Predicting Heart Disease Risk Using Machine Learning

Group 3

Contributors:

Lovet Ndialle

Eugenio Lavarone

Samuel Vierny

Rahul Kachhara Suresh Chandra Kachhara

Course Title:

Professor:

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. Screening and pre-school diagnostic help also dramatically decrease the disease load and results in patients' better quality of life. As more healthcare data emerge, ML provides strong tools to estimate the chances of heart disease by dissecting the intricate relationships of clinical and demographical data. This paper aims to examine how the benefits of applied prognostication can be utilized to generate and apply effective heart disease risk models primarily focusing on supervised and unsupervised ML methodologies. Unified on the grounds of accuracy at the AUC level, 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. Not restricted to the practical aspect, this report also focuses on the use and misuse of ML services in healthcare and their legal challenges comprising data confidentiality and bias. The ultimate goal is to design an ethical, reliable, predictive solution for this essential health care issue (Ahmad & Khan, 2020).

Problem Selection and Relevance

Prediction of heart disease is an important issue in modern health care since the incidence of cardiovascular diseases is growing, and their consequences are critical for people and the healthcare industry. The most important implication of the findings is that early identification of the 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 be of help in to hospitals as they help in focusing the vulnerable groups to take preventive measures. Actuaries can accurately evaluate the likely demographic conditions with an individual and precise coverage programs hence can be provided by insurance corporations (Han & Lee, 2021).

Heart disease prediction ensures that all the people in societies receive adequate measures for preventing the diseases as well as encouraging them to change their lifestyle. 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 one of the major concerns in the 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. They are different clinical and demographical parameters: patient age, cholesterol levels, blood pressure, and perineal 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. The processed dataset has been saved in the name of heart_disease_cleaned.csv and it has used for model training and model evaluation.

Such steps are crucial that the derived dataset is clean and vascular from noises and inconsistencies that affects the creation of Machine learning models. It was also discovered that feature scaling and encoding made it easier to choose and implement different algorithms to increase prediction ratings and model applicability (Deo, 2015).

Model Development and Evaluation

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 enjoyed the best performance level.

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 threw 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. Because of it is highly interpretable, it can be used to determine how each feature influence the predictions However, it is linear and thus 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 particular

ensemble learning technique had successfully managed the imbalance data and gave acceptable performance in both majority and minority classes detection. That it was able to provide the levels of feature importance was a key addition towards the analysis.

Gradient Boosting

This model reached accuracy of 66% and macro F1 score of 0.45 and ROC-AUC of 0.87. Gradient Boosting outperformed the other algorithms in capturing the non-linear relationships though in doing so used more computation time. While itself is slightly lower in performance, the ability to effectively tune the hyperparameters makes it a better candidate for future research further enhancing the model (Ahmad & Khan, 2020).

Unsupervised Learning

The technique of unsupervised learning used in the study but it seeks to classify data which have no labels and K-means clustering was used. This approach intended to segregate 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 decided based on the threshold method and it enumerated three clusters as optimal as they are compact and well separated. However, the silhouette score of 0.167 revealed that there exists only a weak discrimination in the structure of data wherein every cluster is evaluated based on its proximity to other clusters.

Based on the results of cluster analysis, some trends were identified, for example, clusters based on the value of cholesterol and the age distribution of patients, but these trends were not particularly very different from each other, so they did not yield practical information. However, unlike many other supervised learning models, unsupervised learning models such as the K-means algorithm need a lot of preprocessing before results can be obtained such as scaling and feature selection (Deo, 2015; Uddin et al., 2019).

Ethical Implications

Data Privacy and Security

In this study, no private data was captured in the dataset used hence minimizing the risk to individual privacy. However, when it comes to similar models in an operational healthcare environment, protection of patient data that should not be shared is a critical ethical issue. Some of the recommendations include the anonymization of caller data and encrypting the data during storage and through the transfer process. Although, encryption techniques were not used in this study, it is crucial to adhere to privacy and security regulation polices including the HIPAA across the United States and the GDPR across the European Union.

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 unfounded prejudices. To reduce such risks, it is important to make sure that models are tested on diverse population groups. Working with other domains in order to analyze results and continue model improvement contributes to fairness more.

The ethical constructive use of machine learning to increase equitable and fair healthcare outcomes needs constant review and adjustments to lessen the possibility of biased algorithms.

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

The first and continuous requirement can be explained by the subject of healthcare machine learning systems, where shown decisions affect patients. The prediction must be made under such a way that all clinicians and patients who are stakeholders in the system understand how the predictions 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. To bring added interpretability, the results were also integrated with SHAP (Shapley Additive explanations) values that used the concept of game theory to quantify exactly how much each feature contributed to a particular prediction.

The benefits gained from transparent reporting of the model include identifying and explaining the limitations, biases or performance disparities and assumptions made. When deployed, the interactive dashboards can display the forecasts with confidence intervals and more importantly provide 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.

To overcome the concerns about the fairness of the metrics, this study checked the validity of the models on strata subsets of the dataset. However, some issues regarding feature representation were still a problem; for instance, a lack of much data from ethnic minorities. Recommendation of the fairly accurate algorithms and the 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. This embedded nodal knowledge can inform targeted education interventions for patients and practitioners to encourage the consistent, evidenced-based, and safe use of machine learning technologies.

Therefore, missing persons' cases require equity in the access to the numeric and analytic tools that are capable of predicting the missing individual.

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 make the system realistic, moral, an efficient in producing the desired results without compromising on the patient eternal 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 that lead to the disease such as cholesterol levels, ECG results should be provided.

Updates and New Knowledge

For clinical data, the subject domain increases with time, so that the probability of concept drift and model obsolescence that are major disadvantages of CD become considerable. In opposition to this, the deployed system has to include the means to perform updates on a regular basis. Some of these should involve training the model using new patient data to act as update in enhancing its algorithm since the test and norms change with time.

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.

Another necessary correlate is stable logistics, starting with manufacturing and ending with the delivery of ready-made products.

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 have to 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 limitations of data access to only the highly authorized personnel.

The protection of the individual patient can however be boosted by anonymizing patient's data. Although this decreases the detail of the findings some of the methods assure ethical conduct and gain patients' trust.

Bias and Fairness

Socio-spatial bias in machine learning models results in unequal treatment of demography and weakens the ethical framework of the system. Periodic evaluation should determine whether the model is giving equal weightage to genders, age, ethnic origin, and calssifier of society as the majority. To overcome this concern, it is mandatory to have transparency about how the model functions, and the limitations it has to offer to clinicians.

Feedback Loops and Stakeholder Involvement

They need to feedback into it 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 have ways 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. Furthermore, for the purpose of successful deployment, concerns of clinicians need to be tackled and their input integrated into the process.

Cost and Resource Management

Deployment must also bear the outrageous cost that will be incurred by the healthcare providers. The improvement of such models gives higher predictive accuracy, though it increases the time to employ models as well. Optimal solutions must be found between performance and resource utilization. For instance, it is possible to minimize the usage of centralized servers by means of edge computing, and therefore, decrease spending and enhance the speed of reactions.

Professional Ethics in Communication with Patients

Whenever such predictive systems are implemented, patients are bound to ask how their risk levels are being arrived at. Patient and doctor mistrust can be avoided if communication about

the model's part in the diagnostic process is clear and empathetic. The patient should be advised that the model is for the purpose of decision making and not diagnostic conclusive.

Regulatory Approvals

However, regulatory approval should be obtained before deployment at the last stage. Incidentally, deployment of medical AI systems in many regions is only possible after meeting specific safety and effectiveness standards. In preparation of what it takes to seek regulatory approval, it is important to engage regulatory authorities in this process right from the time of product development since this will help in avoiding hitches that can cause delays in the process.

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 a lot under data collection and data preparation category. They obtained the dataset from Kaggle, remain ethical in handling the data properly and 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 done 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.

Although the models explained here give good results, the authors state that more research is required in order to fine-tune and deploy these models into practice. Deployment in a real world 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 much broader context this research highlights the strength of applying machine learning in practice 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 hence create a better future for patients and healthier society.

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