A REPORT ON THE INTERNSHIP PROGRAMME

At

Kerala Institute of Local Administration (KILA)

Mulamkunnathukavu P.O, Thrissur-680 581, Kerala

Submitted to Mahatma Gandhi University Kottayam as part of the Third Semester Course on *Internship Report on Econometric Issue* for the Degree of:

Master of Arts in Econometrics

By:

MELVIN MATHEW

(Register Number: 200011008396)

Department of Economics

T.M. JACOB MEMORIAL GOVERNMENT COLLEGE MANIMALAKUNNU, KOOTHATTUKULAM, ERNAKULAM (DT), KERALA

May 31, 2022

DEPARTMENT OF ECONOMICS
TM JACOB MEMORIAL GOVT.COLLEGE MANIMALAKUNNU
KOOTHATTUKULAM, ERNAKULAM, KERALA

JULY 2022

CERTIFICATE

This is to certify that this internship report submitted by Mr. MELVIN MATHEW, as part of the third semester MA Econometrics programme for the course of Internship Report on Econometrics Issues (EM010306) and is in partial fulfilment for the Degree of Master of Arts in Econometrics (2020-2022), is a record of work done by the candidate. Certified further that to the best of my knowledge the internship report represents an independent work done by the candidate and does not form part of any other thesis or dissertation.

PRINCIPAL	**HEAD OF THE DEPARTMENT**	**SUPERVISOR**
DR. ROY SCARIA	**DR. TOJO JOSE**	**DR. MONISH JOSE** Assistant Professor in Public Administration, KILA

PLACE: KOOTHATTUKULAM
DATE: 25-05-2022

DECLARATION

I, Melvin Mathew, hereby declare that this internship report is prepared and submitted under the guidance of DR. MONISH JOSE, Assistant Professor in Public Administration, KILA, and this is my original work.

I further declare that this Internship report has been submitted in partial fulfilment for the award of my Master of Arts Degree in Econometrics and has not previously formed partly or fully the basis for the award of any degree, diploma, fellowship, associate-ship or other similar title of recognition.

Melvin Mathew

PLACE: KOOTHATTUKULAM
DATE : 31-05-2022

ACKNOWLEDGEMENT

In respect I avail this opportunity to thank all those who have rendered their valuable cooperation, to make this study a success. First and foremost, I thank God Almighty for giving me the opportunity to do this Report. I am grateful for all his blessings and his great care.

I express my sincere thanks to Dr Monish Jose, Assistant professor, KILA and all other staffs of KILA those who helped in my internship.

Also, I'm grateful to Dr Roy Scaria, Principal of TM Jacob Govt College, Koothatukulam, Dr Tojo Jose, Head of the Economics Department for giving the opportunity to work hard and submit this report.

I would like to express my greatest gratitude to the people who have helped & supported me throughout my internship. I thank all our teachers of economics Department of my college for their constant help and support. I also wish to thank my family and all who have directly or indirectly contributed to the completion of my dissertation work successfully. Special thanks to all my friends for their help and encouragement.

Contents

ABSTRACT

Classification and regression trees are prediction models constructed by recursively partitioning a data set and fitting a simple model to each partition. Their name derives from the usual practice of describing the partitioning process by a decision tree. A classification or regression tree can be used to depict a decision tree, which is a prediction model. One of the oldest and most essential methods is the classification and regression tree technique. It's a technique for predicting outcomes based on a set of predictor factors. In data mining, decision trees often create a model used to that predicts the target of the values based on the many input variables. Some of the techniques used in predictive modelling are data mining, machine learning, statistics, and is decision tree learning, also known as induction of decision trees. It is a term used for classification and regression trees describe decision tree algorithms used for that are CART learning tasks. Models are created by recursively dividing the data space and fitting a basic prediction model to each division. Equally a consequence, the dividing may be seen as a decision tree visually. Regression trees are used to model continuous or ordered discrete dependent variables using error prediction, which is measured as the squared difference between actual and predicted values. For dependent variables with an unordered value of the finite number, classification trees are used, the cost of misclassification is used to quantify prediction inaccuracy. Classification and regression trees are machine learning methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Regression trees are for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values.

1.INTRODUCTION

As computing power and statistical insight has grown, increasingly complex and detailed regression techniques have emerged to analyse data. While this expanding set of techniques has proved beneficial in properly modelling certain data, it has also increased the burden on statistical practitioners in choosing appropriate techniques. Arguably an even heavier burden has been placed on non-statistician health practitioners in university, government, and private sectors where statistical software allows for immediate implementation of complex regression techniques without interpretation or guidance.

In response to this growing complexity, a simple tree system, Classification and Regression Tree (CART) analysis, has become increasingly popular, and is particularly valuable in multidisciplinary fields. CART can statistically demonstrate which factors are particularly important in a model or relationship in terms of explanatory power and variance. This process is mathematically identical to certain familiar regression techniques, but presents the data in a way that is easily interpreted by those not well versed in statistical analysis. In this way, CART presents a sophisticated snapshot of the relationship of variables in the data and can be used as a first step in constructing an informative model or a final visualization of important associations. In a large public health project, statisticians can use CART to present preliminary data to clinicians or other project stakeholders who can comment on the statistical results with practice knowledge and intuition. This process of reconciling the clinical and statistical relevance of variables in the data ultimately yields a more well informed and statistically informative model than either a singularly clinical or statistical approach. For example, complex regression models are routinely presented in economics literature with little introduction or explanation as the audience is familiar with these techniques, and are more interested in the specific application of a particular technique or unexpected result. In public health, however, this method of presentation is not motivating for practitioners without statistical expertise and who need to know the mechanism of the health effect to determine clinical relevance or craft an effective intervention. On the other hand, if the data were explained purely by narrative description and anecdote or excluding variables with statistically significant explanatory power without reason it would be interpreted as lacking scientific rigor.

The benefit of CART is to visually bridge interpretation and statistical rigor and facilitate relevant and valid model design.

Within the last 10 years, there has been increasing interest in the use of classification and regression tree (CART) analysis. CART analysis is a tree-building technique which is unlike traditional data analysis methods. It is ideally suited to the generation of clinical decision rules. Because CART analysis is unlike other analysis methods it has been accepted relatively slowly. Furthermore, the vast majority of statisticians have little or no experience with the technique. Other factors which limit CART's general acceptability are the complexity of the analysis and, until recently, the software required to perform CART analysis was difficult to use. Luckily, it is now possible to perform a CART analysis without a deep understanding of each of the multiple steps being completed by the software. In a number of studies, I have found CART to be quite effective for creating clinical decision rules which perform as well or better than rules developed using more traditional methods. In addition, CART is often able to uncover complex interactions between predictors which may be difficult or impossible to uncover using traditional multivariate techniques.

The purpose of this study is to provide an overview of CART methodology, emphasizing practical use rather than the underlying statistical theory.

ABOUT THE INSTITUTE

Kerala Institute of Local Administration (KILA) is an autonomous institution functioning for the Local governments in Kerala. It was registered under the Travancore-Cochin Literary, Scientific and Charitable Societies Act 1955. The Central university of Kerala has recognized it as a Research Centre attached to the Department of International Relations w.e.f 14 July 2014. Ever since its inception in 1990, KILA has been engaged in myriad of capacity building interventions on local governance and decentralization; including training, action, research, publications, seminars and workshops, consultancy, documentation, handholding and information services.

KILA has the mandate of facilitating and accelerating the socio-economic development of the State through strengthening the Local Self Government Institutions (LSGIs). As a Government of Kerala supported nodal agency for training, research and consultancy, KILA is committed to the following objectives:

- Undertake various training programmes for the Elected Representatives and Officials of Rural and Urban Local Governments of Kerala,
- Facilitate and strengthen decentralized planning process,
- Undertake action-oriented research activities,
- Document best practices on local governance for dissemination,
- Organize seminars, workshops and discussions, and
- Formulate policy documents

Area of Expertise

KILA has established its expertise in the following core areas:

- Participatory Planning
- Local Governance and Development
- Urban Governance and Development
- Local Economic Development and Livelihood Promotion

- Participatory Poverty Management
- Child Rights Governance
- Gender and Development
- Inclusive Governance and Development
- Natural Resource Management and Watershed Development
- Financial Management
- Food Security and Sustainable Agricultural Development
- Good Governance and Social Accountability
- Human Development
- Training Skills Development
- Total Quality Management

2. REVIEW OF LITERATURE

Breiman, Friedman, Olshen, and Stone in 1984 under the informative title Classification and Regression Trees, and they open the text with an example from clinical practice seeking to identify high risk patients within 24 hours of hospital admission for a myocardial infarction. This example proved to be particularly relevant – hundreds of studies have since emerged using CART analysis in clinical settings investigating myocardial infarctions.1 Often, these studies have many confounding variables which independently may not strongly predict a given outcome, such as heart attack, but together are important. CART analysis can guide medical researchers to isolate which of these variables is most important as a potential site of intervention. Public health research, especially concerning behavioural factors, often have some intuition regarding the most important predictors, which may explain why the method has been often absent from public health literature. One review of CART in public health has a much more pessimistic rational for the lack of use: "a general lack of awareness of the utility of CART procedures and, among those who are aware of these procedures, an uncertainty concerning their statistical properties."

O. YU, J. C. NELSON, L. BOUNDS and L. A. JACKSON in 2011 studies of community-acquired pneumonia (CAP) that utilize administrative data, cases are typically defined by the presence of a penumonia hospital discharge diagnosis code. However, not all such hospitalizations represent true CAP cases. We identified 3991 hospitalizations during 1997–2005 in a managed care organization, and validated them as CAP or not by reviewing medical records. To improve the accuracy of CAP identification, classification algorithms that incorporated additional administrative information associated with the hospitalization were developed using the classification and regression tree analysis. We found that a pneumonia code designated as the primary discharge diagnosis and duration of hospital stay improved the classification of CAP hospitalizations. Compared to the commonly used method that is based on the presence of a primary discharge diagnosis code of pneumonia alone, these algorithms had higher sensitivity (81–98%) and positive predictive values (82–84%) with only modest decreases in specificity (48–82%) and negative predictive values (75–90%).

David G. T. Denison, Bani K. Mallick and Adrian F. M. Smith in 1998 a stochastic search form of classification and regression tree (CART) analysis (Breiman et al., 1984) is proposed, motivated by a Bayesian model. An approximation to a probability distribution over the space of possible trees is explored using reversible jump Markov chain Monte Carlo methods (Green, 1995).

Wei-Yin Loh, John Eltinge, Moon Jung Cho and Yuanzhi Li in 2019 shows how classification and regression trees and forests can overcome these difficulties and compares them with likelihood methods in terms of bias and mean squared error. The development centers on a component of income data from the U.S. Consumer Expenditure Survey, which has a relatively high rate of item missingness. Classification trees and forests are used to model the unit-level propensity for item missingness in the income component. Regression trees and forests are used to model the conditional mean of the income component. The methods are then used to estimate the mean of the income component, adjusted for item nonresponse. Thirteen methods for estimating a population mean are compared in simulation experiments. The results show that if the number of auxiliary variables with missing values is not small, or if they have substantial missingness rates, likelihood methods can be impracticable or inapplicable. Tree and forest methods are always applicable, are relatively fast, and have higher efficiency than likelihood methods under real-data situations with incomplete-data patterns similar to that in the abovementioned survey. Their efficiency loss under parametric conditions most favorable to likelihood methods is observed to be between 10–25%.

Lester L. Yuan and Amina I. Pollard in 2014 developed a set of classification trees from bootstrapped replicates of the calibration data to explore a broader range of possible trees. They chose a final tree based on its predictive performance with a validation data set. The total N:TP mass ratio was the classification variable selected most frequently from a broad array of biological, chemical, and physical candidate classification variables. Relationships between TP and chl a in the resulting lake classes provided predictions that were substantially more accurate than predictions computed using nutrient ecoregions based on aggregations of Omer Nik Level III ecoregions, but predictions from a random forest model that averaged an ensemble of trees were even more accurate. Thus, the classification approach presented here sacrifices a small amount of predictive accuracy to retain a tree structure that is readily interpretable.

Russell Tronstad in 1995 used to model weekly Los Angeles wholesale prices (1990-93) for twelve different melon types. CART explained more of the variation in melon prices than did an ordinary least squares (OLS) regression with dummy variables. Explanatory variables ranked as the most-to-least important by CART are as follows: week, type of melon, year, size, grade, and shipping container. The most notable price change occurs when prices fall after 13 May.

3. INTERNSHIP PROBLEM

The CART algorithm is an important decision tree algorithm that lies at the foundation of machine learning. Moreover, it is also the basis for other powerful machine learning algorithms like bagged decision trees, random forest, and boosted decision trees. The Classification and regression tree (CART) methodology are one of the oldest and most fundamental algorithms. It is used to predict outcomes based on certain predictor variables.They are excellent for data mining tasks because they require very little data pre-processing. Decision tree models are easy to understand and implement which gives them a strong advantage when compared to other analytical models.The purpose of this study is to provide an overview of CART methodology, emphasizing practical use rather than the underlying statistical theory.

- Focused on CART and Its applications
- CART Application using an example by taking the data from National Family Health Survey (NFHS– 4)
- Data Filtration using STATA 15
- CART analysis using R

4. DATA AND METHODOLOGY

The data used for this study is a primary data set collected from the National Health Survey of India 2011 (NFHS– 4). The 2015-16 National Family Health Survey (NFHS-4), the fourth in the NFHS series, provides information on population, health, and nutrition for India and each state and union territory. For the first time, NFHS-4 provides district-level estimates for many important indicators. All four NFHS surveys have been conducted under the stewardship of the Ministry of Health and Family Welfare (MoHFW), Government of India. MoHFW designated the International Institute for Population Sciences (IIPS), Mumbai, as the nodal agency for the surveys. Funding for NFHS-4 was provided by the United States Agency for International Development (USAID), the United Kingdom Department for International Development (DFID), the Bill and Melinda Gates Foundation (BMGF), UNICEF, UNFPA, the MacArthur Foundation, and the Government of India. Technical assistance for NFHS-4 was provided by ICF, Maryland, USA. Assistance for the HIV component of the survey was provided by the National AIDS Control Organization (NACO) and the National AIDS Research Institute (NARI), Pune.

Four survey questionnaires—household, woman's, man's, and biomarker—were used to collect information in 19 languages using Computer Assisted Personal Interviewing (CAPI). All women age 15-49 and men age 15-54 in the selected sample households were eligible for interviewing. In the household questionnaire, basic information was collected on all usual members of the household and visitors who stayed in the household the previous night, as well as socioeconomic characteristics of the household, water and sanitation, health insurance, and number of deaths in the household in the three years preceding the survey. Two versions of the woman's questionnaire were used in NFHS-4. The first version (district module), which collected information on women's characteristics, marriage, fertility, contraception, reproductive health, children's immunizations, and treatment of childhood illnesses, was fielded in the entire sample of NFHS-4 households. Information on these topics is available at the district, state, and national levels. In the second version of the questionnaire (state module), four additional topics, namely, sexual behaviour, HIV/AIDS, husband's background and women's work, and domestic violence, were also included. This version was fielded in a subsample of NFHS-4 households designed to provide information only at the

state and national levels. The man's questionnaire covered the man's characteristics, marriage, number of children, contraception, fertility preferences, nutrition, sexual behaviour, attitudes towards gender roles, HIV/AIDS, and lifestyle. The biomarker questionnaire covered measurements of height, weight, and haemoglobin levels for children; height, weight, haemoglobin, blood pressure, and random blood glucose for women age 15-49 years and men age 15-54 years, and the collection of finger-stick blood for HIV testing in a laboratory. Questionnaire information and biomarkers were collected only with informed consent from the respondents.

The NFHS-4 sample was designed to provide estimates of all key indicators at the national and state levels, as well as estimates for most key indicators at the district level (for all 640 districts in India, as of the 2011 Census). The total sample size of approximately 572,000 households for India was based on the size needed to produce reliable indicator estimates for each district and for urban and rural areas in districts in which the urban population accounted for 30-70 percent of the total district population. The rural sample was selected 2 through a two-stage sample design with villages as the Primary Sampling Units (PSUs) at the first stage (selected with probability proportional to size), followed by a random selection of 22 households in each PSU at the second stage. In urban areas, there was also a two-stage sample design with Census Enumeration Blocks (CEB) selected at the first stage and a random selection of 22 households in each CEB at the second stage. At the second stage in both urban and rural areas, households were selected after conducting a complete mapping and household listing operation in the selected first-stage units. The figures of NFHS-4 and earlier rounds may not be strictly comparable due to differences in sample size, and NFHS-4 will be a benchmark for future surveys.

The national survey data was filtered to obtain Body Mass Index of women in Kerala. The dependent variable is body mass index and independent variables are type of place of residence, source of drinking water, wealth index for urban/rural, does this household have BPL card, marital status, first systolic reading, BP ever been checked previously, glucose level, weight in kg, height in cm, BMI, woman's highest education. The filtration was conducted using STATA 15 software which enables users to analyse, manage, and produce

graphical visualizations of data. It is widely used in the field of economics, biomedicine, and political science to examine data patterns.

The variable name and details of the data are given below

1.hhid- case identification

2.hv025- type of place of residence

3.hv201- source of drinking water

4.hv270- wealth index for urban/rural

5.sh75- does this household have BPL card

6.hv115_0- current marital status

7.shb18s_02- first systolic reading

8.shb19_02- BP ever been checked previously

9.shb74_02- glucose level

10.ha2_02- women's weight in kg (1 decimal)

11.ha3_02- woman's height in cm (1 decimal)

12.ha40_02- BMI

13.ha66_02- woman's highest education

A sample of the data is given below:

hv025	hv201	hv270a	sh75	hv115_02	shb19_02	shb74_02	ha2_02	ha3_02	ha40_02	ha66_02
urban	Pipes	poorer	yes	divorced	yes	94	431	1490	1941	secondary
rural	Wells	middle	yes	divorced	yes	111	586	1614	2250	secondary
rural	Wells	middle	yes	divorced	yes	101	370	1568	1503	secondary
rural	Pipes	richer	no	divorced	yes	100	370	1496	1653	higher
urban	Public	poorer	yes	divorced	yes	96	472	1488	2132	secondary
urban	Wells	poorer	yes	divorced	yes	113	520	1570	2110	secondary
rural	Wells	middle	yes	divorced	yes	126	526	1678	1866	secondary
rural	Wells	richest	yes	divorced	yes	129	779	1628	2937	higher
rural	Wells	richer	yes	divorced	yes	143	554	1456	2613	secondary
rural	Wells	richest	no	divorced	yes	175	525	1550	2185	higher
rural	Wells	richest	no	divorced	yes	114	617	1580	2470	secondary
urban	Wells	middle	yes	divorced	yes	101	700	1538	2959	secondary
urban	Public	richest	no	divorced	yes	104	745	1720	2518	higher
rural	Wells	richer	yes	divorced	yes	132	623	1510	2732	secondary
rural	Wells	richer	yes	divorced	yes	107	480	1510	2105	higher
rural	Wells	middle	yes	divorced	yes	108	693	1518	3005	secondary
urban	Public	richer	yes	divorced	yes	150	736	1542	3095	higher
urban	Pipes	richest	no	married	yes	109	558	1516	2426	secondary
urban	Public	middle	no	married	yes	105	492	1562	2014	secondary
urban	Pipes	richest	no	married	yes	90	517	1543	2171	higher
urban	Wells	poorer	no	married	yes	219	602	1606	2332	higher
urban	Public	richer	no	married	yes	123	527	1587	2092	higher

urban	Public	richer	no	married	yes	225	575	1593	2264	higher
urban	Public	middle	yes	married	yes	271	546	1568	2221	secondary
urban	Wells	richer	no	married	yes	131	362	1542	1520	secondary
urban	Public	middle	yes	married	yes	140	584	1598	2287	secondary
rural	Public	richest	yes	married	yes	112	642	1551	2669	secondary
rural	Public	poorest	yes	married	yes	182	606	1416	3020	secondary
rural	Public	richest	yes	married	yes	134	531	1652	1946	secondary
rural	Wells	richest	no	married	yes	130	600	1631	2256	secondary
rural	Pipes	richest	no	married	yes	186	513	1532	2184	secondary
rural	Pipes	richer	yes	married	yes	110	562	1532	2395	secondary
rural	Public	richer	no	married	yes	128	532	1530	2273	secondary
rural	Wells	richest	no	married	yes	128	487	1584	1941	secondary
rural	Public	richest	no	married	yes	130	581	1621	2211	secondary
rural	Public	richest	yes	married	yes	99	518	1551	2151	higher
rural	Wells	middle	yes	married	yes	104	501	1511	2192	higher
rural	Wells	richest	no	married	yes	138	491	1553	2034	secondary
rural	Wells	richest	no	married	yes	121	474	1602	1847	higher
rural	Public	richest	no	married	yes	153	541	1496	2415	higher
rural	Wells	middle	yes	married	yes	129	430	1590	1701	higher
rural	Wells	richest	no	married	yes	91	550	1615	2109	higher
rural	Natural Resources	richer	yes	married	yes	137	440	1520	1904	secondary
rural	Wells	richest	no	married	yes	143	560	1590	2215	secondary
rural	Natural Resources	richer	yes	married	yes	106	770	1560	3164	secondary
rural	Wells	richest	no	married	yes	149	490	1460	2299	secondary
rural	Wells	richest	no	married	yes	128	900	1690	3151	higher

rural	Public	richer	yes	married	yes	159	560	1540	2361	secondary
rural	Wells	richest	yes	married	yes	109	588	1526	2523	secondary
rural	Public	richest	no	married	yes	109	443	1516	1925	higher
rural	Wells	richest	no	married	yes	131	675	1603	2625	secondary
rural	Wells	richest	yes	married	yes	159	498	1516	2165	higher
rural	Wells	richest	no	married	yes	130	585	1573	2364	higher
rural	Wells	richest	yes	married	yes	240	697	1565	2846	secondary
rural	Public	richer	yes	married	yes	132	642	1584	2557	secondary
rural	Natural Resources	richest	no	married	yes	141	424	1546	1774	secondary
rural	Public	richer	yes	married	yes	115	663	1572	2683	secondary
rural	Wells	richest	yes	married	yes	105	669	1605	2597	secondary
rural	Pipes	richest	yes	married	yes	137	485	1534	2061	secondary
rural	Public	richer	yes	married	yes	124	522	1416	2603	secondary
rural	Wells	richest	yes	married	yes	110	553	1600	2158	secondary
rural	Pipes	richest	yes	married	yes	123	553	1597	2166	secondary
rural	Wells	richest	no	married	yes	123	517	1539	2181	higher
rural	Wells	richest	no	married	yes	106	549	1495	2454	higher
rural	Public	richest	no	married	yes	86	498	1590	1968	higher
rural	Public	richest	yes	married	yes	218	587	1657	2138	higher
rural	Wells	richest	no	married	yes	113	586	1520	2534	higher
rural	Wells	richest	no	married	yes	173	551	1526	2366	higher
rural	Public	richest	yes	married	yes	98	619	1556	2557	secondary
rural	Wells	richest	no	married	yes	120	444	1540	1872	secondary
rural	Wells	richer	yes	married	yes	120	603	1601	2353	secondary
rural	Wells	richest	no	married	yes	142	454	1542	1909	higher

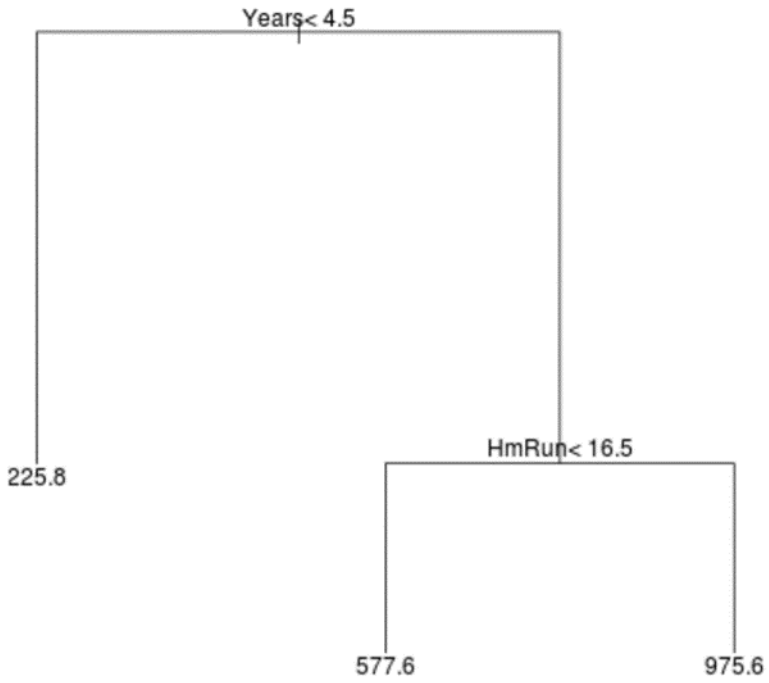
rural	Wells	richer	yes	married	yes	129	387	1579	1552	higher
rural	Wells	richest	no	married	yes	203	431	1520	1863	higher
rural	Pipes	richest	yes	married	yes	144	443	1566	1804	higher
rural	Pipes	richest	yes	married	yes	113	615	1559	2528	secondary
rural	Wells	richest	no	married	yes	105	519	1510	2276	higher
rural	Wells	richer	no	married	yes	128	598	1641	2221	secondary
rural	Public	poorer	yes	married	yes	122	488	1565	1992	secondary
rural	Public	richer	yes	married	yes	151	631	1563	2583	secondary
rural	Public	middle	yes	married	yes	129	432	1538	1826	secondary
rural	Pipes	richer	no	married	yes	170	520	1521	2248	secondary
rural	Public	richest	yes	married	yes	120	460	1491	2069	secondary
rural	Pipes	poorer	yes	married	yes	131	520	1552	2159	higher
rural	Pipes	richer	yes	married	yes	126	500	1591	1975	secondary
rural	Wells	richest	no	married	yes	115	442	1504	1954	higher
rural	Public	middle	yes	married	yes	134	492	1536	2083	higher
rural	Wells	richer	yes	married	yes	106	485	1455	2291	higher
rural	Public	richest	no	married	yes	114	574	1553	2380	secondary
rural	Wells	richest	no	married	yes	129	584	1645	2158	secondary
rural	Wells	richest	yes	married	yes	252	536	1453	2536	secondary
rural	Wells	richest	no	married	yes	94	638	1575	2570	higher
rural	Pipes	richer	yes	married	yes	110	460	1507	2025	secondary

Classification and Regression Tree (CART)

When the relationship between a set of predictor variables and a response variable is linear, methods like multiple linear regression can produce accurate predictive models. However, when the relationship between a set of predictors and a response is highly non-linear and complex then non-linear methods can perform better. One such example of a non-linear method is classification and regression trees, often abbreviated CART. As the name implies, CART models use a set of predictor variables to build decision trees that predict the value of a response variable.

For example, suppose we have a dataset that contains the predictor variables Years played and average home runs along with the response variable Yearly Salary for hundreds of professional baseball players.

Here's what a regression tree might look like for this dataset:



The way to interpret the tree is as follows:

- Players with less than 4.5 years played have a predicted salary of $225.8k.
- Players with greater than or equal to 4.5 years played and less than 16.5 average home runs have a predicted salary of $577.6k.
- Players with greater than or equal to 4.5 years played and greater than or equal to 16.5 average home runs have a predicted salary of $975.6k.

The results of this model should intuitively make sense: Players with more years of experience and more average home runs tend to earn higher salaries.

We can then use this model to predict the salary of a new player.

For example, suppose a given player has played 8 years and averages 10 home runs per year. According to our model, we would predict that this player has an annual salary of $577.6k.

Steps to Build CART Models

We can use the following steps to build a CART model for a given dataset:

Step 1: Use recursive binary splitting to grow a large tree on the training data.

First, we use a greedy algorithm known as recursive binary splitting to grow a regression tree using the following method:

- Consider all predictor variables X_1, X_2, … , X_p and all possible values of the cut points for each of the predictors, then choose the predictor and the cut point such that the resulting tree has the lowest RSS (residual standard error).
- For classification trees, we choose the predictor and cut point such that the resulting tree has the lowest misclassification rate.
- Repeat this process, stopping only when each terminal node has less than some minimum number of observations.

This algorithm is greedy because at each step of the tree-building process it determines the best split to make based only on that step, rather than looking ahead and picking a split that will lead to a better overall tree in some future step.

Step 2: Apply cost complexity pruning to the large tree to obtain a sequence of best trees, as a function of α.

Once we've grown the large tree, we then need to prune the tree using a method known as cost complexity pruning, which works as follows:

- For each possible tree with T terminal nodes, find the tree that minimizes RSS + α|T|.
- Note that as we increase the value of α, trees with more terminal nodes are penalized. This ensures that the tree doesn't become too complex.

This process results in a sequence of best trees for each value of α.

Step 3: Use k-fold cross-validation to choose α.

Once we've found the best tree for each value of α, we can apply k-fold cross-validation to choose the value of α that minimizes the test error.

Step 4: Choose the final model.

Lastly, we choose the final model to be the one that corresponds to the chosen value of α.

Pros & Cons of CART Models

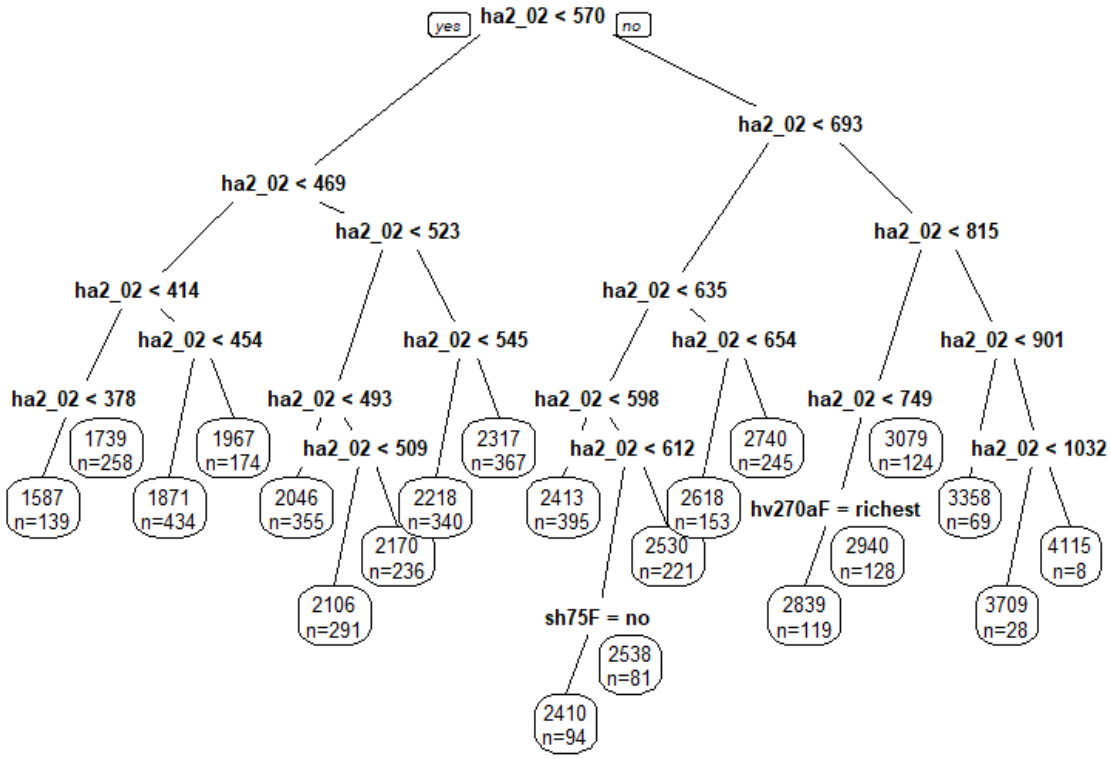
CART models offer the following pros:

- They are easy to interpret.
- They are easy to explain.
- They are easy to visualize.
- They can be applied to both regression and classification problems.

However, CART models come with the following con:

They tend to not have as much predictive accuracy as other non-linear machine learning algorithms. However, by aggregating many decision trees with methods like bagging, boosting, and random forests, their predictive accuracy can be improved.

5.ANALYSIS AND RESULT



Here, the predictor variables are ha2_02, sh75F, hv270aF, along with the response variable. ha40_02(BMI of women above 18 years). From the tree we can intercept that, Women with weight less than 57kg are poor or middle and have bpl card and a BMI ranging from 1587 to 2317.Then, women with weight more than or equal to 57kg are richer or richest, may or may not have bpl card have BMI ranging from 2410 to 4115.

6.Conclusion

Classification and Regression Tree (CART) analysis is a powerful technique with significant potential and clinical utility. Nonetheless, a substantial investment in time and effort is required to use the software, select the correct options, and interpret the results. Nonetheless, the use of CART has been increasing and is likely to increase in the future, largely because of the substantial number of important problems for which it is the best available solution.

Because of its numerous advantages, the regression tree-based technique is widely utilized in the field of analytics. It allows the user to see each stage, allowing him to make a more informed decision. Prioritize the factors of a choice depending on what you believe is of the most importance. In comparison to many other approaches, making judgments based on regression is simple. As you go deeper down the tree, the majority of the unwanted data is filtered away, leaving you with fewer data. The regression tree is simple to construct, and it may be presented to higher authorities in the form of a chart or a simple diagram. CART (Classification and Regression Tree Analysis) is a basic yet effective analytic method for determining the most "important" (in terms of explanatory power) variables. For nursing and other healthcare studies, classification and regression tree analysis provide intriguing potential. The technique is a simple to understand, computationally driven, and practical way for modelling interactions between health-related factors that might otherwise go unnoticed.

The significance of this cannot be emphasized, as undiscovered variables regularly influence patient outcomes in healthcare research. The beauty of this technique is the ability to discover and assess the significance of these elements. In comparison to other machine learning techniques, CART is a strong algorithm that is also very simple to describe. It does not need a lot of processing resources, allowing you to create models quickly. While it is important to avoid overfitting your data, it is a decent method for simple situations. Classification and regression tree analysis is a simple way to define interactions between health-related factors that might otherwise be hidden.

REFERENCES

1. H. Ahn. Tree-structured exponential regression modelling. Biometrical Journal, 2007.
2. H. Ahn and W.-Y. Loh. Tree-structured proportional hazards regression modelling. Biometrics, 1994.
3. L. Breiman. Bagging predictors. Machine Learning, 1996.
4. L. Breiman. Random forests. Machine Learning, 2001.
5. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. CRC Press, 1984.
6. Classification and Regression Trees. by L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone
7. Fifty Years of Classification and Regression Trees by Wei-Yin Loh
8. Classification and regression trees and forests for incomplete data from sample surveys by wei-yin loh, john eltinge, moon jung cho, yuanzhi li
9. https://en.wikipedia.org/wiki/Neural_network
10. https://dhsprogram.com/
11. https://github.com/
12. https://www.jstor.org/

ANNEXURE

Code of R Analysis

```
library(rpart)
library(rpart.plot)

#read data
library(readxl)
internship_2 <- read_excel("E:/k/internship_2.xlsx")
str(internship_2)
data<-na.omit(internship_2)

data$hv025F<-factor(data$hv025)
data$hv201F<-factor(data$hv201)
data$hv270aF<-factor(data$hv270a)
data$sh75F<-factor(data$sh75)
data$hv115_02F<-factor(data$hv115_02)
data$shb19_02F<-factor(data$shb19_02)
data$ha66_02F<-factor(data$ha66_02 )

#height,weight,bpl card,type of residence
tree<-rpart(ha40_02 ~ ha2_02+ha3_02+hv025F+sh75F+hv270aF,data = data, control =
rpart.control(cp = 0.007))
printcp(tree)

bestcp <- tree$cptable[which.min(tree$cptable[,"xerror"]),"CP"]
tree.pruned <- prune(tree, cp = bestcp)

plot(tree.pruned)
text(tree.pruned, cex = 0.8, use.n = TRUE, xpd = TRUE)

prp(tree.pruned, faclen = 0, cex = 0.8, extra = 1)
```