Høgskulen på Vestlandet

ML Assignment #2

Harald Giskegjerde Nilsen

Sindre Kjelsrud

15.11.2023

Table of Contents

1. Describe the problem

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 32% of all deaths worldwide. Over four out of five CVD deaths are due to heart attacks and strokes, and almost 40% of these deaths occur prematurely in people under 70 years of age (World Health Organization, 2021).

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes or already established disease) need early detection and management wherein a machine learning model can be of great help.

Using a diverse range of attributes including age, sex, cholesterol, fasting blood sugar, max heart rate, exercise-induced angina, heart disease and more, we aim to elucidate the complex interplay of these variables in determining the likelihood of a CVD.

We want to enhance our understanding of cardiovascular health and pave the way for a healthier future.

1.1. Scope

About the project

In this digital age there are continuously more and more fields which quantify and store data, making it easier to always have access to the data and easier to share between people. From a medical perspective, all hospitals store information about its patients in large datasets, which allows for easy lookups and easier transfers of medical records between hospitals.

One of the other benefits of having systematically stored data in an indexable way is that it allows for a lot of research to be done on the data. Here it’s possible to draw correlations and see patterns, possibly making the jobs of doctors easier in the future. This is also a backdrop for the following task.

In this project we have a dataset which encapsulates a diverse range of patient data, including age, cholesterol levels, blood pressure, etc. Given this data, we want to try to predict whether a patient is at a larger risk of experiencing a heart attack.

Note that these kinds of predictions are being made by someone with no medical expertise or education, but with a background in computer science. What this means is that important aspects of the data and the results can be undervalued or misinterpreted - or that correlations that actually have no importance can be overstated. A project like this one is best done through interdisciplinary work.

Business objective

Achieving a model that identifies individuals at a high risk of experiencing a heart attack would be substantial, ranging from decreasing mortality rates to reducing healthcare costs for both the healthcare system and patients. Better resource allocation and patient outcomes would be the outcome of accurately predicting the likelihood of a heart attack where preventative measures can be employed.

Existing solutions and similar projects

Obtaining an overview of similar projects within this field is difficult, given how broad the field is and the amount of research being done on similar projects every day. But through a quick Google-search it's possible to read about several projects trying to predict the risk of heart attack given a similar dataset. For example Bhatt et. al. (2023) has a dataset containing patient information for 70,000 patients with 12 attributes, where they manage to achieve an accuracy of 87.28% with cross-validation using a multilayer perceptron classifier.

Today's existing solutions mainly rely on a series of medical tests and subjective evaluations by healthcare professionals. These methods are useful, but they're often time-consuming, expensive and less data-driven.

Stakeholders

- Healthcare providers, such as doctors, nurses, cardiologists, etc.
- Hospital administrators
- Data scientists and ML engineers working on the project
- Patients who are being screened
- Insurance companies

Resources required

- Data scientists and machine learning engineers
- Domain experts in cardiology for feature selection and model evaluation
- High-capacity computational resources for model training and inference
- Data storage solutions

- A team for monitoring and maintaining the system post-deployment

1.2. Metrics

Evaluating the project

Given the similar projects and their results, we aim to achieve a score that is within the same percentile as the similar projects. Takci (2018) uses accuracy as a measure and achieves 84.81 percent correct predictions. However, as we want to take both accuracy and recall into consideration, we will use the F1-score as a measure. Here Bhatt et. al. (2023) achieved a score of roughly 85 percent and Nandal (2022) achieved scores of anywhere from 85 to 89 percent. As we have a limited dataset, limited time and not nearly as much computing power as the previously mentioned projects, we will be greatly happy with a F1-score above 83 percent.

As previously mentioned, to score our models performance, we will use the F1-score. This is because it takes the average mean of the accuracy-score and recall-score. In a scenario where we want to predict if someone is more exposed to suffering a heart attack, we want to try to observe as many correct patients as possible, while at the same time not incorrectly classifying patients as at risk for heart attack. As we see in the scenario described over it directly related to the project's business objective. This is because we will most likely decrease mortality rates by observing more correct patients, as well as getting a byproduct of reducing healthcare costs for both the healthcare system and patients.

2. Data

The dataset being used for this task is obtained through the Data Science-website *Kaggle.com* (Kaggle, u.å.). The dataset contains 918 patient observations, where each patient has 12 features, including the label, which is the row 'HeartDisease'. This column indicates whether the patient is at risk for experiencing a heart attack or not.

The columns in the dataset are either numerical or strings. The columns where the data types are strings need to be handled properly so that a machine learning model can interpret them correctly. This means that when changing the data, we need to take sure we don't skew the data or create new data that misinterprets the original data. There also exists some missing values in the dataset, which needs to be taken care of.

Ethical and privacy concerns

When it comes to ethical and privacy concerns we have taken different measures. In terms of data privacy we've got a dataset that doesn't contain sensitive information from the patients as it has been de-identified. If this was not the case we would have to handle it in compliance with healthcare data standards such as GDPR in Europe. Data bias is also something we had to take into consideration as this is an issue since you need to ensure that the dataset is representative of the diverse population it's intended to serve. A biased dataset could lead to a biased model which may be less effective at, for example, predicting heart attacks in minority groups. One of our solutions for this was to choose a high k-fold for our cross-validation since a higher 'k', for example 10 that we chose, contains fewer data points which reduces the bias.

Text representation

For the columns where it's necessary to convert the data from string to numerical, we have the following columns:

- Sex
- ChestPainType
- RestingECG
- ExerciseAngina
- ST_Slope

When converting sex we only need to transform the values from 'f' and 'm' to 0 and 1. The rest of the columns only have three or four distinct data types, making them ideal for one hot encoding. Here we give each of the distinct values its own column, marking the correct one with a 1 and the rest 0. This will greatly expand the columns in our dataset, but given that it's within a reasonable bound for human interpretation, it's an approach which makes sense.

Handling zeros

In the column 'cholesterol', there are 172 patients with a cholesterol-level of zero, which isn't possible. This means that when the dataset was created, the patients with no registered cholesterol-level, got marked with a zero instead of null as a value. If we proceeded to train the machine learning-model with the dataset that we currently have, the model wouldn't flag any errors and would treat the zeros just like any other column. This isn't ideal, so to address this problem we have to convert the zeros into a value that makes the dataset less skewed and more proportionate to what it would be expected to look like.

Some of the ways we can fix this is: remove the columns with zero, replace the zeros with the mean of all the cholesterol-values or make a function that estimates the cholesterol based on other factors.

3. Modeling

First of all, the task at hand is a classification task, not regression. This means that we won't try to predict a numeric value, but rather use all the information we have to predict either *True* or *False*. Because of that we've got many different models we can test out and explore.

3.1. Model exploration

Given the nature of the problem, several machine learning models are well-suited. The models we've considered exploring are the following:

- ***Logistic Regression:*** A simple model good for an initial baseline which is easy to interpret.
- ***Naive Bayes:*** A simple model good for classification tasks due to its simplicity and efficiency handling large datasets with multiple features.
- ***Support Vector Machine (SVM):*** A more advanced model known to be effective because it can model complex relationships and nonlinearities in the data
- ***Gradient Boost:*** A more advanced model known for high performance, but may require more careful tuning.

3.2. Evaluating performance

After running the model through our pipeline we achieve a F1-score of 0.843, which is much better than we first hoped for.

4. Deployment

4.1. Model deployment

To deploy our model we use a simple gradio-interface to represent our function. The model is available through a form, when after submission the interface will tell you whether you are at risk or not. If the model were to be scaled up, we could deploy it through a web application, giving more users the possibility to use our application.

4.2. Monitoring and maintenance

Machine learning models require monitoring for several reasons, especially in healthcare. Some of these reasons are detection for anomalies or unexpected prediction outcomes, checks for input data to ensure it hasn't changed, and an ongoing evaluation against new, fresh data to check if performance has degraded in terms of F1 score. Maintenance of these reasons can be as easy as sending it to manual reviewing.

4.3. Post-deployment

To improve the system after deployment we can incorporate certain mechanisms and updates. This is done so that the system can evolve in a manner that aligns with advances in both technology and medical research; as well as adapting to the changing nature of healthcare data.

Under you'll see some examples of mechanisms and updates that will help improve the system as we described above.

- ***Feedback loop:*** To collect insights from healthcare providers on the model's performance and prediction reliability.
- ***Model update:*** Newer research or algorithms could help update, or replace, the model to be more efficient.
- ***Feature engineering:*** Additional feature engineering might be necessary depending on performance metrics and user feedback, or even new types of data.
- ***Automated retraining:*** When new data and annotations are accumulated we can retrain the model periodically.

References

- Bhatt, A. et al. (2023) 'Effective Heart Disease Prediction Using Machine Learning Techniques', Algorithms, 16(2), pp. 88. Available at: https://www.mdpi.com/1999-4893/16/2/88 (Accessed: 13 November 2023).

- Takci, H. (2018) 'Improvement of heart attack prediction by the feature selection methods', Turkish Journal of Electrical Engineering and Computer Sciences: Vol. 26: No. 1, Article 1. Available at: https://journals.tubitak.gov.tr/cgi/viewcontent.cgi?article=2005&context=elektrik (Accessed: 13 November 2023).
- Nandal, U. (2022) 'Machine learning-based heart attack prediction', F1000Research, 11, Article 1126. Available at: https://f1000research.com/articles/11-1126 (Accessed: 13 November 2023).
- World Health Organization (2021) 'Cardiovascular diseases (CVDs)', [online] Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (Accessed: 25 October 2023).