IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Implementing Value-Sensitive Machine Learning to Develop a Risk Level Self-Assessment Model for Cervical Cancer

Kaixin Kate Yin

International School of Beijing, 10 An Hua Street, Shunyi District, Beijing, 101300, China; kate.yin@student.isb.bj.edu.cn

ABSTRACT: Cervical cancer begins with cancerous cells in the cervix. Cervical cancer is the third most common cancer worldwide, and 80% of the cases occur in developing countries. The high incidence of this cancer in the developing world is mostly due to a lack of effective screening programs aimed at detecting and treating precancerous conditions. This project aimed to mitigate this issue by developing a self-assessment model based on value-sensitive machine learning. The model would advise if the user should receive a cervical cancer screening based on their lifestyle and disease history. The machine learning model is developed based on dataset with survey responses to enable preliminary assessment on risk level before seeking healthcare resources. The dataset had 858 records; 55 patients had a positive cervical cancer diagnosis, and the remaining 803 patients were healthy. A machine learning approach was adopted, and the samples were divided into two groups randomly as the training and testing groups. 70% (600 patients) of the entire dataset was used to train the machine, and the remaining 30% (258) was assigned as the test dataset. Various classifiers, such as a decision tree, SVM, and logistic classification were also implemented. To evaluate each classification method, a confusion matrix was generated for each method, and classifiers were compared using F1 score and false negative rate. The tradeoff between overall classifier performance and the consequence of false negative rate in this scenario was discussed and an implementation suggestion was provided.

KEYWORDS: Computational Biology and Bioinformatics; Computational Biomodeling; Machine Learning.

Introduction

Cancer stems from normal cells transforming into tumor cells in a multistage process. The chances of survival and treatment for cancer decrease over time; therefore, significant improvements in a patient's chance of survival can be made if diagnoses were made in the early stages and avoiding delays.[1] As in many other diseases, the existence of several screening and diagnosis methods creates a complex ecosystem from a Computer-Aided Diagnosis (CAD) system point of view.[2] However, in developing countries with more limited medical care resources, people may not be able to accurately detect cancer in the early stages, resulting in higher morbidity rates. In addition, the social stigma against women with cervical cancer may be high in developing countries, which can prevent women from seeking medical attention. Therefore, the most critical problems during diagnosis are related to determining the most appropriate screening plan and estimating individual risk for each patient.[3]

To reduce unnecessary screenings and reduce the need for accessing healthcare resources, people who are concerned about their risk of cancer could complete a lifestyle and disease history, which helps determine if they are at risk for developing cancer.[4] Afterward, these patients can seek out screening according to their risk level. Such a risk prediction survey can help developing countries support the targeted group more effectively and reduce the burden on healthcare.[5]

This investigation attempted to develop a risk level self-assessment model using three machine learning methods: decision tree, logistic classification, and support vector machine (SVM). These approaches were chosen due to their renowned high accuracy and efficiency.[6,8] A decision tree model forces the consideration of all possible outcomes of a decision and traces each path to a conclusion. It creates a comprehensive analysis of the consequences along each branch which is suitable for the issue at hand. A cervical cancer dataset contains multiple attributes that need to be taken into account; therefore, all of these factors need to be considered before reaching a conclusion. A logistic classification achieves similar purposes; it estimates the probability of an occurrence of an event based on one or more inputs, which again matches this exploration's objective.[10] A SVM model is known for its kernel trick to handle nonlinear inputs, which could be applied to the case of a cervical cancer dataset that contains Boolean data.[11]

In this study, the cervical cancer risk detection algorithm was built based on a dataset consisting of survey responses and biopsy diagnosis. An algorithm selection was performed based on increasing overall model performance and reducing false negative rates. A discussion on trade-offs between algorithm performance and the consequence of predicting false negative cases was also provided.

Dataset:

The dataset used in this project was collected by Hospital Universitario de Caracas in Caracas, Venezuela. It comprises demographic information, lifestyle, and disease history of 858 patients.[12] There are 35 attributes in total, with 4 types of diagnosis results. The attributes in the dataset are summarized in Table 1.

DOI: 10.36838/v4i1.20

Table 1: Attribute Information.

Feature	Type	Feature	Type
Age	Integer	STDs: pelvic inflammatory disease	Boolean
# of partners	Integer	STDs: genital herpes	Boolean
Age of 1st intercourse	Integer	STDs: molluscum contagiosum	Boolean
# of pregnancies	Integer	STDs: AIDS	Boolean
Smokes	Boolean	STDs: HIV	Boolean
Smokes years	Integer	STDs: Hepatitis B	Boolean
Smokes packs/year	Integer	STDs: HPV	Integer
Hormonal Contraceptives	Boolean	STDs: Number of diagnosis	Integer
Hormonal Contraceptives years	Integer	STDs: Time since first diagnosis	Integer
IUD	Boolean	STDs: Time since last diagnosis	Integer
IUD years	Integer	Dx: Cancer	Boolean
STDs	Boolean	Dx: CIN	Boolean
STDs number	Integer	DX: HPV	Boolean
STDs: condylomatosis	Boolean	Dx	Boolean
STDs: cervical condylomatosis	Boolean	Hinselmann: target variable	Boolean
STDs: vaginal condylomatosis	Boolean	Schiller: target variable	Boolean
STDs: vulvo-perineal condylomatosis	Boolean	Cytology: target variable	Boolean
STDs: syphilis	Boolean	Biopsy: class or target variable	Boolean

Related Work:

Except for the original data, the earliest study conducted based on this dataset was a cost-sensitive classifier, whose accuracy had passed the basic level by a narrow margin.[7] Afterward, more algorithms were developed based on this dataset, such as using two improved support vector machine (SVM) approaches to predict the risk of cervical cancer.[13] Recently, another approach using Firefly Algorithm and Random Forest Classifier was developed.[14] This study was further improved when the synthetic minority oversampling technique (SMOTE) was used to reduce the number of features based on Random Forest classification. Researchers recently managed to achieve an accuracy of 97.25% using a stacked autoencoder with a soft-max layer.

However, all of these studies were optimized for overall classifier performance. There lacks a useful case-sensitive, value-sensitive approach to building classifiers for cervical cancer. Because this dataset was collected in a developing country with highly accessible lifestyle and disease history data, it opens the opportunity to develop a pre-cancer screening, self-assessment tool to raise awareness of women's health. In addition, it gives rise to a possible increase in the early diagnosis rate for women who face challenges for screening for cervical cancer.

In this study, a value-sensitive machine learning approach was implemented. The focus was on the real-world use case and on reducing the false negative rates instead of improving the overall model. The models developed in this study were optimized on false-negative because the project's aim was to develop a pre-screening self-assessment tool, in which false negatives (advising high-risk individuals that there's no need to screen for cervical cancer) leads to much more serious consequence comparing to false positives (advising low-risk individuals to undergo additional screening).

Methods

Value Sensitive Machine Learning:

Value sensitive machine learning is an approach that takes values of ethical importance into account.[15] In this case, human values were taken into account in a well-defined matter throughout the entire modeling process. Quantitative and qualitative data were collected from a survey and the attributes were interpreted under the specific cultural and sociotechnical contexts.[16]

Data Cleaning:

Data dropping and data filling were performed to resolve the missing data in this dataset. First, systematic missing, or cases in which the patients did not provide a response to certain survey questions were identified. This was denoted by Boolean attributes in the dataset, such as whether the patient has a history of STD, or if the patient is under IUD birth control. The missing data could be a result of a reluctance to answer sensitive questions, unknown disease history, difficulty recollecting lifestyle, or no response to survey at all. These data that are missing are completely random, and any filling could bias the original dataset. Thus, the patients who have missing Boolean attributes in any of the survey responses were eliminated. This step produced 728 samples in the dataset.

Regarding the missing Integer attributes, missing responses were filled in with the mean or median of such attributes. Replacing the above with two approximations was a statistical approach of handling missing numerical values. Although this method may add variance to the overall dataset, it was more effective than dropping columns of data. Examples of these attributes include years of smoking, number of pregnancies,

DOI: 10.36838/v4i1.20

number of pregnancies, number of sexual partners, which all have a known correlation with risk for cervical cancer.

Models:

Based on existing work, multiple models were trained using Decision Tree, Logistic Regression and SVM.

Decision tree, as one of the most frequently applied machine learning methods, is trained on a dataset for classification and regression analysis. This model groups the samples into several groups based on a series of questions. The process of classification is like a tree. The root of the tree includes all samples. Then, it divides into several sets of samples using a recursive procedure. The decision tree's key challenge is the selection of the optimal partition attributes, which can be explained by information entropy, gain ratio, or Gini index.[17]

Logistic classification is a binary classification model in which the conditional probability of one of the two possible realizations of the output variable is assumed to be equal to a linear combination of the input variables, transformed by the logistic function.[18]

A SVM is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, it is able to categorize new data points.[19]

First, classifiers optimizing for overall performance with a high F1 score were trained. Afterwards, the best-performing classifier was selected and the selected classifier optimizing for a lower false-negative rate was retrained.

Evaluation:

The dataset was divided into two groups as the training and testing groups randomly. The training dataset was 70% (508 patients) of the cleaned dataset, and the remaining dataset (218 patients) was assigned as the test dataset.

The training dataset was used to train a Decision Tree, an SVM and a Logistic Regression classifier with 10-fold cross-validation. Then, each model was tested on the testing dataset to evaluate its performance.

There are various performance indicators. Since this study only contained two classes, the Percentage of Correctly Classified Instances (PCCI) was used as a performance indicator. In the following expressions, a positive result means that the biopsy test showed positive for cervical cancer and vice versa. The results could be divided into four groups. They were:

• Correctly Classified Class 0 Instances also called as True Negative Class 0 (TNC0)

• Falsely Classified Class 0 Instances also called as False Positive Class 0 (FPC0)

• Falsely Classified Class 1 Instances also called as False Negative Class 1 (FNC1)

• Correctly Classified Class 1 Instances also called as True Positive Class 1 (TPC1)

Results and Discussion

Optimizing for overall performance, the following performance was obtained for the test set:

Decision tree

Using the decision tree method, an F1 score of 92.9% was achieved with the following confusion matrix:

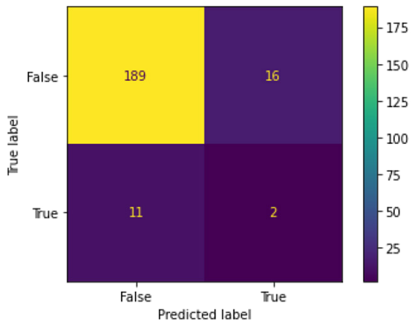

Figure 1: The model generated using the decision tree method with an F1 score of 92.9% and 11 FNC1 cases.

Figure 1 shows the successful identification of 189 TNC0 cases, 2 TPC1 cases, 16 FPC0 cases and 11 FNC1 cases. In cancer screening, it is crucial to avoid false negative cases, also known as Type II errors. Type II errors should be avoided because they could eventually cost a human's life due to mistaken diagnosis.

SVM

Using the SVM method, an F1 score of 92.9% was achieved with the following confusion matrix:

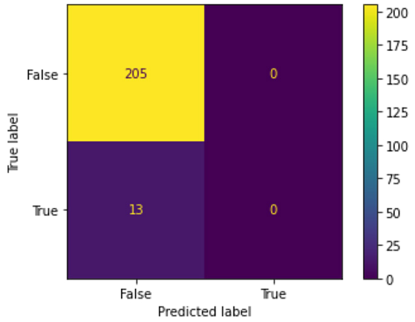

Figure 2: The model generated using the decision tree method with an F1 score of 92.9% and 11 FNC1 cases.

Figure 2 shows the successful identification of 205 TNC0 cases, 0 TPC1 case, 0 FPC0 cases and 13 FNC1 cases.

Though more TNC0 cases were correctly identified through this method, the number of FNC1 cases also increased, which is not ideal.

Logistic Classification

Using the logistic classification method, an accuracy of 93.1% was achieved with the following confusion matrix:

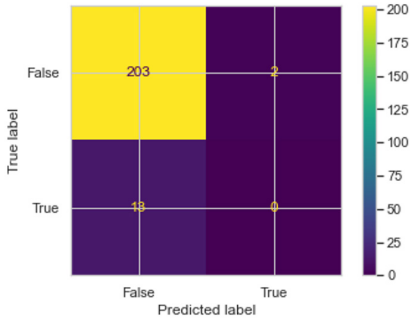

Figure 3: The model generated using the logistic classification method with an F1 score of 93.1% and 13 FNC1 cases.

Figure 3 shows the successful identification of 203 TNC0 cases, 0 TPC1 case, 2 FPC0 cases and 13 FNC1 cases

Though the accuracy improved by 0.2% compared to both

decision tree and SVM, the number of FNC1 cases and FPC0 cases also increased, which is not ideal.

Afterwards, the logistic classification model that would be used for secondary training was chosen because it presented the highest F1 score. A logistic classification classifier was re-trained while optimizing for recall (lowering the false negative rate) and the final classifier was obtained. The final classifier obtained an F1 score of 69.2% and a false negative rate of 0.917% with the following confusion matrix:

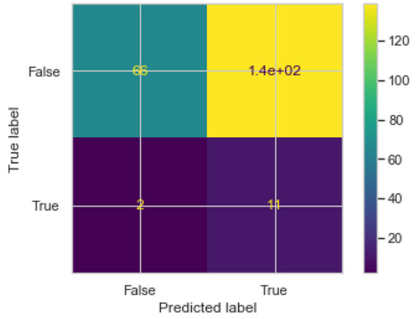

Figure 4: The model generated using the logistic classification method after secondary training with 2 FNC1 cases.

Though the F1 score is much lower than the other two models done using SVM and decision tree, the final classifier using Logistic Regression identifies the least number of FNC1 cases.

Discussion

To evaluate each model's effectiveness, it was more reasonable to assess by examining the number of FNC1 cases rather than the F1 score. FNC1 cases, also known as type II errors, are detrimental to cancer screening because human lives are priceless. Hence, the goal of this risk assessment tool should be to raise awareness and encourage people who are at risk to get tested. As a result, the model should optimize for low FNC1 because if people are at risk of cancer but not aware of their risks, it could cost them their lives.

Though SVM and decision tree models achieve a higher F1 score compared to the logistic classification model (92.9% compared to 69.2%), the confusion matrix of both SVM and decision tree models showed that FNC1 cases were very high: 11 and 13, respectively. On the other hand, the logistic classification model only had 2 FNC1 cases, which was much less than the ones specified by SVM and decision tree models.

Therefore, the logistic classification serves as a more suitable and ideal model designed under the real-world context, successful in getting all people with possible risks to get tested and optimized for the overall medical system efficiency.

Conclusion

It was shown that it is feasible to implement a cervical cancer risk level assessment model using survey responses, enabling self-assessment tools for women in developing countries to perform self-assessment for their risk level before seeking medical resources and biopsy screening. Although the best F1 classification result was obtained by decision tree and SVM, the best model should optimize for a lowering false negative rate due to the detrimental consequences of false negatives in cancer screening. The number of cases classified as false positives indicated that the number of patients with cancer who are not warned. In this study, this number was presented as False Negative (FNC1). For each method, the number of instances classified as false negative was 11, 13, and 2 with decision tree, SVM and logistic classification, respectively. That means that false classified instance rates were 4.26%, 5.04%, and 0.69%, respectively. Because the number of correctly classified instances across all three methods are very similar, the false-negative rates became more critical when determining the best approach. So, in the authors' opinion, the best model for this study is determined to be the logistic classification method.

Acknowledgements

I thank Ms. Zhao and Mrs. Monroe for academic support.

References

1. Cancer Report. February 2017 [cited WHO World Health Organization 25.09.2017]; Available from: http://www.who.int/mediacentre/factsheets/fs297/en/.
2. Ünlerşen, Muhammed & Sabanci, Kadir & Özcan, Muciz. (2017). Determining Cervical Cancer Possibility by Using Machine Learning Methods. International Journal of Latest Research in Engineering and Technology. 3. 65-71.
3. Fernandes, K., J.S. Cardoso, and J. Fernandes, Transfer Learning with Partial Observability Applied to Cervical Cancer Screening, in Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings, L.A. Alexandre, J. Salvador Sánchez, and J.M.F. Rodrigues, Editors. 2017, Springer International Publishing: Cham. p. 243-250.
4. Cancer Prevention and Control. "Cancer Plan Self-Assessment Tool." Centers for Disease Control and Prevention, www.cdc.gov/cancer/ncccp/pdf/cancerselfassesstool.pdf.
5. Nikpour, Maryam & Hajian-Tilaki, Karimollah & Bakhtiari, Afsaneh. (2021). Risk Assessment for Breast Cancer Development and Its Clinical Impact on Screening Performance in Iranian Women [Corrigendum]. Cancer Management and Research. Volume 13. 3079-3080. 10.2147/CMAR.S311176..
6. P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," in IEEE Transactions on Geoscience Electronics, vol. 15, no. 3, pp. 142-147, July 1977, doi: 10.1109/TGE.1977.6498972.
7. Dreiseitl, Stephan, and Lucila Ohno-Machado. "Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review." Journal of Biomedical Informatics, vol. 35, no. 5-6, 2002, pp. 352–359., doi:10.1016/s1532-0464(03)00034-0.
8. Pisner, Derek A., and David M. Schnyer. "Support Vector Machine." Machine Learning, 2020, pp. 101–121., doi:10.1016/b978-0-12-815739-8.00006-7.
9. Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044
10. Castañón, Jorge. "10 Machine Learning Methods That Every Data Scientist Should Know." Medium, Towards Data Science, 5 Sept. 2019, towardsdatascience.com/10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9.
11. Suthaharan S. (2016) Support Vector Machine. In: Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems, vol 36. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7641-3_9
12. Fernandes, Kelwin, Cardoso, S. Jaime and Fernandes, Jessica.

DOI: 10.36838/v4i1.20

"Cervical cancer (Risk Factors) Data Set." UCI, Hospital Universitario de Caracas, 2017, archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29. Accessed 13 Feb. 2021.
13. W. Wu, H. Zhou, Data-driven diagnosis of cervical cancer with support vector machine-based approaches, IEEE Access 5 (2017) 25189–25195.
14. Sawhney, Ramit & Mathur, Puneet & Shankar, Ravi. (2018). A Firefly Algorithm Based Wrapper-Penalty Feature Selection Method for Cancer Diagnosis. 10.1007/978-3-319-95162-1_30.
15. Umbrello, S., van de Poel, I. Mapping value sensitive design onto AI for social good principles. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00038-3
16. Simon, J. & Wong, P.-H. & Rieder, G. (2020). Algorithmic bias and the Value Sensitive Design approach. Internet Policy Review, 9(4). https://doi.org/10.14763/2020.4.1534
17. Rokach, Lior, and Oded Maimon. "Decision Trees." Data Mining and Knowledge Discovery Handbook, Jan. 2005, pp. 165–192., doi:10.1007/0-387-25465-x_9.
18. Maalouf, Maher. (2011). Logistic regression in data analysis: An overview. International Journal of Data Analysis Techniques and Strategies. 3. 281-299. 10.1504/IJDATS.2011.041335.
19. Zhou, Zhi-Hua. "Support Vector Machine." Machine Learning, pp. 129–153., doi:10.1007/978-981-15-1967-3_6.

Author

This is Kaixin Yin, currently a high school senior attending International School of Beijing. I have always been greatly intrigued by the complexity of science, especially bioinformatics. Driven by my passion for programming, I self-studied machine learning and then created a risk level self-assessment model for women in developing countries.