بسم الله الرحمن الرحيم

Sudan University of Science and Technology

Faculty of Graduate Studies

College of science

Department ofApplied Statistics

Classification and Regression Tree-A Case study of AIDS Patients at Antiretroviral TherapycenterKhartoum

(Feb/2014)

نموذج التصنيف والانحدار الشجري- دراسة حالة: مرضي الايدز بمركز الارشاد النفسي والفحص الطوعي بمستشفي الخرطوم

في الفترة (فبراير2014)

A research submitted for partial fulfillment for the degree of M.Sc. in Applied Statistics

Prepared by:

MAYSON ABDALRHEM MOHAMED SOROR

Supervised by:

Dr. AMAL ALSIR ALKHIDIR

March 2014

Dedication

To which ensures suffered hardship for our comfort.

To every mother.

To which nothing was spared for the sake of happiness of his son.

To every father.

To each taught you letter.

To each teacher gave his life to education.

To all my friends.

To Dr .AMAL ALSIR.

Acknowledgement

I thank and appreciate my dear supervisor Dr. AMAL very much for her usual support and helpful gaudiness that she provides to me through these months.

Also I thank my family for their love can they pass to me which make my life full of success and happiness.

Last I thank all my friends for their helping to me all the time.

Mayson .

المستخلص

لغرض التنبؤ بالاصابة بمرض السل الرئوي لمرضي الايدز وجد ان هنالك عدد من المتغيرات المختلفة تؤثر علي الاصابة بالمرض مثل النوع،العمر وتعاطي الكحول وغيرها، وهي متغيرات مختلفة من حيث التصنيف الاحصائي ولايمكن جمعها في نموذج انحدار متعدد واحد.

هدفت الدر اسة لإيجاد نموذج يمكن من خلالة تصنيف المرضي من حيث الاصابة بالمرض من خلال فرض ملائمة نموذج CART لتصنيف المرضي.

وتمكنت الدراسة من إيجاد نموذج للتنبؤ بالاصابة وبعمر المصاب من خلال شجرة التصنيف والانحدار الشجري.

توصي الدراسة بزيادة حجم العينة واحتواء النموذج علي عدد إضافي من المتغيرات ذات الصلة بمرض السل الرئوي، كما توصي بدراسة امراض اخري تصيب مرضي الايدز كالفشل الكلوي وفشل الكبد.

ABSTRACT

The prediction for TB infection for the patient of AIDS disease, we found that many variables influence the infection by it like sex, age and alcohol drinking etc, all these variables different in statistical classification and can't be merge in one multiple regression model.

For that purpose this study aims to found one model to classify the disease by using CART model for prediction.

This study gained a model to predict the infection and the age for the AIDS patients from the classification and regression tree CART.

The study recommended increases the sample size and adding other related variables to the model. Also it recommended to study another diseases which related to AIDS like renal and liver failure.

TABLE OF CONTENT:

Subject **Page**

CHAPTER THREE

CHAPTER FOUR

CHAPTER Five

1.1 Introduction:

The Classification and regression are two important problems in statistics,each deals with the prediction of a response variable y given the values of a vector of predictor variables x's.

The CART is a nonparametric statistical methodology developed for analyzing classification issues either from categorical or continuous independent variables, if the dependent variable is categorical, CART produces a classification tree. When the dependent variable is continuo us, it produces a regression tree.

CART methodology was developed in 1980 by Breiman, Freidman, Olshen and Stone in their paper "Classification and Regression Trees" (1984). For building decision trees, CART uses so-called learning sample - a set of historical data with pre-assigned classes for all observations.

1.2 Research Problem:

To predict pulmonary tuberculosis to AIDS patients, we find that there are many variables affecting it, such as type, age, marital status and others. These variables, including qualitative , quantitative and ordinal Since we can not collect these different variables in Models of multiple regression or logistic. Thenhow to predict variable? Can be done through the use of model CART

1.3 Important of Research:

To Finda new model for prediction instead of using ordinary regression analysis, since CART procedure does not need to satisfy the assumption of the regression analysis, as it represent anew technique to solve the problem of diversity of the independent variables(i.e. continuous variables or categorical variables…..etc.)In the same model.

1.4 Aim:

The general aim of classification and regression tree analysis find a way of using independent variable to partition the observations into homogeneously distributed groups, then use the group to predict the dependent variable. And make use of the accuracy of the CART in the prediction.

The overall aim of AIDS patients to find out positive TB.

1.5 Research Data:

This research has been applied to cross-sectional data were obtained from the centerof AntiretroviralTherapy (HRT) from Khartoum hospital from a sample size of(100)in order to study the following variablesat 10-Feb-2014: Sex, Age, Martial status, alcohol drinking, TB.

1.6 Hypothesis of Research:

Possibility of processing data using CART procedure.

1.7 Statistical Methodology:

In this research used Classification and Regression Tree methodology by SPSS package.

1.8 Previous Study:

1.81. In 2001 the researchers (Marshall RJ) and his study in (A critique is presented of the use of tree-based partitioning algorithms to formulate classification rules and identify subgroups from):-

- clinical and epidemiological data. It is argued that the methods have a number of limitations.
- despite their popularity and apparent closeness to clinical reasoning processes.
- The issue of redundancy in tree-derived decision rules is discussed. Simple rules may be unlikely to be "discovered" by tree growing. Subgroups identified by trees are often hard to interpret or believe and net effects are not assessed.
- These problems arise fundamentally because trees are hierarchical. Newer refinements of tree technology seem unlikely to be useful, wedded as they are to hierarchical structures.

1.8.2 In 2003 to the researchers (Fabio S Aguiar) and his study in (The Classification And Regression Tree (CART) Model To Predict Pulmonary Tuberculosis in Hospitalized Patient) -:

- Decision making in whether to isolate patients with clinical suspicion of TB in tertiary health facilities in countries with limited resource.
- The model cutoff 30 years of age for discriminating TB, particularly in patients with a typical chest x-Ray without dyspnea.
- Also used model for validated of sample of different patients admitted to the same hospital in other period of time to be rebuts model for prediction of TB diagnosis.

1.8.3 In 2008 the researchers (Phelps Mandy)and his study in (Classification and Regression Trees as Alternatives to Regression):-

- linear relationships between the dependent variable(s) and in dependent variables
- describe the use of Classification and Regression Trees (CART) to sidestep the assumptions required by traditional analyses.
- CART has the added benefit of not requiring large sample sizes in order to obtain accurate results, although larger sample sizes are preferred.
- There is also a difference in the goal of CART compared to traditional analyses.
- CART is geared toward prediction, whereas traditional analyses are geared toward developing a specific model for your data. The poster will contain specific information about the procedures underlying CART.

1.9 ResearchStructure:

Chapter One: Introduction, Research Problem, Important of Research, Aim, Research Data, Hypothesis of Research, Statistical Methodology, Previous Study.

Chapter Tow : Introduction, History of AIDS, Definition, Symptoms, Primary infection, Clinical latent infection, Earlysymptomatic HIV infection , Progression to AIDS, Causes, Treatment response.

Chapter Three :Introduction, Principles OF CART Methodology, Classification Tree, BUILDING A CLASSIFICATION TREE, REGRESSION TREES, Advantages and Disadvantages of CART,

Chapter Four : Data Analysis.

Chapter Five : Result And Recommendation.

1.1 2-1: Introduction:

1.2 The first cases of Acquired Immune Deficiency Syndrome (AIDS) were reported in the United States in the spring of 1981. By 1983 the human immunodeficiency virus (HIV), the virus that causes AIDS, had been isolated. Early in the U.S. HIV/AIDS pandemic, the role of substance abuse in the spread of AIDS was clearly established. Injection drug use (IDU) was identified as a direct route of HIV infection and transmission among injection drug users. The largest group of early AIDS cases comprised gay and bisexual men (referred to as men who have sex with men(or MSMs). Early cases of HIV infection that were sexually transmitted often were related to the use of alcohol and other substances, and the majority of these cases occurred in urban, educated, white MSMs.

HIV is a lentivirus, a class of retroviruses, and integrates into the genome in the same manner as other retroviruses. Unlike other retroviruses, which typically bud from the infected cell for a long period of time, HIV can lyze the cell or lie dormant for many years, especially in resting T4 (CD4) lymphocytes; later, a recrudescence of viral production occurs that ultimately destroys the cell. It should be noted that while HIV may disappear from the cells of the circulation, viral replication and budding continues to occur in other tissues. In contrast to T4 lymphocytes, other HIV-infected cells do not show a pronounced cytopathic effect and may exhibit latency or continual budding. Unlike many other retroviruses, no copies of primate lentiviruses are found in the genome of target species and the virus not transmitted through the germ line.

1.3 2-2: History of AIDS:

1.4 The history of HIV and AIDS is a short one and yet since it was first reported, just over thirty years ago; it has become one of the leading causes of death worldwide. With time, as research, investment and commitment into understanding HIV increased, treatment methods evolved and the outcome of people living with HIV improved around the world.

1.5 Moving through our 'AIDS History Topics' you will be guided through the history of AIDS from the 1970s up to 1986, AIDS and HIV in the 1990s, to HIV and AIDS in the new millennium; enabling you to learn how the response to HIV and AIDS has changed over time.

1.6 In the early years of the epidemic, AIDS was unknown and misunderstood, feared, untreatable and often fatal. Years down the line, a virus named HIV was discovered and linked to AIDS. That was the turning point in AIDS history.

1.7 HIV history then took a sharp turn with the development of highly effective antiretroviral drugs which meant that, with access to treatment, people could lead healthy lives with HIV.

1.8 Key historical moments in the history of HIV and the history of AIDS can be explored through the HIV, such as the various advances in HIV treatment over the years; the impact and evolution of legislation; and HIV activism, among other epidemic defining events.

1.9 2-3: Definition:

AIDS is a chronic, potentially life-threatening condition caused by the human immunodeficiency virus (HIV). By damaging your immune system, HIV interferes with your body's ability to fight the organisms that cause disease.[1]

HIV is a sexually transmitted infection. It can also be spread by contact with infected blood, or from mother to child during pregnancy, childbirth or breast-feeding. It can take years before HIV weakens your immune system to the point that you have AIDS.

There's no cure for HIV/AIDS, but there are medications that can dramatically slow the progression of the disease. These drugs have reduced AIDS deaths in many developed nations.

[1] *www.mayoclinic.org/**diseases-conditions/hiv-aids***

1.10 2-4: Symptoms:

The symptoms of HIV and AIDS vary, depending on the phase of infection.

1.112-4-1: Primary infection

The majority of people infected by HIV develop a flu-like illness within a month or two after the virus enters the body. This illness, known as primary or acute HIV infection, may last for a few weeks. Possible symptoms include:

- Fever
- Muscle soreness
- Rash
- Headache
- Sore throat
- Mouth or genital ulcers
- Swollen lymph glands, mainly on the neck
- Joint pain
- Night sweats
- Diarrhea[(2)]

Although the symptoms of primary HIV infection may be mild enough to go unnoticed, the amount of virus in the blood stream (viral load) is particularly high at this time. As a result, HIV infection spreads more efficiently during primary infection than during the next stage of infection.

1.122-4-2: Clinical latent infection

In some people, persistent swelling of lymph nodes occurs during clinical latent HIV. Otherwise, there are no specific signs and symptoms. HIV remains in the body, however, as free virus and in infected white blood cells.
Clinical latent infection typically lasts eight to 10 years. A few people stay in this stage even longer, but others progress to more-severe disease much sooner.

2-4-3: Early symptomatic HIV infection

As the virus continues to multiply and destroy immune cells, you may develop mild infections or chronic symptoms such as:

- Fever

[(2)] www.shaancreations.com/wp-content/uploads/2013/.../HIV-AIDS-SCI.pdf

- Fatigue
- Swollen lymph nodes — often one of the first signs of HIV infection
- Diarrhea
- Weight loss
- Cough and shortness of breath.[3]

1.12.1.1 2-4-4: Progression to AIDS

If you receive no treatment for your HIV infection, the disease typically progresses to AIDS in about 10 years. By the time AIDS develops, your immune system has been severely damaged, making you susceptible to opportunistic infections — diseases that wouldn't trouble a person with a healthy immune system. The signs and symptoms of some of these infections may include:

- Soaking night sweats
- Shaking chills or fever higher than 100 F (38 C) for several weeks
- Cough and shortness of breath
- Chronic diarrhea
- Persistent white spots or unusual lesions on your tongue or in your mouth
- Headaches
- Persistent, unexplained fatigue
- Blurred and distorted vision
- Weight loss
- Skin rashes or bumps

1.12.1.2 2-5: Causes:

Scientists believe a virus similar to HIV first occurred in some populations of chimps and monkeys in Africa, where they're hunted for food. Contact with an infected monkey's blood during butchering or cooking may have allowed the virus to cross into humans and become HIV.

1.12.1.3 2-5-1: How does HIV become AIDS?

HIV destroys CD4 cells — a specific type of white blood cell that plays a large role in helping your body fight disease. Your immune system weakens as more CD4 cells are killed. You can have an HIV infection for years before it progresses to AIDS.

[3]Document From University Of Washington Web site

People infected with HIV progress to AIDS when their CD4 count falls below 200 or they experience an AIDS-defining complication, such as:

- Pneumocystis pneumonia
- Cytomegalovirus
- Tuberculosis
- Toxoplasmosis
- Cryptosporidiosis

1.12.1.4 2-5-2: How HIV is transmitted:

To become infected with HIV, infected blood, semen or vaginal secretions must enter your body. You can't become infected through ordinary contact-hugging, kissing, dancing or shaking hands-with someone who has HIV or AIDS. HIV can't be transmitted through the air, water or via insect bites.

You can become infected with HIV in several ways, including:

- During sex. You may become infected if you have vaginal, anal or oral sex with an infected partner whose blood, semen or vaginal secretions enter your body. The virus can enter your body through mouth sores or small tears that sometimes develop in the rectum or vagina during sexual activity.
- Blood transfusions. In some cases, the virus may be transmitted through blood transfusions. American hospitals and blood banks now screen the blood supply for HIV antibodies, so this risk is very small.
- Sharing needles. HIV can be transmitted through needles and syringes contaminated with infected blood. Sharing intravenous drug paraphernalia puts you at high risk of HIV and other infectious diseases such as hepatitis.
- From mother to child. Infected mothers can infect their babies during pregnancy or delivery, or through breast-feeding. But if women receive treatment for HIV infection during pregnancy, the risk to their babies is significantly reduced.

1.12.1.5 2-6-1: Treatments and drugs:

There's no cure for HIV/AIDS, but a variety of drugs can be used in combination to control the virus. Each of the classes of anti-HIV drugs blocks the virus in different ways. It's best to combine at least three drugs from two different classes to avoid creating strains of HIV that are immune to single drugs. The classes of anti-HIV drugs include:

- Non-nucleoside reverse transcriptase inhibitors (NNRTIs).NNRTIs disable a protein needed by HIV to make copies of itself. Examples include efavirenz (Sustiva), etravirine (Intelence) and nevirapine (Viramune).
- Nucleoside reverse transcriptase inhibitors (NRTIs). NRTIs are faulty versions of building blocks that HIV needs to make copies of itself. Examples include Abacavir (Ziagen), and the combination drugs emtricitabine and tenofovir (Truvada), and lamivudine and zidovudine (Combivir).
- Protease inhibitors (PIs). PIs disable protease, another protein that HIV needs to make copies of itself. Examples include atazanavir (Reyataz), darunavir (Prezista), fosamprenavir (Lexiva) and ritonavir (Norvir).
- Entry or fusion inhibitors. These drugs block HIV's entry into CD4 cells. Examples include enfuvirtide (Fuzeon) and maraviroc (Selzentry).
- Integrase inhibitors. Raltegravir (Isentress) works by disabling integrase, a protein that HIV uses to insert its genetic material into CD4 cells.

1.12.1.6 2-6-2: When to start treatment

Current guidelines indicate that treatment should begin if:

- You have severe symptoms
- Your CD4 count is under 500
- You're pregnant
- You have HIV-related kidney disease
- You're being treated for hepatitis B

1.12.1.7 Treatment can be difficult

HIV treatment regimens may involve taking multiple pills at specific times every day for the rest of your life. Side effects can include:

- Nausea, vomiting or diarrhea
- Abnormal heartbeats
- Shortness of breath
- Skin rash
- Weakened bones
- Bone death, particularly in the hip joints

Co-diseases and co-treatments Some health issues that are a natural part of aging may be more difficult to manage if you have HIV. Some medications that are common for age-related cardiovascular, metabolic and bone conditions, for example, may not interact well with anti-HIV medications. Talk to your doctor

about other conditions you're receiving medication for. There are also known interactions between anti-HIV drugs and:

- Contraceptives and hormones for women
- Medications for the treatment of tuberculosis
- Drugs to treat hepatitis C

1.12.1.8 2-7: Treatment response:

Your response to any treatment is measured by your viral load and CD4 counts. Viral load should be tested at the start of treatment and then every three to four months during therapy. CD4 counts should be checked every three to six months.

HIV treatment should reduce your viral load to the point that it's undetectable. That doesn't mean your HIV is gone. It just means that the test isn't sensitive enough to detect it. You can still transmit HIV to others when your viral load is undetectable.

3-1: Introduction:

Classification and Regression Tree a recursive partitioning method, builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). The classic C&RT algorithm was popularized by Breiman. (Breiman, Friedman, Olshen, & Stone, 1984; see also Ripley, 1996).

A general introduction to tree-classifiers, specifically to the QUEST (Quick, Unbiased, Efficient Statistical Trees) algorithm, is also presented in the context of the Classification Trees Analysis facilities, and much of the following discussion presents the same information, in only a slightly different context. Another, similar type of tree building algorithm is CHAID (Chi-square Automatic Interaction Detector).

3-2: Principles OF CART Methodology:

The most important feature of a classification tree is accuracy. All classification procedures produce errors. The CART procedure does not make any distributional assumptions on covariates. Hence, hypothesis testing is not possible. Confidence in CART's performance based on an assessment of the extent of misclassification it generates from data sets with known class distributions.

To test the predictive accuracy of a tree is to take an independent test data set with known class distributions and run it down the tree and determine the proportion of cases misclassified. Breiman (1984) provide three procedures for estimating the accuracy of tree-structured classifiershave two objectives:

1-Constructing a classification tree, $c(x)$.

2- Finding an estimate of $R * [c(x)]$

Let:

$c(x)$= a tree-structured classifier, where (x) is a vector of characteristics variables.

$R * [c(x)]$ = the classifier's "true" misclassification rate.

L = the learning sample (the sample data from which to construct a classification tree, or a set of historical data with pre-assigned classes for all observation.)

The three estimation procedures below:

3-2-1: Resubstitution Estimate:

This estimates the accuracy of the true misclassification rate, $R * [\, c(x)]$

Build a classification tree, $c(x)$, from the learning sample$L$,and Apply this tree, $c(x)$, to the data set. Let the observations in the sample run down the tree one time,and Compute the proportion of cases that are misclassified. This proportion is the Resubstitutionestimate, $R[c(x)]$, of the true misclassification rate, $R * [c(x)]$

The major weakness of this estimator of the error rate is that it is derived from the same data set from which the tree is built; hence, it underestimates the true misclassification rate. The error rate is always low in such cases.

3-2-2: Test-sample estimate:

Divide the observations in the learning sample, L, into two parts: L_1 and L_2. L_1andL_2 need not be equal. Use L_1 to build the classifier, $c(x)$, and Run observations inL_2 down the tree, $c(x)$,one time.

Compute the proportion of cases that are misclassified. This proportion is the test-sample estimate, $R[c(x)]$, of the "true" misclassification rate, $R*[c(x)]$. In large samples, this estimate provides an unbiased estimate of the true misclassification rate[1].

3.2.3.K-fold cross-validation:

This is the recommended procedure for small samples.

Divide the learning sample, L, into K subsets of an equal number of observations. ($Let\ L_1, L_2, \ldots, L_k$be the subsamples). Construct a classifier, $c(x)$, from the $k-1$ subsamples by leaving out, say, the k^{th} subsample, l_k. Apply the classifier,$c(x)$, to the excluded subsample, L_k, and estimate $R*[c(x)]$ as the proportion of misclassified observations. Denote this estimate as $R\,t_s[c\,k(x)]$, where k denotes k-fold cross-validation, and t_s denotes test sample. By Repeat the above using all subsamples except the subsample L_{k-1}. The subsample L_{k-1} now becomes a test sample. The process above is repeated until every subsample is used once in the construction of $c(x)$and once as a test sample. The result is a series of test sample resubstitution estimates,

$$R^{ts}[c^k(x)], R^{ts}\left[c^{k-1}(x)\right], .., R^{ts}[c^1(x)] \_\_\_\_\_ (1\text{-}3)$$

Add the series of $R^{ts}\,[c^k(x)], R^{ts}[c^{k-1}(x)], .., R^{ts}[c^1(x)]$

generated from the k-fold cross-validation and get an estimate of $R\,[c(x)]$; that is, the k-fold cross-validation estimate $R^{ck}(c)$ of R $[c(x)]$ is given as:

$$R^{ck} = 1, k\sum_{k=1} R^{ts}\,[c(x)] \_\_\_\_\_ (2\text{-}3)$$

which is an average of the error rates from k cross-validation tests.[4]

3.3.Classification Tree:

[4]Yohannes, Yisehac,1999,CLASSIFICATION AND REGRESSION TREES, CART

Three components are required in the construction of a classification tree:

i. Type and format of questions.
ii. Splitting rules and goodness of split criteria.
iii. Class Assignment Rule.[5]

3.3.1.Type and format of questions:

Two question formats are defined in CART:

1.Is$X \leq d$?, if X is a continuous variable and d is a constant within the range of X values.

2.Is $Z = b$?, if Z is a categorical variable and b is one of the integer values assumed by Z .

The number of possible split points on each variable is limited to the number of distinct values each variable assumes in the sample. For example, if a sample size equals n , and if X is a continuous variable and assumes N distinct points in the sample, then the maximum number of split points on X is equal to N . If Z is a categorical variable with m distinct points in a sample, then the number of possible split points on Z equals $(2m–1)–1$.

3.3.2.spliting rules and goodness of split criteria:

This component requires definition of the impurity function and impurity measure.

Let:

$j = 1,2,..., k$ be the number of classes of categorical dependent variables.

then define $p(j/t)$ as class probability distribution of the dependent variable at node t, such that:

$$p(1/t) + p(2/t) + p(3/t) + ... + p(k/t) = 1, \quad j = 1, 2, ..., k. \qquad (3\text{-}3)$$

Let $i(t)$ be the impurity measure at node t. Then define $i(t)$ as a function of class probabilities

[5] pervious Reference

p (1/ t), p(2/ t), p (3/ t), ... Mathematically,

i (t) = φ [p(1/t), p (2/t), . . . , p (j/ t)].

The definition of impurity measure is generic and allows for flexibility of functional forms.

3.3.2.A.Splitting Rules:

There are three major splitting rules in CART:

A.1. Gini splitting rule:

Gini splitting rule is most broadly used rule. It uses the following impurity function *i(t):*

$$i(t) = \sum_{k \neq 1} p(k|t)p(l|t) \_\_\_\_\_ \text{(4-3)}$$

where*k,l 1,...,K* index of the class; *p(k|t)* conditional probability of class *k* provided we are in node *t*.

Applying the Gini impurity function *(4)* to:

$$\arg max \left[i(t_p) - p_{li}(t_l) - p_{ri}(t_r) \right] \_\_\_\_\_ \text{(5-3)}$$

Will get the following change of impurity function:

$$\Delta i(t) = -\sum_{k=1}^{k} p^2[k(t_p) + p_l \sum_{k=1}^{k} p^2(k/t_l) + p_r \sum_{k=1}^{k} p^2(k/t_r) \_\_\_ \text{(6-3)}$$

Therefore, Gini algorithm will solve the following problem:

$$\arg max\, R\, [-\sum_{k=1}^{k} p^2[k(t_p) + p_l \sum_{k=1}^{k} p^2(k/t_l) + p_r \sum_{k=1}^{k} p^2(k/t_r)]$$

$$x_j \leq x_{j,j=1,2,\ldots,m}$$

Gini algorithm will search in learning sample for the largest class and isolate it from the rest of the data. Ginni works well for noisy data.

A.2. Twoing splitting rule

Twoing will search for two classes that will make up together more then 50% of the data. Twoing splitting rule will maximize the following change of impurity measure:

$$\arg max\, R(p_l p_r/4[\sum_{k=i}^{k} p(k/t_l) - p(k/t_r)]^2] \_\_\_\_\_ (7\text{-}3)$$

Although twoing splitting rule us to build more balanced trees, this algorithm works slower than gini rule.[6]

A.3. Linear Combination Splits:

This splitting rule is an alternative to CART's use of a single variable for splitting. It is designed for situations where the class structure of the data appears to depend on linear combinations of variables. In linear combination splits, the question posed for a node split takes the following form:

Is $I_s \alpha_1 x_1 + \alpha_2 x_2 \leq 40$?

For example, is .55 ×consumption+ .05 ×age ≤40?

If the response to the question is "yes," then the case is sent to the left node, and if the response is "no," then the case is sent to the right node.

This rule is valid only for cases with no missing values on predictor variables. Further more, if categorical variables have to be included in the model, they should be converted to sets of dummy variables.

3.3.2 BGoodness of Split Criteria:

Lets be a split at node t.Then, the goodness of split "*s*" is defined as the decrease in impurity measured by

$$\Delta i(s,t) = [i(t) - p_l(i(t_l)) - p_r(i(t_r))] \_\_\_\_\_ (8\text{-}3)$$

Where

s = a particular split.

p_L= the proportion of the cases at node *t* that go into the left child node, t_L.

p_R= the proportion of cases at node t that go into the right child node, t_R,

$i(t_L)$ = impurity of the left child node, and

$i(t_R)$ = impurity of the right child node.

3.3.3 Class Assignment Rule:

[6]Wolfgang H¨ardle,2004,CLASSIFICATION AND REGRESSION TREES.(CART)Theory and Applications

There are two rules for assigning classes to nodes. Each rule is based on one of two types of misclassification costs.

1. **The Plurality Rule**:
   Assign terminal node t to a class for which $p(j|t)$ is the highest. If the majority of the cases in a terminal node belong to a specific class, then that node is assigned to that class. The rule assumes equal misclassification costs for each class. It does not take into account the severity of the cost of making a mistake. This rule is a special case of rule 2**.**

2. **Assign terminal node *t* to a class for which the expected misclassification cost is at a minimum:**
   The application of this rule takes into account the severity of the costs of misclassifying cases or observations in a certain class, and incorporates cost variability into a Gini splitting rule

.

3.4 BUILDING A CLASSIFICATION TREE:

Building tree start by partitioning a sample or the root node into binary nodes based on question of the form

is$X \leq d$?

Where X is a variable in the data set and d is a real number. The goal is to devise a rule that will break up these observations and create groups or binary nodes that are internally more homogenous than the root node.

The tree building process is as follows:

CART split the first variable at all of its possible split points into binary or two child nodes. Cases with a "yes" response to the question posed are sent to the left node and those with "no" responses are sent to the right node. CART then applies its goodness of split criteria to each split point and evaluates the reduction in impurity that is achieved using the formula

$$\Delta i(s,t) = [i(t) - p_l(i(t_l)) - p_r(i(t_r))] \quad (9\text{-}3)$$

This was described earlier.

CART selects the best split of the variable as that split for which the reduction in impurity is highest.

Repeated above processing for each of the remaining variable sat the root node. CART then ranks all of the best splits on each variable according to the reduction in impurity achieved by each split. And it selects the variable its split point that most reduced the impurity of the root or parent node.

CART then assigns classes to these nodes according to the rule that minimizes misclassification costs. CART has a built in Algorithm that takes into account user-defined variable misclassification costs during the splitting process. The default is unit or equal misclassification costs.

Because the CART procedure is recursive, repeatedly all previous processing applied to each nonterminal child node at each successive stage. And continues the splitting process and builds a large tree.

3.5.1 REGRESSION TREES:

The main purpose of CART regression is to produce a tree structured predictor or prediction rule .This predictor serves two major goals: (1) to predict accurately the dependent Variable from the future or new values of the predictor variables;(2)to explain the relationships that exist between the dependent and predictor variables. The CART regression predictor is constructed by detecting the heterogeneity(in terms of variance of the dependent variable)that exists in the data set and then purifying the data set.

CART does this by recursively partitioning a data set into groups or terminal nodes that are internally more homogenous than their castor nodes. At each terminal node, the mean value of the dependent variable is taken as the predicted value. If the objective of a regression tree is explanation, then this is achieved by tracking the paths of a tree to a specific terminal node.

3.5.2 BUILDING A REGRESSION TREE:

The process of constructing a regression tree is similar to that for building a classification tree. Regression tree building centers on three major components:

(1) A set of questions of the form,

$$Is\ X\ \leq\ d?,$$

where X is a variable and d is a constant.

(2) goodness of split criteria for choosing the best split on a variable;

(3) the generation of summary statistics for terminal nodes (unique to a regression tree).

The mechanism for building a regression tree is similar to that for a classification tree. But with a regression tree there is no need to specify priors and misclassification costs. Further more, the dependent variable in a regression tree is numeric or continuous. The splitting criterion employed is the within node sum of squares of the dependent variable and the goodness of a split is measured by the decrease achieved in the weighted sum of squares.

To build regression tree starting with the root node, CART performs all possible splits on each of the predictor variables, applies a predefined node impurity measure to each split, and determines the reduction in impurity that is achieved. and selects the "best" split by applying the goodness of split criteria and partitions the data set into left and right child nodes.

Because CART is recursive, it repeats above processing for each of the nonterminal nodes and produces the largest possible tree. and finally, CART applies its pruning algorithm to the largest tree and produces a sequence of sub-trees of different sizes from which an optimal tree is selected.

3.5.3 Splitting Rules and Goodness of Fit Criteria:

There are two splitting rules or impurity functions for a regression tree. These are

(1) The Least Squares(*LS*)function.

(2) The Least Absolute Deviation (*LAD*) function.

Since the mechanism for both rules is the same, only the *LS* impurity measure will be described. Under the *LS* criterion, node impurity is measured by within node sum of squares, *SS(t)*, which is defined as:

$$SS\ (t) = \sum(y_i - y_t)^2 \quad for\ \ i = \ 1, 2, \ldots, N_t, \_\_\_\_\_ \text{(10-3)}$$

Where:

$y_i\ (t)$= individual values of the dependent variable at node *t*,

$y\ (t)$= the mean of the dependent variable at node *t*. Given the impurity function, *SS(t),* and splits that sends cases to left (t_L) and right (t_R)nodes, the goodness of a split is measured by the function:

$$\emptyset(s, t) = ss(t) - ss(t_r) - ss(t_l) \_\_\_\_\_ \text{(11-3)}$$

Where:

$SS(t_R)$ is the sum of squares of the right child node

$SS(t_L)$ is the sum of squares of the left child node.

The best split is that split for which is the highest. From the series of splits generated by a variable at a node, the rule is to choose the split the results in the maximum reduction in the impurity of the parent node.

An alternative to *SS(t)* is to use the weighted variance of left and right nodes, where the weights are proportions of cases at nodes t_L and t_R:

let $p(t) = N_t/N$ be the proportion of cases at node *t*, and let $s^2(t)$ be the variance of the dependent variable at node *t*. The variance is defined as

$$s^2(t) = 1/N_t(\sum_{i=1}^{Nt} y_i - y_t)^2 \_\_\_\_\_ \text{(12-3)}$$

The goodness of a split is now measured by:

$$\emptyset(s, t) = s^2(t) - [p_l s^2(t_l) + p_r s^2(t_r)] \_\_\_\_\_ \text{(13-3)}$$

The best split is the one for which $\emptyset$(s,t) is the highest or for which the weighted sum of the variances$[p_l s^2(t_l) + p_r s^2(t_r)]$is the smallest.

The procedure successfully separates high values of the dependent variable from its low values and results in left and right nodes that are now internally more homogenous than the parent node. It should be noted that as each split sends observations to the left and right nodes, the mean of the dependent variable in one of the resulting nodes is lower than the mean at the parent node.

3.6. Advantages and Disadvantages ofCART

3.6.1. Advantages of CART:

• CART is nonparametric; therefore this method does not require specification of any functional form.

• CART does not require variables to be selected in advance, CART algorithm will itself identify the most significant variables and eliminate non-significant ones.

To test this property, one can insignificant (random) variable and compare the new tree with tree built on initial dataset. Both trees should be grown using the same parameters (splitting rule and N_{min} parameter).

• CART results are invariant to monotone transformations of its independent variables.

Changing one or several variables to its logarithm or square root will not change the structure of the tree. Only the splitting values (but not variables) in the questions will be different.

• CART can easily handle outliers. Outliers can negatively affect the results of some statistical models, like Principal Component Analysis (PCA) and linear regression. But the splitting algorithm of CART will easily handle noisy data: CART will isolate the outliers in a separate node.

• CART has no assumptions and computationally fast.

• CART is flexible and has an ability to adjust in time.

3.6.1.Disadvantages of CART:

• CART may have unstable decision trees. Insignificant modification of learning sample, such as eliminating several observations, could lead to radical changes in decision tree, increase or decreases of tree complexity, changes in splitting variables and values.

• CART splits only by one variable. This mean, all splits are perpendicular to axis.

- The CART may not catch the correct structure of the data. if can not correctly identify question , because in split question can participate only one variable. In order to capture the datastructure, splitting algorithm will generate many splits (nodes) at the border *of* $x_1 - x_2 \leq 0$ line.

4-1:Data Analysses:

The data is sample size(100) to Sex, Age, Martial status, alcohol drinking, TB.

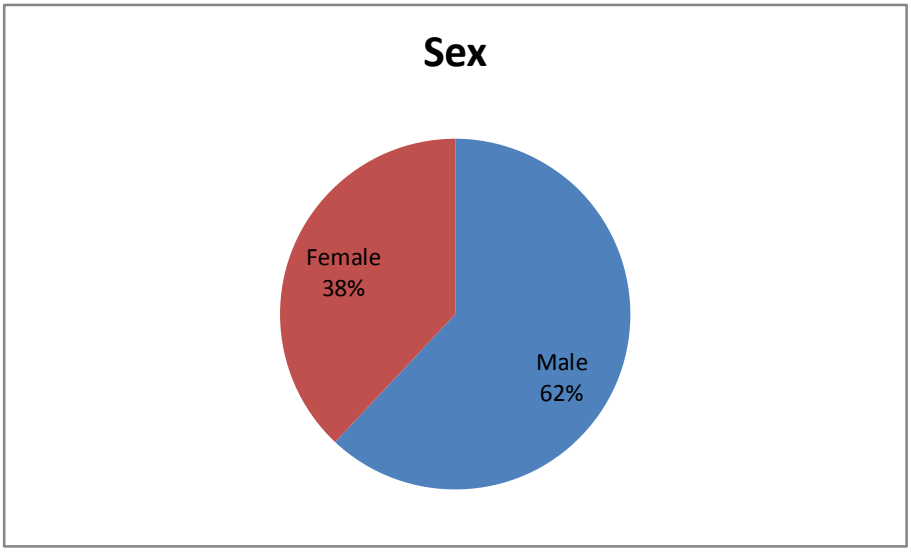
4-2: Summarization of data:

Table(1-4):Sample distribuation acording sex variable:

Sex	Percent
Male	62%
Female	38%
Total	100%

From the table above the majorly are men 62%, while the rest are femalesis 83%

Figure (1-4):Sample distribuation acording sex variable



62% from the sample were male and 38% female which equal to 1/3 from the sample size.

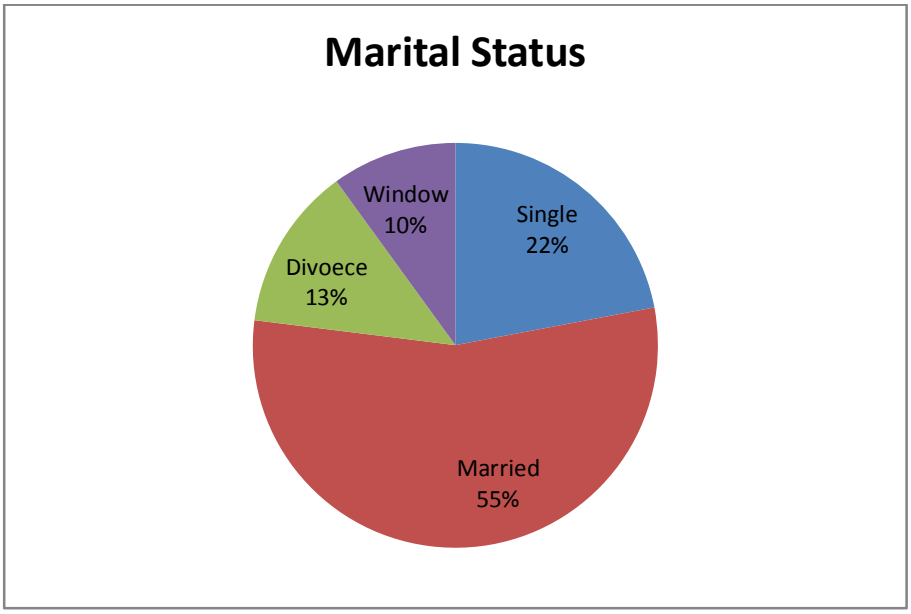
Table(2-4): The Marital Status:

marital status	Percent
Single	22%
Married	55%
Divorce	13%
Widow	10%

Total	100%

From the table above the number of Single in the sample is22 by 22%, Marriedis55 by 55%,Divorceis 13 by 13%, andWidowis 10 by 10%

Figure (2-4):The Marital Status:



55% from the AIDS pationt is married, 25% is single,13% divorced and 10% widwow.

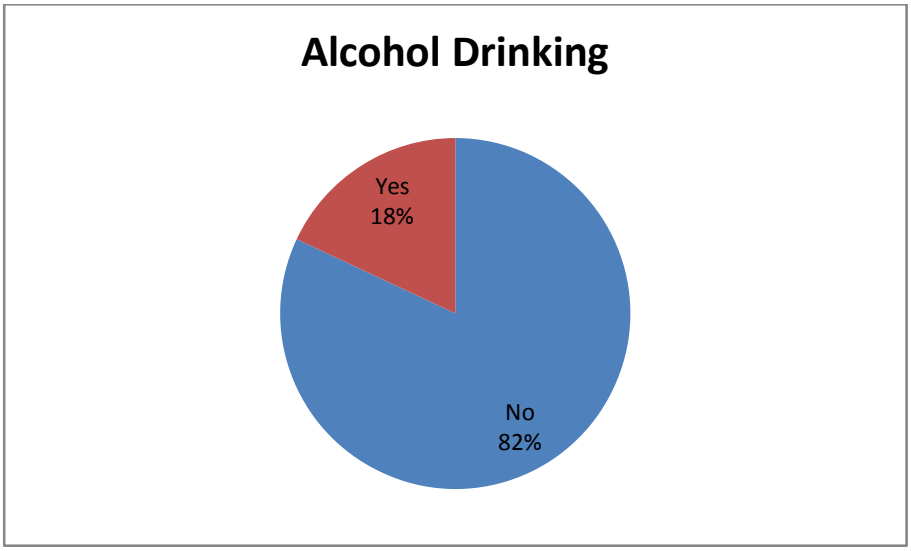
Table(3-4): Sample distribuation acording Alcohol Drinking variable

Alcohol Drinking	Percent
No	82%

Yes	18%
Total	100%

From the table above the number of Alcohol Drinking isNo in the sample is 82 by 82%, andYesis 18 females by 18%

Figure (3-4): Sample distribuation acording Alcohol Drinking variable



For the purpose that many pationt not drinking alcohol we found 18% only from the sample size taking alcohol vs 82% not taking alcohol.

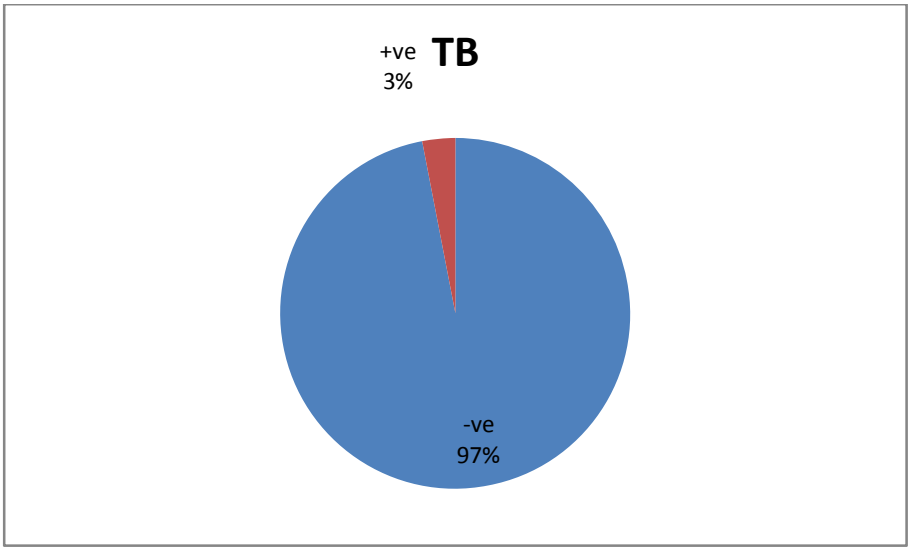
Table(4-4): Sample distribuation acording TB variable

TB	Percent

-ve	97%
+ve	3%
Total	100%

From the table above the number of –veTB in the sample is 97 by 97%, and is 38 +veTB by 83%

Figure (4-4):Sample distribuation acording TB variable



We found that the percentage of TB infection through out the AIDS pation is very little (3%) only, and this is due to the health care that provided form the healte center patient which eliminate the disease.

Sample distribuation acording Age variable:

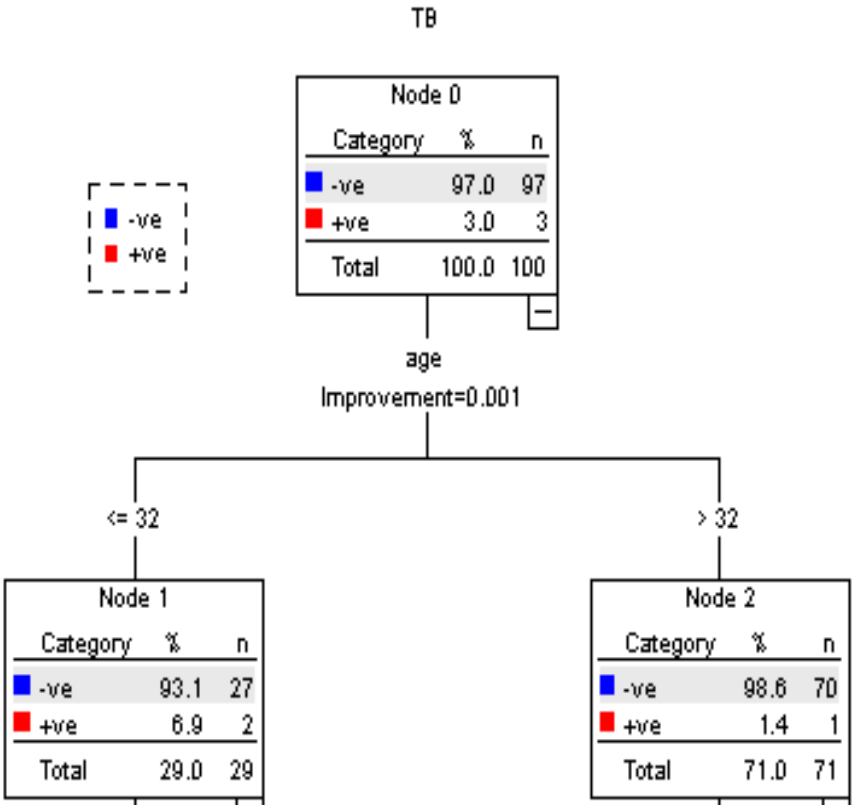
Table(5-4)

Age	Mean	Std.deviation
	38.30	10.354

The averge age of the samle is 38.30 $y_{rs}$with standared diviation 10.354, and that age concerned as the productive age.

4-3:Classification Tree:

Figure (5-4): Level 1



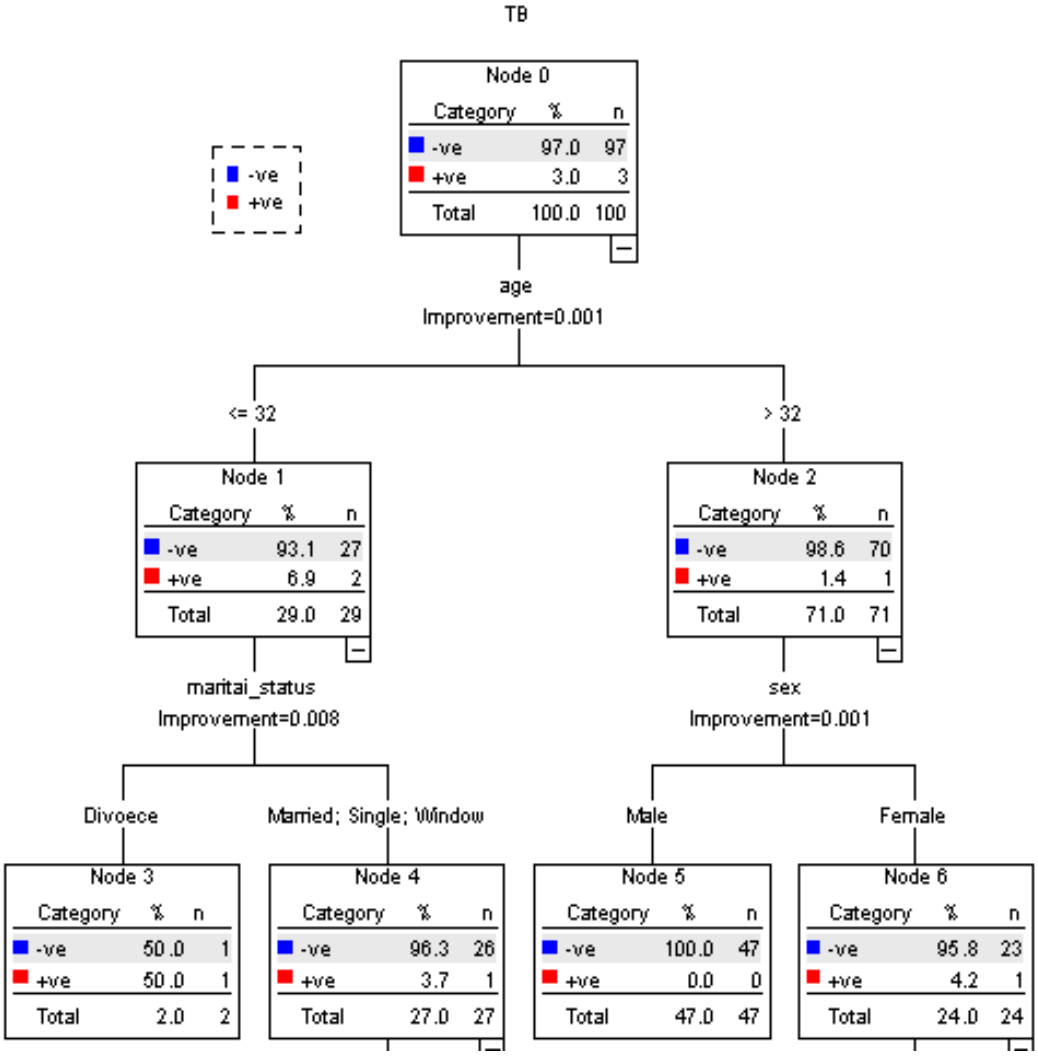
We classify the TB patient according to age variable, we classify two node that the age is either greater than 32 or less than or equal to32.

Table(6-4):Shows the node number 1 and 2:

Age	+ve	-ve
≤32	6.9%	93.1%
>32	1.4%	98.6%

The above table shows that in ≤32 node the TB patient is 6.9% where as in >32 node the patient is 1.4%.

Figure (6-4):Level 2



In level 2 from the classification tree, the marital status variable devided acording to age distribution in level 1 and 3 categories of the marital status variable merge (married, singel and widow) because no significance difference between them.

Table(8-4): Shows the node number 3 and 4:

marital status	+ve	-ve
Divorce	50%	50%
Married-Divorce-Widow	3.7%	96.3%

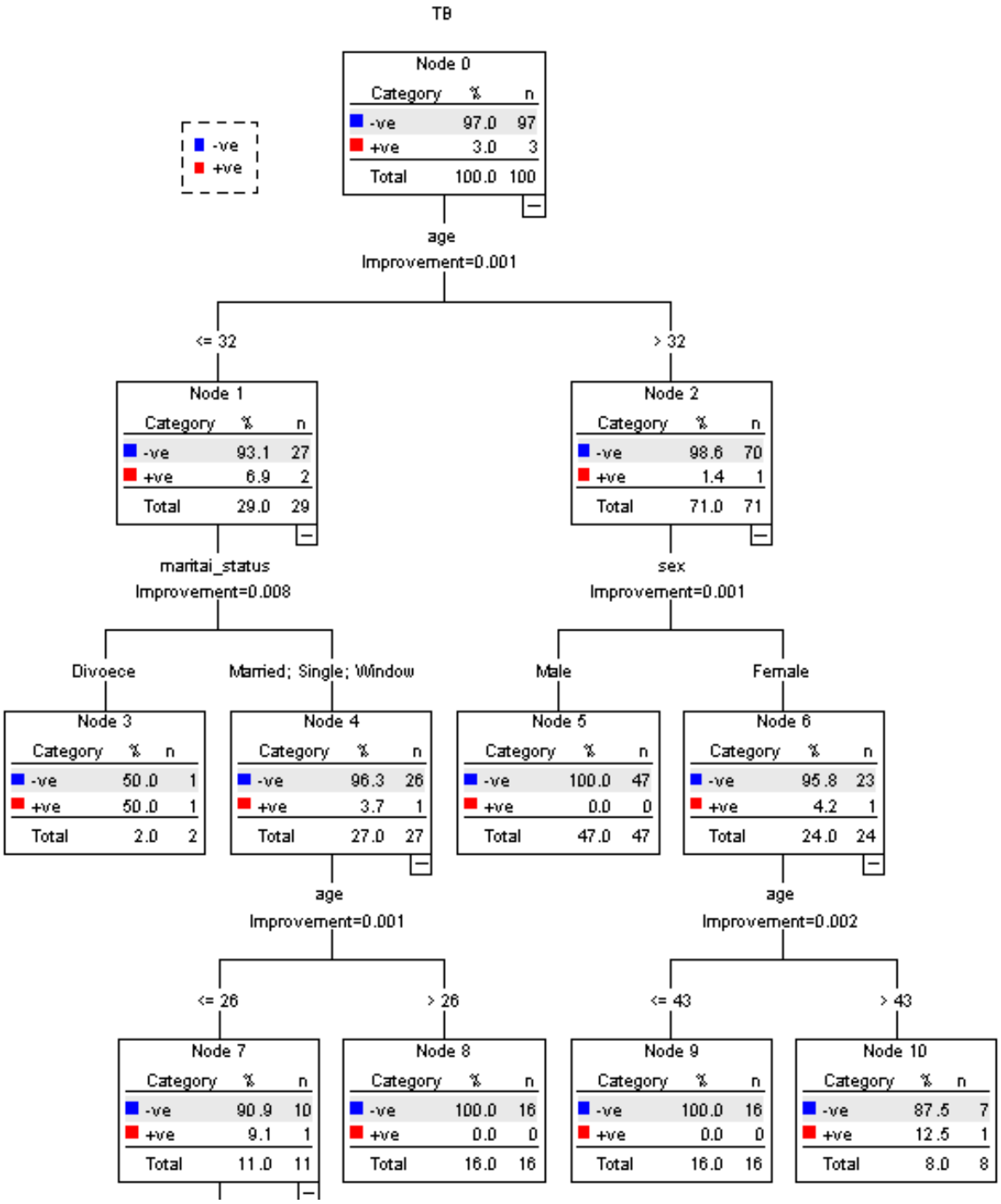
From the above table, I notice that 50%from the divorced are the +ve to TB, where as 3.7% only from married, single and widow is +ve to TB.

Table(7-4)Shows the node number 5 and 6:

Sex	+ve	-ve
Male	0%	100%
Female	4.2%	95.8%

From the above table, I notice that in this level 4.2% female had TB, where no male is affected.

Figure (7-4): Level 3



From figure (7-4):

We classify the age for the second node for the marital status and classify the female for the sex variable.

Table(9-4):Shows the node number 7 and 8:

Age	+ve	-ve
≤26	9.1%	90.9%
>26	0%	100%

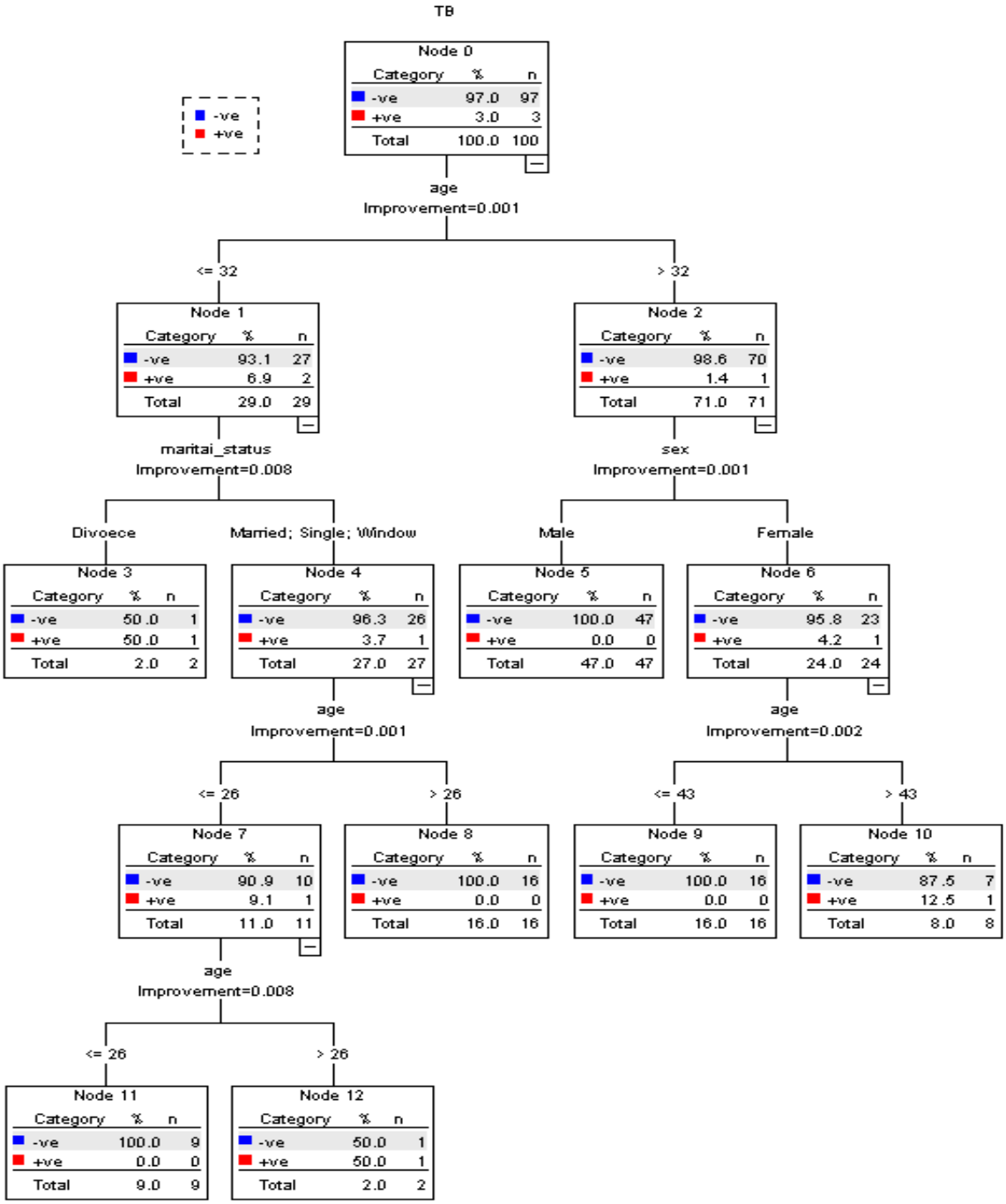
The percentage of infection in age category ≤26 is 9.1%, where as no infection in age category>26.

Table(10-4):Shows the node number 9 and 10:

Age	+ve	-ve
≤43	0%	100%
>43	12.5%	87.5%

The percentage of infection in age category>43 is 12.5%, where as no infection in age category ≤43.

Figure (8-4): Level 4



From figure (8-4):

In the last level from the classification tree, we classify the infection of the disease according to age variable.

Table(11-4):Shows the node number 11 and 12:

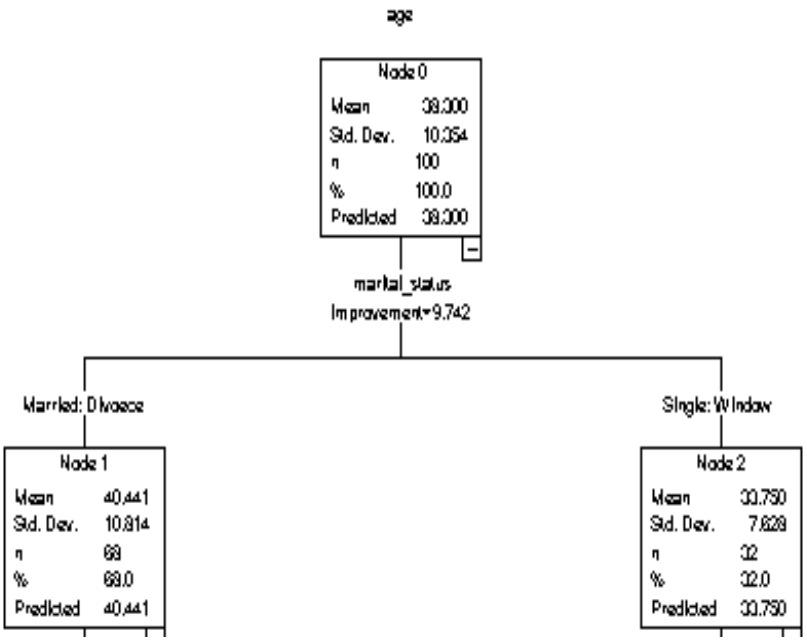
Age	+ve	-ve
≤26	0%	100%
>26	50%	50%

In age catogery >26 we found 50% is TB pationt, and 50%have no TB.

*We notice that alcohol drinking variable not appear in classification tree and that due to no significance inferance to comparison by the other variables

4-4:Regression Tree:

Figure (9-4): Level 1

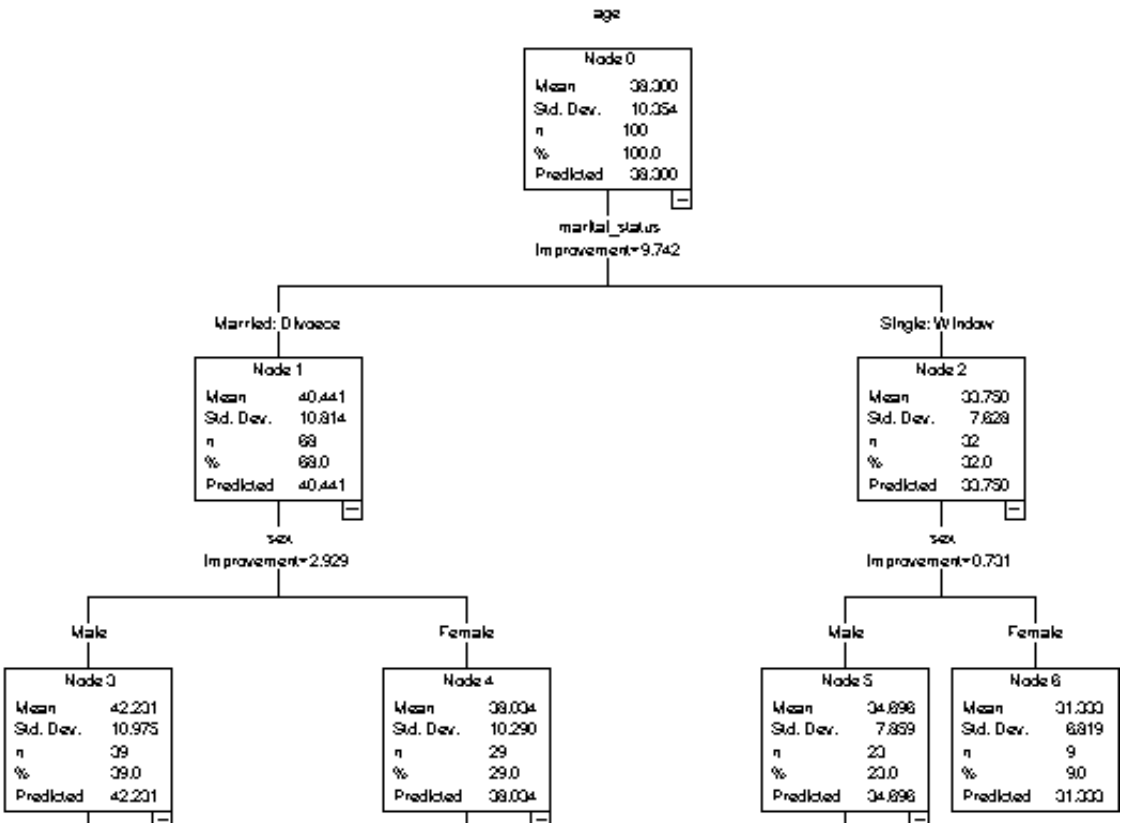


Table(12-4):Shows the node number 1 and 2:

Marital status	Mean	S.D	N	Prediction
Marred, devoice	40.44	10.81	69	40.44
Single, widow	30.75	7.62	32	30.75

I notes that the mean of marred and divorce is40.44, and the S.D equal 10.81. and the mean of single and widow 30.75, with the S.D equal 7.62.

Figure (10-4): Level 2

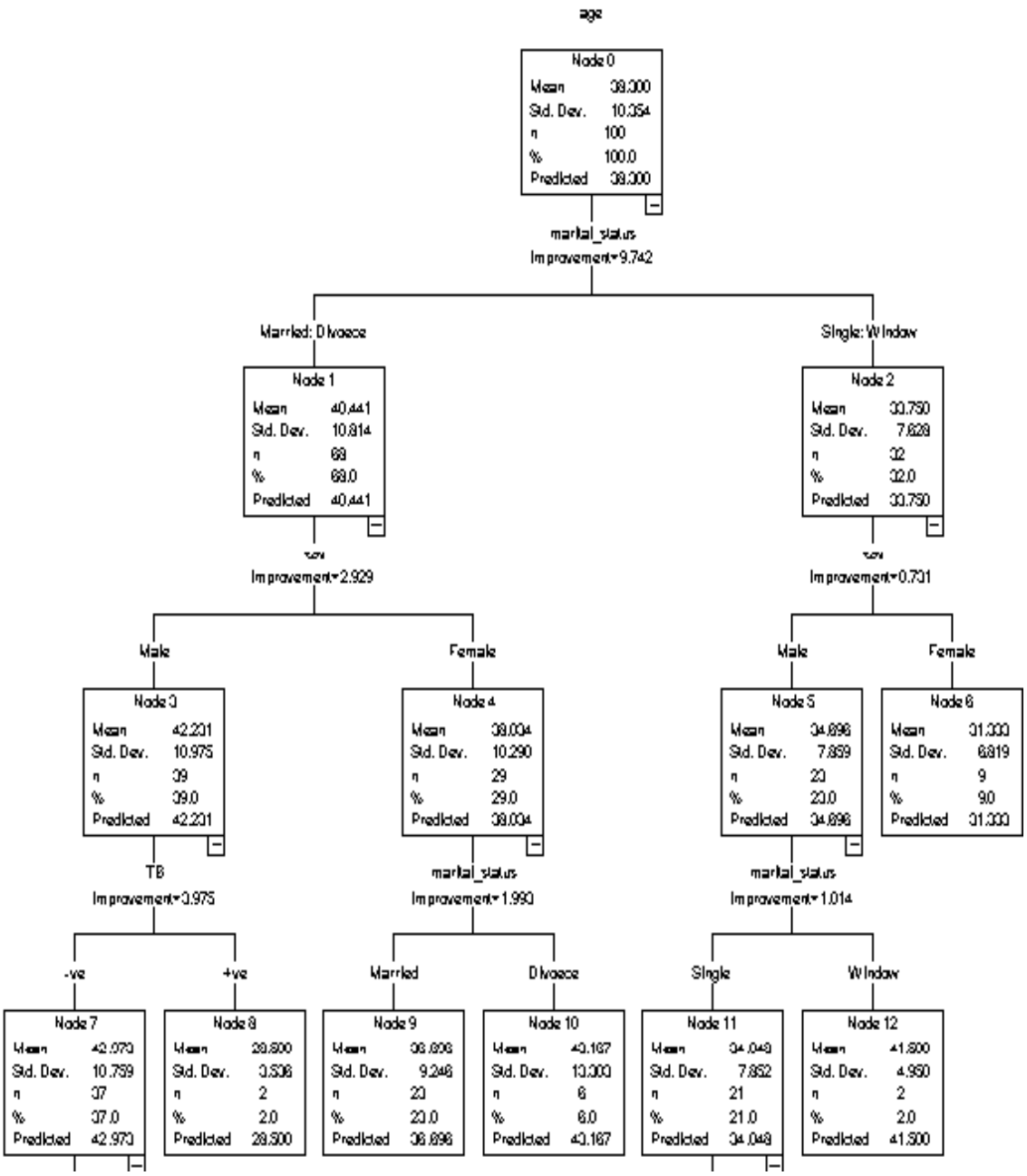


Table(13-4):Shows the node number 3 and 4:

Sex	Mean	S.D	N	Prediction
Male	42.20	10.97	39	42.20
Female	39.00	10.29	29	39.00

From table (13-4), I notes that the mean of male is 42.20, and the S.D equal 10.9and the mean of female is 39.00, and the S.D equal 10.29.

Figure (11-4): Level 3

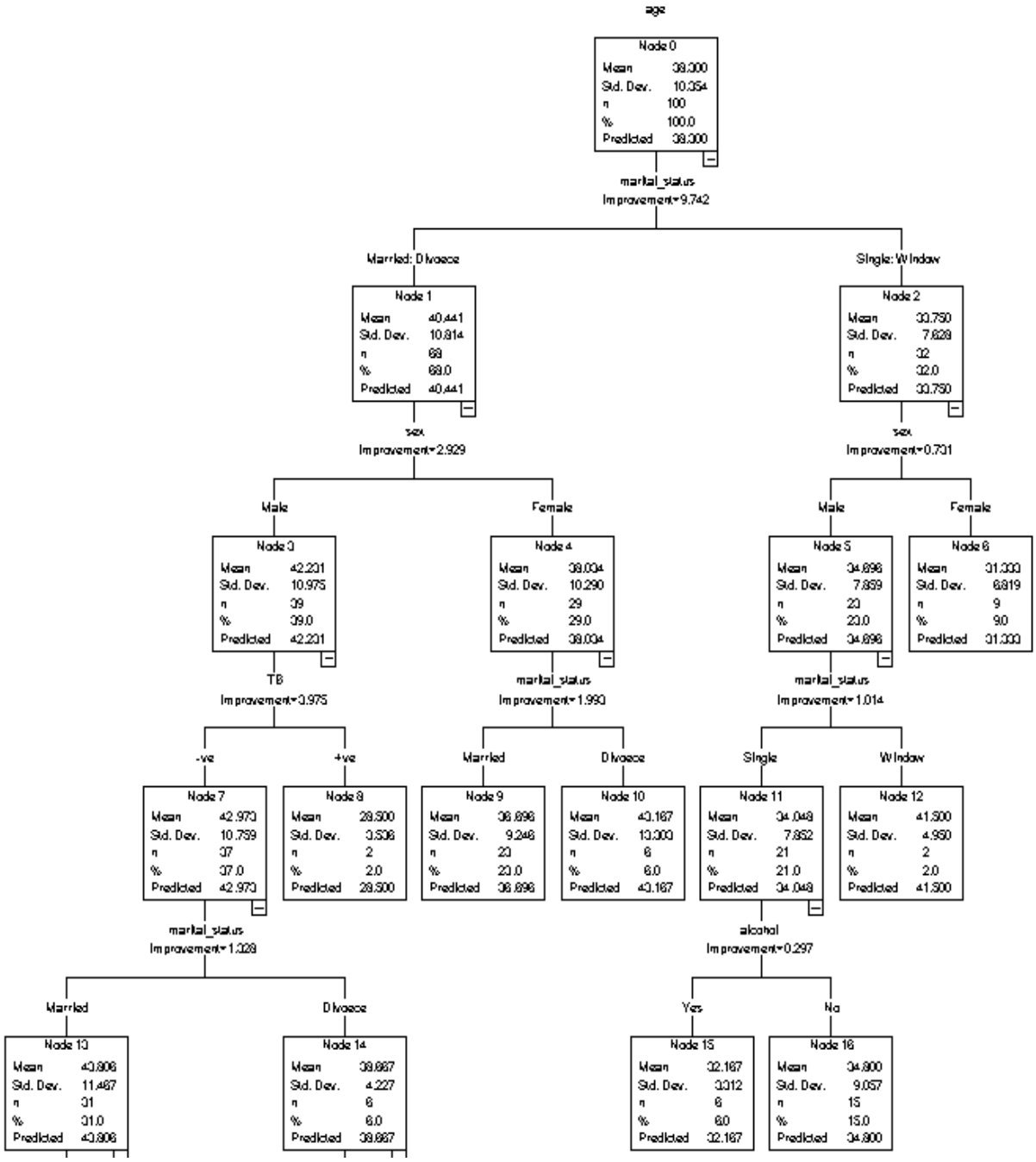


Table(14-4):Shows the node number 7 and 8:

TB	Mean	S.D	N	Prediction
+v	42.97	10.75	37	42.97
-v	29.50	3.50	2	29.50

The mean of +ve TB is 42.97 with S.D 10.75. the mean of -ve TB is 29.50with S.D 3.50.

Figure (12-4):Level 4

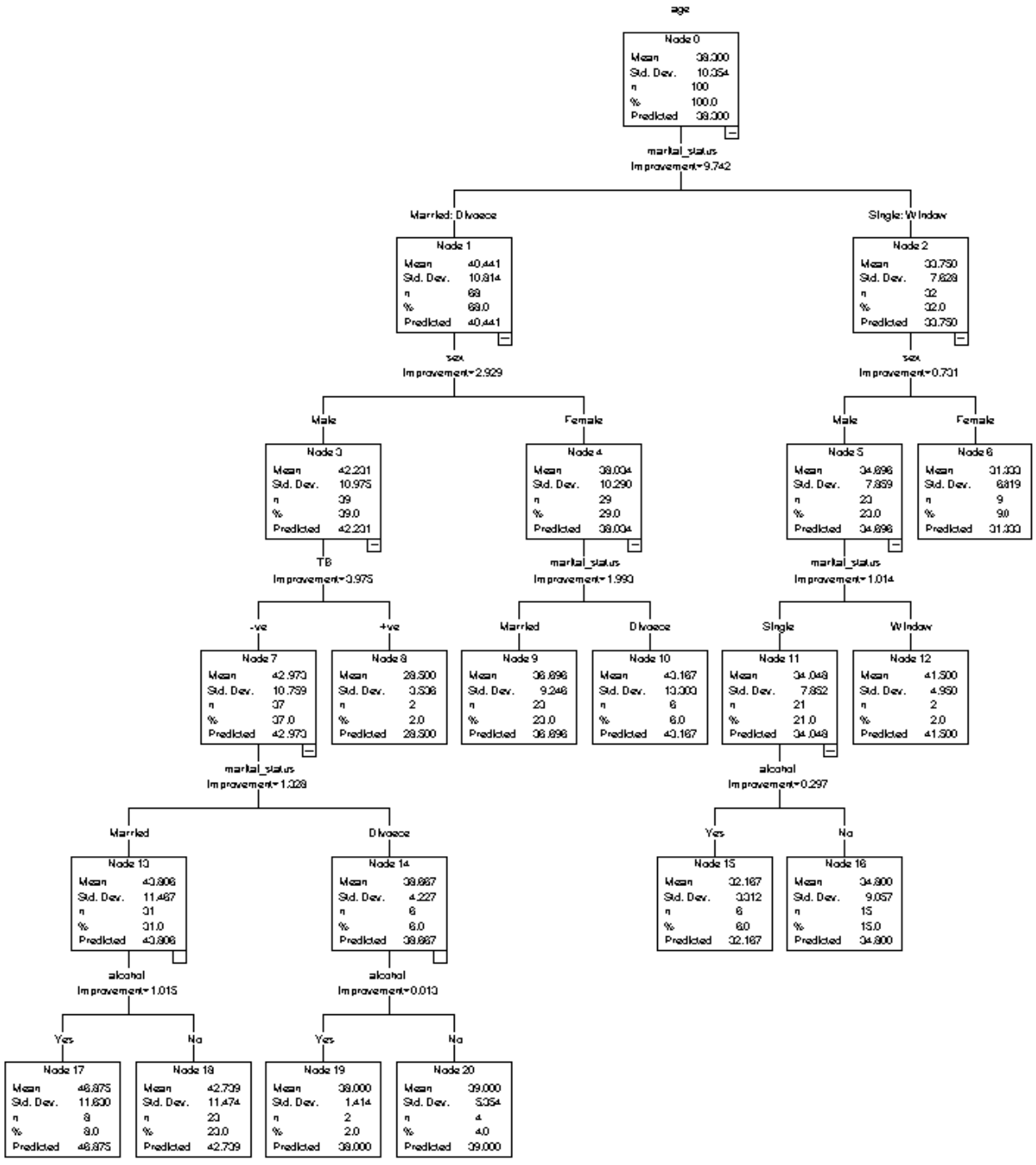


Table(15-4):Shows the node number 7 and 8:

Marital status	Mean	S.D	N	Prediction
Marred	43.80	11.48	31	43.80
Divorce	38.88	4.22	8	38.88

From the above table, I notice that mean of the marred is 43.80 with S.D 11.48. And the mean of Divorce is 38.88 with S.D 4.22.

Figure (13-4): Level 5



Table(16-4):Shows the node number17 and18:

Alcohol Drinking	Mean	S.D	N	Prediction
Yes	48.87	11.60	8	48.87
No	42.70	11.47	20	42.70

From the above table, we notes that the mean of Alcohol Drinking take yes is 48.87 and the S.D equal 11.60.

The mean of Alcohol Drinking take no is 42.70, and the S.D equal 11.47.

Table(1-5): Example of predicting from classification tree

Age	Marital Status	Prediction of TB
26	Married	+ve
40	Divorce	-ve

For example we can predict from classification tree if the patient age is 26 and marital status is married the prediction of TB is +ve. Or the patient age is 40,marital status is divorce the prediction of TB is –ve

Table(2-5): Example of predicting from regression tree

Marital Status	Sex	TB	Prediction of Age
Marred, Divorce	Male	+v	29.50
Marred, Divorce	Male	-v	42.97

For example we can predict from regression tree if the patient Marital status is marred ordivorce,sex is male and TB is +veprediction ofage is 29.50. Or the patient maritalstatus is marred ordivorce and sex is male and TB is -veprediction of age is 42.97.

5-1: Results:

1. We can predict value of the dependent variable in the CART, whether categorical or continuous, depending on other variable.
2. In the classification tree, the dependent variable is (+ve) TB and the independent variables are age, sex and marital status.
3. In the regression tree, the dependent variable is age and the independent variables are TB, sex, alcohol drinking and marital status.
4. For example we can predict from classification tree if the patient age is 26 and marital status is married the prediction of TB is +ve.
5. For example we can predict from regression tree if the patient marital status is marred ordivorce and sex is male and TB is +veprediction of age is 29.50.

5-2: Recommendations:

1. The sample size must be increased to obtain a higher degree of accuracy by pruning the tree of good and choose the contract carefully.
2. Increase the number of variables involved in the model to include other independent variables that are related to the emergence of tuberculosis patients, such as those for housing and the surrounding environment , which directly affect the injury.
3. Apply the same form marked and used to predict pulmonary TB for people living with AIDS has not been applied to any treatment they do not see the symptoms and comparing this model .
4. How to predict the value of the dependent variable in the long-Term.
5. Prediction of other problems that occur to patients such as renal failure and liver failure.

REFERENCES:

1. Cheeseman SH, Havlir D, McLaughlin MM, et al. 1995. Phase I/II evaluation of nevirapine alone and in combination with zidovudine for infection of immunodeficiency virus. Journal of the Acquired Immmune Deficiency Syndromes and Human Retro virology.
2. David Woodm,*A New Type of Data Management System*,(2004),www.tucanatech.com
3. International Food Policy Research Institute, AIDS, Poverty, and Hunger(Challenges and Responses), April,(2005), Washington, D.C.
4. Michael J.A. Berry,*Data Mining Techniques For Marketing, Sales,* and Customer Relationship Management,(2004), Inc., Indianapolis, Indiana, Wiley Publishing.
5. Margaret Henty, Investigating Data Management July, (2008) Http: //www.apsr.edu.au/investigating_data_management.
6. PAOLO GIUDICI & SILVIA FIGINI, Applied *Data Mining for Business and Industry*,(2009),Italy,John Wiley & Sons Ltd.
7. Ranki A, Johansson E, and Krohn K (1988), Interpretation of Antibodies Reacting Solely with Human Retroviral Core Proteins.
8. WHO. (1986) Acquired Immunodeficiency Syndrime (AIDS). WHO/CDC case definition for AIDS. Weekly Epidemiology.
9. http://www.avert.org/history-hiv-and-aids.htm
10. TheHIVAIDSQuestion page.
11. GHAPHIVAidsHandbook
12. http://www.mayoclinic.org/diseases-conditions/hiv aids/basics/definition
13. http://www.ncbi.nlm.nih.gov/books

APPENDEX:

NO	SEX	AGE	MARITAL STATUS	ALCOHOL DRINKINK	TB
1	1	50	2	1	0
2	1	43	2	0	0
3	2	35	2	0	0
4	2	26	2	0	0
5	1	51	2	0	0
6	1	42	1	0	0
7	1	45	2	0	0
8	1	42	1	0	0
9	1	37	3	1	0
10	1	40	2	1	0
11	2	30	2	0	0
12	1	38	4	1	0
13	2	42	2	0	0
14	2	34	4	0	0
15	1	30	2	0	0
16	1	39	3	1	0
17	1	54	1	0	0
18	1	65	2	0	0
19	2	47	4	0	0
20	2	30	4	0	0
21	2	46	2	0	0
22	1	24	1	0	0
23	1	53	2	0	0
24	1	35	1	0	0
25	2	30	4	0	0
26	2	30	4	0	0
27	1	51	2	0	0
28	1	37	1	1	0
29	1	44	1	0	0
30	1	38	1	0	0
31	1	40	2	0	0
32	1	40	3	0	0
33	1	32	1	1	0
34	1	50	2	1	0

Continue

NO	SEX	AGE	MARITAL STATUS	ALCOHOL DRINKINK	TB
35	2	34	2	0	0
36	2	35	2	0	0
37	1	40	2	1	0
38	2	27	2	0	0
39	1	32	3	0	0
40	1	21	1	0	0
41	1	40	2	1	0
42	2	39	3	0	0
43	1	35	2	0	0
44	1	26	2	0	1
45	1	24	2	0	0
46	2	30	2	0	0
47	2	60	2	0	0
48	1	45	4	0	0
49	1	42	2	0	0
50	1	33	1	1	0
51	1	29	1	1	0
52	2	37	2	0	0
53	2	37	3	0	0
54	1	37	1	0	0
55	2	25	2	0	0
56	2	27	4	0	0
57	2	34	4	0	0
58	2	23	4	0	0
59	2	35	3	0	0
60	2	25	2	0	0
61	1	28	1	1	0
62	2	32	2	0	0
63	1	65	2	1	0
64	1	52	2	0	0
65	1	34	1	1	0
66	1	39	3	0	0
67	2	67	3	0	0
68	1	33	1	0	0
69	1	45	3	0	0
70	2	44	2	0	1

Continue

NO	SEX	AGE	MARITAL STATUS	ALCOHOL DRINKINK	TB
71	2	31	2	0	0
72	1	24	2	0	0
73	1	60	2	1	0
74	1	54	2	0	0
75	1	31	3	1	1
76	1	30	1	0	0
77	1	22	1	0	0
78	1	30	2	1	0
79	2	35	2	0	0
80	1	31	2	0	0
81	1	40	2	0	0
82	1	43	2	0	0
83	1	42	2	0	0
84	2	50	2	0	0
85	1	58	2	0	0
86	2	35	2	0	0
87	2	39	2	0	0
88	2	31	3	0	0
89	1	27	1	0	0
90	1	42	2	0	0
91	1	36	2	0	0
92	2	34	2	0	0
93	2	55	2	0	0
94	1	58	2	0	0
95	2	37	2	0	0
96	2	50	3	0	0
97	1	24	2	0	0
98	1	34	1	0	0
99	1	39	1	0	0
100	2	27	1	0	0

Sex: 1 =Male, 2 =Female

MARITAL STATUS:1 =Single, 2 =Married, 3 =Divorce, 4 =Widow

ALCOHOL DRINKINK: 0=No, 1=Yes

TB:0=No, 1=Yes