PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

University of Mohamed El-Bachir El-Ibrahimi - Bordj Bou Arreridj

Faculty of Science and Technology

Department of Electronics

Presented to obtain

THE DIPLOMA OF MASTER

MCIL 5

FIELD: Electronics

Specialty: Electronic Industries

By:

- **CHERFAOUI** Mehdi
- **CHOUITER** Fouad

Entitled

Artificial intelligence application for diabetes prediction

Evaluated on: …………………………..

The evaluation committee is composed of:

Full Name	*Grade*	*Quality*	*Establishment*
DR AIB Samia	MCB	President	Univ BBA
Mme MELIZI Nora	MAA	Examiner	Univ BBA
Dr BEHIH Mohamed	MCB	Supervisor	Univ BBA
Pr BOUTTOUT Farid	Pr	co-Supervisor	Univ BBA

2024/2025

Acknowledgment

First and foremost, all praise and thanks be to almighty ALLAH for giving us health, wisdom, and patience, without him nothing of our work would have been done.

We would like to express our sincere gratitude to our supervisor Mr BEHIH Mouhamed for his suggestions and valuable directions during the writing of this thesis.

We would also like to express our dedication to our parents for their unlimited support, our teachers for helping us until the end, and our friends for their positive encouragement.

We express our deepest appreciation and respect to the esteemed president of the jury and the examiner for the honor they bestow upon us by reviewing this work.

Abstract

This study investigates multiple approaches for the prediction of type 2 diabetes Based on Biometric signs. Three supervised machine learning models (Logistic Regression, Random Forest, and XGBoost) were developed and evaluated based on their predictive accuracy, feature interpretability, and computational performance. Additionally a fuzzy logic system and a rule-based expert system approaches were implemented to simulate human reasoning and clinical decision-making. The models were applied to the Pima Indians Diabetes dataset and tested using a combination of statistical metrics and visual diagnostics. Results show that while machine learning algorithms outperform in terms of raw accuracy, fuzzy and expert systems offer greater transparency and explainability. This work highlights the complementary strengths of data-driven and rule-based systems in the design of intelligent diagnostic tools for healthcare.

Résumé

Cette étude explore plusieurs approches pour la prédiction du diabète de type 2 basée sur des signes biométriques. Trois modèles d'apprentissage automatique supervisé la régression logistique, la forêt aléatoire et XGBoost ont été développés et évalués selon leur précision prédictive, leur interprétabilité et leur performance computationnelle. En parallèle, un système à logique floue et un système expert basé sur des règles ont été mis en œuvre pour simuler le raisonnement humain et la prise de décision clinique. Les modèles ont été appliqués au jeu de données des Pima Indians et testés à l’aide de mesures statistiques et d’analyses visuelles. Les résultats montrent que, bien que les algorithmes d’apprentissage automatique soient plus précis, les systèmes flous et experts offrent une meilleure transparence et une meilleure explicabilité. Ce travail met en lumière la complémentarité entre les systèmes basés sur les données et ceux fondés sur des règles dans la conception d’outils diagnostiques intelligents pour le domaine médical.

الملخص

تتناول هذه الدراسة عدة طرق للتنبؤ بمرض السكري من النوع الثاني بناءً على العلامات الحيوية. تم تطوير وتقييم ثلاثة نماذج تعلم آلي خاضعة للإشراف Logistic Regression، Random Forest، و XGBoost بناءً على دقتها في التنبؤ، وقابلية تفسير خصائصها، وكفاءتها الحاسوبية. بالتوازي، تم تنفيذ نظام يعتمد على Fuzzy Logic ونظام خبير Expert System قائم على قواعد IF-THEN لمحاكاة التفكير البشري واتخاذ القرار الطبي. طُبّقت النماذج على مجموعة بيانات مرضى السكري من Pima واختُبرت باستخدام مؤشرات إحصائية وتحليلات بصرية. أظهرت النتائج أن خوارزميات التعلم الآلي تتفوق من حيث الدقة، بينما توفر الأنظمة المعتمدة على Fuzzy Logic و Expert Systems شفافية وتفسيرًا أفضل. تُبرز هذه الدراسة القوة التكميلية بين الأنظمة القائمة على البيانات وتلك المبنية على القواعد في تصميم أدوات تشخيص ذكية ومفسّرة في مجال الرعاية الصحية.

Table of Contents

List of figure:

General Introduction

Diabetes mellitus is a chronic metabolic disorder affecting millions worldwide, with severe complications if not diagnosed and managed early. The rising prevalence of diabetes has intensified the need for accurate and efficient prediction methods to enable timely intervention. Traditional diagnostic approaches, relying on manual analysis of clinical data, often face limitations in scalability and precision. In recent years, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as powerful tools for predictive healthcare, offering the potential to revolutionize diabetes detection through data-driven insights.

This research explores the application of AI in diabetes prediction, structured into three key chapters. The first chapter provides an overview of AI, its Levels, and its significance in modern healthcare. The second chapter delves into the core algorithms of AI and ML, examining techniques such as supervised learning, decision trees, neural networks, and ensemble methods that enhance predictive accuracy. Finally, the third chapter focuses on diabetes prediction Methods and approaches, presenting experimental results and evaluating their performance using real-world datasets. By leveraging AI, this study aims to develop a robust predictive framework that can assist clinicians in early diabetes identification, ultimately improving patient outcomes and reducing healthcare burdens. The findings contribute to the growing body of research on AI-driven diagnostics, highlighting its potential to transform diabetes care.

Finally, a general conclusion is going to summarize all obtained and reached knowledge and results

Chapter I:

General instructions on artificial intelligence

1. Introduction

This chapter introduces Artificial Intelligence (AI) as a branch of computer science aimed at enabling machines to mimic human intelligence in areas like reasoning, learning, and decision-making. It defines AI and explores its subdomains (Machine Learning, Deep Learning, NLP, etc.), different levels (ANI, AGI, ASI), and healthcare applications, particularly in diagnosis and treatment support like diabetes detection and MRI analysis.

2. Definition of Artificial Inteligent

There is no universal definition of the term artificial intelligence (AI) and it has different meanings depending on the sources and considered domain and subject. Generally, artificial intelligence term is a combination of two words: “Artificial” which indicates something which is not natural and created by humans and “intelligence” which means the ability to understand or think [1]. The development of computer systems or agents that are able to carry out tasks that would require intelligence if completed by people is the subject of artificial intelligence (AI), a multidisciplinary area of computer science. Perception, reasoning, learning, comprehension of normal language, and independent decision-making are frequently included in these tasks. AI systems are made to imitate or duplicate human cognitive processes, allowing robots to comprehend complex data, identify patterns, solve problems, and engage meaningfully with their surroundings, according to scholarly literature and significant surveys. Encyclopaedia Britannica defines AI as follows, Artificial intelligence (AI) is digital computer's or computer-controlled robot's capacity to carry out actions often performed by intelligent entities [2]. According to Oxford English Dictionary, AI is defined as the study and development of computer systems that can copy intelligent human behavior [3].

3. Types of Artificial Intelligence

The term AI refers to three types of artificial intelligence as indicated by Figure I.1, these types or subdomains are: Artificial intelligence level, machine learning, and deep learning.

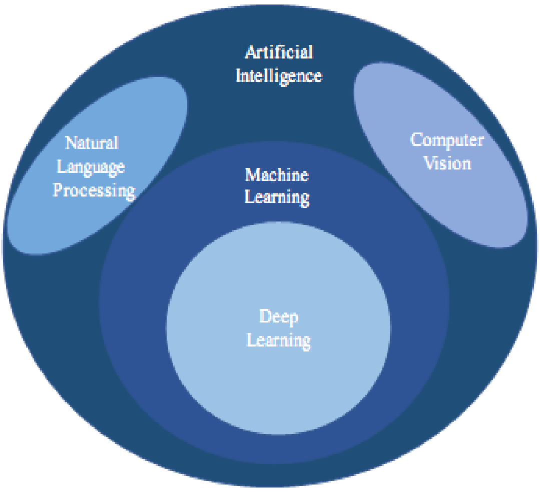


Figure I.1 AI subdomains [1]

3.1 Deep learning

In 2000, Igor Aizenberg was the first to explain deep learning (DL). DL is a subset of machine learning, which falls under artificial intelligence. DL employs algorithms that draw inspiration from the composition and operations of the human brain. As indicated by Figure I.2, the main difference between DL and ML is that DL requires more training time and less testing time. DL is more efficient than machine learning at handling vast amounts of input data to conduct an analysis [4], [5]. A DL performance graph has a great deal of data and is measurable high. DL is a branch of machine learning that takes a nontraditional stance on how artificial intelligence ought to operate. While the other approaches can only perform specific jobs and perform better with data organized in a spreadsheet, it uses a layered architecture of artificial neural networks that are modeled after the neural network of the human brain. Deep learning can be applied to increasingly complicated tasks and analyze high-dimensional data, including sounds and images. [4].

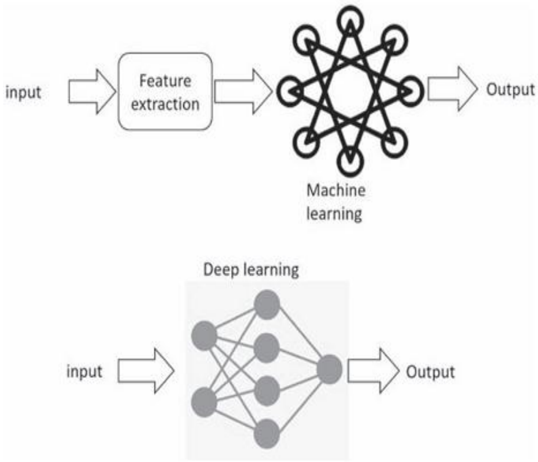


Figure I.2 *Difference between ML and DL [5]*

3.2 Machine Learning

Arthur Samuel first proposed machine learning (ML) in 1959. It is described as the capacity of a machine to learn. Compared to deep learning (DL) models, machine learning (ML) requires less training time. It is better to employ a machine learning method when there is fewer data. Once the data count beyond a certain saturation point, the performance will be diminished. For examination of a vast array of types and unstructured data, a deep learning method is recommended [4].

3.3 Artificial intelligence Levels

As defined in [6], AI level refers broadly to the simulation of human intelligence in machines, encompassing any technique enabling computers to mimic human reasoning, learning, and problem-solving. AI level covers a spectrum of methods and approaches, from expert systems based on explicit rule encoding to data-driven methodologies, e.g.,

1. Expert Systems: rule-based systems that mimic human decision-making (Figure I.3_a),

2. Fuzzy Logic: mathematical approach to handle uncertainty and imprecision (Figure I.3_b),

3. Rule-Based Systems: systems that use rules to reason about data (Figure I.3_c),

4. Planning and Scheduling: systems that plan and schedule tasks and resources (Figure I.3_d),

5. Constraint Programming: systems that solve problems with constraints,

6. Knowledge Representation: Methods for representing knowledge in a machine-readable format, etc.

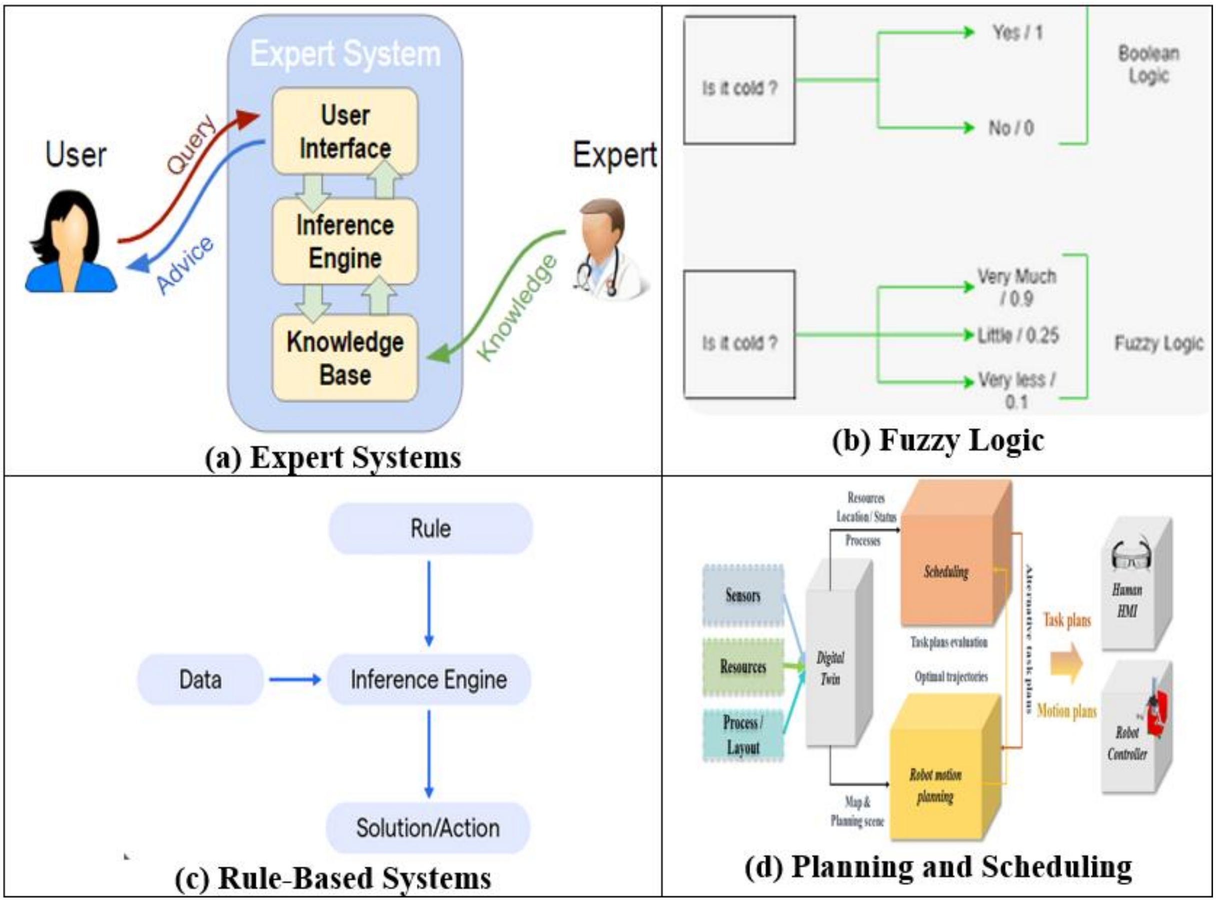


Figure I.3 *Examples of Approach and Methods of AI Level*

3.2.1 Expert System

An expert system is a type of artificial intelligence that mimics the decision-making abilities of a human expert in a specialized field, such as medicine, engineering, or finance. It consists of a knowledge base, which stores facts and rules gathered from human experts, and an inference engine, which processes this information to solve problems or provide recommendations. When a user inputs a query, the system analyzes the data, applies logical reasoning (using methods like forward or backward chaining), and delivers an answer along with an explanation of its reasoning. Expert systems are widely used in diagnostics, troubleshooting, and decision support because they can handle complex tasks efficiently, consistently, and without human bias, making them valuable tools in industries where expert knowledge is crucial but not always readily available.

3.2.2 Fuzzy Logic

The term fuzzy refers to things that are not clear or are vague. In the real world many times we encounter a situation when we can't determine whether the state is true or false, their fuzzy logic provides very valuable flexibility for reasoning. In this way, we can consider the inaccuracies and uncertainties of any situation, the deference between it and Boolean Logic can be clearly shown in figure I.4.

Fuzzy Logic is a form of many-valued logic in which the truth values of variables may be any real number between 0 and 1, instead of just the traditional values of true or false. It is used to deal with imprecise or uncertain information and is a mathematical method for representing vagueness and uncertainty in decision-making.

Fuzzy Logic is based on the idea that in many cases, the concept of true or false is too restrictive, and that there are many shades of gray in between. It allows for partial truths, where a statement can be partially true or false, rather than fully true or false.

Fuzzy Logic is used in a wide range of applications, such as control systems, image processing, natural language processing, medical diagnosis, and artificial intelligence.

The fundamental concept of Fuzzy Logic is the membership function, which defines the degree of membership of an input value to a certain set or category. The membership function is a mapping from an input value to a membership degree between 0 and 1, where 0 represents non-membership and 1 represents full membership.

Fuzzy Logic is implemented using Fuzzy Rules, which are if-then statements that express the relationship between input variables and output variables in a fuzzy way. The output of a Fuzzy Logic system is a fuzzy set, which is a set of membership degrees for each possible output value.

In summary, Fuzzy Logic is a mathematical method for representing vagueness and uncertainty in decision-making, it allows for partial truths, and it is used in a wide range of applications. It is based on the concept of membership function and the implementation is done using Fuzzy rules.

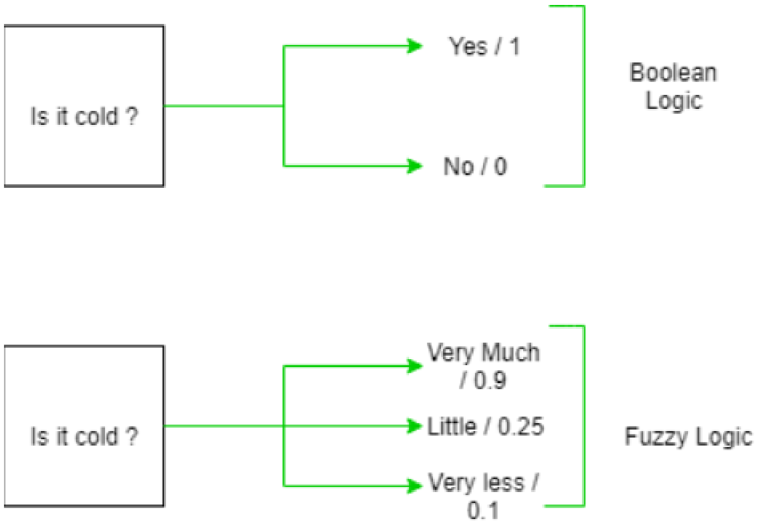


Figure I.4 *Boolean Logic vs Fuzzy Logic*

3.2.3 Case-Based reasoning

Case-Based Reasoning (CBR) is an artificial intelligence approach where problems are solved by recalling and adapting solutions from similar past cases, much like how humans learn from experience. Instead of relying solely on rigid rules or abstract models, CBR stores a collection of previous cases-each containing a problem, its solution, and the outcome-and retrieves the most relevant one when a new problem arises. The system then adjusts the old solution to fit the current situation, tests its effectiveness, and updates its case library with the new experience for future use. This makes CBR particularly useful in fields like medical diagnosis, customer support, and legal reasoning, where past examples often provide practical guidance for handling new but familiar challenges.

3.4 Natural language processing

Artificial intelligence is basically about mimicking human mental abilities. A human mind's most basic capacity is language. They can communicate and comprehend the needs of others because to it. Thus, a lot of scholars have been attempting to improve natural language processing (NLP) in recent years. Comprehending and processing human language is the aim of natural language processing (NLP). It combines language ideas, artificial intelligence, and computer science [5].

NLP has recently emerged as a key component of AI. Voice interfaces, chatbots, and many other technologies have made our lives easier. NLP is separated into levels to depict the events that take place within an NLP system. The linguistic levels of NLP are displayed in Figure I.5.

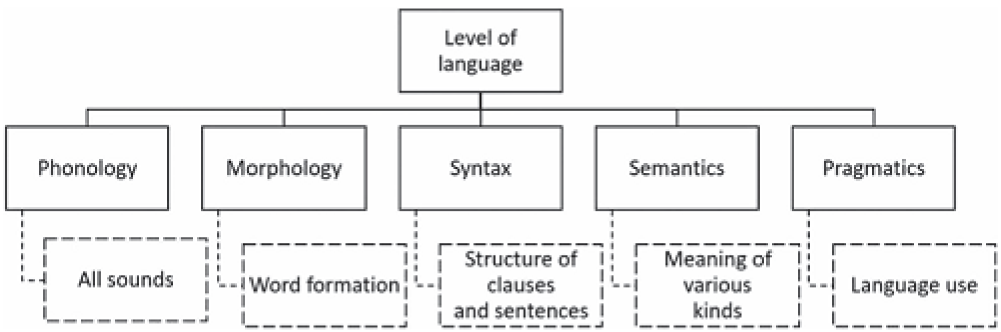


Figure I.5: *Natural Language Processing Levels [6]*

3.5 Computer vision

This makes it possible for computers to perceive, observe, and interpret picture and video data. For instance, computer vision aids in illness diagnosis by identifying patterns in a variety of medical imaging data.

4. Comparative study between AI, ML and DL

ML and DL being the bedfellows of AI. Table I.1 give a comparative study for better understanding.

Tableau I.1: *comparative outlook – AI vs ML vs DL [1].*

Properties	AI	ML	DL
Coined in the year	1956	1959	2000
Coined by	John McCarthy	Arthur Samuel	Igor Aizenberg
Conception	An umbrella term with ML and DL as prominent integrants.	Subset/evolution of AI	Subset/evolution of ML. The conception of how deep is machine learning.
Data	-	Data used is in the structured form	Data used is in the form of an artificial neural network.
Definition	The ability of a machine to mimic human behaviour through a set of rules (algorithm). Basically ability of a machine to think.	The ability of a machine to learn on its own, by its own mistakes and past experience by using statistical methods.	A kind of Machine learning that teaches machines to do what comes naturally to humans by using artificial neural networks (inspired by the structure and function of neurons present in the human brain). A large amount of data is fed until it learns by itself.
Sub categories	3 broad types of AI are – Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI)	3 broad types of ML are – Supervised Learning, Unsupervised Learning and Reinforcement Learning	3 broad categories of deep neural network models are – Pre-trained models, Convolutional Neural Networks (CNN) model, Recursive Neural Networks, and Recurrent Neural Networks (RNN) model
Execution time	-	The training time needed for a machine using the ML technique is less than DL.	The training time needed for a machine using DL is more than ML.
Data dependency	-	For a small amount of data, ML proves to be the best technique to train your machine.	For a large amount of data, DL proves to be the best technique to train your machine.
Efficiency	The efficiency of AI = Efficiency of ML + DL	Less efficient than DL as it can't work for a larger amount of data.	More efficient than ML as it can easily work for larger sets of data.
Data analytics	-	ML consists of thousands of data points.	Big data: millions of data points.

5. Narrow, General and Superintelligent Categories of AI

Modern definitions of AI recognize its broad spectrum ranging from **Narrow AI**, which is specialized for designated tasks, to the more hypothetical forms of **General AI** which could perform any intellectual task a human can, and **Super intelligent AI** which would exceed human intelligence in all respects. These all three categories and links between them are represented by Figure I.6.

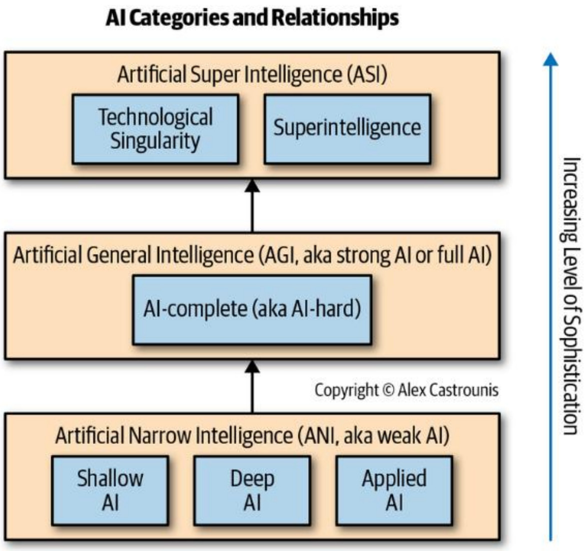


Figure I.6 *Artificial intelligence categories (Narrow, General and Super intelligent) and relationships*

5.1 Narrow artificial intelligence

Narrow AI encompasses systems specialized to perform specific tasks or solve problems within well-defined domains. These systems do not possess consciousness or general reasoning abilities and are constrained by their programming or training data. Narrow AI is characterized by: (1) task-specific: algorithms optimized for clear objectives such as image classification, language translation, or recommendation, (2) absence of generalization: unable to transfer learning or problem-solving beyond their domain, and (3) prevalence: ubiquitous in current industry deployments, including digital assistants, fraud detection, recommendation engines, and medical image analysis.

5.2 General artificial intelligence

General AI refers to intelligence that can learn, understand, and apply reasoning flexibly and adaptively across a wide range of tasks at a level comparable to or exceeding that of an average human. General AI would demonstrate robust reasoning, abstraction, creativity, transfer learning, and potentially even self-awareness. **Key features of** General AI are it can solve novel problems in unfamiliar environments by extrapolating from prior experience, and owns capacity for continual learning and meta-learning, integrating new knowledge dynamically.

5.3 Superintelligent artificial superintelligence

Superintelligent AI denotes AI systems whose intellectual capacities vastly exceed those of the best human minds across virtually all domains, including scientific creativity, general wisdom, and social manipulation. These AI category is capable to rapidly improve its own architecture and algorithms, potentially entering a feedback loop of accelerating intelligence gains. Furthermore, it not just faster or more accurate, but qualitatively more capable in reasoning, planning, strategic thinking, and coordination.

6. Healthcare Applications of Artificial Intelligence

In the medical field, AI uses a collection of algorithms on a large amount of medical data to understand its characteristics and provide information to patients and medical professionals. AI approaches have been employed extensively in medical treatment in recent years because to the development in the availability of vast amounts of medical data. Finding connections between genetic codes, creating virtual surgical robots, or optimizing the effectiveness of a larger healthcare facility, such as a hospital, are all examples of how artificial intelligence is transforming healthcare. Even the question of whether AI will eventually take the place of human is fueled by the extensive usage of AI in healthcare. AI is intended to automate diagnosis processes and support human specialists in making decisions, not to replace human physicians. Instead of replacing humans, technologies created by fusing AI and human intelligence are meant to supplement and improve human processing [3].

6.1 Diabetes Mellitus

Diabetes mellitus is an incurable, metabolic disorder triggered by faulty insulin secretion and insulin resistance along with alterations of protein and lipid metabolism, which results in chronic hyperglycaemia. Prolonged hyperglycaemic conditions lead to glycation of proteins which subsequently leads to secondary pathological manifestations that affect the eyes, kidneys, nerves and arteries . Glycation carried out by monosaccharides damages cells by impairing the function of target proteins, adds to oxidative stress and activates lethal signal transduction pathways. Symptoms of diabetes mellitus include frequent urination, unexplained weight loss, excessive thirst, fatigue, numbness or tingling in the feet and hand tips and dry skin and sores.[7]

6.1.1 Classification of Diabetes Mellitus

Prediabetes is an intermediate state of hyperglycemia with glycemic parameters above normal but below the diabetes threshold. Characterization of the underlying pathophysiology is much more developed in type 1 diabetes mellitus than in type 2 diabetes mellitus. The global disease burden of the major types of diabetes is shown in Figure I.7.

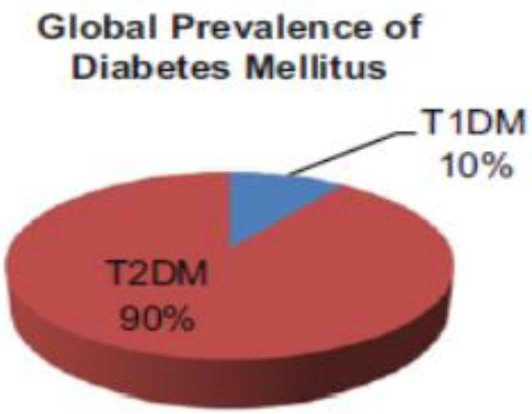


Figure I.7 *Global prevalence of diabetes mellitus.*

It is etiologically classified into three categories:

1. Type 1 diabetes mellitus is an autoimmune disease contributing to approximately 5% of diabetic cases with a high prevalence in adolescents. It is majorly caused due to destruction of pancreatic islet cells via humeral response and T-cell mediated inflammatory response. The presence of autoantibodies such as GAD65 glutamic acid decarboxylase, autoantibodies to insulin, IA2 and IA2β protein tyrosine phosphatase and ZnT8A zinc transporter protein against the pancreatic β cells is the diagnostic of this disease. These individuals are highly susceptible

to ketoacidosis. Risk factors for type 1 diabetes are family history, autoantibodies, environmental factors, dietary deficiencies such as low vitamin D and specific geographic locations such as Sweden and Finland.

2. Type 2 diabetes, also termed as non-insulin-dependent condition that majorly affects the adults (20–79 yr. old), contributing to 95% of prevailing diabetes cases, is associated with obesity and insulin resistance that leads to decreased insulin production overtime. Study suggests that insulin resistance might improve with weight loss and treatment of hyperglycemia but it can rarely be brought to normal levels. There are multiple risk factors for type 2 diabetes and prediabetic individuals: obesity, physical inactivity, age, susceptible race, hypertension, polycystic ovarian syndrome, gestational diabetes mellitus and anomalous cholesterol and triglyceride levels.

3. Gestational diabetes mellitus is observed in 7% of all pregnant women worldwide. It often occurs due to hormonal imbalances that occur during pregnancy, leading to insulin resistance that subsides after pregnancy. Common risk factors for GDM are age, obesity, family history and susceptible race. [7]

6.1.2 Diagnosis of Diabetes Mellitus

Diagnosis of diabetes mellitus is conducted on the basis of plasma glucose levels which comprises of either the fasting plasma glucose (FPG) and the 2-hr plasma glucose (2-hr PG) level during a 75-g oral glucose tolerance test (OGTT) or the A1C test. A random plasma glucose test is performed for individuals showing typical symptoms of hyperglycemia. Table 1 summaries the levels of these diagnostic tests. [7]

Tableau I.1. *List of pathological investigation for diabetes mellitus*

S. no.	Test	Criteria
1.	FPG	≥126 mg/dL (7.0 mmol/L) *Fasting conditions referred here as zero calorie consumption for 8 h or more
2.	2-h PG during OGTT	≥200 mg/dL (11.1 mmol/L)
3.	A1C or HbA1c	≥6.5% (48 mmol/Mol)
4.	RPG	≥200 mg/dL (11.1 mmol/L)

6.1.3 Diabetes Management

The early diagnosis and management of diabetes mellitus is mandatory to prevent lethal complications associated with diabetes. Prolonged diabetes mellitus can lead to neuropathy, nephropathy, retinopathy, hearing impairments, skin infections and cardiovascular vascular diseases such as coronary artery disease, atherosclerosis and heart strokes. [7]

Pharmaceutical Therapy

The basic medication provided to individuals suffering from type 1 diabetes mellitus is insulin treatment to counter this autoimmune disease. In patients where this therapy is inefficacious, beta cell transplants and autoimmune blocking drugs are in clinical trials. Metformin is the most commonly used medication, along with sodium glucose co-transporter 2 inhibitors sold under the names polarizing, dapagliflozin, amylin analogues, glucagon-like peptide 1 receptor agonists (eventide, liraglutide) and di-peptide peptidase-4 (saxagliptin, vildagliptin) inhibitors. Nonglycaemic treatments employ angiotensin-converting enzyme inhibitors, such as ramipril, often used in the cases of patients with nephropathy. [7]

Self-Monitoring

Structured and personalized self-monitoring of blood glucose (SMBG) is an organized method of observing glucose levels that reveals glycemic patterns throughout the day. Presently it is theorized that glycemic variability contributes to diabetes complications independently of glycosylated hemoglobin (HbA1c) levels. To assess diurnal glucose excursions, SMBG has also been established as a useful tool as it helps to monitor diet control and treatment response and in general increases a patient’s understanding of hypoglycemia, thereby reducing their anxiety. Although the pharmacological management of diabetes is sought after and provides several therapeutic opportunities, particularly in the type 2 diabetes mellitus, the changes in the lifestyle are essential: by maintaining proper diet and physical activities, one can reduce obesity associated with this type of diabetes. [7]

6.2 MRI (Magnetic Resonance Imaging)

Artificial Intelligence is transforming Magnetic Resonance Imaging (MRI) in healthcare by improving speed, accuracy, and diagnostic capabilities. AI-powered algorithms can accelerate MRI scans by reconstructing high-quality images from under sampled data, reducing patient scan time while maintaining diagnostic precision. Machine learning models assist in automated image segmentation, identifying tumours, lesions, and anatomical structures with high accuracy, aiding radiologists in early detection of diseases like cancer, strokes, and neurodegenerative disorders. AI also helps in artefact correction, removing motion-related distortions to enhance image clarity. Additionally, predictive AI analyses MRI data alongside patient history to forecast disease progression and recommend personalized treatment plans. By integrating AI, MRI diagnostics become faster, more precise, and less reliant on manual interpretation, ultimately improving patient outcomes.

7. Advantages and disadvantages

7.1 Advantages

- The possibility of human error is significantly decreased with AI's help.
- When it comes to experiments or expeditions where human life may be in danger, like defusing a bomb or conducting any kind of space mission or experiment, AI is a considerably bigger risk-taker.
- AI is capable of performing repetitive activities like email reminders, thank-you notes, document verification, and spam/starred email classification. To put it simply, it can do any odd task that people could become disinterested in.
- Unlike humans, AI is able to work continuously without breaks or vacations.
- AI makes decisions far more quickly than people do.
- The medical field has changed significantly as a result of AI, as it is now possible to identify and treat serious illnesses.

7.1 Disadvantages

- The human element, which is crucial to the healthcare industry, particularly in medical surgery, is diminished when AI is used to support healthcare systems [6].
- AI has raised unemployment since it is faster and more efficient than people.
- Because ongoing advancements require modifications to circuits and hardware, AI has proven to be fairly costly. We are entitled to a substantial amount of money due to frequent modifications.
- Machines are entirely dependent on the data provided and, in contrast to humans, are incapable of independent thought.
- AI-enabled home appliances are making people sleepy and dependent on them.

8. Conclusion

The chapter concludes by highlighting the transformative potential of AI in healthcare, especially for diagnostics and decision support. Despite advantages like speed and consistency, AI also has drawbacks such as high cost and reduced human interaction. The key takeaway is that AI should be seen as a complement to human expertise rather than a replacement.

Chapter II:

Methods and algorithms

1. Introduction

This chapter presents the foundational AI techniques used for classification problems, particularly in healthcare. It starts with Machine Learning categories (supervised, unsupervised, reinforcement), then goes deep into specific classification algorithms: Logistic Regression, Random Forest, and XGBoost. It also explains fuzzy logic and expert systems as alternative AI strategies that simulate human reasoning.

2. Machine Learning Applications

2.1 Supervised Learning

Supervised learning is one of the popular types of ML. The term “supervised” refers to any type of control or guiding. The data used in this case is labeled data that is closely supervised by the relevant output so that changes in the output can be identified. When the ML algorithm uses the same set of data with the same properties, this is referred to as supervised learning. In supervised learning, the algorithm is given a tiny data sheet comparable to the original one to train itself on. The program then attempts to determine a relationship between all of the data presented. It attempts to discover the relationship between the data using the cause-and-effect approach. As a result, the algorithm can continue to learn new things as it grows, The figure II.1 can summarize the explanation. It is further subdivided into two sections: regression and classification. [8]

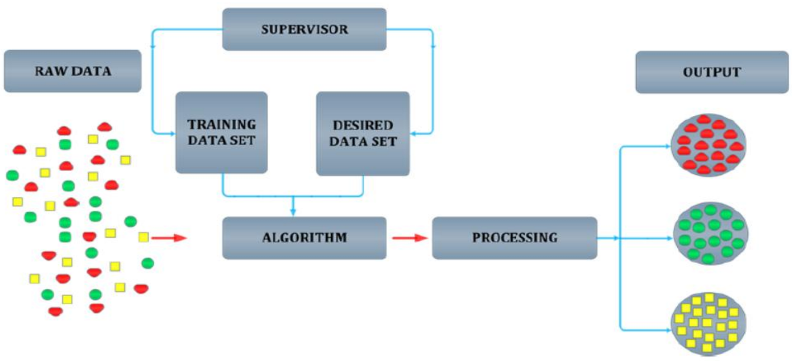


Figure II.1 *Supervised learning*

2.1.1 Regression

It maintains learning from the data sheet supplied and, as a result, provides consistent output for fresh data that is fed to the algorithm. It establishes a link between the dependent and independent variables. It anticipates actual values such as temperature, money, height and price. It is used to forecast weather, the stock market and other events. [8]

2.1.2 Classification

The method receives data as input and uses it to categorize the output in supervised learning. When the input is in the form of categories, it is used to predict the output. A classification model predicts the outcome as labels such as happy or sad, pass or fail, and so on. Classification is regarded as a very successful data mining approach for separating different types of data. The following are the most significant algorithms used in supervised learning: [8]

1) Linear Regression;
2) Logistic Regression;
3) Support Vector Machines SVM;
4) XGBoost;
5) Random Forest;
6) K-Nearest Neibors KNN.

2.2 Unsupervised learning

As a training set for unsupervised learning, unlabeled data is used. As a result, it is machine-readable and does not require human involvement. This makes it ideal for large data sheets. Because no human intervention is necessary, it functions by developing some hidden patterns. Because they are unlabeled, they can be readily included in any dataset by simply modifying the concealed pattern. In comparison to supervised learning, they can tackle more complicated tasks. The algorithm receives input in unsupervised learning, but the output is neither predetermined nor is it rewarded for accurate judgment or output. It cannot be applied immediately to any situation since it has no preset output. Assume that the algorithm is given some photographs of boys and girls, but has no foreknowledge on how to discriminate between them. With no preexisting dataset, it will then cluster the photos into groups based on characteristics, the figure II.2 can give a graphic explanation. It is further divided into two parts: clustering and association. [8]

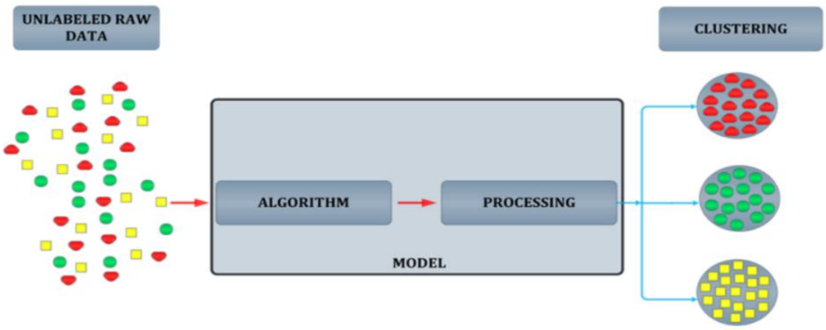


Figure II.2 *Unsupervised Learning*

2.2.1 Clustering

Clustering is a significant unsupervised learning approach. It is the technique of detecting a hidden pattern in unclassified raw data. These algorithms refine the data and, if necessary, construct clusters or groups from it. The produced clusters are of good quality, with extensive intra-class analogy and shallow inter-class analogy.

2.2.2 Association

Association learning algorithms discover relationships between things in a huge data collection. The interdependence of data is determined in this scenario, and mapping is performed accordingly. Market basket analysis is one such association method used by large retailers to discover associations between the store's various items. If a client purchases baby food, they are more likely to purchase baby creams, lotions and diapers.

A few unsupervised learning approaches are mentioned below: [8]

1) K-means clustering;

2) Hierarchical clustering;

3) Principal component analysis.

2.3 Reinforcement learning

The most intriguing of all learning approaches is reinforcement learning. It involves learning the job and determining how to gain the most rewards. It is essentially a trial and error strategy in which the algorithm learns on its own by trying several methods and determining which way gives it the most rewards. Reinforcement learning differs from supervised learning in that it always works with new and fresh data, whereas supervised learning is given tagged data.

A few reinforcement learning approaches are mentioned below: [8]

1) Monte Carlo;
2) Q-learning;
3) Deep Q network.

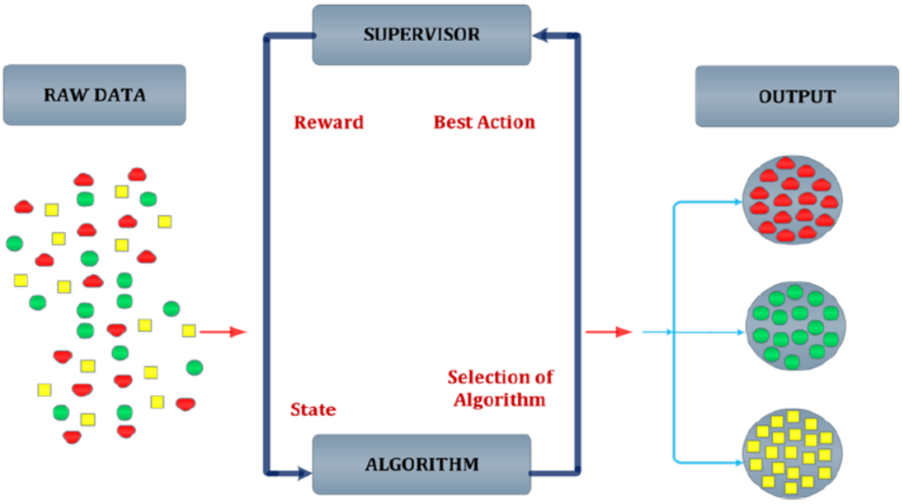


Figure II.3 *Reinforced Learning*

3. Machine Learning Algorithms used for classification problem:

3.1 Logistic Regression

3.1.1 Definition

Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyze the relationship between two data factors. The article explores the fundamentals of logistic regression, its types and implementations.

Logistic regression is used for binary classification where we use sigmoid function that takes input as independent variables and produces a probability value between 0 and 1.

For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 otherwise it belongs to Class 0. It's referred to as regression because it is the extension of linear regression but is mainly used for classification problems. [9]

3.1.2 Types of Logistic Regression

On the basis of the categories, Logistic Regression can be classified into three types:

Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.

Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as “cat”, “dogs”, or “sheep”

Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as “low”, “Medium”, or “High”. [9]

3.1.3 Assumptions of Logistic Regression

We will explore the assumptions of logistic regression as understanding these assumptions is important to ensure that we are using appropriate application of the model. The assumption include:

Independent observations: Each observation is independent of the other. meaning there is no correlation between any input variables.

Binary dependent variables: It takes the assumption that the dependent variable must be binary or dichotomous, meaning it can take only two values. For more than two categories SoftMax functions are used.

Linearity relationship between independent variables and log odds: The relationship between the independent variables and the log odds of the dependent variable should be linear.

No outliers: There should be no outliers in the dataset.

Large sample size: The sample size is sufficiently large [9]

3.1.4 How it works

The logistic regression model transforms the linear regression function continuous value output into categorical value output using a sigmoid function, which maps any real-valued set of independent variables input into a value between 0 and 1. This function is known as the logistic function.[9]

Let the independent input features be:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \quad (II.1)$$

and the dependent variable is Y having only binary value i.e. 0 or 1.

$$y = \begin{cases} 0 & if\ Class\ 1 \\ 1 & if\ Class\ 2 \end{cases} \quad (II.2)$$

then, apply the multi-linear function to the input variables X.

$$z = \left(\sum_{i=1}^{n} \omega_i\, x_i\right) + b \quad (II.3)$$

Here x_i is the ith observation of X, ω_i=$[\omega_1, \omega_2, \omega_3, \cdots, \omega_m]$ is the weights or Coefficient, and b is the bias term also known as intercept. simply this can be represented as the dot product of weight and bias.

$$z = \omega \,.X + b \quad (II.4)$$

Whatever we discussed above is the linear regression.

Now we use the sigmoid function where the input will be z and we find the probability between 0 and 1. i.e. predicted y.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \qquad (II.5)$$

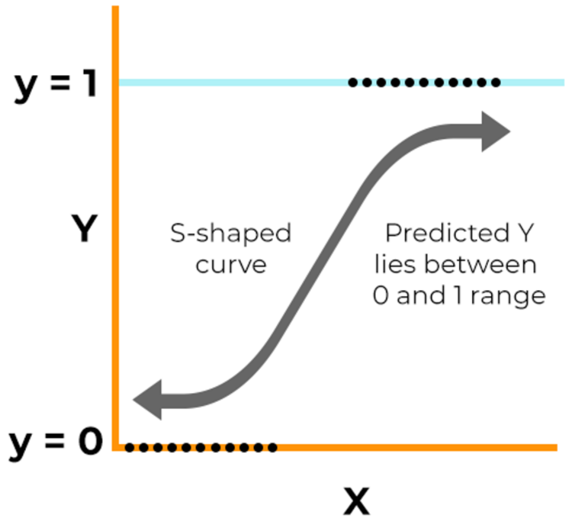


Figure II.4 *Sigmoid Function*

As shown above Figure II.4, the figure sigmoid function converts the continuous variable data into the probability i.e. between 0 and 1.

- $\sigma(z)$ tends towards 1 as $z \to \infty$
- $\sigma(z)$ tends towards 0 as $z \to -\infty$
- $\sigma(z)$ is always bounded between 0 and 1

Where the probability of being a class can be measured as:

$$P(y = 1) = \sigma(z) \qquad (II.6)$$

$$P(y = 0) = 1 - \sigma(z) \qquad (II.7)$$

The odd is the ratio of something occurring to something not occurring. it is different from probability as the probability is the ratio of something occurring to everything that could possibly occur. so odd will be:

$$\frac{p(x)}{1-p(x)} = e^{z} \qquad (II.8)$$

Applying natural log on odd. then log odd will be:

$$\log\left[\frac{p(x)}{1-P(x)}\right] = z \qquad (II.9)$$

$$\log\left[\frac{p(x)}{1-P(x)}\right] = \omega.X + b \qquad (II.10)$$

$$\frac{p(x)}{1-P(x)} = e^{\omega.X+b} \qquad (II.11)$$

$$p(x) = e^{\omega.X+b}.(1-p(x)) \qquad (II.12)$$

$$p(x) = e^{\omega.X+b} - e^{\omega.X+b}.p(x) \qquad (II.13)$$

$$p(x) + e^{\omega.X+b}.p(x) = e^{\omega.X+b} \qquad (II.14)$$

$$p(x)(1+e^{\omega.X+b}) = e^{\omega.X+b} \qquad (II.15)$$

$$p(x) = \frac{e^{\omega.X+b}}{1+e^{\omega.X+b}} \qquad (II.16)$$

then the final logistic regression equation will be:

$$p(X;b;\omega) = \frac{e^{\omega.X+b}}{1+e^{\omega.X+b}} = \frac{1}{1+e^{-\omega.X+b}} \qquad (II.17)$$

3.2 Random Forest (RF)

3.2.1 Definition

RF is also an algorithm-based on supervised/administered learning. The “forest” is a group of decision-making trees, normally trained by the “bagging” method. The common objective of the bagging procedure is to increase the overall result. The advantage of RF is the majority of current ML systems that can be used for both grading and regression problems. RF increases the randomness of the model as the trees grow. It seeks the best distinct feature among a certain number of features instead of finding for the most notable features while splitting a node. This produces a broad range that usually leads to a better model. Therefore, the algorithm for splitting a node takes into account only a random subset of characteristics in the random forest. We can randomize trees by using the random thresholds per feature instead of looking for the best possible threshold. Random forest is type of an ensemble classifier based on the idea of randomization it develops a decision tree having a group of independent and non-identical data. [10]

3.2.2 Key features of Random Forest

- **Handles Missing Data**: Automatically handles missing values during training, eliminating the need for manual imputation.
- Algorithm ranks **features based on their importance in making predictions** offering valuable insights for feature selection and interpretability.
- **Scales Well with Large and Complex Data** without significant performance degradation.
- Algorithm is versatile and can be applied to both classification tasks (e.g., predicting categories) and regression tasks (e.g., predicting continuous values). [10]

3.2.3 Assumptions of Random Forest

- **Each tree makes its own decisions**: Every tree in the forest makes its own predictions without relying on others.
- **Random parts of the data are used**: Each tree is built using random samples and features to reduce mistakes.
- **Enough data is needed**: Sufficient data ensures the trees are different and learn unique patterns and variety.
- **Different predictions improve accuracy**: Combining the predictions from different trees leads to a more accurate final results. [10]

3.2.4 How it works

The random Forest algorithm works in several steps:

- Random Forest builds multiple decision trees using random samples of the data. Each tree is trained on a different subset of the data which makes each tree unique.
- When creating each tree the algorithm randomly selects a subset of features or variables to split the data rather than using all available features at a time. This adds diversity to the trees.
- Each decision tree in the forest makes a prediction based on the data it was trained on. When making final prediction random forest combines the results from all the trees.
    - For classification tasks the final prediction is decided by a majority vote. This means that the category predicted by most trees is the final prediction.
    - For regression tasks the final prediction is the average of the predictions from all the trees.

- The randomness in data samples and feature selection helps to prevent the model from over fitting making the predictions more accurate and reliable like it's shown in figure II.5. [10]

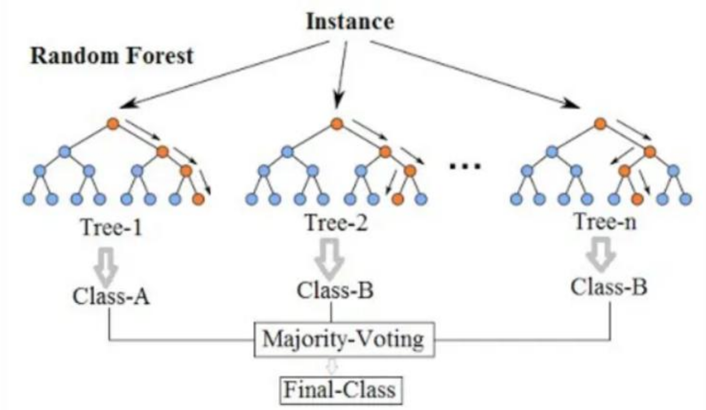


Figure II.5. *Random Forest*

3.3 XGBoost

3.3.1 Definition

In machine learning we often combine different algorithms to get better and optimize results. Our main goal is to minimize loss function for which, one of the famous algorithm is XGBoost (Extreme boosting) technique which works by building an ensemble of decision trees sequentially where each new tree corrects the errors made by the previous one. It uses advanced optimization techniques and regularization methods that reduce over fitting and improve model performance.

In this article, we move beyond the basics and focus on the practical implementation of XGBoost, exploring how to apply it effectively in real-world machine learning tasks. [11]

3.3.2 Parameter

Before jumping to the implementation of XG Boost we need to understand about the parameters used in XGBoost for model optimization.

- **Learning Rate (eta):** An important variable that modifies how much each tree contributes to the final prediction. While more trees are needed smaller values frequently result in more accurate models.

- **Max Depth:** This parameter controls the depth of every tree, avoiding overfitting and being essential to controlling the model's complexity.
- **Gamma**: Based on the decrease in loss it determines when a node in the tree will split. The algorithm becomes more conservative with a higher gamma value, avoiding splits that don't decreases the loss. It help in managing tree complexity.
- **Subsample**: Manages the percentage of data that is sampled at random to grow each tree hence lowering variance and enhancing generalization. Setting it too low could result in under fitting.
- **Colsample By tree**: Establishes the percentage of features that will be sampled at random for growing each tree.
- **N estimators:** Specifies the number of boosting rounds.
- **lambda (L2 regularization term) and alpha (L1 regularization term)**: Control the strength of L2 and L1 regularization respectively. A higher value results in stronger regularization.
- **Min child weight**: Influences the tree structure by controlling the minimum amount of data required to create a new node.
- **Scale pos weight**: Useful in imbalanced class scenarios to control the balance of positive and negative weights. [11]

3.3.3 How it works

XGBoost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models like it's shown in figure II.6.

The models that form the ensemble, also known as base learners, could be either from the same learning algorithm or different learning algorithms. Bagging and boosting serve as two widely used ensemble learners. Though you can apply these techniques with several statistical models, decision trees dominate their usage.

Let's briefly discuss bagging before taking a more detailed look at the concept of gradient boosting. [11]

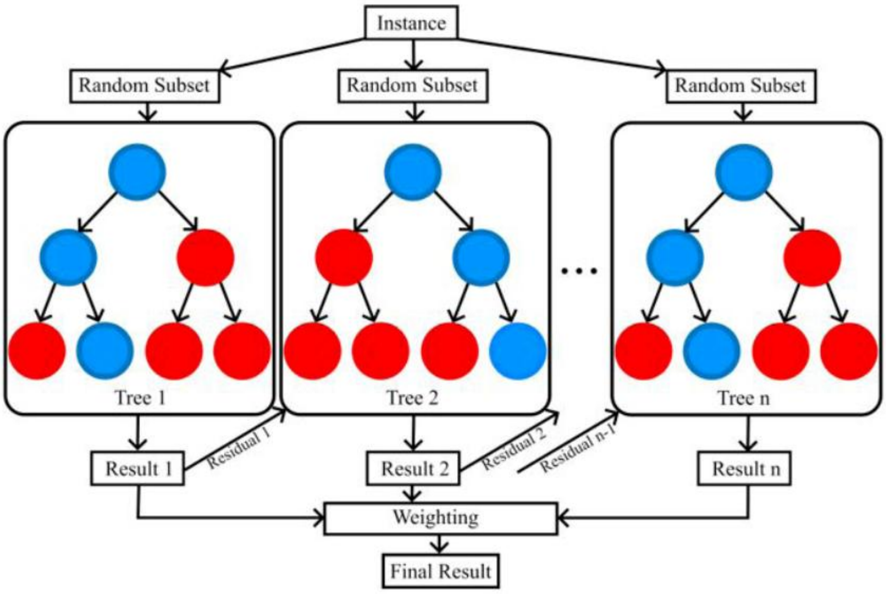


Figure II.6 *XGBoost*

Bagging

While decision trees are one of the most easily interpretable models, they exhibit highly variable behavior. Consider a single training dataset that we randomly split into two parts. Now, let's use each part to train a decision tree in order to obtain two models.

When we fit both these models, they would yield different results. Decision trees exhibit high variance due to this behavior. Bagging or boosting aggregation helps to reduce the variance in any learner. Several decision trees generated in parallel form the base learners of the bagging technique. Data sampled with replacement is fed to these learners for training. The final prediction is the averaged output from all the learners. [11]

Boosting

In boosting, the trees build sequentially so that each subsequent tree aims to reduce the errors of the previous tree. Each tree learns from its predecessors and updates the residual errors. Hence, the tree that grows next in the sequence will learn from an updated version of the residuals.

The base learners in boosting are weak learners in which the bias is high, and the predictive power is just a tad better than random guessing. Each of these weak learners contributes some vital information for prediction, enabling the boosting technique to produce a strong learner by effectively combining these weak learners. The final strong learner brings down both the bias and the variance.

In contrast to bagging techniques like Random Forest, boosting uses trees with fewer splits. Such small trees, which are not very deep, are highly interpretable. You can optimally select parameters like the number of trees or iterations, the learning rate of gradient boosting, and the depth of the tree through validation techniques like k-fold cross-validation. Having a large number of trees might lead to overfitting. So, it is necessary to carefully choose the stopping criteria for boosting. [11]

Gradient Boosting Ensemble Technique

The gradient boosting ensemble technique consists of three simple steps:

- An initial model F0 is defined to predict the target variable y. This model will be associated with a residual (y – F0)
- A new model h1 is fit to the residuals from the previous step
- Now, F0 and h1 are combined to give F1, the boosted version of F0. The mean squared error from F1 will be lower than that from F0:

$$F_1(x) < -F_0(x) + h_1(x)$$

To improve the performance of F1, we could model after the residuals of F1 and create a new model F2:

$$F_2(x) < -F_1(x) + h_2(x)$$

This can be done for ‘m’ iterations, until residuals have been minimized as much as possible:

$$F_m(x) < -F_{m-1}(x) + h_m(x)$$

Here, the additive learners do not disturb the functions created in the previous steps. Instead, they impart information of their own to bring down the errors. [11]

4. Artificial Intelligent approaches used for classification problem

4.1 Fuzzy Logic

4.1.1 Architecture

Its Architecture contains four parts:

- RULE BASE: It contains the set of rules and the IF-THEN conditions provided by the experts to govern the decision-making system, on the basis of linguistic information. Recent developments in fuzzy theory offer several effective methods for the design and tuning of fuzzy controllers. Most of these developments reduce the number of fuzzy rules.
- FUZZIFICATION: It is used to convert inputs i.e. crisp numbers into fuzzy sets. Crisp inputs are basically the exact inputs measured by sensors and passed into the control system for processing, such as temperature, pressure, rpm's, etc.
- INFERENCE ENGINE: It determines the matching degree of the current fuzzy input with respect to each rule and decides which rules are to be fired according to the input field. Next, the fired rules are combined to form the control actions.
- DEFUZZIFICATION: It is used to convert the fuzzy sets obtained by the inference engine into a crisp value. There are several defuzzification methods available and the best-suited one is used with a specific expert system to reduce the error.

4.1.2 Applications

- It is used in the aerospace field for altitude control of spacecraft and satellites.
- It has been used in the automotive system for speed control, traffic control.
- It is used for decision-making support systems and personal evaluation in the large company business.
- It has application in the chemical industry for controlling the pH, drying, chemical distillation process.
- Fuzzy logic is used in Natural language processing and various intensive applications in Artificial Intelligence.
- Fuzzy logic is extensively used in modern control systems such as expert systems.

- Fuzzy Logic is used with Neural Networks as it mimics how a person would make decisions, only much faster. It is done by Aggregation of data and changing it into more meaningful data by forming partial truths as Fuzzy sets.

4.2 Expert Systeme

4.2.1 The importance of Expert System

Expert systems are a game-changer in AI because they:

Preserving Expertise: They capture the knowledge of human experts and store it in a digital format. This ensures that valuable expertise isn't lost when an expert retires or leaves.

Improving Decision-Making: By relying on data and rules, expert systems provide consistent and unbiased recommendations.

Saving Time and Money: They automate tasks that would otherwise require human intervention, reducing costs and increasing efficiency.

Accessibility: Expert systems make expert-level knowledge available to non-experts, democratizing access to specialized information.

For instance, in the 1970s, the MYCIN system was developed to diagnose bacterial infections. While it was never used in real hospitals, it demonstrated how expert systems could assist doctors in making accurate diagnoses.

4.2.2 Reasoning Strategies

Forward Chaining and Backward Chaining, which are two fundamental methods for processing information and solving problems in an expert system:

1. Forward Chaining

This is a data-driven reasoning approach where the system starts with the available facts and applies rules to infer new facts or conclusions like shown on figure II.8. It's typically used to predict outcomes or determine what will happen next. An example given is predicting stock market movements.

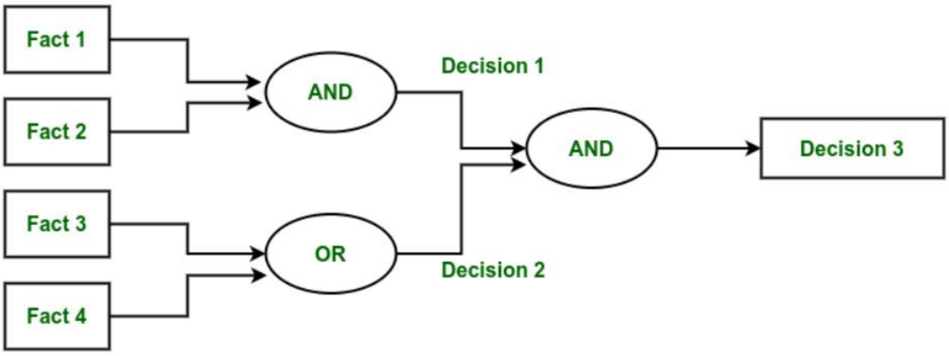


Figure II.7 *Forward Chaining*

2. Backward Chaining

This is a goal-driven reasoning approach where the system starts with a hypothesis or a goal (something to prove) and works backward to determine which facts or conditions would support that conclusion like shown on figure II.9. It's often used to diagnose issues by determining the cause of an observed effect. The examples provided include diagnosing medical conditions like stomach pain, blood cancer, or dengue.

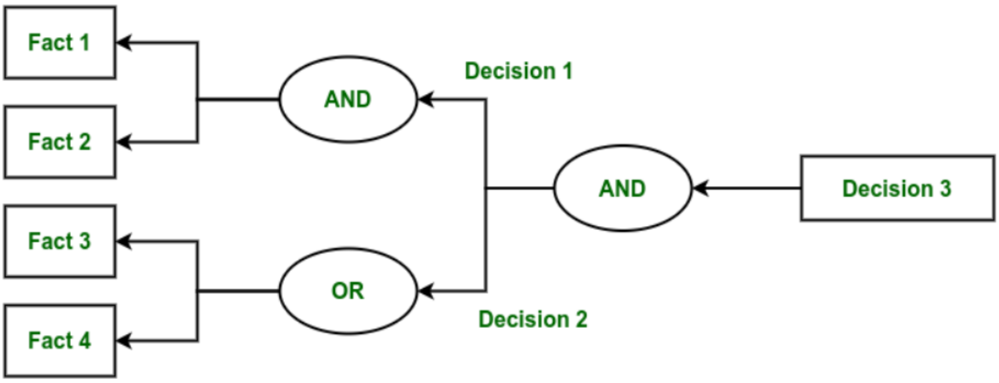


Figure II.8 *Backward Chaining*

4.2.3 How These Components Work Together

Imagine a medical expert system designed to diagnose diseases:

1. Input: A patient reports symptoms like fever, cough, and fatigue through the user interface.
2. Processing: The inference engine analyzes the symptoms using rules from the knowledge base.
3. Output: The system suggests a possible diagnosis, such as pneumonia.
4. Explanation: The explanation module provides a detailed explanation, such as "The diagnosis is based on the presence of fever, cough, and abnormal chest X-ray results."
5. Update: The knowledge acquisition module adds new data, such as recent research on pneumonia treatments, to keep the system up-to-date.

5. Conclusion

The chapter ends by emphasizing that different algorithms have distinct strengths. Machine learning methods offer high accuracy and automation, while fuzzy and expert systems provide more transparency and explainability. Understanding their structure, workflow, and assumptions is essential for choosing the right model in medical diagnostics.

Chapter III:

Diabetes prediction

1. Introduction

This chapter applies the previous knowledge to predict diabetes using three supervised ML algorithms (Logistic Regression, Random Forest, XGBoost) trained on the Pima Indians Diabetes Dataset. It discusses dataset features, preprocessing (e.g., handling missing data, scaling), libraries used, and interprets each model's output with graphs and metrics.

2. Dataset and Input Features

All three machine learning programs are trained on the Pima Indians Diabetes Dataset, a publicly available medical dataset originally published by the National Institute of Diabetes and Digestive and Kidney Diseases. It contains 768 patient records, each with 8 input features and 1 target output. The dataset is commonly used for binary classification tasks, where the goal is to predict whether a person has diabetes (Outcome = 1) or does not (Outcome = 0).

2.1 Pregnancies

Refers to the number of times the patient has been pregnant. Although not a direct indicator of diabetes, it serves as a proxy for reproductive and metabolic history.

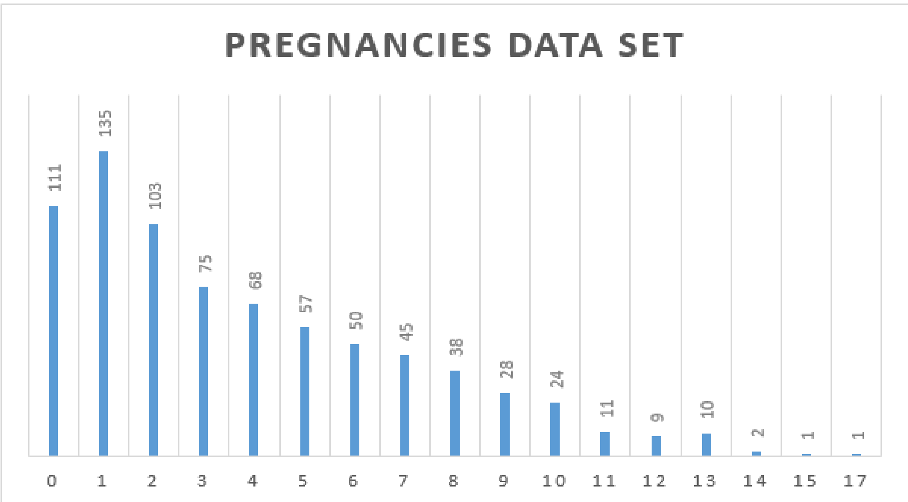


Figure III.1 *Pregnancies Data Set*

Pregnancy induces various hormonal and metabolic changes in the body, and women with a history of multiple pregnancies may be at greater risk of developing gestational diabetes a condition that can progress into type 2 diabetes later in life. Thus, this variable is indirectly connected to insulin resistance and overall metabolic health. Our data of pregnant woman can be shown in figure III.1

2.2 Glucose

Measures the plasma glucose concentration taken two hours after an oral glucose tolerance test. It is a key diagnostic indicator for diabetes. According to the World Health Organization (WHO), a 2-hour plasma glucose level below 140 mg/dL is considered normal, a level between 140 and 199 mg/dL indicates impaired glucose tolerance (pre-diabetes), and a level of 200 mg/dL or above confirms a diagnosis of diabetes. This feature is central to diabetes prediction because elevated blood glucose is one of the most defining characteristics of the disease.The glucose data can be shown in figure III.2 (Note that we don’t have cases over 199)

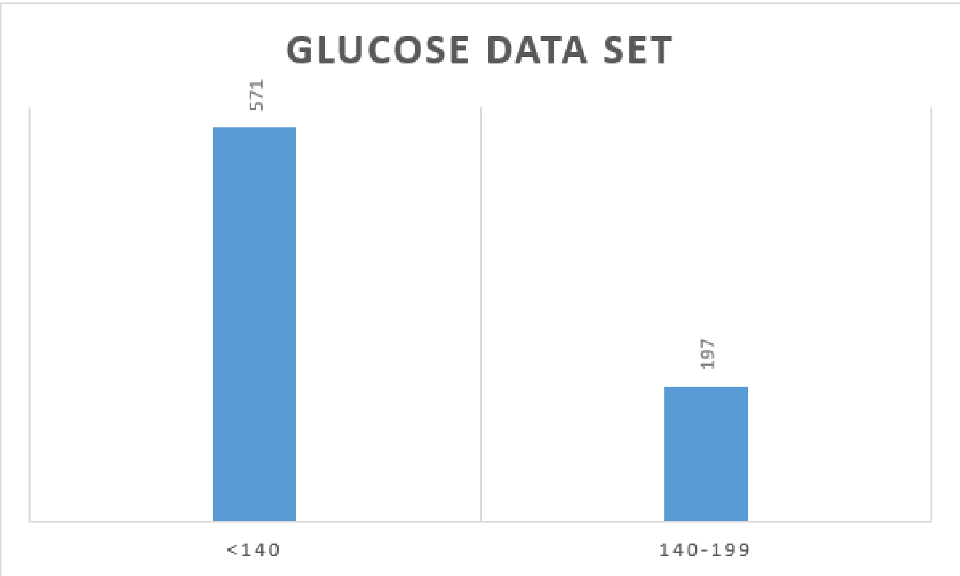


Figure III.2 *Glucose Data Set*

2.3 Blood Pressure

Specifically the diastolic value recorded in this dataset, is the third feature. Diastolic blood pressure measures the pressure in the arteries when the heart is at rest between beats. Persistent high blood pressure (hypertension) is known to be a comorbidity of type 2 diabetes and is often part of the broader metabolic syndrome. Normal diastolic pressure is considered to be below 80 mmHg. Readings between 80–89 mmHg are categorized as Stage 1 hypertension, and readings of 90 mmHg or higher indicate Stage 2 hypertension. The inclusion of this variable reflects the close clinical relationship between cardiovascular health and diabetes risk. The blood pressure data can be shoe in the next figure III.3 in three different intervals

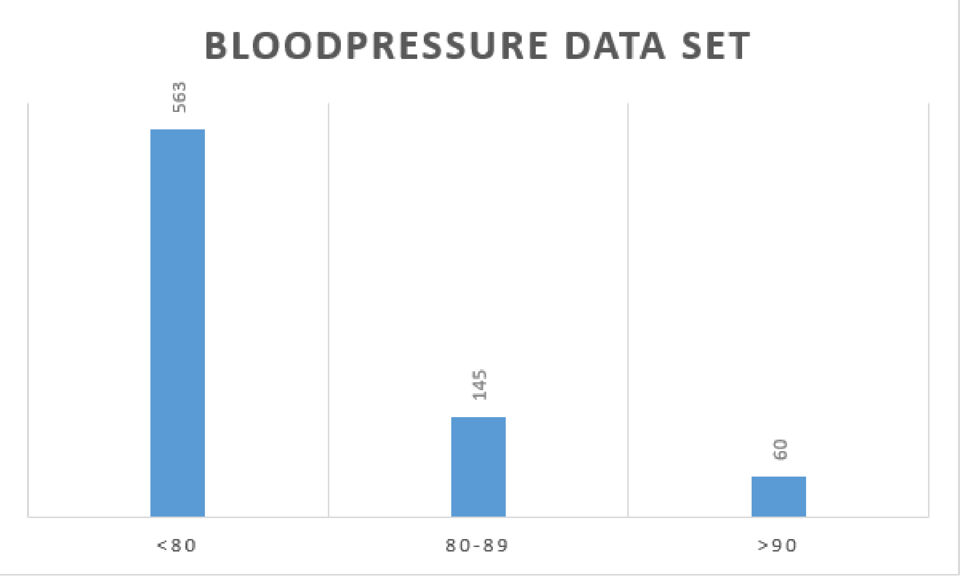


Figure III.3 *Blood Pressure Data Set*

2.4 Skin Thickness

Measures the triceps skinfold in millimeters. This anthropometric measurement is used as a non-invasive method to estimate subcutaneous fat. It provides insights into the patient's nutritional status and overall body fat composition.

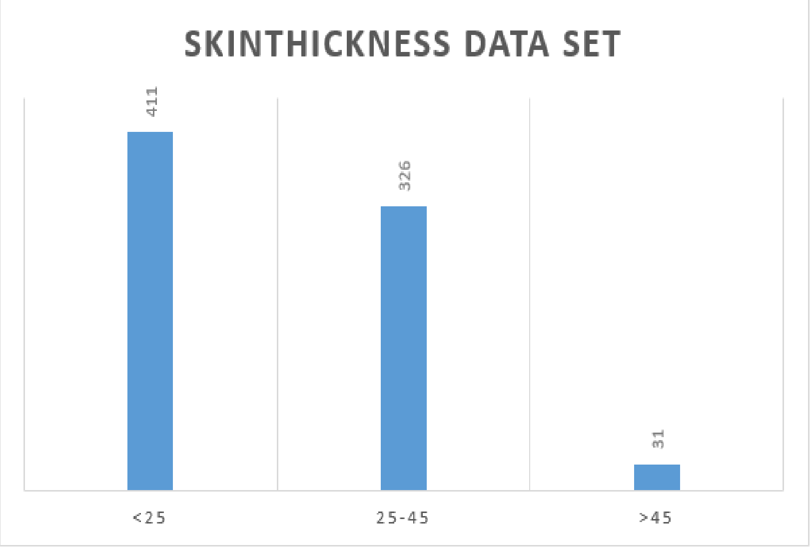


Figure III.4 *SkinThickness Data Set*

Studies have shown that individuals with higher triceps skinfold thickness are more likely to be overweight or obese, both of which are major risk factors for insulin resistance and diabetes. The average value in healthy individuals is around 18.7 mm, with notable differences between genders, typically higher in women., our data can be shown in figure III.4

2.5 Insulin

Represents the serum insulin level measured two hours after a glucose load. This value gives an estimate of how effectively the patient's pancreas is responding to glucose intake. Low insulin values may indicate impaired insulin production, while high values could suggest insulin resistance—a condition in which the body's cells become less responsive to insulin. Both scenarios are important in understanding the progression toward diabetes. Our data are more inconsistent for insulin making it less reliable as an input

2.6 BMI

Calculated as the individual's weight in kilograms divided by the square of their height in meters. BMI is a widely accepted indicator of body fat and a strong predictor of diabetes risk. According to WHO classifications, a BMI of 18.5–24.9 is considered healthy, 25–29.9 is overweight, and 30 or above is obese. Obesity, particularly central obesity, significantly increases the risk of developing type 2 diabetes due to its association with chronic inflammation and insulin resistance.

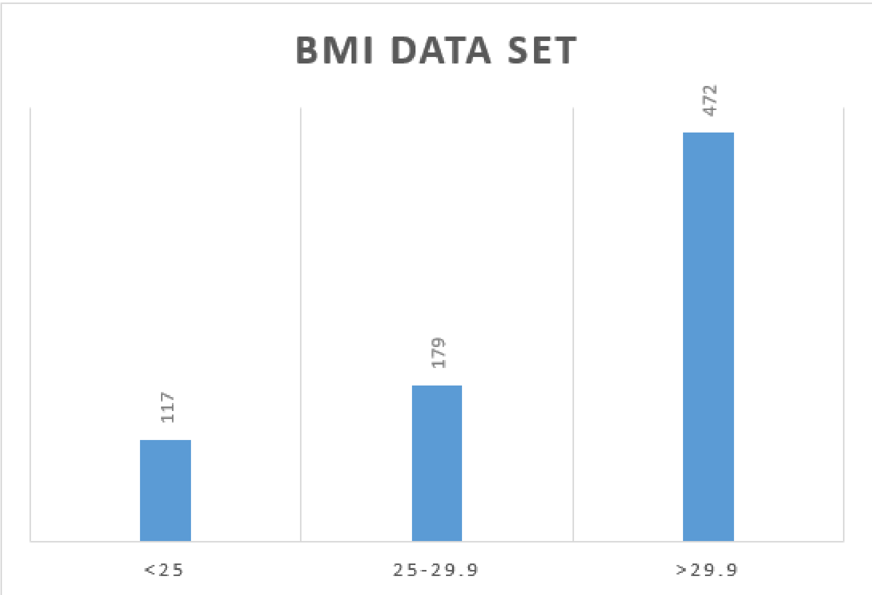


Figure III.5 *BMI Data Set*

2.7 Diabetes Pedigree Function

a score that quantifies the likelihood of diabetes based on the patient's family history. It aggregates information about the presence of diabetes in the patient's relatives and adjusts for factors such as age. Higher values in this function indicate a greater hereditary risk. In the dataset, values typically range from 0.08 to 2.42.

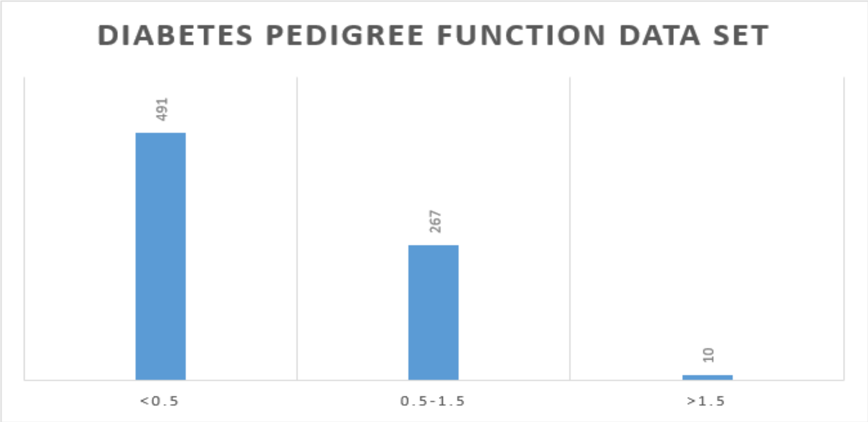


Figure III.6 *Diabetes Pedigree Function Data Set*

2.8 Age

Advancing age is a well-established risk factor for type 2 diabetes. The prevalence of insulin resistance and beta-cell dysfunction increases as individuals grow older. In this dataset, ages range from 21 to 81 years. The correlation between age and diabetes is not only biological but also behavioral, as older adults may be more likely to exhibit sedentary lifestyles and poor dietary habits, both of which contribute to diabetes risk.

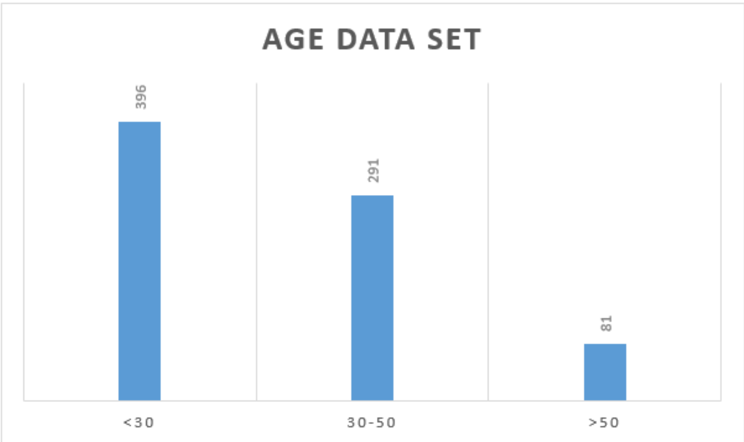


Figure III.7 *Age Data set*

3. Preprocessing

Before feeding data into any machine learning model, it is essential to prepare the dataset through a process known as preprocessing. This step ensures that the data is clean, consistent, and properly formatted so that the model can learn effectively. For the diabetes prediction programs discussed in this chapter, preprocessing involves several crucial tasks: identifying and handling missing values, scaling numerical features, and splitting the data for training and testing.

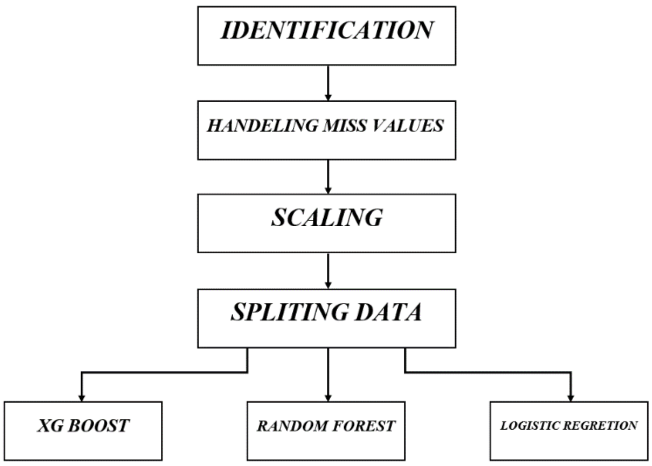


Figure III.8 *Preprocessing Diagram*

3.1 Identifying Missing or Invalid Values

The dataset appears complete at first glance, with no explicit NaN (Not a Number) entries. However, a closer look reveals that certain numeric fields contain zero values, which are medically implausible and therefore treated as implicit missing values.

3.1.1 Features Affected

- Glucose
- BloodPressure
- SkinThickness
- Insulin
- BMI

These features should never naturally have a value of zero for a living person:

- A Glucose level of 0 mg/dL implies immediate death.
- A BloodPressure of 0 mmHg indicates cardiac arrest.
- A BMI of 0 means no body mass, which is not physiologically possible.

3.1.2 Solutions Applied

Two approaches are used in the programs:

1. **Fixed Value Imputation** (used in earlier versions):
   - Missing values are replaced with medically average values, e.g., 140 for glucose, 24.9 for BMI.
   - Simple and fast, but may distort relationships in data.
2. **K-Nearest Neighbors (KNN) Imputation** (used in improved versions):
   - Treats zero values as missing (np.nan), then uses the KNNImputer from scikit-learn.
   - For each missing value, it finds the k=5 most similar rows (neighbors) and estimates the value based on their average.
   - This method considers feature correlations and patient similarity, making it more data-driven and robust.

3.2 Feature Scaling (Normalization)

Feature scaling ensures that all input features contribute equally to the learning process, especially for algorithms that rely on distance or linear weights (e.g., Logistic Regression and KNNImputer).

3.2.1 Method Used: StandardScaler

- Transforms each feature to have mean **= 0** and standard deviation = 1.
- Formula:

$$Z = \frac{x - \mu}{\sigma} \qquad (III.1)$$

Where:

- X : original feature value
- μ : mean of the feature
- σ : standard deviation of the feature

3.2.2 Why It Matters

- **Logistic Regression** is sensitive to feature scales. Without scaling, large values (like Glucose) dominate the model.
- **KNNImputer** uses Euclidean distance to find neighbors, which would be biased if features are on different scales.
- **Random Forest and XGBoost** are tree-based and not scale-sensitive, but we still apply scaling for consistency, comparison, and visualization.

3.3 Splitting the Dataset

To properly evaluate model performance, the dataset is split into training and testing sets.

3.3.1 Strategy

- train_test_split() from scikit-learn is used.
- 70% of data is used for training (X_train, y_train).
- 30% is reserved for testing (X_test, y_test).

3.3.2 Why This Matters

- The training set is used by the model to learn patterns.
- The testing set is **never seen** during training, which ensures **honest evaluation** and simulates how the model will perform on new, real-world data.

4. Interpreting Results

After training the machine learning models on the diabetes dataset, it is essential to assess their **effectiveness** and **reliability**. This is done by analyzing the models' outputs using various **evaluation metrics**, **visual tools**, and **statistical summaries**. These results not only help compare models but also explain how and why predictions were made.

4.1 Accuracy Score

The accuracy score is the simplest and most widely used metric. It measures the percentage of correct predictions out of all predictions made.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Each program prints an accuracy score after prediction:

Tableau III.1 *Accuracy*

Model	**Typical Accuracy (on test set)**
Logistic Regression	~76–78%
Random Forest	~80–84%
XGBoost	~83–86%

Note: A higher accuracy doesn't always mean a better model — especially when the dataset is imbalanced. That's why additional metrics are needed.

4.2 Confusion Matrix

A **confusion matrix** provides a deeper look into the model's prediction behavior by categorizing outcomes into four groups:

Tableau III.2 *Confusion matrix*

	Predicted: No Diabetes (0)	**Predicted: Diabetes (1)**
Actual: No (0)	True Negative (TN)	False Positive (FP)
Actual: Yes (1)	False Negative (FN)	True Positive (TP)

- **True Positives (TP):** Diabetics correctly identified.
- **True Negatives (TN):** Non-diabetics correctly identified.
- **False Positives (FP):** Healthy individuals falsely predicted diabetic.
- **False Negatives (FN):** Diabetics missed by the model.

These matrices are visualized using a **heatmap** (via seaborn), making them easy to interpret.

4.3 Classification Report

The classification_report from sklearn.metrics includes three key metrics per class:

1. **Precision**: How many predicted diabetics were actually diabetic?

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

2. **Recall**: How many actual diabetics did the model identify?

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

3. **F1-Score**: Harmonic mean of precision and recall. A balanced performance measure.

A good model should have high precision and high recall for both classes (0 and 1).

4.4 Feature Importance (Random Forest and XGBoost)

For tree-based models (Random Forest and XGBoost), we can visualize which features were most influential in making predictions.

- Each feature is assigned an importance score based on how much it contributes to reducing classification error across the trees.
- Typically, Glucose, BMI, and Age are the most influential in diabetes prediction.

These are visualized as horizontal bar plots:

- Taller bars = higher influence
- Helps clinicians and data scientists understand the most predictive medical factors

4.5 Model Coefficients (Logistic Regression)

Logistic Regression outputs weights (coefficients) assigned to each feature. These values indicate:

- The **direction** of influence (positive or negative),
- The **magnitude** of effect.

For example:

- A positive coefficient for Glucose means higher glucose increases the probability of diabetes.
- A negative coefficient for BloodPressure might indicate a weak inverse relationship.

These coefficients provide interpretability, making Logistic Regression useful in clinical settings where explainability is critical.

4.6 SHAP Plots (XGBoost and Random Forest)

SHAP (SHapley Additive exPlanations) plots show how each feature contributed to the model's prediction for each individual.

- **Red values** increase the predicted risk of diabetes.
- **Blue values** decrease the predicted risk.
- Helps in **explaining individual predictions** — useful in medical decision-making.

SHAP plots summarize:

- Which features consistently push predictions higher or lower.
- How feature values (low/high) affect predictions.

Example insight: High glucose levels consistently drive predictions toward class 1 (diabetic).

5. Comparative Summary

After implementing and evaluating the three machine learning models, we can compare them across multiple dimensions: accuracy, interpretability, training complexity, handling of data, and practical usability. Each model brings unique strengths, making it more or less suitable depending on the context.

5.1 Summary Comparison Table

Table III.3 *Summary Comparison Table*

Criterion	Logistic Regression	Random Forest	XGBoost
Type	Linear Classifier	Ensemble of Decision Trees (Bagging)	Boosted Decision Trees (Gradient Boosting)
Accuracy	Medium (~76–78%)	High (~80–84%)	Very High (~83–86%)
Speed	Very Fast	Medium	Slower (but optimized)
Interpretability	Very High	Medium	Low (but explainable with SHAP)
Overfitting Control	Regularization (L1/L2)	Bagging & randomness	Regularization, shrinkage, early stopping
Handles Non-Linearity	Poor	Good	Excellent
Feature Importance	Based on coefficients	Gini-based importance scores	Gain-based feature scores, SHAP-friendly
Missing Values	Must be preprocessed	Must be preprocessed	Handled natively, or via imputation
Scaling Required	Yes	No	No
Use Case Suitability	Quick clinical insights, reports	General-purpose accuracy	High-performance systems, competitions

5.2 Interpretation Highlights

- **Logistic Regression** provides transparent results. Each coefficient shows how much a feature contributes to the prediction. It's ideal when explainability and regulatory transparency are required (e.g., clinical guidelines).
- **Random Forest** offers strong performance and is less sensitive to outliers and noise. It handles nonlinear interactions and feature importance well, making it a robust default choice.
- **XGBoost** consistently outperforms other models in accuracy. It's optimized for performance and is widely used in Kaggle competitions and real-world systems. However, it requires more tuning and is less interpretable unless SHAP values are used.

5.3 When to Use Which Model?

Table III.4 *The Models Uses*

Situation	Recommended Model	Reason
You need a simple and interpretable model	Logistic Regression	Easy to understand and explain
You want a strong baseline model with low overfitting	Random Forest	Robust and versatile
You're aiming for maximum performance and can afford complexity	XGBoost	State-of-the-art accuracy
Your dataset has nonlinear patterns or interactions	Random Forest / XGBoost	Tree-based models handle nonlinearities well
You need to understand individual predictions	XGBoost + SHAP	SHAP explains predictions clearly

6. The results of test program

We've made a diabete prediction machine learning program using logistic regression method next we''ll show the results of a test subject:

6.1 Inputs

Next is the test subject inputs that we use in the program

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age	Outcome
1	90	61	20	102	28.1	0.167	25	0

6.2 results

The following is the resulted graphs that helps understand the outcome and the results them self:

Figure III.9 *Glucose Vs Age*

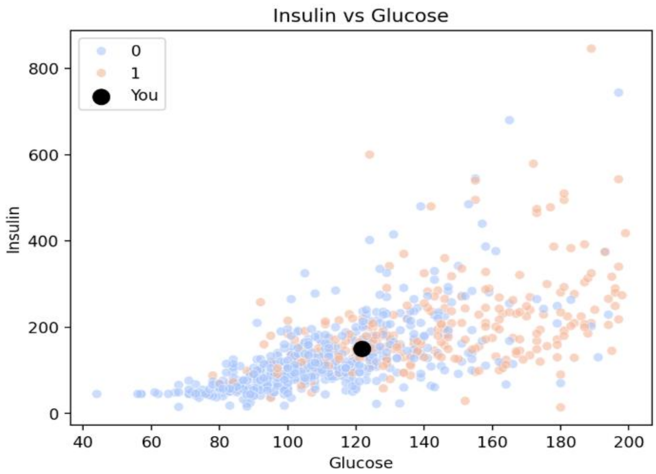


Figure III.11 *Insulin Vs Glucose*

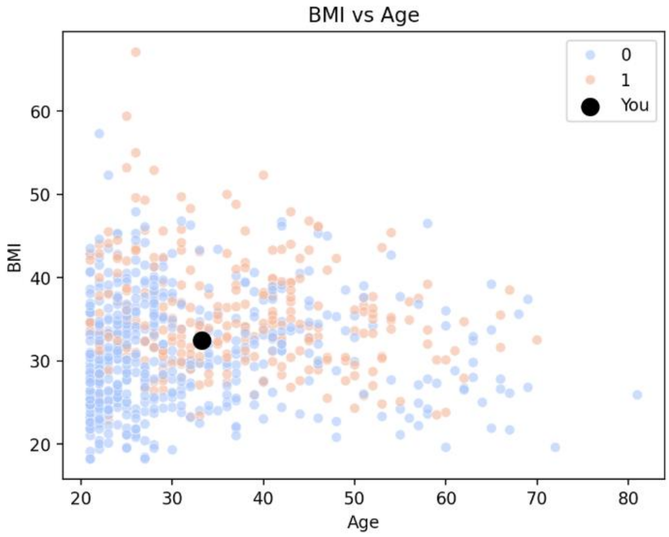
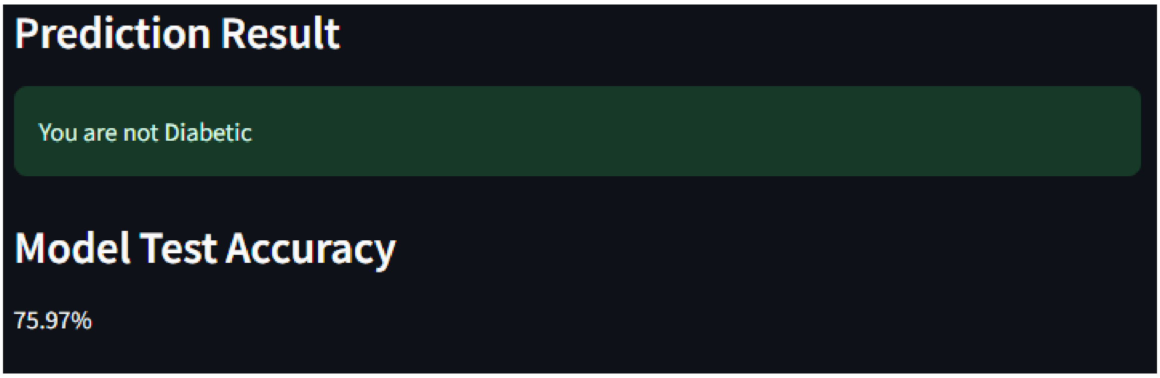


Figure III.10 *BMI Vs Age*

and these are the results



7. Fuzzy Logic and Expert System Approaches

In addition to traditional machine learning models like Logistic Regression, Random Forest, and XGBoost, Fuzzy Logic and Expert Systems offer a rule-based, interpretable approach for reasoning under uncertainty. These methods do not rely on large amounts of training data but instead use human-defined logic and approximate reasoning to make decisions — closely resembling clinical judgment.

7.1 Input Features

To simplify the fuzzy system and keep it interpretable, we typically select 3 to 5 core features based on clinical relevance and importance scores from previous models. In this case, we use:

- **Glucose** — key diagnostic marker for diabetes.

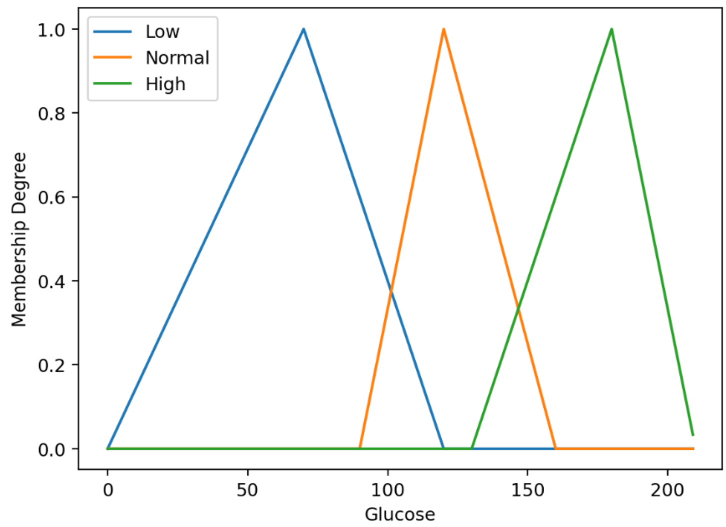


Figure III.12 *Glucose Prediction*

- **BMI** — strong indicator of obesity-related diabetes risk.

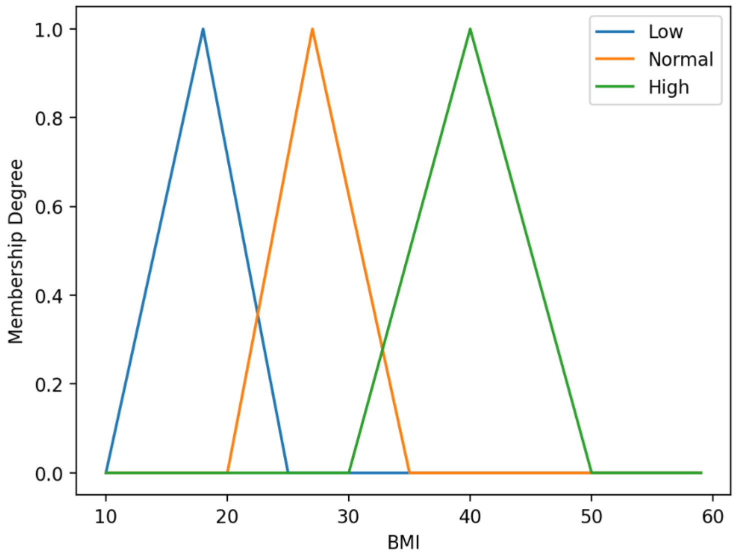


Figure III.13 *BMI Prediction*

- **Age** — risk increases with age.

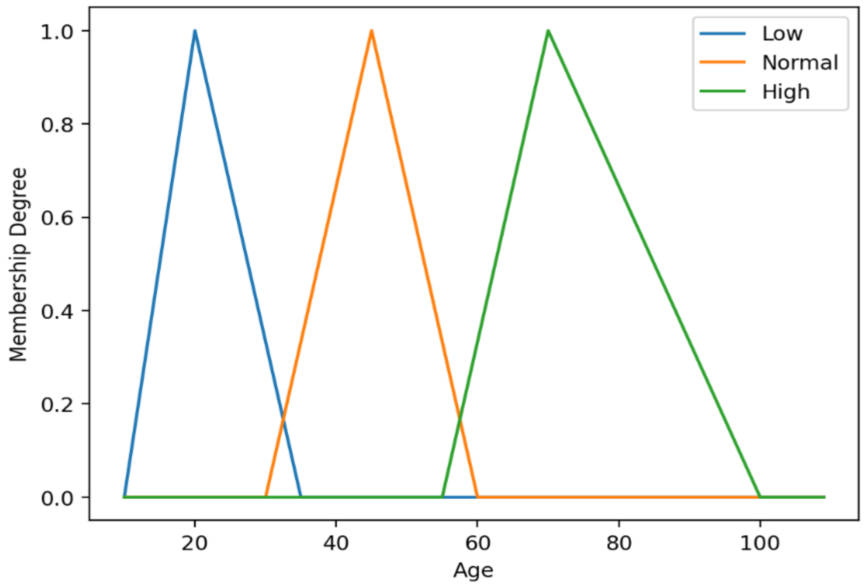


Figure III.14 *Age Prediction*

- **Insulin** — informs about insulin production or resistance.

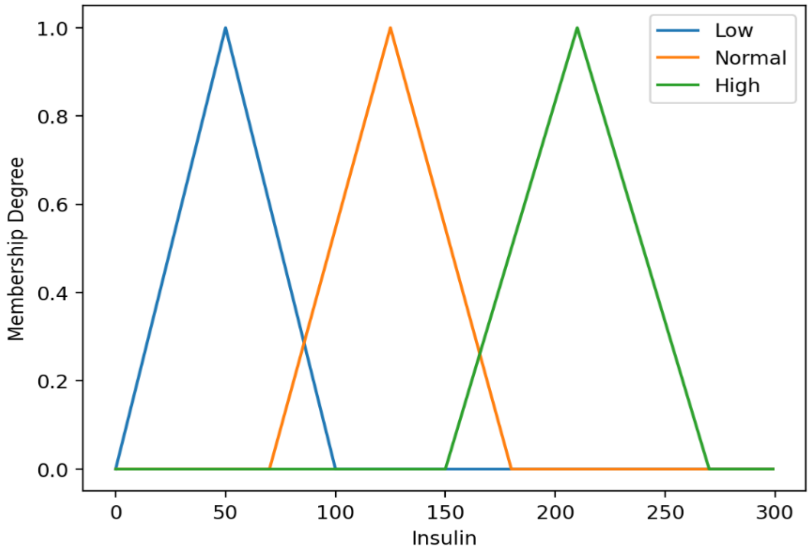


Figure III.15 *Insulin Prediction*

- **Diabetes Pedigree Function** — estimates genetic predisposition.

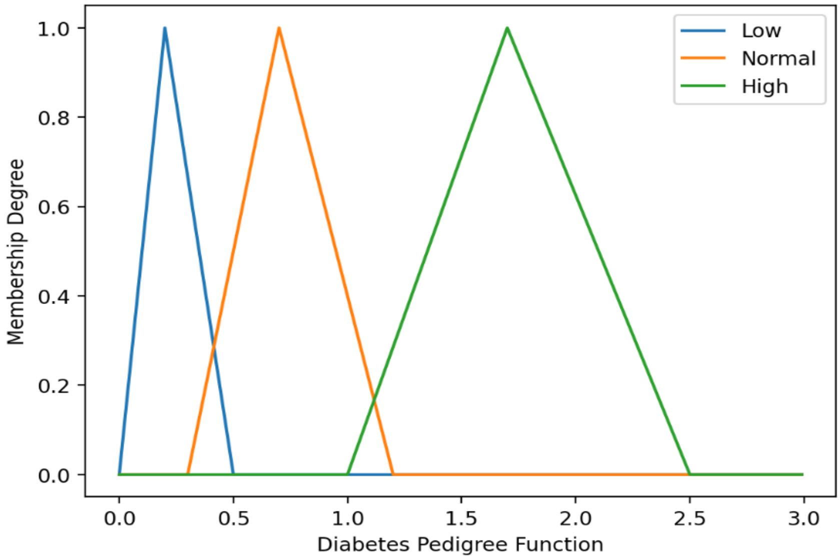


Figure III.16 *Diabetes Pedigree Function Prediction*

Each feature is mapped to fuzzy **sets** (e.g., “Low,” “Normal,” “High”) using membership functions.

7.2 Fuzzy Logic Algorithm

Fuzzy logic mimics human reasoning using the following steps:

1. Fuzzification:
    - Converts crisp numerical inputs into degrees of membership in fuzzy sets.
    - For example: Glucose = 160 may be 0.2 “Normal” and 0.8 “High”.
2. Rule Evaluation (Inference):
    - Rules are written as expert IF-THEN logic, e.g.:

- If Glucose is High AND BMI is Obese, THEN Diabetes Risk is High

3. Aggregation:
    - Combines all rules to form a fuzzy output set.
4. Defuzzification:
    - Converts fuzzy output into a crisp decision, such as a probability or class.

7.3 Expert System Logic

An expert system follows hard-coded rules from clinical guidelines rather than fuzzy sets.

Example rules:

- If Glucose > 200 and BMI > 30 → Predict diabetic.
- If Glucose < 140 and Age < 40 → Predict non-diabetic.

This system doesn’t use degrees of truth but still allows encoding human expertise.

7.4 Output and Interpretation

- The fuzzy system outputs a risk score between 0 and 1.
- Thresholds can be set:
    - 0.0–0.3 → Non-diabetic
    - 0.3–0.7 → At risk

- 0.7–1.0 → Diabetic

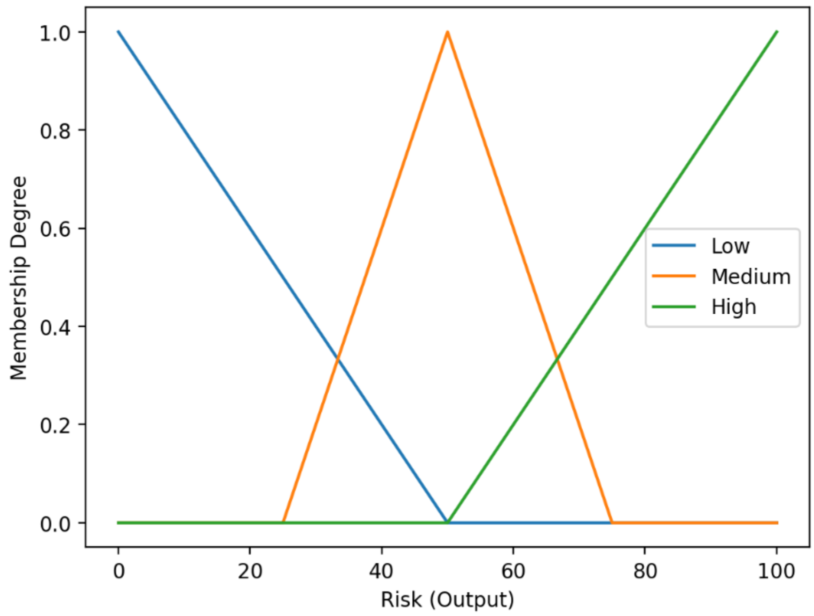


Figure III.17 *Output Prediction*

- Unlike ML models, these decisions are fully explainable using rule contributions.

7.5 Comparison with Machine Learning Models

Table III.5 *Comparaison between Fuzzy logic/Expert System with machine learning*

Feature	Fuzzy Logic / Expert System	Machine Learning Models
Accuracy	Medium (rule-based)	High (data-driven)
Training Data Needed	None	Required
Interpretability	Very High	Varies (high in Logistic Regression)
Adaptability	Manual rule tuning	Auto-optimized
Sensitivity to noise	Low (rules are fixed)	Moderate
Best Use Case	Low-resource, transparent systems	High-accuracy automated pipelines

8. Conlusion

In this chapter, we explored and compared several approaches for predicting diabetes using structured health data. Through the implementation of Logistic Regression, Random Forest, and XGBoost, we demonstrated the strength of modern machine learning techniques in extracting patterns from data and achieving high prediction accuracy. These models are data-driven, scalable, and capable of adapting to complex, nonlinear relationships. However, they often require substantial training data and computational resources, and in some cases, lack transparency.

To complement these models, we introduced Fuzzy Logic and Expert Systems as rule-based alternatives that mimic clinical reasoning. Unlike machine learning, these systems operate on predefined knowledge through human-crafted rules and linguistic variables, making them highly interpretable and suitable for low-resource or decision-support scenarios. While they may not reach the predictive performance of XGBoost or Random Forest in large datasets, they offer valuable clarity and explainability in high-stakes environments like healthcare.

Ultimately, the choice between machine learning and rule-based systems depends on the context. When accuracy and automation are prioritized, data-driven models excel. When transparency and human-aligned logic are critical, expert and fuzzy systems offer strong value. Together, these paradigms form a complementary toolkit for developing intelligent, ethical, and effective diagnostic tools for diabetes and beyond.

General Conclusion

Through this work, we explored and compared two distinct paradigms for diabetes prediction: data-driven machine learning algorithms and rule-based reasoning systems**.**

The machine learning models — Logistic Regression, Random Forest, and XGBoost — demonstrated strong performance in identifying diabetic cases based on numerical input features such as glucose levels, BMI, age, and insulin. Among them, XGBoost provided the highest accuracy, while Logistic Regression offered the best model transparency. Preprocessing techniques like KNN imputation and feature scaling significantly contributed to model robustness and performance.

In contrast, the fuzzy logic and expert system approaches emphasized human-like reasoning, using IF-THEN rules and linguistic variables. These models required no training data but relied on carefully designed knowledge bases and membership functions. Though slightly less accurate than machine learning models, they excelled in interpretability, flexibility, and suitability for low-resource or explainable AI contexts.

The comparison highlights a fundamental trade-off: machine learning provides high accuracy through data, while fuzzy and expert systems deliver interpretability and clinical alignment through rules. Rather than viewing these methods as mutually exclusive, they can be combined into hybrid decision systems, offering both predictive power and human-understandable logic — an ideal approach for responsible AI in healthcare.

References:

[1] Jyoti Prakash Patra & Yogesh Kumar Rathore, "*Applications of Artificial Intelligence in the Healthcare Sector*", Nova, 2023

[2] Bernhard G. Humm, "*Applied Artificial Intelligence*", Leanpub, 2015 – 2022.

[3] Lalit Mohan Goyal, Tanzila Saba, Amjad Rehman, Souad Larabi-Marie-Sainte, "*Artificial Intelligence and Internet of Things Applications in Smart Healthcare*", CRC Press, 2021

[4] Munish Sabharwal, B. Balamurugan Balusamy, S. Rakesh Kumar, N. Gayathri, Shakhzod Suvanov "*Applications of Artificial Intelligence in E-Healthcare Systems*", The Institution of Engineering and Technology, 2022

[5] Hemachandran K, Raul V. Rodriguez, Umashankar Subramaniam and Valentina Emilia Balas "*Artificial Intelligence and Knowledge Processing*", CRC Press, 2024

[6] Mitul Kumar Ahirwal, Narendra D. Londhe, and Anil Kumar "*ARTIFICIAL INTELLIGENCE APPLICATIONS FOR HEALTH CARE*", CRC Press, 2022

[7] Ankur Saxena, Shivani Chandra "*Artificial Intelligence and Machine Learning in Healthcare*", Springer, 2021

[8] Devasis Pradhan, Prasanna Kumar Sahu, Hla Myo Tun, Prasenjit Chatterjee, "Artificial and Cognitive Computing for Sustainable Healthcare Systems in Smart Cities", ISTE, 2024

[9] "*Logistic Regression in Machine Learning*", geeks for geeks, 2025 https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/

[10] "*Random Forest Algorithm in Machine Learning*", geeks for geeks, 2025 https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

[11] "*Implementation of XGBoost (eXtreme Gradient Boosting)*" geeks for geeks, 2025 https://www.geeksforgeeks.org/machine-learning/implementation-of-xgboost-extreme-gradient-boosting/