الجمهورية الجزائرية الديمقراطية الشعبية
PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
وزارة التعليم العالي و البحث العلمي
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
جامعة عمار ثليجي بالأغواط
UNIVERSITY OF AMMAR TELIDJI LAGHOUAT

FACULTY OF SCIENCES
DEPARTMENT OF COMPUTER SCIENCE

MASTER THESIS

Field : Mathematics and Computer Science
Option : Computer Science
Specialization : Information systems and decision

Submitted by :

Abdelaziz Soumia **Ferroudj Chaima**

Theme

Heart Attack Prediction Based on Machine Learning Techniques

Jury members :

Mr.	MAICHA Mohamed EL Habib	MCB	(University of Laghouat)	President
Mr.	CHELLAMA Laradj	MAA	(University of Laghouat)	Examiner
Mr.	GUELLOUMA Younes	MCA	(University of Laghouat)	Examiner
Mr.	BOUAKKAZ Mustapha	Prof	(University of Laghouat)	Advisor

Academic Year 2023/2024

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

Dédicaces

I dedicate this dissertation. . .

This work is dedicated to my parents and dear sisters and brothers . Your support, encouragement, and unwavering love have been the light that has guided me throughout this journey. For my parents, it is your belief in me and the sacrifices you have made that have shaped me into who I am today. Your encouragement gave me the strength to achieve my dreams.
To my sisters **Sabrina ,Adhiba , Wafaa** and brothers, thank you for always being there for me, for understanding, and for being my source of inspiration. Your presence made overcoming all challenges easier.
This work is a tribute to the love and support of my family. Without you, this achievement would not have been possible. Thank you for everything. . . .

Ferroudj Chaima

Dedication

I dedicate this dissertation . . .

To my dearest family,
Words cannot express the depth of my gratitude for your unwavering support, encouragement, and boundless love throughout this journey. To my parents, whose endless sacrifices and faith in me have been my guiding light, thank you for believing in me even when I struggled to believe in myself. To my siblings, for your patience, understanding, and for always being there to lift me up, you have been my pillars of strength. This thesis is a testament to your love and faith, and I dedicate this work to you with all my heart.
To my cherished friend, Amira Belhanafi,
Your invaluable assistance, unwavering support, and steadfast companionship have been my guide during the completion of this thesis. Your encouragement and belief in me have been a source of immense strength and motivation. I am profoundly grateful for your friendship and all that you have done to help me reach this milestone.
To my dear cousin Roukia Dohssi,
Your unwavering support has meant the world to me. Thank you for always being there.
With all my love and deepest appreciation .
. . .

Abdelaziz Soumia

Acknowledgments

First and foremost, we would like to thank Allah, the Almighty and Merciful, who has given us the strength and patience to accomplish this work. We extend our gratitude to our supervisor, Professor **Mustapha Bouakkaz**, for the honor of guiding us. We are deeply thankful for his valuable advice, guidance, and the time he dedicated to help us complete this project, as well as for his availability and direction.

We also express our gratitude to our friend from the Faculty of Medicine, **Hiba Kouader**, for her assistance in the completion of this work, and to our friend **Amira Belhanafi** for her invaluable help with this thesis.

Additionally, we would like to express our sincere thanks to all the teachers who have taught us and whose expertise has supported us in the pursuit of our studies. Finally, we thank everyone who, directly or indirectly, contributed to the completion of this thesis.

Abstract

The emergence of artificial intelligence and its rapid spread has led to a change in many factors, especially in the medical field, so that it has an effective role in exploring diseases, methods of diagnosis, and assisting in health care in various forms. One of the common diseases that pose a great risk to human health is heart disease, especially heart attacks. Many machine learning models have treated heart diseases in general or heart attacks in particular, but with different models. In this thesis, we wanted to experiment with machine learning techniques, especially classification, to classify a person's condition as either having the possibility of a heart attack or not having any possibility of it occurring. So we used logistic regression, SVM and ANN to try to reach the highest possible accuracy to provide Greater effectiveness when using these models. To facilitate the use of these models, we have developed a simple website using Flask that enables the user to enter the necessary factors to predict a heart attack.

Key Words :Machine learning , SVM , Logistic Regression , MLPClassifier

ملخص

ظهور الذكاء الاصطناعي وانتشاره السريع ادى الى تغيير الكثير من العوامل وخصوصا في المجال الطبي بحيث اصبح له دور فعال في استكشاف الامراض وطرق للتشخيص والمساعدة في الرعاية الصحية بأشكال مختلفة . ومن الامراض الشائعة التي لها خطورة كبيرة على صحة الانسان هي امراض القلب وخصوصا النوبات القلبية بحيث يوجد الكثير من نماذج التعلم الالي التي عالجت أمراض القلب بصفة عامة أو النوبات القلبية بصفة خاصة ولكن بنماذج مختلفة

في هذه الاطروحة اردنا تجريب تقنيات التعلم الالي وبالخصوص التصنيف من أجل تصنيف حالة الشخص اما ان يكون لديه احتمال حدوث نوبة قلبية او لا يجود أي احتمال لحدوثها بحيث استعملنا الانحدار اللوجستي ونضام الدعم الالي وتقنيات الشبكة العصبية من اجل المحاولة الى الوصول الى اعلى دقة ممكنة لتقديم فعالية اكبر عند استخدام هذه النماذج ولتسهيل استخدام هذه النماذج قمنا بتطوير موقع ويب بسيط تمكن للمستخدم ادخال العوامل اللازمة من اجل توقع السكتة القلبية

الكلمات المفتاحية :

التعلم الألي , SVM ، الانحدار اللوجستي، MLPClassifier

Résumé

L'émergence de l'intelligence artificielle et sa propagation rapide ont entraîné un changement dans de nombreux domaines, notamment dans le domaine médical, où elle joue un rôle efficace dans l'exploration des maladies, les méthodes de diagnostic et l'aide aux soins de santé sous diverses formes. L'une des maladies courantes qui posent un grand risque pour la santé humaine est la maladie cardiaque, en particulier les crises cardiaques. De nombreux modèles d'apprentissage automatique ont été utilisés pour traiter les maladies cardiaques en général ou les crises cardiaques en particulier, mais avec des modèles différents.

Dans cette thèse, nous avons voulu expérimenter des techniques d'apprentissage automatique, en particulier la classification, afin de classer l'état d'une personne comme ayant la possibilité d'une crise cardiaque ou n'ayant aucune possibilité que cela se produise. Nous avons donc utilisé la régression logistique, les SVM et les réseaux de neurones artificiels (ANN) pour essayer d'atteindre la plus grande précision possible et offrir une plus grande efficacité lors de l'utilisation de ces modèles. Pour faciliter l'utilisation de ces modèles, nous avons développé un site web simple en utilisant Flask qui permet à l'utilisateur de saisir les facteurs nécessaires afin de prédire une crise cardiaque.

Mots clés :Machine learning , SVM , Logistic Regression , MLPClassifier

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

In healthcare, particularly in the field of heart disease prediction, the emergence of artificial intelligence (AI) has marked a significant advancement. This work delves into our contributions to heart attack prediction, utilizing various datasets and machine learning models to achieve notable accuracy. We contextualize our work within the existing research landscape by reviewing relevant work. Our primary objective is to develop a model that can accurately predict the occurrence of heart attacks based on pertinent factors and indicators found in medical datasets. The study is organized as follows:

The first section provides a comprehensive overview of heart attacks, exploring their different types, causes, and potential complications. This foundational knowledge is essential for understanding the medical context and the significance of accurate heart attack prediction

The next section delves into artificial intelligence, explaining the different techniques used in the field.

In the next chapter, we delve into the specifics of the models we used, exploring their design, functionality, and relevance to our study, and their performance in predicting heart attacks, providing a clear understanding of their role in our research.

Finally, we provide a comprehensive summary of our findings, and we present our web application along with insights into potential future research directions and applications in the field of heart attack prediction.

CHAPTER 1

MEDICAL CONTEXT

1.1 Introduction

Heart attacks pose a significant threat to public health, standing as one of the foremost causes of death worldwide. According to the World Health Organization (WHO), cardiovascular diseases, including heart attacks, claim a staggering 18.6 million lives in 2019 alone, comprising roughly 32% of global fatalities. While impacting individuals across all age groups, the likelihood escalates with advancing age, with males generally facing higher susceptibility compared to females. Timeliness is paramount in managing heart attacks, with swift diagnosis and intervention pivotal in mitigating cardiac damage and enhancing survival prospects. By fostering awareness regarding risk factors, symptoms, and the urgency of timely diagnosis and treatment, we can actively contribute to preserving lives and enhancing global health . [1]

1.2 Definition

heart attack, also known medically as a myocardial infarction, is a significant halting of blood supply to a portion of the heart muscle. This is frequently brought on by a blood clot obstructing a coronary artery. The common cause of this obstruction is the buildup of atherosclerotic plaques, which are composed of cholesterol and other materials, inside the coronary arteries. Consequently, the cardiac muscle's ability to receive oxygen and vital nutrients is hindered, which may cause irreversible damage to its cells. [1]

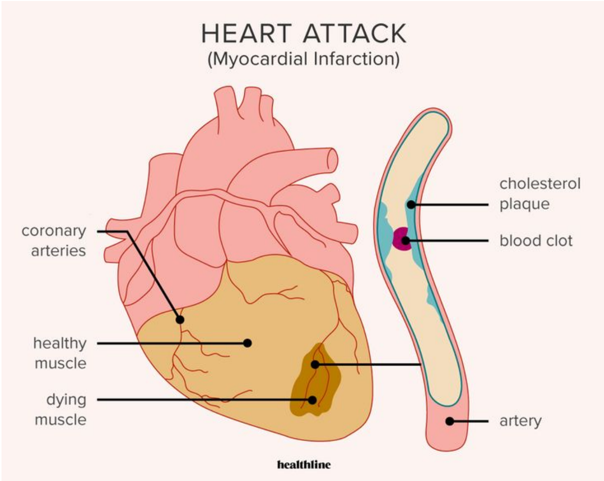

Figure 1.1: heart attack

1.3 Types of heart attack and their characteristics

1.3.1 Myocardial infarction with ST-segment elevation (STEMI)

Is characterized by complete obstruction of a coronary artery, resulting in total blockage of blood supply to the heart. Symptoms associated with this condition include intense and persistent chest pain, shortness of breath, nausea, and excessive sweating. The diagnosis of STEMI is typically confirmed by an electrocardiogram (ECG) revealing characteristic ST-segment elevation. Treatment of myocardial infarction with ST-segment elevation is urgent and aims to restore blood circulation as quickly as possible. Therapeutic options include thrombolysis with medication administration or a more invasive intervention such as angioplasty with stent placement to restore blood flow. [2]

1.3.2 Myocardial infarction without ST segment elevation (NSTEMI)

Occurs when the coronary artery is partially blocked, resulting in a decrease in blood flow to the heart without complete interruption. The symptoms associated with this condition are similar to those of STEMI, although they may be less intense. The diagnosis of NSTEMI is made by an electrocardiogram (ECG) that does not show elevation of the ST segment, but other abnormalities may be present. Treatment of myocardial infarction without ST segment elevation is urgent and may include administration of anti-platelet drugs, aspirin, and careful cardiac monitoring. In some cases, more invasive intervention such as angioplasty with stent installation may be necessary to restore blood flow to the heart muscle. [2]

1.3.3 Silent Myocardial Infarct

Is characterized by its appearance without apparent symptoms, which makes it often detected in subsequent heart tests. Although it may go unnoticed, it presents a risk of potential heart damage despite the absence of obvious symptoms. Therefore, it is crucial to monitor this condition closely and to establish appropriate treatment to prevent possible future complications [2]

1.3.4 Angina pectoris

Is characterized by chest pain resulting from a temporary reduction in the blood supply to the heart. Typical symptoms include chest pain felt as a feeling of pressure or tightening, usually triggered by physical effort or stress. Treatment of angina is aimed at relieving pain and preventing seizures with the help of specific drugs. [2]

1.3.5 Coronary spasm

is manifested by a sudden contraction of a coronary artery, resulting in a sharp decrease in blood flow to the heart. Symptoms associated with this phenomenon include chest pain similar to that of angina, often occurring at rest. Coronary spasm treatment is aimed at preventing spasms and relieving pain, usually through the use of specific medications. [2]

1.4 Causes of heart attack

Myocardial infarction or heart attack occurs when

1. Constriction and obstruction of the coronary arteries by plaques, cholesterol and fat deposits, resulting in the formation of blood clots that block the blood flow to the heart.[3]
2. Arterial wall hardening, known as atherosclerotic disease.[3]
3. Fat accumulation in the coronary arteries.[3]
4. Increased risk with age, especially after 65, and greater predisposition in men than in women.[3]
5. Other risk factors include smoking, high blood pressure, high cholesterol and LDL levels, a diet high in saturated fats (animal meat), obesity, lack of exercise, and diabetes (high blood sugar levels).[3]
6. Family history of heart disease.[3]

7. Muscle spasms of arterial walls and narrowing of arteries.[3]
8. Emotional stress, intense physical effort, exposure to extreme cold temperatures, lifting heavy objects, and drug use (cocaine or amphetamines) are all triggers of heart attack.[3]
9. After a first heart attack, the risk of a new heart attack increases in the future.[3]

To prevent heart attacks, it is recommended to adopt a balanced diet with low fat and sugar intake, exercise regularly, get enough sleep, and manage stress. [3]

1.5 Symptoms of heart attack

Common symptoms of heart attack are:

1. Chest pain or feeling of pressure in the chest.[3]
2. Angina without precursor signs.[3]
3. Sometimes a mild heart attack has no symptoms and is called a "silent heart attack".[3]
4. Preliminary signs of a heart attack include:
    - Chest pain or hypertension.
    - Tightening, burning, feeling of oppression, deafness and pain in the chest for more than 10 minutes.
    - Pain in the left shoulder or left arm, which can go up to the neck or along the jaw.
    - Excessive sweating and dizziness.
    - Muscle weakness.
    - Nausea or vomiting.
    - Sensation of suffocation when inhaling smoke.
    - Anxiety, stress, feeling of imminent danger and depression.[3]
5. However, there are no symptoms for a silent heart attack.

[3]

1.6 The diagnosis of heart attack

Myocardial infarction is often formally diagnosed by a battery of tests and examinations. Typically, patients exhibiting signs of myocardial infarction are admitted to the hospital immediately and get treatment in the coronary artery care unit. A complete physical examination, a measurement of the patient's blood pressure, and a full medical history are the first steps in the diagnosing process. After that, an electrocardiogram, often known as an EKG or ECG, is carried out to assess heart rate and find any irregularities in the heart's blood supply. Blood tests are also carried out to evaluate blood levels of fat and protein, which may signify cardiac muscle injury.

Chromatography or coronary angiography are used to see the coronary arteries and identify any blockage.

[3]

1.7 Treatment of heart attack

Several medical procedures are used to treat myocardial infarction with the goals of preserving heart muscle health and reestablishing blood flow to the heart. Cardiopulmonary resuscitation (CPR) or the use of an automated external defibrillator (AED) should be done right away if the person's heart rate drops or stops. Myocardial infarction patients are often treated with

- Thrombolytic medications, such as tissue plasminogen activator (tPA), streptokinase, or urokinase, which are given to break up arterial blockages within three hours of the heart attack starting.
- The use of analgesics to treat pain, such as meperidine or morphine.
- Giving antihypertensive medications to lower oxygen and blood pressure in the heart, such as beta-blockers, angiotensin-converting enzyme inhibitors (ACE), or calcium channel blockers.
- The delivery of oxygen via nasal tubes.
- The use of anticoagulants to lower the risk of blood clots, such as warfarin, aspirin, or heparin.

[3]

1.8 Problematic

It is imperative that we prioritize the prevention of heart attacks by means of efficacious lifestyle modifications, medical interventions, and public health policies, and guarantee their efficient implementation. Improving early detection, controlling risk factors, and refining treatment protocols are all critical steps in reducing the severity and dangers of heart attacks. , how can we reduce the incidence of heart attacks or reduce their impact and risks on human health? How can we use machine learning and its features to predict them?

1.9 Solution proposed

The aim of this study is to present a solution based on artificial intelligence techniques (machine learning) to predict heart attacks and thus reduce their incidence.

1.10 Conclusion

In this chapter, we have studied heart attacks in detail, including their different types, and we have specifically looked at the potential complications of this condition. Our objective was to present the problem and the solution envisaged for our study. In the next chapter, we will explore the different machine-learning techniques that will be used to meet our goal of predicting heart attacks.

CHAPTER 2

ARTIFICIAL INTELLIGENCE

2.1 Introduction

In the previous chapter, we discussed heart attacks: their causes, symptoms, serious consequences, and treatment. We also proposed a solution for predicting them, based on artificial intelligence(machine learning). This chapter focuses on artificial intelligence and the specific techniques we used in our study.

2.2 Artificial Intelligence

The goal of artificial intelligence (AI) is to build computers and robots that are both more and less capable than humans. Applications with AI capabilities can give information or start actions on their own by interpreting and contextualizing data. AI is now a key component of many technologies, such as voice assistants and smart devices. Companies are automating procedures, expediting decision-making, and improving consumer engagement with chatbots by leveraging techniques like natural language processing and computer vision, which allow computers to understand human language and analyze images.

[4]

2.3 Machine learning

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that focuses on creating computer algorithms that enhance themselves automatically through experience and data utilization. Essentially, machine learning allows computers to learn from data and make decisions or predictions without explicit programming.

At its essence, machine learning involves designing and implementing algorithms that enable these decisions and predictions. These algorithms are built to improve their accuracy and effectiveness over time as they process more data.

In traditional programming, a computer executes a task by following a set of predefined instructions. Conversely, in machine learning, the computer is given a dataset and a task but must figure out how to complete the task based on the provided examples. [5]

2.4 Types of Machine Learning

2.4.1 supervised learning

Using a labeled dataset, supervised learning entails training a machine learning model. This indicates that an output label is associated with every training case. Based on these instances, the model learns to relate inputs to outputs and makes predictions on fresh, unknown data. [5]

2.4.2 Unsupervised learning

Training a model on unlabeled data entails doing unsupervised learning. Deducing the inherent structure from a collection of data points is the aim. Finding patterns, grouping related data points, and reducing the dimensionality of data are common uses for this kind of learning. [5]

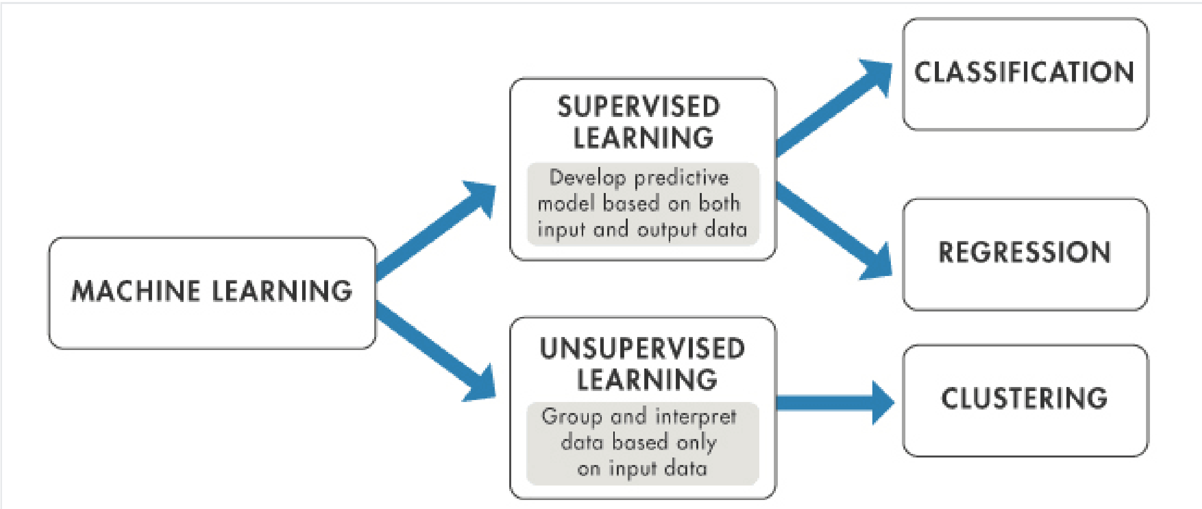

Figure 2.1: supervised and unsupervised learning
[5]

2.4.3 Reinforcement learning

Reinforcement learning is a type of machine learning where an agent learns to make decisions by interacting with its environment. The agent is rewarded or penalized (with points) for the actions it takes, and its goal is to maximize the total reward. Unlike supervised and unsupervised learning, reinforcement learning is particularly suited to problems where the data is sequential, and the decision made at each step can affect future outcomes. [5]

2.5 Deep learning

Deep learning is the branch of machine learning which is based on artificial neural network architecture. An artificial neural network or ANN uses layers of interconnected nodes called neurons that work together to process and learn from the input data.

In a fully connected Deep neural network, there is an input layer and one or more hidden layers connected one after the other. Each neuron receives input from the previous layer neurons or the input layer. The output of one neuron becomes the input to other neurons in the next layer of the network, and this process continues until the final layer produces the output of the network. The layers of the neural network transform the input data through a series of nonlinear transformations, allowing the network to learn complex representations of the input data. [6]

2.5.1 Artificial neural networks

Artificial neural networks are built on the principles of the structure and operation of human neurons. It is also known as neural networks or neural nets. An artificial neural network's input layer, which is the first layer, receives input from external sources and passes it on to the hidden layer, which is the second layer. Each neuron in the hidden layer gets information from the neurons in the previous layer, computes the weighted total, and then transfers it to the neurons in the next layer. These connections are weighted, which means that the impacts of the inputs from the preceding layer are more or less optimized by giving each input a distinct weight. These weights are then adjusted during the training process to enhance the performance of the model[6]

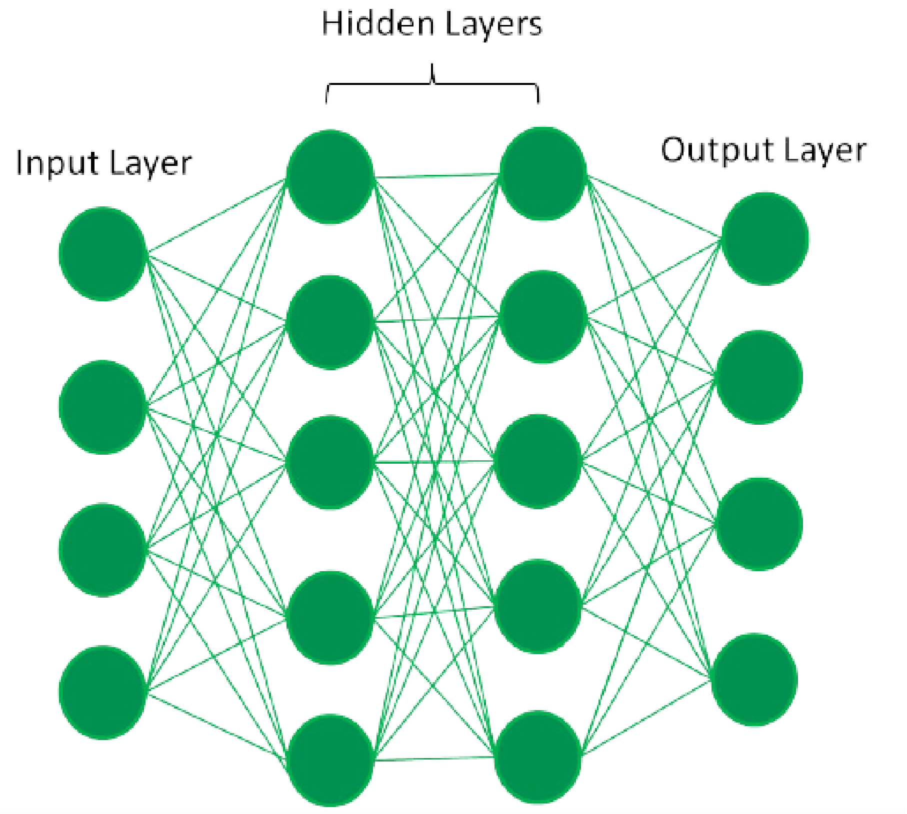

Figure 2.2: Fully Connected Artificial Neural Network
[6]

2.5.2 Types Deep learning

Deep Learning models can automatically learn features from the data, which makes them well-suited for tasks such as image recognition, speech recognition, and natural language processing. The most widely used architectures in deep learning are feedforward neural networks(FNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). [6]

Feedforward neural networks (FNNs)

Are the simplest type of ANNs, with a linear flow of information through the network. FNNs have been widely used for tasks such as image classification, speech recognition, and natural language processing.[6]

Convolutional Neural Networks (CNNs)

Are specifically for image and video recognition tasks. CNNs are able to automatically learn features from the images, which makes them well-suited for tasks such as image classification, object detection, and image segmentation.[6]

Recurrent Neural Networks (RNNs)

Are a type of neural network that is able to process sequential data, such as time series and natural language. RNNs are able to maintain an internal state that captures information about the previous inputs, which makes them well-suited for tasks such as speech recognition, natural language processing, and language translation.[6]

2.6 Difference between Machine Learning and Deep Learning

Machine learning and deep learning both are subsets of artificial intelligence but there are many similarities and differences between them.

Machine Learning	**Deep Learning**
Applies statistical algorithms to learn the hidden patterns and relationships in the dataset.	Uses artificial neural network architecture to learn the hidden patterns and relationships in the dataset.
Can work on a smaller amount of data.	Requires a larger volume of data compared to machine learning.
Better for low-label tasks.	Better for complex tasks like image processing, natural language processing, etc.
Takes less time to train the model.	Takes more time to train the model.
Less complex and easy to interpret the result.	More complex, it works like a black box; interpretations of the result are not easy.
It can work on the CPU or requires less computing power as compared to deep learning.	It requires a high-performance computer with GPU.

Table 2.1: Comparison between Machine Learning and Deep Learning [6]

2.7 Related works

Study	**Year Published**	**Algorithm Used**	**Accuracy**	**Method Used**
Johnson and Brown	2020	SVM	85%	PCA
Patel and Mehta	2020	k-NN	80%	Data Reduction
Lee et al.	2021	Neural Network	92%	Clustering
Zhang et al.	2021	Logistic Regression	83%	Data Normalization
Wang and Chen	2022	Naive Bayes	82%	Data Imputation
Kim and Park	2022	Decision Tree	87%	Association Rules
Gonzalez and Martinez	2023	Gradient Boosting	90%	Data Augmentation
Smith et al.	2024	Random Forest	89%	Feature Selection

Table 2.2: Summary of heart attack prediction studies (2020-2024).

2.8 Conclusion

In summary, this chapter has offered a concise overview of artificial intelligence, covering its various types and techniques pertinent to our study. Additionally, we have briefly examined some related works to inform our research.

CHAPTER 3

MODELS USED

3.1 Introduction

In the previous chapter, we discussed machine learning, defined its types, and explained some important concepts in this field. Now, we will delve into the models we used, how they work, and the results obtained.

3.2 The techniques used for prediction

3.2.1 Logistic Regression

Logistic Regression reigns as one of the most popular especially for classification tasks. This work delves into the intricacies of logistic regression, providing detailed explanations alongside real-world examples to enhance comprehension. Logistic Regression is a type of supervised learning algorithm that uses labeled data to train the model for making predictions. Logistic regression is a widely used statistical algorithm used to model the probability of a binary outcome, such as yes or no, true or false, or 0 or 1 . [7] The logistic regression model can be expressed as:

$$p(Y = 1 \mid X_1, X_2, \ldots, X_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

Where:

- $p(Y = 1 \mid X_1, X_2, \ldots, X_n)$ is the probability that Y equals 1 given the values of $X_1, X_2, \ldots, X_n$.
- e is the base of the natural logarithm (Euler's number).

- $\beta_0, \beta_1, \ldots, \beta_n$ are coefficients that represent the effects of the independent variables on the log-odds of the dependent variable Y.
- β_0 is the intercept term.

Example of Logistic Regression

An illustration of logistic regression: Let's look at the issue of forecasting health insurance purchases according to an individual's age. In this case, the target variable is the individual's decision to obtain health insurance, and the independent variable is their age. We can use logistic regression to tackle this issue. We can generalize the relationship using an S-shaped curve, or sigmoid curve, as shown in Fig. 1. People over 40, for example, have a probability of 0.5, meaning they have an equal chance of getting health insurance. People who have a probability higher than 0.5 are more likely to buy health insurance than people whose probability is lower than 0.5 .[8]

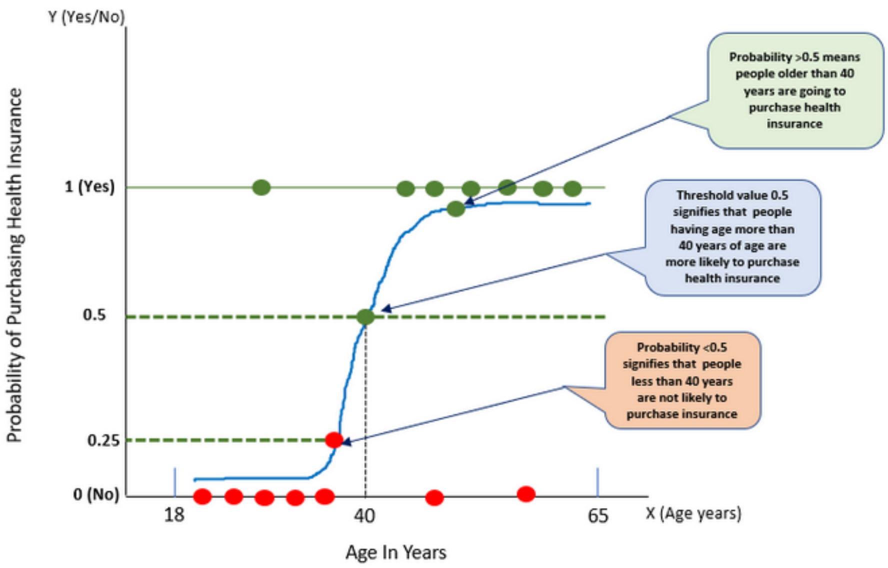

Figure 3.1: example of LR
[8]

[8]

Logistic function and medcine field

The logical function is the foundation of the logistic model. The unique feature of the logistic function is that it always assumes values between 0 and 1, even for values between minus and plus infinity. For instance, determining the factors that influence a disease is a common use in medicine. In this instance, 0 might indicate no disease and 1 could indicate disease.[9]

3.2.2 Types of Logistic Regression

Logistic regression can be classified into three major types as follows:[10]

1. Binary Logistic Regression

When the dependent variable has only two categories. For example, predicting whether a patient is suffering from diabetes or not based on certain medical conditions.[10]

2. Multinomial Logistic Regression

When the dependent variable has more than two categories. For example, predicting whether a banking transaction is safe, fraudulent, or doubtful based on account level and transaction level details.[10]

3. Ordinal Logistic Regression

When the dependent variable has ordered categories. For example, diagnosing a patient as having a high, mild, or severe viral infection based on certain medical conditions.[10]

3.2.3 Support Vector Machine (SVM)

Support vector machines (SVMs) are robust and adaptable supervised machine learning techniques used in regression, outlier detection, and classification. SVMs are typically used in classification problems because of their high level of efficiency in high dimensional spaces. Due to their use of a subset of training points in the prediction function, SVMs are widely used and memory-efficient. [11]

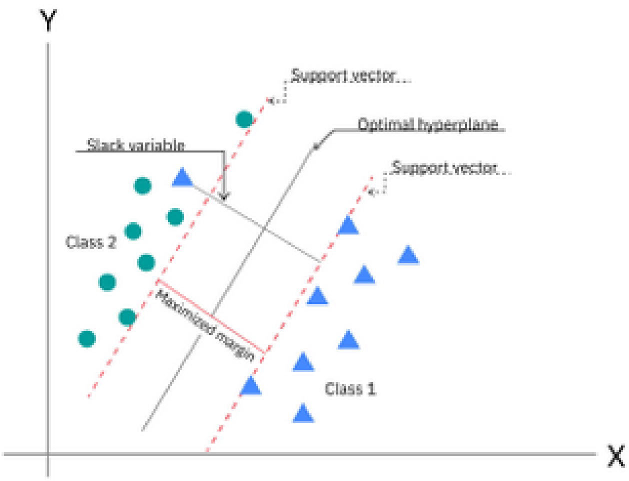

Figure 3.2: SVM
[11]

How SVMs work

How Support Vector Machines (SVMs) Operate: SVMs primarily aim to partition datasets into multiple classes in order to identify a maximum marginal hyperplane (MMH). This can be achieved in two steps:

- Support Vector Machines will first generate hyperplanes that separates the classes in the best way.
- After that it will choose the hyperplane that segregates the classes correctly.[11]

Some important concepts in SVM are as follows:

- **Support Vectors** Support Vectors They may be defined as the data points which are closest to the hyperplane. Support vectors help in deciding the separating line.[11]
- **Hyperplane** The decision plane or space that divides a set of objects having different classes.[11]
- **Margin** The gap between two lines on the closest data points of different classes is called the margin.[11]

3.2.4 Multi-Layer Perceptrons (MLP)

Multi-Layer Perceptrons (MLP) stands out as a pivotal machine learning technique within the realm of artificial neural networks, particularly for classification tasks.

Renowned for its flexibility and efficacy, MLP offers a versatile approach to addressing diverse classification problems, ranging from text classification to image recognition. While conventional linear classifiers may struggle with the complexity of certain datasets, MLPs excel in capturing intricate, non-linear relationships inherent in the data. [12]

Architecture and Working of Multi-Layer Perceptron

Regression and classification are two prominent machine learning tasks that use MLPs. An outline of the composition and operation of an MLP is provided below:

1. **Architecture**
    - **Input Layer**: Neurons in the input layer receive the dataset's features directly. The number of neurons in the input layer corresponds to the total number of features in the dataset, and each neuron in the layer represents a distinct feature.[12]

- **Hidden Layer**: : There could be one or more hidden layers positioned in between the input and output layers. One hyperparameter that can be changed based on the layer is the total number of neurons in each hidden layer. These hidden layers are essential for identifying intricate patterns in the data.[12]
- **Output Layer**: The output layer generates the final predictions or outputs based on the information processed in the hidden layers. The number of neurons in the output layer depends on the requirements of the task:
  - For binary classification, there is typically only one neuron that outputs a probability score.[12]
  - For multi-class classification, the number of neurons equals the number of classes, with each neuron generating a probability score for its respective class.[12]
  - For regression tasks, one neuron produces the continuous projected value.[12]

2. **Working**

- **Initialization**: initialization: The network's neurons' weights (W) Typically these parameters are initialized with modest random numbers.[12]
- **Forward Propagation**: During training, input data is repeatedly passed through the network. Each neuron in a layer receives the weighted sum of inputs from the previous layer, applies an activation function, and transmits the result to the next layer. Activation functions introduce non-linearity to the model, enabling it to learn complex correlations .[12]
- **Loss Calculation**: Loss Calculation: A loss (error) is computed by comparing the network's out-put to the desired target values. Common loss functions include Mean Squared Error (MSE) for regression and Cross-Entropy for classification.[12]
- **Backpropagation**: To minimize the loss, the network adjusts its biases and weights. Backpropagation calculates gradients of the loss with respect to each network parameter. [12]
- **Training**: Forward propagation, loss calculation, and backpropagation are iterated over multiple epochs until the model converges. Hyperparameters such as the learning rate and the number of epochs can be adjusted.[12]
- **Prediction**: Once trained, the MLP uses forward propagation with optimized weights and biases to make predictions on new, unseen data.[12]

CHAPTER 4

OBTAINED RESULTS

4.1 Dataset

In this study, we're investigating heart attack prediction using a dataset called "heart.csv" sourced from Kaggle. This dataset contains various factors related to heart health, like clinical measurements and patient demographics. We aim to understand how different factors influence the likelihood of a heart attack. By studying this data, we aim to create models capable of accurately predicting the risk of heart attacks.

we provide an overview and detailed analysis of the 'heart.csv' dataset

1. **age:** This attribute represents the age of the individual in years. Age is a crucial factor in assessing heart disease risk, as the likelihood of developing heart issues tends to increase with age.

2. **sex:** This attribute indicates the gender of the individual, with binary values representing male (0) and female (1). Gender can influence heart disease risk, with variations in prevalence and symptoms between men and women.

3. **cp:** This attribute represents chest pain type, categorized into different types such as typical angina, atypical angina, non-anginal pain, and asymptomatic. Chest pain is a common symptom associated with various heart conditions.

4. **trtbps:** This attribute refers to resting blood pressure, measured in mm Hg (millimeters of mercury). Elevated blood pressure is a significant risk factor for heart disease and other cardiovascular issues.

5. **chol:** This attribute represents serum cholesterol levels, measured in mg/dl (milligrams per deciliter). High cholesterol levels are associated with an increased risk of developing heart disease.

6. **fbs:** This attribute indicates fasting blood sugar levels, with binary values indicating whether the fasting blood sugar is greater than 120 mg/dl (1) or not (0). Elevated fasting blood sugar levels can indicate diabetes, which is a risk factor for heart disease.

7. **restecg:** This attribute represents resting electrocardiographic results, which may include normal, ST-T wave abnormality, or probable or definite left ventricular hypertrophy. A normal electrocardiographic findings can indicate heart abnormalities.

8. **thalachh:** This attribute represents resting electrocardiographic results, which may include normal, ST-T wave abnormality, or probable or definite left ventricular hypertrophy. A normal electrocardiographic findings can indicate heart abnormalities.

9. **exng:** This attribute indicates exercise-induced angina, with binary values indicating the presence (1) or absence (0) of exercise-induced angina. Angina is chest pain or discomfort caused by reduced blood flow to the heart muscle during physical exertion.

10. **oldpeak:** This attribute represents the ST depression induced by exercise relative to rest. ST depression on an electrocardiogram (ECG) can indicate myocardial ischemia, which occurs when the heart muscle doesn't receive enough oxygen-rich blood.

11. **slp:** This attribute indicates the slope of the peak exercise ST segment, categorized into upsloping, flat, or downsloping. The ST segment on an ECG can provide valuable information about the heart's electrical activity during exercise.

12. **caa:** This attribute represents the number of major vessels (0-3) colored by fluoroscopy. The presence of blockages or narrowing in the major blood vessels can indicate coronary artery disease, a common cause of heart attacks.

13. **thall:** This attribute indicates thalassemia, a blood disorder that affects the production of hemoglobin. Thalassemia can indirectly affect heart health by leading to anemia, which can strain the heart.

14. **output:** This attribute indicates the presence (1) or absence (0) of heart attack. This is the target variable that machine learning models aim to predict based on the other attributes.

4.2 Results

Each of the models that we used provides us with different results in terms of the expected state, and thus in terms of accuracy. Accuracy is one of the metrics used to evaluate classification models. It has the following definition:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

For each model utilized, the following steps were undertaken:

1. We employed the essential libraries, namely NumPy, pandas, and sklearn, for our implementation.
2. We loaded the dataset within the Google Colab environment and proceeded to read it using the pandas library.
3. Simple preprocessing of the data set like checking the missing values and normalizing the features X with StanderScaler
4. Subsequently, we partitioned the data rows into two sets: X for the features and Y for the output.
5. To prepare the data for model training, we scaled it and divided it into training and testing sets. Utilizing the `train_test_split` function, we allocated 20% of the data for testing and 80% for training. Furthermore, we incorporated the parameter `stratify=Y` to ensure balanced class representation across both sets.
6. Select model (LR or SVM, MLP)

4.2.1 Model 1 (Logistic Regression)

Selection of Model

Choosing logistic regression for heart attack prediction can be justified based on several key factors that make it an appropriate and effective method for this kind of binary classification task:

1. Binary Outcome Suitability Logistic regression works best for situations where the result is either one thing or another. When it comes to predicting heart attacks, we're usually looking at whether a patient will have one or not (yes/no), which fits perfectly with how logistic regression deals with this kind of binary outcome.

2. Interpretability Logistic regression models are easy to understand. The coefficients in these models can be directly interpreted as showing how different factors (like age, cholesterol levels, blood pressure, etc.) affect the likelihood of an outcome (such as having a heart attack). This clarity is important in medical settings, where understanding the relationship between risk factors and outcomes is crucial for making informed clinical decisions.

3. Simplicity and Efficiency Logistic regression is easy to implement and runs efficiently on computers. This is particularly useful when working with large datasets, which is common in medical research.

4. Probabilistic Output The output of a logistic regression model is a probability score between 0 and 1, showing the likelihood of a heart attack. This probabilistic interpretation is useful in medical settings, as it allows risk to be quantified and communicated clearly, enabling informed decisions about patient care and preventive measures.

5. Baseline Model Logistic regression often acts as a strong baseline model. Its performance provides a standard for comparing more complex models, such as decision trees, random forests, or neural networks.

Training the Model

1. **Dataset:** For the dataset, we used the HEART dataset with simple processing of the data set by checking the missing values and normalizing the features X with StanderScaler.

2. **Data Splitting:** Divide the dataset into two parts: a training set and a test set, with 80% for training and 20% for testing.

3. **Model Initialization:** Initialize the logistic regression model. This model will learn the relationship between the input features (risk factors) and the binary target variable (heart attack occurrence: yes/no).

4. **Training the Model:** For each patient in the training set, calculate the predicted probability of a heart attack using the logistic function. This function outputs a value between 0 and 1.

$$\text{Predicted Probability} = \frac{1}{1 + e^{-(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

   Here β_i are the coefficients, and x_i are the features.

5. **Model Evaluation:** After training, the performance of the Logistic Regression model is evaluated

   - **Accuracy of Training Data:** 0.8675034867503487 (86.75%)
   - **Accuracy of Testing Data:** 0.8181818181818182 (81.82%)

6. **Deployment:** Once validated, we deploy the model to predict heart attack risk in new patients, helping doctors with decision-making and preventative care.

The pickle library was chosen due to its simplicity in connecting the interface with the model. It offers a straightforward approach to serializing and deserializing machine learning models, allowing for easy integration with the Flask application. Moreover, pickle can load any type of model, making it a versatile choice compared to direct loading methods, especially for models like SVM that don't have direct loading requirements.

4.2.2 Model 2 (Support Vector Machine (SVM))

Selection of Model

1. **Handling Complex Relationships:** SVM can understand complex and non-linear relationships between different risk factors and heart attacks, enabling it to effectively deal with complex medical data.

2. **Ability to Handle High-Dimensional Data:** If there are many predictor variables (features) in the data, SVM can handle them smoothly, making it very useful for predicting heart attack occurrences.

3. **Robustness to Outliers:** SVM is known for its resistance to outliers in the data, ensuring stability in its performance even when dealing with unusual or unexpected data.

4. **Optimal Boundary Determination:** With its ability to choose the optimal boundary between two classes, SVM can provide high classification accuracy for new cases based on its past experiences.

Training the Model

- **Dataset** We used the HEART dataset with simple processing.

- **Feature Scaling:** Divide the dataset into training set and a test set, with 80% for training and 20% for testing.

- **Kernel Selection** We chose the sigmoid kernel because it is compatible with the target of the dataset.

- **Parameter Initialization** SVM model parameters, including the regularization parameter C, were initialized. We chose $C = 1$, and the kernel parameters (sigmoid) were set. These parameters control the trade-off between maximizing the margin and minimizing classification errors.

- **Optimization** The SVM algorithm optimizes the model parameters to maximize the margin between the support vectors (data points closest to the decision boundary) of different classes while minimizing classification errors.

- **Training** The SVM iteratively adjusts the model parameters to minimize the classification error and maximize the margin. This process continues until the model converges to the optimal solution.

Model Evaluation

After training, the performance of the SVM model is evaluated on a separate validation dataset. We obtained these results:

- **Accuracy of Training Data:** 0.8621951219512195 (86.22%)
- **Accuracy of Testing Data:** 0.8048780487804879 (80.49%)

Deployment

Once trained and evaluated, the SVM model can be deployed to make predictions on new data, aiding clinicians in decision-making and preventive care.

4.2.3 Model 3 (MLPClassifier)

Selection of Model

- **Handling Non-linearity:** Heart attack prediction involves numerous risk factors with potentially complex, non-linear relationships. The MLPClassifier, being a type of neural network, can model these non-linearities effectively through its multiple layers and non-linear activation functions.
- **Feature Interactions:** Neural networks, like the MLPClassifier, are adept at capturing interactions between features. This is crucial in medical data where combinations of risk factors (like high cholesterol and high blood pressure together) might significantly influence the risk of a heart attack.
- **Scalability:** The MLPClassifier can handle large datasets with many features, making it suitable for extensive medical datasets with numerous patient records and risk factors.

Training the Model

- **Dataset:** We used the HEART dataset.
- **Data Splitting:** Divide the dataset into two parts: a training set and a test set, with 80% for training and 20% for testing.
- **Model Initialization:**

  - **Architecture Design:** The architecture of the neural network was designed with 2 hidden layers. The first layer contains 6 neurons, and the second layer contains 5 neurons.
  - **Activation Functions:** The sigmoid activation function was chosen for the hidden layers due to its suitability for binary classification tasks.

- **Training:**
  - For each patient in the training set, the input data passes through the network, and each neuron calculates a weighted sum of its inputs and applies the activation function to produce its output.
  - Using optimization algorithms like stochastic gradient descent (SGD) or Adam, the model updates its parameters iteratively. These updates minimize the loss function, bringing the predicted probabilities closer to the actual outcomes observed in the training data.
  - This process continues until the model converges to the optimal solution.

- **Model Evaluation:** The performance of the MLPClassifier model is evaluated on a separate validation dataset.
  - **Accuracy of Training Data:** 0.848780487804878 (84.87%)
  - **Accuracy of Testing Data:** 0.7853658536585366 (78.53%)

- **Deployment:** Once trained and evaluated, the MLP model can be deployed to make predictions on new data, aiding clinicians in decision-making and preventive care

4.2.4 Comparative Analysis of Models for Heart Attack Prediction

Model	**Time (s)**	**Train Acc.**	**Test Acc.**	**Interpret.**
Log. Reg.	1.023	86.75%	81%	Low
SVM	2.749	86.21%	80%	Low
MLP	1.236	84.87%	78%	Low

Table 4.1: Comparison of Models for Heart Attack Prediction (Part 1)

Model	**Comp. Eff.**	**Scalability**	**Overfit.**
Log. Reg.	High	High	Low
SVM	Medium	Medium	Medium
MLP	Low	High	High

Table 4.2: Comparison of Models for Heart Attack Prediction (Part 2)

4.2.5 Logistic Regression Results

From the previous tables, we note that logistic regression excels in all elements in terms of speed of prediction, somewhat high accuracy, ability to scalability, ease of interpretability, no significant Computational efficiency, and lack of overfitting, which gives us great confidence in its results.

4.2.6 SVM Results

We note that SVM also provides high accuracy with fairly good performance in the remaining aspects of prediction speed, Interpretability, Scalability, Computational Efficiency, and overfitting.

4.2.7 MLP Classifier Results:

We note that MLP offers fairly good accuracy, speed in prediction, and high scalability, but it is difficult to interpretability, requires more Computational efficiency and is also vulnerable to overfitting

4.3 How to utilize our system

Our heart attack prediction system can be accessed through an easy-to-use website interface. Users can enter patient data and receive instant heart attack risk assessments. This intuitive approach ensures that healthcare professionals and individuals alike can easily integrate our system into their healthcare practices. Using this interface, the user can enter his data

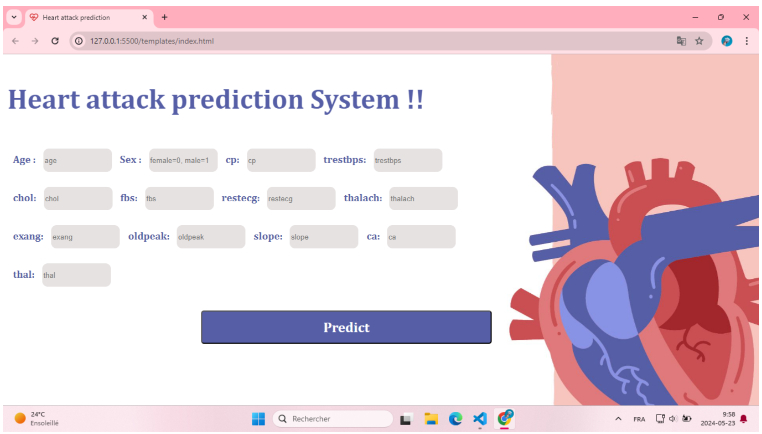

Figure 4.1: interface

This interface represents a case of a person who may have a heart attack

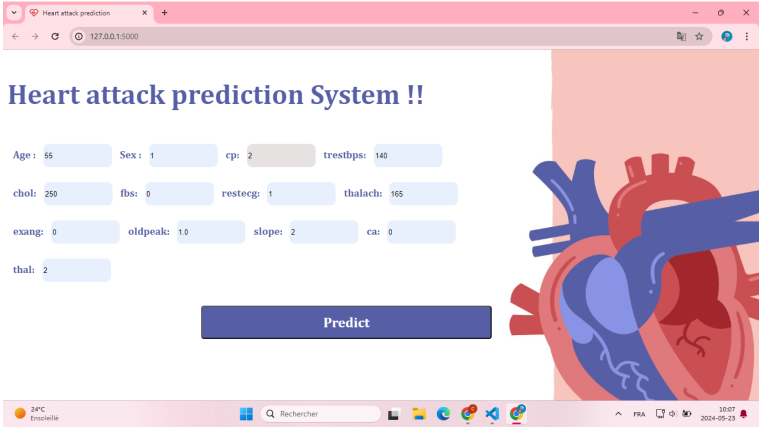

Figure 4.2: input

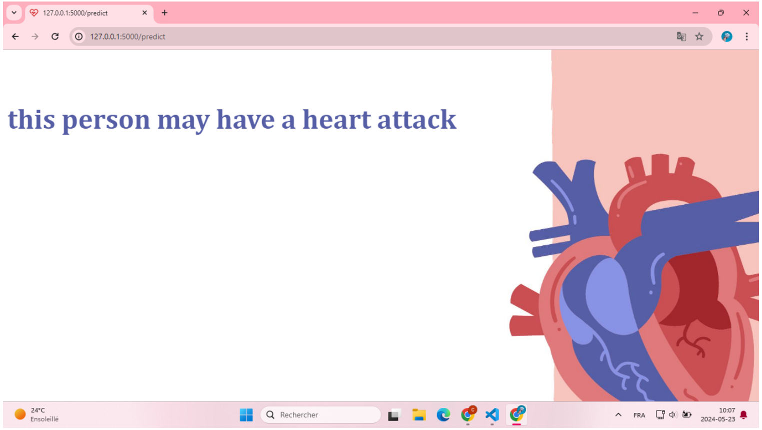

Figure 4.3: result

This interface represents a case of a person who may not have a heart attack.

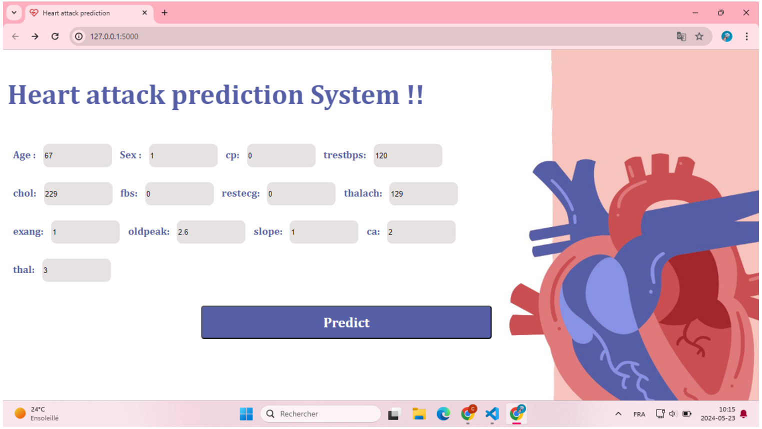

Figure 4.4: input

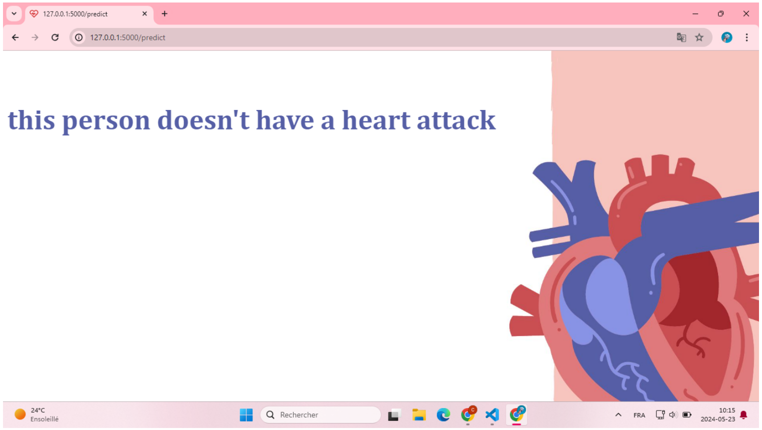

Figure 4.5: result

4.4 Conclusion

this chapter, we made our contribution regarding heart attack prediction, using different models based on machine learning techniques. We provided each model with predictions with fluctuating accuracy values, the highest we obtained was 86%, and developed a simple interface to use these models.

CONCLUSION AND FUTURE PERSPECTIVES

In conclusion, heart attack prediction is a critical aspect of modern healthcare, with the potential to significantly impact patient outcomes and healthcare management. Through this thesis, we explored the application of machine learning techniques, including SVM, MLPClassifier, and logistic regression, in predicting the occurrence of heart attacks.

By leveraging datasets, we developed and evaluated multiple models to accurately predict the probability of a heart attack. Our results showed promising accuracy rates, with the highest accuracy achieved reaching 86,These results underscore the potential of machine learning models to enhance traditional diagnostic methods and improve patient care.

One potential trend is the use of technology for continuous monitoring and early detection of health problems. Working with healthcare providers will be important to ensure the system works well in real-life settings so that valid and usable outcomes are predicted by extracting a complete medical report with machine learning models.

BIBLIOGRAPHY

[1] Mayo Clinic, "Heart attack," 2022. Retrieved from: https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106.

[2] R. N. Sweis and A. Jivan, "Infarctus du myocarde aigu." Édition Professionnelle Du Manuel MSD, February 21 2024.

[3] L. Lu, M. Liu, R. Sun, Y. Zheng, and P. Zhang, "Myocardial infarction: Symptoms and treatments," *Cell Biochemistry and Biophysics*, vol. 72, pp. 865–867, February 1 2015.

[4] Columbia AI, "Artificial intelligence (ai) vs. machine learning." https://ai.engineering.columbia.edu/ai-vs-machine-learning/, October 3 2023. CU-CAI.

[5] M. Crabtree, "What is machine learning? definition, types, tools & more." https://www.datacamp.com/blog/what-is-machine-learning, July 19 2023.

[6] G., "Introduction to deep learning," April 14 2023. GeeksforGeeks.

[7] DATAtab, "Logistic regression • simply explained," 2024. Accessed: 2024-05-19.

[8] WisdomML, "Logistic regression explained: Definition and examples," 2023. Accessed: 2024-05-19.

[9] DataTab, "Logistic regression tutorial." https://datatab.net/tutorial/logistic-regression. Accessed on: 2024-05-19.

[10] WisdomML, "Logistic regression explained: Definition and examples," 2024. Accessed: 2024-06-11.

[11] TutorialsPoint, "Scikit learn - support vector machines," 2023. Accessed: 2024-05-19.

[12] GeeksforGeeks, "Classification using sklearn multi-layer perceptron," 2023. Accessed: 2024-05-19.