International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Cervical Cancer Prediction using Naïve Bayes Classification

Mansvi Girdonia, Ruchi Garg, Pooja Jeyabashkharan, M.S. Minu

Abstract: ***Now a Days cancer is a common diseases among people. And is generally having an obstruction in getting the cancer cured. If the cancer is not predicted in an early stage it can lead to a treacherous death of a person. Our paper describes a classification method known as Naïve Bayes classifier for the prediction of cervical cancer which is a kind of cancer occurring in females. Cervix is a lower portion of uterus in the human female reproductive system. Our algorithm is based on the concept of conditional probability. Naïve Bayes classification algorithm is based on the assumption of independent amount predictors. Bayes classification assumes that the presence of an attribute in a class is unrelated to any other attribute of that class and also if both are related they are independent of the existence of each other. Since all the attributes are contributing independently to the probability. Another feature of this algorithm is that it can be applied to both binary and multi. Further we require less training datasets unlike other algorithms of machine learning. In this algorithm we consider a hypothesis. And we get the probability of hypothesis before we get its evidence and the same hypothesis after getting its evidence. We are using gaussian naïve Bayes algorithm for our prediction algorithm. For this probability calculation we need to consider a large number of attributes such as age of women, Number of intercourse with multiple men , first mating, Number of times of conceiving, Smoking habits, frequency and duration of Smoking(no. of years), Dx, Hinselmann, Schiller, Citology, Biopsy etc. And using these attributes of dataset we calculate the probabilities of occurrence and then finally we use those probabilities for our final predictions, Here we are taking common ratio of training and testing data sets which is 70% and 30%.***

Keywords: Cervical cancer, Naïve Bayes classifier, Cervix.

Manuscript published on 30 April 2019.
* Correspondence Author (s)
M.S. Minu*, CSE Dept. , SRM Institute of Science and technology, Chennai, Tamil Nadu.
Mansvi Girdonia, CSE Dept. , SRM Institute of Science and technology, Chennai, Tamil Nadu.
Ruchi Garg, CSE Dept. , SRM Institute of Science and technology, Chennai, Tamil Nadu.
Pooja Jeyabashkharan, CSE Dept. , SRM Institute of Science and technology, Chennai, Tamil Nadu.

I. INTRODUCTION

Cervical cancer is the most ordinarily occurring cancer predicted in females .The global statistics of occurrence of cancer are as follows 17 million cases of different types of cancer were estimated in 2018 survey. In which 9.8 million of people got died due to cancer world wide. These cancers were mainly occurred due to tobacco intake and smoking. Cervical cancer is commonly occurring cancer in women now a days the cervical cancer statistics are as follows cervical cancer is causing approximately 280,000 deaths every year. And nearly 1 million women a suffering and living with cervical cancer. Nearly 85% of women suffering from cervical cancer are from under developing countries. Here are some countries which have the highest rate of cervical cancer such as Swaziland(75), Malawi(72), Zambia(65) ,Zimbabwe(62) etc. Cervical cancer is This is only happening because the cervical cancer is not predicted on time or is delayed due to some reason. It is proved that if the cervical cancer is predicted in the early stage it can be treated and cured. For the treatment and prediction of cervical cancer a test is recommended to have for women in most of the developed countries known as PAP smear test. Nearly 70% of cervical cancer is caused due to smoking, multiple intercourse partners and HPV (Human Papllio-mavirus),And also a vaccines against HPV is available and is recommended for prevention of cervical cancer. We are having many methods available in machine learning for classification and prediction such as Random forest search, Deep neural networks, Artificial neural networks, K nearest neighbor, Decision tree, linear regression, logistic regression, Naïve Bayes classifier etc. In our paper we are focusing of the naïve Bayes classification algorithm based on the concept of conditional probability. The conditional probability can be explained as the probability of an event that has some relationship to one or more events. The naïve Bayes considers the probability of the hypothesis before the evidence and the probability after getting the evidence and also the probabilities of hypothesis and evidence individually. The Bayes classifier assume that the presence of a feature in a given class is not related to other feature present in that class and if it even related then their existence is independent of each other. The reason why the naïve Bayes classifier is used in this paper is that it is easy to build the algorithm and also very large datasets can be used while working with it. The approach of this algorithm is based on a simple statistical probability concept and hence it is easy to implement the algorithm. There are three types of Naïve Bayes.

They are 1.Gaussian, 2.Multinomial 3.Bernoulli.The Gaussian method is used to classify and it assumes the feature to follow the normal distribution. The multinomial is used for the discrete counts. And the Bernoulli is useful if the feature is having binary vectors such as 0 and 1. We are using gaussian in our paper.

II. LITERATURE SURVEY

Cervical cancer is thereby more chronic in females. Cervical cancer is not often known at the initial stage. Cervical cancer can be due to the increase of cancer cells in the cervix region of the body which then spreads at faster rate. The cancer can globally be lessened by a comprehensive perspective that includes avoidance, early detection, fruitful screening and treatment program. But for that prediction for cervical cancer is very much necessary which is our project. Naïve Bayes classifier can predict if cancer occur in female in the cervix region. Also the Voting technique helps to maintain the level of accuracy in the system.

In this paper [1] ,labelled cumulative Risk Score(CRS), for individual member were determined by an algorithm that counted for time and delay. The constraints set to assign scores, which were programmed using C++.

In this paper[2], soft computing is the branch of AI. Combination of statistical, Probabilistic and optimization technique and along with this we use fuzzy edge detection Methodology. The Neural Network technologies is used in MLP, PNN, RBF and LVQ are applied on MCPs dataset to classify the datasets into: normal and abnormal classes and RBF is higher accuracy.

In paper [3], we have used data structure algorithm which includes decision tree and also used some genetic factors like: Bayesian network is used for the micro satellite dataset. The Bayesian network reveals that IFNG, HPV and ILIOG indirectly determine the diagnosis.

In this paper[4], we have used data mining ,naïve Bayes. Comparison of results of data pop smear test results based on three factors: Literature study and data collection, Determining all characteristics, Data Evaluation and in data evaluation support vector machine gives the poor results.

In this paper[5], by using software cadence 6.0 ,CMOS, Electrical Bioimpedance and the high stability voltage controlled current source (VCCS) for cervical cancer Detection (CCD) applying Electrical Bio-Impedance Spectroscopy(EBS) is made using CFC in EBS applications to detect cervical cancer.

In this paper[6], machine learning ,supervised learning models are of support-vector machines (SVM) along with associated learning algorithms estimate data used for classification and regression analysis .SVM was used to diagnose cervical cancer and consumes high computational time when large dataset is used. In this paper[7],an unconventional method for early cervical cancer detection related to piezoelectric immunosensor was defined. The method defined in this paper involves short examining time for p16INK4a detection. The initial screening of cervical lesion was done using this process. In this paper[8]a process was used to directly Categorize cervical cells – without initial segmentation .It was related to deep features, using convolutional neural networks (ConvNets) a cervical screening which is assisted by automation through liquid-based cytology (LBC) or pap smear is the cancer detection tool used.

In this paper[9],The algorithm used is machine learning, support-vector machines (SVM)approach is to diagnose the cervical cancer. The following methods have been used to diagnose malignant cancer samples ,support vector machine-recursive feature elimination(RFE) and support vector machine-principal component analysis (SVM-PCA).

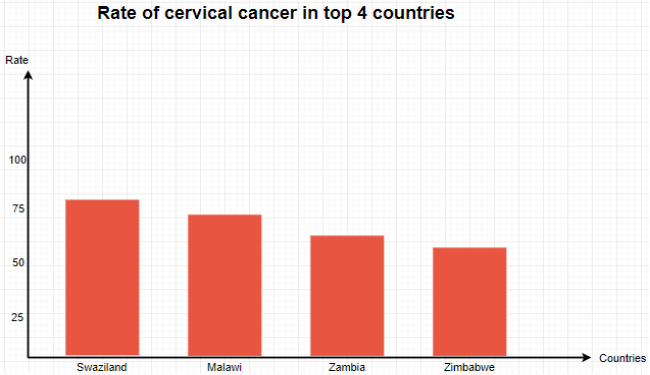

Fig 2.1 Rate of cervical cancer in different countries

III. PROPOSED SYSTEM

In the system we proposed, we are using the naïve Bayes classification to predict the cervical cancer. Naïve Bayes classifiers is the efficient method of prediction in medical applications. In this model, Our procedure is as follows firstly, Our dataset is handled we load our csv file. For loading the csv file we used the loader function and to read the file used the reader function. Then the dataset is split into two subsets- (1)training dataset (2)testing dataset. Training dataset is used for making predictions. Testing dataset is required to evaluate the precision of the model. Then , the summary of training dataset is used while making prediction. The summary consists of the mean standard deviation of each attribute by class value. So, we have two class values and 36 attributes which will makes 72 attributes summaries. Therefore we can split the arrangement into subtasks which are separating calculating the means calculation the standard deviations, summarizing the dataset and summarize the attributes by class. Mean is the average of the data collected. We can use mean as the centre of the gaussian distribution while we calculate the probabilities. Now we determine the standard deviation for attributes of all the values of the class. The square root of variance is called standard deviation. Variance is determined as the mean of different attribute value from mean. The function collects the values for every attribute across the data instances into their lists. Therefore, we are able to calculate mean and also standard deviation for all attributes. Now, next comes summarizing the attributes. We then pull it together by firstly extracting training datasets into the instances that are grouped by class itself. Then we calculate the summaries that we got from training data. Then we make predictions that involves determining probability whether the data instances belong to every class and choosing the class which has the largest probability to be prediction. Further the following procedure can be categorized in 4 subtasks as calculating the gaussian probability density function, calculating the class probability, making a prediction and then estimating the accuracy.

To calculate the Gaussian probability function and to determine probability of the attribute values, mean, standard deviation of the attribute which was estimated from the training data. In the next step we calculate the probability of class. We then calculate the probability of the attribute which belongs to class. Then join the probabilities of attributes for the data instances and get the probability of the whole data instance which belongs to the class. Now we can finally calculate how accurate the model is and make predictions for every data instances of each data for that we used a method say get prediction. This method is used to calculate the predictions based upon the testing dataset. Classification accuracy is the ratios between 0 to 100%. We are using UCI dataset from a hospital consists of habits, health history and demographic details of the patients . The profit of using this technique is that we need very less amount of training data set. We can get good results for large datasets.

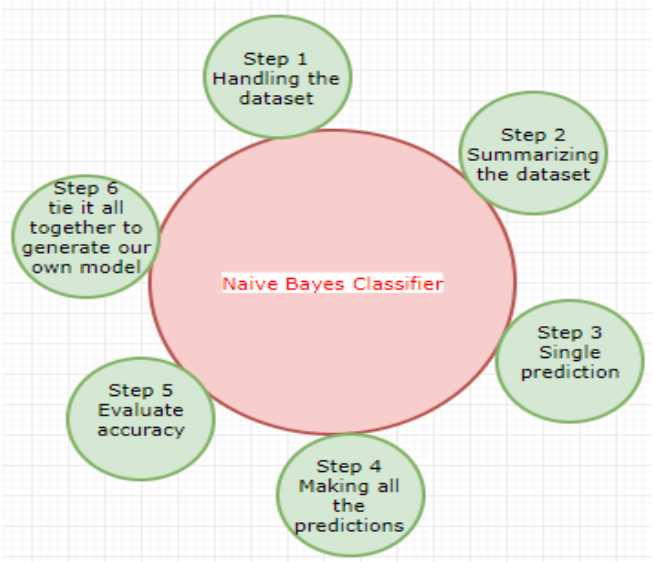

Fig 3.1 Steps of Naïve Bayes Model

IV. FORMULAE

1.Calculating the mean:

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x$$

2.Calculating the variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$$

3.Calculating the standard deviation:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2}$$

4.Calculating gaussian probability density:

$$f(t) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

V. METHODOLOGY

In our algorithm, there are following steps for the implementation of the algorithm. Firstly, we need to import the csv file, math file and random.

1.Data processing:

In this step, we are loading our data into csv file and splitting it into two sets that are training dataset and test dataset.

2.Summarizing the data:

We are summarizing the properties into training set and calculating the mean and standard deviation of the given attributes for calculation the probabilities and for making the predictions.

3. Single prediction:

Using the summaries of dataset we make our first prediction.

4.Generating all predictions:

Using the first prediction follows all the predictions.

5.Evaluation of the accuracy:

After making all the predictions our accuracy is obtained.

VI. ARCHITECTURE DIAGRAM

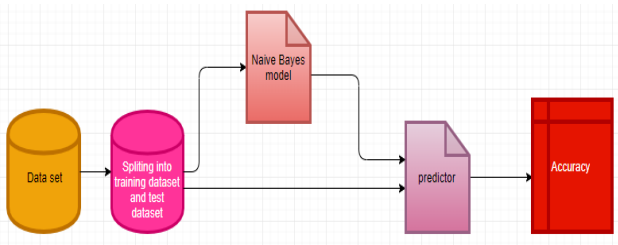

Fig 4.1 Architecture Diagram of Proposed System.

VII. DRAWBACKS

One of the main drawbacks of Naive Bayes classifier is that it proves a very high speculation on the graph of scattered data, that is if any of the two characteristics are not dependent given the outcome class. Therefore, loss of precision. Another drawback is lack of data. For any worth of a characteristic, you needed to analyze a very probable worth by a continuous approach. This can result in probabilities ranges 0-1,which further leads to integer distortion and worsening the outcome. Continuous features turns out to be the third problem. Binning procedure is regularly followed to make them distinct, if it is not taken care,there will be a heavy loss of information.

VIII. CONCLUSION

Cervical cancer is thereby more chronic in females. Cervical cancer is not often known at the initial stage. Cervical cancer can be due to the increase of cancer cells in the cervix region of the body which then spreads at faster rate. The cancer can globally be lessened by a comprehensive perspective that includes avoidance, early detection, fruitful screening and treatment program. But for that prediction for cervical cancer is very much necessary which is our project. Naïve Bayes classifier can predict if cancer occur in female in the cervix region. Also the Voting technique helps to maintain the level of accuracy in the system.

IX. RESULT ANALYSIS

In our model none of the attributes or instances are not dependent on each other. All are independent of each other. There is no loss of precision. We worked with the large amount of data. In which we took the usual split ratio. But we can even work with the different split ratio which gives more better result. In the accuracy and precision of the model. In this model we don't even need a heavy training set data.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Since, It is the mathematical model the understanding is easy in comparision to other model.

Table 9.1 Comparision of our model with other models

INSTANCES	OUR MODEL	OTHER MODELS(approx..)
EFFICIENCY	89%	70%
ACCURACY	87%	85%
VARIATION OF SPLIT RATIO	Possible	Mostly not possible
DATA SET	Large	Small

X. FUTURE WORK

The project mainly deals with the presence of cervical cancer cells found in the female in the lower abdominal area that is cervix . However the future of the project may include the details about the stage in which the one is suffering from. Thereby it will be easy to get proper treatment done.

ACKNOWLEDGMENT

We would like to express my special thanks of gratitude to Our project handler as well as our Dean who gave me the golden opportunity to do this wonderful project on the topic (Cervical Cancer prediction using machine learning), which also helped me in doing a lot of Research and we came to know about so many new things We are really thankful to them. Secondly we would also like to thank our parents and friends who helped me a lot in finalizing this project within the limited time frame.

REFERENCES

1. Komala Rayavarapu Department of CSE Vignan's Foundation for Science, Technology, and Research Guntur, Andhra Pradesh, India komala.rayavarapu@gmail.com Krishna Kishore K.V., MIEEE Dept of CSE Vignan's Foundation for Science, Technology and Research Guntur, Andhra Pradesh, India Kishorekvk_1@yahoo.com," "Prediction of Cervical Cancer using Voting and DNN Classifiers" Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India
2. Sun Y. Park*, "Domain-Specific Image Analysis for Cervical Neoplasia Detection Based on Conditional Random Fields", IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 30, NO. 3, MARCH 2011 867
3. Yulan Wang, Chuan Song, Mengyan Wang, Yonghui Xie, Lan Mi, and Guifang Wang "Rapid, Label-Free, and Highly Sensitive Detection of Cervical Cancer With Fluorescence Lifetime Imaging Microscopy" IEEE JOURNAL OF SELECTED TOPICS IN QUANTUM ELECTRONICS, VOL. 22, NO. 3, MAY/JUNE 2016 6801307
4. Shys-Fan Yang-Mao, Yung-Kuan Chan, and Yen-Ping Chu , "Edge Enhancement Nucleus and Cytoplast Contour Detector of Cervical Smear Images" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 38, NO. 2, APRIL 2008 353
5. Costas Balas, "A Novel Optical Imaging Method for the Early Detection, Quantitative Grading, and Mapping of Cancerous and Precancerous Lesions of Cervix" 96 IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 48, NO. 1, JANUARY 2001
6. Kagan Tumer, *Member, IEEE*, Nirmala Ramanujam, Joydeep Ghosh, and Rebecca Richards-Kortum*reported in the United States alone, in 1995 [1]. "Ensembles of Radial Basis Function Networks for Spectroscopic Detection of Cervical Precancer" IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 45, NO. 8, AUGUST 1998 953
7. N. Ayyanar**,** G. Thavasi Raja, Mohit Sharma and D. Sriram Kuma," Photonic Crystal Fiber Based Refractive Index Sensor for Early Detection of Cancer", DOI 10.1109/JSEN.2018.2854375, IEEE Sensors Journal
8. Ahmad Taher Azar, Shaimaa Ahmed El-Said,"Probabilistic neural network for breast cancer classification", Neural Comput & Applications, Springer,23:1737–1751, 2013.
9. WEN WU AND HAO ZHOU Corresponding author: Hao Zhou (15754603750@163.com) "Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches "
10. K. Hemalatha Research Scholar Dept. of Computer Science Sri Padmavati Mahila Visvavidyalayam Tirupati hemalathakulala@gmail.com, K. Hemalatha Research Scholar Dept. of Computer Science Sri Padmavati Mahila Visvavidyalayam Tirupati hemalathakulala@gmail.com, "An Optimal Neural Network Classifier for Cervical Pap smear Data", 2017 IEEE 7th International Advance Computing Conference (IACC)
11. Endah Purwanti, M. Arief Bustomi, Royan Dawud Aldian,"Classification of Cervical Cells Using Applied Computing Based Artificial Neural Network", Indonesian Scholars Journal – Vol 1, No. 1, August 2013.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)