Classification Techniques for Cardio-Vascular Diseases Using Supervised Machine Learning

John Minou¹, John Mantas¹, Flora Malamateniou², Daphne Kaitelidou³

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 results, 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 cardiovascular diseases.

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.
RESULTS

Classification Techniques for Cardio- Vascular diseases Using Supervised Machine Learning

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 with the lowest metrics is the SVN.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.69139</td>
<td>0.68447</td>
<td>0.68791</td>
<td>0.68857</td>
<td>0.69</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.71929</td>
<td>0.41068</td>
<td>0.52284</td>
<td>0.71929</td>
<td>0.62</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.79454</td>
<td>0.82637</td>
<td>0.81014</td>
<td>0.8436</td>
<td>0.8</td>
</tr>
<tr>
<td>KNN</td>
<td>0.29787</td>
<td>0.1</td>
<td>0.14973</td>
<td>0.82843</td>
<td>0.51</td>
</tr>
<tr>
<td>SVN</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.8301</td>
<td>0.5</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.64285</td>
<td>0.06428</td>
<td>0.11688</td>
<td>0.91102</td>
<td>1</td>
</tr>
</tbody>
</table>

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 oversampling method which is a restriction of this research.

Future work includes testing 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 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 results, the best scores have the Random Forest and Decision Tree classifiers.

REFERENCES

8. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. Journal of arti-


