ORIGINAL PAPER

doi: 10.5455/medarh.2020.74.39-41
MED ARCH. 2020 FEB; 74(1): 39-41
RECEIVED: JAN 02, 2020 | ACCEPTED: FEB 16, 2020

Classification Techniques for Cardio-Vascular Diseases Using Supervised Machine Learning

John Minou[1], John Mantas[1], Flora Malamateniou[2], Daphne Kaitelidou[3]

[1]Health Informatics Laboratory, Faculty of Nursing, National and Kapodistrian University of Athens, Greece

[2]Department of Digital Systems, University of Piraeus, Greece

[3]Department of Health Sciences, Faculty of Nursing, National and Kapodistrian University of Athens, Greece

Corresponding author. John Minou, MSc, PhD Candidate, Faculty of Nursing, National and Kapodistrian University of Athens, Papadiamantopoulou 123A, Greece , E-mail: iominou@yahoo.gr. ORCID ID: HTTP//www.orcid.org/0000-0001-7165-3058

ABSTRACT

Introduction: The World Health Organization has estimated that 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the developed countries are due to cardiovascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients. **Aim:** The aim of this paper is to build and compare classification techniques for cardiovascular diseases. **Methods:** The dataset contained 4270 patients and 14 attributes and it is available on the UCI data repository. The prediction is a binary outcome (event and no event). Variables of each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). **Results:** Different classifiers were tested. The SMOTE technique was used in order to solve the class imbalance. The cross-validation method was used in order to estimate how accurately our predictive models will perform. We evaluate our classifiers by using the following metrics: precision, recall, F1-score, Accuracy, AUC (Area Under Curve). **Conclusions:** Based on the resluts, the best scores have the Random Forest and Decision Tree classifiers.

Keywords: Classification, Cardio vascular diseases, SMOTE, Cross Validation.

1. INTRODUCTION

The World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases (1). Half the deaths in the developed countries is due to Cardiovascular diseases (2).

The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. On the other hand, the data mining approach provides innovation and strategy to replace voluminous information into useful data for achieving a decision. By utilizing information mining systems it needs less investment for the forecast of the sickness with more accuracy and precision (3).

2. AIM

The aim of this paper is to build and compare classification techniques for cardiovasculardiseases.

3. METHODS

The research aim of this paper is to apply and evaluate classification techniques. The classification goal is to predict whether the patient runs a risk of future coronary heart disease (CHD) in the next 10 years. For the supervised classification a dataset was used.

The dataset is publicly available, as a CSV file, on the UCI website and it is from an ongoing cardiovascular study.

It contains 4270 patients and 14 attributes.

What is the difference between variables and attributes, is a potential risk factor. There are both demographic, behavioral and medical risk factors.

The endpoint is defined as a binary outcome: there is or there is not a 10 year risk of coronary heart disease for a patient.

Demographics:

- Sex: male or female.
- Age: Age of the patient.

Behavioral:

- Current Smoker: whether or not the patient is a current smoker.
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day.

Information on medical history:

- BP Meds: whether or not the patient was on blood pressure medication.
- Previous Stroke: whether or not the patient had previously had a stroke.
- Previous Hyp: whether or not the patient was hypertensive.

• Diabetes: whether or not the patient had diabetes.

Information on current medical condition:

• Tot Cholesterol: total cholesterol level.
• Systolic BP: systolic blood pressure.
• Diabetes BP: diastolic blood pressure.
• BMI: Body Mass Index.
• Heart Rate: heart rate.
• Glucose: glucose level.

Target variable to predict:

• 10 year risk of coronary heart disease (CHD) - (binary: "1", means "Yes", "0" means "No").

First, the missing values were removed (4). Aftewards, we examined the dataset for imbalanced data. From the data exploration we noticed that the classes were imbalanced, and the ratio of patients without cardio vascular diseases and patients with cardio vascular diseases was 85:15.

The main motivation behind the need to preprocess imbalanced data before we feed them into a classifier is that typically classifiers are more sensitive to detecting the majority class and less sensitive to the minority class (5). In order to avoid overfitting and data loss the SMOTE oversampling method was used (6). This method generates synthetic data based on feature space similarities between existing minority instances (7). In order to create a synthetic instance, it finds the K-nearest neighbors of each minority instance, randomly selects one of them, and then calculates linear interpolations to produce a new minority instance in the neighborhood (8). After the SMOTE applicaltion we had a ratio of 50:50 balanced data.

Classifiers such as Logistic regression, Naive Bayes Classifier, Decision Tree, K-Means, Support Vector Machine and Random Forest were applied. Metrics such as precision, recall, F1-score, Accuracy, AUC (Area Under Curve) were used to evaluate the performance of the aforementioned classifiers (9). Ten-fold cross-validation was used to assess, and improve the acurracy of our classifiers (10). The implementation was done in Python.

4. RESULTS

According to Table 1, the highest Precision has Decision Tree with 0.79. The worst Precision has SVN with 0. Furthermore, the Decision Tree has the highest Recall, F1-score, Accuracy with 0.82,0.81,0.84 respectively. The highest AUC has the Random Forest. The classifier with the second highest metrics is Logistic Regression. Finally, the classifier withe the lowest metrics is the SVN.

	Precision	Recall	F1-Score	Accuracy	AUC
Logistic Regression	0.69139	0.68447	0.68791	0.68857	0.69
Naïve Bayes	0.71929	0.41068	0.52284	0.71929	0.62
Decision tree	0.79454	0.82637	0.81014	0.8436	0.8
KNN	0.29787	0.1	0.14973	0.82843	0.51
SVN	0	0	0	0.8301	0.5
Random Forest	0.64285	0.06428	0.11688	0.91102	1

Table 1. Classification Results

5. DISCUSSION

Most of the applied classifiers achieved a reasonable performance, except Naive Bayes, KNN and SVN. In general, there is no unique answer for this.

A threshold-based classifier may work well in many applications, but it may be the case that a more complicated system will perform better. It depends on the problem you are dealing with (11-17).

Also these classifiers were applied by using only the SMOTE oversamplig method which is a restriction of this research.

Future work includes testng the classifiers using different oversampling and undersampling methods and compare the results.

6. CONCLUSION

The cross-validation method was used in order to estimate how accurately our prdictive models will perform. We evaluate our classifiers by using the following metrics: precision, recall, F1-score, Accuracy, AUC (Area Under Curve). **Conclusions:** Based on the resluts, the bost scores have the Random Forest and Decision Tree classifiers.

- **Acknowledgments:** This work is partially supported by CrowdHEALTH project, a Horizon 2020 Programme of the European Commission Grant Agreement number: 727560 – Collective wisdom driving public health policies.
- **Author's contribution:** Each author gave substantial contribution in acquisition, analysis and data interpretation. Each author had a part in preparing article for drafting and revising it critically for important intellectual content. Each author gave final approval of the version to be published and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.
- **Conflicts of interest:** There are no conflicts of interest to declare.
- **Financial support and sponsorship:** Nil.

REFERENCES

1. WHO. Global action plan for the prevention and control of NCDs 2013 - 2020. World Health Organization, Geneva. 2013.
2. Cooney MT, Dudina AL, Graham IM. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. J Am Coll Cardiol, 2009; 54: 1209-1227.
3. Japkowicz, N. Assessment metrics for imbalanced learning, in Imbalanced Learning: Foundations, Algorithms, and Applications. Wiley IEEE Press. 2013; 187-210.
4. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms, IEEE Transactions on Knowledge Data Engineering. 2005; 17: 299-310.
5. Elrahman SM, Abraham A. A Review of Class Imbalance Problem, Journal of Network and Innovative Computing 2013; 1: 332-340.
6. Garcia V, Sanchez J S, Mollineda RA. On the effective of preprocessing methods when dealing with different levels of class imbalance. Knowledge-Based Systems. 2012; 25: 13-21.
7. Kerdprasop N, Kerdprasop K. On the Generation of Accurate Predictive Model from Highly Imbalanced Data with Heuristics and replication Technologies, International Journal of Bio-Science and Bio-Technology. 2012; 4: 49-64.
8. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. Journal of arti-

ficial intelligence research. 2002; 16: 321–357.
9. Han J, Kamber M, Pei J. Data Mining Concepts and Techniques. San Francisco. CA: Morgan Kaufmann Publishers, 2011.
10. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Second Edition, Morgan Kaufmann Publishers 2005: 162-169.
11. Kyriazis D, Autexier S, Boniface M, Engen V, Jimenez-Peris R, Jordan B. et al. The CrowHEALTH Project and the Hollistic Health Records: Collective Wisdom Driving Public Health Polices. Acta Inform Med. 2019 Dec; 27(5): 369-373. doi: 10.5455/aim.2019.27.369-373.
12. Magdalinou A, Mantas J, Montandon L, Weber P, Gallos P. Disseminating research Outputs. The CrowdHEALTH Project. Acta Inform Med. 2019 Dec; 27(5): 348-355. doi: 10.5455/aim.2019.27.348-355.
13. Malliaros S, Xenakis C, Moldovan G, Mantas J, Magdalinou A, Montandon L. The Intergrated Holistic Security and Privacy Framework Deployed in CrowdHEALTH Project. Acta Inform Med. 2019 Dec; 27(5): 333-340. doi: 10.5455/aim.2019.27.333-340.
14. Perakis K, Miltiadou D, De Nigro A, Torelli F, Montandon L, Mantas J. et al. Data Sources and Gateways: Design and Open Specification. Acta Inform Med. 2019 Dec; 27(5): 341-347. doi: 10.5455/aim.2019.27.341-347.
15. Wajid U, Orton C, Mogdalinou A, Mantas J, Montandon L. Generating and Knowledge Framework: Design and Open Specification. Acta Inform Med. 2019 Dec; 27(5): 362-368. doi: 10.5455/aim.2019.27.362-368.
16. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Second Edition, Morgan Kaufmann Publishers 2005; 162-169.
17. Minou J, Mantas J, Malamateniou F, Kaleteidou D. Health Professionals Perception About Big Data Technology in Greece. Acta Inform Med. 2020 Dec; 28(1): 48-51. doi: 10.5455/aim2020.28.48-51.