OKAN ÜNİVERSİTESİ

Heart disease prediction using machine learning algorithms.

A paper reviewed by ***Mr. Chahyaandida Ishaya***

Student ID: 253307014

Program Code: 3307010

Instructor: Assoc. Prof. Dr. Evrim Guler

This report is prepared as part of the coursework for **AIE504 – Machine Learning**, Department of Computer Engineering, **MSc in Artificial Intelligence Engineering**, Istanbul Okan University. The aim of the report is to review the article *“Heart Disease Prediction Using Machine Learning Algorithms”* by Harshit Jindal et al. The review focuses on the study’s originality, objectives, methodology, findings, practical relevance, and its contribution to the broader field of machine learning in healthcare.

Summary

The paper proposes the Effective Heart Disease Prediction System (EHDPS), which employs supervised learning machine learning classification algorithms such as logistic regression, k Nearest Neighbors (KNN), and Random Forest to classify patients with heart disease or not. The machine learning classification models are trained on the heart disease database from UCI, which consists of 303 to 304 patient samples with 13 to 14 attributes such as age, sex, chest pain, resting blood pressure, cholesterol, and blood sugar. The performance is said to be 87.5% overall, with KNN giving a result of 88.52%, hence more accurate and cost-effective than previous single systems.

Author(s):
Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, and Preeti Nagrath.

Journal/conference
IOP Conference Series: Materials Science and Engineering (ICCRDA 2020 proceedings).

Year of publication:
2021

Problem/Research Objective

The central research problem is how to predict, with reasonable accuracy and low cost, whether a patient is expected to have heart disease. The central goal is to develop a machine learning-based decision support system that is capable of assigning a patient to "heart disease" vs "no heart disease" based on learning from past patient instances. The secondary goals are to compare the performance of different classifiers on the same database, with a secondary goal of proving that the use of more attributes and multiple models is a way to increase the accuracy of heart disease prediction from previous models.

Methodology used

Data Source & Attributes:

The dataset employed in this research is the UCI heart disease dataset, which has approximately 303 to 304 individuals, with 13 to 14 attributes per instance that fall under the category of medical attributes. Such attributes used in this dataset include demographics (age, gender), symptom-related attributes (chest pain type), physiologic measures (resting blood pressure, serum cholesterol, fasting blood sugar), as well as other indicators for cardiovascular disease, which are considered standard attributes within the UCI heart disease dataset. Every instance is linked with a binary attribute that defines whether an individual has heart disease.

Preprocessing & Data Preparation

The methodology has multiple steps, which include the retrieval of the dataset from the UCI repository, the extraction of the most relevant variables, preprocessing, dividing the dataset, training, as well as evaluation. The preprocessing includes the treatment of missing values, cleaning, as well as normalizing, which is necessary for the KNN algorithm. The dataset is divided to conduct the evaluation on the trained models.

Classification Algorithms

Three supervised learning models have been implemented.

- Logistic Regression: This is a supervised learning algorithm that uses a weighted sum of inputs to predict the probability of heart disease, with the result being a binary classification based on a threshold.

- K Nearest Neighbors (KNN): This is a non-parametric, instance-based learning algorithm, which classifies a patient on the basis of the majority class amongst the top k most similar instances, on the basis of a distance metric in the feature space.

- Random Forest: An ensemble learning method that is a combination of multiple decision trees trained on bootstrapped samples of the dataset, with predicted outputs obtained via majority votes from the trees.

All the classifiers are trained on the same preprocessed training set, which helps make a fair comparison with respect to the test set.

Assessment

Working Principle:

Performance is assessed mainly on the basis of accuracy, which is calculated as the ratio of correctly classified patients in the testing set. The mechanism of the system is quite simple; that is, when a patient's attributes (age, chest pain, blood pressure, cholesterol, etc.) are put into the system, the result is a predicted class, expressed in the form of a binary variable, which helps in taking a call on whether the patient is at risk of heart disease or not. The results are also represented in the form of graphical representations, which helps in identifying the predicted risks with respect to age, resting blood pressure, sex, and type of chest pain.

Key findings or conclusions

An important result here is that KNN and logistic regression result in a better performance on this particular problem compared to random forest, with KNN resulting in the highest value of approximately 88.52% accuracy. The overall system is also able to reach an average accuracy of approximately 87.5%, which is reported to be greater than certain previous heart disease prediction systems that used only one classification technique or fewer attributes. The authors conclude that when machine learning is used on a structured clinical database, it is capable of efficiently predicting heart disease, presenting a cost-effective tool for its early prediction. In addition, the authors point out that the use of a more complete set of attributes from the medical domain and multiple algorithms raises the accuracy level of heart disease prediction.

Critical Analysis

Strengths:

- Motivation and relevance: The manuscript is very motivated by the high prevalence of cardiovascular disease, and preventing cardiovascular disease with the help of early low-cost risk prediction is a highly appealing goal.
- Use of a standard benchmark dataset: The use of a UCI heart disease dataset helps compare the research with past research, which helps embed the research within a known stream of research.
- Multi Algorithm Comparison: Comparison of logistic regression, KNN, and random forest on the same preprocessed dataset is a useful baseline comparison that indicates that different models have different strengths.
- Simple, reproducible pipeline: The overall structure of the work, from data acquisition to splitting, training, and evaluation, is fairly simple and easy to reproduce, following typical ML practice.
- Interpretive visualizations: Graphs displaying the distribution of risks with regard to age, blood pressure, gender, and chest pain can serve as a means of interpreting the results of models on a more manageable level.

Weakness:

- Small dataset and lack of generalizability: The sample size (approximate number of patients is 303 to 304) is small, as it consists of a single public dataset, which lacks validation for external use.
- Limited evaluation metrics: The evaluation is mainly concerned with accuracy, but other metrics such as sensitivity, specificity, precision, recall, ROC AUC, and calibration are not entirely considered, which is a problem when dealing with a medico-administrative classification problem where the cost of false negatives and false positives is different.

- Limited detail on methodology is included: In some aspects of the methodology, specifics such as how the split of train and test sets is determined (hold-out vs. cross-validation), the choice of hyperparameter values for KNN (value of k), and the specifics of the random forest (number of trees, depth), are not elaborated on in detail.
- Lack of comparison with clinical benchmarks: The research fails to compare the performance of the model with a basic clinical risk score or expert assessment, making it difficult to gauge the value addition of the research to the existing clinical solutions.
- Without uncertainty or statistical testing, confidence intervals concerning comparing the performances of models are not specified, casting a shadow of doubt on whether differences (KNN vs. logistic regression, for instance) are significant.

Validity and reliability of the result

The internal validity of the result is adequately supported by the use of a common dataset, rudimentary preprocessing, and a standard training/test split. The use of a single dataset with accuracy as the top criterion, though, obstructs the path to a complete determination of modeling validity within real-world practice, for which class distinction, error cost, and patient variability are paramount. It is a matter of partial support pertaining to the use of common models, but a complete description is impeded, making full reproduction tricky. The external validity is obstructed by a lack of assessment on distinct patients or outside hospitals.

Contribution and relevance

This paper is a contribution to the area of medical data mining, specifically relating to heart disease, in that it gives a real-world application of classical machine learning algorithms on predicting heart disease. It is a useful reminder that even simple models such as logistic regression, KNN, etc. are capable of competitive results on a structured clinical dataset, which is useful in a real-world setting. This is also a useful extension to the literature on the UCI heart disease statistics dataset, which has been investigated by a number of different comparisons involving a variety of different learning systems.

Practical or theoretical value

In terms of real-world application, what this paper demonstrates is how a low-cost, data-informed tool might be used to assist in patient triage with respect to who might be at a higher risk of heart disease. The need for common clinical factors indicates that such a system might be used in a setting that lacks sophisticated means of diagnosis, though a real-world application is not demonstrated within the given paper. In terms of theoretical application, the development of this research does not bring with it a set of novel algorithms, but what is supposed is a teaching tool for how a particular type of supervised classification might be approached within a healthcare setting.

Relevance to your course, research, or real-world problems

In terms of relevance for a machine learning or data mining class, the paper is a solid case study because it includes problem formulation, description of the dataset, preprocessing, choice of models, evaluation, and conclusions within a real-world application area. For a research article on health informatics or machine learning, it is a baseline that can be improved upon with additional models, metrics, or validation in the real-world healthcare area by applying existing patient information to develop a support system that has the potential to decrease cost with faster diagnoses although further validation is necessary before application to healthcare.

Conclusion

On the whole, "Heart disease prediction using machine learning algorithms" is a practical, application-driven research that shows how common supervised learning tasks can be combined within a basic heart disease predicting system. The biggest advantages of the research are obvious motivations, a simple research approach, a common dataset, and comparisons involving multiple algorithms, which together make it a practicable teaching tool, as well as a basic starting point for research on heart disease predictions with ML. The biggest drawbacks are a small dataset size, a small set of evaluation metrics, and a lack of external validation, which are hindrances to making this research a highly rigorous clinical research.

Recommendations

For the purposes of improvement, a number of directions may be proposed:

- Extended evaluation: Future research should include evaluation metrics such as sensitivity, specificity, ROC AUC, precisionrecall curves, and calibration, and analyze the impact of varying thresholds on clinical outcomes.
- Larger and more diverse data sets: Having a multi center or multi country, rather than single center, data set, or combining electronic health record databases, would increase the generalizability of the findings.
- The use of more stringent validation methods, such as k fold cross validation, nested cross validation for hyperparameter tuning, and evaluation on entirely separate, external evaluation sets, would increase the support for robustness.
- Advanced models with interpretability: Evaluations of classical models with modern models such as gradient boosting, deep learning, and interpretability models (like SHAP, LIME, etc.) might bring improvements in performance while also keeping interpretability on track.

- Clinically integrating the system, along with user studies: Such a system, when tested in a real-world setting, involving user studies with doctors, would bring the research from the proof-of-concept stage to a stage that is deployable.