Applying data mining techniques to classify patients with suspected Hepatitis C virus infection

PROJECT REPORT SUBMITTED TO
MANGALORE UNIVERSITY

IN PARTIAL FULLFILMENTS OF THE REQUIREMENTS FOR THE COMPLETION OF MASTERS DEGREE IN STATISTICS

SUBMITTED BY

STEPHIL M.P

M.Sc. IV SEMESTER

UNDER THE GUIDANCE OF

Ms. SUKSHITHA R

DEPARTMENT OF PG STUDIES AND RESEARCH IN STATISTICS MANGALORE UNIVERSITY MANGALAGANGOTHRI-574199

SEPTEMBER-2022

MANGALORE UNIVERSITY

MANGALAGANGOTHRI

DEPARTMENT OF PG STUDIES AND RESEARCH IN STATISTICS

CERTIFICATE

Certified that this is the bonafide record of the project work done by **Ms. Stephil M.P** during the year 2021 - 22 as a part of her M.Sc. (Statistics) fourth semester course.

Reg. No:
201691406124

Project Guide	**Chairman of the department**
(Ms. Sukshitha R)	**(Prof. Ishwara P)**

Place: Mangalagangothri

Date:

ACKNOWLEDGEMENT

First of all, I would like to thank God Almighty for showering his gracious blessings on throughout the project.

I express my profound gratitude to Prof. Ishwara P, Chairman, Department of PG Studies and Research in Statistics, Mangalore University, whole-heartedly for his motivations, support and inspiration.

I am deeply grateful to Ms. Sukshitha R Department of PG Studies and Research in Statistics, Mangalore University, for her continuous help, support and guidance throughout this project.

I would like to express my gratitude to the faculty members Dr. Harsha S Prabhu, Mr.Sathyanarayana, Ms. Prajna R, Ms. Preethi Jayaram Shetty and Ms.Poornima for their constant encouragement.

I thank all my classmates and friends who have helped me directly or indirectly to complete the project successfully.

My special thanks to my parents whose encouragement and support has helped me to rise to this level.

Place:Mangalagangothri Stephil.M.P

Date:

DECLARATION

I, Stephil. M.P hereby declare that the matter embodied in this report entitled **"Applying data mining techniques to classify patients with suspected Hepatitis C virus infection"** *is* a bonafide record of project work carried out by me under the guidance and supervision of Ms. Sukshitha R, Department PG studies and Research in of Statistics, Mangalore University. I further declare that no part of the work contained in the report has previously been formed the basis for the award of any Degree, Diploma, Fellowship or any other similar title or recognition of any other University.

Place: Mangalagangothri (Stephil.M.P)

Date:

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Hepatitis C Virus(HCV) is a virus that causes hepatitis (inflammation of the liver). It is carried and passed to others through the blood and other body fluids.

Different ways the virus is spread include sharing needles with an infected person and being stuck accidentally by a needle contaminated with the virus. Infants born to infected mothers may also become infected with the virus.

Although patients who are infected with hepatitis C virus may not have symptoms, long-term infection may lead to cirrhosis (scarring of the liver) and liver cancer

Hepatitis C Virus(HCV) infection affects more than 170 million people worldwide.1.5% - 3.5% people in India are suffering from the disease. Egypt has the highest prevalence of hepatitis C in the world with prevalence rates reaching 13%-15%. Thus, HCV represents a major public health and economic problem in Egypt.

- The disease results in four stages
- Portal fibrosis
- Periportal fibrosis
- Septal fibrosis
- Cirrhosis

Most people don’t have signs and symptoms in the early stages of primary liver cancer. When signs and symptoms do appear, they may include:
Losing weight without trying

- Loss of appetite
- Upper abdominal pain
- Nausea and vomiting
- General weakness and fatigue
- Abdominal swelling
- Yellowe discoloration of skin and whites of eyes(jaundice)
- White, chalky stools

Risk factors

Factors that increase the risk of primary liver cancer include**:**

- **Chronic infection with HBV or HCV:** Chronic infection with the hepatitis B virus(HBV)or hepatitis C virus (HCV) increase your risk of liver cancer**.**
- **Cirrhosis:** This progressive and irreversible condition causes scar tissue to form in your liver and increase your chances of developing liver cancer.
- **Diabetes:** People with this blood sugar disorder have a great risk of liver cancer than those who don't have diabetes.
- **Non alcoholic fatty liver disease:** An accumulation of fat in the liver increases the risk of liver cancer.
- **Exposure to a flatoxins:** A flatoxins are poisons produced by molds that grow on crops that are stored poorly. Crops, such as grains and nuts, can become contaminated with aflatoxins, which can end up in foods made of these products.
- **Excessive alcohol consumptions:** Consuming more than a moderate amount of alcohol daily over many years can lead to irreversible liver damage and increase your risk of liver cancer.

Treatments

Which treatment is best for you will depend on the size and location of your hepato cellular carcinoma, HCC treatments include:

- **Surgery:** Surgery to remove the cancer and a margin of healthy tissue that surrounds it may be an option for people with early-stage liver cancer shave normal liver function.
- **Liver transplant surgery:** In patients with small tumors and advanced cirrhosis the treatment of choice is liver transplantation. The five year survival in patients with small tumors is 50-60%poorly differentiated tumors that show vascular invasion, and large tumors have a poor prognosis. Although the presence of tumors of both lobes was at any time considered a poor prognosis after liver transplantation, a recent study demonstrated that patients with bilobar disease have the same survival rates as patients with unilobar disease.
- **Destroying cancer cells with heat or cold:** Ablation procedures to kill the cancers cells in the liver using extreme heat or cold may be recommended for people who can't undergo surgery. These procedures include radiofrequency ablation, cryoablation, and ablation using alcohol ormicrowaves.

- **Radiofrequency ablation (RFA):** RFA uses high frequency radio waves to destroy tumor by local heating. The electrodes are inserted into the liver tumor under ultrasound image guidance using percutaneous, laparoscopic or open surgical approach. It is suitable for small tumors(<5 cm).RFA has the best outcomes in patients with a solitary tumor less than 4 cm. Since it is a local treatment and has minimal effect on normal healthy tissue, it can be repeated multiple times. Survival is better for those with smaller tumors. In one study , In one series of 302 patients, the three year survival rates for lesions >5 cm, 2.1 to 5 cm, and < 2 cm were 59, 74, and 91%,respectively. A large randomized trial comparing surgical resection and RFA for small HCC showed similar four-year survival and less morbidities for patients treated with RFA.
- **Cryoablation:** Cryoablation is a technique used to destroy tissue using cold temperature. is not removed and the destroyed cancer is left to be Reabsorbed by the body. Initial results in properly selected patients with unresectable liver tumors are equivalent to those of resection. Cryosurgery involves the placement of a stainless-steel probe into the center of the tumor. Liquidnitrogen is circulated through the end of this device. The tumor -190□for another 15 minutes. After the tumor has thawed, the probe is removed, bleeding is controlled, and the procedure iscomplete. The patient spends the first postoperative night in the intensive care unit and typically is discharged in 3-5 days. Proper selection of patients and attention to detail in performing thecryosurgicalproceduremandatorytoachievegoodresultsandoutcomes.Frequently, cryosurgery is used in conjunction with liver resection, as some of the tumors are removed while others are treated with cryosurgery.
- **Delivering chemotherapy or radiation directly to cancer cells:** Using a catheter that's passed through your blood vessels and into your liver, doctors can deliver chemotherapy drugs or tinny glass spheres containing radiation directly to the cancer cells.
- **Radiation therapy:**Radiation therapy using energy from X- rays or protons may be recommended if surgery isn't an option.A specialized type of radiation therapy ,called stereotacticbody radiotherapy (SBRT),involves ocusing many beams of radiation simultaneously at one point in your body

1.2 Objectives

- Identify the significant risk factors affecting the hepatitis C.
- Predicting survival probability of hepatitis C patients using parametric and non-parametric classification model.
- Identifying better predictive model for classifying hepatitis C patients.

1.3 About the data

To meet the above objectives a secondary data of coronary heart disease is collected from the website www.kaggle.com . The data includes over 615 records and 13 attributes. The variables descriptions are given below,

1. **Category:**(multiclass: "0" means "No","1" means suspected patients,"2" means "Hepatitis", "3" means "Fibrosis", "4" means "Cirrhosis").
2. **Sex:** Male or Female (Nominal)
3. **Age:** Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous) Behavioural
4. **ALB –** Albumin level (Continuous)
5. **ALP -** Alkaline Phosphatase level (Continuous)
6. **ALT -** Alanine transaminase level (Continuous**)**
7. **AST -** Aspartate aminotransferase level(Continuous)
8. **BIL -** Bilaterallevel (Continuous)
9. **CHOL:**Cholestrol level (Continuous)
10. **CHE –**Cholinesterase (Continuous)
11. **CREA –** Creatinine level (Continuous)
12. **GGT -** Gamma glutamyl transferase level (Continuous)
13. **PROT –** Protein level (Continuous)

1.4 Software used

R software

For the data analysis and representation of the data, we have used R software of version 4.1.3. We selected R as it is one of the widely used public domain software for data analysis.

Python programming

The Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter.

Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

Python libraries:

1. Numpy
2. Pandas
3. Seaborn
4. Sklearn
5. SHAP

R and Python are both open-source programming languages with a large community. New libraries or tools are added continuously to their respective catalog. R is mainly used for statistical analysis while Python provides a more general approach to data science.

R and Python are state of the art in terms of programming language oriented towards data science. Learning both of them is, of course, the ideal solution. R and Python requires a time-investment, and such luxury is not available for everyone. Python is a general-purpose language with a readable syntax. R, however, is built by statisticians and encompasses their specific language.

CHAPTER 2

LITRATURE REVIEW

HarelDahari(2010) conducted a study on Meta-Analysis of Hepatitis C Virus Vaccine Efficacy in Chimpanzees indicates an Important for Structural Proteins. The author Studied the effect in patients and chimpanzees that spontaneously cleared hepatitis C virus (HCV) infections demonstrated that natural immunity to the virus is induced during primary infections and that this immunity can be cross protective. he performed a meta-analysis that compared parameters among naïve (n=63), vaccinated (n=53), and rechall enged (n=36) animals, including peak RNA titer post-challenge, timepoints of peak RNA titer, duration of viremia, and proportion of persistent infections.There was no reduction in the rate of HCV persistence in vaccinated animals, compared with naïve animals, when non-structural proteins were included in the vaccine. Vaccines that contained only structural proteins had clearance rates that were significantly higher than vaccines that contained non-structural components.

Amir Hossein KayvanJoo et.al. (2014) in his paper titled "Prediction of hepatitis C virusinterferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms" Utilized feature selection methods (Gini Index, Chi Squared and machine learning algorithms) and other bioinformatics tools to identify genetic determinants of therapy outcome within the entire HCV nucleotide sequence. The researchers used combination of several algorithms, the present study performed a comprehensive bioinformatics analysis and identified several nucleotide attributes within the full-length nucleotide sequences of HCV subtypes 1a and 1b that correlated with treatment outcome. Feature selection algorithms identified several nucleotide features Combination of algorithms utilized the selected nucleotide attributes and predicted HCV subtypes 1a and 1b therapy responders from non-responders with an accuracy of 75.00% and 85.00%, respectively. Based on the identified attributes, decision trees were induced to differentiate different therapy response groups. The study identified new genetic markers that potentially impact the outcome of hepatitis C treatment.

Hend Ibrahim Shousha.et.al.(2018) they studied single nucleotide polymorphism is an etiology-independent predictor of hepatitis C virus (HCV)-related hepatic fibrosis. This study aims to evaluate and compare the prediction accuracy of scoring system like aspartate aminotransferase-to-platelet ratio index and fibrosis-4 index versus data mining for the prediction of HCV-related advanced fibrosis. This retrospective study included 427 patients with chronic hepatitis C. data mining analysis were used to

construct a decision tree by reduced error technique, REPTree algorithm was able to predict advanced fibrosis with sensitivity of 0.7, specificity of 0.7 and receiver operating characteristic (ROC) area of 0.7. The multilayer perceptron (MLP) neural model was selected as the best predictive algorithm with sensitivity of 0.8, specificity of 0.8, and ROC area of 0.8. Thus, the researchers concluded that MLP is better than APRI, FIB-4, and REPTree for predicting advanced fibrosis for patients with chronic hepatitis C.

Monica A. Konerman et.al.(2019) proposed Machine learning algorithms to provide effective ways to build prediction models using longitudinal information given their capacity to incorporate numerous predictor variableswithout compromising the accuracy of the risk predictionthey developed and compared two ML algorithms to predict cirrhosis developmentin a large CHC-infected cohort using longitudinal data.Theresearcher used national Veterans Health Administration (VHA) data to identify CHC patients incare between 2000–2016. Longitudinal models used CS predictors plus longitudinal summary variables(maximum, minimum, maximum of slope, minimum of slope and total variation) betweenenrolment and time zero. Covariates included demographics, labs, and body mass index.Model performance was evaluated using concordance and area under the receiver operating curve (AuROC).Further they implemented-survival-tree based models using longitudinal information are statistically superior to cross-sectional or linear models for predicting development of cirrhosis in CHC, though allfour models were highly accurate.

KhairAhammed et.al. (2020) In this work, a machine learning based model has been proposed that can classify hepatitis C virus infected patient's stages of liver. Researchers gathered the instances of liver fibrosis disease of Egyptian patients from UCI machine learning repository. To balance instances of multiple categories, synthetic minority oversampling methodology was used. Later, different feature selection methods to identify significant features of hepatitis C virus in this dataset. KNN was used to classify hepatitis C virus infected patients. This result has been useful to scrutinize and take decision in hepatitis C virus infectious disease. In this experiment, HCV patient's records was investigated using machine learning techniques to detect the stages of the HCV patient.Top five best sorted results of different classifiers based on various evaluation criteria. In this circumstance, GB, NB and LR shows their accuracy less than 50%, so they are discarded in this section. Thus, SMOTE was applied 8 times in the raw HCV dataset and generated more synthetic instances like existing instances. Then, various feature selection techniques were used to the SMOTE generated HCV patient's instances and produce datasets

CHAPTER 3

METHODOLOGY

2.1 DESCRIPTIVE STATISTICS

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of it. Descriptive statistics are broken down into measures of central tendency and measure of variability or spread. Measures of central tendency include the standard deviation, the minimum and maximum variables, and the kurtosis and skewness.

Measure of central tendency

Mean

The Arithmetic mean is commonly known as average. The average of a given set of numbers is called the arithmetic mean, or simply, the mean of the given numbers. Thus, the arithmetic mean of a group of observations is defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} xi$$

Where the symbol Σ called sigma, stands for summation.

Mode

Mode is a statistical term that refers to the most frequently occurring number found in a set of numbers. The mode is found by collecting and organizing data in order to count the frequency of each result. The result with the highest number of occurrences is the mode of the set.

Measures of Variability

Standard Deviation

Standard deviation is a measure of the dispersion of a set of data from its mean. It is calculated as the square root of variance by determining the variation between each data point relative to the mean. If the data points are further from the mean, there is higher deviation within the data set.

$$\mathbf{S}x = \sqrt{\frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{(n-1)}}$$

n=The number of data points

$\bar{x}$ = The mean of xi

xi= Each of the values of the data.

Measures of Shapes

Kurtosis

Kurtosis is a statistical measure that's used to describe the distribution, or skewness, of observed data around the mean, sometimes referred to as the volatility of volatility. Kurtosis is used generally in the statistical field to describes trends in charts. Kurtosis can be present in a chart with fat tails and a low, even distribution, as well as be present in a chart with skinny tails and a distribution concentrated toward the mean.

Types of Kurtosis

There are three categories of kurtosis that can be displayed by a set of data. All measures of kurtosis are compared against a standard normal distribution, or bell curve.

The first category of kurtosis is a mesokurtic **distribution**. This type of kurtosis is the most similar to a standard normal distribution in that it also resembles a bell curve. However, a graph that is mesokurtic has fatter tails than a standard normal distribution and has a slightly lower peak. This type of kurtosis is considered normally distributed but is not a standard normal distribution.

The second category is a leptokurtic **distribution**. Any distribution that is leptokurtic displays greater kurtosis than a mesokurtic distribution. Characteristics of this type of distribution is one with extremely thick tails and a very thin and tall peak. The prefix of "lepto-" means "skinny," making the shape of a leptokurtic distribution easier to remember. T-distributions are leptokurtic.

The final type of distribution is a platykurtic **distribution**. These types of distributions have slender tails and a peak that's smaller than a mesokurtic distribution. The prefix of "platy-" means "broad," and it is meant to describe a short and broad-looking peak. Uniform distributions are platykurtic.

Skewness

Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

In a normal distribution, the graph appears as a classical, symmetrical “bell-shapedcurve”. The mean, or average, and the mode, or maximum point on the curve, are equal.

1. In a perfect normal distribution (green solid curve in the illustration below), the tails on either side of the curve are exact mirror images of each other.
2. When a distribution is skewed to the left (red dashed curve), the tail on the curve's left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode. This situation is also called **negative skewness**.
3. When a distribution is skewed to the right (blue dotted curve), the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode. This situation is also called **positive skewness.**

Bar chart

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart. A bar graph shows comparisons among discrete categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value. Some bar graphs present bars clustered in groups of more than one, showing the values of more than one measured variable.

Correlation

Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.

➢ A **positive correlation** is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decrease while the other decreases

➢ A **negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other.

We describe correlations with a unit-free measure called the correlation coefficient which ranges from -1 to +1 and is denoted by *r*.

▪ The closer *r* is to zero, the weaker the linear relationship.
▪ Positive *r* values indicate a positive correlation, where the values of both variables tend to increase together.

▪ Negative *r* values indicate a negative correlation, where the values of one variable tend to increase when the values of the other variable decrease.

Data preprocessing

Usually, the data collected from patients' records are not completely clear. Therefore, data cleaning is an essential step for developing machine learning models. First, the ID column was removed. The missing values in our dataset were replaced with the mean value of each variable. The dataset was not balanced, which means that most of the records belonged to the same category. Classification of imbalanced data is biased towards the large categories. The symmetric minority over-sampling technique (SMOTE) was applied to the HCV dataset to facilitate the performance of various classifiers. This method uses the KNN algorithm in creating new synthetic samples to balance the class distribution of the dataset

Applied classification algorithms

Classification models, known as supervised methods, were applied in this study to classify existing data. The dataset was divided into a training set (70%) and test set (30%). Thereafter, each classifier model was trained using the balanced training data. After the training process, the classifier classified patients based on the records in the test set. The performance of the classifiers was evaluated with the test data, and the performance of each model was calculated with different metrics. Model development is done by different classifiers

SMOTE (Synthetic Minority Oversampling Technique)

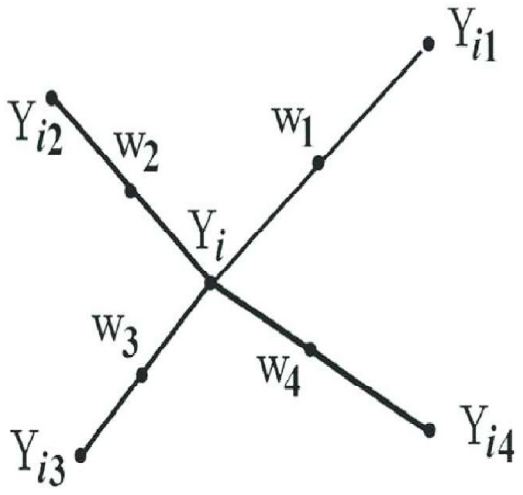
● SMOTE aims to balance class distribution by randomly increasing minority class examples by replicating them.

● SMOTE synthesizes new minority instances between existing minority instances. It generates the **virtual training records by linear interpolation** for the minority class.

● These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbours for each example in the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data.

Synthetic samples generated by taking the difference between the nearest neighbour and feature vector (sample) under consideration. Multiply this difference by a random number among 1 and 0, and add it to the feature vector under consideration. This produces the selection of a random point along the line segment among two distinctive features. The SMOTE algorithm is described in detail below:

- Find the k-nearest neighbours for each sample.
- Select samples randomly from a k-nearest neighbour.
- Find the new samples = original samples + difference * gap (0,1).
- Add new samples to the minority. Finally, a new dataset is created.



The method is shown in figure where Y_i is the point under consideration, Y_{i1} to Y_{i4} are nearest neighbors and w_1 to w_4 the synthetic data generated by the randomized interjection.

Oversampling: This method works with minority class. It replicates the observations from minority class to balance the data. It is also known as up sampling. Similar to under sampling, this method also can be divided into two types: Random Oversampling and Informative Oversampling. Random oversampling balances the data by randomly oversampling the minority class. Informative oversampling uses a pre-specified criterion and synthetically generates minority class observations.

Advantages

- Mitigates the problem of overfitting caused by random oversampling as synthetic examples are generated rather than replication of instances
- No loss of useful information

Disadvantages

- While generating synthetic examples SMOTE does not take into consideration neighboring examples from other classes. This can result in increase in overlapping of classes and can introduce additional noise
- SMOTE is not very effective for high dimensional data

2.2 MULTINOMIAL LOGISTIC REGRESSION

Multinomial logistic regression is used to predict categorical placement in or the probability of category membership on a dependent variable based on multiple independent variables. The independent variables can be either dichotomous (i.e., binary) or continuous (i.e., interval or ratio in scale). Multinomial logistic

regression is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership. Multinomial logistic regression does necessitate careful consideration of the sample size and examination for outlying cases. Like other data analysis procedures, initial data analysis should be thorough and include careful univariate, bivariate, and multivariate assessment. Specifically, multicollinearity should be evaluated with simple correlations among the independent variables. Also, multivariate diagnostics (i.e. standard multiple regression) can be used to assess for multivariate outliers and for the exclusion of outliers or influential cases. Sample size guidelines for multinomial logistic regression indicate a minimum of 10 cases per independent variable (Schwab, 2002).

Multinomial logistic regression is often considered an attractive analysis because; it does not assume normality, linearity, or homoscedasticity. A more powerful alternative to multinomial logistic regression is discriminant function analysis which requires these assumptions are met. Indeed, multinomial logistic regression is used more frequently than discriminant function analysis because the analysis does not have such assumptions. Multinomial logistic regression does have assumptions, such as the assumption of independence among the dependent variable choices. This assumption states that the choice of or membership in one category is not related to the choice or membership of another category (i.e., the dependent variable). The assumption of independence can be tested with the Hausman-McFadden test. Furthermore, multinomial logistic regression also assumes non-perfect separation. If the groups of the outcome variable are perfectly separated by the predictor(s), then unrealistic coefficients will be estimated and effect sizes will be greatly exaggerated. There are different parameter estimation techniques based on the inferential goals of multinomial logistic regression analysis. One might think of these as ways of applying multinomial logistic regression when strata or clusters are apparent in the data. Unconditional logistic regression (Breslow & Day, 1980) refers to the modeling of strata with the use of dummy variables (to express the strata) in a traditional

logistic model. Here, one model is applied to all the cases and the stata are included in the model in the form of separate dummy variables, each reflecting the membership of cases to a particular stata.

Conditional logistic regression (Breslow & Day, 1980; Vittinghoff, Shiboski, Glidden, & McCulloch, 2005) refers to applying the logistic model to each of the stata individually. The coefficients of the predictors (of the logistic model) are conditionally modeled based on the membership of cases to a particular stata. Marginal logistic modeling (Vittingh off, Shiboski, Glidden, & McCulloch, 2005) refers to an

methods explore a relation between two or more predictor (independent) variables and one outcome (dependent) variable. The model describing the relationship expresses the predicted value of the outcome variable as a sum of products, each product formed by multiplying the value and coefficient of the independent variable. The coefficients are obtained as the best mathematical fit for the specified model. A coefficient indicates the impact of each independent variable on the outcome variable adjusting for all other independent variables. The model serves two purposes: (1) it can predict the value of the dependent variable for new values of the independent variables, and (2) it can help describe the relative contribution of each independent variable to the dependent variable, controlling for the influences of the other independent variables.

The four main multivariable methods used in health science are linear regression, logistic regression, discriminant analysis, and proportional hazard regression. The four multivariable methods have many mathematical similarities but differ in the expression and format of the outcome variable. In linear regression, the outcome variable is a continuous quantity, such as blood pressure. In logistic regression, the outcome variable is usually a binary event, such as alive versus dead, or case versus control. In discriminant analysis, the outcome variable is a category or group to which a subject belongs. For only two categories, discriminant analysis produces results similar to logistic regression.

In proportional hazards regression, the outcome variable is the duration of time to the occurrence of a binary "failure "event (for example, death) during a follow-up period of observation. The logistic regression is the most popular multivariable method used in health science (Tetrault, Sauler,Wells, & Conca to, 2008). In this article logistic regression (LR) will be presented from basic concepts to interpretation.

Concepts related to Logistic regression

Logistic regression sometimes called the logistic model or logit model, analyses the

relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve. There are two models of logistic

regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. When the dependent variable is not dichotomous and is comprised of more than two categories, a multinomial logistic regression can be employed. As an illustrative example, consider how coronary heart disease (HCV) can be predicted by the level of serum cholesterol. The probability of HCV increases with these rum cholesterol level. However, the relationship between HVC and serum cholesterol is nonlinear and the probability of HCV changes very little at the lower high extremes of serum cholesterol. This pattern is typical because

probabilities cannot lie outside the range from0 to 1. The relationship can be described as an „S"-shaped curve. The logistic model is popular because the logistic function, on which the logistic regression model is based, provides estimates in the range 0 to 1 and an appealing S shaped description of the combined effect of several risk factors on the risk for an event (Kleinbaum&Klein, 2010).

1. Odds

Odds of an event are the ratio of the probability that an event will occur to the probability that it will not occur. If the probability of an event occurring is p, the probability of the event not occurring is (1-p). Then the corresponding odds is a value given by

$$\text{Odds of } \{\text{Event}\} = \frac{p}{1-p}$$

Since logistic regression calculates the probability of an event occurring over the probability of an event not occurring, the impact of independent variables is usually explained in terms of odds. With logistic regression the mean of the response variable p in terms of an explanatory variable x is modelled relating p and x through the equation $p=\alpha+\beta x$. Unfortunately, this is not a good model because extreme values of x will give values of $\alpha+\beta x$ that does not fall between 0 and 1. The logistic regression solution to this problem is to transform the odds using the natural logarithm (Peng, Lee & Ingersoll, 2002). With logistic regression we model the natural log odds as a linear function of the explanatory variable:

$$logit(y) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \alpha+\beta x \qquad (1)$$

Where p is the probability of interested outcome and x is the explanatory variable. The parameters of the logistic regression are α and β. This is the simple logistic model.

Taking the antilog of equation (1) on both sides, one can derive an equation for the prediction of the probability of the occurrence of interested outcome as

p=P (Y=interested outcome/X= χ, a specific value).

$$= \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

Extending the logic of the simple logistic regression to multiple predictors, one mayconstruct a complex logistic regression as

$$logit(y) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots\ldots\ldots + \beta_k x_k$$

Therefore, p =P (Y=interested outcome/X1=x1,... ,Xk=xk)

$$= \frac{e^{\alpha+\beta_1 x_1+ \beta_2 x_2+\cdots\ldots\ldots\ldots\ldots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+ \beta_2 x_2+\cdots\ldots\ldots\ldots\ldots+\beta_k x_k}} = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+ \beta_2 x_2+\cdots\ldots\ldots\ldots\ldots+\beta_k x_k)}}$$

2. Odds ratio

The odds ratio (OR) is a comparative measure of two odds relative to different events. For two events A and B, the corresponding odds of A occurring relative to B occurring is

$$\text{Odds ratio } \{A \text{ vs. } B\} = \frac{odds(A)}{odds(B)} = \frac{P_A/1-P_A}{P_B/1-P_B}$$

An OR is a measure of association between an exposure and an out- come. The OR represents the odds that an outcome (e.g., disease or disorder) will occur given a particular exposure (e.g. health behavior, medical history), compared to the odds of the outcome occurring in the absence of that exposure.

When a logistic regression is calculated, the regression coefficient (b1) is the estimated increase in the logged odds of the outcome per unit increase in the value of the independent variable. In other words, the exponential function of the regression coefficient (eb1) is the OR associated with a one unit increase in the independent variable. The OR can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome. OR=1indicates exposure does not affect odds of outcome. OR>1 indicates exposure associated with higher odds of outcome. OR>1 indicates exposure associated with higher odds of outcome. OR<1 indicates exposure associated with lower odds of outcome. For example, the variable smoking is coded as 0 (=no smoking) and 1 (=smoking), and the odds ratio for this variable 3.2. Then, the odds for a positive outcome in smoking cases are 3.2 times higher than in non-smoking cases.

Logistic regression is one way to generalize the OR beyond two binary variables (Peng &So, 2002). Suppose we have a binary response variable Y and a binary predictor variable X, and in addition we have

other predictor variables that may or may not be binary. If we use multiple logistic regressions to regress Y on X, then the estimated coefficient ˆ for X is related to a conditional OR. Specifically, at the population level

$$e^{\widehat{\beta}_x} = \frac{P(Y=1|X=1,Z_1,\ldots\ldots,Z_k)/P(Y=0|X=1,Z_1,\ldots\ldots,Z_k)}{P(Y=1|X=0,Z_1\ldots\ldots,Z_k)/P(Y=0|X=0,Z_1,\ldots\ldots,Z_k)}$$

So $e^{\widehat{\beta}_x}$ is an estimate of this conditional odds ratio. The interpretation of $e^{\widehat{\beta}_x}$ is an estimate of the OR between Y and X when the values of $Z_1, \ldots\ldots, Z_k$ are held fixed.

The logistic curve

Logistic regression is a method for fitting a regression curve, y =f(x), when y consists of binary coded (0, 1--failure, success) data. When the response is a binary (dichotomous) variable and x is numerical, logistic regression fits a logistic curve to the relationship between x and y. Logistic curve is an S-shaped or sigmoid curve, often used to model population growth (Eberhardt&Breiwick, 2012). A logistic curve starts with slow, linear growth, followed by exponential growth, which then slows again to a stable rate. A simple logistic function is defined by the formula

$$y = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

This is graphed in Figure 1.

To provide flexibility, the logistic function can be extended to the form

$$y = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

Where α and β determine the logistic intercept and slope. Logistic regression fits α and β, the regression coefficients. Figure 1 shows logistic function when α and β are 0 and 1, respectively.The logistic or logit function is used to transform an 'S'-shaped curve into an approximately straight line and to change the range of the proportion from 0 – 1 to -∞ to ∞

$$logit(y) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \alpha+\beta x$$

Where p is the probability of interested outcome, α is the intercept parameter, β is a regression coefficient, and χ is a predictor.

ASSUMPTIONS OF LOGISTIC REGRESSION

Logistic regression does not require many of the principle assumptions of linear regression models that are based on ordinary least squares method–particularly regarding linearity of relationship between the dependent and independent variables, normality of the error distribution, homoscedasticity of the errors, and measurement level of the independent variables. Logistic regression can handle non-linear relationships between the dependent and independent variables, because it applies a non-linear log transformation of the linear regression. The error terms (the residuals) do not need to be multivariate normally distributed–although multivariate normality yields a more stable solution. The variance of errors can be heteroscedastic for each level of the independent variables. Logistic regression can handle not only continuous data but also discrete data as independent variable

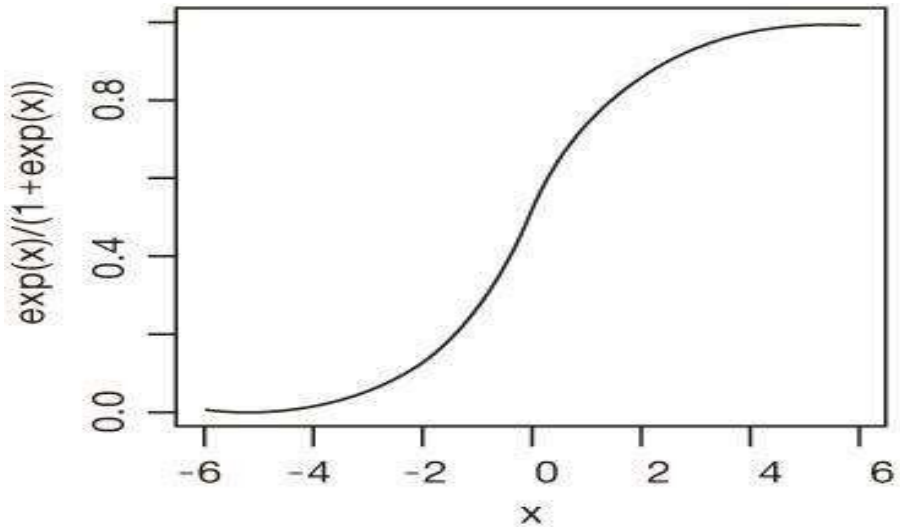


Figure 1: Graph of logistic curve where α=0 and β=1.

However, some other assumptions still apply (Bewick, Cheek, & Ball, 2005; Peng & So, 2002): First, logistic regression requires the dependent variable to be discrete mostly dichotomous. Second, since logistic regression estimates the probability of the event occurring (P(Y=1)), it is necessary to code the dependent variable ccordingly. That is the desired outcome should be coded to be 1. Third, the model should be fitted correctly. It should not be over fitted with the meaningless variables included. Also, it should not be under fitted with meaningful variable not included. Fourth, logistic regression requires each observation to be independent. Also, the model should have little or no multicollinearity. That is, independent variables are not linear functions of each other. Fifth, whilst logistic regression does not require a linear relationship between the dependent and independent variables, it requires that the independent variables are linearly related to the log odds of an event. Lastly, logistic regression requires large sample sizes because maximum likelihood

estimates are less powerful than ordinary least squares used for estimating the unknown parameters in a linear regression model.

STUDY DESIGN OF LOGISTIC REGRESSION

Logistic regression model corresponds to data from either a cross-sectional, prospective, or retrospective case-control study (Hsieh, Bloch & Larsen, 1998). In the cross-sectional studies a random sample is taken from a population, and outcome and explanatory variables are collected simultaneously. The fitted probabilities from a logistic regression model are then estimates of proportions of an outcome in the underlying population. In the prospective studies, a set of subjects are selected and the explanatory variables are observed. Subjects are then followed over some standard period (e.g., a month or a year) or episode (hospital stay) to determine the response outcome. In this case, the fitted probabilities are estimates of the probability of the response outcomes occurring. In the retrospective case-control studies, separate samples of case and control groups are first assembled and potential explanatory variables are collected later often through their recollections. In this case the fitted probabilities do not have a direct interpretation since they are determined by the relative sample sizes for case and control groups. However, odds ratios can be estimated based on logistic regression.

FITTING THE LOGISTIC REGRESSION MODEL:

Although logistic regression model, logit $(y)=\alpha+\beta\chi$ looks similar to a simple linear regression model, the underlying distribution is binomial and the parameters, α and β cannot be estimated using least square method. Instead, the parameters are usually estimated using the method of maximum likelihood of observing the sample values (Menard, 2001). Maximum likelihood will provide values of α and β which maximize the probability of obtaining the data set. It requires iterative computing with computer software. The likelihood function is used to estimate the probability of observing the data, given the unknown parameters (α and β). A "likelihood" is a probability that the observed values of the dependent variable may be predicted from the observed values of the independent variables. The likelihood varies from 0 to 1 like any other probabilities. Practically, it is easier to work with the logarithm of the likelihood function. This function is known as the log-likelihood. Log-likelihood will be used for inference testing when comparing several models. The log likelihood varies from 0 to $-\infty$ (it is negative because the natural log of any number less than 1 is negative).In logistic regression, we observe binary outcome and predictors, and we wish to draw inferences about the probability of an event in the population. Suppose in a population from which we are sampling, each individual has the same probability p that an event occurs. For each individual in our

sample of size n, Yi =1 indicates that an event occurs for the i[th] subject, otherwise, Yi =0. The observed data are $Y_1, \dots\dots, Y_n$ and $X_1, \dots\dots, X_n$

The joint probability of the data is given by

$$L = \prod_{i=1}^{n} \pi(x)^{Yi}(1-\pi(x))^{1-Yi}$$

Natural logarithm of the likelihood is

$$l = log(L) = \sum_{i=1}^{n} Y_i \, log\big(\pi(x)\big) + (n - \sum_{i=1}^{n} Y_i) log\big(1-(x)\big)$$

In which

$$\pi(x) = p(y/x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

To find the parameters that maximizes L. We differentiate L w.r.t α &β and equate into zero.

$$\sum_{i=1}^{n}\big(y_i - \pi(x)\big) = 0 \quad \dots\dots\dots\dots\dots\dots \quad (1)$$

$$\sum_{i=1}^{n} x_i\big(y_i - \pi(x)\big) = 0 \quad \dots\dots\dots\dots\dots\dots \quad (2)$$

In Logistic regression equations (1) & (2) are non- linear in Thus we use special method for their solution these methods are iterative in nature. In similar way we estimated parameters $\beta_1, \beta_2, \dots\dots\dots, \beta_k$ in multivariate logistic regression.

EVALUATION OF A LOGISTIC REGRESSION MODEL

There are several parts involved in the evaluation of the logistic regression model. First, the overall model (relationship between all of the independent variables and dependent variable) needs to be assessed. Second, the importance of each of the independent variables needs to be assessed. Third, predictive accuracy or discriminating ability of the model needs to be evaluated. Finally, the model needs to be validated

1.Overall model evaluation

The likelihood ratio test

Overall fit of a model shows how strong a relationship between all of the independent variables, taken together, and dependent variable is. It can be assessed by comparing the fit of the two models with and without the independent variables. A logistic regression model with the k independent variables (the given

model) is said to provide a better fit to the data if it demonstrates an improvement over the model with no independent variables (the null model). The overall fit of the model with k coefficients can be examined via a likelihood ratio test which tests the null hypothesis

$$H_0 = \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

To do this, the deviance with just the intercept (-2 log likelihood of the null model) is compared to the deviance when the k independent variables have been added (-2 log likelihood of the given model). Likelihood of the null model is the likelihood of obtaining the observation if the independent variables had no effect on the outcome. Likelihood of the given model is the likelihood of obtaining the observations with all independent variables incorporated in the model.

The difference of these two yields a goodness of fit index G, $\chi 2$ statistic with k degrees of freedom (Bewick, Cheek, & Ball, 2005). This is a measure of how well all of the independent variables affect the outcome or dependent variable.

G=$\chi 2$ = (-2 log likelihood of null model) -(-2 log likelihood of given model)

An equivalent formula sometimes presented in the literature is

$$= -2log\frac{likelihood\ of\ \ the\ null\ model}{likelihood\ of\ the\ given\ model}$$

Where the ratio of the maximum likelihood is calculated before taking the natural logarithm and multiplying by -2. The term "likelihood ratio test" is used to describe this test. If the p value for the overall model fit statistic is less than the conventional 0.05, then reject H0 with the conclusion that there is evidence that at least one of the independent variables contributes to the prediction of the outcome.

2. Statistical significance of individual regression coefficients

If the overall model works well, the next question is how important each of the independent variables is. The logistic regression coefficient for the i^{th} independent variable shows the change in the predicted log odds of having an outcome for one unit change in the ith independent variable, all other things being equal. That is, if the ith independent variable is changed1 unit while all of the other predictors are held constant, log odds of outcome is expected to change bi units. There are a couple of different tests designed to assess the significance of an independent variable in logistic regression, such as the likelihood ratio test.

STEPWISE LOGISTIC REGRESSION

Stepwise logistic regression is most often used in situation where the "important" independent variables are not known and associations with the outcome not well understood. In these instances, most studies will

collect many possible independent variables and screen them for significance. Stepwise Logistic Regression offers a fast and effective means of screening a large number of variables, and simultaneously fit a number of logistic regression equations. There are two basic forms of stepwise logistic regression: forward inclusion and backward elimination. In forward logistic regression all independent variables are initially withheld from the model. At subsequent steps in the procedure, those variables determined to be significant are added to the model while all others are withheld. Just opposite occurs in backward logistic regression in which all independent variables are initially included in the model. At subsequent steps in the procedure, those variables determined to be insignificant are eliminated from the model until the remaining variables are all deemed "important". In Stepwise logistic regression selection or deletion of variables from the model is based on a statistical algorithm that checks for "importance" of variables, and either includes or excludes them on the basis of a fixed decision rule. The likelihood ratio chi-square test is used to assess significance in logistic regression since the errors are assumed to follow binomial distribution.

This test assigns p-values to each variable to assess significance. Therefore, the most important variable is the one with the smallest p- value. An important element of Stepwise logistic regression is selection of removal and entry criteria to determine variable significance.

Akaike Information Criteria (AIC)

When several models are available, one compare the models performance based on several likelihood measures which have been proposed in statistical literatures.

The AIC is a measure of the relative quality of a statistical model for a given set of data. AIC provides a means for model selection. AIC is founded on information theory; it offers a relative estimate of the information lost when a given model is used to represent the process that generates the data.

Then AIC penalizes a model with large number of parameters and it is defined as,

$$AIC = -2ln(L) + 2p$$

Where ln(L) denotes the fitted log likelihood and p is the number of parameters. From the given set of data, the preferred model is the one with the minimum AIC value.

Null Deviance and Residual Deviance

Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model onadding independent variables. Lower the value, better the model.

Evaluation matric

CLASSIFICATION TABLE

Actual	Predicted				
	0	1	2	3	4
0	True Class 0	False Class 0	False Class 0	False Class 0	False Class 0
1	False Class 1	True Class 1	False Class 1	False Class 1	False Class 1
2	False Class 2	False Class 2	True Class 2	False Class 2	False Class 2
3	False Class 3	False Class 3	False Class 3	True Class 3	False Class 3
4	False Class 4	False Class 4	False Class 4	False Class 4	True Class 4

Classifier accuracy measures:

- The accuracy of the classifier on a given test set is the percentage of test tuples that are correctly classified by the classifier this can be represented in the table called confusion matrix.
- ✓ **True positives** (TP): These refer to the positive tuples that were correctly labelled by the classifier. Let TP be the number of true positives.
- ✓ **True negatives** (TN): These are the negative tuples that were correctly labelled by the classifier. Let TN be the number of true negatives.
- ✓ **False positives** (FP): These are the negative tuples that were incorrectly labelled as positive. Let FP be the number of false positives.
- ✓ **False negatives** (FN): These are the positive tuples that were mislabelled as negative. Let FN be the number of false negatives.

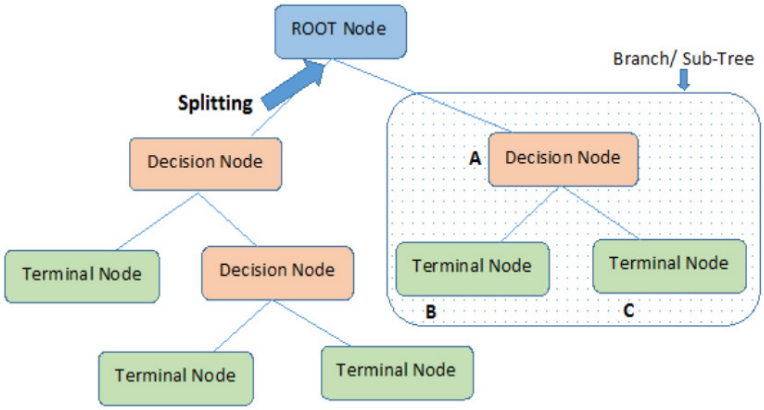
Accuracy = TP+TN/TP+TN+FP+FN

Cross Validation

Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate. When dealing with a Machine Learning task, you have to properly identify the problem so that you can pick the most suitable algorithm which can give you the best score.

2.3 DECISION TREE

A **decision tree** is a flowchart-like tree structure, where each **internal node** (non- leaf node) denotes a test on an attribute, each **branch** represents an outcome of the test, and each **leaf node** (or terminal node) holds a class label. The topmost node in a tree is the **root** node.



Given a tuple *X*, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree.

A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision trees can easily be converted to classification rules.

The construction of decision tree classifiers does not require any domain knowledge or parameter setting,

and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle multidimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand. Decision tree induction algorithms have been used for classification in many application areas such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology. Decision trees are the basis of several commercial rule inductionsystems.

Splitting scenarios

There are three possible scenarios, Let A be the splitting attribute. A has v distinct values, a1, a2,…,av , based on the training data.

A is discrete-valued: In this case, the outcomes of the test at node N correspond directly to the known values of A. A branch is created for each known value, a_j of A and labeled with that value. Partition Dj is the subset of class-labeled tuples in D having value a_j of A.

A is continuous-valued: In this case, the test at node N has two possible outcomes, corresponding to the conditions A split point and A > split point, respectively, where split point is the split-point returned by Attribute selection method as part of the splitting criterion. (In practice, the split-point, a, is often taken as the midpoint of two known adjacent values of A and therefore may not actually be a pre-existing value of A from the training data.) Two branches are grown from N and labelled according to the previous outcomes. The tuples are partitioned such that D1 holds the subset of class labelled tuples in D for which A split point, while D2 holds the rest.

A is discrete-valued and a binary tree must be produced (as dictated by the attribute selection measure or algorithm being used): The test at node N is of the form "A $\in \boldsymbol{S_A}$?," where $\boldsymbol{S_A}$ is the splitting subset for A, returned by Attribute selection method as part of the splitting criterion. It is a subset of the known values of A.

Attribute selection measures

- The attribute selection measure provides a ranking for each attribute describing the given training tuples. The attribute having the best score for the measure is chosen as the splitting attribute for the given tuples.
- If the splitting attribute is continuous-valued or if we are restricted to binary trees, then,

respectively, either a split point or a splitting subset must also be determined as part of the splitting criterion.

- The tree node created for partition D is labelled with the splitting criterion, branches are grown for each out-come of the criterion, and the tuples are partitioned accordingly.
- This section describes three popular attribute selection measures—information gain, gain ratio, and Gini index.

Gini Index

The Gini index is used in CART. Using the notation previously described, the Gini index measures the impurity of D, a data partition or set of training tuples, as

$$\text{Gini (D)} = 1 - \sum_{i=1}^{m} \boldsymbol{P_i^2}$$

Where pi is the probability that a tuple in D belongs to class C_i. The Gini index considers a binary split for each attribute. Let's first consider the case where A is a discrete-valued attribute having v distinct values, a_1, a_2, …..,a_v , occurring in D. To determine the best binary split on A, we examine all the possible subsets that can be formed using known values of A.When considering a binary split, we compute a weighted sum of the impurity of each resulting partition. For example, if a binary split on A partitions D into $\boldsymbol{D_1}$and $\boldsymbol{D_2}$, the Gini index of D given that partitioning is

$$\text{GiniA(D)} = \frac{\boldsymbol{D_1}}{\boldsymbol{D}} \text{ Gini}(\boldsymbol{D_1}) + \frac{\boldsymbol{D_2}}{\boldsymbol{D}} \text{ Gini}(\boldsymbol{D_2})$$

The reduction in impurity that would be incurred by a binary split on a discrete- or continuous-valued attribute A is

$$\Delta Gini = \text{Gini (D)} - Gini\boldsymbol{A}^{(\boldsymbol{D})}$$

The attribute that maximizes the reduction in impurity (or, equivalently, has the minimum Gini index) is selected as the splitting attribute. This attribute and either its splitting subset (for a discrete-valued splitting attribute) or split-point (for a continuous-valued splitting attribute) together form the splitting criterion.

2.4 K-NN CLASSIFICATION

Nearest-Neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it.

When given an unknown tuple, a k-nearest-neighbour classifier searches the pattern space for the k-training tuples that are closest to the unknown tuple.

1. "Closeness" is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, $X_1=(x_{11},x_{12},\ldots\ldots,x_{1n})$ and $X_2=(x_{21},x_{22},\ldots\ldots,x_{2n})$

   $$\text{dist}(X_1, X_2) = \sqrt{\sum_{i=0}^{n}(X1i - X2i)\text{^}2}$$

2. We normalize the values of each attribute before using equation. This helps prevent attributes with initially large ranges from outweighing attributes with initially smaller ranges.
3. Min-Max normalization, for example, can be used to transform a value v of a numeric attribute A to v^0 in the range [0,1] by computing

   $$X_{new} = \frac{(x-min)}{(max-min)}$$

Determination of good value from k

The good value for k, the no. of nearest neighbor can be determined by the experimental starting with k=1 estimate the error rate of classifier.

The process can be repeated each time by implementing k for one or more valuable and so on. The value for k for which error rate i.e minimum may be taken as good value for k.

Another method for finding best value of k is given by,

$$k=\sqrt{number\ of\ training\ tuples}$$

2.5 .Support vector machine:

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. According to the SVM algorithm we find the points closest to the line from both the classes. These points are called support vectors. Now, we compute the distance between the line and the support vectors. This distance is called the margin. Our goal is to maximize the margin. The hyper plane for which the margin is maximum is the optimal hyper plane. Thus SVM tries to make a decision boundary in such a way that the separation between the two classes is as wide as possible.

Support vector machine algorithms that are used for mathematical and engineering problems including for example handwriting digit recognition, object recognition, speaker identification, face detections in images and target detection. Assume that we are given a set S of points xi ∈ Rn with i = 1,2,..., N. Each point xi belongs to either of two classes and thus is given a label yi∈ {-1,1}. The goal is to establish the equation of a hyper plane that divides S leaving all the points of the same class on the same side. SVM performs classification by constructing an N-dimensional hyper plane that optimally separates the data into two categories.

Support Vector Machines, a promising new method for the classification of both linear and nonlinear data. In a nutshell, a support vector machine (or SVM) is an algorithm that works as follows. It uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyper plane (that is, a "decision boundary" separating the tuples of one class from another). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyper plane.

The SVM finds this hyper plane using support vectors ("essential" training tuples) and margins (defined by the support vectors). To explain the mystery of SVMs, let's first look at the simplest case—a two-class problem where the classes are linearly separable. Let the data set D be given as (x_1, y_1), (x_2, y_2),…, (X|D|, y|D|),where Xi is the set of training tuples with associated class labels, yi. Each yi can take one of two values, either +1 or-1 (i.e., yi ε {+1, -1}). To aid in visualization, Generalizing to n dimensions, we want to find the best hyper plane. We will

use the term hyper plane to refer to the decision boundary that we are seeking, regardless of the number of input attributes.

The classification problem can be restricted to consideration of the two-class problem without loss of generality. In this problem the goal is to separate the two classes by a function which is induced from available examples. The goal is to produce a classifier that will work well on unseen examples, i.e. it generalizes well. Here there are many possible linear classifiers that can separate the data, but there is only one that maximizes the margin (maximizes the distance between it and the nearest data point of each class). This linear classifier is termed the optimal separating hyper plane. Intuitively, we would expect this boundary to generalize well as opposed to the other possible boundaries.

The separating line (2 – dimensional hyper plane) on the second picture is a decision plane which divides the objects into two subsets such that in each subset all elements are similar. Among the possible hyper planes, we select the one where the distance of the hyper plane from the closest data points (the "margin") is as large as possible. An intuitive justification for this criterion is: suppose the training data are good, in the sense that every possible test vector is 23 within some radius r of a training vector. Then, if the chosen hyper plane is at least r from any training vector it will correctly separate all the test data. By making the hyper plane as far as possible from any data, r is allowed to be correspondingly large. The desired hyper plane (that maximizes the margin) is also the bisector of the line between the closest points on the convex hulls of the two data sets.

Transforming the Data

• The mathematical equation which describes the separating boundary between two classes should be simple.

• This is why we map the data of input space into feature space. The mapping (rearranging) involves increasing dimension of the feature space.

• The data points are mapped from the input space to a new feature space before they are used for training or for classification.

A separating hyperplane can be written as

$W. X + b = 0$

Where W is a weight vector, namely, $W = \{w_1, w_2,\ldots,w_n\}$; n is the number of attributes; and b is a scalar, often referred to as a bias. To aid in visualization, let's consider two input attributes, A_1 and A_2. Training tuples are 2-D,

e.g., X = (x_1, x_2), where x_1 and x_2 are the values of attributes A_1 and A_2, respectively, for X. If we think of b as an additional weight, w_0, we can rewrite the above separating hyperplane as

$w_0+w_1x_1+w_2x_2 = 0$

Thus, any point that lies above the separating hyperplane satisfies

$w_0+w_1x_1+w_2x_2 > 0$

Similarly, any point that lies below the separating hyperplane satisfies

$w_0+w_1x_1+w_2x_2 < 0$

The weights can be adjusted so that the hyperplanes defining the "sides" of the margin can be written as

$H_1 : w_0+w_1x_1+w_2x_2+1$ for $y_i = +1$,

$H_2 : w_0+w_1x_1+w_2x_2 -1$ for $y_i = -1$

That is, any tuple that falls on or above H_1 belongs to class +1, and any tuple that falls on or below H_2 belongs to class -1. Combining the above two inequalities of Equations we get

$yi(w_0+w_1x_1+w_2x_2)$;

2.6 Naïve Bayes

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers.[6] Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests.

An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification

Probabilistic Model

Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $x=(x_1,x_2,\ldots xn)$ representing some *n* features (independent variables), it assigns to this instance probabilities

$p(C_k/x_1,x_2,..x_n)$

for each of K possible outcomes or classes

The problem with the above formulation is that if the number of features *n* is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. The model must therefore be reformulated to make it more tractable. Using Bayes' theorem, the conditional probability can be decomposed as

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

2.7 Bagging

Bootstrap aggregating, also called bagging (from bootstrap aggregating), is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in

statistical classification and regression. It also reduces variance and helps to avoid over fitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bootstrap Aggregation is a general procedure that can be used to reduce the variance for those algorithms that have high variance.

Theoretically, bagging is defined as follows.

1. Construct a bootstrap sample L ∗i = (Y ∗i , X∗i) (i = 1, . . . , n) according to the empirical distribution of the pairs Li = (Yi , Xi) (i = 1, . . . , n).

2. Compute the bootstrapped predictor ˆθ ∗ n (x) by the plug-in principle;

i.e., ˆθ ∗ n (x) = hn(L ∗ 1 , . . . , L∗ n), where ˆθn(x) = hn(L1, . . . , Ln).

3. The bagged predictor is ˆθn;B(x) = E ∗ [ˆθ ∗ n (x)].

Description of the technique as follows.

Given a standard training set of size *n*, bagging generates *m* new training sets , each of size *n′*, by sampling from *D* uniformly and with replacement. By sampling with replacement, some observations may be repeated in each. If *n′*=*n*, then for large *n* the set is expected to have the fraction (1 - 1/*e*) (≈63.2%) of the unique examples of *D*, the rest being duplicates. This kind of sample is known as a bootstrap sample. Then, *m* models are fitted using the above *m* bootstrap samples and combined by averaging the output (for regression) or voting (for classification).

bootstrap aggregation, is a technique used to reduce the variance of your predictions by combining the result of multiple classifiers modelled on different sub-samples of the same dataset

Here is the equation for bagging:

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$$

CHAPTER 4

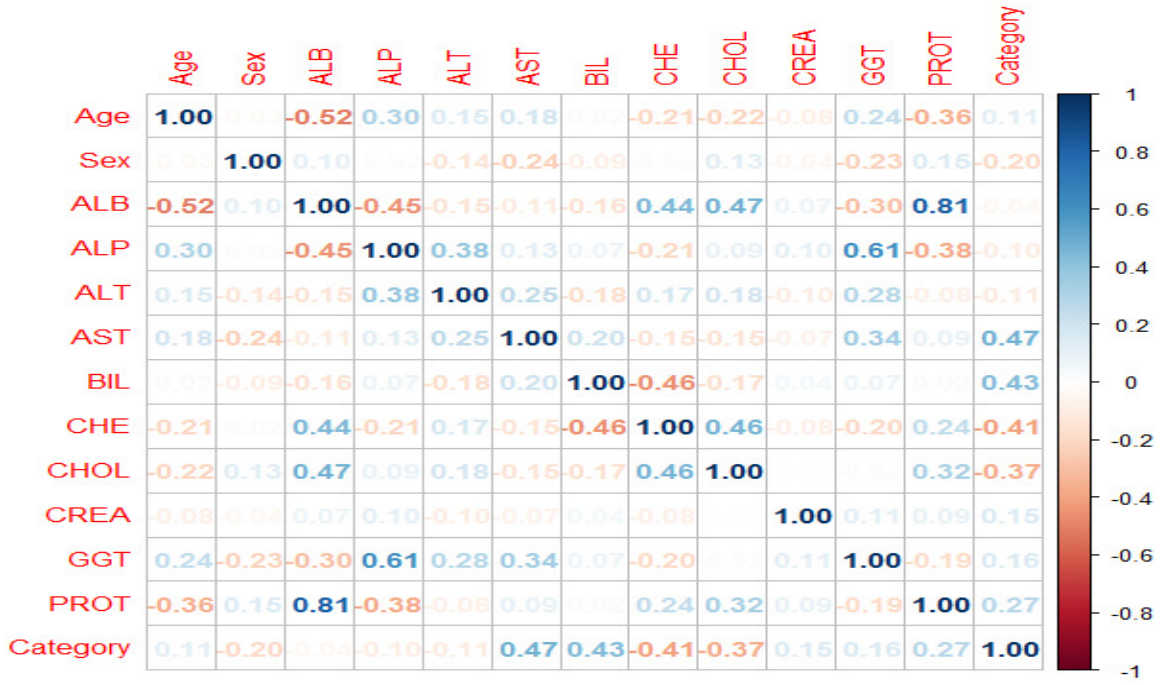
ANALYSIS AND DISCUSSION

DESCRIPTIVE STATISTICS

Variables	Minimum	Maximum	Mean	Median	Skewness	Kurtosis	Standard Deviation
Age	19.00	77.00	49.42	50.00	0.26	2.41	9.93
ALP	11.30	416.60	68.12	66.20	4.74	59.28	25.92
ALT	0.90	3.25	26.58	22.70	6.80	83.92	20.86
ALB	14.9	82.2	36.92	39.45	5.67	89.72	25.67
AST	10.60	324.00	33.77	25.70	5.23	36.42	32.87
BIL	0.80	209.00	11.02	7.10	8.07	82.11	17.41
CHE	1.42	16.41	8.20	8.26	-0.07	4.33	2.19
CHOL	1.43	9.76	4.56	4.98	0.38	3.68	1.12
CREA	8.00	1079.10	81.67	77.00	14.92	271.64	50.70
GGT	4.50	650.90	38.20	22.80	5.92	49.90	54.30
PROT	44.80	86.50	71.89	72.10	-1.06	6.60	5.35

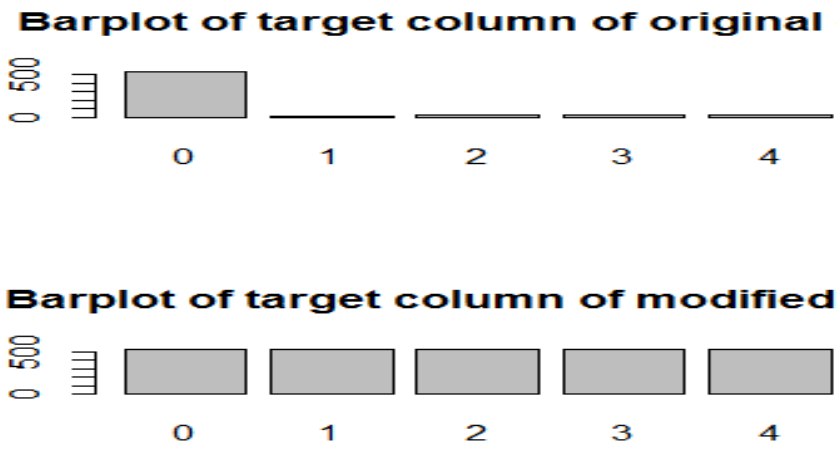
From the above table we observe that kurtosis value is greater than 3 except for Age so Alkaline Phosphatase, Alanine transaminase, Albumin, Aspartate aminotransferase, Bilateral, Cholinesterase, Creatinine, Cholesterol,Gamma glutamyl transferase and Protein is leptokurtic .

Cholinesteraseand Protein are negatively skewed.



We observe that we observe that there is moderate positive correlation between ALB and PROT, GGT and ALP, AST and GGT variables.

Bar Plot



From Bar plot we can observe that all classes are equally distributed by Smote oversampling.

FITTED MULTINOMIAL LOGISTIC REGRESSION MODEL FOR DATA

Coefficients:

	(Intercept)	Age	ALB	ALP	ALT	AST
1	55.282967	-0.05172488	-0.5166515	-0.082116313	0.04683196	0.2594239
2	0.104394	-0.18590339	0.2152807	-0.221713450	-0.04962342	0.2129029
3	12.097183	0.04883783	-0.0556746	-0.242760428	-0.01337645	0.2200959
4	2.517417	0.07317535	-0.4998176	-0.004226382	-0.03655936	0.2038545

	BIL	CHE	CHOL	CREA	GGT	PROT
1	-0.02334303	0.2824766	1.9673579	-0.001209204	0.032040003	-0.97172130
2	0.14392053	0.6287262	0.3039823	-0.043532978	0.042497910	-0.05807117
3	0.07640956	0.6529306	-0.8970854	-0.067156116	0.032529765	-0.09225632
4	0.07271817	-1.8788733	-0.1452176	0.015837203	-0.00255049	0.23428543

Residual Deviance: 709.3724

AIC: 805.3724

Based on wald test we can observe that Age, Sex, ALP, ALB,ALT, AST, BIL, CHE, CREA, GGT and PROT are the significant variables.

Stepwise Logistic Regression

We have applied backward stepwise logistic regression using step() command to select the best subsets. Based on the minimum AIC value we choose the best combination of variables.

Start: AIC=809.3

Category ~ Age + Sex + ALB + ALP + ALT + AST + BIL + CHE + CHOL + CREA + GGT + PROT

	Df	**AIC**
Sex	48	805.3724
<none>	52	809.3022
BIL	48	833.7447
CHOL	48	834.3384
PROT	48	845.7790
CREA	48	846.9484
ALT	48	847.0689
ALB	48	879.7267
AST	48	922.4288
GGT	48	949.6911
CHE	48	956.2723
Age	48	1146.1785
ALP	48	1210.1102

Step: AIC=805.37

Category ~ Age + Sex + ALP + ALT +ALB+ AST + BIL + CHE + CHOL + CREA + GGT + PROT

	Df	AIC
<none>	48	805.3724
CHOL	44	829.7605
BIL	44	830.1757
CREA	44	848.3274
ALT	44	848.3539
PROT	44	850.7082
ALB	44	875.5941
AST	44	919.4445
GGT	44	947.9934
CHE	44	956.4812
Age	44	1143.0810
ALP	44	1207.0979

From step wise regression we observe that Age, Sex, ALP, ALT, AST, BIL, ALB,CHE, CREA, GGT and PROT are the significant variables for class.

CLASSIFICATION USING MULTIPLE LOGISTIC REGRESSSION

CONFUSION MATRIX

Actual	Predicted				
	0	1	2	3	4
0	142	0	5	3	0
1	1	149	0	0	0
2	1	0	126	23	0
3	1	0	25	124	0
4	1	0	6	0	143

From Multiple Logistic classification table we observe that

1. 142 patients are correctly classified under "No-Hepatitis C" category & 8 are misclassified.
2. 149 patients are correctly classified under "Suspected Hepatitis C" category &1 are misclassified.
3. 126 patients are correctly classified under "Hepatitis C" category &24 are misclassified
4. 124 patients are correctly classified under "Fibrosis" category & 26 are misclassified
5. 143 patients are correctly classified under "Cirrhosis" category & 7 are misclassified

Class	0	1	2	3	4
Accuracy	0.97	0.99	0.87	0.89	0.99

• The class 0 accuracy of the MLR model is 0.97, it refers to 97% of the patients in test data are correctly classified as No-Hepatitis C, 99% patients are correctly classified as Suspected Hepatitis C,87% are correctly classified as Hepatitis C ,89% are correctly classified as Fibrosis & 98% patients are correctly classified as Cirhosis.

• Overall accuracy is 91%

DECISION TREE

CONFUSION MATRIX

Actual	Predicted				
	0	1	2	3	4
0	137	0	9	3	1
1	6	142	0	0	2
2	9	0	118	18	5
3	4	0	30	115	1
4	4	1	7	2	136

From Decision tree table we observe that

1. 137 patients are correctly classified under "No-Hepatitis C" category &13are misclassified.

2. 142 patients are correctly classified under "Suspected Hepatitis C" category &8 are misclassified.

3. 118 patients are correctly classified under "Hepatitis C" category &32are misclassified

4. 15 patients are correctly classified under "Fibrosis" category &35 are misclassified

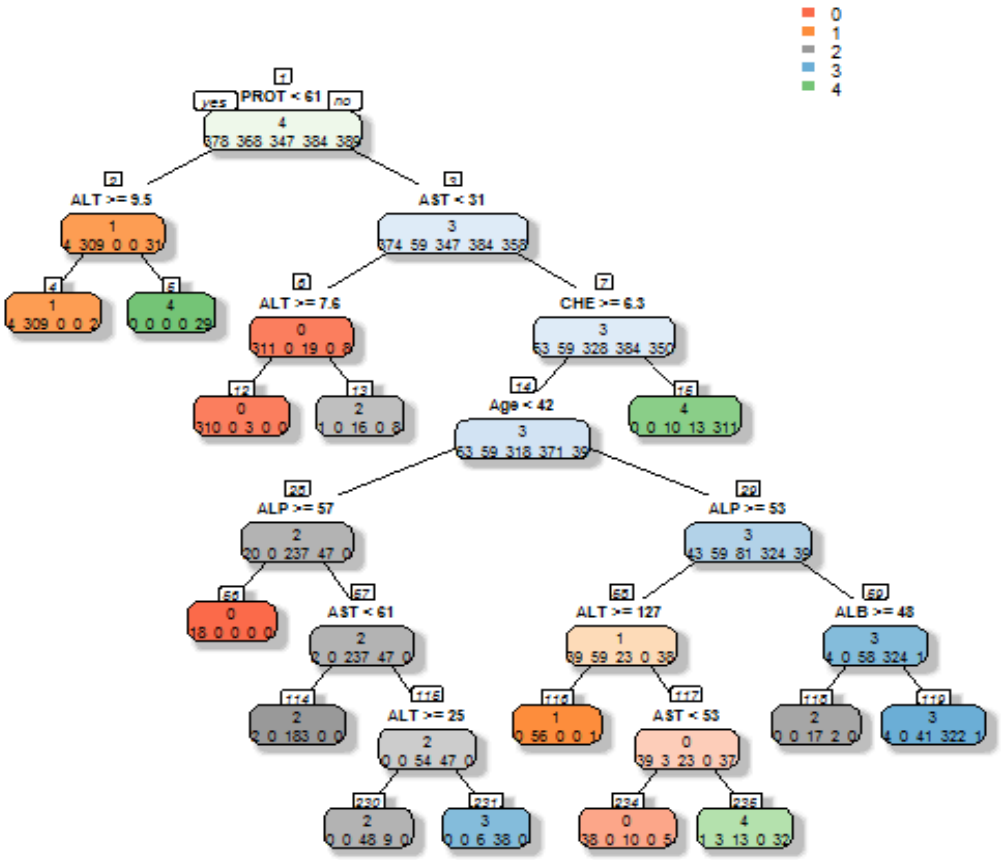
5. 36 patients are correctly classified under "Cirrhosis" category &14are misclassified

Class	0	1	2	3	4
Accuracy	0.91	0.99	0.84	0.90	0.96

• The class 0 accuracy of the Decision tree model is 0.91, it refers to 99% of the patients in test data are correctly classified as No-Hepatitis C, 84% patients are correctly classified as Suspected Hepatitis C,90% are correctly classified as Hepatitis C ,96% are correctly classified as Fibrosis & 98% patients are correctly classified as Cirhosis.

• Overall accuracy is 87%

Decision tree model-based classification



- As seen in the above plot, The CART for predicting diabetes class for patients focuses on Protein, CHE, ALT and AST as the more important variables.

K-NN CLASSIFICATION

CONFUSION MATRIX

Actual	Predicted				
	0	1	2	3	4
0	138	4	5	2	1
1	0	146	0	2	2
2	3	0	144	2	1
3	0	0	8	141	1
4	0	2	5	0	143

From K-NN Classification table we observe that

1. 138 patients are correctly classified under "No-Hepatitis C" category &12 is misclassified.

2. 146 patients are correctly classified under "Suspected Hepatitis C" category &4 are misclassified.

3. 144 patients are correctly classified under "Hepatitis C" category &6 are misclassified

4.141patients are correctly classified under "Fibrosis" category &9 are misclassified

5.143 patients are correctly classified under "Cirrhosis" category &7are misclassified

Class	0	1	2	3	4
Accuracy	0.96	0.99	0.95	0.98	0.96

• The class 0 accuracy of the K-NN Classification model is 0.96, it refers to 96% of the patients in test data are correctly classified as No-Hepatitis C, 99% patients are correctly classified as Suspected Hepatitis C, 95% are correctly classified as Hepatitis C , 98% are correctly classified as Fibrosis & 96% patients are correctly classified as Cirhosis.

• Overall accuracy is 95%

Support Vector Machine

CONFUSION MATRIX

Actual	Predicted				
	0	1	2	3	4
0	143	2	4	0	1
1	1	142	0	5	1
2	1	0	144	0	5
3	2	0	5	143	0
4	1	0	8	0	141

From Support Vector Machinewe observe that

1.143 patients are correctly classified under “No-Hepatitis C” category &7are misclassified.

2.142 patients are correctly classified under “Suspected Hepatitis C” category &7 are misclassified.

3.144 patients are correctly classified under “Hepatitis C” category &6are misclassified

4.142 patients are correctly classified under “Fibrosis” category &7 are misclassified

5.141 patients are correctly classified under “Cirrhosis” category &9are misclassified

Class	0	1	2	3	4
Accuracy	0.98	0.99	0.95	0.98	0.99

• The class 0 accuracy of the Support Vector Machine model is 0.98, it refers to 98% of the patients in test data are correctly classified as No-Hepatitis C, 99% patients are correctly classified as Suspected Hepatitis C,95% are correctly classified as Hepatitis C ,98% are correctly classified as Fibrosis & 99% patients are correctly classified as Cirhosis.

• Overall accuracy is 97%

Naïve Bayes

CONFUSION MATRIX

Actual	Predicted				
	0	1	2	3	4
0	142	0	5	2	1
1	1	143	3	1	2
2	4	2	144	0	0
3	0	1	4	143	2
4	1	2	1	4	142

From Naïve Bayes we observe that

1.142 patients are correctly classified under “No-Hepatitis C” category &8are misclassified.

2.143 patients are correctly classified under “Suspected Hepatitis C”category&7 are misclassified.

3.144 patients are correctly classified under “Hepatitis C” category &6are misclassified

4.143 patients are correctly classified under “Fibrosis” category &7are misclassified

5.142 patients are correctly classified under “Cirrhosis” category &8are misclassified

Class	0	1	2	3	4
Accuracy	0.96	0.98	0.82	0.88	0.89

• The class 0 accuracy of theNaïve Bayes model is 0.96, it refers to 96% of the patients in test data are correctly classified as No-Hepatitis C, 98% patients are correctly classified as Suspected Hepatitis C,82% are correctly classified as Hepatitis C ,88% are correctly classified as Fibrosis & 89% patients are correctly classified asCirhosis.

• Overall accuracy is 85%

Bagging

CONFUSION MATRIX

Actual	Predicted				
	0	1	2	3	4
0	148	2	0	0	0
1	1	146	2	0	1
2	1	3	145	0	1
3	0	0	2	148	0
4	1	0	0	0	149

From Bagging we observe that

1.148 patients are correctly classified under “No-Hepatitis C” category & 2 are misclassified.

2.146 patients are correctly classified under “Suspected Hepatitis C” category & 4 are misclassified.

3.145 patients are correctly classified under “Hepatitis C” category & 5 are misclassified

4.148 patients are correctly classified under “Fibrosis” category & 2 are misclassified

5.149 patients are correctly classified under “Cirrhosis” category & 1 are misclassified

Class	0	1	2	3	4
Accuracy	0.98	0.99	0.97	0.99	0.99

• The class 0 accuracy of the Naïve Bayes model is 0.98, it refers to 98% of the patients in test data are correctly classified as No-Hepatitis C, 99% patients are correctly classified as Suspected Hepatitis C,97% are correctly classified as Hepatitis C ,99% are correctly classified as Fibrosis & 99% patients are correctly classified as Cirhosis.

• Overall accuracy is 98%

Overall Performance:

After fitting various models to this data Balanced Accuracy and overall accuracy is given below.

Models	Balanced Accuracy					Overall Accuracy
	Class 0	Class 1	Class 2	Class 3	Class 4	
Multiple Logistic Regression	0.97	0.99	0.87	0.89	0.99	0.91
Decision tree	0.91	0.99	0.84	0.9	0.96	0.87
K-NN Classification	0.96	0.99	0.95	0.98	0.96	0.95
Support Vector Machine	0.98	0.99	0.95	0.98	0.99	0.97
Naïve Bayes	0.96	0.98	0.82	0.88	0.89	0.85
Bagging	**0.98**	**0.99**	**0.97**	**0.99**	**0.99**	**0.98**

From the above table we observe that Multiple logistic regression, decision tree, K-Nearest neighbours, SVM, Naïve Bayes and Bagging classification models for predicting the HCV conditions. Based on the performance measures we conclude that Bagging performs better than all other model used for modelling HCV status because balanced accuracy of each class and over all accuracy is more than other model.

CHAPTER 5

FINDINGS AND CONCLUSION

In this project several innovative applications of statistical procedure are suggested for the analysis and prediction. Based on the result obtained, the following conclusions were drawn

The models were developed in two main stages, including model fitting and model performance calculation. To evaluate the performance of developed models, a confusion matrix was used for each model. It shows the performance of each model by plotting four possible outcomes: true positive, true negative, false positive, and false negative

1. After pre-processing the data, the dataset included 614 instances. Each patient record was characterized by 13 variables. All variables with their description are presented in table. All variables, except disease category and sex, are numerical. for the continuous variable we computed the mean standard deviation skewness and kurtosis. From the table we observed that except age, all variables are leptokurtic in nature. Based on the skewness we conclude that Cholinesterase and Proteins are negatively skewed and remaining variables are positively skewed. For the categorical variable we constructed bar plot.
2. After the developed models (SVM, KNN,M LR, Bagging, NB, and DT) were developed with the trained dataset, the developed classification models were evaluated with predefined metrics. The final performance of the model was measured as the aggregate of results obtained from the testing set. First, the confusion matrix was created for each classifier. Then, the accuracy for each model were calculated based on the confusion matrix.
3. Since data set contain many variables, we fitted a multivariate logistic regression model. We selected the significant risk factors associated with the HCV patient using Wald test and Stepwise Logistic regression We found that both the method identifies the same 11factorsasthesignificantfactorsnamely:Age,Alkaline phosphatise (ALP), Alanine transaminase(ALT),Aspartatetransaminase(AST),(ALB)Albumin,Bilateral(BIL),CHE(Cholinesterase),Cholesterol(CHOL),Creatinine(CREA),Gammaglutamyltransferase(GGT),Protein(PROT).Fromthelikelihoodratiotestweconcludethattheoverallmodelisgoodfit.

4. To choose the best performing model for use as a proposed mode, the performance of each model but not the average performance received much focus. In addition to other metrics, in comparing the LR, SVM, KNN, Bagging, NB, and DT models using their true positive rate (TPR) and false-positive rate (FPR) values at the optimal cut off points, Bagging had the highest accuracy. This means that Bagging was the better performing model.

REFERENCE

1. Anderson,J.A.(1972) : SeparateSampleLogisticDiscrimination.Biometrika,59,19-35
2. Azzalini,A.&DallaValle,A.(1996),'Themultivariateskew-normaldistribution',Biometrika83(4), 715– 726.
3. Beauducel, A., &Wittmann, W. W. (2005). Simulation study on fit indices in confirmatory factor analysis based on data with slightly distorted simple structure. Structural Equation Modeling: A Multidisciplinary Journal,12(1),4175
4. Bentler,P.M.(1983)," Simultaneous equation systems as moment structure models", Journal of Econometrics 22(1-2), 13–42.
5. Bollen, K. A. (1996b), A limited-information estimator for LISREL models with and without heteroscedastic errors.
6. Chen, F. F. (Chou,C. P., &Bentler, P. M. (1995). Estimates andtests in structural equation modelling. In R. H. Hoyle (Ed.), Structural equation modeling: Concepts,issues, and applications(pp. 37-55):Sage Publications.
7. Cudeck, S. Du Toit& D. Sörbom (Eds.), Structural equation models: Present and future(pp. 139-168). Chicago:Scientific SoftwareInternationalInc2007). Sensitivity of goodnessoffitindexestolackofmeasurementinvariance.StructuralEquationModeling:AMultidisciplinaryJournal,14(3), 464-504.
8. David W. Hosmerand Stanley Lemshow (2000): Applied Logistic Regression, Second Edition,Wiley Series in Probability and Statistics.
9. Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modelling fit indices. Structural Equation Modelling: A Multidisciplinary Journal, 2(2),119-143.
10. Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification non structural equation modelling fit indexes. Structural Equation Modelling: A Multidisciplinary Journal, 6(1), 56-83.
11. Kline, R. B. (2011). Principles and practice of structural equation modeling.3rdedition.NewYork: TheGuilfordPress.61
12. Park, Hyeoun –Ae (2004),'An Introduction to Logistic Regression: From Basic Concept to Interpretation with Particular Attention to Nursing Domain", J Korean AcadNurs Vol.43No.2,154-164.

APPENDIX

```
A=read.csv("hh.csv",header=TRUE);A
table(A$Category)
summary(A)
E=read.csv("hep.csv");E
attach(E)
summary(E)
head(E)
(table(A$Category)/nrow(E))*100
cor(E)
table(E["Category"])

par(mfrow=c(2,1))
barplot(table(A$Category),main="Barplot of target column of original data")
barplot(table(E$Category),main="Barplot of target column of modified data")

##descriptive statistics

#Skewness
library(moments)
s11=A$Age
s12=A$ALP
s13=A$ALT
s14=A$AST
s15=A$BIL
s16=A$CHE
s17=A$CREA
s18=A$GGT
s19=A$PROT
s20=A$CHOL
skewness(s11)
skewness(s12)
skewness(s13)
```

```
skewness(s14)
skewness(s15)
skewness(s16)
skewness(s17)
skewness(s18)
skewness(s19)
skewness(s20)
#kurtosis
kurtosis(s11)
kurtosis(s12)
kurtosis(s13)
kurtosis(s14)
kurtosis(s15)
kurtosis(s16)
kurtosis(s17)
kurtosis(s18)
kurtosis(s19)
kurtosis(s20)

#Standard Deviation
sd(s11)
sd(s12)
sd(s13)
sd(s14)
sd(s15)
sd(s16)
sd(s17)
sd(s18)
sd(s19)
sd(s20)
#mode
m1=(A$Sex)
md1=function(m1){
um1=unique(m1)
```

```
um1[which.max(tabulate(match(m1,um1)))]
}
md1(m1)


m2=(A$Age)
md2=function(m2){
um2=unique(m2)
um2[which.max(tabulate(match(m2,um2)))]
}
md2(m2)


m3=(A$ALB)
md3=function(m3){
um3=unique(m3)
um3[which.max(tabulate(match(m3,um3)))]
}
md3(m3)

m4=(A$ALP)
md4=function(m4){
um4=unique(m4)
um4[which.max(tabulate(match(m4,um4)))]
}
md4(m4)

m5=(A$ALT)
md5=function(m5){
um5=unique(m5)
um5[which.max(tabulate(match(m5,um5)))]
}
md5(m5)
```

```
m6=(A$AST)
md6=function(m6){
um6=unique(m6)
um6[which.max(tabulate(match(m6,um6)))]
}
md6(m6)

m7=(A$BIL)
md7=function(m7){
um7=unique(m7)
um7[which.max(tabulate(match(m7,um7)))]
}
md7(m7)


m8=(A$CHE)
md8=function(m8){
um8=unique(m8)
um8[which.max(tabulate(match(m8,um8)))]
}
md8(m8)

m9=(A$CHOL)
md9=function(m9){
um9=unique(m9)
um9[which.max(tabulate(match(m9,um9)))]
}
md9(m9)


m10=(A$CREA)
md10=function(m10){
um10=unique(m10)
um10[which.max(tabulate(match(m10,um10)))]
```

```
}
md10(m10)

m11=(A$GGT)
md11=function(m11){
um11=unique(m11)
um11[which.max(tabulate(match(m11,um11)))]
}
md11(m11)

m12=(A$PROT)
md12=function(m12){
um12=unique(m12)
um12[which.max(tabulate(match(m12,um12)))]
}
md12(m12)

m13=(A$CHOL)
md13=function(m13){
um13=unique(m13)
um13[which.max(tabulate(match(m13,um13)))]
}
md13(m13)



#data spliting
ds=round(dim(E)[1]*70/100);ds
dt=sample(c(1:dim(E)[1]),ds);dt
train=E[dt,];train
n1=nrow(train);n1
test=E[-dt,];test
n2=nrow(test);n2

table(E$Category)
```

```
summary(E)
#install.packages("corrplot")
library(corrplot)
library(dplyr)
cor=round(cor(E),2)
corrplot(cor,method="number")

##decision tree
library(caret)
library(rpart)
library(rpart.plot)
library(rpart)
Decision_tree=rpart(Category~.,data=train,method="class");Decision_tree
rpart.plot(Decision_tree,shadow.col="gray",nn=TRUE,type=1,extra=1,fallen.leaves=F
ALSE)
pred_decision=predict(Decision_tree,test,type="class")
Table2=table(test$Category,pred_decision);Table2
A1=sum(diag(Table2))/sum(Table2);A1
confusionmatrix2=confusionMatrix(pred_decision,as.factor(test$Category));confusion
matrix2



#KNN
trnlb=train[,"Category"]
testlb=test[,"Category"]
library(class)
#install.packages("gmodels")
library(gmodels)
library(e1071)
library(caTools)
#install.packages("varImp")
library(varImp)
knnpred=knn(train=train,test=test,cl=trnlb,k=10);knnpred
```

```
confusionmatrix4=confusionMatrix(knnpred,as.factor(test$Category));confusionmatri
x4


#NaiveBayes
modelNN=naiveBayes(Category~.,data=train)
predictionNN=predict(modelNN,test,type=("class"))
Table10=table(Actual=test$Category,Predicted=predictionNN)
A7=sum(diag(Table10))/sum(Table10)
confusionmatrix5=confusionMatrix(predictionNN,as.factor(test$Category));confusion
matrix5



#support vector machine
library(e1071)
as.matrix(train)
svm=svm(formula=train$Category~.,data=train,type="C-
classification",kernal="linear");svm
p5=predict(svm,test,type="class");p5
table(p5)
confusionmatrix6=confusionMatrix(p5,as.factor(test$Category));confusionmatrix6


#multinomial model
require(foreign)
require(nnet)
require(ggplot2)
require(reshape2)
#Setting the reference
Level1=relevel(as.factor(train$Category),ref="1")
library(nnet)
multinom_model=multinom(Category~.,data=train,family="multinom")
step(multinom_model)
```

```
exp(coef(multinom_model))
probability.table= round(fitted(multinom_model),4);probability.table
multipred=predict(multinom_model,test,type="class")
tb7=table(test$Category,multipred);tb7
multiacc=sum(diag(tb7))/sum(tb7)
CM7=confusionMatrix(as.factor(multipred),as.factor(test$Category));CM7


#Bagging
library(ipred)
library(rpart)
bag=bagging(as.factor(Category)~.,train)
bag_pred=predict(bag,test)
con6=with(test,table(Category,bag_pred))
A6=sum(diag(con6))/sum(con6)
confusionmatrix6=confusionMatrix(bag_pred,as.factor(test$Category));confusionmati
x6


#SMOTE in Python

importnumpy as np
import pandas as pd
importmatplotlib.pyplot as plt
fromsklearn import tree
fromsklearn.model_selection import train_test_split
fromsklearn.tree import DecisionTreeRegressor
fromsklearn.datasets import make_classification
fromsklearn.metrics import confusion_matrix,accuracy_score,balanced_accuracy_score,classification_report
fromsklearn.svm import SVC
from warnings import filterwarnings
filterwarnings("ignore")
df=pd.read_csv('D:\hepd.csv')
```

```
print(df)
df.head()
df['Category'].value_counts()

cols=['ID','Age','Sex','ALB','ALP','ALT','AST','BIL','CHE','CHOL','CREA','GGT','PROT']
for c in cols:
#print(df[c].describe())
print(c)
print("mean",df[c].mean)
print("var",df[c].var())
print("sd",df[c].std())
df.info()
cols=df.columns.to_list()
cols.remove('Category')
(df['Category'].value_counts()/len(df))*100
fromimblearn.over_sampling import SMOTE
X=df[cols]
y=df['Category']
sm = SMOTE(random_state=42)
X_res, y_res = sm.fit_resample(X, y)
data=pd.DataFrame(X_res,columns=X.columns)
data['Category']=y_res
data['Category'].value_counts()
data
data.head(100)
X=data[cols]
y=data['Category']
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=25,shuffle=True,random_state=42)
data['Category'].value_counts()
data.to_csv ('D:\hepcdata', index=True)
```