CERTIFICATION

This project titled, **DEVELOPMENT OF A PREDICTIVE MODEL FOR THE CLASSIFICATION OF THE SURVIVAL OF HEPATITIS PATIENTS USING DECISION TREES ALGORITHM,** prepared and submitted by **OWIE-OBAZEE ESOSA of** matriculation number 16010301024 in fulfilment of the requirements for the degree of **BACHELOR OF SCIENCE (Computer Science)** is hereby accepted.

_____ _____

Prof. P.A. Idowu **Date**

(Supervisor)

_____ _____

Dr. I. O. Akinyemi **Date**

(Head of Department)

Accepted as partial fulfilment of the requirements for the degree of BACHELOR OF SCIENCE (Computer Science)

_____ **(Signature and Date)**

Prof. A. I. Akinwande

Dean, College of Basic and Applied Sciences

DEDICATION

This project work is dedicated to God Almighty.

ACKNOWLEDGEMENT

I owe much gratitude to God Almighty who gave me the wisdom, knowledge, understanding, strength, divine help and provision from the commencement of this project work to its completion.

I specially appreciate my supervisor Prof. P.A.Idowu who took keen interest in my project work and guided me all along, and taking the pains to ensure the successful completion of this project work.

I will like to acknowledge the Head of Department Computer Science and Mathematics Dr. I.O. Akinyemi, and offer deep gratitude for the efforts, constant encouragement, guidance and support. I also appreciate all the members of staff of the Department of Computer Science: Dr. Alaba O. B., Dr. (Mrs.) Kasali F.A., Mr.J.A Balogun , Dr. Idowu P.A., Dr. (Mrs.) F.Oladejo , Mr. Ebo I.O and others to mention but a few..

I heartily thank my parents Mr and Mrs Owie-Obazee., and my wonderful siblings, thank you all for your moral and financial support. I am grateful for all the investments into my education and future.

I sincerely appreciate my friends and all Mountain Top University colleagues for their help and support during the period of working on this project. I say God bless you all.

TABLE OF CONTENT

CERTIFICATION i

DEDICATION ii

ACKNOWLEDGEMENT iii

TABLE OF CONTENT iv

LIST OF TABLES vii

LIST OF FIGURES viii

ABSTRCT 1

CHAPTER ONE 2

INTRODUCTION 2

1.1 Background of Study 2

1.2 Statement of Problem 4

1.3 Aim and Objectives of the Study 4

1.4 Methodology of the Study 5

1.5 Justification of the Study 5

1.6 Scope and Limitations of the Study 6

1.7 Organization of Thesis 6

CHAPTER TWO 7

LITERATURE REVIEW 7

2.1 The Liver 7

2.1.1 Hepatitis 7

2.1.2 Hepatitis C virus disease 8

2.1.3 Hepatitis C infection and transmission 9

2.1.4 Hepatitis C diagnosis and treatment 14

2.2 Predictive Modeling 16

2.2.1 Types of predictive models 16

2.2.2 Developing a predictive model 17

2.2.3Validating a predictive model 18

2.2.4Assessing the performance of predictive model 19

2.3 Machine Learning 20

2.31 Supervised machine learning algorithms 20

2.4 Decision Trees Algorithms 23
2.4.1 CHAID (Chi-Squared automatic interaction detector) 24
2.4.2 CART (Classification and regression trees) 24
2.4.3 ID3 (Iterative dichotomiser 3) 25
2.4.4 C4.5 25
2.4.5 C5.0 25
2.5 Related Works 26
CHAPTER THREE 29
RESEARCH METHODOLOGY 29
3.1 Introduction 29
3.2 Data Identification and Collection 29
3.2.1 Identification of variables monitored during follow-up 30
3.2.2 Data collection of variables monitored 30
3.3 Data-Preprocessing 32
3.4 Formulation of Predictive Model for Hepatitis C Survival 35
3.4.1 Decision tree 36
3.4.2 Decision trees algorithm used 37
3.5 Performance Evaluation 37
CHAPTER FOUR 41
RESULTS AND DISCUSSIONS 41
4.1 Introduction 41
4.2 Results of the Description of the Data Collected 41
4.2.1 Results of the attributes in collected dataset 47
4.2.2 Results of the missing attributes in collected dataset 48
4.2.3 Results of data preprocessing of the data collected 49
4.3 Results of the Formulation and Simulation of Predictive Model 53
4.3.1 Results of model development using percentage split technique 54
4.4 Discussion of Results 59
CHAPTER FIVE 66
SUMMARY, CONCLUSIONS AND RECOMMENDATIONS 66
5.1 Summary 66
5.2 Conclusions 66
5.3 Recommendations 67
References 68

APPENDICES **71**

LIST OF TABLES

Table		Page
2.1:	Performance Characteristics for a Predictive Model	22
2.2:	Comparison of Decision Trees Algorithm	31
3.1:	Identified variables for determining Liver disease	39
4.1:	Description of the Distribution of Patients in Selected Dataset	48
4.2:	Description of Nominal Attributes among Selected Dataset	49
4.3:	Description of Numeric Attributes among Selected Dataset	52
4.4:	Identification of Variables with Missing Values	56
4.5:	Performance Evaluation Results of Model Simulation	72

LIST OF FIGURES

Figures	Page
3.1: arff file containing identified attributes	41
3.2: Diagram of a Confusion Matrix	45
4.1: Bar Chart Plot of Variables with Missing Values	57
4.2: Description of the dataset	58
4.3: Results of Various Percentage Split Technique Used	63
4.4: Results of Various Percentage Split Technique Used	65
4.5: Results of Various Percentage Split Technique Used	67
4.6: Diagram of Decision Trees generated by C5 algorithm	70

ABSTRCT

Hepatitis C is a viral infection that causes liver inflammation, sometimes leading to serious liver damage. Globally, an estimated 71 million people have chronic Hepatitis C virus infection and, although research in the area is ongoing, there is currently no effective vaccine against Hepatitis C. It has also been noted that the poor prediction of hepatitis at various health institutions has also led to mass infection, hence this study. The aim of this study is to develop a model that will aid medical experts and novice alike in the classification of the survival of patients with hepatitis C virus (HCV) under treatment so as to mitigate the onset of untimely death based on information assessed from patients with HCV.

In order to achieve the aims and objectives identified for this study, C4.5 Decision Trees Algorithm was used to formulate the prediction model for the survival of Hepatitis Disease based on the data collected The model was simulated using a Waikato Environment for Knowledge Analysis (WEKA) software and The model was validated based on accuracy, sensitivity, false alarm rate and precision using the data collected.

The classification model developed in this study can be integrated into health information Systems in order to complement electronic health records systems which collect information about the identified variables and can be processed by the classification model for the identification of the clinical outcome of patients to whom treatment is provided.

CHAPTER ONE

INTRODUCTION

1.1 Background of Study

Hepatitis C is a viral infection that causes liver inflammation, sometimes leading to serious liver damage. Until recently, its treatment required weekly injections and other oral medications that many infected people could not take because of other health problems or unacceptable side effects. Hepatitis C infection (HCV) disease is a significant reason for liver infections related grimness and mortality worldwide and it represents a major public health problem HCV can spread parentally through both transfusion and contact with tainted blood and its items, intravenous medication utilizing, tainting during clinical techniques and need of regard for well-being safety measures.

In spite of a declining frequency of new contamination, the weight of infection, both regarding mortality and regarding cost, is relied upon to increment over the one decade from now and HCV contamination will be a potential reason for grimness and mortality and for the need of liver transplantation later on. During the initial infection people often have mild or no symptoms and occasionally a fever, dark urine, abdominal pain, and yellow tinged skin occurs. The virus persists in the liver in about 75% to 85% of those initially infected while earlier on chronic infection typically has no symptoms (Te and Jensen, 2010).

Over several years nonetheless, it often leads to liver disease and sometimes cirrhosis while in some cases, individuals with cirrhosis will develop difficulties such as liver failure, liver cancer, or dilated blood vessels in the esophagus and stomach (WHO, 2016). HCV is spread mainly by blood-to-blood contact associated with intravenous drug use, poorly sterilized medical

equipment, needle-stick injuries in healthcare, and transfusions (NAID, 2016). The World Health Organization (WHO) estimates that 170 million persons are infected with HCV worldwide and 3 to 4 million new infections occur each year, making it one of the top public health problems in the world (Bernalet al., 2015). With a prevalence of 5.3% and an estimated 32 million people infected with HCV, Sub Saharan Africa has the highest burden of the disease in the world.

Other WHO regions with a high prevalence of HCV include Eastern Mediterranean (prevalence 4.6%) and Western Pacific (prevalence 3.9%). There is no vaccine against hepatitis C. Prevention includes harm reduction efforts among people who use intravenous drugs and testing donated blood while chronic infection can be cured about95% of the time with antiviral medications such as sofosbuvir or simeprevir. Peginterferon and ribavirin were earlier generation treatments which had a cure rate of less than 50% and greater side effects and getting access to the newer treatments however can be expensive. Those who develop cirrhosis or liver cancer may need a liver transplant and it is the major reason for liver transplantation, though the virus usually recurs after transplantation (Basra *et al.*, 2011).

Data mining involves the use of machine learning algorithms to the identification of unnoticed patterns in large datasets using computers (Mitchell, 1997). Depending on the type of input data, machine learning algorithms can be separated into supervised and unsupervised learning. In supervised learning, input data comes with a recognized class structure (Mohri *et al.*, 2012). In unsupervised learning, input data does not have a recognized class structure, and the task of the algorithm is to reveal a structure in the data (Sugiyama, 2015). This input data is known as training data. Given their capacity to incorporate numerous predictor

variables without compromising the accuracy of the survival prediction, machine learning algorithms provide efficient ways to build prediction models using longitudinal information. The algorithm is usually tasked with creating a model that can predict one of the properties by using other properties. After a model is created, it is used to process data that has the same class structure as input data.

1.2 Statement of Problem

Globally, an estimated 71 million people have chronic Hepatitis C virus infection and, although research in the area is ongoing, there is currently no effective vaccine against Hepatitis C. Half of people with Hepatitis C virus do not know that they are infected, mainly because they have no symptoms – which can take decades to appear. Over the years, it has been discovered that the prediction for the survival of Hepatitis C is constantly getting worse and there has been continuous research the identification of the factors that are related to the prediction for the survival of hepatitis. It has also been noted that the poor prediction of hepatitis at various health institutions has also led to mass infection, hence this study.

1.3 Aim and Objectives of the Study

The aim of this study is to develop a model that will aid medical experts and novice alike in the classification of the survival of patients with hepatitis C virus (HCV) under treatment so as to mitigate the onset of untimely death based on information assessed from patients with HCV.

The specific research objectives are to

i. elicit knowledge on the variables monitored during the treatment of Hepatitis C patients and collect relevant data;

ii. formulate the model for the survival of hepatitis disease;

iii. simulate the model formulated in (ii); and

iv. validate the model.

1.4 Methodology of the Study

In order to achieve the aims and objectives identified for this study, the following methods will be applied.

a. Structured interview used to elicit information on the variables monitored among Hepatitis patients;

b. C4.5 Decision Trees Algorithm to formulate the prediction model for the survival of Hepatitis Disease based on the data collected in (a);

c. The model will be simulated using a Waikato Environment for Knowledge Analysis (WEKA) software; and

d. The model will be validated based on accuracy, sensitivity, false alarm rate and precision using the data collected in (a).

1.5 Justification of the Study

A lot of targets have been set by the Nigerian health sector in regards to providing essential health services to all citizens. It is therefore essential to improve the quality of decisions affecting treatment options in order to reduce disease mortality rates in Nigeria. Predictive models for Hepatitis C survival classification will help identify the most relevant variables for survival and thus allow medical experts to concentrate on a smaller but important set of variables during clinical observation of HCV patients receiving treatment.

1.6 Scope and Limitations of the Study

This study is limited in scope to the use of secondary dataset that was collected from a public online repository. The study focused on the classification of survival for a limited number of years and does not consider the problem as a regression problem which assesses the probability of survival.

1.7 Organization of Thesis

The first chapter of this thesis has been presented in this section. Chapter two consists of the review of related works surrounding, Hepatitis C prevalence, treatment and survival modeling, machine learning and the application of machine learning algorithms in disease survival modeling. Chapter three presents the methodology that was applied in this study including materials and methods, procedures and evaluation criteria. Chapter four presents the results and discussion of the outcome of the methods applied in this study. Chapter five presents the summary, conclusion and recommendations of the study.

CHAPTER TWO

LITERATURE REVIEW

2.1 The Liver

The liver is the largest solid organ and the largest gland in the human body it is located in the right region of the abdomen a little above the stomach. It performs so many functions such as detoxification, production of albumin, synthesis of angiotensinogen, in blood filtration and so many more. An organ as complex as the liver can experience a range of problems in accordance to lives daily activities or behavioral pattern, the consequences can be dangerous or even fatal. Examples of liver disease include: Cirrhosis, Hepatitis, Alcoholic liver disease, fatty liver disease etc. (Elaine and Luo, 2018). Majority of these diseases are of bacterial or viral origin such as the hepatitis a, hepatitis b, hepatitis c diseases caused by the hepatitis a, hepatitis b, hepatitis c virus respectively. It should be noted that these diseases can pose serious health challenges if unmanaged efficiently.

2.1.1 Hepatitis

Hepatitis is an inflammation of the liver. The condition can be self-limiting or can progress to fibrosis (scarring), cirrhosis or liver cancer (World Health Organization, 2019). Hepatitis infections are the most widely recognized reason for hepatitis on the planet yet different contaminations, harmful substances (for example liquor, certain medications), and autoimmune diseases can likewise cause hepatitis. There are 5 fundamental hepatitis infections, alluded to as types A, B, C, D and E (Bernal and Wendon, 2013). These 5 kinds are of most significant concern as a result of the weight of illness and demise they cause and the potential for outbreaks and epidemic spread. Specifically, types B and C lead

to relentless sickness in a huge number of individuals and, together, are the most widely recognized reason for liver cirrhosis and malignant growth.

Hepatitis A and E are regularly brought about by ingestion of contaminated food or water. Hepatitis B, C and D typically happen because of parental contact with tainted body liquids. Normal methods of transmission for these infections includes receiving of tainted blood or blood items, obtrusive clinical strategies utilizing defiled gear and for blood transmission from mother to infant during childbirth, from relative to kid, and furthermore by sexual contact. Hepatitis A, B, and D are preventable with immunization while medications may be used to treat chronic cases of viral hepatitis (Basra, 2011).

There is no specific treatment for NASH; however, A healthy lifestyle is important including physical activity, a balanced diet and weight loss. In certain cases, autoimmune hepatitis may be treated with drugs to suppress the immune system, or a liver transplant may also be an option. Reasons for hepatitis can be isolated into the accompanying significant classes: infectious, metabolic, ischemic, autoimmune, genetic, and others. Infectious agents include viruses, microscopic organisms, and parasites. Poisons, medications, liquor, and non-alcoholic fatty liver infection are metabolic reasons for liver injury and inflammation. Autoimmune and hereditary reasons for hepatitis involve hereditary predispositions and tend to influence trademark populaces.

2.1.2 Hepatitis C virus disease

Hepatitis C virus (HCV) is a virus that infects liver cells and causes liver inflammation. Hepatitis C is an infectious disease caused by the hepatitis C virus (HCV) that mainly affects the liver (Ryan and Ray, 2004). During the initial infection people often have mild or no symptoms (CDC, 2016). Occasionally A fever, dark urine, abdominal pain, and yellow tinged skin occur. The virus persists

in the liver in about 75% to 85% of those initially infected. Early on chronic infection typically has no symptoms (CDC, 2016). The strategies that HCV utilizes to parasitize its hosts make it formidable enemy.

Therapeutic interventions need considerable sophistication to counter its progress. It is estimated that 3–4 million people are infected with HCV each year. Some 130–170 million people are chronically infected with HCV and at risk of developing liver cirrhosis and/or liver cancer. More than 350 000 people die from HCV related liver diseases each year. It is not spread by superficial contact (WHO, 2015). It is one of five known hepatitis viruses: A, B, C, D, and E (NIDDKD, 2012). Diagnosis is by blood testing to look for either antibodies to the virus or its RNA and testing is recommended in all people who are at risk.

There is no vaccine against hepatitis C (Webster *et al*., 2015). Prevention includes harm reduction efforts among people who use intravenous drugs and testing donated blood (WHO, 2015). Chronic infection can be cured about 95% of the time with antiviral medications such as sofosbuvir or simeprevir (WHO, 2015). Peginterferon and ribavirin were earlier generation treatments which had a cure rate of less than 50% and greater side effects (Kim, 2016). Getting access to the newer treatments however can be expensive. Those who develop cirrhosis or liver cancer may require a liver transplant (Rosen, 2011). Hepatitis C is the leading reason for liver transplantation, though the virus usually recurs after transplantation.

2.1.3 Hepatitis C infection and transmission

Hepatitis C infection causes acute symptoms in 15% of cases (Maheshwari and Ray, 2008). Symptoms are generally mild and vague, including a decreased appetite, fatigue, nausea, muscle or joint pains, and weight loss and

rarely does acute liver failure result (Bailey, 2010). Most cases of acute infection are not associated with jaundice (Springer, 2011). The infection resolves spontaneously in 10–50% of cases, which occurs more frequently in individuals who are young and female (Bailey, 2010). About 80% of those exposed to the virus develop a chronic infection (Nelson et al., 2011). This is defined as the presence of detectable viral replication for at least six months. Most experience minimal or no symptoms during the initial few decades of the infection (Ray and Thomas, 2009).

Chronic hepatitis C can be associated with fatigue and mild cognitive problems and this infection after several years may cause cirrhosis or liver cancer (Forton *et al.,* 2005). Late relapses after apparent cure have been reported, but these can be difficult to distinguish from re-infection (Nicot, 2004). The hepatitis C virus (HCV) is a small, enveloped, single-stranded, positive-sense RNA virus [5]. It is a member of the Hepacivirus genus in the family Flaviviridae (Ray and Thomas, 2009). There are seven major genotypes of HCV, which are known as genotypes one to seven (Nakano *et al.*, 2011). The genotypes are divided into several subtypes with the number of subtypes depending on the genotype.

In the United States, about 70% of cases are caused by genotype 1, 20% by genotype 2 and about 1% by each of the other genotypes (Wilkins *et al*., 2010). Genotype 1 is also the most common in South America and Europe. The half-life of the virus particles in the serum is around 3 hours and may be as short as 45 minutes (Pockros, 2011). In an infected person, about 10virus particles are produced each day (Lerat and Hollinger, 2004). In addition to replicating in the liver the virus can multiply in lymphocytes.

Intravenous drug use (IDU) is a major risk factor for hepatitis C in many parts of the world. According to Xia et al. (2008), out of 77 countries reviewed, 25(including the United States) were found to have prevalence of hepatitis C in the intravenous drug user population of between60% and 80%. Twelve countries had rates greater than80%. It is believed that ten million intravenous drug users are infected with hepatitis C; China (1.6 million), the United States (1.5 million), and Russia (1.3 million) have the highest absolute totals (Nelson et al., 2011). Occurrence of hepatitis C among prison in mates in the United States is 10 to 20 times that of the occurrence observed in the general population; this has been attributed to high-risk behavior in prisons such as IDU and tattooing with unsterilized equipment (Imperial, 2010).

Shared intranasal drug use may also be a risk factor (Moyer, 2013). Sexual transmission of HCV has been controversial. It is believed that HCV can be transmitted sexually, but that it is inefficient -- meaning, it is not easy or likely to pass the virus during sex. On the other hand, HCV infection is very efficient when it is passed from the blood of one person to the blood of another person, such as when people share needles for drug use. The frequency of HCV transmission between monogamous sex partners is very low according to most studies. However, the likelihood of sexual transmission of HCV is increased under any of the following circumstances: having multiple lifetime sex partners, engaging in rough sex such as anal sex.

Having a history of a sexually transmitted disease, having sex with a prostitute or intravenous drug user, having sex during menstruation or whenever blood is present. When counseling patients regarding sexual transmission, the following issues may be relevant: For discordant couples, with one HCV-positive

partner and one HCV-negative partner, the negative partner should be regularly screened for HCV infection. For discordant couples in long-term monogamous relationships, a change in sexual practices is not necessary (e.g., if they have not been using condoms, they do not have to start using condoms). For patients who have new or multiple partners, HIV infection, or high-risk sexual behaviors, it is recommended that they use condoms and exercise caution regarding potential blood exposure to help reduce the chance of HCV infection. For HCV-negative patients who have a new HCV-positive partner or engage in high-risk behaviors with a partner of unknown HCV status, regular screening is recommended. (U.S Department of veterans affairs, 2020).

Blood transfusion, transfusion of blood products, or organ transplants without HCV screening carry significant risks of infection (Wilkins et al., 2010). The United States instituted universal screening in 1992 and Canada instituted universal screening in 1990 (Marx, 2010; Day et al., 2009). This decreased the risk from one in 200 units to between one in 10,000 to one in 10,000,000 per unit of blood (Ponde, 2011). This low risk remains as there is a period of about 11–70 days between the potential blood donor's acquiring hepatitis C and the blood's testing positive depending on the method while some countries do not screen for hepatitis C due to the cost (Ponde, 2011).

Those who have experienced a needle stick injury from someone who was HCV positive have about a 1.8% chance of subsequently contracting the disease themselves (Wilkins *et al.*, 2010). The risk is greater if the needle in question is hollow and the puncture wound is deep. There is a risk from mucosal exposures to blood, but this risk is low, and there is no risk if blood exposure occurs on intact skin (Alter, 2007). Hospital equipment has also been documented as a

method of transmission of hepatitis C, including reuse of needles and syringes; (U.S Department of veterans affairs, 2020).multiple-use medication vials; infusion bags; and improperly sterilized surgical equipment, among others (Alter, 2007).

Tattooing is associated with two to threefold increased risk of hepatitis C (Jafari *et al.*, 2010). This can be because of either inappropriately sanitized equipment or pollution of the dyes being utilized. Tattoos or piercings performed either before the mid-1980s, underground, or on the other hand nonprofessionally are of specific worry, since sterile strategies in such settings might be inadequate. The danger likewise gives off an impression of being more prominent for bigger tattoos (Jafari *et al.*, 2010). It is estimated that nearly half of prison inmates share unsterilized tattooing equipment. It is rare for tattoos in a licensed facility to be directly associated with HCV infection (CDC, 2012).

Personal-care items, for example, razors, toothbrushes, and manicuring or pedicuring tools can be sullied with blood. Sharing such things can conceivably prompt presentation to HCV (Lock et al., 2006). Suitable alert should be taken with respect to any ailment that outcomes in dying, for example, cuts and bruises. HCV isn't spread through easygoing contact, for example, embracing, kissing, or sharing eating or cooking utensils (CDC, 2012). Mother-to-youngster transmission of hepatitis C happens in under 10% of pregnancies and there are no measures that modify this danger (Lam et al., 2010). It isn't clear when transmission happens during pregnancy, yet it might happen both during gestation and at delivery (Ponde, 2011). A long labor is related with a more serious danger of transmission. There is no proof that breast-feeding spreads HCV; in any case,

to be mindful, a contaminated mother is encouraged to abstain from breastfeeding if her nipples are broken and bleeding, or if her viral loads are high (Mast, 2004).

No Identifiable Source of Contamination as per the Centers for Disease Control and Prevention, infusion drug use represents roughly 60% of all HCV contaminations in the United States, while other known exposures represent 20-30%. Roughly 10% of patients in most epidemiological investigations, notwithstanding, have no recognizable source of contamination. HCV introduction in these patients might be from a number of extraordinary methods of transmission, including vertical transmission, and parental transmission from clinical or dental strategies before the accessibility of HCV testing. There is no definitive information to show that people with a background marked by presentations, for example, intranasal cocaine use, inking or body puncturing are at an expanded danger for HCV contamination dependent on these introductions exclusively. It is accepted, nonetheless, that these are likely methods of HCV HCV acquisition in the absence of adequate sterilization techniques. (U.S Department of veterans affairs, 2020).

2.1.4 Hepatitis C diagnosis and treatment

The hatching time frame for hepatitis C is fourteen days to a half year. Following starting disease, around 80% of people don't display any indications. The individuals who are intensely indicative may display fever, exhaustion, decreased appetite, sickness, spewing, stomach pain dull pee, dim shaded defecation, joint pain and jaundice (yellowing of skin and the whites of the eyes). The infection is also frequently undiagnosed in those people who continue to acquire chronic HCV infection because the infection remains asymptomatic until decades after infection when symptoms develop secondary to serious liver damage. HCV infection is diagnosed in 2 steps:

i. Screening for anti-HCV antibodies with a serological test distinguishes individuals who have been contaminated with the infection.

ii. If the test is positive for anti-HCV antibodies, a nucleic acid test for HCV ribonucleic acid (RNA) is required to confirm chronic infection as approximately 30 per cent of people infected with HCV spontaneously clear the infection via a strong immune response without treatment. They will also test positive for anti-HCV antibodies, though they are no longer contaminated.

Early diagnosis may avoid health complications that may result from infection and avoid transmission of the virus. WHO suggests screening for individuals that may have an elevated risk of infection. Populations at increased risk of HCV infection include: individuals who infuse drugs, individuals who use intranasal drugs, beneficiaries of contaminated blood items, children born to mothers infected with HCV, individuals with sexual accomplices who are HCV – positive, individuals with HIV disease, detainees or recently imprisoned people and individuals who have had tattoos or piercings. About 2.3 million individuals of the assessed 36.7 million living with HIV internationally have serological proof of past or present HCV infection. On the other hand, among all HIV – infected people, the prevalence of anti-HCV was6.2%.

Liver illnesses represent a significant reason for morbidity and mortality among people living with HIV. Hepatitis C does not generally require treatment as the immune system in a few individuals will clear the infection, and a few people with chronic infection don't generate liver damage. At the point when treatment is important, the objective of hepatitis C treatment is fix. The fix rate relies upon a few elements including the strain of the infection and the type of

treatment given. The standard of care for hepatitis C is evolving quickly. Sofosbuvir, daclatasvir and the sofosbuvir/ledipasvir combination are a piece of the favored regimens in the WHO rules, and can accomplish fix rates above 95%. These medicines are much more effective, safer and better-tolerated than the older therapies.

Access to HCV treatment is improving, yet stays limited. In 2015, of the 71 million people living with HCV infection universally 20% (14 million) knew their finding. 7.4% of those analyzed (1.1 million) were begun on treatment in 2015. In 2016, 1.76 million people were additionally treated in bringing the global coverage of hepatitis C curative treatment to 13%Much should be done all together for the world to accomplish the 80%treatment objective by 2030.

2.2 Predictive Modeling

Predictive modeling, also called predictive analytics, is a mathematical process that seeks to predict future events or outcomes by analyzing patterns that are likely to forecast future results. The goal of predictive modeling is to answer this question: “Based on known past behavior, what is most likely to happen in the future Accurate predictive models can inform patients and physicians about the future course of an illness or the risk of developing illness and thereby help guide decisions on screening and/or treatment (Waijee *et al*, 2013).

2.2.1 Types of predictive models

Machine learning has been previously used to predict behavior outcomes in business, such as identifying consumer preferences for product based on prior purchasing history. A number of different techniques to develop predictive algorithms exist, using a variety of prediction analytic tools/software and have been described in detail in literature (Waijee *et al.*, 2010; Siegel *et al*., 2011).

Some examples include neural networks, support vector machines, decision trees, naïve Bayes etc. Decision trees, for example, use techniques such as classification and regression trees, boosting and random trees to predict various outcomes.

Machine learning algorithms, such as random-forest approaches have several advantages over traditional explanatory statistical modeling, such as lack of a predefined hypothesis, making it less likely to overlook unexpected hypothesis (Liaw *et al.,* 2002). Approaching a predictive problem without a specific causal hypothesis can be quite effective when many potential predictors are available and when there are interactions between predictors, which are common in engineering, biological and social causative processes. Predictive models using machine learning algorithms may therefore facilitate the recognition of important variables that may otherwise not be initially identified (Waijee *et al.,* 2010).

2.2.2 Developing a predictive model

The first step in developing a predictive model, when using traditional regression analysis, is selecting relevant candidate predictor variables for possible inclusion in the model; however, there is no consensus for the best strategy to do so (Royston *et al.,* 2009). A backward-elimination approach starts with all candidate variables, hypothesis tests are sequentially applied to determine which variables 17 should be removed from the final model, whereas a full-model approach includes all candidate variables to avoid potential over-fitting and selection bias. Previously reported significant predictor variables should typically be included in the final model regardless of their statistical significance but the number of variables included is usually limited by the sample size of the dataset (Greenland, 1989). Inappropriate selection of variables is an important and common cause of poor model performance in this situation. As described above,

variables selection is less of an issue using machine learning techniques given that they are often not solely based on predefined hypothesis (Ibrahim *et al.*, 2012). There are several other important issues relating to data management when developing a predictive model, such as dealing with missing data and variable transformation (Kaambwa *et al.*, 2012; Waijee *et al.*, 2013).

2.2.3Validating a predictive model

To be useful as a predictive model, it must not only have predictive ability in the cohort of derivation but also in a validation cohort (Hemingway *et al.*, 2009). A model's performance may differ substantially between derivation and validation cohorts for several reasons including over-fitting of the model, missing important predictor variables, and inter-observer variability of predictors leading to measurement errors (Altman *et al.*, 2009). Subsequently model execution in the induction dataset might be excessively idealistic and isn't an assurance that the model will perform similarly well in another dataset. Unfortunately, the majority of published prediction research focuses solely on model derivation, and validation studies are scarce scant (Toll *et al*., 2008; Altman *et al*., 2009).

Validation can be performed using internal and external validation. A common approach to internal validation is to split the data into two portions – a training set and validation set. If splitting the dataset is not possible given the limited available data, measures such as cross validation or bootstrapping can be used for internal validation (Steyerberg et al., 2010). However, internal validation nearly always yields optimistic results given that the derivation and validation dataset are very similar (as they are from the same dataset). Although external validation is more difficult as it requires data collected from similar sources in a different setting or a different location, it is usually preferred to internal validation (Steyerberg et al., 2001).

2.2.4Assessing the performance of predictive model

When assessing model performance, it is important to remember that explanatory models are judged based on the strength of associations, whereas predictive models are judges solely based on their ability to make accurate predictions. The performance of a predictive model is assessed using several complementary tests, which assess overall performance; calibration, discrimination, and reclassification (Steyerberg *et al.*, 2010). Performance characteristics should be determined and reported for both the derivation and validation datasets. The overall model performance can be measured using R^2, which characterizes the degree of variation in risk explained by the model (Gerds *et al.*, 2008). The adjusted R^2 has been proposed as a better measure, as it accounts for the number of predictors and helps preventing over-fitting. Brier scores are similar measure of performance, which are used when the outcome of interest is categorical instead of continuous (Czado *et al.*, 2009).

Calibration is the difference between observed and predicted event rates for groups of dataset and is assessed using the Hosmer-Lemeshow test (Hosmer *et al.*, 1997). Discrimination is the ability of a model to distinguish between records which do and do not experience an outcome of interest, and it is commonly assessed using the Receiver Operating Characteristics (ROC) curves (Heagrty *et al.*, 2005). However, ROC analysis alone is relatively insensitive for assessing differences between good predictive models (Cook, 2007); therefore, several relatively novel performance measures have been proposed. The net reclassification improvement and integrated discrimination improvement are measures used to assess changes in predicted outcome classification between two models (Pencina *et al.*, 2012).

2.3 Machine Learning

Machine learning, by its definition, is a field of computer science that evolved from studying pattern recognition and computational learning theory in artificial intelligence. It is the learning and building of algorithms that can learn from and make predictions on data sets. These procedures operate by construction of a model from example inputs in order to make data-driven predictions or choices rather than following firm static program instructions. There are several applications for machine learning, the most significant of which is predictive modeling. Every instance (records/set of fields or attributes) in any dataset used by machine learning algorithms is represented using the same set of features (attributes/independent variables). The features may be continuous, categorical or binary. If instances are given with known labels (the corresponding target outputs) then the learning is called supervised, in contrast to unsupervised learning, where instances are unlabeled (Jain *et al.*, 1999).

Supervised classification is one of the tasks most frequently carried out by Social Intelligent Systems. Thus, a large number of techniques have been developed based on Artificial Intelligence (Logic-based techniques, perceptron-based techniques) and Statistics (Bayesian networks, Instance-based networks). The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is known (Kotsiantis *et al.*, 2006).

2.31 Supervised machine learning algorithms

Supervised learning entails learning a mapping between a set of *input variables* (features/attributes) labeled $\mathcal{X}$ and an *output variable* $\mathcal{Y}$ (where j is the

number of records (cases)) and applying this mapping to predict the outputs for unseen data (data containing values for $\mathcal{X}$ but no$\mathcal{Y}$. Supervised machine learning is the most commonly used machine learning technique in engineering and medicine. In supervised machine learning paradigm, the goal is to infer a function, *f*:

$$f: \ \mathcal{X} \rightarrow \mathcal{Y} \tag{2.1}$$

This function, f is the model inferred by the supervised ML algorithm from a sample data or training set $\mathcal{A}_j$composed of pairs of (inputs (X_i) and output(Y_i)) such that $X_i \varepsilon\, \mathcal{X}$ and $Y_i \varepsilon\, \mathcal{Y}$:

$$\mathcal{A}_j = ((X_1, Y_1), (X_2, Y_2), \quad \ldots . (X_n, Y_n)\, \epsilon\, (\mathcal{X} \times \mathcal{Y})^n \tag{2.2}$$

Typically, for regression problems, $\mathcal{X} \subset \mathbb{R}^d$ (where d is the dimension (or number of features) of the vector,$\mathcal{X}$) and $Y_n \ \epsilon \ \mathbb{R}$; for classification problems $\mathcal{X}\ and\ Y_i$ are discrete while for binary classification $Y_i\, \epsilon\, \{-1, +1\}$. In the statistical learning framework, the first fundamental hypothesis is that the training data are independently and identically generated from an unknown but fixed joint probability distribution function $P(X, Y)$. The goal of the learning algorithm is to find a function, f attempting to model the dependency encoded in $P(X, Y)$ between the input, X and the output, Y. $\mathcal{H}$will denote the set of functions where the solution, f is sought such that $f\, \epsilon\, \mathcal{H}$where $\mathcal{H}$ is the set of all possible functions, *f*.

The second fundamental concept is the notion of error or *loss* to measure the agreement between the prediction *f(X)* and the desired output *Y*. A loss *(or cost)* function, *L* is introduced to evaluate this error (see equation 2.3):

$$L: \quad \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+ \tag{2.3}$$

The choice of the loss function *L(f(X), Y)* depends on the learning problem being solved. Loss functions are classified according to their regularity or singularity properties and according to their ability to produce convex or non-convex criteria for optimization. In the case of pattern recognition, where $Y = \{-1, +1\}$, a common choice for *L* is the misclassification error which is measured as follows:

$$L(f(\boldsymbol{X}), y) = \frac{1}{2}|f(\boldsymbol{X}) - y| \tag{2.4}$$

This *cost* is singular and symmetric. Practical algorithmic considerations may bias the choice of *L*. For instance, singular functions may be selected for their ability to provide *sparse* solutions. For unsupervised learning the problem may be expressed in a similar way using the loss function defined in equation (2.5) and defined in equation (2.6):

$$L_u: \quad \mathcal{Y} \rightarrow \mathbb{R}^+ \tag{2.5}$$

$$L_u\big(f(\boldsymbol{X})\big) = -\log\big(f(\boldsymbol{X})\big) \tag{2.6}$$

The loss function *L* leads to the definition of the risk for a function *f.* also called the *generalization error*:

$$\boldsymbol{R}(f) = \int L(f(\boldsymbol{X}), y)\, dP(\boldsymbol{X}, y) \tag{2.7}$$

In classification, the objective could be to find the function f in$\mathcal{H}$that minimizes R(*f*). Unfortunately, it is not possible because the joint probability P(x, y) is unknown. From a probabilistic point of view, using the input and output random variable notations X and Y, the risk can be expressed in equation (2.8) which can be rewritten in two expectations:

$$R(f) = \mathbb{E}\big(L(f(X), Y)\big) \tag{2.8}$$

$$R(f) = \mathbb{E}[\mathbb{E}(L(f(X), Y)|X)] \tag{2.9}$$

The expression in equation (2.9) offers the opportunity to separately minimize $[\mathbb{E}(L(f(X), Y)|X) = \boldsymbol{x}]$ with respect to the scalar value of f (**x**). The resulting function is the Bayes estimator associated with the risk R. The learning problem is expressed as a minimization of R for any classifier f. As the joint probability is unknown, the solutions inferred from the available training set $\mathcal{A}_n = \big((\boldsymbol{x}_1, y_1), \ldots\ldots\ldots\ldots (\boldsymbol{x}_n, y_n)\big)$. There are two ways to address the problem. The first approach, called generative based, tries to approximate the joint probability P(X, Y), or P(Y|X)P(X), and then compute the Bayes estimator with the obtained probability. The second approach, called discriminative-based, attacks the estimation of the risk R (f) head on.

2.4 Decision Trees Algorithms

The theory of a decision tree has the following parts: a root node is the starting point of the tree; branches connect nodes showing the flow from question to answer. Nodes that have child nodes are called interior nodes. Leaf or terminal nodes are nodes that do not have child nodes and represent a possible value of target variable given the variables represented by the path from the root. The rules are inducted by definition from each respective node to branch to leaf (Chaurasia *et al*., 2012).

The basic idea of decision tree analysis is to split the given data set into subsets by recursive portioning of the parent node into child node based on the homogeneity of within node instances or separation of between-node instances with respect to target variables. For each node, attributes are examined and the splitter is chosen to be the attribute such that after dividing the nodes into child nodes according to the value of the attribute variable, the target variable is

differentiated to the best using algorithm. Because of this, there is the need to distinguish between important attributes, and attributes which contribute little to overall decision process. Splitting points attribute variables and values of chosen variables are chosen based on a selection and splitting criteria dependent on the chosen decision trees algorithm used.

2.4.1 CHAID (Chi-Squared automatic interaction detector)

It is a fundamental decision tree learning algorithm which is easy to interpret, easy to handle and can be used for classification and discovery of relations between variables (Kass, 1980). CHAID is an extension of the AID (Automatic Interaction Detector) and THAID (Theta Automatic Interaction Detector) procedures. CHAID deals on principal of adjusted significance testing. After detection of interaction between variables it selects the best attribute for splitting the node which made a child node as a collection of homogeneous values of the selected attribute. The method can handle missing values. It does not imply any pruning method.

2.4.2 CART (Classification and regression trees)

This was proposed by Breiman et al. (1984) constructs binary trees which is likewise allude as Hierarchical Optimal Discriminate Analysis (HODA). CART is a non-parametric decision tree learning technique, which, depending on whether the dependent variable is categorical or numerical, respectively. The word binary means that a node can only be divided into two classes in a decision tree. The attribute with the largest reduction in impurity is used f or splitting the node's records. CART accepts data with numerical or categorical values and also handles missing attribute values. It uses cost-complexity pruning and also generate regression trees.

2.4.3 ID3 (Iterative dichotomiser 3)

ID3 was developed by Quinlan (1986). The information gain approach is typically used in the decision tree method to evaluate the appropriate property for each node of the decision tree created. Thus, we can select the attribute with the highest information gain (entropy reduction in the level of maximum) as the test attribute of current node. In this way, the information needed to classify the training sample subset obtained from later on partitioning will be the smallest. That is to say, the use of this property to partition the sample set contained in current node will make the mixture degree of different types for all generated sample subsets reduce to a minimum. Therefore, the use of such an information theory approach will effectively reduce the required dividing number of object classification.

2.4.4 C4.5

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason C4.5 is often referred to as a statistical classifier (Xiaolianget al., 2009). The algorithm C4.5 uses information gain as a splitting criterion. Data with categorical or numerical values may be accepted. To handle continuous values, it generates threshold and then divides attributes with values above the threshold and values equal to or below the threshold. The C4.5algorithm can handle missing values easily. Since missing attribute values are not used by C4.5 in gain calculations.

2.4.5 C5.0

C5.0 is the C4.5 algorithm extension, which is also the ID3 extension. It is the classification algorithm used in the Big Data Set. For speed, memory, and performance, it is better than C4.5. The model C5.0 works by dividing the sample

based on the field that provides the maximum gain of information. Based on the largest information gain area, the C5.0 model will split samples. Afterwards the sample subset that is derived from the former split will be split. Until the sample subset can not be broken, the process will continue and is normally according to another field. Finally looking at the lowest level split, all sample subsets that do not make a major contribution to the model will be rejected. The multi value attribute and the missing attribute from the data set are easily managed by C5.0 (Nilimaet al., 2012).

The Hunt's Algorithm which is applied by decision trees is used to generate a decision tree from top to bottom using a divide and conquer approach. The algorithm requires the use of an attribute selector to identify which attribute is needed to divide the dataset thus generating a tree-node. Hunt's algorithm maintains an optimal split for every stage of data splitting and node generation according to some threshold value in a greedy fashion (Bala, 2004).

2.5 Related Works

There has been a contribution of a number of related work on predictive modeling in the area of hepatitis C diseases. Some of these studies focused on the area of the suvival of hepatitis diseases alongside the recurrence of liver failure following liver transplants and other related area of concerns on the body of knowledge. Following is a review of a number of related works on the subject matter of the development of predictive models for diseases using various machine learning algorithms.

Idowu *et al.* (2017), developed a predictive model for the classification of pediatric HIV/AIDS patients' survival using decision trees algorithms. The study used data collected from 216 pediatric patients collected from 2 tertiary hospitals

in south-west Nigeria. The dataset collected consisted of 11 discrete attributes including the survival class. The predictive model for the survival of HIV/AIDS was formulated using C4.5 decision trees algorithm. The results of the study showed that following the formulation of the predictive model for HIV/AIDS survival, 3 attributes were selected and used to grow the decision tree. The attributes identified was used by the decision trees algorithm to generate a tree with four rules and a classification rate of 99.07%. The rules provided can be used by experts to have a better understanding of the relationship between the attributes and HIV/AIDS survival.

Idowu *et al.* (2015), developed a predictive model for the survival of pediatric sickle cell disease (SCD) using clinical variables. The predictive model was developed using a fuzzy logic based model using three (3) clinical variables. The fuzzy logic model applied the triangular membership function for the formulation of the input and output variables following which the centroid method was chosen as the defuzzification technique. The model developed was not validated using live dataset collected from hospitals. Relevant variables for SCD survival could have been easily identified using feature selection methods.

Ganda *et al.* (2013) performed a comparative analysis of the predictive models developed for predicting the survival of heart failure using unsupervised machine learning algorithms. K-means clustering algorithm was used to classify the survival of heart failure patients into two (2) groups following the application of correlation-based feature selection algorithm to select relevant variables for heart failure survival. The performance of k-means clustering algorithm was improved following the use of relevant variables identified compared to using all the variables identified.

Agrawal *et al.* (2012), developed a predictive model for the classification of the survival of the survival of lung cancer patients. Data for the study was collected from the Surveillance, Epidemiology and End Results (SEER) Program containing patients' data for survival of 6 months, 9 months, 1 year, 2 year and 5 years consisting of 13 input variables. Different decision trees algorithms were used for the formulation of the predictive model, such as: C4.5 decision trees, random forest, Decision Stump and alternating decision trees. The decision trees algorithms used had accuracies of 73.61%, 74.45%, 76.80%, 85.45% and 91.35% for the 6 months, 9 months, 1 year, 2 year and 5 years survival dataset.

Yasin *et al.* (2011), worked on the development of a predictive model for the classification of the diagnosis of hepatitis C disease using machine learning algorithms. The dataset of the study was collected from the University of California UCI Data Repository which consisted of 155 records consisting of 15 binary attributes, 5 continuous attributes and a class attribute. The data-processing involved the use of data normalization and the removal of missing values following which feature selection was applied for the identification of the most relevant features among the initially identified 20 attributes. The results showed that the predictive model developed using the dataset with reduced attribute (37% of initially identified attributes) had a performance of 89% which outperformed that of the model developed using the initially identified attributes.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, the methodology applied to this research work is clearly defined. The chapter starts with a description of the framework for the research methodology, which explains the series of steps required: starting from data identification and collection, model formulation and performance evaluation of the developed predictive models. Before the formulation of the model using machine learning algorithms, filter-based feature selection methods were used in identifying the relevant features for predicting the survival of Hepatitis C disease. In addition, the selected machine learning algorithms selected for the formulation of the predictive model were presented alongside a description of their respective loss/cost functions used in the model formulation process. Finally, the tools of performance evaluations were presented alongside the simulation environment chosen for the study.

3.2 Data Identification and Collection

This section highlights the process involved in identifying the data containing the variables monitored during the follow up of Hepatitis C patients. Each variable name was identified and properly defined with their respective units defined. The method of data collection was also clearly stated showing from whom the data was

collected and the instruments of data collection from the data source alongside the identification of the different survival classes in the dataset.

3.2.1 Identification of variables monitored during follow-up

Following the review of literature in the body of knowledge of Hepatitis C disease, a number of features were identified to be monitored during the follow-up of Hepatitis C patients receiving treatment. The variables monitored (which were identified in related literature) were pared to the variables monitored by Hepatologist attending to hepatitis C patients

3.2.2 Data collection of variables monitored

Following consent by the medical director of the hospitals, the data required for the development of the predictive model for the survival of hepatitis C patients receiving follow-up treatment were collected. There was no need for consent forms since the patients were not required to partake in the study rather; electronic data containing information about each patient excluding their personal information (e.g. names, address, hospital ID, contact number etc.) were collected from the health records. The data collected was stored in spreadsheet format and collected suing a flash drive following the identification of the variables monitored during follow-up of Imagined treatment. For the purpose of handling the problem as a classification problem, the target class (output variable) was determined using three labels, namely: survived, not survived and censored. The table 3.1 identifies the variables used for predicting the survival of hepatitis c.

i. **Survived**: refers to the hepatitis C patients that lived up to or more than the estimated survival time and are either dead or alive (vital status);
ii. **Not Survived**: refers to the hepatitis C patients that did not live up to the estimated survival time and are dead; and

Censored: refers to the hepatitis C patients that were lost during follow-up due to one reason or the other – the patients' survival time is less than the estimated survival time and they are still alive.

Table 3.1: Identified variables for determining Liver disease

S/N	Variable Names	Labels
1.	**Age**	Nominal
2.	**Sex**	Nominal – Male and Female
3.	**Steroid**	Nominal
4.	**Antiviral**	Nominal
5.	**Fatigue**	Nominal
6.	**Malaise**	Nominal
7.	**Anorexia**	Nominal
8.	**Liver big**	Nominal
9.	**Liver firm**	Nominal
10.	**Spleen palpable**	Nominal
11.	**Spider**	Nominal
12.	**Ascites**	Nominal
13.	**Varices**	Nominal
14.	**Bilirubin**	Numeric
15.	**Alk phosphate**	Numeric
16.	**Sgot**	Numeric
17.	**Albumin**	Numeric
18.	**Protime**	Numeric
19.	**Histology**	Nominal
20.	**Survival**	Nominal

The pseudo-code in the following paragraph was used in assigning a target class (Survived, Not Survived and Censored) to each patient's records using the values of the vital status and the survival time of each patient.

If (Survival time >= n)

*Then Survival class = "**Survived**"*

Else if ((Survival time < n) AND (Vital Status = "Alive"))

*Then Survival class = "**Censored**"*

Else

*Survival Class = "**Not Survived**".*

End if.

Where n is the time in days

Following the identification of the target survival class, the records with the target class identified as censored were all removed from the original dataset. This was due to the fact that the study is only concerned with the patients who have been followed up for treatment and have lived up to the estimated time except they died during the course of receiving treatment at the hospital. The variables monitored are believed to contain variables relevant to predicting the survival of hepatitis C disease.

3.3 Data-Preprocessing

Following the collection of data for the patients alongside the values of the 19 attributes alongside the survival of hepatitis disease, the data collected was checked for

the presence of error in data entry including misspellings and missing data. The data was transformed into the attribute file format (.arff) for the purpose of the development of the predictive model for the survival of hepatitis disease using the simulation environment. Figure 3.1 shows a screenshot of the format of the .arff used for model development in the Waikato Environment for Knowledge Analysis (WEKA) – a light-weight java application composed of a suite of supervised and unsupervised machine learning tools.

The arff file is composed of three parts, namely: the relation name, attribute names and the dataset. The relation name section which contains the tag *@relation hepatitis_survival*, used to identify the name of the relation (or file) that contains the data needed for simulation. This section is located at the first line of the file and the tag 'name' following @relation must always be the same as the file name else the file loader of the simulation environment will cease to open the file. This section is followed in the next line by the attribute names section.

The attribute names section which contains the tag @attribute *attribute_name label* was used to identify the attributes that describe the dataset stored in the .arff file needed for simulation. Each attribute name alongside its labels is stated following the @relation tag on each line. The label can be a set of values inserted between brackets or a descriptor (e.g. date, numeric etc.). The last attribute is identified as the target class (survival of hepatitis disease) while the previous attributes are the variables for the risk of liver disease.

The data section which contains the tag @data followed in the next line by the values of the attributes for each record of the survival of hepatitis disease separated by a comma. Each value was listed on a row for each record in the same order as the attributes were listed in the attribute names section. The values inserted into each

record must be the same values defined in each respective attribute; if there is an error in spelling or a label not defined is inserted then the file loader of the simulation environment will fail to load the file. The dataset collected for the purpose of the development of the predictive model for the risk of liver disease was stored in arff in

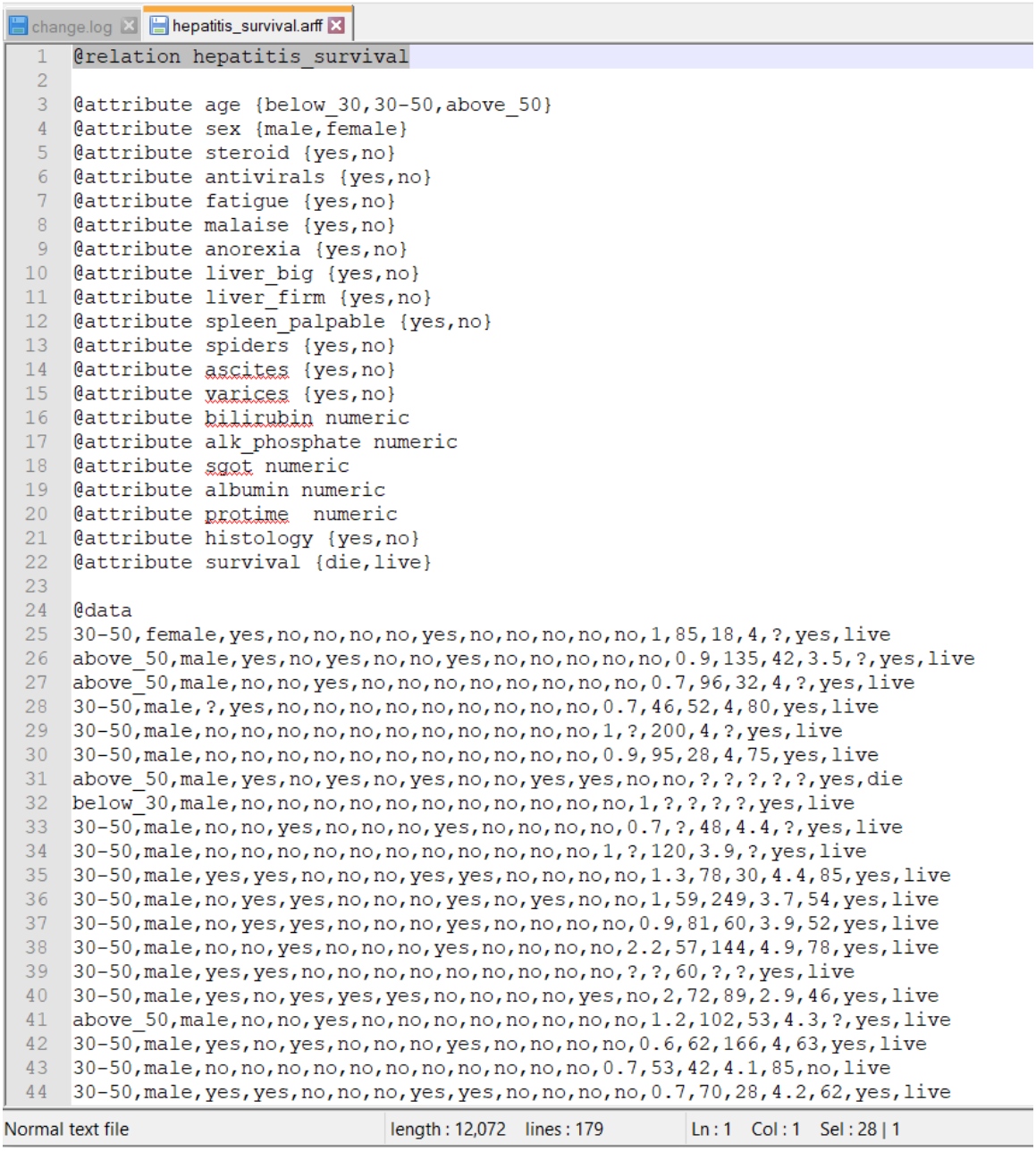

Figure 3.1: arff file containing identified attributes

the name *hepatitis_survival.arff* while the number of attributes listed in the attribute section were 20 including the target attribute. Following this, the values of the survival for the record of the patients considered for this study was provided.

3.4 Formulation of Predictive Model for Hepatitis C Survival

Following the identification of the most relevant and predictive variables (prognostic factors) for Hepatitis C survival, the next phase if the formulation of the predictive model for Hepatitis C survival using the identified variables. Mathematical expressions called mapping functions were used to express the process of model development (and loss function) following which the description of the selected supervised machine learning (SML) algorithms selected for the purpose of this study.

The training dataset *S* which consisted of the original features identified at the point of data identification and collection is represented by X_i, where i is the number of features existing in the original dataset of patients whose record were collected (number of Hepatitis C survival cases). If *n* datasets are selected for training the predictive model using a supervised machine learning to formulate the model using the relevant variables using the mapping:

$$\boldsymbol{\varphi}: \boldsymbol{X_{ik}} \rightarrow \boldsymbol{Y_k};$$

$$defined\ as\ \boldsymbol{\varphi}(\boldsymbol{X_{ik}}) = \boldsymbol{Y_k}\ for\ all\ patients, k \tag{3.1}$$

Where $X_{ik}$the set of attributes, i for patient, k and Y_k is the survival class of patient, k. Hence, the decision trees algorithm determine the best fit for $\varphi \in \mathbb{H}$ (the set

of all possible mappings) based on the minimization of the loss function defined for the decision trees algorithm as the mapping below:

$$\mathbb{L}: \mathbb{Y} \times \mathbb{Y} \rightarrow \mathbb{R}^{+};$$

$$defined\ as\ \mathbb{L}(\boldsymbol{Y_{a,}}, \boldsymbol{Y_p}) \tag{3.2}$$

Where $\mathbb{R}^{+}$ is a positive real number and $Y_{a,}, Y_p$ are the actual and predicted values of the target class (HCV survival) respectively. Hence, the optimal predictive model is formulated when$\lim_{n \to i} \mathbb{L}_n = 0$. Hence, the classification for the survival of Hepatitis C patients is thus:

$$\mathbb{L}(\boldsymbol{Y_{a,}}, \boldsymbol{Y_p}) = \begin{cases} correct\ classification; & = 0 \\ incorrect\ classifcation; & \neq 0 \end{cases} \tag{3.3}$$

3.4.1 Decision tree

The theory of a decision tree has the following parts: a root node is the starting point of the tree; branches connect nodes showing the flow from question to answer. Nodes that have child nodes are called interior nodes. Leaf or terminal nodes are nodes that do not have child nodes and represent a possible value of target variable given the variables represented by the path from the root. The rules are inducted by definition from each respective node to branch to leaf**.** Given a set X_{ij} of *j* number of cases, the decision trees algorithm grows an initial tree using the divide-and-conquer algorithm as follows:

i. If all the cases in X_{ij} belong to the same class or X_{ij} is small, the tree is a leaf labeled with the most frequent class inX_{ij}.

ii. Otherwise, choose a test based on a single attributeX_i with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition X_{ij} into corresponding subsetsaccording to the outcome for each case, and apply the same procedure recursively to each subset.

3.4.2 Decision trees algorithm used

For this study, the C4.5 decision trees algorithm was used for the formulation of the predictive model for the diagnosis of hypertension due to its advantages over the ID3 decision trees algorithm due to its ability to: handle continuous and discrete attributes, handle missing values, handle attributes with differing costs and prune trees after creation. The two criteria used by the C4.5 decision trees in developing its decision trees are presented in equations (3.4) and (3.5) defined as the information gain and the split criteria respectively. Equation (3.4) is used in determining which attribute is used to split the dataset at every iteration while equation (3.5) is used to determine which of the selected attribute split is most effective in splitting the dataset after attribute selection by equation (3.4).

$$IG(X_i) = H(X_i) - \sum_{t \epsilon T} \frac{|t|}{|X_{ij}|} \cdot H(X_i) \tag{3.4}$$

Where:

$$H(X_i) = -\sum_{t \epsilon T} \frac{|t, X_i|}{|X_{ij}|} \cdot \log_2 \frac{|t, X_i|}{|X_{ij}|}$$

$$Split(T) = -\sum_{t \epsilon T} \frac{|t|}{|X_{ij}|} \cdot \log_2 \frac{|t|}{|X_{ij}|} \tag{3.5}$$

T is the set of values for a given attribute X_i.

3.5 Performance Evaluation

In order to evaluate the performance of the supervised machine learning algorithms used for the classification of the survival of hepatitis C, there was the need to plot the results of the classification on a confusion matrix (Figure 3.6). A confusion matrix is a square which shows the actual classification along the vertical and the

predicted along the vertical. All correct classifications lie along the diagonal from the north-west corner to the south-east corner also called True Positives (TP) and True Negatives (TN) while other cells are called the False Positives (FP) and False Negatives (FN).

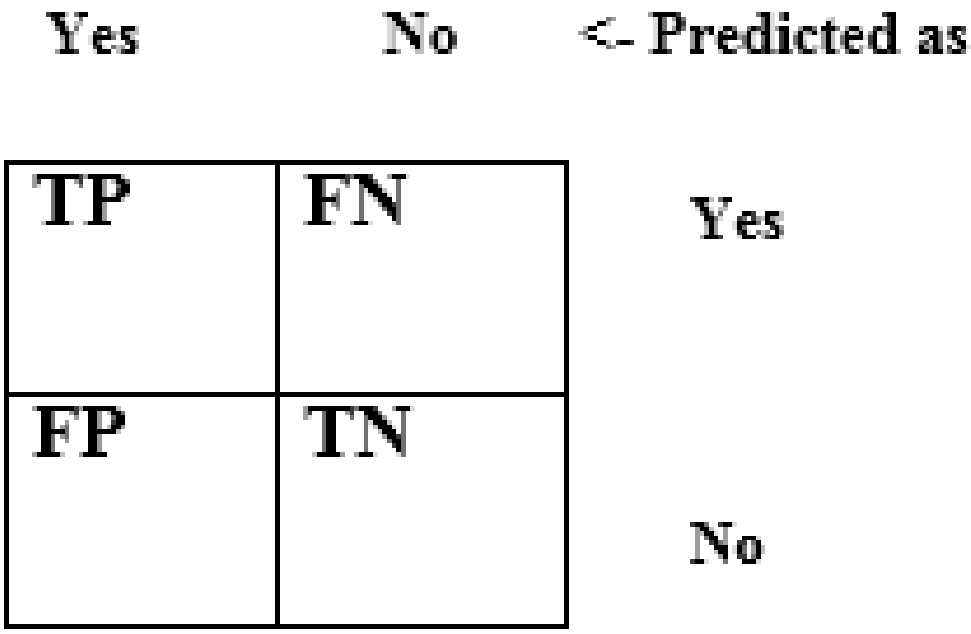

Figure 3.6: Diagram of a Confusion Matrix

In this study, the Yes cases are considered as the cases of patients that survived while the while the No cases are the cases of patients that did not survival Hepatitis C. The definitions of the features of the confusion matrix are presented as follows: True positives (TP) are correctly classified Yes cases, False positives (FP) are incorrectly classified No cases, True negatives (TN) are correctly classified No cases and False negatives (FN) are incorrectly classified Yes cases.

The true positive/negative and false positive/negative values recorded from the confusion matrix can then be used to evaluate the performance of the prediction model. A description of the definition and expressions of the metrics are presented as follows:

a. True Positive (TP) rates (sensitivity/recall) – proportion of positive cases correctly classified.

$$TP\ rate_{Yes} = \frac{TP}{TP + FN} \tag{3.6a}$$

$$TP\ rate_{No} = \frac{TN}{FP + TN} \tag{3.6b}$$

b. False Positive (FP) rates (1-specificity/false alarms) – proportion of negative cases incorrectly classified as positives.

$$FP\ rate_{Yes} = \frac{FP}{FP + TN} \tag{3.7a}$$

$$FP\ rate_{No} = \frac{FN}{TP + FN} \tag{3.7b}$$

c. Precision – proportion of predicted positive/negative cases that are correct.

$$Precision_{Yes} = \frac{TP}{TP + FN} \quad (3.8a)$$

$$Precision_{No} = \frac{TN}{TN + FP} \quad (3.8b)$$

d. Accuracy – proportion of the total predictions that was correct.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.9)$$

CHAPTER FOUR

RESULTS AND DISCUSSIONS

4.1 Introduction

This section presents the results and discussion of the development of a predictive model for the classification of the survival of patients with hepatitis disease. The description of the dataset was done using frequency distribution tables for the nominal/discrete attributes while summary statistics of the numeric attributes was done by identifying the mean, minimum, maximum and standard deviation of the data set distribution. The study also presents the results of the formulation and simulation of the predictive model using the percentage split using 5,10,15,20,25, 30,35,40,45 and 50 percent testing dataset. The results of the comparison of the various simulations used for the development of the predictive model was also presented based on the accuracy, precision, True Positive (TP) rate and False Positive (FP) rate of each model developed. The results of the tree extracted using the selected decision trees algorithm with the best performance was also presented alongside a description of the rules extracted from the decision tree

4.2 Results of the Description of the Data Collected

Table 4.1 shows a description of the distribution of the survival of Hepatitis disease among the patients selected for this study. The results show that majority of the patients were alive which constituted about 79% of the dataset while the remaining 21% consisted of those patients whom were dead. Following the description of the distribution of the survival class of the patients, the attributes used to describe each patient in the dataset were categorized into two (2), namely: nominal and numeric

Variable	Label	Frequency	Percentage (%)
Hepatitis Survival Class	**Die**	32	20.6
	Alive	123	79.4
	Total	**155**	**100.0**

Table 4.1: Description of the Distribution of Patients in Selected Dataset

Table 4.2: Description of Nominal Attributes among Selected Dataset

Variable	Label	Frequency	Percentage (%)
Age (in Years)	**Below 30**	25	16.1
	30 – 50	87	56.1
	Above 50	43	27.8
Sex	**Male**	139	89.7
	Female	16	10.3
Steroid	**Yes**	76	49.0
	No	78	50.3
Antivirals	**Yes**	24	15.5
	No	131	84.5
Fatigue	**Yes**	100	64.5
	No	54	34.8
Malaise	**Yes**	61	39.3
	No	93	60.0

Anorexia	**Yes**	32	20.6
	No	122	78.7
Big Liver	**Yes**	25	16.1
	No	120	77.4
Firm Liver	**Yes**	60	38.7
	No	84	54.2
Palpable Spleen	**Yes**	30	19.4
	No	120	77.4
Spiders	**Yes**	51	32.9
	No	99	63.9
Ascites	**Yes**	20	12.9
	No	130	83.9
Varices	**Yes**	18	11.6
	No	132	85.2

Histology	**Yes**	85	54.8
	No	70	45.2

Table 4.3: Description of Numeric Attributes among Selected Dataset

Variable Name	Minimum	Maximum	Mean	Standard Deviation
Bilirubin	0.3	8.0	1.43	1.212
Alkaline Phosphate	26.0	295.0	105.33	51.508
Sgot	14.0	648.0	85.89	89.651
Albumin	2.1	6.4	3.82	0.652
Protime	0.0	100.0	61.85	22.875

attributes. The distribution of the nominal and numeric attributes are presented in Table 4.2 and Table 4.3 respectively using frequency distribution tables.

4.2.1 Results of the attributes in collected dataset

The results presented in Table 4.2 show the distribution of the nominal attributes among the attributes in the selected dataset. Based on this result, it was observed that majority of the patients were between 30 and 50 years old owing for a proportion of 56% followed by patients whom were above 50 years owing for a proportion of about 28%. In addition, to information about the age group of the patients, it was also observed that majority of the patients were male owing for a proportion of 89% of the dataset distribution. The results of the clinical variables also showed that majority of patients had no antivirals (85%), had fatigue (65%), had no malaise (60%), had no big liver (77%), had no firm liver (54%), had no palpable spleen (77%), had no spiders (64%), had no ascites (84%), had no Varices (85%) and have histology (55%).

The results presented in table 4.3 shows the distribution of the numeric attributes among the selected attributes based on their mean, minimum, maximum and standard deviation (or spread of data). The results of the Bilirubin content showed that the dataset contains a distribution with a mean of 1.43 and standard deviation of 1.212 which reflected a distribution between 0.3 and 8.0. The results of the Alkaline Phosphate content showed that the dataset contains a distribution with a mean of 105.33 and standard deviation of 51.508 which reflected a distribution between 26.0 and 295.0. The results of the Sgot content showed that the dataset contains a distribution with a mean of 85.89 and standard deviation of 89.651 which reflected a distribution between 14.0 and 648.0.

The results of the Albumin content showed that the dataset contains a distribution with a mean of 3.82 and standard deviation of 0.652 which reflected a distribution between 2.1 and 6.4. The results of the Protime showed that the dataset contained a distribution with a mean of 61.85 and standard deviation of 22.875 which reflected a distribution between 0.0 and 100.0. The results showed that the greatest variation among the patients was reflected by the values of the Sgot, Alkaline Phosphate and Albumin contents unlike the smaller variations exhibited by Bilirubin and Albumin contents.

4.2.2 Results of the missing attributes in collected dataset

The results of the identification of the variables with missing values are presented in Table 4.4 which shows the number of records with missing values alongside their proportion among the 155 dataset records. Figure 4.1 shows the graphical plot of the distribution of the variables with missing values using bar charts. Table 4.4 shows that the variable with the highest number of missing values is the Protime content owing for a proportion of 43% followed by the values of Alkaline Phosphate owing for a proportion of 19% alongside Albumin content with a proportion of 10% with missing values of Firm and big Liver owing for a proportion of about 7% each. Also, equal number of missing values for Steroids, fatigue, malaise and Anorexia were observed owing for a proportion of 0.6% which were variables with the least number of missing values. Another class contains equal proportion of 3.2% for the missing values of attributes such as: palpable Spleen, Spiders, Ascites and Varices. The distribution of the number of missing values found in the dataset is presented using Bar charts as shown in Figure 4.1.

4.2.3 Results of data preprocessing of the data collected

Following the collection of the dataset required for this study from the University of Chicago, Illinois (UCI) Machine Learning Repository, the data was preprocessed as an attribute file format (.arff). Figure 4.2 shows a description of the

Table 4.4: Identification of Variables with Missing Values

Variable name	Missing values	Percentage (%)
Steroids	1	0.6
Fatigue	1	0.6
Malaise	1	0.6
Anorexia	1	0.6
Big Liver	10	6.5
Firm Liver	11	7.1
Palpable Spleen	5	3.2
Spiders	5	3.2
Ascites	5	3.2
Varices	5	3.2
Bilirubin	6	3.9
Alkaline Phosphate	29	18.7
Sgot	4	2.6
Albumin	16	10.3
Protime	67	43.2

E

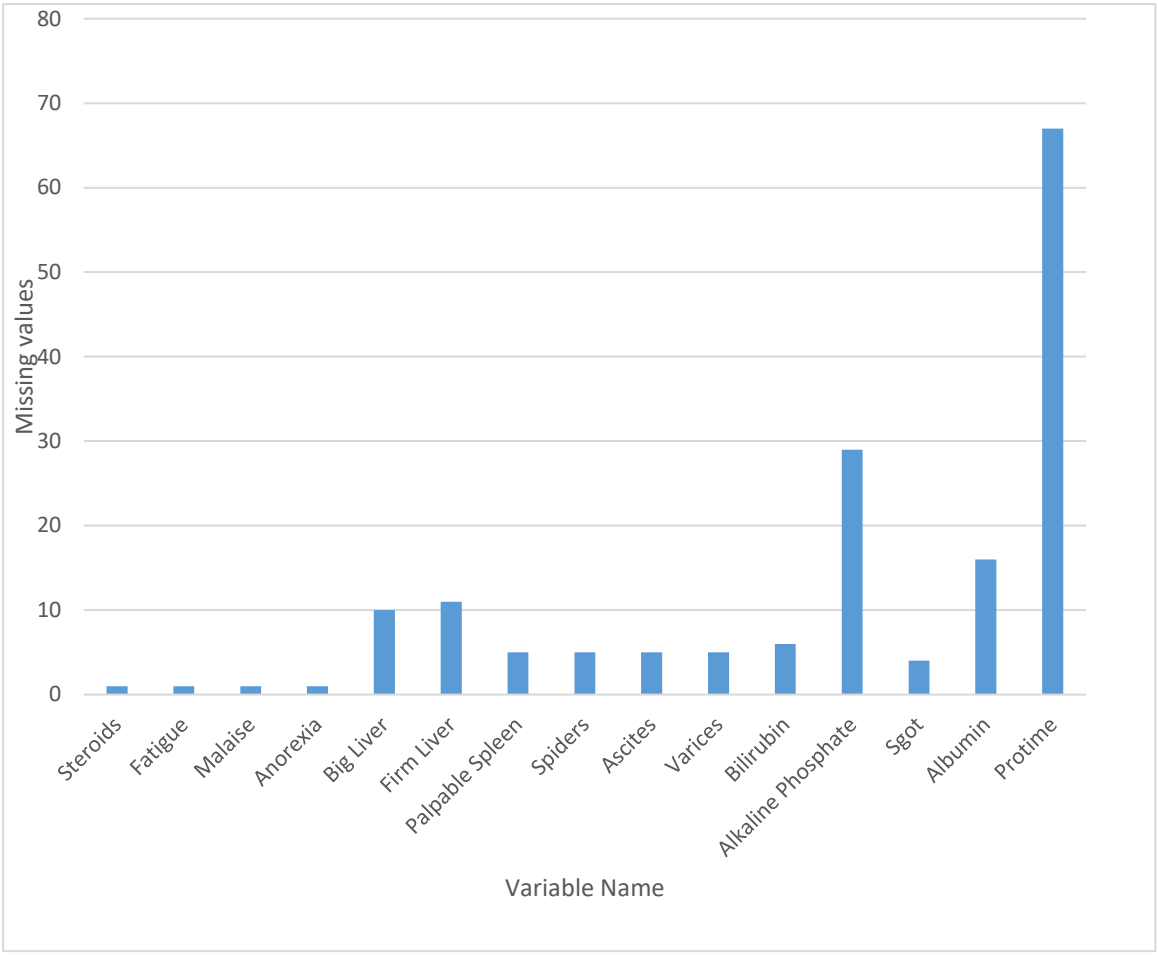

Figure 4.1: Bar Chart Plot of Variables with Missing Values

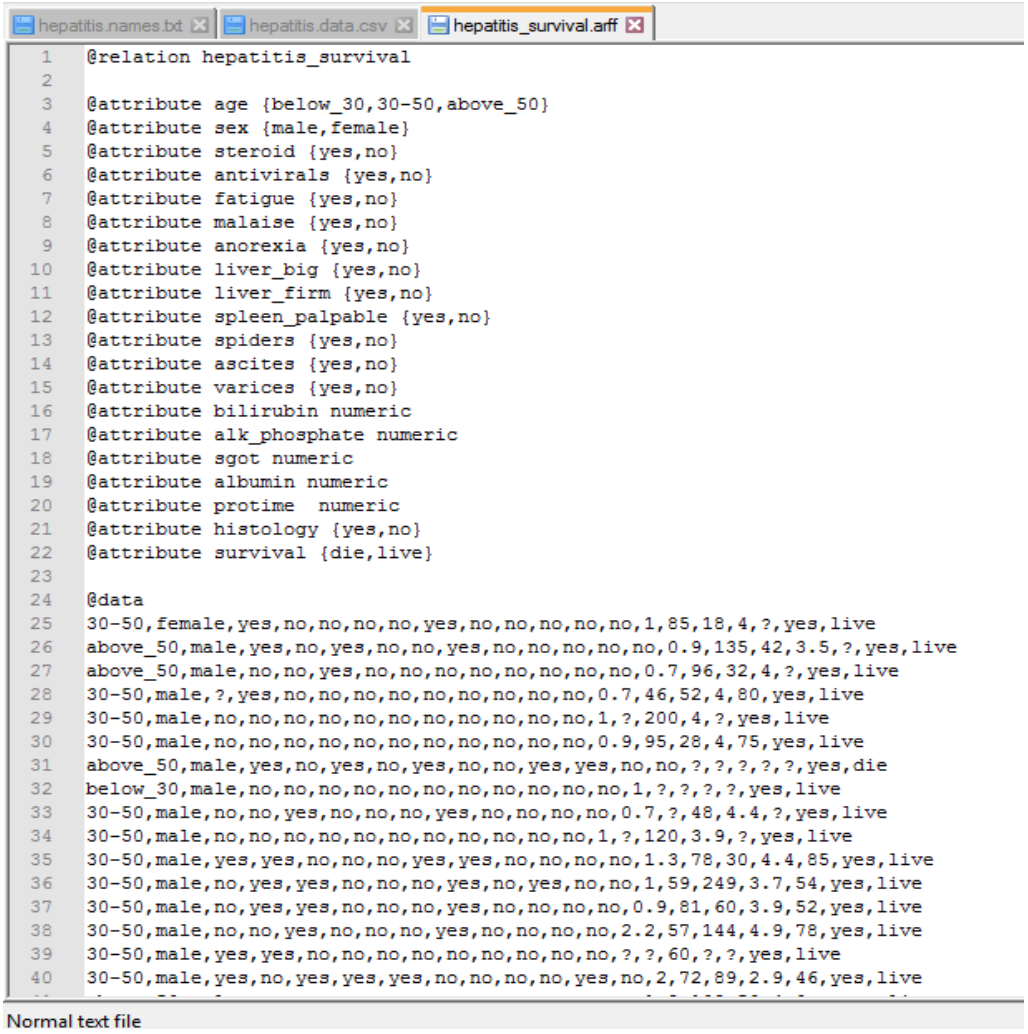

Figure 4.2: Description of the dataset

presentation of the contents of the data selected for this study. The first section of the files identifies the relation, just as we have a relational table in an SQL database consisting of values for a number of attribute pair. Unlike in an SQL database, the last

attribute defined in an .arff file format is the target variable, which in this study is the survival of Hepatitis disease.

Following the first section of the .arff file format of the data collected, the next section consists of the declaration of the attributes used to define the data collected for this study. Each line of declared attribute consists of the name of the attribute alongside the possible values of the attribute enclosed in brackets if the attribute is nominal else a numeric declaration is made using the string “numeric”. As stated in the previous paragraph, the dataset contains 20 attributes which includes the target class attribute as the last attribute row with the possible values “Die” and “Live”.

The final section of the .arff file format shows the values of the attribute pair for each patient for whose data is stored in the dataset. The data section contains 155 rows of an attribute pair of 20 attributes with the last attribute identifying the survival class of the patients identified on the row. The data presented in the last section contains 20 attribute values separated by comma for each patient whose survival class for Hepatitis disease was identified.

4.3 Results of the Formulation and Simulation of Predictive Model

This section presents the results of the process of the formulation of the predictive model using C4.5 decision trees algorithm and the simulation of the model using the WEKA software. The results of the formulation of the decision trees model for the classification of the survival of Hepatitis disease was done using a structural hierarchical trees structure which presented each node in the tree as an attribute selected among the initially identified attributes of the original dataset. The tree generated from the dataset for the classification of Survival was converted to IF-THEN statements which can also be used to implement an expert system for decision support of clinical outcome of patients under study or medication. The model was formulated using the

J48 classifier the native class for implementing the C4.5 decision trees algorithm on WEKA.

The process of simulation using the WEKA software was done using 1 form of training technique. The simulation involved the use of a percentage training process of using 90%, 95%, 80%, 85%, 70%, 75%, 60% 65%, 50%and 45% for training. The results of the simulation process for the training technique used was presented using a 2 by 2 confusion matrix such that the sum of values of rows correspond to total actual cases while the sum of the values of columns correspond to total predicted cases. The confusion matrices were then used to evaluate the performance of the classification models based on the selected performance metrics defined.

4.3.1 Results of model development using percentage split technique

The results of the model simulation process using the percentage split technique involved a process of model development by using a larger percentage of the dataset for training the model (training data) and a lower percentage for testing the model (testing data). By using a percentage of 90%, 95%, 80%, 85%, 70%, 75%, 60% 65%, 50%and 45% for training a model, the results of the correct and incorrect classifications made by the C4.5 decision trees algorithm is presented in Figure 4.3, Figure 4.4, Figure 4.5 using 5%,10%,15% ,20%,25%, 30%,35%, 40%,45 and 50% respectively of dataset for testing the developed model.

The results of using 95% for training the model is displayed on the confusion matrix shown in Figure 4.3 (top left) which shows that 0 Die cases and 8 Live cases were used for testing. The model developed using 95% of the dataset for training was able to correctly classify 0 Die cases and 6 Live cases but misclassified 0 Die case as Live and 2 Live cases as Die which also presented an accuracy of 75% for 6 correct out of all 8 actual cases.

The results of using 90% for training the model is displayed on the confusion matrix shown in Figure 4.3 (top right) which shows that 1 Die case and 14 Live cases were used for testing. The model developed using 90% of the dataset for training was able to correctly classify the 1 Die case and 12 Live cases but misclassified 2 Live cases as Die which also presented an accuracy of 86.7% for 13 correct out of all 15 actual cases.

The results of using 85% for training the model is displayed on the confusion matrix shown in Figure 4.3 (bottom left) which shows that 4 Die cases and 19 Live cases were used for testing. The model developed using 85% of the dataset for training was able to correctly classify 0 Die cases and 14 Live cases but misclassified 4 Die case as Live and 5 Live cases as Die which also presented an accuracy of 60.9% for 14 correct out of all 23 actual cases

The results of using 80% for training the model is displayed on the confusion matrix shown in Figure 4.3 (bottom right) which shows that 6 Die cases and 25 Live cases were used for testing. The model developed using 80% of the dataset for training was able to correctly classify 5 Die cases and 17 Live cases but misclassified 1 Die case as Live and 8 Live cases as Die which also presented an accuracy of 71% for 22 correct out of all 31 actual cases.

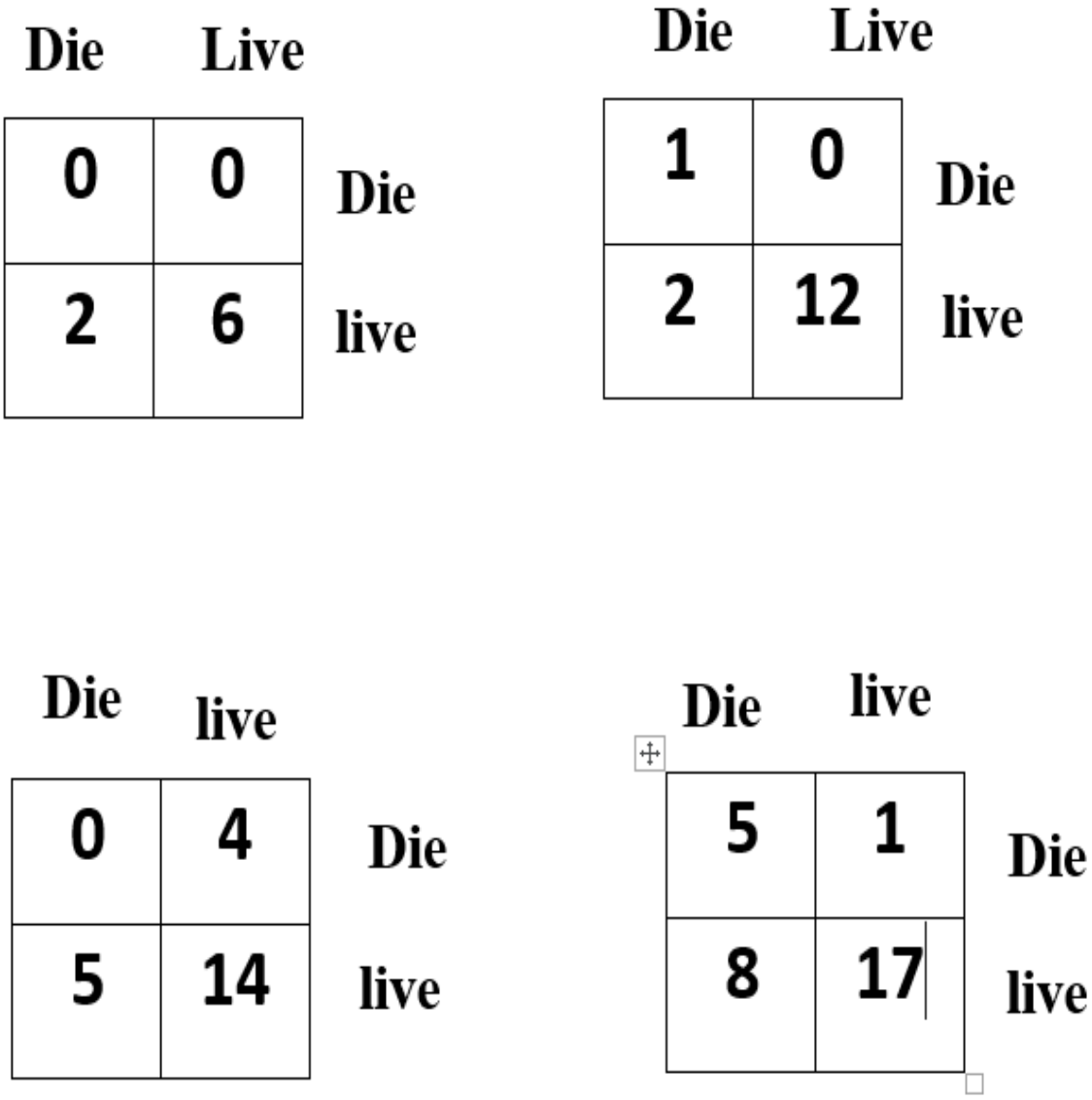

Figure 4.3: Results of Various Percentage Split Technique Used

The results of using 75% for training the model is displayed on the confusion matrix shown in Figure 4.4 (top-left) which shows that 8 Die cases and 31 Live cases were used for testing. The model developed using 75% of the dataset for training was able to correctly classify 3 Die cases and 28 Live cases but misclassified 5 Die case as Live and 3 Live cases as Die which also presented an accuracy of 79.5% for 31 correct out of all 39 actual cases.

The results of using 70% for training the model is displayed on the confusion matrix shown in Figure 4.4 (top-right) which shows that 9 Die cases and 37 Live cases were used for testing. The model developed using 70% of the dataset for training was able to correctly classify 4 Die cases and 35 Live cases but misclassified 5 Die case as Live and 2 Live cases as Die which also presented an accuracy of 84.8% for 39 correct out of all 46 actual cases.

The results of using 65% for training the model is displayed on the confusion matrix shown in Figure 4.4 (bottom-left) which shows that 9 Die cases and 45 Live cases were used for testing. The model developed using 65% of the dataset for training was able to correctly classify 3 Die cases and 40 Live cases but misclassified 6 Die case as Live and 5 Live cases as Die which also presented an accuracy of 79.6% for 39 correct out of all 54 actual cases.

The results of using 60% for training the model is displayed on the confusion matrix shown in Figure 4.4 (bottom-right) which shows that 10 Die cases and 52 Live cases were used for testing. The model developed using 60% of the dataset for training was able to correctly classify 2 Die cases and 48 Live cases but misclassified 8 Die case as Live and 4 Live cases as Die which also presented an accuracy of 80.6% for 50 correct out of all 62 actual cases.

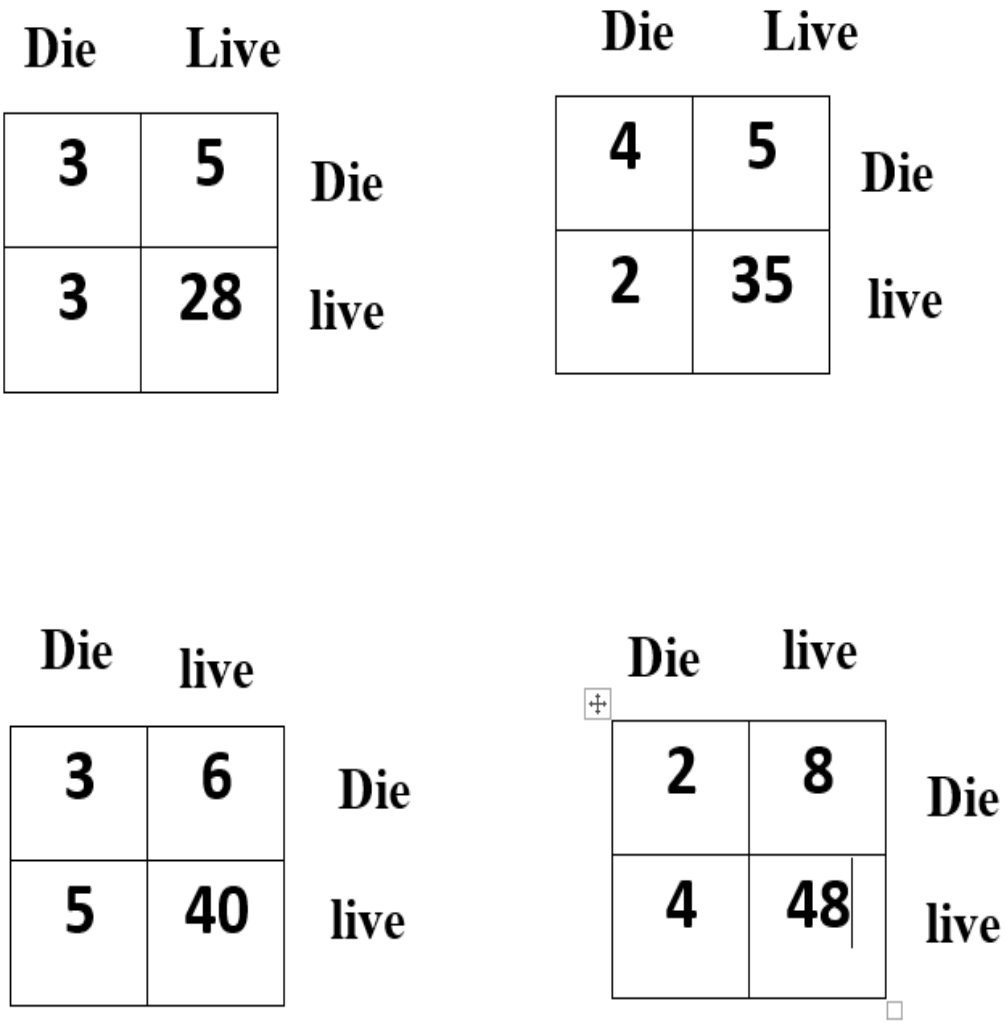

Figure 4.4: Results of Various Percentage Split Technique Used

The results of using 55% for training the model is displayed on the confusion matrix shown in Figure 4.5 (left) which shows that 11 Die cases and 59 Live cases were used for testing. The model developed using 55% of the dataset for training was able to correctly classify 1 Die cases and 59 Live cases but misclassified 10 Die case as Live and 0 Live cases as Die which also presented an accuracy of 85.7% for 600 correct out of all 70 actual cases.

The results of using 50% for training the model is displayed on the confusion matrix shown in Figure 4.5 (right) which shows that 12 Die cases and 65 Live cases were used for testing. The model developed using 50% of the dataset for training was able to correctly classify 1 Die case and all 65 Live cases but misclassified 11 Die cases as Live cases which also presented an accuracy of 85.7% for 66 correct out of all 77 actual cases.

4.4 Discussion of Results

Based on the results presented earlier regarding the formulation and simulation of the results of this study, this section presents the discussion of the results presented. The results of the use of the percentage split and the k-fold cross validation technique for training the model provided a decision tree at the end of the simulation using C4.5 decision trees algorithm. The decision trees generated by the C4.5 decision trees algorithm is presented in Figure 4.6. The decision trees generated had as its nodes, attributes selected from the initially identified 19 attributes in the dataset collected for this study. Among the initially identified variables in this study, the variables used by the decision trees algorithm in generating the classification model for the survival of Hepatitis disease are: Presence of Ascites, Present Age of Patient, Presence of Spiders,

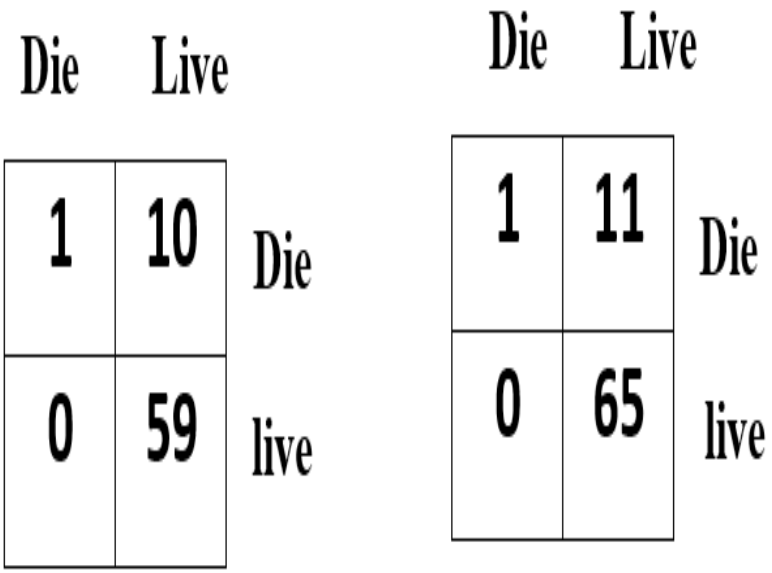

Figure 4.5: Results of Various Percentage Split Technique Used

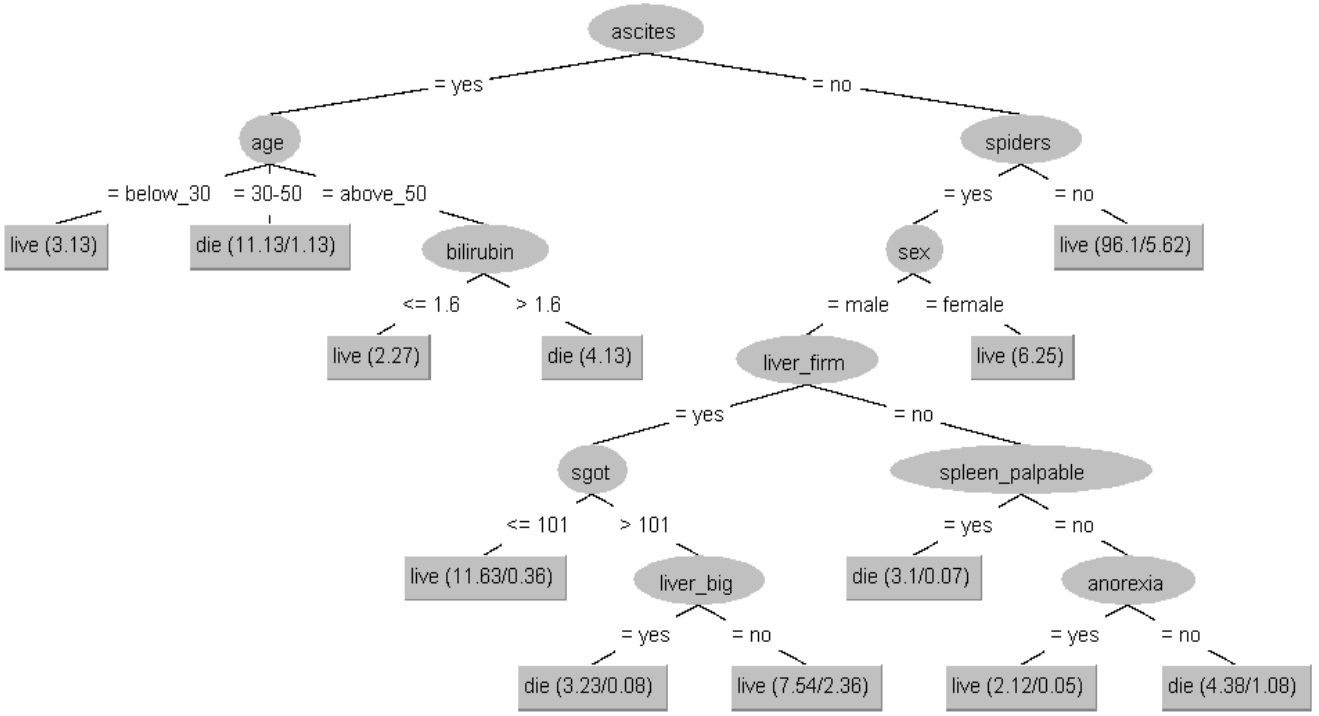

Figure 4.6: Diagram of Decision Trees generated by C4.5 Decision Trees Algorithm

Bilirubin content, Sex of the Patient, Firm Liver, Palpable Spleen, Big Liver and Presence of Anorexia.

These variables were used as the nodes (oval shape) generated by the decision trees algorithm that was used to build the tree which was composed of 12 rules identified by the 12 leaf nodes (square shape) at the terminal point of the decision trees at the bottom. The rules were interpreted from the tree using IF-THEN rules to trace the relationship between children nods from parent nodes all the way to the terminal nodes called the edges where the consequent part of is found. Following is a description of the rules extracted from the decision trees generate in this study. The extracted rules can be used to guide the clinical decision making process taken by experts in the prognosis of the outcome of hepatitis patients.

i. IF (Ascites=Yes) AND (Age=below 30) THEN (Survival=Live);
ii. IF (Ascites=Yes) AND (Age=30-50) THEN (Survival=Die);
iii. IF (Ascites=Yes) AND (Age=Above 50) AND (Bilirubin=<=1.6) THEN (Survival=Live);
iv. IF (Ascites=Yes) AND (Age=Above 50) AND (Bilirubin=>1.6) THEN (Survival=Die);
v. IF (Ascites=No) AND (Presence of Spiders=No) THEN (Survival=Live);
vi. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Female) THEN (Survival=Live);
vii. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Male) AND (Firm Liver=Yes) AND (Sgot=<=101) THEN (Survival=Live);
viii. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Male) AND (Firm Liver=Yes) AND (Sgot=>101) AND (Big Liver=Yes) THEN (Survival=Die);

ix. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Male) AND (Firm Liver=Yes) AND (Sgot=>101) AND (Big Liver=No) THEN (Survival=Live);

x. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Male) AND (Firm Liver=No) AND (Palpable Spleen=Yes) THEN (Survival=Die);

xi. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Male) AND (Firm Liver=No) AND (Palpable Spleen=No) AND (Presence of Anorexia=Yes) THEN (Survival=Live); and

xii. IF (Ascites=No) AND (Presence of Spiders=Yes) AND (Sex=Male) AND (Firm Liver=No) AND (Palpable Spleen=No) AND (Presence of Anorexia=No) THEN (Survival=Die).

Following the presentation of the decision tree that was generated in this study for the classification of the survival of patients with Hepatitis disease, the presentation of the discussion of the evaluation of the performance of the classification model based on the training techniques adopted in this study. Table 4.6 shows a summary of the evaluation of the performance of the classification model developed using C4.5 decision trees algorithm via the percentage split techniques.

Using the percentage split, it was observed from the results that the best performance was achieved by using 90% of the total dataset records for training and using 10% for testing the performance of the model. It was observed that the model had an accuracy y of 86.7% as a result of 13 correct classifications out of the 15 cases presented in the testing dataset. The results also showed that as the proportion of testing dataset was increasing from 5% to 50%, the accuracy of the model dropped to 61% using 85% for training, which increased to 70% using 80% for training, increased to 80% using 75%, increased to 85% using 70% for training, dropped to 80% using 65%

Table 4.6: Performance Evaluation Results of Model Simulation

Dataset	Correct/Total	Accuracy (%)	TP rate	FP rate	Precision	Area under ROC
95% Training	6/8	75	0.750	?	1.000	?
90% Training	13/15	86.67	0.867	0.010	0.956	1.000
85% Training	14/23	60.9	0.609	0.872	0.643	0.395
80% Training	22/31	70.97	0.710	0.196	0.836	0.777
75% Training	31/39	79.5	0.795	0.517	0.777	0.698
70% Training	39/46	84.78	0.848	0.457	0.834	0.652
65% Training	43/54	79.6	0.796	0.574	0.796	0.722
60% Training	50/62	80.6	0.806	0.683	0.773	0.786
55%Training	60/70	85.7	0.857	0.766	0.878	0.738
50% Training	66/77	85.7	0.857	0.774	0.878	0.712

for training, increased to 80.7% using 60% for training, increased to 85.7% using 55% for training and increased to 85.7% using 50% for training.

It was observed from the results that the best performance for the percentage split was achieved using 90% and 50% of the dataset records for training the model. However, using the 90% for training, it was observed that a better FP rate was achieved owing for an average of 1% of actual cases misclassified compared to using 50% for training. The results of the 90% for training also showed that an average of 87% of actual cases were correctly classified and an average of 96% of predicted cases were also correctly classified.

It was also observed from the results that on a general note, the percentage split with the best performance between 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55% and 50% training dataset was the decision trees model developed using the 90% training dataset records. Therefore, using the decision trees model for the classification of the survival of Hepatitis patients, clinical experts are able to make credible decision about patients.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary

This study identified variables that were related to the survival of Hepatitis patients receiving treatment and also collected relevant data from an online repository provided by the University of Chicago, Illinois (UCI) Machine Learning Repository. The study preprocessed the data collected from the repository for the purpose of formatting the dataset in order to be complaint with the tools proposed in this study. The study formulated a classification model for the survival of Hepatitis B patients using the C4.5 decision trees algorithm via a percentage split technique. The classification model was simulated using the Waikato Environment for Knowledge Analysis (WEKA). The classification models developed using the C4.5 decision trees algorithm via the percentage split training technique were compared based on their accuracy, TP rate, FP rate and precision.

5.2 Conclusions

The study concluded that based on the variables that were identified in the dataset collected for this study, the C4.5 decision trees algorithm was formulated using a selected number of variables as the nodes of the generated tree. The study concluded that out of the initially identified 19 variables, the variables finally selected in their order of importance on the tree are: Presence of Ascites, Present Age of Patient, Presence of Spiders, Bilirubin content, Sex of the Patient, Firm Liver, Palpable Spleen, Big Liver and Presence of Anorexia. The study concluded from the decision tree generated that 12 rules were extracted using IF-THEN rules. The study also concluded from the results that using the 90% of dataset records for model building via the

percentage split technique provided better classification results and lower misclassification results compared to other percentage split technique used. The study concluded that using a lesser number of variables for the classification of the survival of Hepatitis B patients on treatment will improve clinical decision making made by medical experts.

5.3 Recommendations

The classification model developed in this study can be integrated into health information Systems in order to complement electronic health records systems which collect information about the identified variables and can be processed by the classification model for the identification of the clinical outcome of patients to whom treatment is provided. The variables identified by the decision trees algorithm for building the classification model are the most relevant among the initially identified variables and thus are more likely related to hepatitis survival compared to the other variables.

References

Agrawal, A., Misra, S., Narayanan, R., Polepeddi&Choudhary, A. (2012). Lung Cancer Survival Prediction using Ensemble data Mining on SEER Data. Journal of *Scientific Programming 20*: 29 – 42.

Basra, G., Basra, S. and Parupudi, S. (2011). Symptoms and Signs of Acute Alcoholic Hepatitis. *World Journal of Hepatology 3*(5): 118 – 120.

Basra, S. (2011). Definition, epidemiology and magnitude of alcoholic hepatitis. *World Journal of Hepatology 3(*5): 108– 113.

Bernal W. & Wendon J. (2013). Acute Liver Failure. *New England Journal of Medicine 369*(26): 2525–2534.

Bernal, W., Lee, W.M., Wendon, J., Larsen, F.S. and Williams, R. (2015). Acute Liver Failure: A Curable Disease by 2024? *Journal of Hepatology 62*(1): 112 – 120.

Centre for Disease Control, CDC (2016). Hepatitis C FAQs for Health Professionals. CDC. January 8, 2016. Retrieved on the 4th of February, 2018.

Cox, D.R. (1972). Regression models and Life Tables. *Journal of Stat. Soc. Serv. 34*: 187.

Cruz, J.A. and Wishart, D.S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics 2*: 59 - 75

Dienstag, J.L. (2015). Acute Viral Hepatitis. In Kasper, D., Fauci, A., Hauser, S., Longo, D., Jameson, J. and Loscalzo, J. *Harrison's Principles of Internal Medicine*, 19th Edition.. New York, NY: McGraw-Hill.

Dimitoglou, G., Adams, J.A. and Jim, C.M. (2012). Comparison of the C4.5 and a naïve bayes classifier for the prediction of lung cancer survivability. *Journal of Computing 4*(8): 1 – 12.

Hall, M.A. (1999). *Correlation-based Feature Selection for Machine learning*. PhD Thesis of the University of Waikato, Hamilton, New Zealand.

Idowu, P.A., Aladekomo, T.A., Agbelusi, O., Alaba, O.B. and Balogun J.A. (2017). Survival Model for Pediatric HIV/AIDS Patient Using C4.5 Decision Trees Algorithm. *International Journal of Child Health and Human Development 10*(2): 143 – 155

Idowu, P.A., Aladekomo, T.A., Williams, K.O. and Balogun J.A. (2015). Predictive Model for Likelihood of Survival of Sickle-Cell Anaemia (SCA) among Pediatric Patients using Fuzzy Logic. *Transactions on Networks and Communications 3*(1): 31 – 44

Kaplan, E.L. and Meier, P. (1958). Non-Parametric estimation from incomplete observation. *Journal of American Statistical Association53*: 457

Mitchell, T. (1997). *Machine Learning*, McGraw Hill, New York.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning 1*: 81-106

Rutherford, A. & Dienstag, J.L. (*2016). Viral Hepatitis. In Greenberger, N.J., Blumberg, R.S. and Burakoff, R. Current Diagnosis & Treatment: Gastroenterology, Hepatology, & Endoscopy*. 3rd Edition New York, NY: McGraw-Hill.

Te, H.S. and Jensen, D.M. (2010). Epidemiology of hepatitis B and C viruses: AGlobal Overview. *Journal of Clinical Liver Disease 14*(1):1 – 21.

Waijee, A.K., Higgings, P.D.R. and Singal, A.G. (2013). A Primer on Predictive Models. *Clinical and Translational Gastroenterology 4*(44): 1 – 4.

World Health Organization, WHO (2016). Hepatitis C Fact sheet. Retrieved online on 4th February, 2018.

Yildirim, P. (2015). Filter-Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease. *International Journal of Machine Learning and Computing 5*(4): 258 – 263

APPENDICES

@relation hepatitis_survival

@attribute age {below_30,30-50,above_50}

@attribute sex {male,female}

@attribute steroid {yes,no}

@attribute antivirals {yes,no}

@attribute fatigue {yes,no}

@attribute malaise {yes,no}

@attribute anorexia {yes,no}

@attribute liver_big {yes,no}

@attribute liver_firm {yes,no}

@attribute spleen_palpable {yes,no}

@attribute spiders {yes,no}

@attribute ascites {yes,no}

@attribute varices {yes,no}

@attribute bilirubin numeric

@attribute alk_phosphate numeric

@attribute sgot numeric

@attribute albumin numeric

@attribute protime numeric

@attribute histology {yes,no}

@attribute survival {die,live}

@data

30-50,female,yes,no,no,no,no,yes,no,no,no,no,no,1,85,18,4,?,yes,live

above_50,male,yes,no,yes,no,no,yes,no,no,no,no,no,0.9,135,42,3.5,?,yes,live

above_50,male,no,no,yes,no,no,no,no,no,no,no,no,0.7,96,32,4,?,yes,live

30-50,male,?,yes,no,no,no,no,no,no,no,no,no,0.7,46,52,4,80,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,1,?,200,4,?,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,0.9,95,28,4,75,yes,live

above_50,male,yes,no,yes,no,yes,no,no,yes,yes,no,no,?,?,?,?,?,yes,die

below_30,male,no,no,no,no,no,no,no,no,no,no,no,1,?,?,?,?,yes,live

30-50,male,no,no,yes,no,no,no,yes,no,no,no,no,0.7,?,48,4.4,?,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,1,?,120,3.9,?,yes,live

30-50,male,yes,yes,no,no,no,yes,yes,no,no,no,no,1.3,78,30,4.4,85,yes,live

30-50,male,no,yes,yes,no,no,no,yes,no,yes,no,no,1,59,249,3.7,54,yes,live

30-50,male,no,yes,yes,no,no,no,yes,no,no,no,no,0.9,81,60,3.9,52,yes,live

30-50,male,no,no,yes,no,no,no,yes,no,no,no,no,2.2,57,144,4.9,78,yes,live

30-50,male,yes,yes,no,no,no,no,no,no,no,no,no,?,?,60,?,?,yes,live

30-50,male,yes,no,yes,yes,yes,no,no,no,no,yes,no,2,72,89,2.9,46,yes,live

above_50,male,no,no,yes,no,no,no,no,no,no,no,no,1.2,102,53,4.3,?,yes,live

30-50,male,yes,no,yes,no,no,no,yes,no,no,no,no,0.6,62,166,4,63,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,0.7,53,42,4.1,85,no,live

30-50,male,yes,yes,no,no,no,yes,yes,no,no,no,no,0.7,70,28,4.2,62,yes,live

below_30,female,no,yes,yes,no,no,no,no,no,no,no,no,0.9,48,20,4.2,64,yes,live

below_30,male,no,no,yes,yes,yes,yes,yes,yes,yes,no,no,1.2,133,98,4.1,39,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,1,85,20,4,100,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,0.9,60,63,4.7,47,yes,live

below_30,female,yes,yes,no,no,no,no,no,no,no,no,no,0.4,45,18,4.3,70,yes,live

below_30,male,yes,no,yes,yes,no,no,no,no,no,no,no,0.8,95,46,3.8,100,yes,live

30-50,male,yes,yes,yes,yes,yes,no,yes,no,yes,no,no,0.6,85,48,3.7,?,yes,live

above_50,female,no,no,yes,no,no,no,yes,no,yes,no,no,1.4,175,55,2.7,36,yes,

above_50,male,yes,no,yes,no,no,yes,yes,no,no,no,no,1.3,78,25,3.8,100,yes,live

above_50,male,yes,yes,yes,yes,no,no,no,no,no,no,no,1,78,58,4.6,52,yes,live

30-50,male,yes,yes,yes,yes,no,no,yes,no,no,no,no,2.3,280,98,3.8,40,yes,die

above_50,male,yes,no,yes,yes,no,?,?,no,no,no,no,1,?,60,?,?,yes,die

30-50,female,no,yes,yes,yes,yes,no,no,no,no,no,no,0.7,81,53,5,74,yes,live

below_30,female,yes,no,no,no,no,no,yes,no,no,no,no,0.5,135,29,3.8,60,yes,live

30-50,male,no,no,yes,no,no,no,no,no,no,no,no,0.9,58,92,4.3,73,yes,live

30-50,male,no,no,yes,no,no,no,no,no,yes,no,no,0.6,67,28,4.2,?,yes,die

below_30,male,no,no,yes,yes,yes,no,no,yes,no,no,no,1.3,194,150,4.1,90,yes,live

below_30,female,yes,no,yes,yes,yes,yes,yes,yes,yes,no,no,2.3,150,68,3.9,?,yes

30-50,male,yes,no,no,no,no,no,no,no,no,no,no,1,85,14,4,100,yes,live

above_50,male,no,no,yes,yes,no,no,yes,yes,yes,yes,no,0.3,180,53,2.9,74,no,live

above_50,male,yes,yes,no,no,no,no,no,no,no,no,no,0.7,75,55,4,21,yes,live

below_30,male,no,no,no,no,no,?,?,?,?,?,?,4.6,56,16,4.6,?,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,1,46,90,4.4,60,yes,live

above_50,male,yes,no,yes,no,no,no,no,no,no,no,no,0.7,71,18,4.4,100,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,?,?,86,?,?,yes,live

below_30,male,no,no,yes,yes,no,no,no,no,no,no,no,0.7,74,110,4.4,?,yes,live

30-50,male,yes,no,no,no,no,no,yes,no,yes,no,no,0.6,80,80,3.8,?,yes,live

below_30,female,no,no,yes,yes,no,no,yes,no,no,no,no,1.8,191,420,3.3,46,yes,

30-50,male,yes,no,no,no,no,no,no,yes,no,no,no,0.8,85,44,4.2,85,yes,live

30-50,male,no,yes,yes,yes,yes,no,no,no,yes,no,no,0.7,125,65,4.2,77,yes,live

30-50,male,yes,no,no,no,no,no,no,no,no,no,no,0.9,85,60,4,?,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,1,85,20,4,?,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,no,0.6,110,145,4.4,70,yes,live

30-50,male,no,yes,yes,no,no,no,yes,yes,no,no,no,1.2,85,31,4,100,yes,live

30-50,male,no,no,yes,no,no,no,no,no,no,no,no,0.7,50,78,4.2,74,yes,live

30-50,male,yes,no,yes,yes,yes,no,no,no,no,no,no,0.8,92,59,?,?,yes,live

30-50,male,yes,no,?,?,?,?,?,?,?,?,?,?,?,?,?,yes,live

30-50,male,no,yes,no,no,no,no,no,no,no,no,no,0.7,52,38,3.9,52,yes,live

above_50,male,no,yes,yes,yes,no,yes,yes,no,no,no,no,1,80,38,4.3,74,yes,live

30-50,female,yes,no,yes,yes,no,no,no,yes,no,no,no,1,85,75,?,?,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,0.7,26,58,4.5,100,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,0.7,102,64,4,90,yes,live

30-50,male,no,no,yes,yes,yes,no,no,no,yes,no,yes,3.5,215,54,3.4,29,yes,live

30-50,male,yes,no,no,no,no,yes,yes,yes,no,no,no,0.7,164,44,3.1,41,yes,live

30-50,male,no,no,yes,yes,no,no,no,no,no,no,no,0.8,103,43,3.5,66,yes,live

below_30,male,no,no,no,no,no,no,no,no,no,no,no,0.8,?,38,4.2,?,yes,live

above_50,male,yes,no,no,no,no,no,no,no,no,no,no,0.7,62,33,3,?,yes,live

above_50,male,no,no,yes,yes,yes,no,no,no,yes,yes,no,4.1,?,48,2.6,73,yes,die

30-50,male,no,no,yes,no,no,no,no,no,no,no,no,1,34,15,4,54,yes,live

30-50,male,yes,no,yes,yes,no,no,no,no,no,no,no,1.6,68,68,3.7,?,yes,live

below_30,male,no,no,no,no,no,no,no,no,no,no,no,0.8,82,39,4.3,?,yes,live

30-50,male,yes,no,yes,yes,no,yes,yes,no,yes,no,no,2.8,127,182,?,?,yes,die

above_50,male,no,no,yes,yes,yes,?,?,?,?,?,?,0.9,76,271,4.4,?,yes,live

30-50,male,yes,no,yes,yes,yes,no,yes,no,no,no,no,1,?,45,4,57,yes,live

above_50,male,no,no,no,no,no,no,no,no,no,no,no,1.5,100,100,5.3,?,yes,live

30-50,male,yes,yes,yes,yes,no,no,no,no,no,no,no,1,55,45,4.1,56,yes,live

above_50,male,no,no,yes,no,no,yes,yes,yes,yes,no,no,2,167,242,3.3,?,yes,die

30-50,female,yes,yes,no,no,no,no,yes,no,no,no,no,0.6,30,24,4,76,yes,live

30-50,male,yes,no,yes,no,no,yes,yes,no,yes,no,no,1,72,46,4.4,57,yes,live

below_30,male,no,no,no,no,no,no,no,no,no,no,no,0.7,85,31,4.9,?,yes,live

below_30,male,no,no,yes,yes,yes,no,no,no,no,no,no,0.8,?,14,4.8,?,yes,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,0.7,62,224,4.2,100,yes,live

30-50,male,yes,no,no,no,no,no,no,no,no,no,no,0.7,100,31,4,100,yes,live

above_50,female,yes,no,yes,yes,no,no,no,?,?,?,?,1.5,179,69,2.9,?,yes,live

above_50,female,no,no,yes,yes,no,no,yes,no,yes,no,no,1.3,141,156,3.9,58,yes,live

below_30,male,yes,no,yes,yes,yes,no,yes,no,no,no,no,1.6,44,123,4,46,yes,live

30-50,male,yes,no,yes,yes,no,no,no,yes,no,no,yes,0.9,135,55,?,41,no,die

30-50,male,no,no,yes,yes,yes,no,yes,no,yes,yes,yes,2.5,165,64,2.8,?,no,die

30-50,male,yes,no,yes,yes,yes,no,yes,no,yes,yes,yes,1.2,118,16,2.8,?,no,die

30-50,male,yes,no,yes,yes,yes,yes,yes,no,no,no,no,0.6,76,18,4.4,84,no,live

above_50,female,yes,no,yes,no,no,yes,yes,yes,yes,no,no,0.9,230,117,3.4,41,no,live

30-50,male,yes,no,yes,yes,yes,no,no,yes,yes,no,yes,4.6,?,55,3.3,?,no,die

30-50,male,no,no,no,no,no,?,?,no,no,no,no,1,?,60,4,?,no,live

above_50,male,yes,no,no,no,no,no,no,no,no,no,no,1.5,?,69,2.9,?,no,live

above_50,male,yes,no,yes,yes,no,no,yes,yes,yes,no,no,1.5,107,157,3.6,38,no,die

30-50,male,yes,yes,yes,yes,yes,yes,yes,no,no,no,no,0.6,40,69,4.2,67,no,live

30-50,male,yes,no,yes,yes,no,no,yes,no,yes,no,no,0.8,147,128,3.9,100,no,live

30-50,male,yes,no,yes,yes,no,yes,yes,no,yes,no,no,3,114,65,3.5,?,no,live

30-50,male,no,no,no,no,no,no,no,no,yes,no,yes,2,84,23,4.2,66,no,die

above_50,male,yes,no,yes,no,no,yes,yes,yes,yes,no,no,?,?,40,?,?,no,live

30-50,male,yes,no,yes,yes,no,no,yes,no,yes,yes,yes,4.8,123,157,2.7,31,no,die

below_30,male,no,no,no,no,no,no,no,no,no,no,no,0.7,?,24,?,?,no,live

below_30,male,yes,no,yes,no,no,no,yes,no,no,no,no,2.4,168,227,3,66,no,live

above_50,male,yes,no,yes,yes,yes,no,yes,yes,yes,no,yes,4.6,215,269,3.9,51,no,live

30-50,male,no,no,yes,yes,no,no,yes,no,no,yes,yes,1.7,86,20,2.1,46,no,die

below_30,male,no,no,no,no,no,no,no,no,no,no,no,0.6,?,34,6.4,?,no,live

30-50,male,yes,no,yes,no,no,?,?,yes,yes,yes,no,1.5,138,58,2.6,?,no,die

30-50,male,yes,no,yes,yes,yes,no,no,no,no,no,no,2.3,?,648,?,?,no,live

above_50,male,yes,yes,no,no,no,yes,yes,no,no,no,no,1,155,225,3.6,67,no,live

30-50,male,yes,no,yes,yes,no,no,no,no,no,yes,no,0.7,63,80,3,31,no,die

below_30,male,no,no,no,no,no,no,yes,yes,no,no,no,0.7,256,25,4.2,?,no,live

30-50,male,yes,yes,yes,yes,no,no,no,no,yes,no,no,0.5,62,68,3.8,29,no,die

above_50,male,yes,no,yes,no,no,no,no,no,no,no,no,1,85,30,4,?,no,live

30-50,male,yes,no,yes,no,no,no,yes,yes,no,no,no,1.2,81,65,3,?,yes,live

30-50,male,yes,no,no,no,no,no,no,no,no,no,no,1.1,141,75,3.3,?,no,live

above_50,female,no,no,yes,no,no,no,no,no,no,no,no,3.2,119,136,?,?,no,live

below_30,male,yes,no,yes,no,no,no,no,no,no,no,no,1,?,34,4.1,?,no,live

above_50,male,no,no,no,no,no,no,no,no,no,no,no,1,139,81,3.9,62,no,live

above_50,male,yes,no,yes,yes,no,?,?,no,yes,no,no,?,?,?,?,?,no,die

above_50,male,no,no,yes,no,no,yes,yes,no,no,no,no,3.2,85,28,3.8,?,no,live

above_50,male,yes,no,yes,yes,yes,yes,yes,no,yes,no,no,2.9,90,153,4,?,no,die

below_30,male,yes,no,yes,yes,yes,no,no,no,yes,yes,no,1,160,118,2.9,23,no,live

30-50,male,no,no,no,no,no,no,no,yes,no,no,no,1.5,85,40,?,?,no,live

30-50,male,yes,no,yes,no,no,no,no,no,yes,no,no,0.9,?,231,4.3,?,no,live

above_50,male,no,no,no,no,no,no,yes,yes,yes,no,no,1,85,75,4,72,no,live

30-50,female,no,no,yes,yes,yes,yes,yes,no,yes,no,no,0.7,70,24,4.1,100,no,live

below_30,male,no,no,yes,yes,yes,?,?,no,yes,yes,no,1,?,20,4,?,no,live

above_50,male,no,no,yes,no,no,no,yes,yes,no,yes,yes,2.8,155,75,2.4,32,no,die

above_50,male,yes,no,yes,yes,no,no,no,no,no,yes,no,1.2,85,92,3.1,66,no,live

above_50,male,yes,no,yes,yes,no,no,no,no,yes,yes,no,4.6,82,55,3.3,30,no,die

above_50,male,no,no,no,no,no,no,no,no,no,no,no,1,85,30,4.5,0,no,live

30-50,male,yes,no,yes,yes,yes,no,no,yes,no,no,no,8,?,101,2.2,?,no,die

30-50,male,no,no,yes,yes,yes,no,yes,no,yes,no,no,2,158,278,3.8,?,no,live

above_50,male,no,yes,yes,no,no,no,yes,no,no,no,no,1,115,52,3.4,50,no,live

30-50,male,yes,no,no,no,no,no,yes,no,no,no,no,0.4,243,49,3.8,90,no,die

below_30,male,no,no,yes,no,no,yes,yes,yes,yes,yes,yes,1.3,181,181,4.5,57,no,live

above_50,male,no,no,no,no,no,yes,yes,no,yes,no,no,0.8,?,33,4.5,?,no,live

30-50,male,no,no,no,no,no,no,yes,no,yes,no,yes,1.6,130,140,3.5,56,no,live

30-50,male,no,no,yes,yes,no,no,yes,no,yes,yes,yes,1,166,30,2.6,31,no,die

30-50,male,no,yes,no,no,no,no,no,no,no,no,no,1.3,85,44,4.2,85,no,live

30-50,male,yes,no,yes,yes,yes,yes,yes,no,yes,no,yes,1.7,295,60,2.7,?,no,live

above_50,male,yes,no,yes,yes,no,?,?,yes,no,yes,no,3.9,120,28,3.5,43,no,die

above_50,male,no,no,yes,no,no,no,yes,yes,yes,no,yes,1,?,20,3,63,no,live

30-50,male,yes,no,yes,yes,no,no,no,yes,yes,no,no,1.4,85,70,3.5,35,no,die

30-50,male,no,no,yes,yes,yes,no,no,no,yes,yes,no,1.9,?,114,2.4,?,no,die

30-50,male,yes,no,yes,no,no,no,no,no,no,no,no,1.2,75,173,4.2,54,no,live

30-50,male,no,no,yes,no,no,no,yes,yes,yes,no,yes,4.2,65,120,3.4,?,no,die

above_50,male,yes,no,yes,yes,yes,?,?,?,?,?,?,1.7,109,528,2.8,35,no,die

below_30,male,yes,no,no,no,no,no,?,no,no,no,no,0.9,89,152,4,?,no,live

30-50,male,no,no,no,no,no,no,no,no,no,no,no,no,0.6,120,30,4,?,no,live

30-50,male,no,no,yes,yes,yes,no,no,no,yes,yes,yes,7.6,?,242,3.3,50,no,die

30-50,male,no,no,yes,no,no,no,yes,no,no,no,no,0.9,126,142,4.3,?,no,live

above_50,male,yes,no,yes,yes,no,yes,yes,no,yes,no,no,0.8,75,20,4.1,?,no,live

above_50,female,yes,no,yes,no,no,no,no,yes,yes,no,yes,1.5,81,19,4.1,48,no,live

30-50,male,no,no,yes,no,no,no,no,yes,yes,yes,no,1.2,100,19,3.1,42,no,die