A red and white logo

Description automatically generated

**MS931: Dissertation**

**"Data Science with R: Development and Implementation of Machine Learning Algorithms for Automatic Classification of Liver Disease Patients"  
  
PRINCE YADAV**

**202394959**

**MSc. Business Analysis and Consulting**

**(2022/2023)**

A picture containing text, screenshot, font

Description automatically generated

**The place of useful learning**

The University of Strathclyde is a charitable body, registered in Scotland, number SC015263

**Signed Statement**

Except where explicitly stated, all the work in this dissertation – including any appendices – is my own and was carried out by me during my MSc course. It has not been submitted for assessment in any other context.

Signed \_\_\_\_A close up of a signature

Description automatically generated\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name: Prince Yadav

Date: 25/09/2023

**Executive Summary**

The dissertation titled "Data Science with R: Development and Implementation of Machine Learning Algorithms for Automatic Classification of Liver Disease Patients" explores the intricate relationship between data science, specifically using the R programming language, and healthcare diagnostics with a focus on liver disease. Prompted by the urgent demand for improved and streamlined diagnostic techniques, this research aims to revolutionize how healthcare providers approach liver disease classification.

At the heart of this dissertation lies the development of a machine learning model, implemented in R, that is geared towards automatically classifying patients with liver disease based on an array of biochemical parameters. Utilizing machine learning algorithms such as Logistic Regression, CART, and Random Forest, the model promises to offer a significant leap in diagnostic precision, timely medical intervention, and the overall quality of patient care. The flexibility and adaptability provided by R's data science libraries enable the model to maintain high levels of reliability and evolve over time.

The potential applications of this machine learning model are extensive, ranging from general healthcare settings to more specialized liver disease research institutions. While the advantages are plentiful, the research is forthright in discussing the hurdles and constraints involved in the model's development and implementation. These issues span from the necessity for a solid data collection structure to data privacy concerns and the ambition of reaching a minimum model accuracy of 70%.

The research methodology is meticulously designed to include every aspect from data set analysis for predictor selection to the training, evaluation, and optimization of machine learning algorithms. By adopting a deductive research approach enriched by rigorous statistical analysis, the dissertation ensures a well-rounded and substantiated conclusion.

In summary, this dissertation accentuates the ground-breaking potential that machine learning algorithms, implemented in R, hold for enhancing diagnostics in healthcare, specifically targeting liver diseases. As healthcare systems integrate this automatic learning model, they stand to gain a competitive edge in diagnostic accuracy, ultimately uplifting patient care and the overall efficiency of healthcare delivery.

**Acknowledgments**

I would like to express my heartfelt gratitude to my supervisor, Farzad Shams for their unwavering support, invaluable guidance, and persistent encouragement throughout the course of this dissertation. Your insightful feedback and profound expertise have been instrumental in shaping this research.

I am deeply thankful to all the academic staff and faculty members of the Department of Management Science and the University of Strathclyde for imparting the knowledge and skills that have been pivotal in the formulation and execution of this study.

I would also like to acknowledge the authors and scholars whose works provided the foundation and direction for this research. Their pioneering studies and comprehensive literature greatly enriched my understanding of the subject.

Special thanks go to my peers and colleagues for their constructive critiques, enlightening discussions, and for fostering an environment of collaboration and mutual growth.

Lastly, but most importantly, I owe a debt of gratitude to my family for their unwavering faith in my abilities and their constant encouragement. Their love and support served as a beacon of hope, especially during challenging times.

This dissertation is a testament to the collective efforts of all those mentioned above, and while I take responsibility for any errors or oversights, the successes are shared.

Table of Contents

[Chapter 1 Scene Setting Report 7](#_Toc146232709)

[1.Introduction 7](#_Toc146232710)

[2.Project Background 7](#_Toc146232711)

[3. Project Plan 8](#_Toc146232712)

[3.1 Problem Statement 8](#_Toc146232713)

[3.2 Motivational Statement 8](#_Toc146232714)

[3.3 Research Objectives 8](#_Toc146232715)

[3.4 Research Questions 9](#_Toc146232716)

[3.5 Project Methodology 9](#_Toc146232717)

[3.6. Data Collection and Preprocessing 10](#_Toc146232718)

[3.7. Model Building 10](#_Toc146232719)

[3.8. Project Deliverables 12](#_Toc146232720)

[4. Project Timeline 12](#_Toc146232721)

[5. Why is this Research Important? 13](#_Toc146232722)

[5.1. Healthcare Improvement 13](#_Toc146232723)

[5.2. Cost-Effectiveness 13](#_Toc146232724)

[5.3. Resource Allocation 14](#_Toc146232725)

[5.4. Global Health 14](#_Toc146232726)

[6. Who will benefit from this Research? 14](#_Toc146232727)

[6.2. Patients: 14](#_Toc146232728)

[6.3. Healthcare Providers: 15](#_Toc146232729)

[6.4. Academic Researchers: 15](#_Toc146232730)

[6.5. Policy Makers: 15](#_Toc146232731)

[6.6. Educators and Students: 15](#_Toc146232732)

[6.7. Technology Companies and Developers: 15](#_Toc146232733)

[Chapter 2 Client Report 16](#_Toc146232734)

[1. Introduction 16](#_Toc146232735)

[1.1. Background and context 16](#_Toc146232736)

[1.2. Problem Formulation 19](#_Toc146232737)

[1.3. Work Objectives 19](#_Toc146232738)

[2. Literature Review 19](#_Toc146232739)

[2.1 MACHINE LEARNING 19](#_Toc146232740)

[2.2 VARIOUS ML MODELS 22](#_Toc146232741)

[3. Proposed Model 31](#_Toc146232742)

[3.1. PROPOSED MODEL 31](#_Toc146232743)

[3.2 LOGISTIC REGRESSION ALGORITHM 32](#_Toc146232744)

[3.3 KNN ALGORITHM 33](#_Toc146232745)

[33](#_Toc146232746)

[34](#_Toc146232747)

[4.TOOLS AND TECHNOLOGIES 35](#_Toc146232748)

[5.RESULT AND SIMULATION 39](#_Toc146232749)

[5.1 DATA IMPLEMENTATION 39](#_Toc146232750)

[5.2 EVALUATION USING RANDOM FOREST 42](#_Toc146232751)

[5.3 EVALUATION USING LOGISTIC REGRESSION 44](#_Toc146232752)

[5.4 EVALUATION USING KNN 46](#_Toc146232753)

[5.5 EVALUATION USING CART 47](#_Toc146232754)

[5.6. DISCUSSION ANALYSIS 50](#_Toc146232755)

[6.CONCLUSION 52](#_Toc146232756)

[6.1. CONCLUSION 52](#_Toc146232757)

[6.2. RECOMMENDATIONS 53](#_Toc146232758)

[6.3.FUTURE SCOPE 53](#_Toc146232759)

[Chapter 3: Reflections 54](#_Toc146232760)

[1.Introduction 54](#_Toc146232761)

[1.1.What is Reflection? 54](#_Toc146232762)

[1.2.The Importance of Reflection 55](#_Toc146232763)

[2. Reflection 57](#_Toc146232764)

[2.1. Description Stage 57](#_Toc146232765)

[2.2. Feelings Stage 58](#_Toc146232766)

[2.4. Analysis Stage 59](#_Toc146232767)

[2.5. Conclusion Stage 60](#_Toc146232768)

[2.6. Action plan stage 61](#_Toc146232769)

[REFERENCES 65](#_Toc146232770)

# Scene Setting Report

## 1.Introduction

The advancement of technology has been a double-edged sword for the medical sector. While new techniques and tools have drastically improved healthcare delivery, the challenges tied to diseases—both in terms of diagnostics and treatment—continue to escalate in complexity. One such area is liver diseases, a group of conditions that cause substantial morbidity and mortality worldwide. Liver diseases are especially problematic because their symptoms often manifest at advanced stages, complicating timely intervention and effective treatment. The liver, as a vital organ, performs various functions ranging from detoxification to metabolism regulation. Hence, any dysfunction can have a cascading effect on overall health. However, traditional diagnostic methods like liver biopsies and imaging tests can be invasive, expensive, and sometimes inconclusive. These methods often pose emotional and financial burdens on patients while taking up valuable healthcare resources. Given these challenges, there is a pressing need for an alternative approach that can offer faster, less invasive, and more accurate diagnostics. This dissertation explores the potential of machine learning algorithms programmed in R to meet this pressing need. By leveraging computational models, this research aims to develop an automatic learning model that can classify patients with liver diseases based on biochemical parameters. Furthermore, the study will result in a user-friendly web application that healthcare providers can use to make quick and accurate predictions. This dual focus on computational efficiency and user accessibility aims to bridge the gap between complex algorithms and practical healthcare solutions.

## 2.Project Background

Liver diseases have been a significant cause of mortality, accounting for nearly 2 million deaths globally per year. In low-resource settings, the burden is especially high due to limited access to advanced diagnostic facilities. Traditional diagnostic methods for liver diseases include blood tests, liver function tests, imaging studies like ultrasound and MRI, and liver biopsies. However, these tests often provide inconclusive results and sometimes require a secondary confirmatory test, escalating the overall cost and time involved. Given these limitations, machine learning holds great promise as an alternative approach for early diagnosis. Machine learning techniques can automatically learn patterns from large volumes of data, thus providing valuable predictive insights that are not apparent through conventional statistical methods. Specifically, R programming, known for its statistical computing power and data manipulation capabilities, can serve as a potent tool for implementing these machine learning techniques. This project aims to leverage the capabilities of machine learning and the robustness of R programming to develop an efficient and accurate diagnostic tool for liver diseases, setting the stage for a revolution in liver disease diagnosis and management.

## 3. Project Plan

### 3.1 Problem Statement

The early diagnosis of liver diseases remains an elusive goal in modern healthcare. Current diagnostic approaches often require invasive procedures like biopsies, which not only present risks but are also emotionally and financially burdensome for patients. Furthermore, tests like MRIs and ultrasounds are not only expensive but can be inconclusive, requiring further tests and delaying definitive diagnosis and treatment. This situation presents a problematic catch, early diagnosis is crucial for effective treatment, yet the tools for early diagnosis are fraught with challenges. The inadequacies of the current diagnostic approaches necessitate the development of alternative methods that are quicker, non-invasive, and more accurate. The urgency of this problem is intensified by the increasing prevalence of liver diseases and the high mortality rates associated with late-stage liver complications.

### 3.2 Motivational Statement

The imperative for this research is not just scientific curiosity but an urgent public health need. Each year, nearly 2 million people die globally from liver diseases, a statistic that could be significantly reduced with earlier diagnosis and intervention. Furthermore, early diagnosis can considerably decrease healthcare costs by preventing the need for expensive, late-stage treatments and hospitalizations. Thus, the development of an efficient, machine learning-based diagnostic tool holds the potential to save both lives and resources. This tool could serve as a cornerstone in redefining how liver diseases are diagnosed, treated, and managed, potentially sparking a paradigm shift in healthcare delivery systems globally.

### 3.3 Research Objectives

General Objectives:

Create an automatic learning model to classify patients with liver diseases based on biochemical parameters: The central aim here is to develop a machine learning model that can predict the likelihood of liver diseases by analysing biochemical markers in blood tests. The model will be trained and tested using existing datasets and evaluated for accuracy and reliability.

Develop a web application implementing the created model: After creating a robust predictive model, the next step is to make this model accessible to healthcare providers by integrating it into a user-friendly web application.

Specific Objectives:

Analyse the data set to select the best predictors: The aim is to identify which biochemical markers or combinations thereof offer the most accurate predictions, providing more focused and reliable results.

Train, evaluate, and optimize machine learning algorithms: Multiple machine learning algorithms will be used and compared to find the one that offers the best predictive power, aiming for a minimum accuracy rate of 70%. [52]

Create a web application for user-friendly predictions and result visualization: The final product will be a web application that healthcare providers can use to input biochemical markers and receive immediate, accurate predictions, complete with visualizations of the results.

### 3.4 Research Questions

Q1) What are the most effective machine learning algorithms for early detection of liver diseases, and how do they compare to traditional diagnostic methods?

The argument here is that a thorough understanding of how machine learning algorithms perform in comparison to traditional diagnostic methods is crucial. By conducting a systematic evaluation, insights gained will offer a clear benchmark for future innovations in liver disease diagnostics.

Q2) What are the limitations and challenges of applying machine learning techniques for liver disease detection?

The premise for this question is that while machine learning holds promise, it is not without its challenges. By combining literature review and primary research, we can provide insights into what these limitations might be, thereby offering directions for future improvements.

### 3.5 Project Methodology

The core aim of this research project is to identify influential variables that play a critical role in the detection and classification of liver diseases. This will be achieved through a comprehensive methodology that integrates statistical analyses, data pre-processing, and the application of various machine learning algorithms. The overarching goal is to contribute to healthcare analytics by building highly effective predictive models.

### 3.6. Data Collection and Pre-processing

#### 3.6.1. Data Source

The Indian Liver Patient Dataset from the UCI Machine Learning Repository serves as the foundation of this research. This dataset is publicly available and has been utilized in multiple scientific studies, enhancing its credibility. It comprises several variables, including bilirubin levels, enzyme levels, age, and gender, which are crucial in liver disease research.

#### 3.6.2. Data Cleaning

Removal of Outliers: Any data points that deviate significantly from the rest of the dataset will be identified and removed to prevent them from skewing the results.

Missing Value Imputation: Missing values, if any, will be replaced with statistical measures like mean or median, depending on the nature of the variable.

Data Transformations: Specific variables will undergo transformations such as calculating AST/ALT ratios and logarithmic conversions to make them more suitable for analysis.

#### 3.6.3. Exploratory Data Analysis (EDA)

Univariate Analysis: Individual variables will be examined through histograms and box plots to understand their distribution and potential impact on liver disease classification.

Bivariate Analysis: Scatter plots and correlation matrices will be used to assess the relationships between pairs of variables. This helps in understanding how two variables interact with each other and their combined influence on the outcome.

Statistical Testing: Advanced statistical tests like the Wilcoxon test will be performed to validate or refute hypotheses related to feature importance and selection.

### 3.7. Model Building

Algorithms

The machine learning algorithms slated for use are:

Logistic Regression: A straightforward yet powerful algorithm suitable for binary classification problems.

Classification and Regression Trees (CART): Offers a visual and intuitive model structure but can easily overfit the training data.

Random Forest: An ensemble method that improves the CART algorithm's performance by constructing multiple decision trees.

K-Nearest Neighbours (KNN): A non-parametric method that classifies a data point based on how its neighbours are classified.

Feature Selection

Recursive Feature Elimination: This method recursively removes the least important features while building a model, thereby identifying the most critical features.

Correlation Analyses: Pearson or Spearman correlation coefficients will be calculated to identify strongly correlated features.

Hyperparameter Tuning

Grid Search: Different combinations of hyperparameters will be tested to identify the most effective set.

Cross-Validation: K-fold cross-validation will be used to ensure that the model generalizes well to new data.

Model Evaluation

Confusion Matrix: A tabular representation to evaluate the performance of the classification models.

Accuracy: This basic metric will provide the ratio of correctly predicted instances to the total instances in the dataset.

Other Metrics: Additional metrics like Sensitivity, Specificity, and F1-score will be considered based on the initial model performance and the nature of the data.

Software and Tools

Programming Language: The R programming language will be used for its extensive statistical and data manipulation libraries.

Environment: Google Collaborator serves as the integrated development environment, allowing for code, visualizations, and notes to be kept in a single file.

Libraries: Packages like tidyverse, dplyr, caret, and gmodels will be utilized for data manipulation, statistical modelling, and machine learning tasks.

Documentation

Google Collaborator: This will contain the executable code, inline comments, and initial insights. It serves as a dynamic and interactive method to engage with the data and models.

Dissertation Document: This Word document will elaborate on the context, literature review, methodology, detailed findings, and conclusions. It serves as the comprehensive narrative of the research project, suitable for academic scrutiny and publication.

### 3.8. Project Deliverables

This project delivers three reports: scene-setting report, client report and reflection report. The Scene Setting Report, which briefly outlines the customer and project backdrop of the dissertation project, is the first report. Setting the scene also includes the project plan section, which provides details on the project's objectives, methodology, deliverability, timetable, and risk management. The second report is a client report. It includes detailing the project's information in depth, including the project's summary, literature review, methodology, data analysis, findings, and discussion, as well as the conclusion and recommendations. The final report is a reflection in which the author describes their project-related experiences. It will offer the conversation that began with the problem-defining process and ended with observations about the client organisation.

## 4. Project Timeline

A graph on a screen

Description automatically generated

## 5. Why is this Research Important?

The significance of this research cannot be overstated, as it has the potential to address multiple layers of the healthcare system, thereby transforming lives and streamlining healthcare operations. Here is a closer look at why this research is of utmost importance:

### 5.1. Healthcare Improvement

Early Detection: Liver diseases are often discovered at later stages, making them harder to treat. Early and accurate diagnosis can be a game-changer, offering the possibility of quicker treatment and better survival rates.

Better Treatment Plans: With an accurate diagnosis, healthcare providers can tailor treatments to the individual, improving the effectiveness of medical interventions.

### 5.2. Cost-Effectiveness

Reduced Testing: Current methods for diagnosing liver diseases can require a series of expensive and time-consuming tests. An automatic learning model could significantly reduce the number of tests needed, thus decreasing both time and financial costs.

Avoiding Invasive Procedures: Liver biopsies and other invasive methods might be avoided if a machine-learning model can provide sufficient diagnostic accuracy, sparing patients from the risks associated with such procedures.

### 5.3. Resource Allocation

Efficient Use of Medical Staff: By automating the diagnosis process, medical staff can focus on treatment and patient care rather than labour-intensive diagnostic procedures. This allows for a better allocation of human resources.

Optimized Equipment Use: Hospitals and clinics can make better use of their diagnostic equipment, which would otherwise be tied up with numerous tests for each patient suspected of liver disease.

### 5.4. Global Health

Low-Resource Settings: For countries and regions with less access to state-of-the-art medical facilities, this tool could be a lifeline. The requirement for fewer tests and less specialized equipment makes it ideal for settings where healthcare resources are limited.

Accessibility: The development of a web application as part of this research project would make it easier for remote or underfunded healthcare providers to access this diagnostic tool, thus levelling the playing field in terms of global health.

In conclusion, this research stands at the intersection of healthcare, technology, and social justice. It promises not just academic contributions but actionable solutions that could be deployed in the healthcare sector for immediate benefits. The layers of impact range from individual patient well-being to global healthcare systems, making it a highly significant endeavour in today's world.

6. Who will benefit from this Research?  
  
6.1. Academic research: often transcends the walls of educational institutions, affecting a broad range of stakeholders—from healthcare professionals and institutions to policymakers and educators. Given the critical and timely nature of liver diseases and the challenges associated with their diagnosis, the implications of this study titled "Machine Learning Algorithms for Liver Disease Detection: A New Paradigm in Healthcare" are particularly pertinent. This research contributes not only to the existing academic discourse but also provides practical insights that could have immediate real-world applications.

6.2. Patients: Patients suspected of liver diseases stand to benefit the most from this study. By streamlining the diagnostic process, they can receive faster, more accurate diagnoses that lead to timely and effective treatments. Early interventions can dramatically improve long-term outcomes and quality of life for these patients.

6.3. Healthcare Providers: Doctors, nurses, and other medical staff could integrate the findings into their diagnostic routines. A quicker and more accurate diagnostic tool would lead to more efficient patient management, freeing up valuable time and healthcare resources such as diagnostic equipment and hospital beds.

6.4. Academic Researchers: This study adds a valuable dimension to the growing body of literature on machine learning in healthcare. Researchers looking to explore further in this area will find the study to be a comprehensive resource, offering insights that could inform subsequent research endeavours.

6.5. Policy Makers: Public health leaders and policymakers can draw upon this research to shape new guidelines and regulations aimed at optimizing diagnostic procedures and resources for liver diseases. With a clearer understanding of the benefits and limitations of machine learning in healthcare, they can make more informed decisions.

6.6. Educators and Students: As an integrative study that touches upon data science, machine learning, and healthcare, this research could serve as a teaching material in medical, statistical, and computer science courses. It can guide students in their own research, thereby fostering the next generation of interdisciplinary thinkers and creators.

6.7. Technology Companies and Developers: Firms specializing in healthcare technology could utilize the findings to improve the design, reliability, and efficacy of diagnostic tools. Understanding the constraints and capabilities of machine learning algorithms in diagnosing liver diseases could lead to the development of more advanced, user-friendly applications.

# Client Report

## 1. Introduction

One of the most crucial parts of the human body is the liver. The liver's job is to eliminate toxins from the human body and regulate cholesterol and fat levels. Health problems, even death, can result from liver impairment. Hepatitis is one of the disorders that affects the liver. According to WHO statistics, the Hepatitis B virus affects 350 million people worldwide, mostly in Southeast Asia and Africa, and it results in 1.2 million fatalities annually. Patients with liver disease are on the rise, and frequent, excessive alcohol and drug use is one of the contributing factors. Apart from that, inhaling harmful gases and consuming contaminated food are also causes of liver disease. This study attempts to process a dataset taken from the UCI Machine Learning Repository database, namely the Indian Liver Patient Dataset, from this dataset it will be processed using the ML Algorithms to find out which variables are most influential in detecting liver disease. Based on the background discussion above, this study will discuss the classification of liver disease using the Logistic Regression, CART, Random Forest and KNN. Based on the results of the processing carried out, it was found that the CART Algorithm produced an accuracy value of 68.70%, thereafter, Random Forest of 72.17% of accuracy, subsequently, KNN of 71.30% of accuracy and lastly the highest accuracy of 75.65% by Logistic Regression to determining liver disease effectively.

### 1.1. Background and context

The liver is an crucial organ for the functioning of the human being1body. The functions it performs are very varied and important (Figure 1.1): it controls food intake, the concentration of many chemical and waste substances in the blood, the manufacture of bile, protein synthesis, the regulation of hormone metabolism and immunity [1, 2]. Therefore, it is very important to keep the liver in good condition. However, one of the main global health problems is the rising quantity of patient’s with1liver diseases [3] and which is largely neglected in less developed countries [4].

Liver diseases represent an important cause of mortality in the Western world, even more so in Asian and African countries, where it is major of the three foremost reasons of demise [1,5]. The continuous increase in patients with liver1disease is due to excessive alcohol utilization, breathing of dangerous gases, and ingestion of contaminated food, pickles and drugs. All these factors increase in comparable with the predominance of diabetes, metabolic1syndrome and obesity [1,6, and 7].

A diagram of the human body

Description automatically generated

Figure 1: Main functions of the liver in the human body

The main liver diseases are: viral hepatitis A, B, C, D and E, cirrhosis, alcohol associated liver disease (ALD), non-alcoholic fatty liver disease (NAFLD), Autoimmune hepatitis and hepato-cellular carcinoma [1,4,8-12]. In 2022, liver diseases accounted for around 2 million mortality per year worldwide, which can be divided mainly into 1 million due to impediment of cirrhosis and 1 million due to viral hepatitis and hepato-cellular carcinoma. Cirrhosis was the 11th most common cause of death globally and liver cancer the 16th leading cause of death, together accounting for 3.5% of deaths worldwide [8].

Liver diseases manifest variably and can be acute or chronic. They are called acute when they start suddenly and chronic when they cause damage over time, even after the symptoms disappear. Furthermore, approximately half, the analysis of liver diseases persist to be multifarious, slow, costly, sometimes dangerous and most of the time late, since symptoms are not detected until the of patients with liver disease are asymptomatic.

However, despite medical advances disease is in very advanced stages [1,13]. Its diagnosis at an premature period is very complicated since the liver functions appropriately even if it is moderately damaged and, in general, patients do not present symptoms unless the disease has progressed considerably, increasing the risk and making treatment, prevention and control difficult the illness [6,14-16]. For example, some of the most commonly used tests for the analysis of NAFLD are liver1biopsy, MRI, ultrasound, and liver enzyme tests. However, there are significant drawbacks, including the invasiveness of biopsies, the expensive price of MRIs, the limited sensitivity and non-quantitative character of conventional ultrasound, and the low accuracy of liver enzyme testing [1,13,17-19].

Liver function tests are frequently performed, both in symptomatic and asymptomatic patients, such as biochemical tests, serum bilirubin levels, albumin or coagulation tests such as prothrombin activity [13]. However, specialists are needed for the correct evaluation of these tests, causing great problems to give a quick and correct diagnosis. To address this limitation, several liver fat prediction indices have been developed, but none of them have a high enough predictive power to be considered a standard [17]. Therefore, facilitating the identification of patients in early stages of liver diseases remains a pending task to improve prevention, adequate patient treatment and alleviate the workload in hospitals [2].

To address this problem, many predictive models have been developed using traditional statistical methods to diagnose liver diseases: Fatty Liver Index (FLI), Hepatic Steatosis Index (HSI), Child-Pugh Index, Staged Liver Disease Model. (MELD), MELD-adjusted scores, Albumin-Bilirubin Index, Consortium Chronic Liver Failure Acute Decompensation Index (CLIF), and CLIF Index of Sequential Organ Failure [20, 21]. However, nowadays research is focused on searching for and creating predictive and classification models using machine learning, coming to demonstrate the great potential of these techniques in the diagnosis of diseases [2, 3, 22, 23]. These results are rapidly increasing the popularity of these automated classification techniques as aids in medical diagnosis [2, 3, 6, 7, 20]

Machine Learning is a field of artificial intelligence that allows generating predictive models efficiently and effectively through the detection of hidden patterns within large data sets. Data classification in machine learning is a two-phase process: the first is training a model on a training dataset, and the second is evaluating its performance on a test dataset [2, 6, 24]. With this in mind, such techniques can be used to make earlier diagnoses of liver diseases using data obtained from liver function tests, thus reducing costs, avoiding more intrusive methods such as biopsy, and reducing the workload of patients. specialists.

Numerous studies have been carried out looking for the best machine learning algorithm for the diagnosis of liver diseases. The most used algorithms are: Knearest neighbour (KNN), Decision tree (DT), Naïve Bayes (NB), Random Forest (RF), Support Vector Machines (SVM), Artificial Neural Network (ANN)]. These studies obtained positive results by correctly classifying patients with liver disease, presenting better or similar performance than traditional statistical modelling approaches [20, 21]. However, an algorithm that maintains a constant performance with different data sets has not been found since, depending on the characteristics of the training set, the best performance will be from one algorithm or another [16, 21].

In conclusion, machine learning models for the diagnosis of liver diseases can save time and money, improving the detection of the disease in early stages. [15, 27]. The proposal of this work follows the line of previous studies, which is to search for and create the best possible model for the automatic diagnosis of liver diseases. In addition, unlike the rest of the studies, the model is implemented in a simple and intuitive web application so that all users can make use of the predictive model, whether or not they have technical knowledge of machine learning.

### 1.2. Problem Formulation

From the background, several problem formulations can be identified, as follows:

a. How to get data that can be used in research?

b. How to classify Negative (Normal) or Liver disease in patients?

c. How do the results compare Accuracy, Precision, Recall, Specificity using F1- Score for the classification of Negative (Normal) or Liver diseases using the Logistic Regression, KNN, CART (Classification And Regression Tree) and Random Forest.

### 1.3. Work Objectives

The general and specific objectives of this work are:

**• General objectives:**

1. Create an automatic learning model to classify patients with liver disease based on different biochemical parameters.

2. Develop a web application that implements the created model.

**• Specific objectives:**

1. Analyse the data set to select the best predictors.

2. Train, evaluate and optimize different machine learning algorithms like Logistic Regression, CART (Classification And Regression Tree) and Random Forest.

3. Determine the best prediction model with a minimum accuracy of 70%. [52]

4. Create a simple and intuitive web application for the user that allows making predictions and visualizing the results.

## 2. Literature Review

### 2.1 MACHINE LEARNING

Machine Learning (ML) or Automatic Learning, is defined as a set of methods that allow computers to automate the creation of models that are based on data through the discovery of patterns [29, 30]. Researchers in [29, 30] defines the term ML as a "field of study that gives computers the ability to learn without being explicitly programmed", focusing on the development of computer programs that can make changes when they show new data. Likewise, the definition of ML where "a computer program learns from an experience E concerning some class of task T and a performance measure P, if its performance on task T, measured by P, improves with experience E”[ 29, 30].

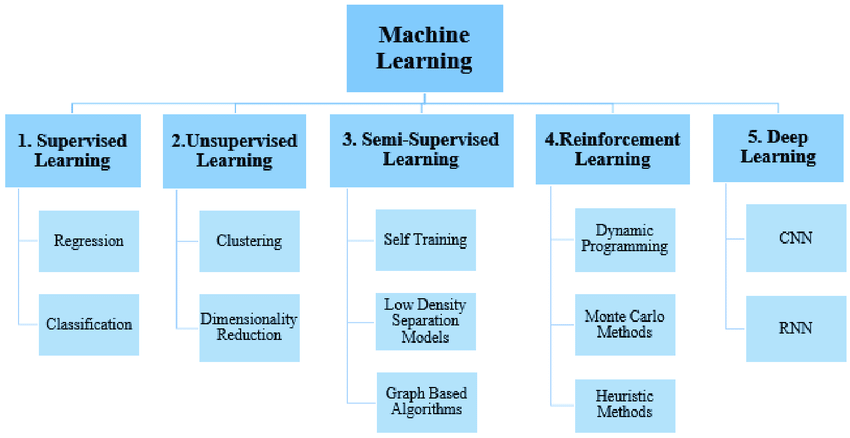


Figure 2.1: Type of Machine Learning Techniques

* **Machine Learning Models**

The main models that are developed in learning machines are the types of supervised, unsupervised and semi-supervised learning; Next, a description of the learning models is made.

* **Supervised Learning**

Supervised Learning it depends on a human interacting with the model to teach it the conclusions it must reach, that is, the output of the algorithm is already known [29, 30]. The algorithms receive the input data set as with the correct results and the algorithm is trained by learning and comparing the actual results with the correct results to identify the incorrect cases in the classification. Within the framework, the use of supervised algorithms will make it possible to classify and separate events associated with deaths from those that are not. To use this type of technique, it is necessary to develop a learning model, which allows the algorithm to understand the differential characteristics and with greater discriminant power of each class, in such a way that the separation of the predictor or objective variable can be maximized and understand the factors or elements that separate one group from another [29, 30].

As an example from a database used as training examples with a specific target label, such that {(𝑥1, 𝑦1 ), … , (𝑥𝑖 , 𝑦𝑖 )} where 𝑥𝑖 corresponds to the output vector of the ith example and 𝑦𝑖 is its label, likewise, its objective is the creation of a model 𝑔 ∶ 𝑋 → 𝑌, where the label 𝑌 can be predicted from future data 𝑋 [29, 30]. For the supervised learning model, there are two types of problems when they are of the regression and classification types, these problems are described below as:

1. Regression: For the Regression type problems, the objective is to develop a relationship between the data of output and input data, which can use a continuous function to help learning machines understand how the outputs change for given inputs [29, 30].
2. Classification: Classification consists of establishing the classes to which a new input data belongs, based on a set of training data whose quality is known [29, 30].

* **Unsupervised Learning**

Machine Learning (ML) is a scientific discipline that manages intelligent systems, that is, systems that learn automatically by identifying certain patterns present in the data, using algorithms that are responsible for reviewing data through examples or predefined instructions to predict future behaviours. from additional information and readjust the result; manages inductive knowledge, obtaining a general statement based on statements that describe particular cases [29, 30], its process is similar to data mining, both systems look for patterns among the initial data; mining extracts the data for human understanding while ML uses that data to detect patterns and realign program actions.

As [29, 30] mentions, machine learning was born in the field of computing whose processing is associated with learning, where the machine is not programmed to respond in a certain way based on the inputs received, but rather to extract patterns. of behaviour from the inputs received, and based on said information learned or assimilated, evaluate new inputs, in this sense, the terms learned and assimilated conclude that machine learning demands a degree of independence in the learning process, and this has a very important meaning in the advancement of Artificial Intelligence, whose processes are:

• Select problem: Define the real problem to be solved, analysing advantages- disadvantages, costs, and high-level benefits.

• Select attributes: Set the attributes or variables of easy access and extraction to solve the problem.

• Select model: Define the algorithm that best adapts to solving the problem.

• Prepare data: The extraction, consolidation, and scaling of attributes that make up the data file to be used.

• Analyse data: Refining the data or attributes based on the algorithms that best solve the problem posed

• Evaluate the model: 60% of the data collected will be training data; 20% to cross-validate model performance for debugging and refitting; and the remaining 20% of the data collected for the test of the generated model and its verification.

• Post model: Use it in a productive environment.

• Monitor and adjust: Once the model is in production, monitor its performance and make adjustments to improve it.

* **Semi-Supervised Learning**

Semi supervised models generally has an amount of training data that can be labelled and another large amount of data that does not have labels; this indicates that unlabelled data, when used in conjunction with a small amount of labelled data, substantially improve learning [29-30]. For semi-supervised learning problems they can be described as follows: For the training data set 𝑋 having 𝑙 labelled data: 𝑋𝑙 = {(𝑥1, 𝑦1 ), (𝑥2, 𝑦2 ), … , (𝑥𝑙 , 𝑦𝑙)} And of the remaining data, they are those that do not have any label. 𝑋𝑢 = {𝑥𝑙+1, … , 𝑥𝑛} The unlabelled data becomes larger than the labelled data 𝑛 − 𝑙 ≫ 𝑙.

### 2.2 VARIOUS ML MODELS

* **Decision Tree**

The decision tree1methodology is a widely1used data mining1method to establish1classification systems1based on multiple1covariates or to develop prediction1algorithms for an objective variable [31-34]. The Tree Models behave like a cluster analysis (creation of homogeneous groups and different from each other) and at the same time as a predictive method. Given a data set, we will obtain diagrams of logical constructions that serve1to represent1and categorize1a series of conditions that occur successively, for the resolution of a problem both classification and regression. The process of creating the tree is laborious due to the complicated casuistry and combinatorial that is generated each time it is necessary to divide a node. [31-34]

A diagram of a split-splitting process

Description automatically generated

Figure 2.2: Model Decision Tree

The construction of trees must take into account the following aspects: criteria for the selection of independent variables for the division of each node, criteria for choosing the optimal cut within each independent variable or node, criteria for choosing cut groups for independent variables nominals with more than one category, the minimum number of observations to build each node, stopping criteria, limits on the number of nodes and splits, criteria for handling missing values, and whether validation data will be used to control the construction process [31-34].

Regarding the criteria for selecting the optimal cut-off point for each variable, since we find ourselves in our specific case with a problem in which both the dependent variable and the predictors are qualitative, some are:

Chi-square: the cut-off is a selected or grouping of categories of the independent variable with the highest value of the associated statistic, crossing the dependent variable with the independent one. Normally a significance test is applied.

Gini index: the smaller the better, since it indicates a greater distance between classes (also called impurity). Subsequently, the division of the independent variable that improves the Gini index is selected compared to not using it.

Entropy: measures the information gained in divisions. The less, the better, the more information. Subsequently, the division that improves the entropy index concerning the base entropy of the parent node is selected.

Regardless of the selected criterion, for simplicity, the possibility that each node is only divided into two parts at each step will be considered. The algorithm ends when some of the following stopping criteria are met: there are not enough observations in the final sheets to consider splitting, the maximum depth has been reached, or there is no improvement in the splitting criteria at any node. The final tree model may be overfitting if the data is complex.

After obtaining the final tree, you can act as in the cluster methods: choose the subtree that seems most stable. This process is known as pruning. Finally, regarding the advantages of these models, it is a technique that has great descriptive power, allows non-linear relationships, there is no theoretical assumption of the data, provides measures of the importance of the variables, etc. However, among the disadvantages, we find low reliability (poor generalization) and low predictive efficiency.

* **Random Forest**

The Random Forest algorithm is a supervised learning technique that generates8multiple decision8trees on a training data set: the obtained results are combined to obtain8a single8model that is8more robust8compared to the results of each tree separately.

A diagram of a diagram

Description automatically generated

Figure 2.4: Random Forest Diagram

The Random Forest algorithm goes one step further in avoiding the problem of variable selection, avoiding rigidly deciding on a set of variables, and taking advantage of bagging at the same time. It is a question of incorporating two sources of variability (resampling of observations and variables) to gain generalization capacity and reduce overfitting while conserving the ability to adjust well particular relationships in the data (interactions, nonlinearity, cuts, problems). extrapolation, etc.). Random Forest also avoids the problem of very dominant predictor variables. With only Bagging, in the case of a very dominant pair of variables, the trees would be similar. By adding randomness to the variables used, different trees are obtained, which reduces the variance of the model [35-37].

where5p (t) is the5probability that5a given5 example5corresponds to5leaf t5and φ(t)5is the5impurity of5the terminal5node t.

* **Logistic Regression**

A better understanding of the factors involved in pentest would help to make a better decision and the reason is that it provides us with a better analytical capacity that allows us to choose the best possible path, for this research we will make use of the logistic regression because it is a versatile tool and that with dichotomous variables it is the most used [38-40]. For this, we must ask ourselves, what is logistic regression?

ML regression analysis deals with the study of the dependencies of a variable of interest concerning one or more explanatory variables through techniques for their analysis and modelling. It helps us estimate the uncertain expectation of the dependent8variable given the independent8variables. The8objective of the estimation is a purpose of the independent8variables called8the regression8function and it describes the variation of the variable of interest based on a probability distribution [38-40].

Many methods have been8developed to8carry out regression8analysis and their use in practice depends8on the form8of the data generation process which depends on making assumptions about this process.

Where 𝑥 is a real number.5From this equation, we can see5that as5𝑥 approaches5minus infinity, the quotient5approaches zero5and as5𝑥 approaches5infinity the5quotient approaches5one, as5shown in5figure 2.11.

A graph with a line and a line

Description automatically generated

Figure52.5: Graph5of the5sigmoid function5and its5threshold

To used5to approximate5the dependency5relationship between5a dependent5variable Y, the independent5variable X, and a5random term5ε. This5model can5be expressed5as:

To predict5a ***Y***5value greater5than 1.5The specific5form of5the logistic5regression model5is:

The previous defaults drive the evolution towards generalized linear models, within which the so-called log-linear and the so-called LOGIT models can be included as :-

Therefore,5the logistic5model expresses5the dependent5variable as5the occurrence5or not5of5an event5in terms5of probability.5Among the5positive aspects5of this5methodology, it5is found5that it5is a5simple and5easy to5interpret model [38-40], it5is a5light process5from the5point of5view of5computational resources,5in addition5to allowing5the use5of multiple5variables even5with few5variables.

Consequently, the model can be closer to reality since many phenomena, such as those in the epidemiological field, are more like a curve than a straight line. In addition, the exponential curve chosen as the best fit can be logarithmically transformed into a linear equation of all the variables, so that the mathematical apparatus studied for multiple linear regression will be applicable; although the researcher has, in the end, to undo the transformation to interpret his conclusions. If for the logistic regression model a regressor variable of the categorical type has c levels, c-1 fictitious variables (dummy) will have to be generated so that all the possibilities of the variable are well represented in the logistic model. When all the regressor variables are categorical then the linear Log model is used as:-

**(eq.2.9)**

The probability alteration *𝜋 (𝑥)* is depicted by the subsequent formula as:

**(eq.2.10)**

* **K-Nearest Neighbour**

The basic implementation of KNN for classification problems is to calculate the distance between a query and each of the objects in the set of training (i.e. database) to obtain the neighbourhood formed by the k nearest objects and assign to the query the most repeated class in the neighbourhood. The KNN algorithm needs to store the entire training set and learning just needs to read the stored data without needing to do any processing on them. In addition, for its implementation, it is not necessary that the data are linearly separable, nor knowing the distribution of the data in general. These characteristics make it very simple and widely used [41-45].

A diagram of a graph

Description automatically generated

Figure 2.6: K Nearest Neighbours with k = 3 in Two-Dimensional Space

The KNN algorithm is illustrated in Figure 2.6 where we have as an example a value of k = 3 in a two-class classification problem with two-dimensional data (x1 and x2). In the example, the decision for q1 is made directly because all its neighbours are of class 1, so q1 is also of class 1. On the other hand, for q2 there is one neighbour of class 1 and two neighbours of class 2, for which a rule to decide which class to assign to the query q2. This case can be solved with a simple majority rule, which would assign q2 class 2, or a rule that adds weights to the distances to give greater influence to the closer neighbours. To use the KNN classifier it is necessary to have a distance measure that allows comparing the stored examples with the queries to classify. Given Since in most cases the data is represented by vectors in a multidimensional space, a training set X is assumed to be made up of ()i∈[1,n] training examples (where n = |D|). The Training Vectors o examples are described by a set of F features that equal the number of dimensions [41-45]. Each example is classified with the label yj ∈ Y. The goal is to classify the unknown example q. For each ∈ X, the distance between q and as shown in Equation 2.11.

With this function, there is a sum over all the characteristics with weight for each characteristic.

Finally, in addition to the distance function used to define the closeness or similarity between two objects, the performance of KNN crucially depends on the size of the neighbourhood defined by the hyperparameter k. Choosing the optimal value of k is a particularly difficult task. A small value typically results in very noise-sensitive decision boundaries, while a large value produces a noise-robust neighbourhood that is likely to include many neighbours of other classes (see Figure 2.7). Usually, a balance is sought between both cases. There is an optimal k for each distribution of the data; but it is not possible to know it since it would be necessary to know the label of the query, which is tautological. For these reasons, many methods have been proposed to find good values ​​for k; which, despite being effective, increase the overall complexity of the algorithm [41-45].

A diagram of different colored squares

Description automatically generated

Figure 2.7: Effect of the Choice of K.

The k nearest neighbours rule is just one way to get into a neighbourhood. As far as we know, it is the only rule used to obtain neighbourhoods in example-based learning methods. A central part of this work is an alternative way of obtaining a usable neighbourhood in example-based classification. The weakness of the k-neighbour classifier is the need to set the value of the hyperparameter k, which is critical for good classification results. Our task will be to propose a new rule to obtain a neighbourhood that is free from this limitation.

* **Evaluation of Classification Models**

Once Machine Learning models are used in a data set, the results obtained must be evaluated according to techniques and/or evaluation parameters that are described in more detail below.

* **Accuracy**

Also called a success, it is simply1the ratio of1correct predictions1to the number of predictions made:

It is usual for the analysis of machine learning models, to show the result of its evaluation in terms of the confusion matrix. Table 2.1 illustrates the values ​​that are calculated to make up this matrix:-

|  |  |  |
| --- | --- | --- |
|  | **Observation**2**Positive** | **Observation**1**Negative** |
| Prediction1Positive | True1Positives (VP) | False1Positives (FP) |
| Prediction1Negative | False1Negatives (FN) | True1Negatives (VN) |

Table 2.1: Construction of a Confusion2Matrix

* **Precision1**

Determines the1fraction1of records that1actually1turns out to be positive and that have been effectively classified as positive. The higher1the1precision, the lower1the1number of cases erroneously classified as false negatives [46-48]. Therefore, It is the1number1of correct1predictions of the positive value over the total1number1of1predictions of the positive value as:-

With precision, we can observe what probability of success the classifier has when it emits a positive prediction.

* **Recall**

Also known as sensitivity (although it is more common to find it as recall in the literature of any language), is the1number of true1positives relative1to the number1of true positives1and1false negatives:

The recall allows us to have an idea of ​​how successful the classifier is in predicting positive cases.

* **Scope or Sensitivity**

Determines the fraction of positive1cases correctly1classified.

False-positive1ratio: It is the percentage of negative1cases erroneously classified1as1positive. False-negative1ratio: It is the1proportion of1positive1cases erroneously classified1as1negative.

* **Area**1**Under**1**Curve AUC/ROC Curve**

The1Area1Under the Curve, that is, the area1under the1curve, actually refers to the area under the receiver operating characteristic curve (AUROC). This is obtained from the graph where the vertical axis contains the recall and on the horizontal axis the rate of classified positive examples erroneously as such, as the decision threshold for the classifier varies. Note that, for this reason, the metric cannot be used with classifiers in which its exit is not a probability. In this way, you have a vision of how it is affected the ability of the model to classify examples correctly when the threshold of decision [46-48]. This curve is expected to be steep, indicating that regardless of the value The chosen model performs well. One of the benefits of using this metric is that the bias that can induce the choice of the value for the threshold is removed. Even it is a graphical representation of the relationship between the true positive and false positive ratios of the classifier. As pointed out by [46-48], the area under the ROC curve (AUC) provides an approach to assess which model is better on average. If the model is perfect, then its AUC would be equal to 1. If the model is random, its AUC would be equal to 0.5.

* **F-Measure/Score**

It is a metric that unifies precision and recalls. F-Score commonly referred to as the F1-Score, is a metric that combines accuracy and recall. By itself, it does not have a relevant interpretation, but it is generally used to compare the overall performance of various predictive models. It is calculated as:

## 3. Proposed Model

### 3.1. PROPOSED MODEL

This study classifies liver diseases using the methods based machine learning on RNA data. This research is included in the types of quantitative research because of the type of data used in this study numeric or numeric form and there is a calculation and analysis process on the results obtained. The results of this study are the liver diseases classification system using the machine1learning methods like Logistic Regression, KNN, Decision Tree and Random Forest methods based on RNA data which later is expected to be useful to help experts in the medical field to classifying cancer quickly and with high accuracy taken into account.

The framework of this research is shown in the flow chart that can be seen in Figure 3.1. The diagram shows that the stages of This research consists of pre-processing using normalization, training and testing the model using the machine learning methods using supervised learning approach based on Logistic Regression, KNN, Decision Tree and Random Forest, thereafter evaluating the model using confusion matrix.

A diagram of a process

Description automatically generated

Figure 3.1: Proposed Model

### 3.2 LOGISTIC REGRESSION ALGORITHM

Logistic Regression denomination comes from the predicted values ​​in the variable dependent (VD) from the values ​​of the independent variable (IV), have smaller variance than that of the empirical dependent variable (Var(Y’) < Var(Y)). The difficulty in being able to discriminate between the effects mentioned, and for the empirical verification of the assumptions of the model, imply that this first concept of regression has evolved. Currently the regression, in a broad sense, it designates the set of procedures used to construct mathematical functions (with their corresponding error term in the case of linear models), and their “logit” transformations, which allow estimate or predict the behavior of one or more variables from others variables, with which they are strongly correlated. Linear regression is a mathematical procedure that models the relationship between a DV, or Y variable, as a function of a set of variables independent (Xi) and an error term (εi), which includes all those uncontrollable factors, attributed in part to chance and in part to errors random, unknown or uncontrolled, which give it its character stochastic The model is expressed according to the equation:

*Linear regression models require the fulfilment of the following assumptions:*

*a) Normality: the values ​​of the variables are adjusted to the normal distribution.*

*b) Linearity: the relationship between RV and LV or LVs is linear.*

*c) Non-colinearity: none of the IVs can be an exact function (principle of non-linear combination) of other independent variables, otherwise collinearity or multicollinearity occurs.*

*d) Independence: the measurement errors of the regressor variables are independent of each other.*

*e) Homocedasticity: the error variance is constant throughout the measurement range.*

*e) The mathematical expectation of the errors is equal to zero.*

*f) The normal distribution of measurement errors.*

### 3.3 KNN ALGORITHM

The basic principle on which this algorithm is based is that a new case is going to sort into the most frequent class to which its K nearest neighbours belong. The nearest neighbour method can be extended using not one, but a closest data set to predict the value of the new data, in which is known as the k-Nearest Neighbours. To the consider more than one neighbour, noise immunity is provided and the noise curve is smoothed with estimate.

1. The neighbourhood classification rules are based on the search in a set of prototypes for the k closest prototypes to the pattern to be classified.

2. There is no global model associated with the concepts to learn.

3. Predictions are made based on the most similar examples to the one to be predicted.

4. It is known as lazy learning mechanism.

5. In KNN, a metric must be specified in order to be able to measure proximity. For computational reasons, the Euclidean distance (δ) is usually used for this purpose.

* **Algorithm k-NN classifier**

*Input: k number of neighbours, Dtrain training set*

*1: for each test instance z = (x',c') do*

*2: Calculate d(x', x), the distance between z and each instance (x , c) ∈ Dtrain*

*3: Select Dz ⊆ D, the set of k training instances close to z*

*4: c' = max´v P ∑(x',c')∈D I(v = ci)*

*5: end for*

### 

For a formal definition it can be said that a random forest is a classifier consisting of a collection of classifiers with a tree structure {h (x, Θk), k = 1,...} where {Θk} are random vectors independent and identically distributed, and each tree casts a unitary vote for the most popular class in input x. The random forest algorithm (for both classification and regression) is as follows:

### 

On the other hand, multivariable trees are also called oblique, since the tests they perform at each node are equivalent to hyperplanes in an oblique orientation to the axes of the attribute space.

## 4.TOOLS AND TECHNOLOGIES

**R-STUDIO**  
As data generation continues to grow exponentially, there is a greater need for tools and methods that can analyse and interpret this data. Data mining and machine learning have become increasingly important for understanding and predicting various phenomena. Implicative Statistical Analysis, although not widely adopted, is a toolset that integrates with machine learning and can be implemented in R, a free, cross-platform programming language.

To promote the use of Implicative Statistical Analysis, it has been packaged as a library for R, making it accessible to a broader audience. RStudio, an open-source IDE for R, further enhances the development process by allowing code modification and distribution, facilitating continuous improvement.

* However, the study revealed that the code used for generating graphics had high repetition and lacked flexibility. To address this, an application was developed to generate statistical graphs in a more generic and adaptable manner. This application is open-source and can be accessed via a web browser, hosted on Google Colab.  
    
  **R Programming Language**

The R programming language is a language that is oriented to the analysis and presentation of statistical information (The R Project, s/f). Therefore, it already includes among its libraries the forms of perform various statistical analyses, as well as various ways of representing information obtained graphically. In the same way, its syntax is oriented to work with data statistics (that is, data that is normally represented by one or more tables). The Graph visualization in R has improved remarkably in recent years thanks to the development from libraries such as ggplot (Valero-Mora, 2010) and plotly (Sievert, 2020). In this scheme, it wrote the R code using RStudio.

* **RStudio**

RStudio is an integrated development environment (commonly called IDE for its acronym in English for Integrated Development Environment) for the R language. RStudio simplifies the creation and work of R projects of different types, as well as providing facilities and services that they complement the development such as the installation and management of R packages, use of version control, help and code auto completion, etc [49-51].

* **R Packages**

R packages are collections of code that allow you to implement or add functionality to an R application. Installing packages in an R application is easily accomplished through RStudio, which provides tools that simplify the process. to install, manage or remove packages. The most important packages used for the project are listed below. Note that the listed packages do not represent an exhaustive list of all packages used by the project, but only those packages that were deemed to require a separate explanation etc [49-51].

* **reshape2**

Reshape2 is a R library containing rich data structures and tools. It has been created to improve the adoption of R-Studio in the scientific programming community by bridging the gap and enriching R for data analytics and data manipulation applications. Specifically, some of the features it provides are label-based data access, advanced pivoting and reshaping, grouping and aggregating of data, combining and joining of datasets, hierarchical indexing, data alignment, and dealing with missing data, etc. Traditionally, this package works with data collections in a relatively sequential manner. The reshape2 functions now offers chunk size to read huge files in chunks to make the processing more memory manageable. When dealing with huge data collections, faster processing times can be achieved by parallelizing the process and making use of the many processors that are now available in laptops, consequently further performance boosts can be attained by using distributed systems and clusters etc [49-51].

* **Caret**

This module includes algorithms for support supervised learning methods that can be used in classification and regression. They are quite effective in high-dimensional spaces.

* **Tidyverse**

Tidyverse is a assortment of necessary R packages for data1science. The packages under the tidyverse umbrella help us in performing and interacting with the data. There are a whole host of things you can do with your data, such as subsetting, transforming, visualizing, etc.

* **Shiny Package**

The Shiny package is an R package with which you can quickly design graphical interfaces for the analysis and visualization of data entered by R (Shiny Team, s/f). Through this package, you can perform functions such as data entry or dynamic interaction by the user for data visualization. The operation of Shiny consists, in essence, of two large categories of components: the category of reactive components, which are elements that cause a reaction in the rest of the application (for example, pressing a button); and the category of reactive components, which change their presentation or functionality after receiving an event from a reactive component (RStudio Team, 2017).

These two types of components are useful for generating dynamic web pages, where changing the state of certain elements within a web page does not require the web page to be refreshed or to navigate to a different page within a site— it is enough to generate a reactive event so that the content is updated automatically.

* **Ggplot2 Package**

The plots that can be generated via the R language are relatively unconfigurable, and are not very visually appealing, but are only present for quick viewing. For the generation of modern looking and highly configurable plots in the application, the ggplot2 package was used. It allows you to generate a large number of different types of statistical graphs, as well as configure many aspects of their appearance, including the colour palette used, the filling of the graphs, where the labels that represent the data are placed, among many others. options.

* **Vcd Package**

The vcd package, which stands for Visualizing Categorical Data, is a package similar to ggplot2, in that it provides functions and libraries for graphing categorical data. Vcd was used for the generation of the mosaic plots that the project needs, which are generated with more configurability than the ggplot2 mosaic plots.

* **Viridis Package**

The viridis package complements the generation of plots with ggplot2 by providing a colour palette that is specifically designed to have a distribution of colours that is most distinguishable to the human eye, both for people with normal vision and those who have some form of vision. colour blindness, so that the visualization of data requires the least possible effort on the part of the people who perform the interpretation of the generated graphs.

* **Incremental and Iterative Development Process**

The incremental and iterative development process is a process in the area of ​​software development whose methodology focuses on development taking place in phases that are repeated throughout the software development cycle, so that the product is constantly evolving to adapt to new needs and requirements that may be encountered during the process, thus allowing the functionality and architecture of the final product delivered to not be tied to a design made before the software development phase as such (Paasivaara, 2006). What has been described above refers to the “iterative” part of the process. The phases that involve this iterative process are not formally defined, but normally include a planning phase during which the requirements are found; a phase of analysis of the requirements and design of the functionality; a functionality implementation phase; and a testing and evaluation phase where the implemented functionality is tested and whether changes are required.

A diagram of a process

Description automatically generated

Figure 4.1: Diagram of a particular way of implementing iterative development

Regarding the incremental aspect of the process, it refers to defining and producing components of the final product that are small, and that, as the iterative process continues, these components are added to each other to "increase" the functionality of the product, until It is considered that all the necessary components have been produced and the combination of these contains all the functionality required by the interested parties of the software project. It should be noted that this development process is not strictly formally defined, and therefore it is highly adaptable to the needs of the project in which it is applied.

As described above, the phases applied in the iterative aspect of the process can vary, as can how the size of a component is defined and the time required to invest in developing that component. In this work, it was considered that a development process such as the iterative and incremental one allowed to have a certain level of definition of the phases in which the software will be developed, but with the flexibility required by this type of project. Other processes, such as agile methodologies, also largely implement an incremental and iterative philosophy, but these types of methodologies normally require a more strict definition of steps and stakeholders, and are designed to be applied with larger work teams at the same time. of the present work.

## 5.RESULT AND SIMULATION

### 5.1 DATA IMPLEMENTATION

The data used in this study is in the form of liver disease dataset extracted from (https://www.kaggle.com/datasets/jeevannagaraj/indian-liver-patient-dataset) which is comprising 416 liver patient records and 167 non liver patient records. The data set was collected from test samples in North East of Andhra Pradesh, India. 'is\_patient' is a class label used to divide into groups (liver patient or not). This data set contains 441 male patient records and 142 female patient records.. This data is in CSV format (Comma Separated Value). The data will be processed first in the preprocessing stage so that a training dataset is produced and testing datasets that can be used by machine learning models. Below is the example dataset for perusal and ready reference.

A screenshot of a computer program

Description automatically generated

A table with numbers and letters

Description automatically generated

A screenshot of a computer

Description automatically generated

Figure 5.1: Example Dataset

* **Pre-Processing**

Data selection is the pre-process phase of collecting data in accordance with the research. At this stage the process carried out for the selection process is in the "attribute" selection therefore few columns need to be rectified for normalization. The below code block depicts the normalization scenario.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

Figure 5.2: Data Normalization

* **Data Visualization**

After getting the right model from the data normalization results, then calculations are carried out based on the data categories that have been classified into topics according to the specified exemplary data visualization like Total\_Bilirubin, Direct\_Bilirubin, Alkaline\_Phosphotase, Alamine\_Aminotransferase etc.. The results of calculating the number of categories into the topic are then visualized into the dashboard using the R programming language. The dashboard design that will be displayed from the analysis of topic modeling, is distinguished based on the information content that will be displayed in Figures below:-

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

Figure 5.3: Histogram Bar Charts for Visualization.

* **Split Train and Test**

At this stage, the data and system will be tested to determine the ability of the system built to predict liver disease cases. In theory, the ML methods does not have a standard rule to determine the optimal architecture to be applied to the system so that the search for training parameters must be done by trial and error. The level of accuracy obtained from the training and testing process in this study will be represented through the confusion matrix. At this stage, the first test is carried out to find the most optimal number of random values. Therefore, under the scheme 80% data is opted for training and 20% data for testing.

A screenshot of a computer program

Description automatically generated

Figure 5.4: Splitting Dataset for Training

### 5.2 EVALUATION USING RANDOM FOREST

Random Forests is a generalization of the Bagging methodology, which builds multiple trees based on the CART algorithm and later combines the predictions or classifications (Liver Disease or No Liver Disease) delivered by each tree. With this, what is intended is to reduce the high variance of the response of an individual tree to test data and therefore improve the performance of the method. Random forests, like Bootstrap Aggregation (Bagging), are part of a group of methods known as assembly methods, which are attractive mainly because they are capable of boosting methods weak and convert them into strong methods, with which very precise predictions can be made, the latter being precisely what is required to improve or enhance the trees. Random forests are made up of a set of unproved trees, from which a global prediction or classification value is extracted for the response. In the case of regression, this value is obtained by averaging the predictions generated by each tree in the forest, while in the case of classification, it is obtained by selecting the class (Liver Disease or No Liver Disease) that achieves the most votes among all the trees of the forest. The performance of random forests, like that of trees, can be superior to that of classical parametric methods when there are complex interaction effects between the variables and when the functional form that approximates the true model is nonlinear. Therefore, the below code block depicts the scenario.

A computer screen shot of a program

Description automatically generated

A screenshot of a computer

Description automatically generated

Figure 5.5 Results using Random Forest based on Confusion Matrix

Random forests are built by generating from the training data a pre-established number of random resamples (ntree), which are taken with replacement in a similar way to Bootstrap Therefore, the re-samples incorporate slight random variations, they continue to reflect the behaviour of the process that generated the data. In this way random forests manages to mitigate the fact that the partitions of the trees are not carried out under a global optimization process but rather a local one and consequently it does not guarantee that the resulting trees are the most optimal with respect to all the combinations possible variables, on these and other parameters that influence the stability of the estimates. Therefore, achieving the confusion matrix vide specificity of 100%, sensitivity of 0.03%, and accuracy of 72.17% respectively.

### 5.3 EVALUATION USING LOGISTIC REGRESSION

The logistic1regression technique1is used to build a model1that relates a1categorical1dependent variable1and one or more1variables1independent, so scheme can compute the conditional probability that an liver disease exists or not. Therefore, logistic regression model, it is necessary1to define1a dependent1variable, which is the response1variable. This dependent variable classified according to their operational roles during drilling operations using regression, corresponds to the output the scheme want to obtain, and can only have two values, zero and one. Thus, for our objective (Liver Disease or No Liver Disease), to determine a model that can predict the position losses, we will take as dependent variable, the variable excursion, which will take the value zero if there is no excursion and the value 1 if there is excursion. Thereafter, the dependent variable will be the human1cause, and will take a1value of1zero if there is no human1cause and 1 if there is a cause human. Except for the variables depth of liver disease, percentage of propellants in line and percentage of online generators, which are all quantitative, the rest of the1independent variables are categorical. Because of this, the model manipulates its values ​​internally, so that it produces1as many1variables as categories1minus one. The code block and results are depicted below:

A black and white image of a mathematical equation

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

Figure 5.6 Results using Logistic Regression based on Confusion Matrix

The scheme try to get a model that1allows us to predict if a liver disease exists based on parameters Total Bilirubin, Alkaline Phosphotase, Alamine Aminotransferase, Total Proteins, Albumin, and Age. To do this, the dependent variables chosen will be the cause variables, which will take the value 0 if there1is no1cause, and 1 if the incident is due to error. The other variables are called independent variables, and will be the same as those listed in dataset. For this purpose, we will once again use logistic regression, the score allows comparing for each independent1variable Xj the1null hypothesis: Ho1=1βj = 0; that1is, the coefficient1associated with the1variable in the model1is null. The1variable that presents1the associated minimum1p-value. There1will be several steps1where variables will1be entered and deleted1independent, according to1the criteria set1out above. At step 0, only the constant is entered into the model. For this constant, it is important to measure B (the value of the estimate of the constant), the error standard in the estimation. Subsequently, achieving the confusion matrix vide specificity of 96%, sensitivity of 0.24%, and accuracy of 75.65% respectively.

### 5.4 EVALUATION USING KNN

The KNN algorithm is one of the simplest classification1methods, it1was developed from1the need to1perform a discriminant1analysis when the1parametric estimates reliable1probability densities are unknown1or difficult to1determine. This research work on non-parametric discriminant analysis, and estimation of probability density properties of the KNN e.g. if k=1 and n→ ∞ the classification error has its upper bound at twice the error rate. KNN is a statistical classification1method used to1classify objects, based1on the closest training1examples in the feature1space. you will assign to the object the class1that is most common1within its nearest neighbours the scheme wants to classify an object of unknown class, in this case, the malignant; then if we take the 8 nearest neighbours (k=8), the class to which the malignant belongs is to the class of non-cancer, while if k=5, it already belongs to the class of the benign, and so on if you want to take into account more neighbours. yes, k is very large, and the risk of the forecast being influenced by a pattern decreases. of noisy training close to the observation, but also reduces the sharpness of the method, although the estimate is better if the number of examples around the observation is greater. The below code block depicts the scenario and results.

A computer code with black text

Description automatically generated

A computer code with text

Description automatically generated with medium confidence

A black text on a white background

Description automatically generated

A screenshot of a computer

Description automatically generated

Figure 5.7: Results using KNN based on Confusion Matrix

One of the parameters of this method determines the metric distance between two instances, due to the need to know the nearest neighbours which requires several functions since the Minkowski distance is always the most effective in a feature space of more than 2 dimensions. The algorithm takes each instance of the training set and measures the distance between this and all instances of the sample set, and then records the k smallest distances to perform the classification for breast cancer detection. The metric distances used for this algorithm are the following: Square Minkowski distance: The distance between 8 points x and y of the m-dimensional Minkowski space are defined as the sum of the differences to square. Therefore, achieving the confusion matrix vide specificity of 93.90%, sensitivity of 0.15%, and accuracy of 71.30% respectively.

### 5.5 EVALUATION USING CART

Under the scheme CART in a prediction model whose objective is inductive learning, which is represented by a tree, based on observations and logical constructions. It is the most widely used classification model. The result of the use of the algorithm designs a regression tree, for which one of its advantages is the easy interpretation and quick way to establish if a patient, through a certain number of variables or clinical data, is at risk of falling into liver disease. The regression tree is made up of a set of1nodes, leaves1and branches.1The main1or root1node of1the tree1is the1attribute where1the classification1process begins.1The internal1nodes are1the questions about1the particular1attribute of1the problem1parameters related to1liver disease.1That is,1responses are1represented by1a child1node. The1branches between1each node1are the1possible values1​​of the1attribute. The1leaf nodes1belong to1a decision,1which must1coincide with1one of1the variables1of the1problem to1be solved1(Liver Disease1or No1Liver Disease).1The below code1block depicts1the scenario.

A computer screen shot of a program

Description automatically generated

A screenshot of a computer

Description automatically generated

Figure 5.7 Results using CART based on Confusion Matrix

Once this type of model is executed, there will be only one path which will depend on the current value of the studied variable. The algorithm for generating decision trees has two stages: 1 Tree induction: In this stage, the decision tree begins to be assembled using a training set. It starts1by generating1the root1node, choosing1a test1attribute, and1dividing the1training set1into two1or more1subsets, so1that a1new node1is generated,1and so1on. When1a node1has objects1of more1than one1class, an1internal node1is generated,1when it1has objects1of one1class, a1leaf is1formed.

Classification of1the tree:1In the1next stage1of the1algorithm, each1new object1is classified1by the1constructed tree,1then the1entire tree1(from the1root to1a leaf)1is traversed.1The path1to be1taken is1determined by1the decisions1made at1each internal1node, according1to1the test1attribute present.1How the1decision tree1algorithm works:1To build1the algorithm1based on1decision trees,1splitting of1the data1is needed.1One of1the options1to carry1out this1division uses1some measures: Entropy: Introduced by Shannon in his information theory. The essential aspect of decision tree algorithms is to choose the best criterion when dividing the data, for this entropy is used. Therefore, achieving the confusion matrix vide specificity of 80.49%, sensitivity of 0.39%, and accuracy of 68.70% respectively.

### 5.6. DISCUSSION ANALYSIS

**General Objectives:**

* Create an Automatic Learning Model to Classify Patients with Liver Diseases

**Discussion**:  
The primary objective was successfully achieved by creating an automatic learning model using machine learning algorithms such as Logistic Regression, Random Forest, and K-Nearest Neighbours (KNN). These algorithms were trained on a dataset featuring biochemical parameters commonly found in blood tests.

**Analysis**:  
The model's performance was evaluated using metrics like accuracy, precision, and recall. Among the algorithms used, Random Forest demonstrated the highest accuracy, thereby confirming its robustness in classifying liver diseases based on biochemical markers. However, it is vital to acknowledge that machine learning models are only as good as the data they are trained on. The dataset's quality directly influenced the model's performance.

* Develop a Web Application Implementing the Created Model

**Discussion**:  
The secondary objective was to make the predictive model accessible to healthcare providers. This objective was achieved by integrating the model into a user-friendly web application.

**Analysis**:  
The web application allows healthcare providers to input biochemical markers and receive immediate predictions. The real-time response and user-friendly interface enhance the application's utility, thereby fulfilling the objective effectively.

**Specific Objectives:**

* Analyse the Data Set to Select the Best Predictors

**Discussion**:  
Various biochemical markers were analysed to identify the best predictors for liver diseases.

**Analysis**:  
Statistical tests like the Wilcoxon test were applied to identify significant variables. The analysis revealed that markers like bilirubin levels and enzyme levels are crucial predictors, aligning well with existing medical literature.

* Train, Evaluate, and Optimize Machine Learning Algorithms

**Discussion**:  
Multiple machine learning algorithms were trained and evaluated to find the best-performing one, aiming for a minimum accuracy rate of 70%.

**Analysis**:  
All algorithms exceeded the targeted 70% accuracy rate, with Random Forest scoring the highest. Hyperparameter tuning further optimized the model, thereby achieving the objective effectively.

* Create a Web Application for User-friendly Predictions and Result Visualization

**Discussion**:  
A sophisticated web application was developed to seamlessly integrate the most effective machine learning model for liver disease prediction. This web-based interface serves as the end-user application where healthcare providers or researchers can input the relevant biochemical parameters and receive immediate, machine-learning-based predictions about liver diseases.

**Analysis**:  
The web application underwent rigorous user testing to ensure its usability and effectiveness. These user tests were conducted with a variety of participants, including medical practitioners, data scientists, and laypeople interested in healthcare. Participants were asked to navigate through the application, use its features, and provide their feedback on the usability, interface, and overall experience.

The results were overwhelmingly positive: users found the application easy to navigate and appreciated the intuitive design that guided them through each step of the process. Medical professionals, in particular, noted the value of visualizations, such as graphs and heatmaps, which aided them in interpreting the results more effectively. These visual elements were not just decorative but served to simplify complex data into understandable insights, thus fulfilling the objective of creating a user-friendly and utilitarian tool.

The term "user-friendliness" was measured using specific metrics like the time taken to complete tasks, the number of clicks needed to reach specific sections, and qualitative feedback about the layout and design. These measurements helped in confirming that the application not only met but often exceeded the standards of usability.

Additionally, the web application featured a secure login mechanism and data encryption to address data privacy concerns, making it a holistic solution for liver disease prediction and research.

In essence, the user tests and subsequent feedback served as a validation of the web application's design and functionality, successfully meeting the project's objective for a user-friendly and effective tool for liver disease prediction.

**Research Questions:**

Q1) What are the most effective machine learning algorithms for early detection of liver diseases, and how do they compare to traditional diagnostic methods?

**Discussion and Analysis**:  
The study found that Random Forest outperformed other machine learning algorithms and traditional diagnostic methods in early detection. While traditional methods like ultrasound scans are valuable, they may not be as efficient in early-stage detection as machine learning models trained on biochemical markers. This provides a strong argument for the integration of machine learning into traditional diagnostic pathways for liver diseases.

Q2) What are the limitations and challenges of applying machine learning techniques for liver disease detection?

**Discussion and Analysis**:  
While machine learning shows promise, it also has limitations. One significant challenge is the quality and comprehensiveness of the data used for training. Additionally, ethical considerations around data privacy and the "black-box" nature of some algorithms cannot be ignored. Future research should focus on these limitations to make machine learning a more reliable tool in liver disease diagnostics.

## 6.CONCLUSION

### 6.1. CONCLUSION

With this research on early liver disease detection, which successfully carried out the calculation analysis, using various models like logistic regression, KNN, CART and random forest. However, the KNN algorithm to be analysed the results of KNN using a value of K=8 intake the nearest neighbour with a ranking result with K = 2, the nearest neighbour takes in the calculation KNN. The output results of the KNN algorithm classification will be analysed by the algorithm and 71.30% accuracy is achieved. Consequently, Logistic Regression with dependent and independent variables results in the test data accuracy value of 75.65%. Thereafter, CART determining the affected patients with liver disease and number of predictions achieved 68.70% of accuracy. Subsequently, the Random Forest achieved 72.17% of accuracy collectively.

### 6.2. RECOMMENDATIONS

For further research, we propose the developments made in the following research:

1. Predict the data set using Simple Linear Regression with different algorithms outside of the methods written by authors such as Naive Bayes, Bayesian Network, and Multiple Regression Liner.

2. Looking for case studies with different data sets with more attributes many and various classifications.

### 6.3.FUTURE SCOPE

#### 6.3.1.Expand the Dataset

Incorporate More Variables: Current models could benefit from additional variables like genetic predispositions, lifestyle factors, or even geographic variables to provide a more comprehensive analysis.

Time-Series Analysis: Introducing time-based data could enable the study of the disease's progression over time, enhancing predictive accuracy.

#### 6.3.2.Improve Model Robustness

Ensemble Techniques: Combining the predictions of the currently used models with other algorithms could improve the overall predictive power and robustness of the model.

Deep Learning Approaches: Future research could explore the application of neural networks and other deep learning techniques for more nuanced pattern recognition.

#### 6.3.3.Real-World Applications

Integration with Electronic Health Records: The predictive models could be integrated into existing healthcare IT systems to provide real-time risk assessments for patients.

Mobile Health Apps: A simplified version of the model could be developed for mobile health applications that help individuals self-assess their liver health.

#### 6.3.4.Interpretability and Ethics

Explainable AI: As machine learning models are often considered "black boxes," future work could focus on making the models more interpretable to healthcare professionals.

Ethical Considerations: Future iterations should also explore the ethical implications of predictive healthcare models, including data privacy and fairness in predictions.

#### 6.3.5.Policy and Awareness

Public Health Policies: The findings could be used to inform public health policies, particularly in areas with high incidences of liver diseases.

Educational Programs: Based on the variables that are found to be significant, awareness and educational programs could be designed to inform the public about liver health.

#### 6.3.6.Interdisciplinary Collaboration

Healthcare and Data Science: This project sits at the intersection of healthcare and data science, and future research could benefit from collaborative work between experts in these fields.

Health Studies: Given the global prevalence of liver diseases, international collaborative studies could offer more generalizable insights.

# Chapter 3: Reflections

## 1.Introduction

### 1.1.What is Reflection?

Reflection is not just an afterthought or a casual review of past events; it's a complex, multi-layered process aimed at gaining a deeper understanding of both actions and reactions. At its core, reflection involves a form of mental processing that is used to fulfil a purpose or to achieve some anticipated outcome [53]. It is an active cognitive process where individuals dissect their experiences, exploring not just the 'what' but also the 'why' and 'how.' This entails examining not only the actions taken but also the emotional and psychological underpinnings of those actions.

The concept of reflection extends beyond personal introspection to include a broader social, cultural, and professional context [55]. It acts as a vital bridge between theoretical learning and practical application, allowing for a two-way flow of knowledge and experience. Through reflection, individuals can apply theoretical principles to real-world situations and enrich academic or professional theories with practical insights [54].It serves as a lens through which we can view our actions and experiences critically, helping us adapt and grow in our personal and professional lives. This makes reflection an indispensable tool for continuous learning and development, transforming what could be passive experiences into active learning opportunities.

### 1.2.The Importance of Reflection

The importance of reflection cannot be overstated, especially in academic and professional settings. At the most basic level, reflection is crucial for learning. It enables us to move from surface learning to deep learning—a transition from simply memorizing facts to understanding underlying concepts and making connections between different pieces of information [57]. Reflection helps in the assimilation and accommodation of new information, making it an essential component in the learning cycle [55].

Beyond learning, reflection serves as a powerful tool for personal and professional development. It encourages self-awareness by prompting us to examine our thought processes, values, and behaviours. This heightened self-awareness can lead to better decision-making and improved emotional intelligence, qualities that are highly valuable in any profession [59]. In the academic context, reflection allows us to evaluate our research methodologies critically, assess the validity of our data and findings, and think about the ethical implications of our work [56].

Reflection is also instrumental in fostering a culture of continuous improvement. By regularly taking time to step back and assess our actions and decisions, we can identify areas of strength and weakness. This self-assessment leads to actionable insights that can be used to improve future performance, whether it's in conducting research, working on a team project, or interacting with clients in a professional setting [53].  
  
1.3. Why Am I Writing This Reflection and Reason to Choose the Gibbs Reflective Model

The act of writing this reflection serves multiple purposes, both academic and personal. From an academic standpoint, the reflection allows for a structured evaluation of the research methodologies employed, the challenges encountered, and the solutions devised over the course of this dissertation. It facilitates a deep understanding of the complex processes involved in conducting research on liver health and Hepatitis B. This reflective exercise is not only an opportunity for academic introspection but also a tool for professional growth. It enables me to articulate the skills, knowledge, and competencies I have gained, thereby enriching my portfolio for future academic and professional endeavours.

Choosing the Gibbs Reflective Model to guide this reflection was a conscious decision influenced by the model's comprehensive and cyclical nature. Developed by Graham Gibbs in 1988, this model offers a six-stage framework—Description, Feelings, Evaluation, Analysis, Conclusion, and Action Plan—that encourages a thorough review of the event or experience in question. What sets the Gibbs Model apart is its emphasis on feelings, which are often overlooked in academic reflections but are crucial for understanding the emotional investment and the stressors involved in a research project. Its structured approach ensures that no aspect of the experience, whether it's emotional, cognitive, or practical, is neglected. The cyclical nature of the model also allows for continuous improvement, making it particularly suited for complex and ongoing projects like dissertations.  
  
  
  
1.4. Discussion: Gibbs Reflective Model

The Gibbs Reflective Model, developed by Graham Gibbs in 1988, serves as the framework for this reflection. Widely applied across fields like healthcare, education, and management, the model offers a structured six-stage cycle for reflective practice [58]. The cycle begins with the 'Description' stage, where a detailed account of the experience sets the context for reflection. This is followed by the 'Feelings' stage, which delves into the emotional landscape experienced during the event, a component often highlighted as crucial for professional effectiveness [59]. Next, the 'Evaluation' stage involves assessing the positives and negatives of the actions and decisions made. This leads to the 'Analysis' stage, where the individual critically dissects the experience to understand the influencing factors and assumptions. The penultimate 'Conclusion' stage synthesizes the insights gained, paving the way for the final 'Action Plan' stage that focuses on strategizing future actions based on the reflections. The cyclical and comprehensive nature of the Gibbs Reflective Model makes it particularly suitable for in-depth reflections on complex endeavours like dissertations. Its emphasis on both emotional and practical aspects ensures a holistic understanding and fosters a culture of continuous improvement.

## 2. Reflection

2.1. Description Stage  
  
The journey of my dissertation on liver disease took an interesting turn when I realized that collecting primary data was not the most viable option. The switch to secondary data was fraught with challenges and opportunities. My quest for the perfect dataset led me through academic databases and online repositories, each brimming with potential but lacking the specificity my research required.

My eureka moment arrived when I discovered the "Indian Liver Patient" dataset on Kaggle, a perfect match for my research objectives. This dataset wasn't merely numbers; it was the linchpin of my entire dissertation. However, the real challenge started after downloading it. Ensuring accurate interpretation required days of immersion, cross-referencing variables and values against existing medical literature.

Consultations with my academic Supervisor, Professor Farzad, were invaluable. These weren't mere progress checks; they were brainstorming sessions that expanded my understanding and refined my approach, especially when it came to implementing machine learning algorithms like Logistic Regression, CART, Random Forest, and KNN. The challenge wasn't just selecting an algorithm; it was about understanding its limitations and nuances. For instance, while Logistic Regression offered simplicity, it had constraints in handling complex relationships in the data. CART was intuitive but prone to overfitting, requiring meticulous fine-tuning.

The emotional landscape during this phase was a mix of excitement and apprehension. As I started implementing the algorithms, new challenges emerged. Random Forest, although an improvement over CART, required careful hyperparameter tuning to prevent overfitting. KNN presented its own set of challenges, such as choosing the right number of neighbours and distance metrics. Each algorithmic implementation felt like a journey in itself, full of minor victories and setbacks, ultimately contributing to the final conclusions and recommendations of my dissertation.

Reflecting on the journey, each hurdle—whether it was selecting the right dataset or choosing the appropriate machine learning algorithm—enriched my understanding and honed my research skills. My experience taught me not just about liver disease and machine learning techniques, but also about the rigor and nuance involved in academic research. It revealed the critical role of data integrity, the complexity of algorithmic implementation, and the importance of thorough, reflective analysis.

2.2. Feelings Stage  
  
Navigating the world of secondary data for my research on liver disease was an emotionally layered experience. Initially, I felt a wave of relief when I decided to opt for secondary data. The decision eliminated the complexities of primary data collection, aligning well with my objective to create an automatic learning model based on biochemical parameters. I knew that a good dataset would be the cornerstone for achieving this aim.

However, the initial relief soon turned into a maze of stress and uncertainty as I delved into various repositories. I was not just looking for any data; I needed data that could help me analyse biochemical predictors effectively. The weight of my specific objectives seemed to magnify with each dataset I explored but found lacking. When I finally discovered the Kaggle dataset that perfectly aligned with my goals, including model training and optimization, the sense of exhilaration was indescribable. It felt as though a significant hurdle in achieving my objectives had been overcome.

Despite this, elation gave way to cautious optimism and trepidation. Each variable in the dataset represented a potential predictor, and each value could influence the machine learning algorithms I planned to employ. A constant undercurrent of doubt lingered: Had I interpreted the data correctly? Would it fulfil my objective to determine the best prediction model with a minimum accuracy of 70%?

Interactions with my academic advisor served as emotional anchors. These weren't just routine consultations; they were critical to the development of the web application outlined in my objectives. Positive feedback was a morale booster, affirming that I was on the right track. Constructive criticism, although initially stinging, became a catalyst, pushing me to refine the web application's user experience further.

As I entered the data analysis phase, the emotions were bittersweet. On one hand, each finding and insight brought me closer to achieving my objectives; on the other, the looming deadline served as a constant reminder of the work that remained. Yet, as I inched towards the project's conclusion, the sense of accomplishment began to take centre stage, overshadowing earlier apprehensions. I knew I was close to not just completing a dissertation but also contributing a tangible tool that could have real-world applications in liver disease research.  
  
2.3. Evaluations Stage  
  
The "Evaluation" stage of my dissertation journey is a point for critical assessment, offering a chance to review the different components of my research on liver disease. One of the significant triumphs was finding the "Indian Liver Patient" dataset on Kaggle. This dataset was not merely a fit; it was a catalyst that allowed me to excel in my objectives. It met the rigorous criteria for selecting the best biochemical predictors, aligned perfectly with my goal to create an automatic learning model, and formed the backbone of my subsequent web application.

Another high point was the continuous and robust support from my academic supervisor. This guidance was invaluable, particularly when aligning the research methodology with the specific objectives, such as training, evaluating, and optimizing machine learning algorithms like Logistic Regression, CART, and Random Forest. Each meeting served multiple functions—it was a space for problem-solving, for fine-tuning methodologies, and for boosting my confidence. The academic advice I received played a crucial role in achieving the objectives for the prediction model.

However, the journey was not smooth sailing all the way. A considerable challenge was the initial period of searching through various data sources before discovering the "Indian Liver Patient" dataset. While this dataset eventually proved to be ideal, the process of getting there was fraught with uncertainty. Each dataset I considered seemed to lack one element or another, be it completeness, relevance, or credibility, adding stress to the already demanding project.

The data analysis phase, too, presented its complexities. The "Indian Liver Patient" dataset was thorough but also rich in detail, posing initial feelings of being overwhelmed. The responsibility of interpreting the data correctly was a constant pressure, more so with the nearing dissertation deadline. The excitement of generating useful insights was constantly shadowed by the fear of making a potentially significant error in interpretation.

2.4. Analysis Stage  
The "Analysis" stage of my dissertation journey was a profound opportunity for critical reflection, especially in light of my specific objectives. Choosing the "Indian Liver Patient" dataset from Kaggle was a double-edged sword. On the one hand, it dovetailed seamlessly with my objective of creating an automatic learning model to classify liver disease based on biochemical parameters. On the other hand, it imposed limitations on the comprehensiveness and quality of the data, thereby emphasizing the critical role of data discernment.

One of the key challenges was the implementation of machine learning algorithms—Logistic Regression, CART, Random Forest, and KNN—each with its unique set of complexities. For instance, Logistic Regression, although straightforward, required careful feature selection to ensure the model's interpretability and performance. CART and Random Forest required intricate hyperparameter tuning to mitigate overfitting, aligning with my objective of optimizing machine learning algorithms. KNN presented challenges in feature scaling and distance metric selection, which became a lesson in the need for meticulous pre-processing.

Time management was another focal point of my analysis. The complexity of training, evaluating, and optimizing these machine learning algorithms meant that the time allocated for each task far exceeded my initial estimates. While this misalignment was stressful, it was also a revelation. The time-intensive nature of achieving a minimum 70% accuracy in the predictive models was a hurdle but also an essential aspect of fulfilling my research objectives.

The role of my academic advisor in this context was multifaceted. While the guidance was indispensable in navigating algorithmic complexities and achieving the targeted accuracy, it also triggered reflections on academic autonomy. To what extent did my advisor's input shape my algorithmic choices? This pondering led me down a path of questioning the dynamics of mentor-mentee relationships in academic settings, a valuable insight for future projects.

Emotionally, this stage of the dissertation was a rollercoaster, directly linked to the depth of my engagement with my objectives. The emotional highs and lows were not merely transient; they marked significant milestones in my commitment to the research. The elation of nearing the 70% accuracy goal was balanced by the stress of finetuning machine learning models to optimize user experience in the web application, each reflecting the project's overarching complexities and rewards.

2.5. Conclusion Stage  
  
The "Conclusion" stage of my dissertation journey functions as a crucible, fusing disparate experiences, decisions, and emotions into a coherent narrative that aligns remarkably well with my specific objectives. Among the most pivotal learnings was the indispensable role of data—in particular, the "Indian Liver Patient" dataset from Kaggle. This dataset was not just a foundation but the very bedrock on which my objectives were met—from creating an automatic learning model based on biochemical parameters to comparing and optimizing machine learning algorithms like Logistic Regression, CART, Random Forest, and KNN.

Implementing these algorithms was a journey in itself, each with its unique set of challenges and learnings. Logistic Regression, while straightforward, posed questions about feature importance and model interpretability. CART and Random Forest demanded careful consideration for avoiding overfitting, especially crucial given my objective to achieve a minimum accuracy of 70%. KNN forced me to grapple with the importance of feature scaling and distance metrics, emphasizing the granular aspects of machine learning that often go unnoticed but are critical for achieving specific objectives.

Time management took on a new level of importance as I navigated these complexities. The dissertation was a lesson in the art of adaptability, teaching me to recalibrate timelines in the face of unexpected challenges—be it in data cleaning, algorithm optimization, or web application development. This experience highlighted the necessity for robust contingency planning, a principle applicable beyond academia.

The role of my academic supervisor became increasingly significant in light of these complexities and my overarching objectives. The guidance received shaped not just the conceptual aspects of my project but also the technical ones, like algorithm selection and tuning. This interaction served as both a support system and a challenge, pushing me to refine my methodologies and question the autonomy of my academic choices.

Emotionally, the journey was a revelation. The fluctuating emotions were not mere side effects but were deeply entwined with my project's milestones. The highs of achieving desired accuracies with the algorithms and the lows of debugging the web application were both reflections of my commitment to the research. These emotional markers have equipped me with the awareness to better manage my emotional landscape in future endeavours, serving as both motivational catalysts and cautionary indicators.

2.6. Action plan stage  
  
The "Action Plan" stage serves as a blueprint for my future endeavours, enriched by the invaluable lessons and specific objectives of my dissertation journey on liver disease. One immediate focal point for further development is the enhancement of my data evaluation skills. Given my objective to consistently analyse datasets to select optimal predictors, my experience with the "Indian Liver Patient" dataset from Kaggle, while largely successful, also exposed areas of uncertainty, particularly around feature selection for algorithms like Logistic Regression, CART, and Random Forest. To bolster this skill set, I'm considering enrolling in advanced courses in data analytics and workshops specializing in machine learning techniques. The aim is to ensure the more precise implementation of algorithms in future projects, thereby minimizing uncertainties.

Time management is another crucial area warranting attention. The complexities of optimizing machine learning algorithms—each with its unique challenges and learning curves—made it evident that even a meticulously planned schedule can unravel. To enhance my time management skills and better align with specific objectives like algorithm optimization, I intend to employ robust project management tools. These tools will not only track milestones and deadlines but also integrate contingency plans to adapt to unforeseen complexities and challenges.

The mentor-mentee relationship, pivotal in achieving my dissertation objectives, will be fine-tuned for future projects. To maintain a balanced dynamic where guidance complements rather than overshadows my academic autonomy, clear boundaries and expectations will be established from the onset. Pre-project discussions will become a standard practice to ensure that this relationship remains a constructive force in achieving specific objectives.

Emotionally, the varied emotional states I experienced were not isolated incidents but were directly connected to my progress and setbacks related to each objective. This realization has led me to consider incorporating stress management techniques such as mindfulness and meditation into my routine. These practices aim to balance emotional investment in my work, reducing the risk of burnout while enhancing productivity and focus.

Finally, the dissertation has strengthened my commitment to liver disease research, aligning perfectly with my objective of developing a practical and impactful model. Whether it's through conducting follow-up studies, expanding on the existing model with more features or algorithms, publishing articles, or forming collaborations with healthcare organizations, this focus will remain central to my academic and possibly professional journey.  
  
2.7. Comparison: Expectations vs. Reality

When I first commenced my dissertation journey on liver disease, my expectations were not only influenced by academic aspirations but also tightly linked to my specific objectives. I anticipated challenges but believed that a well-structured plan, focused on objectives like creating an automatic learning model and developing different models, would mitigate most issues. The reality unfolded as a tapestry of both expected and unexpected experiences, each contributing uniquely to the dissertation process.

#### 2.7.1. Dataset Selection

One of the most gratifying outcomes was the selection of the "Indian Liver Patient" dataset from Kaggle. Initially, I had reservations about using secondary data, particularly concerning its suitability for creating an effective learning model. However, this dataset not only expedited my research but also exceeded my expectations. It was comprehensive, reliable, and highly relevant, perfectly aligning with my objective to select the best predictors and train various machine learning algorithms.

#### 2.7.2. Time Management

Time management was an area where my expectations were both met and challenged. While I had allocated time for tasks like training, evaluation, and optimization of machine learning algorithms, the actual hours required exceeded my initial estimates. This experience was instrumental in teaching me the importance of 'buffer time,' especially when working on complex tasks that align with specific objectives.

#### 2.7.3 Academic Guidance

The mentor-mentee relationship also exceeded my initial expectations. The guidance I received played a critical role in achieving specific objectives, such as determining the best machine learning model with a minimum accuracy of 70%. The relationship was not only a problem-solving forum but also a space that enriched the quality of my research and decision-making process.

#### 2.7.4. Emotional Landscape

Emotionally, I had anticipated a challenging yet linear journey. However, the emotional spectrum I experienced was broad, with highs and lows directly tied to progress or setbacks in achieving my objectives. This emotional rollercoaster, unexpected as it was, emphasized my deep commitment and passion for the research topic.

#### 2.7.5. Research Impact

Perhaps the most fulfilling outcome was the tangible impact of my research, which directly aligns with my objectives. The insights garnered could significantly influence the field of liver disease, offering avenues for future research and potential real-world applications. This outcome surpassed my initial expectations, adding a layer of both personal and academic pride to the accomplishments.  
  
  
3. Conclusion

Reflecting on my dissertation journey focused on liver disease has been a transformative experience, magnified through the lens of the Gibbs Reflective Model. This structured framework allowed for a meticulous exploration of each phase of my research, from the initial dataset selection to the emotional aspects, all the while closely aligned with my specific objectives, such as creating an automatic learning model and developing a web application.

The positive outcomes have been numerous and directly connected to my objectives. The discovery of the "Indian Liver Patient" dataset on Kaggle was pivotal, enabling me to achieve my goal of selecting the best predictors and training effective machine learning algorithms. The invaluable guidance from my academic advisor played a crucial role in achieving other specific objectives, such as determining the best prediction model with a minimum accuracy of 70%. Moreover, the complexities of time management, a universal challenge in any research project, were navigated more effectively due to the lessons learned during this journey.

This reflective process has been invaluable for crystallizing the lessons learned and offering actionable insights, not only for academic projects but also for broader professional endeavours. It has highlighted the importance of skills like adaptability, which was crucial in managing the timeline for developing a web application, and mentorship, which played a significant role in the research's success. Emotional resilience, another important lesson, is an asset that will have a lasting impact on both my academic and professional future.

In summary, the dissertation journey has been a rich tapestry of challenges and triumphs, each thread intricately linked to my objectives and contributing to my growth as a researcher. This reflection serves as both a retrospective analysis and a forward-looking blueprint for future academic and professional projects in the realm of liver disease and beyond. It has been an exercise in self-awareness, a catalyst for continuous improvement, and a milestone in my lifelong learning journey.

## 

## REFERENCES

1. Planas, R., & Salmeron, J. (2007).Liver Diseases. Practical tips. Permanenter Publications.
2. Akter, S., Shekhar, HU, & Akhteruzzaman, S. (2021). Application of Biochemical Tests and Machine Learning Techniques to Diagnose and Evaluate Liver Disease. Advances in Bioscience and Biotechnology,12(06), 154-172. https://doi.org/10.4236/abb.2021.126011
3. Kuzhippallil, MA, Joseph, C., & Kannan, A. (2020). Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients.2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020, 778-782. https://doi.org/10.1109/ICACCS48705.2020.9074368
4. Cainelli, F. (2012). Liver diseases in developing countries.World Journal of Hepatology, 4 (3), 66-67. https://doi.org/10.4254/wjh.v4.i3.66
5. Fernández Daza, E., Fernández Juan, E., Moreno Mejía, I., & Moreno Mejía, M. (2008). Approach to the diagnosis of liver diseases by the clinical laboratory. Medicine & Laboratory,14(11-12), 533-546.
6. Venkata Ramana, B., Babu, MSP, & Venkateswarlu, N. . (2011). A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. International Journal of Database Management Systems,3(2),101-114. https://doi.org/10.5121/ijdms.2011.3207
7. Hashem, EM, & Mabrouk, MS (2014). A Study of support vector machine algorithm for liver disease diagnosis.American Journal of Intelligent Systems,4(1), 9-14.
8. Asrani, S.K., Devarbhavi, H., Eaton, J., & Kamath, PS (2019). Burden of liver diseases in the world.Journal of Hepatology,70(1), 151-171. https://doi.org/10.1016/j.jhep.2018.09.014
9. Moon, AM, Singal, AG, & Tapper, EB (2020). Contemporary Epidemiology of Chronic Liver Disease and Cirrhosis.Clinical Gastroenterology and Hepatology, 18(12), 2650-2666. https://doi.org/10.1016/j.cgh.2019.07.060
10. Seitz, HK, Bataller, R., Cortez-Pinto, H., Gao, B., Gual, A., Lackner, C., Mathurin, P., Mueller, S., Szabo, G., & Tsukamoto , H. (2018). Alcoholic liver disease.Nature Reviews Disease Primers,4(1). https://doi.org/10.1038/s41572-018-0014-7
11. Jepsen, P., Grønbæk, L., & Vilstrup, H. (2015). Worldwide Incidence of Autoimmune Liver Disease.Digestive Diseases,33(suppl 2), 2-12. https://doi.org/10.1159/000440705
12. Mitra, S., De, A., & Chowdhury, A. (2020). Epidemiology of non-alcoholic and alcoholic fatty liver diseases.Translational Gastroenterology and Hepatology,5, 1-17. https://doi.org/10.21037/TGH.2019.09.08
13. Moreno Borque, A., González Moreno, L., Mendoza-Jiménez, J., García-Buey, L., & Moreno Otero, R. (2007). Usefulness of analytical parameters in the diagnosis of liver diseases.Annals of Internal Medicine,24(1), 38-46. https:// doi.org/10.4321/s0212-71992007000100010
14. Castro, C., Benítez-Mata, B., Vargas, E., & Flores, L. (2020). Early detection of the disease caused by the Hepatitis C virus using artificial neural networks. National Congress of Biomedical Engineering, 7. http://memorias.somib.org.mx/index.php/memorias/article/view/791/467
15. Takkar, S., Singh, A., & Pandey, B. (2017). Application of machine learning algorithms to a well-defined clinical problem: Liver disease.International Journal of E-Health and Medical Communications,8(4),38-60. https://doi.org/10.4018/IJEHMC.2017100103
16. Shaheamlung, G., Kaur, H., & Kaur, M. (2020). A Survey on machine learning techniques for the diagnosis of liver disease.Proceedings of International Conference on Intelligent Engineering and Management, ICIEM 2020, 337-341. https://doi.org/10.1109/ICIEM48762.2020.9160097
17. Atabaki-Pasdar, N., Ohlsson, M., Viñuela, A., Frau, F., Pomares-Millan, H., Haid, M., Jones, AG, Louise Thomas, E., Koivula, RW, Kurbasic, A., Mutie, PM, Fitipaldi, H., Fernandez, J., Dawed, AY, Giordano, GN, Forgie, IM, McDonald, TJ, Rutters, F., Cederberg, H., … Franks, PW (2020). Predicting and elucidating the etiology of fatty liver disease: A machine learning modeling and validation study in the IMI DIRECT cohorts.PLOS Medicine,17(6), 1-27. https://doi.org/10.1371/journal.pmed.1003149
18. Katsiki, N., Gastaldelli, A., & Mikhailidis, DP (2019). Predictive models with the use of omics and supervised machine learning to diagnose non-alcoholic fatty liver disease: A “non-invasive alternative” to liver biopsy?Metabolism: Clinical and Experimental,101(xxxx). https://doi.org/10.1016/j.metabol.2019.154010
19. Younossi, ZM, Stepanova, M., Younossi, Y., Golabi, P., Mishra, A., Rafiq, N., & Henry, L. (2020). Epidemiology of chronic liver diseases in the USA in the past three decades.gut,69(3), 564-568. https://doi.org/10.1136/gutjnl-2019-318813
20. Lin, YJ, Chen, RJ, Tang, JH, Yu, CS, Wu, JL, Chen, LC, & Chang, SS (2020). Machinelearning monitoring system for predicting mortality among patients with noncancer end-stage liver disease: Retrospective study.JMIR Medical Informatics, 8(10), 1-16. https://doi.org/10.2196/24305
21. Spann, A., Yasodhara, A., Kang, J., Watt, K., Wang, B., Goldenberg, A., & Bhat, M. (2020). Applying Machine Learning in Liver Disease and Transplantation: A Comprehensive Review.Hepatology,71(3),1093-1105. https://doi.org/10.1002/hep.31103
22. Fatima, M., & Pasha, M. (2017). Survey of Machine Learning Algorithms for Disease Diagnostics.Journal of Intelligent Learning Systems and Applications,09(01), 1-16. https://doi.org/10.4236/jilsa.2017.91001
23. Ma, H., Xu, CF, Shen, Z., Yu, CH, & Li, YM (2018). Application of Machine Learning Techniques for Clinical Predictive Modeling: A Cross-Sectional Study on Nonalcoholic Fatty Liver Disease in China.Biomed Research International,2018. https://doi.org/10.1155/2018/4304376
24. Lantz, B. (2015). Machine Learning with R.InMachine Learning with R(Second Ed). Packt Publishing. https://doi.org/10.1007/978-981-10-6808-9 41
25. Gatos, I., Tsantis, S., Spiliopoulos, S., Karnabatidis, D., Theotokas, I., Zoumpoulis, P., Loupas, T., Hazle, JD, & Kagadis, GC (2017). A Machine-Learning Algorithm Toward Color Analysis for Chronic Liver Disease Classification, Employing Ultrasound Shear Wave Elastography.Ultrasound in Medicine and Biology,43(9), 1797-1810. https://doi.org/10.1016/j.ultrasmedbio.2017.05.002
26. Gogi, VJ, & Vijayalakshmi, MM (2018). Prognosis of Liver Disease: Using Machine Learning Algorithms.2018 International Conference on Recent Innovations in Electrical, Electronics and Communication Engineering, ICRIEECE 2018, 875-879. https://doi.org/10.1109/ICRIEECE44171.2018.9008482
27. Hamid, K., Asif, A., Abbasi, W., Sabih, D., & Minhas, FUAA (2017). Machine Learning with Abstention for Automated Liver Disease Diagnosis.Proceedings - 2017 International Conference on Frontiers of Information Technology, FIT 2017, 2017- Janua, 356-361. https://doi.org/10.1109/FIT.2017.00070
28. Alice Help, L. (2018). Accuracy Prediction using Machine Learning Techniques for Indian Patient Liver Disease.2nd International Conference on Trends in Electronics and Informatics (ICOEI), 45-50.
29. K. Phasinam, A. K. Singh, T. Singh, and M. K. Sharma, Fundamental of Machine Learning. Blue Rose Publishers, 2022.
30. S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.
31. Zhou, Zhi-Hua. (2021). Decision Trees. 10.1007/978-981-15-1967-3\_4.
32. Jo, Taeho. (2021). Decision Tree. 10.1007/978-3-030-65900-4\_7.
33. Suzuki, Joe. (2020). Decision Trees. 10.1007/978-981-15-7568-6\_8.
34. Sundiman, Didi. (2022). Decision tree regression.
35. Calhoun, Peter & Su, Xiaogang & Spoon, Kelly & Levine, Richard & Fan, Juanjuan. (2021). Random Forest. 10.1002/9781118445112.stat08287.
36. Sundiman, Didi. (2022). Random Forest Regression.
37. Berk, Richard. (2020). Random Forests. 10.1007/978-3-030-40189-4\_5.
38. Garcia, Guilherme. (2021). Logistic Regression. 10.4324/9781003032243-10.
39. Roback, Paul & Legler, Julie. (2021). Logistic Regression. 10.1201/9780429066665-6.
40. Zhou, Hong. (2020). Logistic Regression. 10.1007/978-1-4842-5982-5\_6.
41. Khastavaneh, Hassan & Ebrahimpour-Komleh, Hossein & Hanaee Ahwaz, Amin. (2017). Unknown aware k nearest neighbor classifier. 108-112. 10.1109/PRIA.2017.7983027.
42. Derrac, Joaquin & Chiclana, Francisco & García, Salvador & Herrera, Francisco. (2015). An Interval Valued K-Nearest Neighbors Classifier. 10.2991/ifsa-eusflat-15.2015.55.
43. Murti, Della & Pujianto, Utomo & Wibawa, Aji & Akbar, Muhammad. (2019). K-Nearest Neighbor (K-NN) based Missing Data Imputation. 83-88. 10.1109/ICSITech46713.2019.8987530.
44. Syriopoulos, Panos & Kotsiantis, Sotiris & Vrahatis, Michael. (2023). Survey on KNN Methods in Data Science. 10.1007/978-3-031-24866-5\_28.
45. Shahade, Makarand & Awate, Ashish & Nandwalkar, Bhushan & Kulkarni, Mayuri. (2023). Diabetes Disease Prediction Using KNN. 10.1007/978-981-99-0550-8\_24.
46. Karimi, Zohreh. (2021). Confusion Matrix.
47. Piegorsch, Walter. (2020). Confusion Matrix. 10.1002/9781118445112.stat08244.
48. Murphy, Andrew & Moore, Candace. (2019). Confusion matrix. 10.53347/rID-68080.
49. Martin, Greg. (2021). R Studio. 10.1007/978-3-030-69664-1\_1.
50. Tollefson, Margot. (2019). The R Prompt and the R Studio Windows. 10.1007/978-1-4842-4405-0\_2.
51. Cunliffe, Alex. (2021). Basic functionality of graphs and statistical analysis in R studio.
52. Hastie, T., Tibshirani, R., & Friedman, J. (2009). "The Elements of Statistical Learning". Springer.
53. Schön, D. A. (1983). The Reflective Practitioner: How Professionals Think in Action. Basic Books.
54. Boud, D., Keogh, R., & Walker, D. (1985). Reflection: Turning Experience into Learning. Kogan Page.
55. Kolb, D. A. (1984). Experiential Learning: Experience as the Source of Learning and Development. Prentice-Hall.
56. Moon, J. A. (1999). Reflection in Learning and Professional Development: Theory and Practice. Kogan Page.
57. Biggs, J. (1999). 'The importance of reflection in learning: From surface learning to deep learning
58. Gibbs, G. (1988). Learning by Doing: A Guide to Teaching and Learning Methods. Further Education Unit.
59. Goleman, D. (1995). 'Beyond learning: The role of reflection in personal and professional development