Texas A&M University - Commerce

Department of Computer Science

Heart Disease Detection Using Machine Learning

Swetha Paspunuri

Supervisor: Derek Harter, Ph.D.

A report submitted in partial fulfilment of the requirements of
Texas A&M University - Commerce for the degree of
Master of Science in *Computer Science*

February 4, 2024

Declaration

I, Swetha Paspunuri, of the Department of Computer Science, Texas A&M University - Commerce, confirm that this is my own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly.

I give consent to a copy of my report being shared with future students as an exemplar.

I give consent for my work to be made available more widely to members of TAMUC and public with interest in teaching, learning and research.

Swetha Paspunuri
February 4, 2024

Abstract

In Contemporary world Cardiovascular diseases have seen a rise, even affecting newborn. Detecting heart-related diseases prior is vital because it helps doctors start treatment sooner, leading to better results for patients and less strain on healthcare resources. With more and more people facing heart problems, it's crucial to have advanced predictive tools. Using the abundant data available in Cardiology, our project aims to integrate the technology into health care for predictive modelling. The primary goal of this project is to develop an efficient heart disease prediction system using various machine learning models to predict Coronary Artery Disease with utmost precision and effectiveness. We employed a dataset consisting of necessary patient information from online sources to train and validate our models. The first step is Cleaning and preprocessing data that allow us to find key patterns for training the models. The various machine learning models used are Logistic Regression, Random Forest, Naive Bayes. We evaluate these models using important measures like precision, which tells us how accurate positive predictions are; recall, which shows how well the models capture all actual positive cases; and the F1 score, which balances both precision and recall which can be assessed to provide the health sectors with a more efficient approach to detect heart related diseases.

Keywords: Logistic Regression, Random Forest, Naive Bayes , F1 score,Precision.

Acknowledgements

An acknowledgements section is optional. You may like to acknowledge the support and help of your supervisor(s), friends, or any other person(s), department(s), institute(s), etc. If you have been provided specific facility from department/school acknowledged so.

Contents

List of Figures

List of Tables

List of Abbreviations

SMPCS	School of Mathematical, Physical and Computational Sciences

Chapter 1

Introduction

Today, heart problems are a major health concern affecting individuals worldwide. Many people are suffering from heart issues like heart disease, heart failure, and irregular heartbeats nowadays. Heart problems can affect people, not just physically but also emotionally. Those with heart conditions often find it difficult to live normally and face many difficulties. Additionally, the financial side of managing heart problems adds an extra layer of challenges. Spotting heart-related issues early is crucial. It helps healthcare professionals to step in quickly, enhance patient outcomes, and ease the strain on healthcare resources. Early detection allows for timely intervention, potentially preventing the progression of heart conditions.To address the need for early detection, our project focuses on developing a machine learning model capable of accurately identifying the presence of heart diseases. In this endeavor, we utilize a heart-related issue dataset from (Janosi and Detrano, 1988), sourced from the online repository UC Irvine. The data undergoes thorough cleaning and pre-processing to extract useful information essential for training the machine learning model. The machine learning algorithms employed, as highlighted by (Sharma et al., 2020), include Logistic Regression, Naïve Bayes, and Random Forest classification. These algorithms have demonstrated effectiveness in detecting coronary artery disease by evaluating outputs based on various factors such as resting blood pressure, serum cholesterol, maximum heart rate achieved, and more. Furthermore, our project aims not only to detect heart-related issues but also to contribute valuable insights to the broader field of cardiovascular health. By leveraging advanced algorithms, we seek to ensure the effective prediction of heart-related problems, potentially revolutionizing the early diagnosis and management of cardiovascular conditions.

1.1 Background

Our project focuses on addressing the issue of cardiovascular diseases in today's world, affecting everyone irrespective of their age. The primary motivation behind our work is to detect the heart-related diseases as early as possible. This identification helps doctors to start the treatment sooner, to improve patient results and effectively using the healthcare resources. For this, our project focuses on integrating technology into healthcare by using the abundant data in cardiology for predictive modeling. The primary goal is to develop an efficient heart disease prediction system by concentrating on predicting Coronary Artery Disease with precision and effectiveness. So, we use different machine learning models such as Logistic Regression, Random Forest, and Naive Bayes. These models play a crucial role in predicting and understanding heart-related issues. The

initial phase involves cleaning and preprocessing the data to extract necessary patterns required for training the models. Our project becomes significant as it can give better resources to doctors for finding and handling heart problems early on. We want to help make hearts healthier by explaining some crucial ideas and ways to use them in a simple way.

1.2 Research Question

How can machine learning models, specifically Logistic Regression, Naïve Bayes, and Random Forest classification algorithms, be effectively utilized to develop a heart disease prediction system for early detection of Coronary Artery Disease, with a focus on improving patient outcomes and contributing to advancements in cardiovascular health?

1.3 Aims and objectives

Aims: To develop and implement an advanced heart disease prediction system, utilizing machine learning models for early detection of Coronary Artery Disease, with the ultimate goal of enhancing patient outcomes and contributing to the ongoing global efforts in cardiovascular health. **Objectives:** The Objectives include to Perform data cleaning and preprocessing, apply Logistic Regression, Naïve Bayes, and Random Forest algorithms, integrate technology into healthcare, provide valuable resources to healthcare professionals, and innovate existing approaches to address gaps, thereby achieving early identification and proactive management of heart-related issues.

1.4 Solution approach

The solution approach involves a thorough step-by-step method designed to create an advanced system for predicting heart disease, specifically focusing on early detection of Coronary Artery Disease (CAD), with the ultimate goal of achieving our defined aim and objectives. The project commences with the acquisition of a dataset related to cardiac issues obtained from the research conducted by Janosi and Detrano, which is accessible through UC Irvine. Following that, we carefully clean and process the data to find important patterns needed for training our machine learning models. We incorporate technology into healthcare by using different smart algorithms, like Logistic Regression, Naïve Bayes, and Random Forest classification. These specific algorithms are chosen because they are proficient in effectively predicting Coronary Artery Disease (CAD), by taking into account key factors such as resting blood pressure, serum cholesterol, and maximum heart rate achieved. The models are evaluated using specific metrics like precision, recall, and the F1 score to provide us with the understanding of how well they are performing.

1.4.1 A subsection 1

You may or may not need subsections here. Depending on your project's needs, add two or more subsection(s). A section takes at least two subsections.

1.4.2 A subsection 2

Depending on your project's needs, add more section(s) and subsection(s).

A subsection 1 of a subsection

The command \subsubsection{} creates a paragraph heading in LaTeX.

A subsection 2 of a subsection

Write your text here...

1.5 Summary of contributions and achievements

Describe clearly what you have done/created/achieved and what the major results and their implications are.

1.6 Organization of the report

Describe the outline of the rest of the report here. Let the reader know what to expect ahead in the report. Describe how you have organized your report.

Example: how to refer a chapter, section, subsection. This report is organised into seven chapters. Chapter 2 details the literature review of this project. In Section 3...

Note: Take care of the word like "Chapter," "Section," "Figure" etc. before the LaTeXcommand \ref{}. Otherwise, a sentence will be confusing. For example, In 2 literature review is described. In this sentence, the word "Chapter" is missing. Therefore, a reader would not know whether 2 is for a Chapter or a Section or a Figure.

Chapter 2

Literature Review

The literature review looks into the current state of cardiovascular disease research, highlighting its global impact across different age groups. Many recent studies show worry about heart problems like disease, failure, and abnormal rhythms, which hurt people's bodies and emotions. Identifying the health issues sooner could help the doctors to respond fast, improve the patient results, effectively utilize the health care resources. [4] Janosi and Detrano's (1988) heart-related issue dataset from UC Irvine serves as a valuable resource for our project, showing how existing data can be helpful in medicine. [5] The study by Sharma et al. (2020) reinforces the significance of employing machine learning algorithms, specifically Logistic Regression, Naïve Bayes, and Random Forest classification, for accurate detection of coronary artery disease. Studies show that things like how high your blood pressure is at rest, how much cholesterol you have in your blood (serum cholesterol), and how fast your heart beats during exercise are all important for figuring out if someone might have heart disease. Our project mainly focuses on integrating technology into healthcare, aligning with the existing literature. [6] Looking at past research on heart disease prediction shows what are the necessary improvements and different ways to predict and understand heart problems. By addressing these gaps, our project helps doctors find and treat heart problems as soon as possible, which is btter for patients and hospitals. The literature review emphasizes our significant impact in contributing to the ongoing efforts to combat cardiovascular diseases and underscores the need for advanced predictive tools to address this global health challenge.

2.1 Example of "risk" of unintentional plagiarism

Using other sources, ideas, and material always bring with it a risk of unintentional plagiarism. **MUST**: do read the university guidelines on the definition of plagiarism as well as the guidelines on how to avoid plagiarism (**?**).

2.2 Critique of the review

Describe your main findings and evaluation of the literature.

2.3 Summary

Write a summary of this chapter

Chapter 3

Methodology

We mentioned in Chapter 1 that a project report's structure could follow a particular paradigm. Hence, the organization of a report (effectively the Table of Content of a report) can vary depending on the type of project you are doing. Check which of the given examples suit your project. Alternatively, follow your supervisor's advice.

3.1 Examples of the sections of a methodology chapter

A general report structure is summarised (suggested) in Table 3.1. Table 3.1 describes that, in general, a typical report structure has three main parts: (1) front matter, (2) main text, and (3) end matter. The structure of the front matter and end matter will remain the same for all the undergraduate final year project report. However, the main text varies as per the project's needs.

3.1.1 Example of a software/Web development main text structure

Notice that the "methodology" Chapter of Software/Web development in Table 3.2 takes a standard software engineering paradigm (approach). Alternatively, these suggested sections can be the chapters of their own. Also, notice that "Chapter 5" in Table 3.2 is "Testing and Validation" which is different from the general report template mentioned in Table 3.1. Check with your supervisor if in doubt.

3.1.2 Example of an algorithm analysis main text structure

Some project might involve the implementation of a state-of-the-art algorithm and its performance analysis and comparison with other algorithms. In that case, the suggestion in Table 3.3 may suit you the best.

3.1.3 Example of an application type main text structure

If you are applying some algorithms/tools/technologies on some problems/datasets/etc., you may use the methodology section prescribed in Table 3.4.

Table 3.1: Undergraduate report template structure

Frontmatter		Title Page Abstract Acknowledgements Table of Contents List of Figures List of Tables List of Abbreviations
Main text	Chapter 1	Introduction
	Chapter 2	Literature Review
	Chapter 3	Methodology
	Chapter 4	Results
	Chapter 5	Discussion and Analysis
	Chapter 6	Conclusions and Future Work
	Chapter 7	Refection
End matter		References Appendices (Optional) Index (Optional)

Table 3.2: Example of a software engineering-type report structure

Chapter 1	Introduction	
Chapter 2	Literature Review	
Chapter 3	Methodology	
		Requirements specifications
		Analysis
		Design
		Implementations
Chapter 4	Testing and Validation	
Chapter 5	Results and Discussion	
Chapter 6	Conclusions and Future Work	
Chapter 7	Reflection	

3.1.4 Example of a science lab-type main text structure

If you are doing a science lab experiment type of project, you may use the methodology section suggested in Table 3.5. In this kind of project, you may refer to the "Methodology" section as "Materials and Methods."

Table 3.3: Example of an algorithm analysis type report structure

Chapter 1	Introduction	
Chapter 2	Literature Review	
Chapter 3	Methodology	
		Algorithms descriptions
		Implementations
		Experiments design
Chapter 4	Results	
Chapter 5	Discussion and Analysis	
Chapter 6	Conclusion and Future Work	
Chapter 7	Reflection	

Table 3.4: Example of an application type report structure

Chapter 1	Introduction	
Chapter 2	Literature Review	
Chapter 3	Methodology	
		Problems (tasks) descriptions
		Algorithms/tools/technologies/etc. descriptions
		Implementations
		Experiments design and setup
Chapter 4	Results	
Chapter 5	Discussion and Analysis	
Chapter 6	Conclusion and Future Work	
Chapter 7	Reflection	

Table 3.5: Example of a science lab experiment-type report structure

Chapter 1	Introduction	
Chapter 2	Literature Review	
Chapter 3	Materials and Methods	
		Problems (tasks) description
		Materials
		Procedures
		Implementations
		Experiment set-up
Chapter 4	Results	
Chapter 5	Discussion and Analysis	
Chapter 6	Conclusion and Future Work	
Chapter 7	Reflection	

3.2 Example of an Equation in LaTeX

Eq. 3.1 [note that this is an example of an equation's in-text citation] is an example of an equation in LaTeX. In Eq. (3.1), s is the mean of elements $x_i \in \mathbf{x}$:

$$s = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{3.1}$$

Have you noticed that all the variables of the equation are defined using the **in-text** maths command $.$, and Eq. (3.1) is treated as a part of the sentence with proper punctuation? Always treat an equation or expression as a part of the sentence.

3.3 Example of a Figure in LaTeX

Figure 3.1 is an example of a figure in LaTeX. For more details, check the link:
wikibooks.org/wiki/LaTeX/Floats,_Figures_and_Captions.
Keep your artwork (graphics, figures, illustrations) clean and readable. At least 300dpi is a good resolution of a PNG format artwork. However, an SVG format artwork saved as a PDF will produce the best quality graphics. There are numerous tools out there that can produce vector graphics and let you save that as an SVG file and/or as a PDF file. One example of such a tool is the "Flow algorithm software". Here is the link for that: flowgorithm.org.

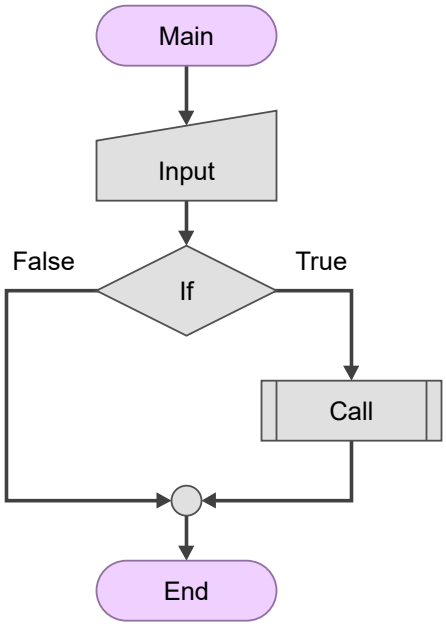

Figure 3.1: Example figure in LaTeX.

3.4 Example of an algorithm in LaTeX

Algorithm 1 is a good example of an algorithm in LaTeX.

Algorithm 1 Example caption: sum of all even numbers

Input: $\mathbf{x} = x_1, x_2, \ldots, x_N$
Output: $EvenSum$ (Sum of even numbers in $\mathbf{x}$)

1: **function** EvenSummation($\mathbf{x}$)
2: $EvenSum \leftarrow 0$
3: $N \leftarrow length(\mathbf{x})$
4: **for** $i \leftarrow 1$ to N **do**
5: **if** $x_i \mod 2 == 0$ **then** ▷ check if a number is even?
6: $EvenSum \leftarrow EvenSum + x_i$
7: **end if**
8: **end for**
9: **return** $EvenSum$
10: **end function**

3.5 Example of code snippet in LaTeX

Code Listing 3.1 is a good example of including a code snippet in a report. While using code snippets, take care of the following:

- do not paste your entire code (implementation) or everything you have coded. Add code snippets only.
- The algorithm shown in Algorithm 1 is usually preferred over code snippets in a technical/-scientific report.
- Make sure the entire code snippet or algorithm stays on a single page and does not overflow to another page(s).

Here are three examples of code snippets for three different languages (Python, Java, and CPP) illustrated in Listings 3.1, 3.2, and 3.3 respectively.

```python
import numpy as np

x  = [0, 1, 2, 3, 4, 5] # assign values to an array
evenSum = evenSummation(x) # call a function

def evenSummation(x):
    evenSum = 0
    n = len(x)
    for i in range(n):
        if np.mod(x[i],2) == 0: # check if a number is even?
            evenSum = evenSum + x[i]
    return evenSum
```

Listing 3.1: Code snippet in LaTeX and this is a Python code example

Here we used the "\clearpage" command and forced-out the second listing example onto the next page.

```java
public class EvenSum{
    public static int evenSummation(int[] x){
        int evenSum = 0;
        int n = x.length;
        for(int i = 0; i < n; i++){
            if(x[i]%2 == 0){ // check if a number is even?
                evenSum = evenSum + x[i];
            }
        }
        return evenSum;
    }
    public static void main(String[] args){
        int[] x  = {0, 1, 2, 3, 4, 5}; // assign values to an array
        int evenSum = evenSummation(x);
        System.out.println(evenSum);
    }
}
```

Listing 3.2: Code snippet in LaTeX and this is a Java code example

```cpp
int evenSummation(int x[]){
    int evenSum = 0;
    int n = sizeof(x);
    for(int i = 0; i < n; i++){
        if(x[i]%2 == 0){ // check if a number is even?
            evenSum = evenSum + x[i];
      }
    }
    return evenSum;
}

int main(){
    int x[]  = {0, 1, 2, 3, 4, 5}; // assign values to an array
    int evenSum = evenSummation(x);
    cout<<evenSum;
    return 0;
}
```

Listing 3.3: Code snippet in LaTeX and this is a C/C++ code example

3.6 Example of in-text citation style

3.6.1 Example of the equations and illustrations placement and reference in the text

Make sure whenever you refer to the equations, tables, figures, algorithms, and listings for the first time, they also appear (placed) somewhere on the same page or in the following page(s). Always make sure to refer to the equations, tables and figures used in the report. Do not leave them without an **in-text citation**. You can refer to equations, tables and figures more them once.

3.6.2 Example of the equations and illustrations style

Write **Eq.** with an uppercase "Eq" for an equation before using an equation number with (\eqref{.}). Use "Table" to refer to a table, "Figure" to refer to a figure, "Algorithm" to

refer to an algorithm and "Listing" to refer to listings (code snippets). Note that, we do not use the articles "a," "an," and "the" before the words Eq., Figure, Table, and Listing, but you may use an article for referring the words figure, table, etc. in general.

For example, the sentence "A report structure is shown in **the** Table 3.1" should be written as "A report structure is shown **in** Table 3.1."

3.7 Summary

Write a summary of this chapter.

Note: In the case of **software engineering** project a Chapter "**Testing and Validation**" should precede the "Results" chapter. See Section 3.1.1 for report organization of such project.

Chapter 4

Results

The results chapter tells a reader about your findings based on the methodology you have used to solve the investigated problem. For example:

- If your project aims to develop a software/web application, the results may be the developed software/system/performance of the system, etc., obtained using a relevant methodological approach in software engineering.
- If your project aims to implement an algorithm for its analysis, the results may be the performance of the algorithm obtained using a relevant experiment design.
- If your project aims to solve some problems/research questions over a collected dataset, the results may be the findings obtained using the applied tools/algorithms/etc.

Arrange your results and findings in a logical sequence.

4.1 A section

...

4.2 Example of a Table in LaTeX

Table 4.1 is an example of a table created using the package LaTeX"booktabs." do check the link: wikibooks.org/wiki/LaTeX/Tables for more details. A table should be clean and readable. Unnecessary horizontal lines and vertical lines in tables make them unreadable and messy. The example in Table 4.1 uses a minimum number of liens (only necessary ones). Make sure that the top rule and bottom rule (top and bottom horizontal lines) of a table are present.

Table 4.1: Example of a table in LaTeX

Bike		
Type	Color	Price (£)
Electric	black	700
Hybrid	blue	500
Road	blue	300
Mountain	red	300
Folding	black	500

4.3 Example of captions style

- The **caption of a Figure (artwork) goes below** the artwork (Figure/Graphics/illustration). See example artwork in Figure 3.1.
- The **caption of a Table goes above** the table. See the example in Table 4.1.
- The **caption of an Algorithm goes above** the algorithm. See the example in Algorithm 1.
- The **caption of a Listing goes below** the Listing (Code snippet). See example listing in Listing 3.1.

4.4 Summary

Write a summary of this chapter.

Chapter 5

Discussion and Analysis

Depending on the type of project you are doing, this chapter can be merged with "Results" Chapter as " Results and Discussion" as suggested by your supervisor.

In the case of software development and the standalone applications, describe the significance of the obtained results/performance of the system.

5.1 A section

Discussion and analysis chapter evaluates and analyses the results. It interprets the obtained results.

5.2 Significance of the findings

In this chapter, you should also try to discuss the significance of the results and key findings, in order to enhance the reader's understanding of the investigated problem

5.3 Limitations

Discuss the key limitations and potential implications or improvements of the findings.

5.4 Summary

Write a summary of this chapter.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Typically a conclusions chapter first summarizes the investigated problem and its aims and objectives. It summaries the critical/significant/major findings/results about the aims and objectives that have been obtained by applying the key methods/implementations/experiment set-ups. A conclusions chapter draws a picture/outline of your project's central and the most signification contributions and achievements.

A good conclusions summary could be approximately 300–500 words long, but this is just a recommendation.

A conclusions chapter followed by an abstract is the last things you write in your project report.

6.2 Future work

This section should refer to Chapter 4 where the author has reflected their criticality about their own solution. The future work is then sensibly proposed in this section.

Guidance on writing future work: While working on a project, you gain experience and learn the potential of your project and its future works. Discuss the future work of the project in technical terms. This has to be based on what has not been yet achieved in comparison to what you had initially planned and what you have learned from the project. Describe to a reader what future work(s) can be started from the things you have completed. This includes identifying what has not been achieved and what could be achieved.

A good future work summary could be approximately 300–500 words long, but this is just a recommendation.

Chapter 7

Reflection

Write a short paragraph on the substantial learning experience. This can include your decision-making approach in problem-solving.

Some hints: You obviously learned how to use different programming languages, write reports in LaTeXand use other technical tools. In this section, we are more interested in what you thought about the experience. Take some time to think and reflect on your individual project as an experience, rather than just a list of technical skills and knowledge. You may describe things you have learned from the research approach and strategy, the process of identifying and solving a problem, the process research inquiry, and the understanding of the impact of the project on your learning experience and future work.

Also think in terms of:

- what knowledge and skills you have developed
- what challenges you faced, but was not able to overcome
- what you could do this project differently if the same or similar problem would come
- rationalize the divisions from your initial planed aims and objectives.

A good reflective summary could be approximately 300–500 words long, but this is just a recommendation.

Note: The next chapter is "**References**," which will be automatically generated if you are using BibTeX referencing method. This template uses BibTeX referencing. Also, note that there is difference between "References" and "Bibliography." The list of "References" strictly only contain the list of articles, paper, and content you have cited (i.e., refereed) in the report. Whereas Bibliography is a list that contains the list of articles, paper, and content you have cited in the report plus the list of articles, paper, and content you have read in order to gain knowledge from. We recommend to use only the list of "References."

References

Janosi, Andras, S. W. P. M. and Detrano, R. (1988), 'Heart Disease', UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52P4X.

Sharma, V., Yadav, S. and Gupta, M. (2020), Heart disease prediction using machine learning techniques, *in* '2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)', pp. 177–181.

Appendix A

An Appendix Chapter (Optional)

Some lengthy tables, codes, raw data, length proofs, etc. which are **very important but not essential part** of the project report goes into an Appendix. An appendix is something a reader would consult if he/she needs extra information and a more comprehensive understating of the report. Also, note that you should use one appendix for one idea.

An appendix is optional. If you feel you do not need to include an appendix in your report, avoid including it. Sometime including irrelevant and unnecessary materials in the Appendices may unreasonably increase the total number of pages in your report and distract the reader.

Appendix B

An Appendix Chapter (Optional)

...