JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CERVICAL CANCER PREDICTION USING MACHINE-LEARNING TECHNIQUES

Ms J Fathima Kaleema
Department of Computer Science,
TBAK College for Women,
Kilakarai, Tamilnadu, India.

Abstract: Cervical cancer is the fourth most common cancer among women around the world. The objective of this study is to provide a comparative study to predict the cervical cancer dataset. The extraction involved over 38 attributes here used three different machine learning algorithms (XGBoost, Decision Tree, Logistic Regression) has been applied on four different medical tests (Biopsy, Cytology, Hinselmann, and Schiller) as four different target variables. The disease cannot be identified in the early stage. The result showed that the performance of EML outperforms other classifiers after evaluation. In this paper it exposes the classifiers can effectively achieve the best performance with the least number of highly important attributes.

Keyword: Machine Learning, XGBoost , Decision Tree, Logistic Regression

I INTRODUCTION

Cervical cancer is a type of cancer that begins in the cervix which connects the uterus and vagina. This cancer can affect the deeper tissues of their cervix and may be spread to other parts of their body after breast cancer, colorectal cancer, lung cancer and skin cancer as per [1].Many young women become infected with multiple types of human papillomavirus, which then can increase their risk of getting cervical cancer in the future and it may spread to other parts of their body . Most of the women who don't have early abnormal changes who do not have regular examinations are at high risk for localized cancer by the time they're age of 40 and for invasive cancer by age of 50.The target of this paper to diagnos the four different variables hinselmann, schiller, citology, and biopsy. This paper examines to diagnose the biopsy

Cervical cancer dataset collected from kaggle dataset and used google colab the platform for the purpose of coding for dataset prediction. The methodology involves use of supervised learning algorithms and classification technique like Decision Tree, Logistic Regression and XGBoost Classifier with Dimensionality Reduction technique.

II LITERATURE REVIEW

Early detection of cervical cancer is a way to avoid the death rate because most of the women affected from cervical cancer .They didn't identify the cervical cancerous in abnormal state this is main cause of an increase the female death rate in this world[2].

WEN WU "Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine- Based Approaches", Department of Blood Transfusion, Jinan Military General Hospital, Jinan, China Year: 2017 [3]. In this paper it claimed that some set of data to the way of increasing the possibility of developing cervical cancer. An advantage of this paper is to SVM(Support vector Machine) to introduce the diagnosis the cervical cancer. It supports to improve the SVM method, support vector machine-recursive feature elimination and support vector machine-principal component analysis (SVM-PCA), are proposed to diagnose the cancer samples.

H. Teame et al., "Factors associated with cervical precancerous lesions among women screened for cervical cancer in Addis Ababa, Ethiopia: A case control study," PLoS ONE 13(1) e0191506, 2018[4]. Age play the vital role in increasing the risk of developing the cervical cancer. In this paper, most of the women infected with HPV during sexual transmitted infections were associated with precancerous lesion.

DATA COLLECTION

The data was collected from the kaggle Machine Learning repository and available in online. The dataset contains 858 instances and 36 different features

No	Attribute	No	Attribute
1.	Age	19.	STDs:pelvic inflammatory disease
2.	Number of sexual partners	20.	STDs:genital herpes
3.	First sexual intercourse	21.	STDs:molluscum contagiosum
4.	Num of pregnancies	22.	STDs:AIDS
5.	Smokes	23.	STDs:HIV
6.	Smokes (years)	24.	STDs:Hepatitis B
7.	Smokes (packs/year)	25.	STDs:HPV
8.	Hormonal Contraceptives	26.	STDs: Number of diagnosis
9.	Hormonal Contraceptives (years)	27.	STDs: Time since first diagnosis
10.	IUD	28.	STDs: Time since last diagnosis
11.	IUD (years)	29.	Dx:Cancer
12.	STDs	30.	Dx:CIN
13.	STDs (number)	31.	Dx:HPV
14.	STDs:condylomatosis	32.	Dx
15.	STDs:cervical condylomatosis	33.	Hinselmann
16.	STDs:vaginal condylomatosis	34.	Schiller
17.	STDs:vulvo-perineal condylomatosis	35.	Citology
18.	STDs:syphilis	36.	Biopsy

III DATA PRE-PROCESSING

It is a technique that is used to convert raw data into numeric dataset .In this dataset some data having NaN (Not a Numeric) values are replaced with numerical values with use of digit transformation using sklearn techniques. After converting dataset feeding it to the algorithm. For getting better result from the Machine Learning algorithm, the ML algorithm not support null values so there is need to preprocess for medical dataset which has major attribtues, The dataset was split into Train dataset and Testing dataset.In this process 20% of test dataset and 80% of trained dataset

IV METHODOLOGY

i) XGBoost

Extreme Gradient Boosting or XGBoost is a decision tree based ensemble ML Algorithm that is used in the library of gradient boosting Some of the major benefits of XGBoost are that its highly scalable/parallelizable, easy to visualize, quick to execute, and typically outperforms other algorithms are used to more regularized model formalization, to control over-fitting, which gives it better performance.

ii) DECISION TREE

Decision Tree is a supervised learning technique that will be used for both classification and regression problems, but mostly it is preferred to solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. This algorithm compares the values of root attribute with the root (dataset) attribute and, based on the comparison, it follow the branch of the next node.

iii) LOGISTIC REGRESSION

Logistic regression is also one of the most popular Machine Learning algorithms, It's comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

V RESULTS AND DISCUSSION

The experimental test results on the proposed model are explained the predictive performance of XGBoosting, Logistic Regression and decision tree algorithm is analyzed by the performance metrics of accuracy.

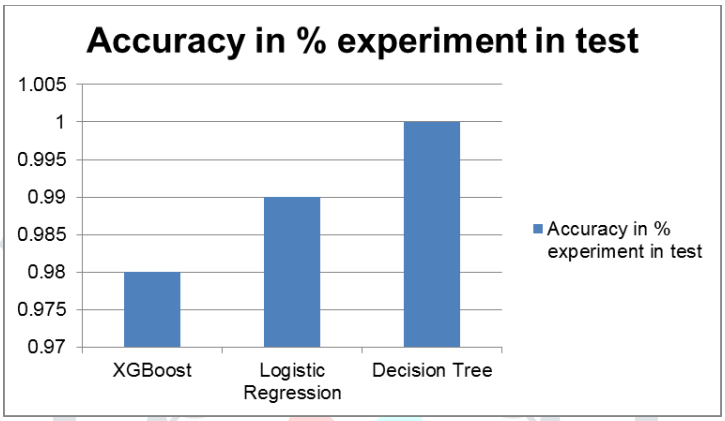
PREDICTIVE ACCURACY ANALYSIS

The predictive performance of the proposed model is experimented on the training set. The predictive accuracy of the proposed model is shown in Figure 1. Moreover, the accuracy for XGBoosting, Logistic Regression and Decision Tree for cervical cancer classification test is given in Table 1 it shows the test accuracy for comparison of three algorithms. Fig 1 shows the result of accuracy in the percentage of experiment test

Table 1: Accuracy of XGBoosting, Logistic Regression and decision tree

Learning Algorithm	Accuracy in % experiment in test
XGBoost	**0.98**
Logistic Regression	**0.99**
Decision Tree	**1.0**

Figure 1 Accuracy after pre-processing



VI CONCLUSION

In this paper, it proposed cervical cancer prediction model with XGBoosting Logistic regression and decision tree algorithm on cervical cancer dataset collected form kaggle data repository. The proposed model solves the problem of biased classification on imbalanced observation by non-ensemble algorithm through ensemble classifier namely the decision tree. The predictive performance of the proposed model is evaluated by employing different performance metrics such as accuracy on the test set. The result of performance analysis reveals that the decision tree algorithm has better performance to comparing between logistic regression and XGBoosting algorithms. Hence, the Decision tree algorithm is a better performance on prediction of the majority class and poor performance on the minority class.

REFERENCE

[1] American Cancer Society, Cancer Facts for Women, Available at: https://www.cancer.org/healthy/find-cancer-early/womens-health/cancer-facts-for-women.html(Accessed date 25.07.2020).

[2] W. J. Koh et al., "Cervical cancer, version 2.2015," JNCCN Journal of the National Comprehensive Cancer Network, vol. 13, no. 4. Harborside Press, pp. 395–404, 01-Apr-2015.

[3] W. Wu and H. Zhou, "Data-driven diagnosis of cervical cancer with support vector machine-based approaches," IEEE Access, vol. 5, pp. 25189–25195, Oct. 2017.

[4] H. Teame et al., "Factors associated with cervical precancerous lesions among women screened for cervical cancer in Addis Ababa, Ethiopia: A case control study," PLoS ONE 13(1) e0191506, 2018

[5] S. Sharma, "Cervical Cancer stage prediction using Decision Tree approach of Machine Learning",*International Journal of Advanced Research in Computer and Communication Engineering* vol. 5, Issue 4,2016.

[6]D. N. Punjani, K. H. Atkotiya, "Cervical Cancer Prediction using Data Mining",*International Journal for Research in Applied Science & Engineering Technology (IJRASET)* , vol. 5, Issue XII, 2017.

[7] S. Sharma, "Cervical Cancer stage prediction using Decision Tree approach of Machine Learning",*International Journal of Advanced Research in Computer and Communication Engineering* vol. 5, Issue 4,2016.

[8] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13–17, pp. 785–794, San Francisco, CA, USA, August 2016.

[9] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, 2013.

[10] Y. M. S. Al-Wesabi, A. Choudhury, and D. Won, "Classification of cervical cancer dataset," in *Proceedings of the 2018 IISE Annual Conference and Expo*, pp. 1456–1461, Orlando, FL, USA, May 2018.

[11] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, "Machine learning for assisting cervical cancer diagnosis: an ensemble approach," *Future Generation Computer Systems*, vol. 106, pp. 199–205, 2020.

[12]"Cervical cancer," 2020, https://www.nccc-online.org/hpvcervical-cancer/cervical-cancer-overview/.

[13] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, "Machine learning for assisting cervical cancer diagnosis: an ensemble approach," *Future Generation Computer Systems*, vol. 106, pp. 199–205, 2020.