Predicting Heart Disease Risk – Asssessment 3 Team
Project Submission - Machine Learning

Group 5

Contributors:

Lovet Ndialle

Eugenio Lavarone

Samuel Vierny

Rahul Kachhara Suresh Chandra Kachhara

Course Title:

Machine Learning

Professor:

Moez Ali

Report: Predicting Heart Disease Risk Using Machine Learning

Introduction

Cardiovascular disease remains a paramount threat to public health due to being amongst the main killer diseases across the globe. Early screening and diagnostics reduce the disease burden and improve patient quality of life. As more healthcare data emerges, ML provides strong tools to estimate the chances of heart disease by dissecting the intricate relationships of clinical and demographic data. This paper explores how supervised and unsupervised ML models can predict heart disease risk. Focusing on maximizing model accuracy, particularly in terms of AUC scores, this work defines the factors that ensure the creation of the most suitable model for implementation for real-world applications, using such models as Logistic Regression, Random Forest, and Gradient Boosting. This report also examines ethical challenges, including data confidentiality and bias. The goal is to design an ethical, reliable, predictive solution for this essential health care issue (Ahmad & Khan, 2020).

Problem Selection and Relevance

Heart disease prediction is vital as cardiovascular diseases are increasing and pose significant health and economic consequences. The most important implication of the findings is that early identification ofthe at-risk persons can occur with a view of minimizing the mortality rates and costs of the health services (Rajkomar et al., 2019).

The above models can help hospitals focus on vulnerable groups and help in taking preventive measures. Actuaries can accurately evaluate the likely demographic conditions with precise coverage programs for individuals, which can be provided by insurance companies (Han & Lee, 2021).

Heart disease prediction encourages lifestyle changes and preventive care. Further, using machine learning in this context minimizes clinical evaluation thereby improving the diagnosis accuracy while minimizing the resources needed. This research work is relevant to major concerns in ICT, healthcare, and society, as it is used to explain how the ML poses the society to work towards solving various health challenges affecting the globe (Han & Lee, 2021).

Data Collection and Preparation

Data for this study was obtained from Kaggle, an open data repository, specifically https://www.kaggle.com/datasets/abdmental01/heart-disease-dataset. There are various clinical and demographic parameters: patient age, cholesterol levels, blood pressure, and exercise-induced angina, which are important for heart disease prognosis.

The data cleaning and preprocessing steps were implemented in order to improve the quality of data. We managed missing values using the imputation procedure, and outliers using statistical methods. Gender and chest pain type, for instance, were converted from categorical data into numerical data; numerical data, on the other hand, were normalized. Such steps are crucial to ensure the derived dataset is clean and free from noise and inconsistencies that affect the creation of Machine learning models Feature scaling and encoding improved model performance and applicability (Deo, 2015).

Supervised Learning Models

Supervised learning where the training data is labeled was the focus of all methods used in this study. Three models were developed and evaluated: Of these three classifiers; Logistic Regression, Random Forest, and Gradient Boosting, Random Forest showed the best performance.

Logistic Regression

The accuracy of this model was 68.48%, the macro F1-score was 0.44 when using the binary approach, and the ROC-AUC estimate was 0.88. While Logistic Regression provided vital results with the majority classes, it poorly classified the minority classes, hence the unbalanced performance of the method across different categories of heart diseases. However, because the model is linear it cannot capture higher-order interactions well.

Random Forest

However, it was Random Forest that demonstrated the best results in general with 69.02% of accuracy, 0.47 of macro F1-score, and 0.88 of ROC-AUC. It was clear that this ensemble learning technique had successfully managed the imbalance data and gave acceptable performance in both majority and minority classes detection. Its ability to provide feature importance levels was a key contribution towards the analysis.

Gradient Boosting

This model achieved accuracy of 66% and macro F1 score of 0.45 and ROC-AUC of 0.87. Gradient Boosting captured non-linear relationships well, though it required more computational time. Despite slightly lower performance, its tunable hyperparameters make it promising for future research (Ahmad & Khan, 2020).

Unsupervised Learning

K-means clustering was used to classify unlabeled data. This approach is intended to separate patients into unique categories in terms of clinical characteristics.

K-means algorithm used in this study divided the dataset into number of clusters where intra-cluster variance was minimized. The number of clusters was chosen using a threshold-based method, 'the elbow method', resulting in three optimal clusters.
However, the silhouette score of 0.167 revealed a weak cluster separation in the data, wherein every cluster is evaluated based on its proximity to other clusters.

Data Privacy and Security

In this study, no private data was captured in the dataset used hence minimizing the risk to individual privacy. Some of the recommendations include the anonymization of patient data and encrypting the data during storage and through the transfer process.

Algorithmic Bias

Machine learning models can inadvertently reflect biases in their training data, leading to unfair predictions. For instance, underrepresentation of certain demographics, such as minority groups or specific age brackets, can skew results.

In this study, measures such as class balancing were applied to address imbalances in the dataset. Re-sampling techniques, such as oversampling, ensured that minority classes were better represented in the training process. While these efforts improved class balance, fully eliminating bias remains a challenge, especially given the inherent disparities in healthcare data.

Regarding negative effects, biased information in an algorithm can be a serious problem where

healthcare is employed: a patient can be diagnosed with a disease or prescribed different medicine due to the model's inherent biases. Testing on diverse populations and regular reviews are vital to reduce algorithmic bias.

Transparency and interpretability.

The interpretability of machine learning models is crucial in healthcare, where predictions influence critical decisions. Transparent models like Logistic Regression were used in this study to provide insights into how features such as cholesterol and blood pressure contribute to predictions.

To enhance the interpretability of more complex models, such as Random Forest, feature importance scores were analyzed. Future implementations could integrate advanced interpretability tools like SHAP (Shapley Additive Explanations) to quantify the impact of each feature on individual predictions. While SHAP was not directly applied in this study, its integration in real-world deployments can improve model trustworthiness.

It is essential, especially in healthcare, that machine learning systems be interpretable, as their decisions directly affect patients. The prediction must be made under such a way that all clinicians and patients who are stakeholders in the system understand how thepredictions are made.

As it was observed, interpretable models like Logistic Regression offer explanations about information importance, including cholesterol levels and blood pressure, as factors that directly affect predictions. Regarding other decision types, variable importance measures and plots together with different Random Forest extensions were used to explain decisions for the more complicated models.

When deployed, the interactive dashboards can display the forecasts with confidence intervals and more importantlyprovide explanations aiding clinicians to cross-check the results besides medical knowledge.

Transparency helps fill the gap between technology and clinical practice by creating trust in machine learning systems while working to ensure they are incorporated responsibly into healthcare practice.

Fairness and Equity

There is, as a result, a need for possessing fairness and equity in the administration of machine learning in health care. They should work well for all patients they can encounter of different ages, gender, ethnic backgrounds or any other status.

Fairness checks were performed on stratified subsets, though data from ethnic minorities remained limited. Subsequent subgroup analysis aided in discovering the unfairness.

In real-world application, vast testing of models must be done to test how they do not worsen inequity in accessing or receiving proper health care. Community collaborations and increasing heterogeneity of involves can improve the balance of considerations.

Societal Impacts

The implications that arise from using machine learning to predict heart diseases in the society are exciting and have their drawbacks. On one side, we have a perfectly legitimate use of predictive models to provide individuals with information that allows them to act preventively and increase their chances of leading a healthy life. At the same time, the work with automated systems can have a number of negative outcomes, for example, anxiety or an excessive number

of prescribed treatments.

In setting recommendation algorithms into health care delivery systems, practitioners' judgment must be checked and balanced by artificial intelligence or well-developed systems to prevent or minimize the reliance of clinicians on the recommendation algorithms rather than using them as tools to assist in developing their own conclusions.

Notes and Recommendations by Implementation Area

Implementation of the heart disease risk prediction model in actual clinical practice scenario brings questions and tasks of different levels of difficulty. These are crucial to making the system realistic, moral, and efficient in producing the desired results without compromising on the patienteternal and the clinical relevance.

Organization Integration with Clinical Workflows

The implementation needs to fit into currently established health systems with reference to integration of EHRs. Patient data is kept in EHR systems, and they are an ideal environment for models to run. Integration makes the outcomes available to the healthcare providers at a time they are making decisions on what precision is needed. To do this, the application programming interfaces (APIs) may allow the model and the EHRs to interact in real-time.

This means that there must be an easy-to-understand visual representation or front end that people are able to use. Clinicians should be able to understand what exactly the model gave them, with reasons as to why that particular result was obtained. For instance, if the model shows that a patient is at high risk of coming down with heart disease, information on other features thatlead to the disease such as cholesterol levels, ECG results should be provided.

Scalability and Infrastructure

Scalability is a key attribute in many software systems as organizations strive to cater for increasing complexity of customer requirements and ever-demanding customer needs In any software system, infrastructure is another important characteristic that is central for supporting scalability.

One major concern, therefore, must be scalability, because the systems are likely to be applied to healthcare facilities that cater for a vast population of patients. A cloud-based architecture is sustainable in that it provides the means for expanding capacity and usage of the system while allowing many healthcare providers to access the system from multiple locations while data and processing can be securely stored and completed from off-site locations.

Data Privacy and Security

The information that pertains to patients is extremely sensitive, that is why such factors as privacy and security should be among the leading priorities in any deployment. Technological systems must consider regulations including the Health Insurance Portability and Accountability Act (HIPAA) for the United States and the General Data Protection Regulation (GDPR) in the EU. It is recommended that data is protected when it is being transmitted and when it is stored. Furthermore, it will be necessary to intersect access control with the limitationsof data access to only the highly authorized personnel.

Data anonymization is recommended, although this can decrease the detail of the finding and reduce patients' trust.

Feedback Loops and Stakeholder Involvement

Different groups need to provide feedback in order to be incorporated in the system, which is an

important aspect that has to form part of the deployment strategy. Both clinicians and patients should haveways of feeding back the validity, operational usefulness, or outcomes of the model back to the collaborators or developers. This can help inform subsequent refinements to the system, which can help to maintain the systems relevance to the users.

Cost and Resource Management

Deployment requires balancing predictive accuracy with resource efficiency, such as minimizing server usage through edge computing to reduce costs and enhance response times.

Professional Ethics in Communication with Patients

Clear communication is essential to ensure patients understand that the model supports clinical decision-making but does not provide conclusive diagnoses.

Regulatory Approvals

Regulatory approval is essential before deployment, and early collaboration with authorities can help avoid delays and ensure compliance with safety and effectiveness standards.

Contribution of Team Members

The successful completion of this project relied on the collaborative efforts of the team members, each of whom brought their unique expertise and dedication to various aspects of the assignment:

Lovet Ndialle: Lovet Ndialle also contributed significantly to data collection and data preparation category. They obtained the dataset from Kaggle, remain ethical in handling the data properlyand performed a variety of preprocessing where they dealt with missing values, outliers, normalization, and encoding of features.

Samuel Vierny: Developing and evaluating the supervised learning models that were used by Samuel Vierny who mainly concentrated on model's construction and optimization. The methods they used include Logistic Regression, Random Forest and Gradient Boosting the hyperparameters they tuned were accuracy, F1-score, and ROC-AUC.

Eugenio Iavarone: Eugenio Iavarone contribution was made in uncovering unsupervised learning approach. I used the K-means clustering in an attempt to capture the pattern in the dataset and produced statements in regards to patient grouping. I also assessed the clustering outcomes and debated about disadvantages of this method in the framework of heart disease prediction.

Rahul Kachhara: In the assessment of ethical considerations, responsibilities, and obligations Rahul K took the initiative. All the teams provided extensive sections on data privacy and security, several contributed to the section on the 'black box' problem encompassing the lack of transparency and interpretability of the algorithms used in big data, analysis of bias, fairness and equity issues, and what the impact of big data and analytics means to society.

Conclusion

In this paper, we have developed a machine learning model to estimate the likelihood of heart diseases to the attendants and those involved in giving out health advice. Out of all the performed models, the Random Forest algorithm was the most accurate and had the best stability rate. Nevertheless,how such models are implemented must address the following ethical concerns: data privacy, fairness of algorithms, and model interpretability.

In a real-world setting, proper deployment would entail designing interfaces for ease of use, incorporation of the model in real world clinical scenarios and more so, the on-going evaluation of the model in various clinical contexts.

In a much broader context this research highlights the strength of applying machine learning inpractice of medicine. Thus, with appropriate use of technology it is possible to increase the identifying of heart disease at an early stage, decrease the costs of health treatment, and hencecreate a better future for patients and healthier society.

Word Count: 2410

References

1. Chollet, F. (2018). Deep learning with Python. Manning Publications.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
3. Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787. https://doi.org/10.1016/j.jbankfin.2010.06.001
4. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774). https://doi.org/10.48550/arXiv.1705.07874
5. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. The New England Journal of Medicine, 375(13), 1216-1219. https://doi.org/10.1056/NEJMp1606181
6. Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
7. Yeo, C. K., Ho, T. T., & Goh, J. M. (2019). Leveraging machine learning for early detection of heart disease. International Journal of Medical Informatics, 129, 233–243. https://doi.org/10.1016/j.ijmedinf.2019.06.002