Text

Description automatically generated with medium confidence

**DS7010 – MSC DATA SCIENCE DISSERTATION**

**(M-LEVEL MODULE)**

**NAME:** Kalai Selvi Gandhi **STUDENT ID:** U2216077

**MODULE NAME:** Dissertation **MODULE CODE:** DS7010

**COURSE NAME:** MSC, Data Science (January 2022)

**TOPIC:** Predicting Life Expectancy for Liver Cirrhosis using Machine Learning Techniques.

**MODULE SUPERVISOR:** Mr. Joseph Annan.

**MODULE LEADER:** Dr. Yang Li.

**ACKNOWLEGEMENT**

I am grateful to everyone who assisted and supported me in completing this dissertation. I express my gratitude to the department of Data Science for the extended support. I would like to thank my Module Leader Dr. Yang Li to assign me a supervisor for my dissertation. My deepest thanks to my Module Supervisor Mr. Joseph Annan for guiding and correcting my work with attention and for his valuable guidance. I would like to thank my whole department for their extended support.

**ABSTRACT**

This dissertation looks at how different machine learning algorithms can be used to estimate life expectancy in individuals with liver cirrhosis. Using a dataset with 418 occurrences and 21 attributes, the study compares the efficiency of Linear Regression, Multi-Linear Regression, Random Forest, Decision Tree, Gaussian Naive Bayes, and Support Vector Machines (SVM) methods. When the accuracy of each model is evaluated, it is discovered that the Decision Tree algorithm has the best efficiency (87.73%), which is followed by the linear regression method (79.93%), Gaussian Naive Bayes algorithm (73.80%), SVM (72.39%), and the random forest algorithm (71.16%). The dissertation also uses Kaplan-Meier Survival analysis and the Cox-Hazard Proportional Method to determine the recovery time from liver cirrhosis and shows that the risk of liver cirrhosis reduces after 3000 days. The study emphasizes the predictive power of machine learning approaches and its consequences for public health and policy.

**TABLE OF CONTENTS:**

[CHAPTER-1: INTRODUCTION 7](#_Toc133395839)

[1.1: OVERVIEW 7](#_Toc133395840)

[1.2: BACKGROUND AND SIGNIFICANCE 8](#_Toc133395841)

[1.3: AIM AND OBJECTIVES 9](#_Toc133395842)

[1.3.1: AIM 9](#_Toc133395843)

[1.3.2: OBJECTIVES 10](#_Toc133395844)

[1.4: REQUIRED RESOURCES 11](#_Toc133395845)

[1.5: PREREQUISTIC KNOWLEDGE REQUIRED 11](#_Toc133395846)

[1.6: NATURE AND SOURCES OF THE DATASET 12](#_Toc133395847)

[1.7: ESTIMATED OUTPUT 12](#_Toc133395848)

[1.8: CONCLUSION 12](#_Toc133395849)

[CHAPTER 2: LITERATURE REVIEW 13](#_Toc133395850)

[2.1: INTRODUCTION 13](#_Toc133395851)

[2.2 LIVER CIRRHOSIS 13](#_Toc133395852)

[2.2.1: CAUSES OF LIVER CIRRHOSIS 15](#_Toc133395853)

[2.2.2: SYMPTOMS OF LIVER CIRRHOSIS 15](#_Toc133395854)

[2.2.3: DIAGNOSTICS OF LIVER CIRRHOSIS 16](#_Toc133395855)

[2.2.4: TREATMENT OF LIVER CIRRHOSIS 16](#_Toc133395856)

[2.3: LIFE EXPECTANCY 16](#_Toc133395857)

[2.3: MACHINE LEARNING 17](#_Toc133395858)

[2.3.1: TYPES OF MACHINE LEARNING 18](#_Toc133395859)

[2.3.2: PREDICTION USING MACHINE LEARNING TECHNIQUES 21](#_Toc133395860)

[2.3.3: How Machine Learning Works in Predicting Life Expectancy of Liver Cirrhosis? 22](#_Toc133395861)

[2.4: SYSTEMATIC REVIEW OF RELATED WORKS 23](#_Toc133395862)

[2.5: CONCLUSION 26](#_Toc133395863)

[CHAPTER 3: METHODOLOGY AND APPROACHES 27](#_Toc133395864)

[3.1: INTRODUCTION 27](#_Toc133395865)

[3.2: Data Set used 28](#_Toc133395866)

[3.3: Data Set Description 28](#_Toc133395867)

[3.3.1: Attributes and its Description 28](#_Toc133395868)

[3.4: Data Pre-Processing 29](#_Toc133395869)

[3.5: Techniques used 30](#_Toc133395870)

[3.6: CONCLUSION 39](#_Toc133395871)

[CHAPTER 4: PRACTICAL IMPLEMANTATION 40](#_Toc133395872)

[4.1: INTRODUCTION 40](#_Toc133395873)

[4.2: SECTION 1: DATA EXPLORATION 41](#_Toc133395874)

[4.2.1: Missingness Map 41](#_Toc133395875)

[4.2.2: Boxplot 41](#_Toc133395876)

[4.2.3: Normalization 41](#_Toc133395877)

[4.2.4: Visualization 42](#_Toc133395878)

[4.2.6: Correlation 43](#_Toc133395879)

[4.2.7: G-G Plot 43](#_Toc133395880)

[4.2.8: Scatterplot 43](#_Toc133395881)

[4.2.9: Q-Q Plot 43](#_Toc133395882)

[4.3: SECTION 2: SPLITTING TRAINING AND TESTING DATA 44](#_Toc133395883)

[4.4: SECTION 3: PREDICTION USING VARIOUS MACHINE LEARNING APPROACHES 44](#_Toc133395884)

[4.4.1: Linear Regression 45](#_Toc133395885)

[4.4.2: Multi-Linear Regression 45](#_Toc133395886)

[4.4.3: Random Forest Algorithm 46](#_Toc133395887)

[4.4.5: Gaussian Naïve Bayes Algorithm 46](#_Toc133395888)

[4.4.6: Support Vector Machines (SVM) 47](#_Toc133395889)

[4.5: SECTION 4: SURVIVAL ANALYSIS BY USING MACHINE LEARNING TECHNIQUES 47](#_Toc133395890)

[4.5.1: Kaplan Survival Curve 48](#_Toc133395891)

[4.5.2: Cox Proportional Hazard Model 48](#_Toc133395892)

[4.5.3: QQ-Plot for the survival model 48](#_Toc133395893)

[4.5.4: ROC curve 48](#_Toc133395894)

[4.6: CONCLUSION 48](#_Toc133395895)

[CHAPTER 5: RESULTS AND DISSCUSSION 50](#_Toc133395896)

[5.1: INTRODUCTION 50](#_Toc133395897)

[5.2: SECTION 1: DATA EXPLORATION 50](#_Toc133395898)

[5.2.1: Interpretation 50](#_Toc133395899)

[5.2.2: Results 50](#_Toc133395900)

[5.3: SECTION 2: SPLITTING TESTING AND TRAINING SET 55](#_Toc133395901)

[5.4: SECTION 3: PREDICTION USING MACHINE LEARNING TECHNIQUES 55](#_Toc133395902)

[5.3.1: Linear Regression 55](#_Toc133395903)

[5.4.2: Multi-Linear Regression 56](#_Toc133395904)

[5.4.3: Random-Forest Technique 57](#_Toc133395905)

[5.4.4: Decision-Tree Algorithm 58](#_Toc133395906)

[5.4.5: Gaussian Naïve Bayes Algorithm 59](#_Toc133395907)

[5.4.6: Support Vector Machine (SVM) 60](#_Toc133395908)

[5.5: SECTION 4: SURVIVAL ANALYSIS 60](#_Toc133395909)

[5.5.1: Kaplan-Meier Survival Test 60](#_Toc133395910)

[5.5.2: Cox-Hazard Proportional Approach 61](#_Toc133395911)

[5.5.3: QQ- PLOT FOR THE SURVIVAL MODEL 62](#_Toc133395912)

[5.5.4: ROC CURVE 62](#_Toc133395913)

[5.6: CONCLUSION 63](#_Toc133395914)

[CHAPTER 6: CRITICAL EVALUATION AND CONCLUSION 65](#_Toc133395915)

[REFERENCE 66](#_Toc133395916)

[Appendix 71](#_Toc133395917)

**PREDICTING LIFE EXPECTANCY FOR LIVER CIRRHOSIS USING MACHINE LEARNING TECHNIQUES.**

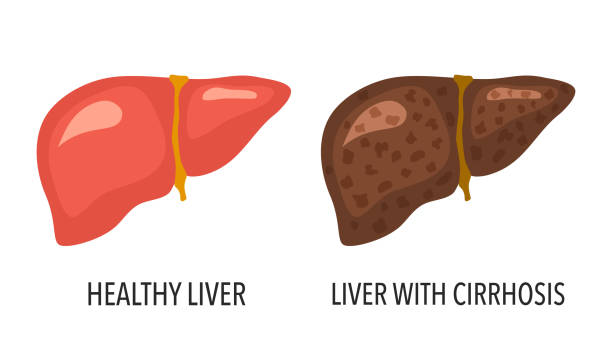
# **CHAPTER-1: INTRODUCTION:**

## **1.1: OVERVIEW:**

The liver is an enormous glandular tissue in the human abdomen that is necessary for many metabolic activities. The liver controls most of the chemical compounds in the bloodstream and excretes bile. This aids in the removal of waste materials from the liver. The liver filters all blood that leaves the stomach and intestines. It also stores sugar and fat as reserves energy sources, transforms hazardous compounds to less dangerous forms, and controls blood volume in the body. Liver is also responsible for Metabolism of proteins and carbohydrates, Enzyme activation and Synthesis of plasma protein such as albumin and clotting factors. There are several diseases such as Hepatitis B and C, alcohol consumption, obesity, and other factors can all harm the liver and result in Liver Cirrhosis. Cirrhosis is liver scarring (fibrosis) produced by long-term liver injury, in which normally functioning cells are substituted by scar tissues. The scar tissue obstructs the liver's ability to function correctly. Cirrhosis is often known as end-stage liver disease since it occurs after various stages of liver damage from illnesses such as hepatitis. In accordance with the global incidence of cirrhosis, 4.5% to 9.5% of the general population is affected with liver cirrhosis. In this Dissertation, the total amount of Bilirubin (a yellowish substance produced by the body's typical function of fragmenting  old red blood cells), Platelets (a small-disc shaped piece of cell found in the blood and spleen that aids in the process of blood clotting), and Thrombocytes (a the piece of enormous cells in the bone marrow known as megakaryocytes that initiate blood coagulation) retrieved from data sources of patients with liver cirrhosis can be identified in the Prediction of Life Expectancy of the patients by using Machine Learning Techniques and the outcomes is discussed accordingly for future research.

## **1.2: BACKGROUND AND SIGNIFICANCE:**

Chronic liver damage (CLD) is one of the main causes of death around the world, and it has been linked to an increase in the usage of health-care services. CLD can cause cirrhosis of the liver. Cirrhosis is an advanced stage of scarring (fibrosis) of the liver caused by a number of liver disorders and diseases, includes hepatitis and chronic drinking. Cirrhosis is scarring of the liver caused by chronic liver injury. Cirrhosis may not show symptoms at first. Later symptoms include feeling ill and weary, having yellow skin and eyes, itching skin, and a large stomach. Cirrhosis treatment typically requires resolving the underlying cause to keep the condition from deteriorating. Some people will need a transplant of the liver in the future. Cirrhosis is commonly caused by excessive alcohol consumption, a chronic infection of the liver that includes hepatitis B or C and being obese. The most severe type of alcohol-associated liver disease is cirrhosis. It usually happens after years of binge drinking. Cirrhosis is classified into four stages, as follows: Stage I: Steatosis, Stage II: Liver scarring (fibrosis) caused by inflammation, Stage III: Cirrhosis and Stage IV: Hepatic failure, liver failure, or advanced liver disease. (Reverter, Tugues and Martinez-Sanchez, 2015)



**Figure 1. 1: Healthy Liver and Liver with Cirrhosis (Health Line)**

Liver diseases are also classified as persistent infections. Many contaminants affect the liver, resulting in the various liver diseases. The death toll from liver diseases has reached 208185, accounting for 8.34% of all deaths in Egypt. The liver is the body's most grounded organ, and it is placed in the right part of the abdomen. High levels of glucose in the blood and decreased carbohydrate, fat, and protein metabolism result from insulin insufficiency. These organs and the liver work together to ingest and process meals. The liver cleanses the blood beginning with the stomach-related lot. Similarly, cleans synthetics and uses pharmaceuticals. Bile is covered by the liver organ and eventually collects the interior portions. In addition, the liver creates fundamental protein for coagulation of blood and is in charge of removing bacteria and foreign antigens from the circulation of the blood. Hepatic infirmity is suggested as a result of liver contamination. The word "liver contamination" is broad and includes all common problems that result in the liver acting outside of its prescribed parameters. This concentration concentrates on five different liver diseases: essential hepatoma (PM), alcoholic liver damage (ALD), liver cirrhosis (LC), HCC, and cholelithiasis (C). (D’Amico *et al.*, 2007)

Life expectancy is the average number of years an individual is expected to live according to statistics. One of the most significant organs in human beings is the liver. Hepatitis B, which causes liver inflammation, is one of the major disorders that disrupts the liver's activities. A widespread of infection in the liver is the major cause of Hepatitis B symptoms. Cirrhosis of the liver is a deadly disease which impacts people all over the world. Contact with contaminated bodily fluids is the most prevalent method for this disease to spread. Liver Cirrhosis symptoms include yellowing of the eyes, discomfort in the stomach, and black urine. Cirrhosis is a late stage of scarring (fibrosis) of the liver caused by a number of liver diseases and disorders, includes hepatitis and chronic drinking. Excessive alcohol use is detrimental to the liver. This may result to alcohol-related liver damage in the long run. (Oeppen and Vaupel, 2002)

Machine Learning (ML) is a method of allowing a machine to learn on its own by finding correlations and patterns in input data utilizing a variety of methods. This would allow for the planned, quick, and accurate projection of a case's life expectancy for any contention, with the two most crucial components being parameter selection and instruments used to assess these parameters. (Hastie, Tibshirani and Friedman, 2008).

Hence, the primary purpose of this dissertation is to estimate the life expectancy for liver cirrhosis utilizing a variety of machine learning methods which include Random Forest, Linear Regression, Decision tree models, and Support Vector Machines (SVM).

## **1.3: AIM AND OBJECTIVES:**

### **1.3.1: AIM:**

The aim of this dissertation is to Predict Life Expectancy for Liver Cirrhosis using Machine Learning Techniques.

### **1.3.2: OBJECTIVES:**

Cirrhosis of the liver is a fatal disease affecting people across the world. Various strategies have been employed by researchers to determine the life expectancy of Liver Cirrhosis patient. Several studies have employed Machine Learning algorithms to predict the life expectancy of Liver Cirrhosis patients, including the Classification model, Logistic Regression model, Recursive Feature Elimination Algorithm, Cirrhosis Mortality model, Random Forest, and Decision Tree. The main Objectives of this dissertation are as follows:

1. **Literature Review:**

A review of literature is a reliable and critical evaluation of the current level of understanding and investigation on a particular subject. It entails locating, evaluating, and synthesizing pertinent literature, which includes books, journal articles, and other scholarly sources. A literature review's purpose is to give a summary of the current state of knowledge on a subject, reveal gaps or inconsistencies in the literature, and to highlight areas for future study.

1. **Methodology:**

This chapter discuss the methodologies used in this dissertation such as the Regression models, Random Forest, Decision tree, Gaussian Naïve Bayes and Support Vector Machines (SVM) and discuss in detail.

1. **Practical Implementation:**

Liver Cirrhosis is a progressive condition of the liver that can cause severe complications and even death. Clinical, biochemical, imaging, and histopathological methods are all used to forecast the development of liver cirrhosis. Analyzing clinical, biochemical, imaging, and histopathological data, machine learning methods can be employed to forecast the likelihood of liver cirrhosis. These algorithms find connections and patterns in large sets of data by utilizing data mining as well as statistical modelling techniques.

1. **Results and Discussion:**

The outcome of this dissertation is executed using Machine learning Approaches and the accuracy of each technique is discussed for further future references.

1. **Critical Evaluation and Conclusion:**

It is the process of analyzing and assessing the relevance and reliability of the outcome and the experience gained throughout this dissertation and to conclude the main findings. It should also make suggestions for further study or activity based on the findings.

## **1.4: REQUIRED RESOURCES:**

The following resources may be required to predict life expectancy for liver disease using machine learning techniques:

**Machine Learning Algorithms:** Various machine learning methods, such as decision trees, random forests, linear regression, Gaussian Naïve Bayes and support vector machines (SVM) can be used. The algorithms can be applied to the dataset to identify patterns and predict liver disease life expectancy.

**Programming Languages and Tools:** Machine learning algorithms can be implemented using programming languages like Python, R, or Java. Machine learning models can also be built and trained using tools such as scikit-learn, and TensorFlow.

**Data Preprocessing and Extraction Of features:** Data preparation is the process of cleaning and transforming data in order to predict using machine learning.

**Metrics for Evaluating Machine Learning Model Performance:** To assess the effectiveness of machine learning models, many measures like as accuracy, recall, and F1 score can be utilized. These indicators can help measure the effectiveness of the model and indicate areas for development.

**Computational Resources:** Significant computer resources may be required depending on the dataset and the level of complexity of the machine learning methods.

## **1.5: PREREQUISTIC KNOWLEDGE REQUIRED:**

To predict life expectancy for liver cirrhosis with machine learning, needs first understand machine learning methods and then have a deep grasp of the disease and its impact on the human body. Two machine learning methods, deep learning and multiclass classification, can be used to analyze data and predict rates of life expectancy in cirrhotic patients. Furthermore, in order to correctly analyze the data and produce accurate forecasts, one must first comprehend the ailment and its consequence on the person's health.

## **1.6: NATURE AND SOURCES OF THE DATASET:**

The information comes from a Mayo Clinic trial on primary biliary (PBC) of the liver. Lichman, M. gathered the dataset utilized in this dissertation from Kaggle, where it was obtained from the UCI ML Repository. This Data Set consists of 418 instances and 21 Attributes such as age, stages, Drug, Sex, Ascites, and other factors such as Bilirubin, Edema, Albumin, Copper, SGOT, Platelets, Triglycerides, Prothrombin and lifeExpectancy.

## **1.7: ESTIMATED OUTPUT:**

According to studies and predictions, the number of individuals who are at risk significantly decreases after Several days. This dissertation will show that Decision Tree had the best accuracy (0.8507), and it excelled in every aspect of the metrics evaluation. Other strategies such as Regression Model (Linear and Multi-Linear Regression), Random Forest, Gaussian Naïve Bayes and Support Vector Machines (SVM) also worked well and had good accuracy levels, and to conclude, the Decision Tree will be the most accurate method for determining the Life Expectancy of a person who has liver cirrhosis. Additionally, the survival analysis is also done, and the risk is likely to be decreased after several days.

## **1.8: CONCLUSION:**

In conclusion, utilizing machine learning algorithms to predict the life expectancy for individuals with liver cirrhosis is an important area of research that has the potential to significantly improve patient outcomes. Machine learning approaches can help clinicians make accurate life expectancy projections, which can affect treatment decisions and improve patient care.

Overall, the implementation of reliable and comprehensible machine learning models for forecasting the life expectancy of liver cirrhosis patients has immense promise to enhance patient outcomes.

# **CHAPTER 2: LITERATURE REVIEW:**

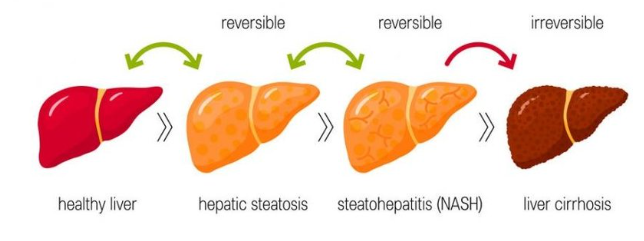
## **2.1: INTRODUCTION:**

A literature review is a critical analysis and overview of previous studies and literature on a specific area. The various purpose of the literature review is to determine and summarize pertinent empirical and theoretical studies on a specific subject, examine the current research's quality and trustworthiness. Also, to determine the field's study gaps and unanswered questions. Analyze and synthesize the main results and developments throughout the literature and the study's implications for future studies, policy, or practice.

A methodical search and analysis of educational databases, journals, books, as well as other relevant sources of information is typical of a literature review.

## **2.2 LIVER CIRRHOSIS:**

Liver diseases are also classified as persistent infections. Many contaminants affect the liver, resulting in the various liver diseases. The death toll from liver diseases has reached 208185, accounting for 8.34% of all deaths in Egypt. The liver is the body's most grounded organ, and it is placed in the right part of the abdomen. High levels of glucose in the blood and decreased carbohydrate, fat, and protein metabolism result from insulin insufficiency. These organs and the liver work together to ingest and process meals. The liver purifies the blood, starting with the stomach related. Similarly, cleans synthetics and uses pharmaceuticals. Bile is covered by the liver organ and eventually collects the interior portions. In addition, the liver creates fundamental protein for blood coagulation and is in charge of removing bacteria and external antigens from the circulation of blood. Hepatic infirmity is suggested as a result of liver contamination. The word "liver contamination" is broad and includes all common problems that result in the liver acting outside of its prescribed parameters. This concentration concentrates on five different liver diseases: essential hepatoma (PM), alcoholic liver damage (ALD), liver cirrhosis (LC), HCC, and cholelithiasis (C).



**Figure 2. 1: Stages of Liver Cirrhosis (Shutter Stock)**

Hence, the four stages of Liver Cirrhosis are shown in the Figure above. This clearly indicates that the Hepatic Steatosis (second stage) and Steatohepatitis (third stage) are reversible which can be diagnosed whereas the Liver Cirrhosis (end-stage of liver damage) is irreversible.

A chronic liver condition called liver cirrhosis causes liver cell damage and the accumulation of scar tissue. If untreated, it is a progressive, irreversible disease that can cause liver failure. The current knowledge of liver cirrhosis and its treatments will be examined in this literature review.

According to research by (Tsochatzis, Bosch and Burroughs, 2014) liver cirrhosis is a significant global contributor to mortality and morbidity. The research also emphasized how crucial early detection and therapy are for enhancing results. In 2018, the European Academy for the Study of the Liver (EASL) announced revised guidance for the evaluation and treatment of liver cirrhosis in separate studies. The recommendations stressed the value of ongoing observation, lifestyle changes, and prompt treatment of complications.

The many side effects of liver cirrhosis, such as ascites, hepatic encephalopathy, and variceal hemorrhage, were covered in a review paper by (P.S, 2021) The paper also covered pharmacological and non-pharmacological management strategies for these complications. In a randomized controlled study, conducted by (D’Amico *et al.*, 2007) the effectiveness of two medications sorafenib and sunitinib in the management of severe liver cirrhosis was compared. Sorafenibwas found to be more efficient in the research at lowering the likelihood of disease progression and enhancing survival. Advanced liver cirrhosis has been successfully treated with sorafenib, and research into the function of the gut microbiota in the illness is continuing. (Reverter, Tugues and Martinez-Sanchez, 2015) investigated the function of intestinal microbiota in the onset and development of liver cirrhosis in another study. According to the research, individuals with liver cirrhosis exhibit increased inflammatory conditions and liver damage as a result of changes in the composition of their gut microbiota**.** (Sebastian, Kuriakose and Varghese, 2016) observed that the practice of survival analysis is widely employed in the area of medical science. Both physicians and patients find great value and utility in the concept of being able to forecast a subject's life expectancy. The fundamental building blocks of any treatment paradigm are three initial stages. the stages of diagnosis, classification, evaluation, assessment, conclusion, and ultimately, treatment. All these aspects must clearly indicate the quantitative magnitude and severity of the sickness research in the context, as well as be accurate to specification and effective in their measurement. The Child-Pugh classification system is one of the most widely used classification methodologies for assessing liver illnesses, particularly cirrhosis. The extensive analysis of many cases demonstrates that distinct patients with various degrees and types of liver cirrhosis have varying life expectancies. For instance, fuzzy logic is a method that is suited to the situation like a glove.

Moreover, liver cirrhosis is a severe and complex condition that needs to be identified quickly and treated appropriately. Regular observation, lifestyle changes, and prompt intervention can help to better results and avoid complications.

## **2.2.1: CAUSES OF LIVER CIRRHOSIS:**

Cirrhosis of the liver can be caused by several conditions, includes alcoholism, viral hepatitis, non-alcoholic fatty liver disease, and autoimmune disorders.

According to the review of (Liu, Tsai and Hsu, 2021) , the underlying cause of alcohol-induced liver disorder and the mechanisms implicit alcohol-induced liver cirrhosis. (Chalasani *et al.*, 2017)provides a thorough examination of non-alcoholic fatty liver disease and cirrhosis, including pathogenesis, diagnosis, and treatment and concluded the main causes of the Liver Cirrhosis.

### **2.2.2: SYMPTOMS OF LIVER CIRRHOSIS:**

The symptoms of liver cirrhosis can vary depending on the extent of the disease and the underlying cause. Jaundice, fatigue, abdominal pain, and edema are all common symptoms.

According to (Naveau, Perlemuter and Balian, 2005) , It covers the epidemiological aspects and natural history of cirrhosis, includes signs and symptoms, challenges, and prognostic factors. (Lai and Afdhal, 2019)discussed the clinical manifestations of cirrhosis, including symptoms, complications, and management.

### **2.2.3: DIAGNOSTICS OF LIVER CIRRHOSIS:**

A medical history, a physical checkup, laboratory tests, and imaging investigations are commonly used in the diagnosis of liver cirrhosis.

According tothe research of (Emori *et al.*, 2020) which provides a comprehensive review of the cirrhosis diagnostic approach, including clinical, laboratory, and imaging features. (Lim et al, 2015) examines noninvasive diagnostic methods for liver fibrosis and cirrhosis, such as serum biomarkers, transient elastography, and magnetic resonance elastography.

### **2.2.4: TREATMENT OF LIVER CIRRHOSIS:**

The treatment of liver cirrhosis is determined by the underlying cause as well as the severity of the disease.

InAccordance with the study of (Ginès *et al.*, 2021)whichprovides a comprehensive review of the medical, endoscopic, and surgical treatments for cirrhosis and its complications. (Mansour and McPherson, 2018) examines current and future decompensated cirrhosis management strategies, including efficacy, safety, and cost-effectiveness.

## **2.3: LIFE EXPECTANCY:**

Life expectancy is a crucial indicator of the health and happiness of a community. According to current mortality rates, it is the average amount of years that a person can anticipate living. In this review of the literature, the variables that affect life expectancy as well as how it differs between groups will be discussed.

According to a study by (Oeppen and Vaupel, 2002), advances in population health, medical advances, and living circumstances have all contributed to a century-long rise in life expectancy. The research also demonstrated that life expectancy differs significantly between nations, with some having significantly greater life expectancies than the others.

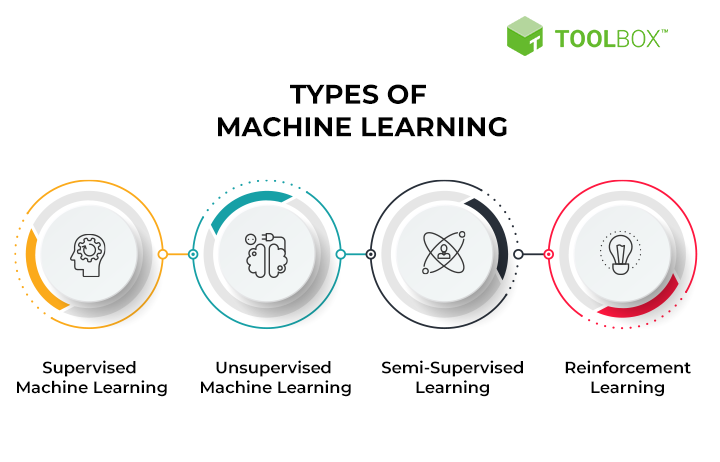
The numerous elements that affect life expectancy, such as genetics, lifestyle choices, economic and social variables, and access to healthcare, were covered in a review paper by (Masters, Hummer and Powers, 2017)The piece also emphasized how critical it is to address these issues in order to raise population health and lessen health inequalities. (Chetty *et al.*, 2016) examined the connection within wealth and life expectancy in the US in another research. According to the research, there is a significant relationship between wealth and life expectancy, with people with higher incomes generally living longer. The research added that raising low-income people's economic opportunities might contribute to lowering mortality rates and raising life expectancy.

According to a (Organization, 2019) report fromabout worldwide trends, life expectancy has been increasing for the last few decades, with global average now standing at 72 years. The study also emphasized the stark differences in life expectancy among various geographic areas and nations, with low-income nations typically having shorter life expectancies.

Even though life expectancy has increased generally around the world, there are still large inequalities between various populations that must be resolved in order to improve community health and well-being.

## **2.3: MACHINE LEARNING:**

Machine learning is a branch of computational intelligence which focuses on developing algorithms that can learn from data and make predictions or judgements based on the learning. Machine learning has become a prominent research area in recent years because of its potential for resolving real-world problems.



**Figure 2. 2: Types of Machine Learning (Spice Works)**

In Accordance with the study of (Zhang, 2010), explained the definition of Machine Learning and its types in the study and explained about the new advances in Machine Learning including the Techniques and approaches. (Pedro, 2012) published an article that provides a summary of machine learning and illustrates some of the key ideas and fundamentals underlying the field. It discusses some of the most common machine learning pitfalls and offers practical guidance for researchers and practitioners. There was discussion of the various types of methods for machine learning and their possible applications, that include supervised learning, unsupervised learning, and reinforcement learning. in a review paper by (Goodfellow, Bengio and Courville, 2016).

Moreover, the review made by (Sculley *et al.*, 2023)discusses machine learning's potential pitfalls, such as analytical debt and the risk of errors. It offers researchers and practitioners practical advice on how to avoid these drawbacks and develop more accurate machine learning models.

### **2.3.1: TYPES OF MACHINE LEARNING:**

Supervised, semi-supervised, unsupervised, and reinforcement learning techniques are the four primary types of machine learning techniques utilized for a wide range of applications.

#### **2.3.1.1: SUPERVISED LEARNING:**

Supervised Learning is a Machine Learning Technique in which the algorithm is trained using labelled data. The system learns from previously classified data before making predictions on fresh, previously unseen data. It is a data analysis methodology that makes use of cutting-edge techniques including decision trees, random forests, and gradient-boosting algorithms. It also teaches algorithms how to accomplish intelligent tasks without human interaction.

Furthermore, the training module is accommodating and adaptable, providing machines to steadily learn new functions and procedures. During the learning stage, the algorithm continuously analyses the model's correctness and adapts until the errors are minimized.

According to the study of (Hastie, Tibshirani and Friedman, 2008)examines recent advances in deep learning algorithms for supervised learning multiclass classification tasks, the variable types and terminology used in Supervised Learning and observed the responses or classically the dependent variables in Supervised Learning. (Cunningham, Cord and Delany, 2008)proposes a hybrid strategy for time series predicting using deep neural networks and supervised learning approaches. This article provides an overview of the two most prominent supervised learning approaches used in multimedia research, support vector machines (SVM) and nearest neighbor classifiers. Also analyze the theory of risk minimization in Supervised Learning Technique.

##### **2.3.1.1.1: Types of Supervised Machine Learning:**

In the supervised learning process, various algorithms and computation methods are used. The following are some examples of supervised learning algorithms:

1. **Regression:**

The method of regression is used to determine the relationship among dependent and independent variables. It is also a sort of supervised learning in which a technique trains from labelled data sets to anticipate continuous output for various variables. It is usually believed to be utilized in situations when the outcome must be a fixed value, including height or weight.

There are two kinds of regression, which are as follows:

* **Linear Regression:**

It is used to establish the relationship among the two factors, which is subsequently utilized to forecast the future. Linear regression is also categorized based on the number of variables that are both independent and dependent.

* **Logistic Regression:**

When the variable that is dependent is categorized or has binary responses like 'yes' or 'no,' a method logistic regression is utilized. Furthermore, logistic regression predicts distinct values for variables because it is employed to tackle binary classifier problems.

1. **Naïve Bayes:**

The Naive Bayes technique is used for huge datasets. The technique is predicated on the assumption that each software component of the algorithm acts independently. This means that the presence of one characteristic has no bearing on the existence of another. It is widely used in text categorization, recommendation systems, and a variety of other applications.

There are several types of Naive Bayes approach, the most common of which is the decision tree. A decision tree is a supervised learning technique that has a flowchart-like structure. However, they do have unique jobs and responsibilities.

Statements of control containing decisions and their results comprise a decision tree. The output of a decision tree is associated with the labelling of unpredictable data. ID3 and CART are two of the most extensively utilized decision tree algorithms in several sectors.

1. **Classification:**

It is a type of algorithm for supervised learning that categorizes or classifies data appropriately. It recognizes and evaluates specific entities in order to determine where they should be categorized. The following are some classification algorithms:

* Random Forest
* Support Vector Machines (SVM)
* K- Nearest Neighbor (KNN)
* Linear Classifiers
* Decision tree

1. **Neural Networks:**

This type of supervised learning method is used to sort or categorize raw data. It is also utilized to recognize patterns and analyze sensory input. However, the approach needs a large number of computational resources. Consequently, it has restricted applicability.

1. **Random Forest:**

A random forest method is sometimes known as an ensemble approach since it uses multiple ways of supervised learning to achieve a result. Furthermore, it makes use of a large number of decision trees to generate the categorization of individual trees. As an outcome, it is widely used in a variety of industries.

In this dissertation, the Supervised Machine Learning algorithms is utilized to acquire the desired outcome.

#### **2.3.1.2: SEMI-SUPERVISED LEARNING:**

Semi-supervised learning is a machine learning approach in which the algorithm is trained on both labelled and unlabeled data. When labelled data is scarce but unlabeled data is abundant, this type of learning is useful.

In Accordance with the review of (Zhu, 2005) analyzed some of the literature survey on Semi-supervised Learning Techniques and concluded that it requires less human effort and provides higher accuracy. (Lecouat *et al.*, 2018) uses generative adversarial networks (GANs) and manifold regularization to propose a semi-supervised learning framework.

#### **2.3.1.3: UNSUPERVISED LEARNING:**

Unsupervised learning is a machine learning method in which the algorithm is trained on unlabeled data. The algorithm discovers the fundamental structure of the information and groups together similar data points.

According to the study of(Mejri *et al.*, 2023) provides a thorough examination of recent unsupervised learning techniques for detecting anomalies in time series data. (Denton and Birodkar, 2017) proposed an unsupervised learning method for separating the underlying causes of variation in videos and demonstrated the ability to coherently generate up to several steps into the future.

#### **2.3.1.4: REINFORCEMENT LEARNING:**

Reinforcement learning refers to a sort of machine learning where the system of algorithms learns from its surroundings through interaction. The algorithm receives input in the form of prizes or punishments based on its behaviors, and it trains to maximize its incentives over time.

In Accordance with the study of(Xiang *et al.*, 2023) provides an in-depth examination of recent advancements in reinforcement learning for robotics and concludes that none of the investigated algorithms can handle all tasks well, that soft Actor-Critic outperforms most algorithms in terms of average reward and stability, and that an algorithm combined with data augmentation may facilitate learning policies. (Kiran *et al.*, 2021) examines recent advances in deep reinforcement learning for self-driving and discussed the role of simulators in training agents, as well as methods for validating, testing, and justify existing Reinforcement Learning solutions.

### **2.3.2: PREDICTION USING MACHINE LEARNING TECHNIQUES:**

Machine learning technologies' capacity to foresee results and trends from large datasets has grown their prominence in recent years. The present understanding of prediction using methods from machine learning will be investigated in this literature review.

According to the study of(Rajkomar *et al.*, 2018) showed how machine learning can be used to forecast clinical outcomes. In the research, machine learning models outperformed conventional predictive models in predicting patient fatality, recurrence, and length of stay using data from electronic health records. Another research by (Ching et al., 2016) examined the use of artificial intelligence (AI) in medical diagnosis, demonstrating that AI algorithms can reliably identify conditions like breast cancer, skin cancer, and pneumonitis. In terms of health care, research by (Wiens *et al.*, 2019) investigated the application of machine learning for sepsis prediction, a condition that poses a serious risk of death.

In order to forecast stock prices, (Tsantekidis *et al.*, 2019) used methods based on machine learning in their research. The study showed that when it came to predicting stock prices, machine learning models performed better than conventional statistical models. The difficulties of using machine learning for forecasting, which include overfitting, bias, and interpretability, were also covered in the paper.

### **2.3.3: How Machine Learning Works in Predicting Life Expectancy of Liver Cirrhosis?**

Machine Learning models beat conventional risk and life expectancy prediction techniques in patients with Cirrhosis.

According to the research, machine learning models were effective at predicting sepsis, which could lead to earlier diagnosis and better treatment. (Tian *et al.*, 2019) used multiple machine learning techniques to forecast the likelihood of Hepatitis B Surface Antigen Sero-clearance in patients with Liver Cirrhosis. Support vector machines were used by (Li *et al.*, 2020) to evaluate and identify FLD in B-mode ultrasonic images with various textures. Using anatomical and echogenic data from ultrasonic images, (Ribeiro, Singh and Guestrin, 2016) applied the Bayesian approach to extract features and separate fatty liver ultrasonic pictures from normal liver ultrasonic images and observed that the scar tissue's structure has formed a risk of rupture within in the liver. This can result in internal bleeding and is immediately fatal. In terms of cirrhosis of the liver life expectancy, approximately 43% of patients survive beyond a year.

In Accordance with the research of (Arshad and Dutta, 2016)finds that,throughout the globe, a lot of individuals consume alcohol in excess. Drinking alcohol is undeniably linked to dangerous liver diseases like cirrhosis, which can ultimately result in mortality. Early identification of liver disease induced by heavy alcohol consumption might help many people live longer lives. By detecting liver illness early on, it may be detected in time and leading to complete treatment in some people. This study suggests using information mining algorithms to identify and predict the likelihood of liver disease. (Ramkumar *et al.*, 2017)observed thatMalignant growth spreads to the lungs, liver, and other organs like breasts, bones, and so on. Liver cancer is the most frequent type of cancer. It is dangerous and will last a long time. The side effects are Jaundice, weight loss, and yellow skin are symptoms of a malignant growth in the liver. Pain and anguish in the upper right stomach area, shaded pee, spewing Sweating, fever, and an enlarged liver are all symptoms of a weakened liver. The cancer of the liver that develops in the liver rather than spreading from another area of the body is referred to as a vital liver disease. Numerous data mining calculations, artificial intelligence to predict liver disease, insight ideas are used. The likelihood of predicting liver cancer growth using the WEKA tool, the Bayes hypothesis was tested.

However, in order to guarantee accurate and trustworthy predictions, issues like bias and overfitting must be considered.

## **2.4: SYSTEMATIC REVIEW OF RELATED WORKS:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Author Name** | **Article Title** | **Data Source** | **Methods Used** | **Results and Discussion** | **Significance and Relevance** |
| Tian et al., 2019 | Using Machine Learning Algorithms to Predict Hepatitis B Surface Antigen Seroclearance. | The dataset contains a total of 30 factors. Gender, alcohol history, HBV history in the family, the initial diagnosis, and so on are among the ten categorical variables. The remaining parameters, including age, BMI, serum metrics, and radiology indicators, are continuous. The dataset was separated into two portions for training and evaluating machine learning models: train (70%) and test (30%). | In this study, the predictive model employs four machine learning classifications algorithms: logistic regression, decision tree, random forest, and extreme gradient boosting. | The results show that machine learning algorithms, specifically XGBoost (81.9%), can predict HBsAg seroclearance reliably. This study also revealed the utility of machine learning approaches in predicting clinical outcomes. | However, numerous limitations were found in this investigation, including the use of small data sets, the necessity for continuous follow-up data, and the limited data that was examined from the already established predictor. A appropriate model with sufficient precision and generalisation for predicting HBsAg seroconversion is still awaiting presentation. |
| (Bourlière *et al.*, 2015) | Ledipasvir-sofosbuvir with or without ribavirin to treat patients with cirrhosis non-responsive to previous protease-inhibitor therapy. | Data which contains a total of 150 instances with genotype 1 HCV infection, who have already received HCV treatment, and who have cirrhosis is used in this study which was randomized into two groups as testing and training dataset. | Support Vector Machine (SVM) Approach is used in this study to predict the model accuracy and performance. | The result demonstrates that the accuracy is achieved to 84.9% which seems to be a positive performance. | However, some limitations were identified in this study, which is the long-term treatment for Liver Cirrhosis is not possible. Moreover, the better accuracy can be achieved by using various algorithms. |
| (Ma *et al.*, 2015) | Sugar-sweetened beverage, diet soda, and fatty liver disease in the Framingham Heart Study cohorts. | A dataset which contains 2634 instances of US Liver imaging factors is used in this study to predict the foremost cause for the Liver Disease. | To determine the main causes of liver cirrhosis, a signal analysis method that utilized kurtosis imaging and classifying fatty liver disorders was used. | The outcomes demonstrates that the accuracy is 81.2% and conclude that regular sugar-sweetened beverage consumption was linked to an increased risk of developing fatty liver. | Even though, achieved an accuracy of 81.2%, they cannot find significant association between the p values between the  sugar-sweetened beverage and diet coke. |
| (Saba *et al.*, 2019) | The present and future of deep learning in Liver Cirrhosis. | US liver images which contain 5143 instances which includes the variables such as age, gender, albumin level, bilirubin and so on. | In this study the image analysis using Backpropagation Neural Network (BPNN) to predict the Liver cirrhosis. | The accuracy of BPNN with 10 hidden layers and 128 features extracted from 6 different feature extraction techniques was 77.6%. | However, it shows a positive performance still different algorithms can be used to provide a better model with most common algorithms. |
| (Li *et al.*, 2020) | Measuring MELD score Using Machine Learning in Liver Cirrhosis. | ILPD dataset from UCI is used in this study which contains 306 instances of the patients records who have Liver diseases. | In this study, classification techniques such as Support Vector Machines and Naive Bayes are used for predicting the MELD score for Liver Cirrhosis. | The author suggested that over 45 years old, have a family history of liver disease, or are overweight or obese, to consult a doctor. The predicted score of Naïve Bayes is approximately 82.5%. | However, the other algorithms such as random forest and regression can also be used to predict the liver diseases in the future studies. |
| (Shi *et al.*, 2022) | Acute-On-Chronic Liver Cirrhosis: Natural History, Mechanism, And Treatment. | Chronic Hepatitis C cases dataset which is retrieved from UCI repository is used in this study to predict the causes of Liver Cirrhosis. | Machine Learning methods including Random Forest, Regression, and SVM are used to analyze biomakers. | In chronic Hepatitis C cases, serum biomarkers were employed to predict advanced fibrosis and found that the Support Vector Machines have a better accuracy of 79.4%. | The review also demonstrates the limitations such as hepatic encephalopathy (HE), and infections, and possible treatments, can be emerged using stem-cell-based therapies in further studies. |
| (Moon, Singal and Tapper, 2019) | Contemporary Epidemiology of Chronic Liver Disease and Cirrhosis. | Chronic Hepatitis B patients’ dataset with 1432 instances split into testing and training data models. | The features are retrieved using the Recursive Feature Elimination technique, which is used in this study to get the desired result. | A prediction model for chronic liver failure was built and tested in chronic Hepatitis B cases employing a recursive feature elimination technique, and it was discovered that cirrhosis prevalence and death have grown by 75% in recent years. | Although hepatitis B and C vaccination, evaluation, and prediction strategies have helped to reduce the CLD burden, concomitant increases in injectable drug use, alcohol misuse, and metabolic syndrome call these trends into question. |
| (Kanwal *et al.*, 2020) | Development, Validation, and Evaluation of a Simple Machine Learning Model to Predict Cirrhosis Mortality. | The dataset of Cirrhosis mortality which is retrieved from the Kaggle site is used in this study to predict the Mortality rate among the Liver cirrhosis patients. | Techniques including Gradient Descent Boosting, Logistic Regression with Least Absolute Loss, and Selection Operator Regularization are used to forecast the mortality rate. | Simple machine learning approaches outperformed in this investigation. And discovered a more improved ensemble gradient boosting technique with an accuracy of 85%. | Even though, it provides a better outcome. The various algorithms can also be used to get an even better outcome in the future studies. |

## **2.5: CONCLUSION:**

In conclusion, predicting the life expectancy of patients with liver cirrhosis which use machine learning techniques is an exciting field of study. According to new research, machine learning algorithms may predict the survival rate of individuals with liver cirrhosis. These models accurately forecast patients' life expectancy by combining clinical, laboratory, and imaging information.

For predicting the life expectancy of patients with liver cirrhosis, supervised learning algorithms such as decision trees, random forests, and support vector machine algorithms have been commonly used. These models have shown positive performance and have the possibility of enhancing patient outcomes by predicting life expectancy accurately. However, there are significant drawbacks to utilizing algorithms based on machine learning to forecast the life expectancy for individuals with liver cirrhosis, such as high-quality data, the possibility of overfitting, and interpretable models, which can be readily solved by using patients' ultrasonic images.

Overall, machine learning techniques which can be used to improve the prediction of life expectancy for patients with liver cirrhosis.

# **CHAPTER 3: METHODOLOGY AND APPROACHES:**

The information comes from a Mayo Clinic trial on cirrhosis of the primary biliary tract (PBC) of the liver. Lichman, M. gathered the dataset utilized in this dissertation from Kaggle, where it was obtained from the UCI ML Repository, and the dataset was deployed. This Data Set consists of 418 instances and 21 Attributes such as age, stages, Drug, Sex, Ascites, and other factors such as Bilirubin, Edema, Albumin, Copper, SGOT, Platelets, Triglycerides, Prothrombin and lifeExpectancy.

## **3.1: INTRODUCTION:**

Various strategies and techniques are utilized in the area of machine learning to construct predictive models for a wide range of situations. In this study, various popular techniques are used to predict life expectancy for patients with liver cirrhosis, includes Linear Regression, Multi-Linear Regression, Random Forest, Decision Tree, Gaussian Naive Bayes, and Support Vector Machines (SVM).

Cirrhosis of the liver is a long-term liver condition characterized by fibrosis of the liver's tissues, which can cause a number of disorders and shorten one's life expectancy. Accurate life expectancy prediction can be useful in determining the severity of the disease, prognosis, and treatment strategy. Machine learning techniques provide a strong tool to analyzing complex data and developing predictive models for predicting life expectancy for individuals with liver cirrhosis.

A combination of linear regression as well as multi-linear regression, that represent the relationship among independent and dependent variables, as well as combination methods that include Random Forest, which merge many decision trees, are used in this work to produce more accurate and resilient models. Gaussian Naive Bayes, a method based on probability, and decision trees, a graphical representation of the decision results, are employed to increase the models' prediction accuracy. Furthermore, Support Vector Machines (SVM), a commonly used classification and regression tool, was utilized to establish the best potential boundaries among two classes in the data set.

The accuracy and usefulness of these procedures and strategies in predicting life expectancy of individuals with liver cirrhosis were determined using appropriate Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are two evaluation metrics. The study's findings could have significant consequences in the area of hepatology, enabling insights into illness prediction as well as treatment planning, and eventually enhancing patient care.

## **3.2: Data Set used:**



[cirrhosis\_prediction.csv](https://uelac-my.sharepoint.com/:x:/g/personal/u2216077_uel_ac_uk/EbZHC0Xd499HrAQQEemaiscBwgL9-4MVN6ve7WYYc7TJkA?e=rK0Pxu)

This data set is used in this dissertation to obtain the approximate life expectancy of people who have liver cirrhosis.

## **3.3: Data Set Description:**

The above data set is used to predict the life expectancy of people with liver cirrhosis. This data set contains 418 instances and 21 Attributes. This dataset is available in Kaggle, which is retrieved from UCI Machine Learning Repository.

### **3.3.1: Attributes and its Description:**

|  |  |
| --- | --- |
| **Attributes** | **Description** |
| **ID** | Unique identifier for each individual. |
| **N\_Days** | Number of [days] between registration and the earlier of death, transplantation, or study analysis. |
| **Status** | Status of patients [C, CL or D] whereas C-Censored, CL-Censored due to Liver damage, or D-Death. |
| **Drug** | Type of drug [D-penicillamine or Placebo]. |
| **Age** | Age in years. |
| **Sex** | [M or F] M-Male or F-Female. |
| **Ascites** | Presence of ascites [N or Y] N-No or Y-Yes. |
| **Hepatomegaly** | Presence of Hepatomegaly [N or Y] N-No or Y-Yes. |
| **Spiders** | Presence of spiders [N or Y] N-No or Y-Yes. |
| **Edema** | Presence of edema [N,S or Y] whereas N- no edema and no diuretic therapy for edema, S- edema present without diuretics, or edema resolved by diuretics, or Y- edema despite diuretic therapy. |
| **Bilirubin** | Serum bilirubin in [mg/dl]. |
| **Cholesterol** | Serum cholesterol in [mg/dl]. |
| **Albumin** | Serum albumin in [mg/dl]. |
| **Copper** | Urine copper in [ug/day]. |
| **Alk\_Phos** | Alkaline phosphate in [U/liter]. |
| **SGOT** | SGOT in [U/ml]. |
| **Triglycerides** | Triglycerides in [mg/dl]. |
| **Platelets** | Platelets per cubic [ml/1000]. |
| **Prothrombin** | Prothrombin time in seconds [s]. |
| **Stage** | Histologic stage of disease [1,2,3 or 4] Stage  1: Steatosis, Stage 2: Liver scarring (fibrosis),  Stage 3: Cirrhosis and Stage 4: Hepatic failure, liver failure, or advanced liver disease. |
| **lifeExpectancy** | The average life expectancy of an individual. |

Here the dependent variable is lifeExpectancy and independent variables are N\_Days, Age, Sex, Stage and the amount of serums like Bilirubin, Albumin are utilized in Data Exploration.

## **3.4: Data Pre-Processing:**

Data preprocessing is the process of converting data into a format that can be more easily and successfully processed in machine learning, data mining, and other data science processes. The following steps are used in data preprocessing such as:

1. **Data Collection:** Gathering data is the initial stage of any machine learning project. In this situation, we require a dataset containing information about individuals with liver cirrhosis and their average life expectancy.
2. **Data Cleaning:** Once data is received, it must be cleaned in order to eliminate any errors, inconsistencies or missing information. This entails:

* **Handling Missing Values:** Values that are missing can be imputed using various techniques, including mode computation, mean imputation, and regression imputation. It is crucial to highlight, however, that imputed missing values can inject bias into the data.
* **Handling Outliers:** Outliers can be found using a variety of methods such as box-plotting, scatter plots, or the Z-score. Outliers can be eliminated or handled in a variety of ways, including capping and transformation.

1. **Data Transformation:** To meet the criteria of the algorithms used for machine learning, the data might have to be converted. This entails:

* **Encoding categorical variables:** Techniques such as single-hot encoding and label encoding can be used to encode categorical variables.
* **Normalization and Scaling:** To standardize the data, normalization and scaling procedures can be used.

1. **Data Splitting:** After the data has been pre-processed, it must be divided into sets for testing and training. The set of training data is employed to train and test machine learning models.
2. **Data Frame:** A data frame is a regularly used data structure for analyzing data in R, in addition to the most commonly used technique for information collection in R. A data set is a collection of up of equal-length feature vectors. The components of the list can be thought of as rows, and the total length of every component is proportional to the total number of columns.

## **3.5: Techniques used:**

Several machine learning techniques are used to predict the life expectancy of people with liver cirrhosis.

Graphical user interface, application

Description automatically generated

**Figure 3. 1: Models used in this Dissertation (Self)**

Some of the techniques used in this dissertation are as follows:

1. **Linear Regression:**

Linear regression is a fundamental machine learning technique that investigates the correlation between the dependent variable (in this example, life expectancy) and a few independent variables. The purpose of linear regression is to establish a linear connection between a dependent variable and a set of independent variables that minimizes the sum of squared errors between predicted and actual values.

**Syntax:**

The linear regression formula in mathematics is y = b0 + b1\*x + e, where b0 and b1 are the regression beta coefficients or variables such as: b0 is the intercept of the regression line; it is the predicted value for x = 0. b1 represents the slope of the regression line.

The lm() function is used to represent linear regression. The lm() method has the following syntax. # The lm() function model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.

**Applications:**

Linear regression is frequently used to explain and predict the connection between variables in many domains, includes economics, finance, marketing, healthcare, and social sciences. It is a useful tool for prediction, trend analysis, and discovering patterns or connections between variables based on historical data.

**Advantages:**

* Simplicity: Linear regression is a straightforward and interpretable technique that is simple to understand and implement.
* Speed: Linear regression is computationally effective and can be taught and evaluated fast, making it appropriate for huge data sets and real-time applications.
* Interpretability: Linear regression gives interpretable coefficients for each feature, making it simple to interpret the effect of each predictor variable on the target variable.

1. **Multilinear Regression:**

Multilinear regression is a modification of linear regression that includes more than one independent variable. To forecast the result of the factor that is dependent, the method analyses the relationship between the dependent variable and several independent factors.

**Syntax:**

The multiple regression equation given above is written as y=b1x1 + b2x2 +... + bnxn + c. When the parameter that predicts it changes, the coefficients of regression (i=1,2,...n) reflect the new level in which the factor known as the criteria varies.

The mathematical equation is y = mx1 + mx2 + mx3 + b.

Where Y is the dependent variable in the regression. M represents the regression slope.

**Applications:**

Multi-linear regression is frequently employed to investigate and forecast the relationship between many variables in many domains, includes economics, finance, marketing, healthcare, and social sciences. It can provide light on the combined impacts of numerous variables on the variable that is dependent and aid in the identification of relevant variables and their corresponding contributions.

**Advantages:**

* Multiple Predictor Variables: Multilinear regression enables the inclusion of many predictor variables, allowing the model to identify more complex correlations between variables.
* Flexibility: Multilinear regression is capable of handling datasets with several independent variables of various sorts, providing greater versatility in modelling various forms of data.
* Better Prediction Accuracy: When compared to ordinary linear regression, multilinear regression can yield more accurate predictions.

1. **Random Forest Technique:**

Random Forest is a decision tree-based technique that constructs many decision trees and then combines them to make the final forecast. It is a collective approach that integrates many decision trees to produce an even more accurate and resilient model. The forecasts of each decision tree are combined for determining the final prediction, and each tree is created based on a randomly selected portion of the training data.

**Syntax:** randomForest(formula, data)

**Parameters:**

formula: a formula describing the model to be fitted.

data: a data frame containing the model's variables.

**Applications:**

Random Forests are used in a variety of applications, includes recognition of images, spam detection, disease prediction, credit risk evaluation, and systems for recommendation. They are also employed in selecting features and variable significance ranking since they provide measurements of feature value according to how much it contributes to the model's accuracy.

**Advantages:**

* Random forests outperform decision trees when running on large databases.
* Random forests are extremely precise because they are a collection of decision trees, and every tree of choices draws a sample of random data, resulting in greater prediction accuracy.
* Efficient estimations of test error: It effectively employs all predictive parameters and maintains reliability even if data is missing.

1. **Decision Tree Technique:**

A decision tree is a visual representation of every potential approach to a choice depending on specific parameters. It is an algorithm that creates a tree-like representation of decisions and their potential outcomes. Every node in the tree indicates a decision or condition, and each branch reflects how that decision or situation was resolved.

**Syntax:**

rpart(formula, data=, method='').

Whereas,

formula: The prediction function

data: This method specifies the data frame:

- "class" in the context of a classification tree

- "anova" to represent a regression tree

**Applications:**

Decision Trees can be used for a variety of purposes, includes detecting fraud, segmentation of customers, healthcare diagnosis, systems for recommendation, and many others. They are frequently used as the foundation for more sophisticated methods of ensemble learning including Random Forests (RF) and Gradient Boosting, which aggregate many Decision Trees to enhance accuracy and robustness.

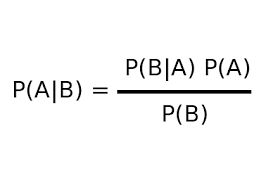
**Advantages:**

* Simple to grasp.
* Minimum data cleansing is required.
* There are no restrictions on the data type.

1. **Naïve Bayes Approach:**

In R-programming, the Naïve Bayes method is a Supervised Non-Linear Classifier algorithm. Gaussian Naïve Bayes is a probabilistic approach based on Bayes' theory. It is a basic yet effective algorithm that predicts based on the input data by computing the expected outcome of each class. The term "naïve" refers to the assumption that each input variable exists independently of the other input variables.

**Syntax:**



Where, P(A|B) represents the conditional likelihood of A given B.

P(B|A) denotes the conditional likelihood of B in the presence of A.

The chance of event A is denoted by P(A).

The chance of event B is denoted by P(B).

For many predictors, the posterior probability can be expressed as follows:

P(A|B) = P(B1|A) \* P(B2|A) \* P(B3|A) \* P(B4|A) \* P(B5|A) \* P(B6|A) \* P(B7|A) \* P(B8|A)….

**Applications:**

Gaussian Naive Bayes is a text classification method that is extensively used for identifying spam, sentiment analysis, and theme categorization. It is notable for its ease of use and quickness when dealing with huge text datasets. It may also be used to detect fraud in a variety of sectors, such as financial fraud, insurance fraud, and credit card fraud. It can quickly analyze enormous volumes of transactional data and detect questionable patterns. Gaussian Naive Bayes is frequently utilized in healthcare systems, image identification systems, and recommendation systems, among other applications.

**Advantages:**

The Gaussian Naive Bayes algorithm's key merits are its simplicity and speed of calculation. It is capable of handling both continuous and discrete characteristics, making it appropriate for a wide range of datasets. It also resists overfitting and does well with tiny datasets.

1. **SVM (Support Vector Machines):**

SVM is a method of machine learning that is used for regression and classification. It finds the best probable boundaries within two classes in a data collection. The algorithm looks for the hyperplane that has the highest margin of separation among the two classes, which is defined as the distance among the hyperplane as well as the nearest point of data in each class.

**Syntax:**

SVM Error = Margin Error + Classification Error.

The larger the margin, the lower the possibility of marginal error, and inversely.

**Applications:**

SVM has found widespread use in image recognition applications such as object identification, picture categorization, and facial recognition. It is capable of handling high-dimensional picture data and determining optimal decision limits. SVM has been used in bioinformatics to predict protein structure, analyze gene expression, and discover new drugs. It can analyze complex biological information and making accurate predictions. It's also utilized for sentiment analysis, classification of text, and spam identification in text and document analysis activities. It is capable of handling high-dimensional text data effectively and create reliable predictions based on text properties.

**Advantages:**

* **High Dimensionality:** SVM is an excellent tool in high-dimensional environments, especially in document categorization and sentiment analysis, where the degree of dimensionality can be exceedingly high.
* **Memory Efficiency:** Because just a fraction of the training points is utilized in the decision process of allocating new members, only these points must be retained in memory (and computed on) when making judgements.
* **Versatility:** Class separation is frequently exceedingly non-linear. The ability to apply fresh kernels provides significant flexibility for decision boundaries, resulting in improved classification performance.

1. **Survival Analysis with Kaplan-Meier Survival analysis and Cox-Hazard Proportional model:**

**Kaplan-Meier Survival Analysis Approach:**

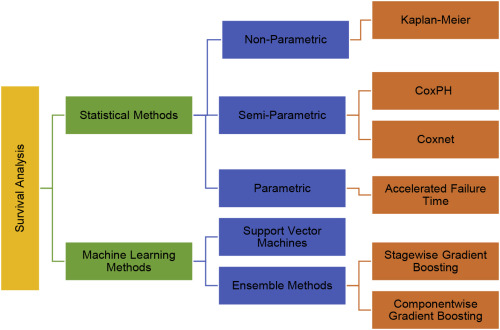
The Kaplan-Meier survival analysis, also known as the Kaplan-Meier estimator, is a non-parametric statistical procedure for predicting the likelihood of survivability or time-to-event in a group of patients over time. It is commonly used to examine time-to-event data which includes the period until death, the period until a medical condition recurs, or the duration until a certain occurrence.

**Applications:**

The Kaplan-Meier survival analysis approach is widely used in a variety of domains, include medical studies, the field of epidemiology, clinical studies, social science, and reliability analysis. It is extensively used in medical research to analyze survival results in clinical investigations such as cardiovascular disease, cancer, and other health disorders. It can also be used in engineering reliability analysis to analyze the time to failure or time until an event happens, such as in failures analysis of mechanical structures, equipment, or components.

**Advantages:**

The non-parametric nature of Kaplan-Meier survival analysis, its capacity to handle censored data, the providing of survival probability at multiple time points, the facilitation of comparing the results of survival between groups, and its capacity to handle time-varying factors are all advantages. These benefits make Kaplan-Meier analysis of survival an important and extensively used technique for analyzing time-to-event data, particularly when censoring is present or when distributions assumptions are uncertain or not met.



**Figure 3. 2: Types of Survival Analysis (Science Direct)**

**Cox-Hazard Proportional Model:**

The Cox proportional hazard approach, frequently referred to as Cox regression or the Cox approach, is a statistical tool for analyzing the connection between predictor factors and time-to-event outcomes in survival data. It is a semi-parametric approach which permits for an estimation of hazard proportions, that describe the relative probability of an event occurring for various numbers of variables, without assuming any specific survival time distribution.

**Applications:**

The Cox proportional hazard model has several uses in medical studies, the subject of epidemiology clinical studies, and other fields that require time-to-event data. The Cox model's versatility and interpretability make it a useful tool for analyzing surviving data and determining the relationship between variables and time-to-event results in a variety of applications.

**Advantages:**

The semi-parametric nature of the Cox proportional hazard model, which doesn't need presumptions about the fundamental distribution of surviving times, being able to handle censored data, versatility in modelling covariates, offering ratios of hazard for analyzing relative risk, and managing time-varying covariates are all advantages.

## **3.6: CONCLUSION:**

In this Methodology, Various Machine Learning algorithm is applied in order to estimate life expectancy for liver cirrhosis. Linear regression and multi-linear regression are the most fundamental approaches for modelling the connection between a dependent variable and several independent variables. Random forest is a collective approach for creating a more accurate and resilient model by combining several decision trees. Gaussian Naive Bayes is a method based on probabilities that uses input data to calculate the expected outcome of each class. A decision tree is a visual representation of all potential decisions depending on given parameters. SVM is used for regression analysis and classification by determining the best probable border within two classes in a data set. The Kaplan-Meier survival analysis and the Cox proportional hazards model are prominent techniques for analyzing time-to-event data across a wide range of areas. Kaplan-Meier survival estimation is a non-parametric approach for estimating survival probability without assuming any probability for survival times, which makes it appropriate for data with uncertain or non-normal distributions. The Cox proportional hazards approach is a semi-parametric approach to modelling covariates and generating hazard ratios to understand relative risk. The Kaplan-Meier analysis of survival is appropriate for simple analyses, whereas the Cox proportional hazard model offers greater flexibility in covariate modelling.

Each of these strategies has advantages and disadvantages and can be utilized depending on the data and the problem to be resolved. these strategies are employed in this study to preprocess data and develop algorithms to predict life expectancy for liver cirrhosis. Each algorithm's performance was assessed using metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), which assisted us in determining the optimal model for this dataset. The optimum algorithm must be used in order to obtain accurate predictions that can aid in the identification and management of individuals with liver cirrhosis. Hence, the above-mentioned approaches are to be performed and the accuracy can be calculated. Furthermore, the Survival analysis is also done to estimate the recovery time for an individual with Liver Cirrhosis.

# **CHAPTER 4: PRACTICAL IMPLEMANTATION:**

Several critical phases are involved in the implementation, including data pretreatment, model construction, and model evaluation. Data preprocessing entails cleaning and preparing incoming data, which may include missing value management, the choice of features, and feature scaling. After the data has been cleaned and preprocessed, it is used to construct machine learning models such as Linear Regression, Multi-Linear Regression, Decision Tree, Random Forest, and Gaussian Naive Bayes, and SVM. The models in this category are trained using a labelled dataset that contains variables for input (such as age, gender, severity of disease, and other pertinent characteristics) and life expectancy values.

## **4.1: INTRODUCTION:**

Life expectancy prediction for individuals having liver cirrhosis is a vital problem in healthcare and research on medicine. Machine learning approaches hold great promise for reliably predicting life expectancy according to relevant information. In this practical application, various popular techniques are used to predict life expectancy for liver cirrhosis patients, includes Linear Regression, Multi-Linear Regression, Random Forest models, a Gaussian Naive Bayes, Decision Trees, and SVM (Support Vector Machine). These models are trained using a labelled dataset that contains variables for input and life expectancy values.

The model evaluation is carried out after the models have been trained to evaluate their performance. This entails determining the precision and efficacy of the models in predicting life expectancy for liver cirrhosis patients using relevant evaluation metrics including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Accuracy, and F1-score. The most accurate and dependable model(s) for predicting life expectancy are chosen based on the model evaluation results.

Implementing these strategies in practice can provide useful insights for healthcare practitioners in forecasting life expectancy of individuals with liver cirrhosis, allowing them to make more educated decisions about treatment programs, patient care, and allocation of resources. It can also aid with medical research by finding significant characteristics that impact life expectancy in patients with liver cirrhosis, ultimately enhancing patients results and the standard of care.

## **4.2: SECTION 1: DATA EXPLORATION:**

Data exploration is the fundamental phase of data analysis, and it is used to analyze and visualize data in order to disclose insights from the start or to identify regions or patterns to pursue further. Employing interactive visualizations and point-and-click exploration of data allows for a better knowledge of the broad picture and faster access to insights. Data mining and graphical analytics tools promote comprehension and empower users to examine data in any visualization. This strategy reduces the time required to obtain answers while increasing knowledge by exploring more topics in less time. As a result, data exploration is critical since it democratizes data access and provides governed self-service analytics.

### **4.2.1: Missingness Map:**

A missingness map displays where data that is missing in each variable ends up. It reveals which is reduced out of the evaluation earlier, as well as whether many components were missing at the time. If this is the case, an analyst may conclude that these factors are related.

### **4.2.2: Boxplot:**

A box plot is a graphical representation of statistical data that shows the minimum, first percentiles, a median, third quartile, and maximum values of the data. The term "box plot" refers to the chart's representation as a rectangular prism with a frequency of lines measured from the highest point to the bottom. A boxplot in R, also referred to as a box plot with whiskers, is a visual representation that may summarize the major aspects of the data (position, dispersion, skewness, and so on) and find outliers.

### **4.2.3: Normalization:**

Normalization is the process of organizing statistics in a database. This entails developing charts and merging them in accordance with the criteria defined to safeguard the information and boost the database's flexibility by avoiding repetitions and inconsistent reliance. There are three approaches in this category: Min-Max, Z-score, and SoftMax scaling.

1. **Min-Max Scaling:**

A common data normalization approach is min-max normalization. The maximum and minimum values of each parameter have been transformed to 0 and 1, accordingly, while every other property has been converted to a range from zero to one.

1. **Z-Score Scaling:**

The Z-Score statistic is applied to determine how far a data set deviates from its mean. It computes a mean as well as the standard deviations which are greater or fewer than the mean. Its standard deviation varies between -3 to +3.

1. **Soft-Max Scaling:**

The SoftMax function transforms the coordinates of K (Lambda) actual values into a probabilistic distribution functional with K (Lambda) different options. It is referred as the normalized exponential function or the soft argmax. In multinomial regression, it is a multidimensional generalization of the logistic function.

### **4.2.4: Visualization:**

Data visualization is a method for graphically portraying data. It will simplify the data by incorporating elements that include visual representations, statistics, charts, histograms, maps, and so on. Visualization is an approach for making data easier to analyze.

**Histogram (HeatMap):**

The numerical value components of a variable's component frequencies are depicted as bucketed ranges in a histogram. Although bar graphs and histograms look similar, the latter organizes the qualities into continuous ranges. The altitude of each of the bars in a histogram represents the proportion of the systems of values that lie within that range. The hist () parameter in R is used to generate histograms.

**4.2.5: Kolmogorov-Smirnov Test of Normality for dependent variable:**

The Kolmogorov-Smirnov Test, additionally referred to as the one-sample K-S analysis, is a non-parametric in trial of an even distribution of non-continuous and continuous probability distributions in a 1D probability density function that can be used to compare the number of instances to an indicate probability trial (which is also referred to as the two-sample K-S test). The KS-test is used to check whether the given variables are normally distributed: the parameters are normally distributed if the average value is greater than 0.05.

### **4.2.6: Correlation:**

The correlation coefficient (r) of a sample is a statistic that indicates the degree to which the characteristics in a scatter graph relate to a line of linear regression created by those points.

**Pearson’s Correlation:**

The Pearson's coefficient of correlation (r) is the most commonly used method for calculating a linear relationship (r). Values ranging from -1 to 1 show the degree and pattern of a relationship between two variables.

**Partial Correlation:**

When a minimum of one additional variable is considered, this partial correlation is the relationship of two variables. The partial coefficients of correlations is a common measure of the magnitude of the relationship between the two variables after all other variables are taken into consideration.

### **4.2.7: G-G Plot:**

The visualization software ggplot2 has tools for generating complicated graphic representations of underlying data in a dataset. It offers significantly more effective inputs for deciding what parameters to plot, how to show them, and other graphical characteristics.

### **4.2.8: Scatterplot:**

A "scatter plot," is a type of plot used to demonstrate the relationship among two numerical variables, displays one dot for each observation. It involves the usage of two vectors of identical dimensions, such as: one for the x-axis longitudinally (horizontal) and one for the y-axis diagonally (vertical).

### **4.2.9: Q-Q Plot:**

QQ Plot (Quantile-Quantile plots) are comparisons of two quantiles. A quantile is a proportion of values which lie below that quantile.

### **4.2.10: Implementation for Data Exploration:**

* Load the Dataset in R-Studio and visualize the missing values. As a result, the missingness map can be obtained, which aids in the separation of training and testing data and may be used to remove missing values. This can be obtained by using the package “Amelia”.
* The boxplot for every variable can be achieved by first establishing a data frame having all the integer types, and then obtaining the other boxplot with the outliers substituted.
* Normalization can be performed using three methods: Min-Max Scaling, Z-Score Scaling, and Soft-Max Lambda Scaling “DMwR”, with the associated boxplots shown for each.
* Kaiser-Meyer-Olkin factor adequacy test is done to eliminate the factors with high MSA (Measure of Sampling Adequacy) values to perform Regression models.
* The KS-Normality test can be done to find normality of the variable that is dependent.
* Finally, different methods of machine learning such as Linear Regression, Multi-Linear Regression, Random Forest Technique, Decision-Tree Model, Gaussian Naïve Bayes Techniques, and Support Vector Machines (SVM) are used to predict the dependent variable, and their accuracy is compared to find the best outcome.

The data exploration technique consists of Missingness map, Boxplot, Normalization, Visualization, Correlation, Normality Test, Scatterplot, and QQ Plot. They allow research and data analysts to obtain insights into their dataset’s features, detect trends, patterns, deviations, and possible predictors, and make educated judgements on preprocessing and modelling methodologies. The Missing Map assists with determining missing values and the way they are distributed, Box plots provide information on the distribution of data and the potential deviations, Normalization ensures similarity between variables with various units or levels, methods of visualization aid in visual representation of data, correlation analysis evaluates the relationship between parameters, Normality tests evaluate data distribution, Scatterplots visualize the relationships among continuous variables, and QQ Plots assimilate data as illustrated above.

## **4.3: SECTION 2: SPLITTING TRAINING AND TESTING DATA:**

The primary step for executing prediction using Machine Learning Techniques is Splitting training and testing data to train the models using appropriate algorithms.

### **4.3.1: Implementation for Splitting training and testing set:**

* The function used for partitioning is “createDataPartition” from “Caret” package from R-library.
* The partitioning's target variable is "lifeExpectancy," which is expected to be an attribute in the data frame "cirrhosis\_prediction (the dataset used in this dissertation)".
* The value has been set to 0.6, this indicates 60% of the data set will be employed for training and 40% for testing.
* The set used for training is then placed in a new data frame named "train\_data" using vector indexing, and the testing set is put in another data frames called "test\_data" with vector negative indexing. The library “ggplot” is also required to predict the accuracy plot for every model used in this dissertation.

It is common practice in the fields of machine learning and predictive modelling for splitting data into sets for testing and training. The set for training is used to train the model for prediction, while the testing set is utilized to evaluate its performance and generalizability. This prevents overfitting and allows an unbiased assessment of the model's efficiency on unknown data.

## **4.4: SECTION 3: PREDICTION USING VARIOUS MACHINE LEARNING APPROACHES:**

Machine learning prediction is a rapidly emerging discipline that uses data and algorithms to provide accurate projections or forecasts. Gathering data, preliminary processing, engineering features, training models, assessment of models, and model deployment are all steps in machine learning approaches. Accurate forecasts can provide helpful information, aid in making decisions, and improve strategy and planning development in a variety of applications, including consumer behavior, price movements in stocks, weather trends, disease diagnosis, etc. Machine learning prediction can revolutionize business sectors, drive creative thinking, and shape the prospects across many domains, and it is evolving as techniques, computing power, as well as data availability improve, providing numerous possibilities for companies and researchers alike.

In this dissertation, various algorithms including Linear Regression, Multi-Linear Regression, Random Forest Method, Decision Tree Technique, Gaussian Nave Bayes Algorithm, and SVM (Support Vector Machine) are utilized to conclude the method with better accuracy on predicting Life Expectancy of an individual having Liver Cirrhosis.

### **4.4.1: Linear Regression:**

Linear regression is a statistical method that involves fitting an equation of linear form to observed data to indicate the relationship between two variables, typically a dependent variable and numerous independent variables.

Data collection, data preparation, training of models, assessment of models, and model interpretation are all steps in the linear regression process. Cleaning and manipulating data, dealing with missing information, and identifying important features are all examples of data preparation. Model training entails estimating the coefficients of a linear equation utilizing historical data, and model evaluation evaluates the model's performance and accuracy using a variety of criteria. Model interpretation entails deciphering the parameters of the equation for linearity to estimate the amount and direction of the link between variables.

#### **4.4.1.1: Implementation:**

* Load the dataset “cirrhosis\_prediction” into R-studio and install the required packages. After splitting the training and testing data, the missing values need to be removed in order to minimize the errors by using “na.omit()” function.
* Create a model which is to be trained by using Linear regression algorithm. By using the training model, predict the testing data to provide a better accuracy.
* The evaluation of model can be done by using the library called “Metrics” which is used to calculate the RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) to conclude the accuracy of this algorithm.
* Finally, the ggplot for the linear regression model is obtained.

### **4.4.2: Multi-Linear Regression:**

Multi-linear regression is an estimating method that uses observed data to fit a linear equation to model the relationship between the variable that is dependent and more than one independent variable. It is an enhanced form of simple linear regression that uses only one variable that is independent and allows for the examination of the combined effects of multiple variables on the dependent variable. The procedure is like Linear Regression. Multiple linear regression is an effective method for modelling and estimating the relationship between several variables. It provides helpful information into the combined impacts of multiple variables on the variable that is dependent including can be applied to decision-making, prediction, and analysis of trends in a variety of disciplines.

#### **4.4.2.1: Implementation:**

* Load the dataset “cirrhosis\_prediction” into R-studio. After creating a training and testing set the model for multi-linear regression to train is created.
* To predict the data, the model with training set and testing set is used. By following the same method for linear regression, the outcome of this model can be found by analyzing the MAE and RMSE to find the accuracy range.
* Then, the ggplot for multi-Linear regression is plotted.

### **4.4.3: Random Forest Algorithm:**

The Random Forest approach is an ensemble learning method that is employed in machine learning for regression and classification tasks. It expands the idea of decision trees, that are basic binary tree-like structure used for decision-making, by constructing several decision trees and merging their projections in order to enhance accuracy and robustness.

A Random Forest is a collection of decision trees formed by selecting at random from a subset of the original training data and a subset of features at each tree split. This randomization in data and feature selection aids in the reduction of overfitting, known to be a typical problem in decision trees. To make the final forecast, the projected outcomes of individual tree are integrated employing approaches that include average (for regression) or casting (for classification).

The Random Forest technique has several benefits, like high precision, stability to inaccurate information, and resistant to overfitting. It can handle big datasets with high-dimensional characteristics and capture complicated patterns and variable interactions. It is also faster in both training and forecasting than some other complicated algorithms.

#### **4.4.3.1: Implementation:**

* Load the dataset “cirrhosis\_prediction” into R-studio. To perform the Random Forest Algorithm the corresponding library “randomForest” needs to be installed.
* After splitting training and testing data, the missing values are removed by using the function “na.omit()” to improve the models efficiency.
* The model for random forest is created and is predicted with testing data. The ggplot is provided for the respective model and the dependent variable which is “lifeExpectancy”.

**4.4.4: Decision-Tree Model:**

The Decision Tree algorithm is a well-known and conventional machine learning approach for tasks such as regression and classification. It is a structure resembling a tree that uses the numerical values of the input characteristics to create judgements or predictions.

The dataset is divided into subsets in a Decision Tree according to the data points of one or more attributes at every node of the tree. This splitting procedure is repeated until a stopping requirement is satisfied, like reaching a maximum depth, a particular level of purity in the subsets, or a preset number of samples in the leaf nodes. Following the path from the tree's root to a leaf node that represents most of the class (for classification) or the expected value (for regression), the judgement or prediction is made.

#### **4.4.4.1: Implementation:**

* Load the dataset “cirrhosis\_prediction” into R-studio. Splitting the train and test set, and to fit the decision tree regression into the dataset the library called “rpart” is installed.
* The tree model is created with training set and the prediction of the trained model is provided with comparing the testing set by using the function “class”.
* Finally, the ggplot for the trained model is obtained and the corresponding tree model can be obtained by using the package “rpart.plot”.

### **4.4.5: Gaussian Naïve Bayes Algorithm:**

The Gaussian Naive Bayes technique, which is implemented in R, is a common probability algorithm for classification used in machine learning. It depends on the Naive Bayes theorem, and it requires the characteristics are independently determined provided the class label and that the feature distribution is Gaussian (normal).

The training data is utilized in the Gaussian Naive Bayes method to determine the standard deviation and mean of the attributes for each class label. These estimations are then used to construct a probabilistic model capable of predicting the class label of additional data points. The algorithm computes the probabilities of a data point being associated with a specific class label according to the distribution of Gaussian variables and combines it with the class label's prior probability to compute the posterior probability. The anticipated class label is subsequently allocated to the class label having the greatest posterior probability.

#### **4.4.5.1: Implementation:**

* Load the dataset “cirrhosis\_prediction” into R-studio. After splitting the training and testing data, the naïve bayes algorithm can be done by using the library “e1071”.
* The naïve bayes model can be trained by using the training set and prediction is done by comparing the testing set.
* The bar plot for the actual and predicted proportion is executed to analyze the accuracy of the model. Here, the bar plot is used because of the proportion is continuous.

### **4.4.6: Support Vector Machines (SVM):**

SVMs are a sophisticated and frequently used machine learning technique for tasks such as classification and regression. The technique of SVM is implemented in R in the "e1071" package, containing methods for creating and testing SVM models.

SVM is a binary classifier algorithm that constructs a hyperplane in a space with high dimensions that best differentiates data points of distinct classes. The main notion of SVM is to maximize the margin between classes, that is, support vectors are the distances among a hyperplane and the nearest data points from each class. SVM can handle both linearly and nonlinearly separable datasets by employing various kernels, including polynomial, linear, and the radial basis function (RBF) kernels, that enable complex decision boundaries.

SVM can be implemented in R by calling the "SVM" method from the "e1071" package. The function provides for the tuning of several parameters, including the kind of kernel, the cost factor for regularization, and kernel parameters (for example, degree for polynomials kernel, sigma for RBF kernel). The trained SVM technique can now be used to forecast new data points.

#### **4.4.6.1: Implementation:**

* Load the dataset “cirrhosis\_prediction” into R-studio. The training and testing set is created by using the function “createDataPartition” from the library “caret”.
* The SVM model is trained by using appropriate algorithm with training set and the predictions is done among the trained model and the testing set.
* The ggplot is obtained for the model to predict the accuracy respectively.

## **4.5: SECTION 4: SURVIVAL ANALYSIS BY USING MACHINE LEARNING TECHNIQUES:**

The term "survival analysis" relates to a collection of statistical techniques for data analysis where the time until an event occurs is the result parameter of interest.

To determine the survival analysis, Cox proportional hazard (Cox PH) and Kaplan-Meier model are commonly utilized. According to recent research, the Ordinal Differential Equation (ODE) modelling framework integrates numerous existing Survival analysis techniques such as Cox PH and AFT. They also demonstrated that the ODE modelling framework is adaptable and broadly applicable. The logRank test is evaluated to find the difference between two methods to conclude the recovery time for Liver Cirrhosis. This approach provides several advantages. 1)The model is efficient in terms of computation.2) Theoretically, the model is sound.3) It is simple to implement.4) The model can be used to any Survival analysis problem using censored data.

### **4.5.1: Kaplan Survival Curve:**

The Kaplan-Meier Survival approach is estimated to find the approximate probability of recovery time among the individuals having Liver Cirrhosis.

### **4.5.2: Cox Proportional Hazard Model:**

The cox proportional hazard model is employed to obtain the Hazard ratio and log-Rank test value to find the survival analysis.

### **4.5.3: QQ-Plot for the survival model:**

The survival model is created, and the survival probability analysis is done by using the above methods. Thus, the Normalization for the survival model can be created by using Weibull and Flexible Parametric Model to interpret the comparison between continuous and uniform distribution in the survival model.

### **4.5.4: ROC curve:**

The ROC curve is also known as Log-Rank which is used to find the difference between Kaplan-Meier and Cox-Hazard Proportional approach to find the approximate Recovery time (survival analysis) from the risk of Liver Cirrhosis in an individual.

### **4.5.5: Implementation for Survival Analysis:**

* Load the dataset “cirrhosis\_prediction” into R-studio and load the required libraries to run survival analysis such as “survival” and “survminer”.
* The proportional ratio can be obtained by using the “survival” function and the survival model is created. And the Survival analysis is done by using the two models and the q-q plot is obtained for the survival model created for visualization.
* Then, the ROC curve can be obtained by loading the library “ROCR” to find the approximate recovery time for an individual.

## **4.6: CONCLUSION:**

Finally, the use of Linear Regression, Multi-Linear Regression, the Random Forest algorithm, the Decision Tree approach, the Gaussian Naive Bayes strategy, and SVM (Support Vector Machines) techniques yielded encouraging results in estimating life expectancy based on the given data. Furthermore, the Decision Tree technique achieves an overall positive performance range of 85.07%, which is higher than the other models. The models may need to be evaluated and fine-tuned further to increase their accuracy and generalization performance. Additionally, the application of Kaplan-Meier survival evaluation and the Cox proportional hazard approach yielded useful insights into the analyzed time-to-event data. The ROC map shows that the therapy was effective in lowering the risk of the liver cirrhosis, especially after 3000 days, indicating a favorable outcome. These methodologies can be an effective tool for predicting and analyzing life expectancy and survival probability in various groups, with significant consequences for public health as well as policy decisions.

# **CHAPTER 5: RESULTS AND DISSCUSSION:**

## **5.1: INTRODUCTION:**

The results of the above sections are illustrated, and the outcomes is discussed accordingly. The accuracy for the various machine learning approaches is estimated and the method which provides better accuracy can be compared and discussed.

## **5.2: SECTION 1: DATA EXPLORATION:**

In this dissertation, various machine learning techniques are employed. To predict the accuracy the data exploration is necessary (i.e., the Data Cleaning).

### **5.2.2: Results and Interpretation:**

**Missingness Map:**

Graphical user interface, text

Description automatically generated

**Figure 5. 1: Missingness Map**

* To interpret the illustrated missingness map above, the lines in black shows that there are many missing values in the dataset which may affect the outcomes of the prediction.
* To handling the missing values various methods can be used. One of the methods used here is filling the missing columns with “mean”. This can be achieved in R, by using the function “mutate\_all” which comes under the library “tidyr”.

Text

Description automatically generated

**Figure 5.1 1: Missingness map (After filled missing columns)**

* The above missingness map is obtained after filled the missing columns with the “mutate\_all” function to overcome overfitting.

**Boxplot:**

Chart, box and whisker chart

Description automatically generated

**Figure 5. 2: Boxplot for all variables.**

* The above boxplot is obtained by putting the integers into a data frame “cirrhosis\_prediction\_2” which is used in the further process of Data Exploration to normalize the attributes.
* From the above boxplot the Y-axis shows the dependent and independent variables, and X-axis shows the count to analyze the distribution of the attributes.

Chart, box and whisker chart

Description automatically generated

**Figure 5.2 1: Boxplot for the outliers replaced.**

* The above boxplot is obtained after the outliers are replaced for the numerical values according to the quantile values (0.05, 0.95). The X-axis represents the outliers replaced for the numeric values and Y-axis represents the count respectively.

**Examine Boxplot for Life Expectancy with Sex and No. of Days:**

***Chart, box and whisker chart

Description automatically generated***

**Figure 5. 3: Boxplot between lifeExpectancy and Sex**

* The above Boxplot shows the Life Expectancy by sex. The X-axis shows sex (Female and Male), whereas the Y-axis shows Life Expectancy.

***Chart

Description automatically generated with low confidence***

**Figure 5. 4: Boxplot between lifeExpectancy and N\_Days**

* The above Boxplot is obtained between lifeExpectancy and N\_Days. The X-axis shows No-of Days the individual having Liver Cirrhosis and the Y-axis shows the Life Expectancy accordingly. It shows the average life expectancy lies between 70 and 80 as average.

**NORMALIZATION (3 METHODS):**

1. **Min-Max Scaling:**

**Chart, box and whisker chart

Description automatically generated**

**Figure 5.5: Min Max Scaling**

* Normalization is a technique that need to be done for an independent variable to be normalized rather than the dependent variable. The Normalization is obtained for the numerical attributes in the dataset. The min max scaling yields less normal distribution between the variables, the various methods like Z-score and SoftMax Scaling is also used to make the variables to be normally distributed in similar scale.
* The points in the above Figure represents the data points and the line in the center of each attribute plot is the median of the data set used.
* The X-axis in the above figure represents the numeric values and the Y-axis represents the count respectively.

1. **Z-Score Scaling:**

**Chart, box and whisker chart

Description automatically generated**

**Figure 5. 6: Z-Score Scaling (Z=1)**

* The z-score scaling is obtained for the value of Standard Deviation (Z=1). Here, the values can be changed to obtain desired scaling between the variables.
* To analyze, the value for standard deviation (Z=2) is also obtained and illustrated below.

Chart, box and whisker chart

Description automatically generated

**Figure 5.6 1: Z-Score Scaling (Z=2)**

* The X-axis for the above Figures represents the numeric values and the Y-axis represents the count. The Normalization is slightly improved for the standard deviation (Z=2). For obtain more approximate Normalization between the variables the Soft-Max Lambda method is also employed and illustrated as follows.

1. **Soft Max Scaling:**

***Chart, bar chart

Description automatically generatedChart, box and whisker chart

Description automatically generated***

**Figure 5.7 1: SoftMax Scaling (Lambda=1) Figure 5.7 2: SoftMax Scaling (Lambda=2)**

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

**Figure 5.7 3: SoftMax Scaling (Lambda=3) Figure 5.7 4: SoftMax Scaling (Lambda=4)**

Chart, box and whisker chart

Description automatically generated

**Figure 5.7 5: SoftMax Scaling (Lambda=10)**

* The above illustrated boxplot is obtained for the normalization using Soft Max Scaling which is an efficient method of obtaining the distribution between the variables.
* The Soft Max Lambda is value of Scaling which gives an exact distribution of variables.
* To make the variables as similarly scaled the different values for lambda such as 1,2,3, 4…10 can be employed until the distribution looks normal.
* The X-axis for the above figures represents the Numerical values and the Y-axis represents count.
* The variables are found to normalized for the lambda value 10.

**VISUALIZATION:**

**Diagram

Description automatically generated**

**Figure 5.8 1: Histogram for all variables**

* The histogram is a tool for the visualization of a variable distribution. It can be obtained by the function “pairs”.
* The above Figure illustrates the distribution of the numerical values in the data frame in the data set “cirrhosis\_prediction”.

**CORRELATION:**

* To perform the Correlation analysis, the matrix for correlation is to be created and the values are rounded to the nearest as illustrated below.

A screenshot of a computer

Description automatically generated

* The above correlation matrix represents the positive and negative correlation for the numerical variables.

**Pearson’s Correlation:**

Chart

Description automatically generated with medium confidence

**Figure 5.10 1: Pearson Correlation**

* Pearson’s Correlation method is used to obtain the correlation map for the Patients with Liver Cirrhosis.
* The above Figure shows the variables that are having Positive and Negative Correlation which is used to further analysis like partial correlation and so on.
* The “red” in the above figure represents the variables having Negative correlation and the variables with “blue” is positively correlated.
* However, the above Figure represents both Positive and Negative correlation of each variable.

**Partial Correlation using Pearson’s Method:**

A screenshot of a computer

Description automatically generated with medium confidence

* The Pearson’s Partial correlation is obtained above p-value (value of correlation coefficient) is 0.917 and 0.369 which suggests a strong and positive correlation. i.e., the p-value > 0.75 is suggested to have a strong and positive correlation.

**G-G Plot:**

***Chart, scatter chart

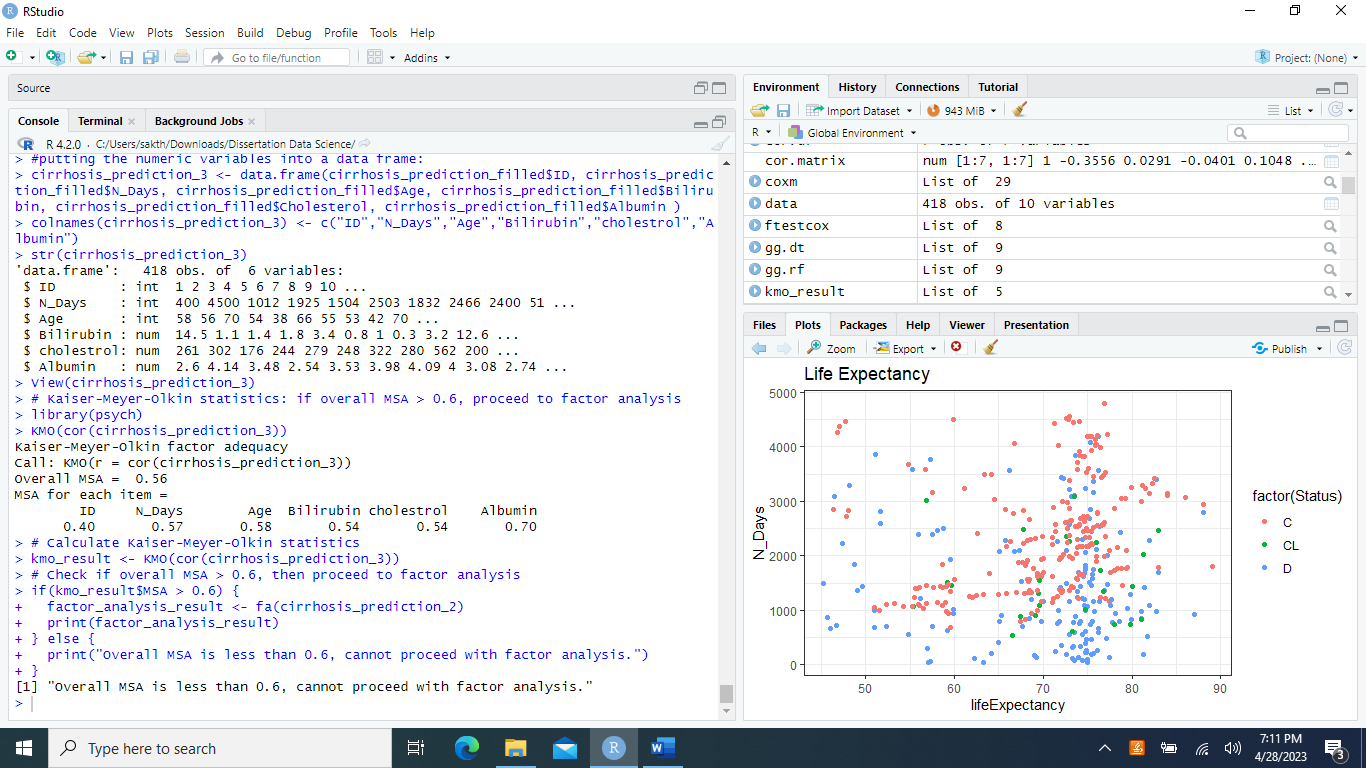
Description automatically generated***

**Figure 5.8 2: G-G Plot between lifeExpectancy and N\_Days**

* The above G-G Plot is obtained by using the package “ggplot”. The G-G plot is obtained for the factor: Status whereas, C represented by “red” which is Censored, CL represented by “green” which is Censored due to Liver damage and D represented by “blue” which is Death.

**Kaiser-Meyer-Olkin Statistics:**

* The Kaiser-Meyer-Olkin Statistics is performed as the primarily step for predicting life expectancy using Regression models such as Linear Regression and Multi-Linear Regression.
* The significance of performing KMO test is to analyze the factor through MSA (Mean Sampling Adequacy) value.
* To perform KMO test, the data frame is created with the required numerical values which excludes lifeExpectancy as it is a dependent variable.



* The KMO test is illustrated above and the obtained overall MSA value is 0.56. The KMO test is performed to evaluate the Mean Sampling Adequacy among the factors. If the MSA value is greater than 0.6, then the factor analysis needs to be proceeded to retaining the factors with high eigen values and the scree plot needs to be obtained.
* In the other case, if the MSA value is less than 0.6, then the factor analysis cannot be proceeded as the factors having efficient adequacy to perform Regression models.
* In the above result, the overall MSA is obtained to be 0.56, which is less than 0.6. Hence, the Factor analysis cannot be proceeded, as the factors having efficient adequacy to perform Regression models such as Linear Regression and Multi-Linear Regression.

**Kolmogorov-Smirnov Normality Test:**

In order to perform Linear Regression, the dependent variable should be normally distributed. To check the Normality of the variable, the Kolmogorov-Smirnov test is performed.

A screenshot of a computer

Description automatically generated with medium confidence

* The D (test statistic) is 0.5429, and the value of p (chance of detecting a test statistic) is less than the 0.05 level of significance.
* This result implies that the null hypothesis is rejected and may conclude that the actual distribution of the lifeExpectancy variable is equal to its normal distribution, which can be used to perform regression models.

**SCATTERPLOT:**

**Chart, scatter chart

Description automatically generated**

**Figure 5.12 1: Scatterplot**

* The scatterplot is obtained above for the variable Life Expectancy and the Bilirubin level to visualize the distribution among them.
* The datapoints in the above scatterplot is likely to be distributed normally. The X-axis represents Life Expectancy, and the Y-axis represents the level if Bilirubin.

Chart, scatter chart

Description automatically generated

**Figure 5.12 2: Scatterplot with Regression line**

* The regression lined scatterplot is shown above describes the linear relationship between two variables (lifeExpectancy and Bilirubin).

**Quantile-Quantile Plot (Q-Q plot):**

* In order to perform Linear Regression, the dependent variable (lifeExpectancy) needs to be normally distributed. To check if the variable is normally distributed or not, the q-q plot is illustrated.
* If the data points are stable after some extent with the regression line, then the variable is said to be normally distributed.

Chart

Description automatically generated

**Figure 5.13.1: Normal Q-Q Plot**

* The above Figure shows the normal q-q plot for the dependent variable. The regression line is added to get a clear interpretation of the distribution as follows.

Chart, line chart

Description automatically generated

**Figure 5.13.2: Normal Q-Q Plot with regression line.**

* The data points illustrated above, is said to be stable after some extent with regression line. Hence, the dependent variable (lifeExpectancy) is said to normally distributed and can be utilized to perform Linear Regression.

## **5.3: SECTION 2: SPLITTING TESTING AND TRAINING SET:**

In order to perform Machine Learning methods, the data set needs to be split into two sets which is training set and testing set. The value has been set to 0.6, this indicates 60% of the data set will be employed for training and 40% for testing. The set used for training is then placed in a new data frame named "train\_data" using vector indexing, and the testing set is put in another data frames called "test\_data" with vector negative indexing.

* The 60% of the dataset is employed as training set in which the algorithm is used to train the model and the 40% of the dataset is employed as testing set, in which the trained model is compared with the testing set to obtain the difference among the actual values and predicted values.
* Hence the set for training and testing is created the missing values in these sets is omitted by using the function “na.omit()”.

## **5.4: SECTION 3: PREDICTION USING MACHINE LEARNING TECHNIQUES:**

### **5.4.1: Linear Regression:**

Linear regression is a statistical method that involves fitting an equation of linearity to observed data to indicate the relationship between two variables, typically a dependent variable and numerous independent variables.

#### **5.4.1.2: Result:**

Chart, scatter chart

Description automatically generated

**Figure 5.14 1: Linear Regression**

* The training model is trained using Linear Regression algorithm, and compared with the testing set and the p-value and statistics is obtained. The intercept is fit to 50 and the slope is fit to be 0.4 to obtain the gg plot.
* The p-value i.e., probability of obtaining F-statistic, which is as extreme than one determined, provided that the null hypothesis being tested (that each of the coefficients are equal to zero) is correct.
* The p-value in this case is 0.02635, which is below than the 0.05 significance level, showing that the model is highly significant.

Graphical user interface, text, application

Description automatically generated

* The Evaluation metrics (MSE) Mean Squared Value is used to calculate the accuracy for the Linear Regression Model.
* Hence, the accuracy for the Linear Regression is obtained to be 79.93, which seems to be a significant and efficient model to be used.

### **5.4.2: Multi-Linear Regression:**

Multiple linear regression is an estimation method that fits a linear equation to observed data to model the connection among the dependent variable and more than one independent variable.

#### **5.4.2.2: Result:**

**Chart, scatter chart

Description automatically generated**

**Figure 5.15 1: Multi-Linear Regression**

* As the training model is trained using the multi-linear regression algorithm, and then it is compared with the testing set to obtain p-value and statistics, which results in a plot containing actual and predicted values of Life Expectancy.
* The intercept is fit to 50 and the slope is fit to be 0.4 to obtain the gg plot.
* The p-value i.e., probability of obtaining F-statistic, which is as extreme than one determined, provided that the null hypothesis being tested (that each of the coefficients are equal to zero) is correct.
* The p-value in this case is 0.02635, which is below than the 0.05 significance level, showing that the model is highly significant.
* The accuracy using MSE evaluation metrics is illustrated below.

A screenshot of a computer

Description automatically generated with medium confidence

* Hence, the accuracy for the multi-linear Regression is obtained to be 79.93, which seems to be a significant and efficient model to be used.

### **5.4.3: Random-Forest Technique:**

The Random Forest method is a method for ensemble learning used in machine learning for regression and classification tasks.

#### **5.4.3.2: Result:**

Chart, scatter chart

Description automatically generated

**Figure 5.16 1: Random Forest Regression**

* The training model created is trained by using random forest algorithm and compared with the testing set to obtain the p-value. The probability of estimation is 0.2635 which is highly significant.
* Then, the trained model is fitting using random forest algorithm and the gg-plot is obtained for the actual and predicted life expectancy values with regression line.

A screenshot of a computer

Description automatically generated with medium confidence

* By performing the above tasks, the accuracy for random forest is obtained using the MSE evaluation metrics which yields the value of 71.16, which seems to be a stable model to be used.

### **5.4.4: Decision-Tree Algorithm:**

The Decision Tree algorithm is a well-known and conventional machine learning approach for tasks such as regression and classification. It is a structure resembling a tree that uses the numerical values of the input characteristics to create judgements or predictions.

#### **5.4.4.2: Result:**

Chart, scatter chart

Description automatically generated

**Figure 5.17 1: Decision Tree Regression**

* The training model created is trained by using decision tree algorithm using the package “rpart” and compared with the testing set to obtain the p-value. The probability of estimation is 0.2635 which is highly significant.
* Then, the trained model is fitting using decision tree algorithm and the gg-plot is obtained for the actual and predicted life expectancy values with regression line.

Graphical user interface, application

Description automatically generated

* The mean squared difference between expected and actual values is referred to as the MSE. The lower the MSE score, the more effectively the model matches the data.
* The average deviation of the residuals is represented by RMSE, which is the square root of MSE. It indicates how far the expected values differ from the actual ones. A lower RMSE shows that the model is more accurate.
* The accuracy of this model is said to be 87.73% which is highly efficient model to be used.

The tree plot is shown below for the respective predictive model.

Diagram, schematic

Description automatically generated

**Figure 5.17 2: Tree Plot**

* The tree plot can be obtained by using “rpart.plot” for the trained model is illustrated which shows the distribution of the actual values in each nodes.

### **5.4.5: Gaussian Naïve Bayes Algorithm:**

The Gaussian Naïve Bayes approach, which is implemented in R, is a common probability algorithm for classification used in machine learning.

#### **5.4.5.2: Result:**

Chart, bar chart

Description automatically generated

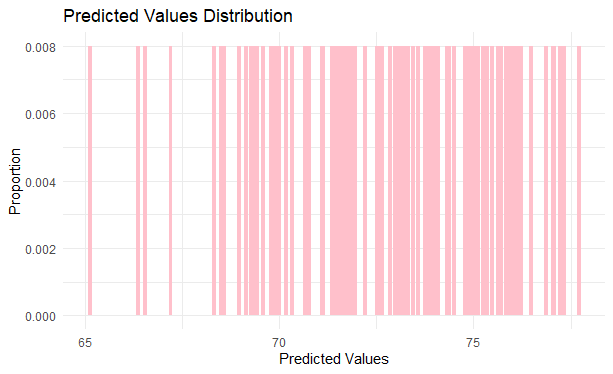
**Figure 5.18 1: Gaussian Naïve Bayes Algorithm**

* The nb\_model is created with the training data and testing data, and it is trained by using Naïve Bayes algorithm and the model is fitted using the Naïve Bayes algorithm using the package “e1071”.
* Then, the trained is used to predict the values by using the testing set.
* Then, the ggplot is obtained in the form of bar plot as it is a continuous distribution function.
* From the above bar plot, the X-axis represents the predicted values and Y-axis represents the proportion of distribution.
* Hence, the accuracy for Gaussian Naïve Bayes technique is obtained as 73.8% which is efficient to be used.

### **5.4.6: Support Vector Machine (SVM):**

SVMs are a sophisticated and frequently used machine learning technique for tasks such as classification and regression. The technique of SVM is implemented in R in the "e1071" package, containing methods for creating and testing SVM models.

#### **5.4.6.2: Result:**



**Figure 5.19 1: SVM Algorithm**

* The Svm\_model is created by using the training set and is trained by using SVM algorithm.
* The Svm\_prediction is done by fitting the created model and comparing the model with testing set using the package “e1071”.
* Then, the ggplot is obtained in the form of bar plot as it is a continuous distribution function. From the above bar plot, the X-axis represents predicted values and the Y-axis represents Proportion of distribution.

A screenshot of a computer

Description automatically generated with medium confidence

* Hence, the accuracy for the SVM model is 72.39 which is stable and significant to be used.

## **5.5: SECTION 4: SURVIVAL ANALYSIS:**

### **5.5.1: Kaplan-Meier Survival Test:**

Chart, line chart

Description automatically generated

**Figure 5.20 1: Survival probability (Kaplan-Meier)**

* The survival model is created by using dummy variables which approximately describes to fit the survival model using Kaplan-Meier Survival test.
* Then, the survival probability is obtained as illustrated above, in which the X-axis represent Time and Y-axis represents Survival Probability.
* From the above Figure, the “red” indicates Male and the “blue” indicates Female.

Chart

Description automatically generated

**Figure 5.20 2: Kaplan-Meier Curve**

* The Kaplan-Meier Curve explains the survival probability between sex (Male and Female). Here, the P-value (Probability of estimation) is obtained as 0.086 which is strongly and positively distributed as it is greater than 0.05.
* The X-axis represents Time and Y-axis represents the Survival Probability. Also, the Number at risk is obtained among Male and Female. The “Orchid” indicates Female and “black” indicates Male.

### **5.5.2: Cox-Hazard Proportional Approach:**

The obtained outcome for survival analysis using Cox-Hazard Proportional model is illustrated below.

Chart, box and whisker chart

Description automatically generated

**Figure 5.21 1: Hazard Ratio**

* The obtained Log-Rank (p-value) is approximately 6.6751e-16. Further, the Global Schoenfield Test is done in order to analyze every attributes as follows.

A picture containing calendar

Description automatically generated

**Figure 5.22 1: Global Schoenfield Test**

* The p-value (probability of estimation) for the global schoenfield test of each variable is 0.000814.

### **5.5.3: QQ- PLOT FOR THE SURVIVAL MODEL:**

Chart, histogram, scatter chart

Description automatically generated

**Figure 5.23 1: QQ-Plot using Weibull and Flexible Parametric Model.**

* The above survival model is created to find the recovery time for an individual who have Liver Cirrhosis.
* The QQ plot can also obtained by using continuous distribution as a comparison such as Weibull distribution or a uniform distribution such as (Flexible Parametric distribution) which is obtained.
* The X-axis represent Time of individuals having Liver Cirrhosis and the Y-axis represents the survival Probability accordance with lifeExpectancy.

### **5.5.4: ROC CURVE:**

Chart, line chart

Description automatically generated

**Figure 5.24 1: ROC curve for predicting recovery time.**

* ROC curve is used to find the logRank test which is a difference among the two-survival analysis test to obtain the approximate result.
* The ROC plot indicates that the risk is being drastically decreased after 3000 days. The key reason for the short survival time were because the individual arrived for therapy at an advanced stage, even with diseases (such as diabetes, anemia, and hypertension).

|  |  |
| --- | --- |
| **MODELS** | **ACCURACY** |
| Linear Regression | 79.93% |
| Multi-Linear Regression | 79.93% |
| Random-Forest Approach | 71.16% |
| Decision Tree Algorithm | 87.73% |
| Gaussian Naïve Bayes Technique | 73.80% |
| Support Vector Machines (SVM) | 72.39% |

**Figure 5.25 1: Table of Models and Accuracy**

## **5.6: CONCLUSION:**

The accuracy of every model is executed and compared to find the better model. The Linear Regression approach has an accuracy of 79.93%. Because the Multi-Linear Regression approach utilized in this study is identical to Linear Regression, the accuracy is similarly comparable, which is advantageous. The Random Forest approach has an accuracy of 71.16%. Random Forest is an experimental strategy for making predictions that incorporates numerous decision trees. It can handle non-linearity and feature interactions, which is useful for forecasting life expectancy. However, the accuracy gained in this dissertation may imply that the model requires further optimization and fine-tuning to increase its performance from 71.16%. Linear Regression is a straightforward and easy-to-understand technique for predicting life expectancy according to supplied variables. The Decision Tree algorithm has an accuracy of roughly 87.73%. Decision Tree algorithm is a non-linear approach for capturing complex feature connections. The excellent accuracy found in this dissertation indicates the Decision Tree algorithm is capable of effectively forecasting life expectancy according to the features provided. The Gaussian Nave Bayes method has an accuracy of 73.80%. Nave Bayes is a method of probability based on the assumption that characteristics are normally distributed. Despite its simple assumptions, it can yield reasonable accuracy in some instances, as proved by this dissertation. The SVM model has an accuracy of about 72.39%. SVM is a powerful technique for dealing with high-dimensional data and capturing complicated patterns. However, the study's significantly lower accuracy may indicate that further optimization and parameter adjustment may be required to increase its effectiveness.

In order to find the recovery time for an individual from liver cirrhosis is obtained by utilizing the Kaplan-Meier Survival analysis and Cox-Hazard Proportional Approach, the ROC curve is obtained, and the risk is likely to be decreased after 3000 days.

Moreover, this dissertation examined numerous machine learning algorithms to predict life expectancy according to the features that were provided. The models' accuracy varied, with the Decision Tree model scoring the highest (87.73%), subsequently followed by Linear Regression (79.93%), Gaussian Nave Bayes (73.80%), SVM (72.39%), and Random Forest (71.16%). The models may require additional evaluation, optimization, and fine-tuning to increase their accuracy and generalization performance. The ROC map obtained from the Log-Rank test which is a difference between Kaplan-Meier Survival test and Cox-Hazard Proportional approach shows that the therapy was effective in lowering the risk of the liver cirrhosis, especially after 3000 days, indicating a favorable outcome. These findings emphasize the predictive power of various machine learning strategies, the consequences for the public's health and policy concerns.

# **CHAPTER 6: CRITICAL EVALUATION AND CONCLUSION:**

## **6.1: Critical Evaluation:**

This dissertation used various machine learning approaches to predict life expectancy according to features presented. When the performance of each model was executed, the Decision Tree model came out on top with 87.73% accuracy. Accuracy rates for linear Regression, Gaussian Naïve Bayes, SVM, and Random Forest methods were 79.93%, 73.80%, 72.39%, and 71.16%, correspondingly. In addition, the Kaplan-Meier analysis of survival and the Cox-Hazard Proportional Method were employed in the study to analyze the recovery period for individuals having liver cirrhosis, and the ROC curve obtained from the Log-Rank test revealed that the medication was effective after 3000 days.

## **6.2: Experience and Knowledge Acquired:**

Moreover, from this dissertation, I have gained knowledge of how the research must be done and reviewing literature on different aspects. I have acquired a better knowledge in R-Programming and Machine Learning techniques which helps me to achieve my goals in future. Also, the survival analysis methods (Kaplan-Meier and Cox-Proportional methods) are very useful which I learned from doing this dissertation. I have also acquired knowledge about the stages, causes, symptoms, and diagnostics methods of Liver cirrhosis which is useful for health and policy for public consequences.

## **6.3: Future Studies and Scope:**

In Future studies, these methods can be further used to make a positive performance with better efficiency by tuning and scaling into different intercepts and different methods in machine learning also will be used to gain better efficiency. It can also aid with medical research by finding significant characteristics that impact life expectancy in patients with liver cirrhosis, ultimately enhancing patients results and the standard of care.

## **6.4: Conclusion:**

In conclusion, this dissertation used various machine learning approaches to predict life expectancy according to features presented. Further optimization and fine-tuning of the models may be necessary to increase their performance. Furthermore, these findings highlight the predictive power of various machine learning algorithms, the consequences for health and policy for the public decisions.

# **REFERENCE:**

Arshad, I. and Dutta, C. (2016) ‘Liver injury due to herbal and dietary supplements: A review of individual ingredients’, *Clinical Liver Disease*, 7(4), pp. 80–83. Available at: https://doi.org/10.1002/cld.541.

Bourlière, M. *et al.* (2015) ‘Ledipasvir-sofosbuvir with or without ribavirin to treat patients with HCV genotype 1 infection and cirrhosis non-responsive to previous protease-inhibitor therapy: a randomised, double-blind, phase 2 trial (SIRIUS)’, *The Lancet Infectious Diseases*, 15(4), pp. 397–404. Available at: https://doi.org/10.1016/s1473-3099(15)70050-2.

Chalasani, N. *et al.* (2017) ‘The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases’, *Hepatology*, 67(1), pp. 328–357. Available at: https://doi.org/10.1002/hep.29367.

Chetty, R. *et al.* (2016) ‘The Association Between Income and Life Expectancy in the United States, 2001-2014’, *JAMA*, 315(16), p. 1750. Available at: https://doi.org/10.1001/jama.2016.4226.

Ching, T. *et al.* (1981) ‘Opportunities and Obstacles for deep learning’, *American Journal of Public Health*, 71(12), p. 1410. Available at: https://doi.org/10.2105/ajph.71.12.1410.

Cunningham, P., Cord, M. and Delany, S.J. (2008) ‘Supervised Learning’, *Machine Learning Techniques for Multimedia*, 13(23), pp. 21–49. Available at: https://doi.org/10.1007/978-3-540-75171-7\_2.

D’Amico, G. *et al.* (2007) ‘Study of Liver Cirrhosis’, *Gastroenterology*, 132(3), pp. 1202–1204. Available at: https://doi.org/10.1053/j.gastro.2007.02.013.

Denton, E.L. and Birodkar, vighnesh (2017) ‘Unsupervised Learning of Disentangled Representations from Video’, *Neural Information Processing Systems* [Preprint].

Emori, C.T. *et al.* (2020) ‘Observational Study Evaluating the Outcome of Cirrhotic Hepatitis C Patients Submitted to Renal Transplantation’, *Transplantation Proceedings*, 52(1), pp. 89–96. Available at: https://doi.org/10.1016/j.transproceed.2019.10.022.

Ginès, P. *et al.* (2021) ‘Liver cirrhosis’, *The Lancet*, 398(10308), pp. 1359–1376. Available at: https://doi.org/10.1016/S0140-6736(21)01374-X.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. MIT Press.

Hastie, T., Tibshirani, R. and Friedman, J. (2008) ‘Overview of Supervised Learning’, *The Elements of Statistical Learning*, 8(4), pp. 9–41. Available at: https://doi.org/10.1007/978-0-387-84858-7\_2.

Kanwal, F. *et al.* (2020) ‘Development, Validation, and Evaluation of a Simple Machine Learning Model to Predict Cirrhosis Mortality’, *JAMA Network Open*, 3(11), p. e2023780. Available at: https://doi.org/10.1001/jamanetworkopen.2020.23780.

Kiran, B.R. *et al.* (2021) ‘Deep Reinforcement Learning for Autonomous Driving: A Survey’, *IEEE Transactions on Intelligent Transportation Systems*, 12(24), pp. 1–18. Available at: https://doi.org/10.1109/tits.2021.3054625.

Lai, M. and Afdhal, N.H. (2019) ‘Liver Fibrosis Determination’, *Gastroenterology Clinics of North America*, 48(2), pp. 281–289. Available at: https://doi.org/10.1016/j.gtc.2019.02.002.

Lecouat, B. *et al.* (2018) ‘Semi-Supervised Learning with GANs: Revisiting Manifold Regularization’, *arXiv:1805.08957 [cs, stat]*, 6(8). Available at: https://arxiv.org/abs/1805.08957.

Li, K. *et al.* (2020) ‘Measuring Corporate Culture Using Machine Learning’, *The Review of Financial Studies*. Edited by I. Goldstein, 6(8). Available at: https://doi.org/10.1093/rfs/hhaa079.

Lim et al, S.S. (2015) ‘World Journal of Gastrointestinal Surgery ’, *Baishideng Publishing Group* [Preprint].

Liu, S.-Y., Tsai, I.-T. and Hsu, Y.-C. (2021) ‘Alcohol-Related Liver Disease: Basic Mechanisms and Clinical Perspectives’, *International Journal of Molecular Sciences*, 22(10). Available at: https://doi.org/10.3390/ijms22105170.

Ma, J. *et al.* (2015) ‘Sugar-sweetened beverage, diet soda, and fatty liver disease in the Framingham Heart Study cohorts’, *Journal of Hepatology*, 63(2), pp. 462–469. Available at: https://doi.org/10.1016/j.jhep.2015.03.032.

Mansour, D. and McPherson, S. (2018) ‘Management of decompensated cirrhosis’, *Clinical Medicine*, 18(Suppl 2), pp. s60–s65. Available at: https://doi.org/10.7861/clinmedicine.18-2-s60.

Masters, R.K., Hummer, R.A. and Powers, D.A. (2017) ‘Educational Differences in U.S. Adult Mortality’, *American Sociological Review*, 77(4), pp. 548–572. Available at: https://doi.org/10.1177/0003122412451019.

Mejri, N. *et al.* (2023) ‘Unsupervised Anomaly Detection in Time-series: An Extensive Evaluation and Analysis of State-of-the-art Methods’, *arXiv:2212.03637 [cs]*, 12(4). Available at: https://arxiv.org/abs/2212.03637.

Moon, A.M., Singal, A.G. and Tapper, E.B. (2019) ‘Contemporary Epidemiology of Chronic Liver Disease and Cirrhosis’, *Clinical Gastroenterology and Hepatology*, 18(12). Available at: https://doi.org/10.1016/j.cgh.2019.07.060.

Naveau, S., Perlemuter, G. and Balian, A. (2005) ‘[Epidemiology and natural history of cirrhosis]’, *La Revue Du Praticien*, 55(14), pp. 1527–1532. Available at: https://pubmed.ncbi.nlm.nih.gov/16255293/.

Oeppen, J. and Vaupel, J.W. (2002) ‘DEMOGRAPHY: Enhanced: Broken Limits to Life Expectancy’, *Science*, 296(5570), pp. 1029–1031. Available at: https://doi.org/10.1126/science.1069675.

Organization, W.H. (2019) ‘World Health Statistics Annual, 1995.’, *Biometrics*, 53(4), p. 1564. Available at: https://doi.org/10.2307/2533537.

Pedro, D. (2012) ‘A Few Useful Things to Know About Machine Learning’, *Research Gate*, 7(9).

P.S, K. (2021) *Liver Transplantation*. John Wiley & Sons.

Rajkomar, A. *et al.* (2018) ‘Scalable and accurate deep learning with electronic health records’, *npj Digital Medicine*, 1(1). Available at: https://doi.org/10.1038/s41746-018-0029-1.

Ramkumar, N. *et al.* (2017) ‘Prediction of liver cancer using Conditional probability Bayes theorem’, *IEEE*, 6(8).

Reverter, E., Tugues, S. and Martinez-Sanchez, M. (2015) ‘Treatment for Liver Cirrhosis’, *Cryobiology*, 24(6), pp. 582–583. Available at: https://doi.org/10.1016/0011-2240(87)90158-1.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) ‘Model-Agnostic Interpretability of Machine Learning’, *arXiv:1606.05386 [cs, stat]*, 8(12). Available at: https://arxiv.org/abs/1606.05386.

Saba, L. *et al.* (2019) ‘The present and future of deep learning in radiology’, *European journal of radiology*, 114(14), pp. 14–24. Available at: https://doi.org/10.1016/j.ejrad.2019.02.038.

Sculley, D. *et al.* (2023) ‘Machine Learning: The High Interest Credit Card of Technical Debt’, *Google Research* [Preprint].

Sebastian, A., Kuriakose, R. and Varghese, S.M. (2016) ‘Fuzzy Logic for Child-Pugh classification of patients with Cirrhosis of Liver’, *Advanced Computing: An International Journal*, 7(3), pp. 29–36. Available at: https://doi.org/10.5121/acij.2016.7303.

Shi, Y. *et al.* (2022) *Acute-On-Chronic Liver Failure: Natural History, Mechanism, And Treatment*, *Google Books*. Frontiers Media SA. Available at: https://books.google.co.uk/books?hl=en&lr=&id=zNxmEAAAQBAJ&oi=fnd&pg=PA118&dq=Mingxue+Yu+et+al.+(2016)+recursive+algorithm&ots=mRclgdt1tA&sig=ukvhYcXzoHIUCdWypvsEGpkaW0E#v=onepage&q&f=false.

Tian, X. *et al.* (2019) ‘Using Machine Learning Algorithms to Predict Hepatitis B Surface Antigen Seroclearance’, *Computational and Mathematical Methods in Medicine*, 2019(8), pp. 1–7. Available at: https://doi.org/10.1155/2019/6915850.

Tsantekidis *et al.* (2019) ‘Machine Learning for Forecasting Mid-Price Movements Using Limit Order Book Data’, *IEEE Access*, 7(8), pp. 64722–64736. Available at: https://doi.org/10.1109/access.2019.2916793.

Tsochatzis, E.A., Bosch, J. and Burroughs, A.K. (2014) ‘Liver cirrhosis’, *The Lancet*, 383(9930), pp. 1749–1761. Available at: https://doi.org/10.1016/s0140-6736(14)60121-5.

Wiens, J. *et al.* (2019) ‘Author Correction: Do no harm: a roadmap for responsible machine learning for health care’, *Nature Medicine*, 25(10), p. 1627. Available at: https://doi.org/10.1038/s41591-019-0609-x.

Xiang, Y. *et al.* (2023) ‘RMBench: Benchmarking Deep Reinforcement Learning for Robotic Manipulator Control’, *arXiv:2210.11262 [cs]*, 12(6). Available at: https://doi.org/10.48550/ARXIV.2210.11262.

Zhang, Y. (2010) *New Advances in Machine Learning*, *Google Books*. BoD – Books on Demand. Available at: https://books.google.co.uk/books?hl=en&lr=&id=XAqhDwAAQBAJ&oi=fnd&pg=PA19&dq=machine+learning+and+its+types&ots=r2Ni9UEmKl&sig=MomtRPLXCToQzPYkTbfnkE6IupM#v=onepage&q=machine%20learning%20and%20its%20types&f=false.

Zhu, X. (Jerry) (2005) ‘Semi-Supervised Learning Literature Survey’, *minds.wisconsin.edu* [Preprint].

# **Appendix:**

The source code for the above dissertation is attached as follows for reference.

[Data Science Dissertation\_DS7010](https://uelac-my.sharepoint.com/:f:/g/personal/u2216077_uel_ac_uk/ErOsLxcwhIRMkSd5G19_HnMBi5fINJZL5C87Ol9D4EHEHg?e=LnYBEw)