EasyChair Preprint

№ 1538

Predicting Cancer Disease Using KNN, J48 and Logistic Regression algorithm.

Ariful Islam Bhuiyan, Tajul Islam, Taufik Akunjee, Rafiqul Islam and Md. Hashikul Islam

September 19, 2019

Predicting Cancer Disease Using KNN, J48 and Logistic Regression algorithm.

Abstract **— In Bangladesh, cervical cancer is the 2nd numerous well-known cancer in women. It is evaluated that every year 11,956 unique possibilities of cervical cancer exist identified in Bangladesh and 6582 women die of the disease. The impartial of this investigation is to examine the motivational features to work with cancer patients, their consequences in job well-being among Portuguese healthcare specialists and to recognize the position of socio demographic and industrial variables on motive and job contentment. Weka medium has been used for mapping the accuracy of the cancer disease dataset including 20 sorts of cancer. Weka tool is utilized to estimate the exactness of the cancer disease dataset including 20 sorts of cancer. In (KNN) k Nearest Neighbour the accuracy is 97.1%, J48 accuracy is 97.8% and Logistic Regression accuracy is 98.2%.**

Keyword— different cancers; classification; KNN; j48; Logistic Regression

I. INTRODUCTION

A. Background

In most maximum people's minds, there is no scarier capability than that of cancer. Cancer is frequently considered an untreatable, unbearably unpleasant illness with no cure. Despite traditional, this impression of cancer may be, it exists wonderful and over-generalized. Cancer is unquestionably a serious and probably life-threatening sickness. For example, it is the principal purpose of death in Americans under the age 85 and the second leading effect of extinction in older Americans. There will exist 1.5 million modern trials of cancer incidence in the United States arriving year, and over 570,000 extinction because of it not involving basal and squamous skin cancers which are not published but could add another two million cases per year (ACS, 2010) [15]. However, it is a misunderstanding to think that completely shapes of cancer are untreatable and deadly. The probity of the object is that there are varied types of cancer, many of which can today be adequately interpreted to reduce, decrease or delay the impression of the condition on patient's lives. While a determination of cancer may still leave patients appearing dependent and out of control, in many cases today there is cause for hope rather than hopelessness [1].

B. Motivation

The data mining procedure has shifted one primary methodology for computing utilization in pharmaceutical informatics and is growing very rapidly in the restorative department due to its progress in the order and prognostication algorithms that support experts in judgment building. Improvement in data mining utilization and its associations are displayed in the areas of message management in healthcare institutions, strength informatics, outpatient care, and monitoring arrangements, assistive technology, large-scale drawing study to knowledge descent and computerized classifications of anonymous groups. Several algorithms connected with data mining. Having significantly helped to guess medical data more clearly, by selecting disordered data from normal data, for maintaining decision-making as generously as visualization and association of hidden complicated connections among diagnostic features of different patient groups [2]. It is quite difficult to get a writing that is correlated with cancer epidemic forecast. The remarkable part of this paper is considered 20 sorts of cancer prediction. Still, now there are several types of cancers which are unknown to us. Awfully modifications may exist notable to opportunity or disclosure to a cancer-causing substances.

C. Contribution

Currently, most of the doctors for recognizing the type of cancers create to make surgical biopsy. But most of them trust that biopsy is an extremely critical task and must be dissuaded as much as feasible. Therefore, introducing a smart method that can help doctors to recognize the type of cancer and escape unnecessary surgical biopsy would be beneficial for both patients and doctors. The main purpose of the paper is to trace how exactly these data mining algorithms can prophesy the possibility of rotation of the disease within the patients based on significant explained

parameters. The statement summits the redaction of various shows that classification algorithms on the dataset. Applying several data mining procedures and find a better result for cancer.

D. Paper Organization

This paper contains six sections. Section I contains Introduction, section II provide related work, section III provide methodology, section IV will provide result and analysis, section V will provide conclusion. The references have been attached in the last segment.

II. RELATED WORK

Scientists are learning and investigate about cancer every day. They discover effective ways to treat and prevent cancer.

Britta Weigelt, Frederick L Baehner and Jorge S Reis-Filho [3] declared that the improvement of gene exposure microarray technologies one decade ago has a passionate influence on the scientific society. The capacity to explore the evolution of thousands of genes in single research has discovered a new way for initial and translational experiments in breast cancer and bestowed the probability of responding questions that previously could not even be asserted. The use of 'quantitative assessment' of genes more than histopathology worldly measurement of tumor properties would offer a more accurate imposition of the incessant tumor biology that recognize clinical effect in breast cancer patients.

Shajahan et al [4] operated on the exercise of data mining procedure to pattern breast cancer data using decision trees to forebode the existence of cancer. Data collected held 699 evidence (patient records) with 10 characteristics and the output class as either effective or maleficent. The input used held sample code number, clump thickness, cell size and shape uniformity, cell growth and other results anatomical examination. The outcome of the supervised learning algorithm used displayed that the random tree algorithm had the absolute accuracy of 100% and the error rate of 0, while CART had the lowest accuracy with a rate of 0.0258%.

Sudhir D. Sawarkar et al [5] in their research they used SVM and ANN on the WBC data. The consequence of SVM and ANN prognosis models were formed relatively more appropriate than the human existing. The 97% high accuracy of these propercy patterns can be used to accept the decision to escape biopsy.

Ritu Chauhan et at [6] concentrate on clustering algorithms like HAC and K-Means in which, HAC is used on K-Means to differentiate the abundance of clusters. The characteristic of the cluster is developed, if HAC is applied to K-Means.

Charles Edeki et al [7] suspects that none of the data mining and statistical understanding algorithms employed to breast cancer dataset outmatch the others in a distinct procedure that it could be exposed the optimal algorithms and none of the algorithms accomplished poorly as to be dispelled from future prophecy pattern in breast cancer survivability tasks.

III. METHODOLOGY

A. Data Collection

In the topic of predicting cancer, we gather the data of cancer suffers. Establishing a dataset, we several data have existed collect. Doctors assist us in establishing an experiment. Absolute data of this paper have existed obtained from cancer victims from Khulna Medical College Hospital. These data are validated because we are collected from actual victims. There are 2000 data and it contains 101 attributes and 1 class attributes. 101 attributes include symptoms and some tests part of cancers and 1 class attribute which is embodied types of cancer. The cancer disease dataset format is .csv format.

We expected those who are curious to predict cancer disorder they may be beneficial by this paper.

B. Data organization

As cancer is a congenital illness, it gives birth to no specific duration or period to strike. In my particular knowledge, we have noticed some cancer victims who existed attacked at a period of 45 or more. One of my friends Purobi Moni, lives in Baghpara, Khulna and his father died 3years ago undergoing cancer. Furthermore, some victims, conserving met them in Khulna Medical College Hospital awfully as Bijoy Manik is 59, Mohona Purobi (39) who exists undergoing in prostate cancer and others like lung and blood cancer etc.

Data verified

All data of this paper have been validated by a Cancer Specialist Dr. Mrinal Kanti Sarkar, Khulna Medical College Hospital, Khulna. He has organized some editions about cancer disease like brain cancer, prostate cancer, breast cancer and lung cancer. In those books, he retains communicated that limit data about cancer disease have existed acquired from the Internet.

Directory of some cancers symptoms and their tests.

1. Lung Cancer

Symptoms:

Weight Loss	Hoarseness	Cough	Chest Pain	Bone Pain	Coughing up Blood	Shortness of Breath	Headache
p	n	n	p	n	p	p	n

Tests:

CT Scan	X-Ray	Sputum Cytology	Biopsy
p	p	n	p

Result: Lung Cancer

2. Kidney Cancer

Symptoms:

Fatigue	Swelling in Leg	Fever	Weight Loss	Blood Urine	Loss of Appetite	Side Pain	Anemia
p	n	p	p	p	p	p	p

Tests:

Ultrasound	CT Scan	MRI	Blood Test	Urine Test	IVP	Renal Arteriogram	Biopsy
p	P	p	p	p	n	p	p

Result: Kidney Cancer

3. Liver Cancer

Symptoms:

Abdominal Pain	Fatigue	Weight Loss	Vomiting	Jaundice	Loss of Appetite	Weakness	Swelling in the Abdomen	Chalky Stool
p	p	p	p	n	p	p	n	p

Tests:

Ultrasound	CT Scan	MRI	Laparoscopy	Blood Test	Biopsy
n	p	p	n	p	p

Result: Liver Cancer

4. Prostate Cancer

Symptoms:

Fatigue	Frequently Urine	Weight Loss	Blood Urine	Bone Pain
p	p	p	n	p

Tests:

Ultrasound	DRE	PSA	Biopsy
p	n	p	p

Result: Prostate Cancer

5. No Cancer

Symptoms:

Abdominal Pain	Vomiting	Loss of Appetite	Fatigue	Weakness	Indigestion	D. Swallowing	Diarrhea
p	n	p	n	p	p	p	n

Tests:

CT Scan	PET	Ultrasound	Barium Swallow	MRI	Biopsy
p	n	n	p	n	n

Result: No Cancer

We are an experiment in nearly 20 cancers, which exists we understood to all. Here, some cancers indicate their symptoms and tests. If the biopsy outcome is positive the sufferer is sure that he is a cancer victim. But the biopsy outcome is negative then he must be confident that he is not a cancer victim.

C. Data Statement

Cancer is when unique separations distance in an unruly direction. Numerous cancers may have finally pushed to the limit out into more tissues. Here 20 sorts of cancer work in exemplified which we adequately as the known. Naturally, we survey the symptoms of the cancer than we survey the tests which are a must for cancer revelation. The symptoms are Fever, Weight Loss, Abdominal Pain, Jaundice, Rash. These warnings frequently suggest cancer but it can be a justification for another syndrome. In the first spotlight, people constantly do not anticipate that they have cancer. It withstands a long duration to investigate that the victims are sufferings from cancer.

Brain cancer remembers some symptoms which are extremely familiar to another illness. CT scan and MRI the crucial portion which assists a doctor to observe that the victims have some irregularities in their brain. After the biopsy exists the easy platform cancer illness. For identifying blood cancer blood test influences a crucial function. Bone Marrow Aspiration furthermore assists to recognize blood cancer but the biopsy can assist too obvious about it.

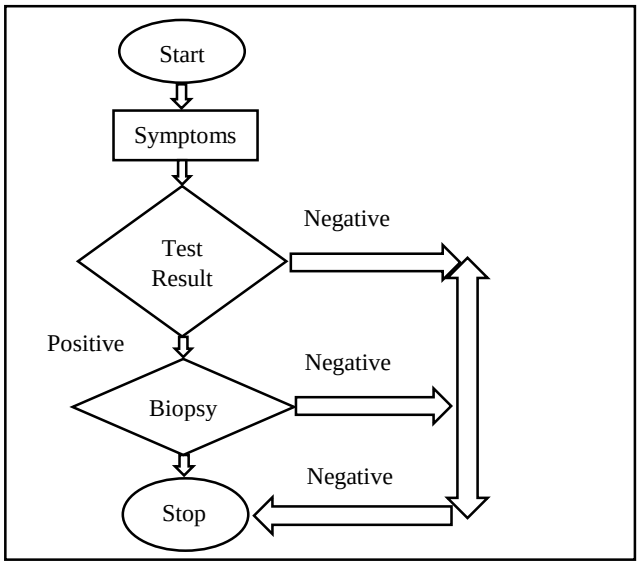

Fig 1: Dataflow Diagram of Predicting Cancer

In figure 1 the strategy is to get turned on with the beginning. Symptoms are suggested as well as the training component. If any symptoms are true accordingly it indicated the outcome. If the test portion implies true then the biopsy is the performed, otherwise the procedure is ended. The test portion and the biopsy both display negative then the procedure ended.

D. Data Mining Process:

The data mining strategy helps a huge abundance of datasets. It also anticipates where the dataset will exist. It is expended for prediction, clustering, classification. There are two fundamental category training and testing.

- Training is the portion where the dataset is trained by input with ordinary output.
- Testing is a portion where the trial of the declaration is expressed in fig 2.

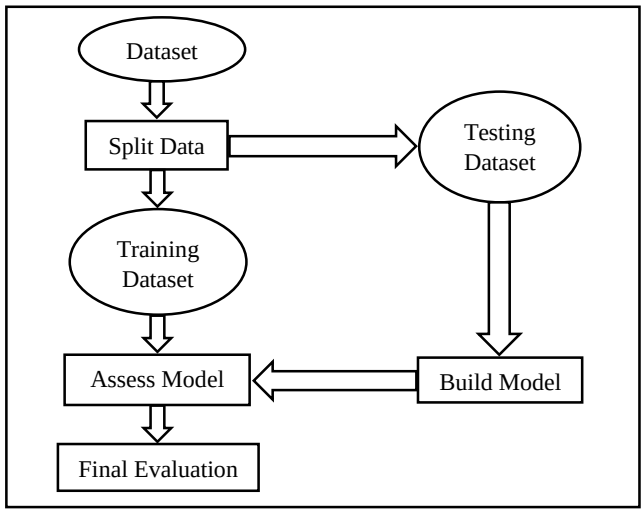

Fig 2: Data Mining Diagram

IV. RESULT AND ANALYSIS

A. Classifiers

Classification is the technic of predicting the class of certain data points. We take some training set of data. Weka makes it simple to find the classifiers. Weka is an algorithm for data mining tasks. It contains some tools like clustering, classification, regression algorithm [8]. We use windows 10 operating system and weka 3.8 version.

(KNN) K Nearest Neighbour used for classification and regression problems. However its additional widely borrowed in succession technics. We use it because its relief to deduce output, less calculation time and its predictive power [9].

The pseudo code of KNN algorithm:

1. Input: P, a set query points and S, a set of reference point;
2. Output: A list of M reference points for each investigation point;
3. For each investigation aspect p ∈ P do
4. Compute distances between p and all s ∈ S;
5. Sort the computed distances;
6. Select k-nearest reference points corresponding to k smallest distances; [14]

J48 algorithm we used this paper because this algorithm improving the detection accuracy and the accomplishment of the technic. This algorithm also show the better accuracy and more effective performance [10].

The pseudo Code of J48 algorithm:

1. Create a root node L;
2. IF (P belongs to same category C)
   {leaf node = L;
   Mark L as class C;
   Return L;}
3. For i =1 to n
   {Calculate Information_gain (Ig);}
4. ta = testing attribute;
5. L.ta = attribute having highest information_gain;
6. if (L.ta == continuous)
   {find threshold;}
7. For (Each P in splitting of P)
8. if (P is empty)
   {child of L is a leaf node;}
   else
   {child of L = dtree P}
9. calculate classification error rate of node L;
10. return L; [12]

Regression is a bunch of process for relationship among the statistical variables. This algorithm is widely manipulated to prediction and forecasting. It is likewise used for understand the independent and the dependent variables. Many technics have been carrying out by this regression algorithm [11].

The pseudo Code of Logistic Regression:

1. Start at the root node.
2. For each ordered variable V,
3. Perform a chi-squared sample of autonomy of each V, variable versus L on the data in the node and compute its significance probability.
4. Choose the variable V × associated with the V, that has the smallest significance probability.
5. Find the split set {V × ∈ M ×} that minimize the sum of Gini indexes and use it to slash the node into two child nodes.
6. If a stopping benchmark is reached, exit.
7. Otherwise, apply notches 2-5 to each minor node.
8. Cut back the fence with the CART procedure. [13]

We made an effort to expect cancer syndrome implementing three categories of algorithm and locate reasonable accuracy. We use the windows 10 and Weka 3.8 version. So our major responsibility is to discover the accuracy of the three classifier algorithm. We analyze 20 types of cancers accuracy, error rate, recall, specifity, precision and f-force. We using 10 folds cross validation and three classification algorithm, weka gives us a confusion matrix. Confusion matrix gives us VO, VU, IO, IU values.

$$\text{Accuracy}=\frac{VO+VU}{O+U} \quad \text{(i)}$$

$$\text{Error rate}=\frac{IO+IU}{O+U} \quad \text{(ii)}$$

$$\text{Recall}=\frac{VO}{O} \quad \text{(iii)}$$

$$\text{Specificity}=\frac{VU}{U} \quad \text{(iv)}$$

$$\text{Precision}=\frac{VO}{VO+IO} \quad \text{(v)}$$

$$\text{F-Score}=\frac{2\times precision\times recall}{precision+recall} \quad \text{(vii)}$$

Here, VO = Valid Optimistic
VU = Valid Unfavorable
IO = Inaccurate Optimistic
IU = Inaccurate Unfavorable
O = Optimistic
U = Unfavorable

Table I. KNN USING 10-FOLDS CROSS-VALIDATION

Class (Cancer)	*Accuracy*	*Error Rate*	*Recall*	*Specificity*	*Precision*	*F-Score*
Lung	97.2%	2.2 %	98.4 %	97.2%	93.8%	96.1 %
Fallopian Tube	97.1%	2.8 %	99.7 %	98.1%	100%	100 %
Gallbladder	98.7%	2.1 %	98.6 %	92.6%	100%	100 %
Head and Head	97.1%	2.9 %	100 %	98.1%	95.4%	97.6 %
Hypopharynx	98.3%	2.7 %	98.9 %	98.3%	99.3%	100 %
Kaposi Sarcoma	96.1%	4.1 %	99.8 %	98.1%	98.9%	99.4 %
Kidney	97%	3.3 %	99.4 %	97.9%	95.4%	97.7 %
Leukemia	94.2%	5.6 %	99.2 %	96.6%	97.2%	98.6 %
Liver	99.7%	0.2 %	99.5 %	99.1%	95%	96.9 %
Esophageal	99.2%	0.6 %	98.3 %	99.4%	92.9%	96.3 %

Lymphoma-Hodgkin	99.6%	0.3 %	100 %	99.7%	96%	98%
Mesothelioma	97.3%	2.8 %	99.1 %	98.2%	91.2%	94.9 %
Chordoma	98.2%	2.6 %	98.7 %	98.1%	94.2%	97%
Prostate	97.8%	1.9 %	92.1 %	98.2%	88%	93.1 %
Stomach	99.1%	0.7 %	98.8 %	99.1%	98.8%	95.3 %
Blood	99.7%	0.2 %	93.1 %	97.8%	81%	86.6 %
Brain	98.8%	1.1 %	95.4 %	97.7%	86.6%	90.7 %
Colorectal	97.7%	2.7 %	97.8 %	96.1%	92.9%	95.3 %
Pancreatic	97.8%	2%	99%	98.6%	91.2%	95%
No Cancer	98.4%	2.2 %	72.8 %	98.1%	93.8%	82%

Subsequently plopping the dataset in Weka, we subsist to exploit 10-fold cross validation, accordingly, we preserve to accomplish a consequence. In the consequence, there is placed confusion matrix. From the matrix, we have to obtain VO, VU, IO, IU, O, U significances. We put those significances in the proper equation and have to obtain the desired consequence for Table I. This Table I implies the significances of accuracy, error rate, recall, specificity, precision and F-score using KNN classifier algorithm.

Table II. J48 USING 10-FOLDS CROSS-VALIDATION

Class (Cancer)	*Accuracy*	*Error Rate*	*Recall*	*Specificity*	*Precision*	*F-Score*
Lung	97.8%	2.2%	96.8 %	97.2%	98.4%	97.1%
Fallopian Tube	97.2%	2.8%	93.8 %	98.1%	95.6% %	96.7% %
Gallbladder	92.3%	7.2% %	99.6 %	92.6%	98.5% %	98.6%
Head and Head	97.4%	2.9%	98.4 %	98.1%	94.4%	97.5%
Hypopharynx	97.9%	2.7%	98.9 %	98.3%	97.3%	98.7%
Kaposi Sarcoma	97.7%	2.1%	98.8 %	98.1%	97.8%	98.4%
Kidney	97.8%	3.3%	98.4 %	97.9%	98.8%	96.7%
Leukemia	97.2%	1.6%	99.2 %	96.6%	97.7%	97.8%
Liver	98.7%	1.2%	99.5 %	99.1%	95.8%	97.8%
Esophageal	100%	0%	100 %	100%	97.7%	97.8%
Lymphoma-Hodgkin	98.6%	1.3%	98.3 %	97.7%	96.8%	98.7%
Mesothelioma	96.3%	3.8%	94.1 %	95.2%	95.5%	97.9%
Chordoma	94.2%	5.6%	95.7 %	96.1%	96.6%	97.8%
Prostate	99.8%	0.1%	98.1 %	99.2%	98.7%	97.7%
Stomach	99.4%	0.7%	97.8 %	98.9%	98.9%	96.7%
Blood	97.7%	1.2%	99.1 %	97.8%	98.9%	86.8%
Brain	97.8%	2.1%	98.4 %	95.7%	98.6%	98.9%
Colorectal	97.9%	1.7%	96.8 %	98.8%	96.9%	98.9%
Pancreatic	96.8%	3.7%	99%	98.3%	96.2%	97.8%
No Cancer	92.3%	7.2%	99.7 %	98.6%	98.6%	98.9%

Subsequently plopping the dataset in Weka, we subsist to exploit 10-fold cross validation, accordingly, we preserve to accomplish a consequence. In the consequence, there is placed confusion matrix. From the matrix, we have to obtain VO, VU, IO, IU, O, U significances. We put those significances in the proper equation and have to obtain the desired consequence for Table II. This Table II implies the significances of accuracy, error rate, recall, specificity, precision and F-score using J48 classifier algorithm.

Table III. LOGISTIC REGRESSION USING 10-FOLDS CROSS-VALIDATION

Class (Cancer)	*Accuracy*	*Error Rate*	*Recall*	*Specificity*	*Precision*	*F-Score*
Lung	98.2%	1.7 %	98.9%	99.2%	98.8%	97.1 %
Fallopian Tube	99.1%	0.8 %	93.4%	93.1%	98%	98.8 %
Gallbladder	98.9%	1.1 %	93.6%	98.6%	97.9%	99.3 %
Head and Head	98.9%	1.1 %	90%	96.1%	98.4%	96.6 %
Hypopharynx	99.6%	0.2 %	94.9%	98.8%	98.4%	97.5 %
Kaposi Sarcoma	98.8%	1.1 %	96.8%	96.7%	99.5%	97.4 %
Kidney	98.9%	1.3 %	99.7%	97.8%	99.8%	96.5 %
Leukemia	98.9%	1.6 %	99.5%	98.9%	98.7%	95.6 %
Liver	98.8%	1.2 %	95.5%	98.7%	95.9%	98.9 %
Esophageal	98.9%	1.6 %	96.3%	98.6%	96.8%	96.6 %
Lymphoma-Hodgkin	97.7%	2.3 %	95.3% %	98.3%	96.9%	98.6 %
Mesothelioma	97.4%	0.6 %	96.3%	99.5%	98.6%	99.6 %
Chordoma	98.5%	0.4 %	99.9%	97.6%	96.9%	97.8 %
Prostate	95.5%	4.9 %	97.3%	99.2%	98.7%	99.8 %
Stomach	96.6%	3.7 %	94.8%	99.1%	98.9%	98.3 %
Blood	99.7%	0.2 %	95.1%	97.4%	98.9%	95.6 %
Brain	97.8%	2.1 %	96.4%	97.3%	96.2%	96.7 %
Colorectal	98.9%	1.1 %	97.7%	99.1%	96.9%	97.3 %
Pancreatic	98.8%	2.1 %	99.9%	97.6%	97.2%	96.4 %
No Cancer	99.5%	0.5 %	99.4%	98.6%	98.7%	96.8 %

Subsequently plopping the dataset in Weka, we subsist to exploit 10-fold cross validation, accordingly, we preserve to

accomplish a consequence. In the consequence, there is placed confusion matrix. From the matrix, we have to obtain VO, VU, IO, IU, O, U significances. We put those significances in the proper equation and have to obtain the desired consequence for Table III. This Table III implies the significances of accuracy, error rate, recall, specificity, precision and F-score using Regression classifier algorithm.

Table IV. FINAL RESULT

Class	*Accuracy*	*Error Rate*	*Recall*	*Specificity*	*Precision*	*F-Score*
KNN	97.1%	2.9 %	93.7 %	97%	93.8%	93.4 %
J48	97.8%	2.1 %	98%	97.9%	98.1%	98%
Regression	98.2%	1.7 %	99.7 %	98%	99.6%	99.7 %

From Table IV shows the final result analysis. In this table, the average rate of accuracy, error rate, recall, specificity, precision, F-score have been shown. Here the average rate of accuracy of KNN is 97.1%. The average rate of accuracy J48 is 97.8% and for Regression is 98.2%. The error rate of KNN is 2.9%. The error rate of J48 is 2.1% and for Regression it is 1.7%. The recall of KNN is 93.7%. The recall of J48 is 98% and for Regression it is 99.7%. The specificity of KNN is 97%. The specificity of J48 is 97.9% and for Regression it is 98%. The precision of KNN is 93.8%. The precision of J48 is 98.1% and for Regression it is 99.6%. The F-score of KNN is 93.4%. The F-score of J48 is 98% and for Regression it is 99.7%. From this, it is apparent that Regression provides an accurate explanation than the distinct two classifier algorithm. So, here Regression is reasonably than different two classification algorithms.

KNN < J48 < Logistic Regression

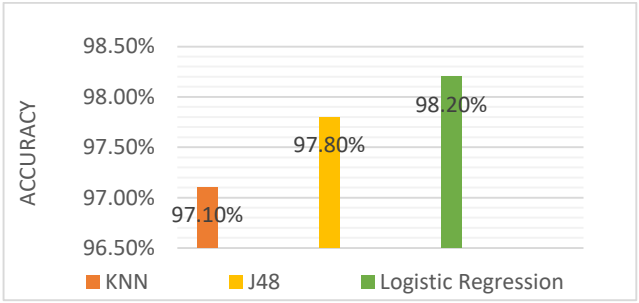

Fig 3: Graph of Accuracy

In fig. 3 KNN is represented by orange, J48 is represented by yellow and Logistic Regression is represented by green color in their accuracy level.

V. CONCLUSION

This paper mainly made on medical dataset that is count the cancer disease. From the analysis we can say that main reason for the cancer is gulping liquor and smoking and many other things. Here we discuss about three algorithm. Its identity the Confusion matrix and this matrix gives us the result of grouping algorithm. Here we discuss about twenty varieties of cancer that are lung cancer, prostate cancer, head and neck cancer, blood cancer, brain cancer etc. that variety of cancer assists us to collect the accuracy, error rate, recall, specificity, precision and f-force. Tables and other data helps to create the clear perception. It also helps a comparison between the three algorithm.

In the future, we will attempt to expand more innovation to a massive refinement, also attempt to expand new prototypes prediction and survivability.

REFERENCES

[1] R.V. Hoch, D.A. Thompson, R.J. Baker, and R.J. Weigel, "GATA-3 is expressed in association with estrogen receptor in breast cancer", Int. J. Cancer (Pred. Oncol.), Vol. 84, pp.122-128, 1999.

[2] MAQC Consortium, "The MicroArray Quality (MAQC)-II study of common practices for the development and validation of microarray-based predictive models", Nat Biotechnol., Vol 28, No. 8,pp. 827-838, 2010.

[3] Britta Weigelt, Frederick L Baehner and Jorge S Reis-Filho "The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade," Journal of Pathology, J Pathol 2010; 220: 263-280, Published online 19 November 2009 in Wiley InterScience (www.interscience.wiley.com), DOI: 10.1002/path.2648.

[4] Shajahan, S.S., Shanthi, S. & ManoChitra, V. (2013). "Application of data mining techniques to model breast cancer data." International Journal of Emerging Technology and Advanced, 3(11), 3622-369.

[5] Sudhir D., Ghatol Ashok A., Pande amol P(2006). "Neural Network aided Breast Cancer Detection and Diagnosis." 7th WSEAS International Conference on Neural Networks, 2006.

[6] Ritu Chauhan "Data clustering method for Discovering clusters in spatial cancer databases" International Journal of Computer Applications (0975-8887) Volume 10-No.6, November 2010.

[7] Charles Edeki "Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability" Mediterranean Journal of Social Sciences Vol 3 (14) November 2012, ISSN: 2039-9340.

[8] Towardsdatascience "Machine Learning" Available: towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623 [Accessed: 3.9.2019]

[9] Wikipedia, the free encyclopedia "K-Nearest Neighbours algorithm" Available: https://en.m.wikipedia.org/wiki/K-nearest_neighbours_algorithm [Accessed: 3.9.2019]

[10] Wikipedia, the free encyclopedia "C4.5 algorithm" Available: https://en.m.wikipedia.org/wiki/C4.5_algorithm [Accessed: 3.9.2019]

[11] Analyticsindiamag "Top 6 Regression algorithms used data mining applications" Available: www.analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/ [Accessed: 3.9.2019]

[12] Google "Pseudo code of J48 algorithm" Available: https://images.app.goo.gl/3egYcCuQ7zHxUTt88 [Accessed: 3.9.2019]

[13] Google "Pseudo code of Logistic Regression" Available: https://images.app.goo.gl/xSbFk8Gk9zzgt39W9 [Accessed: 3.9.2019]

[14] Google "Pseudo code of K-nearest neighbours" Available: https://images.app.goo.gl/5pqeLBGbJ7YHbcy39 [Accessed: 3.9.2019]

[15] Silva, T. Oliveria, J. Neves, and P. Novais, "Treating Colon Cancer Survivability Prediction as a Classification Problem," 2016.