ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 11, November 2016

ORAL CANCER DETECTION USING DATA ANALYSIS

C. venkata Prasad
M-Tech Student, Department of CSE
JNTUA College of Engineering, Ananthapuramu

Dr. A. P. Siva Kumar
Assistant Professor, Department of CSE
JNTUA College of Engineering, Ananthapuramu

***Abstract*--Oral cancer also called oral cavity cancer. It had developed in any part of the mouth. The causes for oral cancer are heavy tobacco usage and more alcohol usage. Symptoms for this cancer are a lump or a white patch or red patch on the inside of the mouth. In our proposed process, we are detecting the oral cancer by analyzing and constructing decision tree to the data set. The data set contain the data about the oral cancer patients. This process is cost effective and also it is easy process. In the present system we use R programming to detect the oral cancer.**

***Index Terms*— dataset, oral cancer, rpart.**

I. INTRODUCTION

Every year new cases of oral cancer are growing in the world. The tobacco usage is the main cause for the oral cavity cancer. This oral cancer can form in any place of the mouth like tongue, throat etc. In the world, most of the male people suffering from oral cancer than female people. The family history is also one of the causes for oral cancer. In the early stage of oral cancer white and red patches found in mouth, the teeth are become loose and pain will come when swallowing. In the advanced stage of oral cancer bleeding will done in the mouth, a lump will found in the mouth and also the voice would not clear.

When oral cancer had effected, the cells and lymph node which are in the mouth are damaged. First the oral cancer begins in cells which are in the mouth, and then it will grow from those cells. These cells will make tissues. These effected cells, will generate new cells those are also damaged cells. So it forms a tumor in mouth. To control the oral cancer first stop the tobacco usage, smoking cigarettes and drinking alcohol. In this present work oral cancer will identified by data analysis. This data analysis is done by using the r programming. In the proposed system, r studio tool was used for r programming.

II. DATA AND PROCESS

A. Data set

In the present work considered 93 patients information, who are suffering from oral cancer. This data set contains 8 attributes those are like age, site of the tissue, smoking, drinking and etc. In this dataset all patients had been classified in to two group's namely early stage and advanced stage. Dataset contains the data all about the oral cancer patients which are already affected.

This data set contains the information about the oral cancer patients. In this data set all the patients are in early stage or in advanced stage. Some members have middle age but those are in advanced stage of oral cancer. Because their tobacco usage attribute is yes.

To perform operations on the dataset utilizing a few methodologies among the characteristics r programming procedure is the best way to deal with perform measurable operation on dataset. R writing computer programs is utilized to play out the measurable operations on dataset. Presently R writing computer programs is generally utilized as a part of numerous stages in this world. Significantly R writing computer programs is utilizing as a part of the vehicle business, climate anticipating, and organization financial spending plan examining purposes. Indeed, even in bio-medicinal classification likewise R programming assumes an essential part. By utilizing this R programming, we can perform diverse operations on the dataset with the assistance of numerical operations .There are a great deal of scientific operations to perform investigation on the dataset. By utilizing R-Studio with R programming operations. This R-studio is utilized to play out the measurable operation on the datasets likewise we can spare the entire venture for future reference.

Age	Sex	Site	Clinical classification	Lymph node metastasis	History of OLK	Smoking	Drinking
64	F	G	Advanced stage	Y	N	N	Y
75	M	G	Early stage	N	N	N	N
58	M	P	Early stage	N	N	Y	Y
76	M	F	Advanced stage	Y	N	Y	Y
76	M	B	Advanced stage	Y	N	N	Y
55	F	G	Early stage	N	N	N	N
62	M	F	Early stage	N	N	N	N
68	F	G	Early stage	N	N	N	N
70	F	T	Advanced stage	Y	N	N	N
63	M	G	Early stage	N	N	N	N
63	M	G	Early stage	N	N	N	N
52	F	T	Advanced stage	Y	Y	N	N
52	F	T	Advanced stage	Y	Y	N	N
58	M	G	Early stage	N	N	Y	Y
74	F	T	Early stage	N	Y	N	N
60	F	T	Early stage	N	N	N	N
79	M	G	Early stage	N	N	N	N
83	M	B	Early stage	N	N	Y	N
79	M	B	Early stage	N	N	N	N
60	F	T	Early stage	N	N	N	N
79	M	G	Early stage	N	N	N	N
56	M	G	Early stage	N	N	Y	Y
60	F	T	Early stage	N	N	N	N
73	F	T	Advanced stage	N	Y	N	N
21	M	T	Early stage	N	N	Y	N
51	M	T	Early stage	N	Y	Y	Y

Fig 1: Dataset

B. Process

In this work, R-studio tool used to analyze the dataset, which is run by using r programming. This programming is one of the best programming for analysis of data and also gives best results. In this process, first the dataset import by the import option which is in r studio tool. R programming is very useful for statistical analysis of dataset. In the present process, construct the decision tree by using the predefined packages which are in r studio. In the Present work r 3.3.0 version had used to analyzing the dataset.

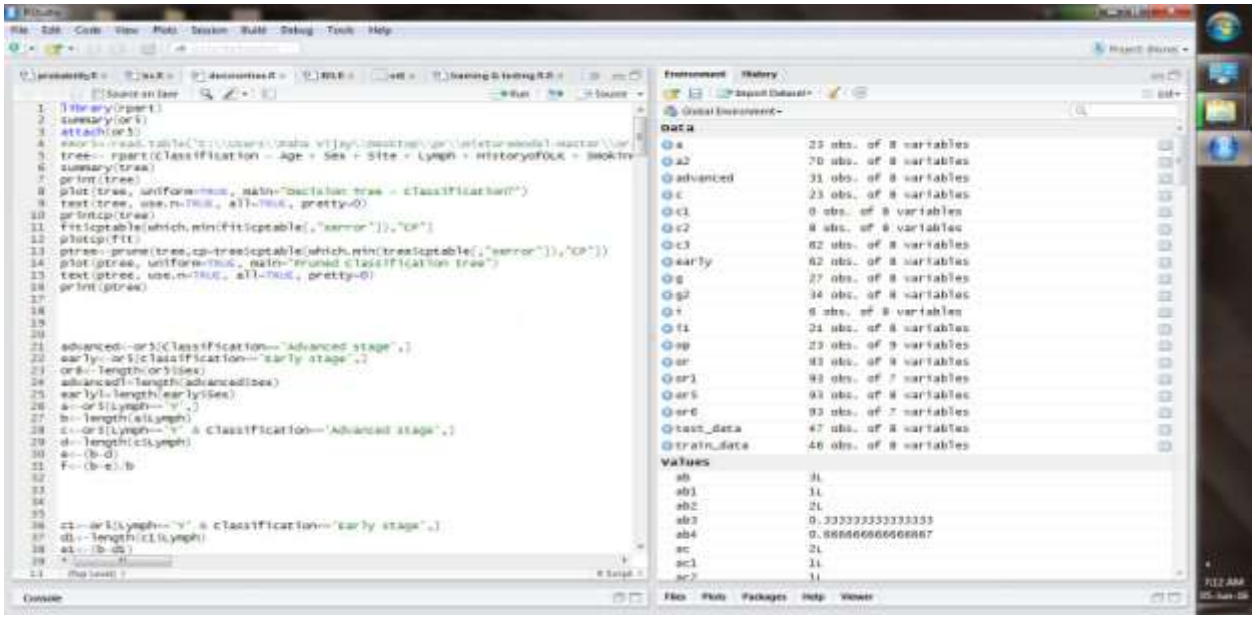

Fig 2: R-studio

In the proposed process first find out the probability of all attributes based on the classification attribute, which is the patient is in early stage or in advanced stage.

$$probability\left(\frac{A}{B}\right) = \frac{no.\,of\ favorable\ outcomes}{total\ no.\,of\ outcomes}$$

By using the above formula find the probability of all attributes. Next find out the variable importance for the attributes which are in dataset. This variable importance is used to find which attribute has highest priority and which is next to divide in to sub parts. After finding the variable importance, the attribute lymph node got highest priority so this node acts as root node and then the tree constructed based on other attributes. The variable priority sequence in the present work is lymph node, site, age, sex and smoking. By using the following r program probability will calculated.

```
r=length((or5[(Age<67.5 & Age<69 & Age>=66.5 &
Site=='F' & Lymph=='N') |
          (Age<67.5 & Age<69 & Age>=66.5 &
Site=='G' & Lymph=='N') |
          (Age<67.5 & Age<69 & Age>=66.5 &
Site=='P' & Lymph=='N'),])$Age)
```

```
r1=length((or5[(Age<67.5 & Age<69 & Age>=66.5
& Site=='F' & Lymph=='N' &
          Classification=='Advancedstage')  |
(Age<67.5 & Age<69 & Age>=66.5 & Site=='G' &
Lymph=='N' &
          Classification=='Advanced  stage')  |
(Age<67.5 & Age<69 & Age>=66.5 & Site=='P' &
Lymph=='N' &
          Classification=='Advanced
stage'),])$Age)
r2=length((or5[(Age<67.5 & Age<69 & Age>=66.5
& Site=='F' & Lymph=='N' & Classification=='Early
stage') |
          (Age<67.5 & Age<69 & Age>=66.5 &
Site=='G' & Lymph=='N' & Classification=='Early
stage') |
          (Age<67.5 & Age<69 & Age>=66.5 &
Site=='P' & Lymph=='N' &  Classification=='Early
          stage'),])$Age)
s=r1/r
s1=r2/r
```

in this r,r1,r2,,s1 are variables gives probability of attribute.

In this process the root node is divided in to two parts, one part contains nodes which have lymph node attribute as ‘yes’ and all are In advanced stage. Second part contains nodes which have lymph node attribute as ‘yes’ and all are in early stage.

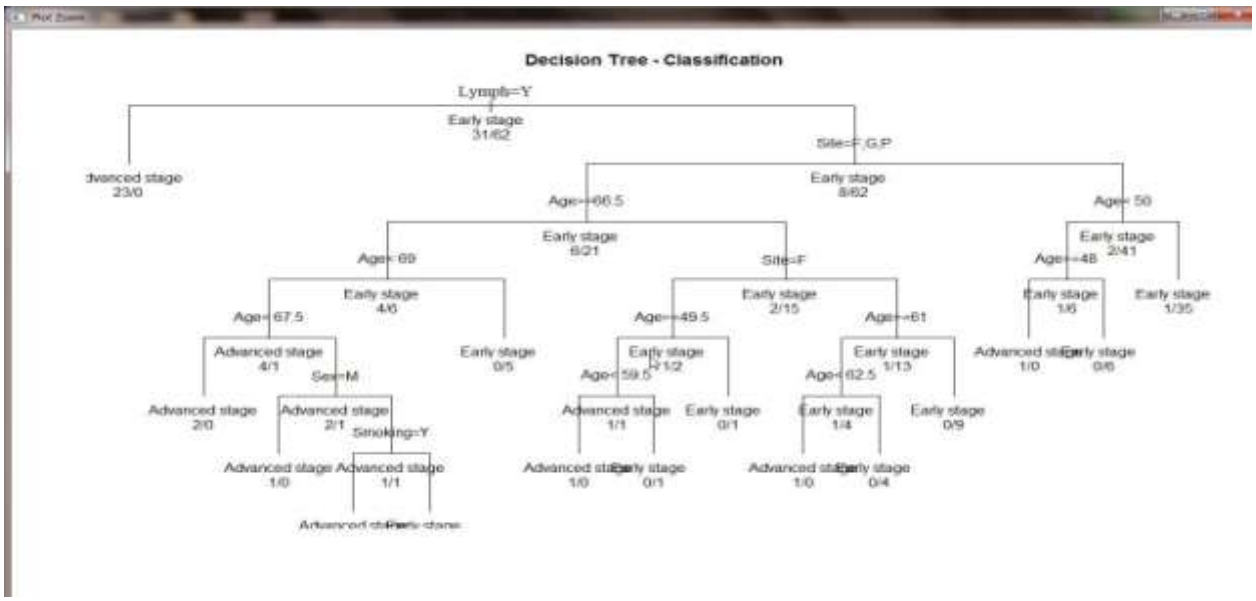

Fig 3: Decision Tree- classification

In the proposed work r part package had used to construct decision tree. In this r part package r part function for classification, this function contain classification attribute and remaining attributes as parameter variables.

The decision tree contain root node as lymph node and this is divided into two parts as early stage and advanced stage, next at the advanced stage node by taking decision two parts will generated. Like that all nodes are created in the decision tree. The constructed decision tree is as follows.

III. RESULT

The decision tree contains the attributes which are in the dataset. This decision tree used to find out the stage of the cancer of new patient early stage or advanced stage. For finding the stage of oral cancer first collects the information which is in the dataset. And then based on the decision tree identify whether that patient oral cancer is in early stage are in advanced stage.

IV. CONCLUSION AND FUTURE WORK

In the proposed system the constructed decision tree used to find out the stage of the oral cancer. In this work 93 patients information had used, in the future work the decision tree will constructed by considering large dataset and also prune the tree for better results.

REFERENCES

[1] K. Kourou, P.Exarchos, Costas Papaloukas and I.Fotiadis, Senior Member, IEEE. " A Bayesian Network-base Approach for Discovering Oral Cancer Candidate Biomackers" .- 978-1-4244- 9270-1/15/2015 IEEE

[2] "Nonparametric Network Design and Analysis of Disease Genes in Oral Cancer Progression" K. Kalantzaki, E. S. Bei, K. P. Exarchos, M. Zervakis, Member, IEEE, M. Garofalakis, Member, IEEE, and D. I. Fotiadis, Senior Member, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 18, NO. 2, MARCH 2014

[3] "Dynamic Bayesian Networks in Modelling Cellular Systems: a Critical Appraisal on Simulated Data" Fulvia Ferrazzi1,2, Paola Sebastiani3, Isaac S Kohane2, Marco F Ramoni2, Riccardo Bellazzi1 Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06) 0-7695-2517-1/06 $20.00 © 2006 IEEE.

[4] "R Programming tutorial site" Available:http://www.tutorialspoint.com/r/ .

[5] "Git hub website" Available:http://github.com/R datasets.

[6] "UCI Machine learning website" Available:http://archive.ics.uci/edu/ml/datasets/

AUTHORS

C Venkata Prasad obtained B.Tech degree in Computer Science Engineering from Yogivemana Univerity, Kadapa, A.P,India. Currently pursuing M.Tech in Computer Science from Jawaharlal Nehru Technological University Anantapur College of Engineering, JNT University, Anantapur, A.P, India, during 2014 to 2016.His research interests include Data Analytics and Data Mining.

Dr A.P.SivaKumar is currently working as an Assistant Professor in Computer Science at JNTUA College of Engineering, JNT University, Anantapur, A.P, India. He received his Ph.D degree in Computer Science And Engineering from JNT University, anantapur,A.P, India. He received the bachelor's degree in 2002 and the Master's degree in 2004, both from JNTU Hyderabad, A.P, India. He has around 10 years of experience as a Lecturer/Research and Development with strong analytical background in the education sector. His research interests are Cross Lingual Information retrieval and Natural Language Processing