Cancer Rate in India in 2015 & 2020- Analysis and Prediction

Reshma P
P G Scholar, Amal Jyothi College of Engineering
Kanjirappally, Kerala.
reshmapraveen58@gmail.com

Ms. Nimmy Francis
Asst. Professor, Amal Jyothi College of Engineering
Kanjirappally, Kerala.
nimmyfrancis@amaljyothi.ac.in

Abstract: **Cancer rate in India is increasing day by day. This could be due to changing life styles which includes food habits, uncontrolled changes in climatic conditions, etc. One of the main fact is the increase in number of alcohol consumption in both males and females. An analysis on cancer rate enables to predict the current situation of people in India , which enables to resolve the current situation. Here we use machine learning algorithm to analyze the cancer rate in India. The steps included in analyzing the dataset are :1)dataset collection 2) dataset preprocessing 3) dataset Classification. Data set analysis is carried out using Weka Tool where the dataset are collected from online repositories of actual cancer patients.**

Key word: **Machine Learning, Dataset, Weka Tool, Classification.**

I. INTRODUCTION

Data mining enables to extract refined data result from a volume of data which gives an accurate value on the data search. Several machine learning algorithms are used for mining data .Machine learning classifies each data into supervised ,unsupervised, semi-supervised and reinforcement learning In this research paper ,the aim is to analyze the cancer rate in 2015 and 2020 and predict the current situation of the people. Here we use Weka tool to analyze the dataset.[1]

II. STEPS INCLUDED IN DATA ANALYZING

Weka (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining. The main steps included in analyzing data includes:

- Dataset Preparation
- Dataset Preprocessing
- Dataset Classification

In Dataset preparation we create an excel sheet with th following attributes: Cancer Number ,Number of Cancer Patients, Rank of Each Type of Cancer, Death Number ,Death Rank and te corresponding class for each year 2015 and 2020.

The next step is to create the corresponding CSV file of the Dataset. CSV(Comma Separated Value)[2] format will lay the data in a table of rows and columns and a comma is used to separate values on a row.

After creating the CSV file , we will convert the corresponding file into arff file(Attribute Relation File Format)[3] , where a header is used that provides metadata about the data types in the columns. The arff file will be processed in the Weka tool for data processing and classification.

The classified data will use the Precision and Recall Values to compare and analyze the cancer rate in 2015 and 2020.

- Precision = TP /(TP + FP)[4]
- Recall = TP/(TP+FN)[5]

III IMPLEMENTATION

Implementation Steps

DOI: 10.5281/zenodo.6183287
ISBN: 978-93-5607-317-3 *@2022 MCA, Amal Jyothi College of Engineering KanjiraKottayamppally,*

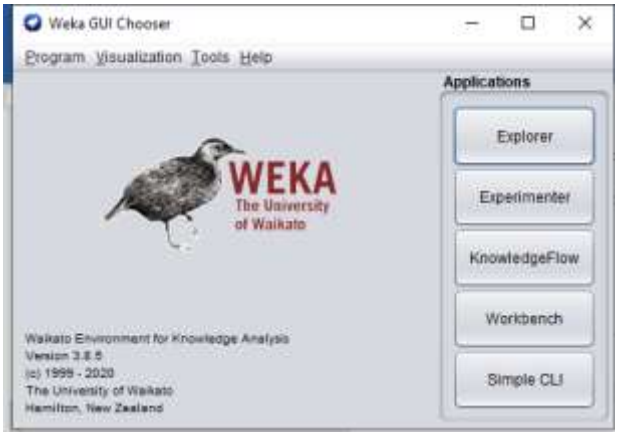

Figure1:Download and Install Weka Tool

	Cancer	Number	Rank	Death Number	DeathRate	Class
2	Breast	155000	1	76000	1	severe
3	Lip	120000	2	74260	3	severe
4	Cervixuteri	122844	3	60708	2	severe
5	Lung	113535	4	56212	4	severe
6	Oesophagus	43180	5	28342	5	severe
7	Stomach	30225	6	23250	6	severe
8	Lukaemia	27419	7	15392	7	severe
9	Ovary	21701	8	10077	9	severe
10	NonHodykinlymp	30828	9	16390	12	severe
11	Liver	24743	10	15793	8	severe
12	Larynx	24687	11	12660	11	severe
13	Prostate	18540	12	9783	14	moderate
14	Colon	66841	13	8236	13	moderate
15	Brain	31460	14	15656	10	moderate
16	Hypopharynx	28489	15	9443	20	moderate
17	Rectum	73260	16	16149	15	moderate
18	Bladder	55096	17	11154	21	moderate
19	Oropharynx	20617	18	11703	17	moderate
20	Thyroid	36432	19	2195	25	moderate
21	Gallbladder	19570	20	14736	16	moderate

Figure 2: Prepare Dataset of cancer patients in 2015 and convert it into a csv file.

	Cancer	Number	Rank	DeathNumber	Deathrate	Class
2	Breast	178361	1	90408	1	severe
3	Lip	135923	2	75230	3	severe
4	Cervixuteri	123907	3	77348	2	severe
5	Lung	72510	4	66279	4	severe
6	Oesophagus	65180	5	58342	5	severe
7	Stomach	60222	6	53253	6	severe
8	Lukaemia	48419	7	35392	7	severe
9	Ovary	45701	8	32077	9	severe
10	NonHodykinlymphoma	35828	9	26390	12	severe
11	Liver	34743	10	33793	8	severe
12	Larynx	34687	11	21860	11	severe
13	Prostate	34540	12	16783	14	moderate
14	Colon	31646	13	19236	13	moderate
15	Brain	31460	14	26656	10	moderate
16	Hypopharynx	28489	15	13443	20	moderate
17	Rectum	28260	16	16349	15	moderate
18	Bladder	21096	17	11154	21	moderate
19	Oropharynx	20617	18	12703	17	moderate
20	Thyroid	20432	19	4895	25	moderate
21	Gallbladder	19570	20	14736	16	moderate

Figure 3: Remaining Dataset of 2015

22	Kidney	22861	21	9897	22 moderate
23	CorpusUteri	14413	22	4285	23 moderate
24	Multiplemyelom	14641	23	1556	18 mild
25	Pancreas	12642	24	9153	19 mild
26	Penis	9677	25	2760	26 mild
27	Hodgkinlymphon	9221	26	1513	28 mild
28	Salivary glands	6350	27	1227	24 mild
29	Nasopharynx	5697	28	2148	27 mild
30	Vagina	3518	29	1723	30 mild
31	Anus	5452	30	776	29 mild
32	Testis	3681	31	852	34 mild
33	Melanomaofskin	3916	32	996	31 mild
34	Vulva	3347	33	984	32 mild
35	Mesothelioma	1709	34	843	33 mild
36	Kaposisarcoma	54	35	23	35 mild

Figure 4: Prepare Dataset of cancer patients in 2020 in an excel sheet and convert it into a csv file.

22	Kidney	16861	21	9897	22 moderate
23	CorpusUteri	10413	22	6385	23 moderate
24	Multiplemyeloma	14641	23	12556	18 mild
25	Pancreas	12642	24	12153	19 mild
26	Penis	10677	25	4780	26 mild
27	Hodgkinlymphoma	9221	26	3513	28 mild
28	Salivary glands	7850	27	5127	24 mild
29	Nasopharynx	5697	28	4148	27 mild
30	Vagina	5518	29	2723	30 mild
31	Anus	5452	30	2776	29 mild
32	Testis	4681	31	1252	34 mild
33	Melanomaofskin	3918	32	2296	31 mild
34	Vulva	3447	33	1894	32 mild
35	Mesothelioma	1709	34	1543	33 mild
36	Kaposisarcoma	66	35	43	35 mild

Figure5:Convert the corresponding csv file to arff file

DOI: 10.5281/zenodo.6183287
ISBN: 978-93-5607-317-3 *@2022 MCA, Amal Jyothi College of Engineering KanjiraKottayamppally,*

Figure6:Convert the corresponding csv file to arff file

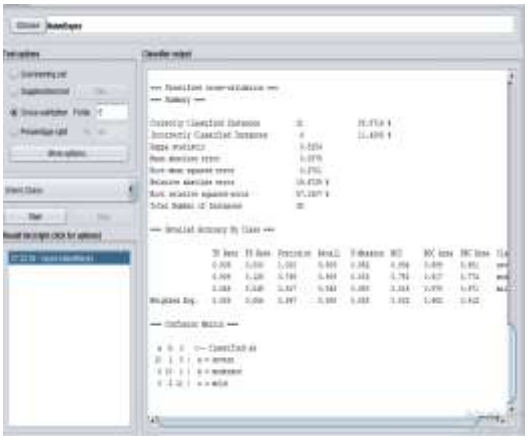

Figure 8 :Apply Naive Bayes to the corresponding dataset of cancer patients in 2015

Figure 7:Run 2015 arff files in weka tool

Figure 9 :Run 2020 arff files in weka tool

DOI: 10.5281/zenodo.6183287
ISBN: 978-93-5607-317-3 *@2022 MCA, Amal Jyothi College of Engineering KanjiraKottayamppally,*

Accuracy Measures

Year	TP Rate	FP Rate	Precision	Recall	Class
2015	0.909	0.000	1.000	0.909	Severe
	0.909	0.125	0.769	0.909	Moderate
	0.846	0.045	0.917	0.846	Mild
2020	0.818	0.000	1.000	0.818	Severe
	1.000	0.167	0.733	0.846	Moderate
	0.846	0.000	1.000	0.917	Mild

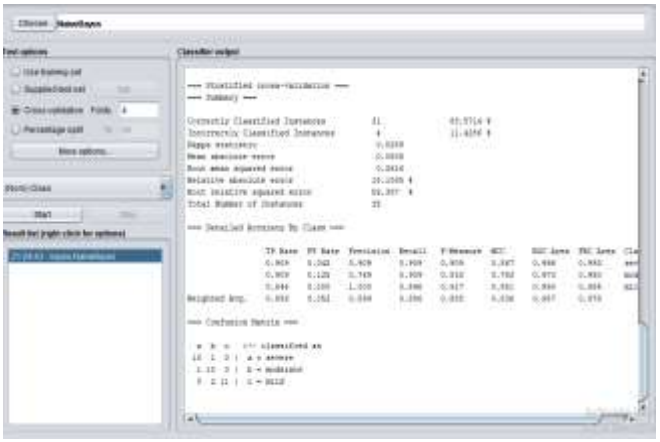

Figure 10:Appy Naïve Bayes to the corresponding dataset of Cancer patients in 2020

IV ANALYSIS

Here , Weka Tool is used for analyzing the cancer rate among the people in the year 2015 and 2020. By using Naive Bayes in estimating the cancer rate enabled us to determine the current situation of cancer graph.

Summary for Cancer Rate in 2015

Correctly Classified Instances	31(88.5714%)
Incorrectly Classified Instances	4(11.4286%)
Kappa Statistic	0.884
Mean Absolute Error	0.0875
Root Mean Squared Error	0.2701
Relative Absolute Error	19.6729%
Root Relative Squared Error	57.1507%
Total Number of Instances	35

Summary For Cancer Rate in 2020

Correctly Classified Instances	31(88.5714%)
Incorrectly Classified Instances	4(11.4286%)
Kappa Statistic	0.8289
Mean Absolute Error	0.0811
Root Mean Squared Error	0.2597
Relative Absolute Error	18.2386%
Root Relative Squared Error	54.9354%
Total Number of Instances	35

DOI: 10.5281/zenodo.6183287
ISBN: 978-93-5607-317-3 *@2022 MCA, Amal Jyothi College of Engineering KanjiraKottayamppally,*

V. CONCLUSION

The Result of analyzing cancer rate in the years 2015 and 2020 shown that there is a sight variation in the cancer rate which may be due to the increased medical facilities and medical care. But it is evident that there is always an increase in number of cancer patients day by day.

The analysis have been carried out with the help of Naïve Bayes Classification which showed the result as :

Weighted Avg in:	TP Rate	FP Rate	Precision	Recall
2015	0.886	0.056	0.897	0.886
2020	0.886	0.052	0.916	0.886

REFERENCES

[1]https://searchenterpriseai.techtarget.com/definition/machine-learning-ML

[2] https://www.howtogeek.com/348960/what-is-a-csv-file-and-how-do-i-open-it/

[3]https://www.cs.waikato.ac.nz/ml/weka/arff.html#:~:text=An%20ARFF%20(Attribute%2DRelation%20File,the%20Weka%20machine%20learning%20software.

[4]https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/

[5]https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/

DOI: 10.5281/zenodo.6183287
ISBN: 978-93-5607-317-3 *@2022 MCA, Amal Jyothi College of Engineering KanjiraKottayamppally,*