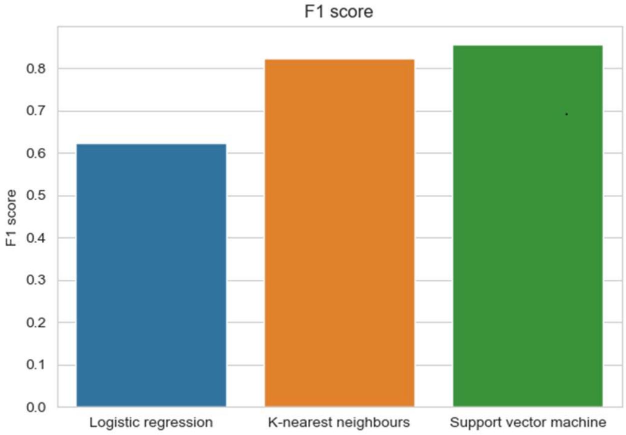

Fig. 3. The area under the ROC curve of the three algorithms.

In this article we based the comparison also on the time elapsed in the training of each model. Unlike the other metrics, support vector machine performance is very low in comparison with logistic regression and k-nearest neighbor when it comes to time elapsed during the training of the model. The table1 shows the time elapsed for the training of each model.

Model	Time Elapsed
Logistic Regression	00:00:05.69
K-Nearest Neighbors	00:00:14.25
Support Vector Machine	00:35:32.88

Table 1. The time elapsed of each algorithm.

6. Conclusion

The heart is a vital organ of the body; unfortunately, nowadays an important amount of death is caused by heart diseases. Cardiovascular diseases are one of the major challenges of healthcare worldwide. Prevention and management of cardiovascular diseases require a pervasive and comprehensive system for recording data. Information of patient records is one of the most important data, which must be classified for an easy and fast treatment process.

The goal of the current study was to provide the best reliable algorithm for predicting cardiovascular diseases. The logistic regression algorithm have the worst performances but also the shortest time of training, in the other hand K-nearest neighbor algorithm have a good performance and the training time was short too, finally Support vector machine algorithm have the best performances yet the training time was too long. Consequently, the optimal algorithm for our problem is **K-nearest neighbor**.

However, it is vital to note that the success rate of these models is dependent on a number of parameters, and selecting one strategy as the best is insufficient. The type of data in the database, the number of properties/risk factors, the size of the database, the low amount of missing (null) data, and access to proper and correct data all help to improve the performance of algorithms.

This study can be further extended by improving the algorithms results by using other algorithms like self-adaptive neural networks. These results can be exploited in creating technologies for the accurate prediction of heart disease in hospitals. It can improve the capabilities of traditional methods and decrease human error, by this way we make a contribution to the science of medical diagnosis and analysis.

References

1. https://www.who.int/health-topics/cardiovascular-diseases/

2. https://www.worldlifeexpectancy.com/morocco-coronary-heart-disease

3. Charlotte Andersson, Matthew Nayor, Connie W. Tsao, Daniel Levy, Ramachandran S. Vasan, Framingham Heart Study: JACC Focus Seminar, 1/8,Journal of the American College of Cardiology, Volume 77, Issue 21, Pages 2680-2692 (2021).

4. Manhar, M. A., Soesanti, I., & Setiawan, N. A.; A Improving Feature Selection on Heart Disease Dataset With Boruta Approach. Journal FORTEI-JEERI,1(1),41-48(2020) https://doi.org/10.46962/forteijeeri.v1i1.6

5. Kursa, Miron B., Jankowski, Aleksander, and Rudnicki, Witold R. 'Boruta – A System for Feature Selection': pp. 271 – 285. 1 Jan. (2010),

6. Arul Jothi, K., Subburam, S., Umadevi, V., & Hemavathy, K. Heart disease prediction system using machine learning. Materials Today: Proceedings. (2021)

7. J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), pp. 1-5, doi: 10.1109/ICCPCT.2016.7530265. (2016)

8. Khemphila and V. Boonjing, "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients," 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), pp. 193-198, doi: 10.1109/CISIM.2010.5643666. (2010),

9. https://www.kdnuggets.com/2017/10/top-10-machine-learning-algorithms-beginners.html last accessed 2021/04/21.

10. Gupta, N., Ahuja, N., Malhotra, S., Bala, A., & Kaur, G. Intelligent heart disease prediction in cloud environment through ensembling. Expert Systems, 34(3), e12207.doi:10.1111/exsy.12207. (2017)