Predicting the Likelihood of Developing Cervical Cancer According to Patient Behavior Using a Random Forest Classifier

Kyli McKee

December 2020

Abstract

The use of machine learning in the field of medicine has started to increase rapidly. Specifically, supervised machine learning techniques have had the biggest impact on diagnostic and prognostic methods. This study analyzes a data set that includes 72 patients and looks at 19 different behavioral features. Each patient is assigned a 0 or a 1 according to whether they did or did not develop cervical cancer respectively. Using this data set, a supervised learning algorithm was created to predict a patient's likelihood of developing cervical cancer according to their behavior. By utilizing machine learning techniques such as supervised k-folds cross-validation and a random forest classifier, the study was able to create a prediction model that is 92.86% accurate. Additionally, the most important behavioral features that play a role in the development of cervical cancer were identified; these features are empowerment_abilities, motivation_strength, and behavior_personalHygine. The supervised learning algorithm developed in this study can be utilized in clinical settings around the world to increase the early detection and prevention of cervical cancer. Overall, the implementation of machine learning has the opportunity to create widespread change in the medical field and improve patient outcomes.

Keywords: Machine learning, Cervical cancer, Random forest classification, Stratified K-fold cross-validation

1 Introduction

Machine learning is a discipline that combines both computer science and statistics to analyze how computers learn from data. This discipline is driven by the need to determine relationships from large sets of data that are often complex and multifactorial [2]. Due to its wide variety of applications, machine learning is becoming increasingly utilized in many different fields, especially research and medicine. Specifically in medicine, machine learning has made significant improvements in medical specialties such as personalized medicine, surgery, therapeutics, radiology, oncology, and many more [4].

Machine learning is separated into two broader categories: supervised and unsupervised learning. Supervised learning works by providing the computer with features and their desired outcomes. With the inputs and outputs the computer is tasked with finding links between the two groups. These specific learning algorithms are capable of performing regression (outputs are continuous values) and classification (outputs are discrete values). In medicine, supervised learning techniques are most commonly utilized to determine a patient's diagnosis and prognosis. Unsupervised learning is given a set of features that do not have corresponding outputs. This type of machine learning tasks the computer with determining whether or not there are any patterns present within the dataset. As opposed to classification and regression, unsupervised learning is primarily focused on dimensionality reduction and clustering the data [4].

The concept of utilizing machine learning in medicine has only grown significant support over the last decade. With technology improving so rapidly, specifically the improvement of computational resources along with data storage and sharing, the implementation of machine learning in the medical field has become more feasible. Even though there have been significant strides made to slowly introduce machine learning practices into medicine, there are still many drawbacks and factors that need to be addressed. Some aspects of machine learning that still need improvements are the security of the data, limiting the number of false classifications, limiting statistical inference, along with other drawbacks. Overall, we are seeing a shift in the medical field as machine learning is progressively getting introduced and used in practice [1].

As the number of clinical datasets are increasing, there has been increased usage of machine learning techniques for diagnostic and prediction tasks [8]. This study is aimed at developing a supervised learning algorithm that is capable of predicting a patient's risk of developing cervical cancer. Cervical cancer is the cause of 77% of deaths in women worldwide. Many of these cases and deaths are occurring in developing countries. What is unique about cervical cancer is that it has a prolonged pre-invasive period which contributes to the fact that it is preventable and diagnosable in early stages of the disease. Additionally, there are many behavioral factors that are associated with the development of cervical cancer [3]. Factors such as personal hygiene, eating habits, attitude, social support, and empowerment are some of the behavioral aspects that are thought to play a role in the development of cervical cancer. The present study was conducted in order to determine the behavioral features that are the most relevant in predicting the development of cervical cancer, and to create a predictive model based on the identified relevant features. This model will contribute significantly to the early detection and prevention of cervical cancer, which will improve the overall patient outcomes.

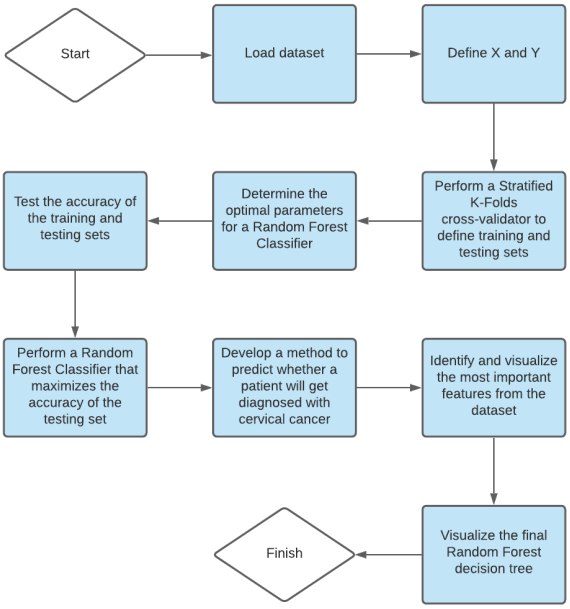

Figure 1: Study Flowchart
This image is an illustration of the steps performed to complete this study.

2 Methods

This study is designed to identify the most important behavioral features that contribute to the development of cervical cancer. Additionally, it aims at creating a predictive model to determine if a patient's behavior will lead to the development of cervical cancer. In order to accomplish the desired outcomes of this study, machine learning techniques are utilized to create a supervised learning algorithm. The techniques that are utilized in this study are a stratified K-folds cross-validator and a random forest classifier. The mechanism of the study is included below (Figure 1). This flowchart illustrates a brief summary of the steps involved in creating the machine learning classification algorithm that was developed during this study.

2.1 Technology

This project was designed and executed using Python version 3.8.3.

2.2 Cervical Cancer Data Set

The data set utilized in this study was titled "Cervical Cancer Behavior Risk Data Set" and was obtained from the University of California, Irvine's (UCI) website [7]. The data set analyzes 72 patients and 20 variables were measured for each patient. The first 19 variables are multivariate and are categories that define specific behavioral attributes. These 19 behavioral categories are behavior_sexualRisk, behavior_eating, behavior_personalHygine, intention_aggregation, intention_commitment, attitude_consistency, attitude_spontaneity, norm_significantPerson, norm_fulfillment, perception _vulnerability, perception _severity, motivation _strength, motivation _willingness, socialSupport _emotionality, socialSupport _appreciation, socialSupport_instrumental, empowerment_knowledge, empowerment_abilities, and empowerment_desires. For each individual behavioral variable, the patient is ranked on a scale of 1 to 15 according to how severe their behavior is. The final measured variable is ca_cervix and it is univariate. For this category each patient is given a score of 0 or 1 according to whether they did not develop or they did develop cervical cancer respectively. Of the 72 patients included in the data set, 21 of them developed cervical cancer and the remaining 51 did not.

2.3 Separation of Features and Classes

The data set had to be split into features and classes in order to proceed with the project. Before doing so, the data was converted to a numpy array because as an array the data takes up less space and it is more compatible with various different libraries. Once converted into an array, both X and Y were defined. X is defined as the features of the data, meaning it includes the data from the 19 behavioral categories listed above. Y is defined as the class, the ca_cervix column data, that states whether or not a patient developed cervical cancer.

2.4 Stratified K-fold Cross-Validation

Once X and Y were defined, a stratified K-fold cross-validation resampling technique was applied to the data. When utilizing a stratified K-fold cross-validation, the data is split into k equal parts. The training and testing data sets are then formed from these separate, equal parts of the data. The training set is equal to k-1, and the remainder of the data is used to test and validate the model. This resampling process is repeated k times in order to increase the accuracy and generalizability of the model [6]. The model in this study employed a k-value of five because it is one of the most common k-values used in this method. This selection resulted in five different training and testing sets. The model's performance for each set was evaluated and the final model accuracy was determined by averaging the accuracy achieved during each of the five iterations.

2.5 Random Forest Classifier

In this study a random forest classifier was used to determine the most important features from the data set and create a prediction test that can determine whether or not a patient will develop cervical cancer. Random forests are a very common supervised classification method. A random forest contains groups of decision trees that are each created by a bagging algorithm, an algorithm that creates many subsamples of the data set. Each tree uses different features to create a class prediction. All of the individual trees together form a forest that predicts a class according to the output that appears the most frequently. By utilizing multiple trees as opposed to one decision tree, the chance of overfitting the data significantly decreases, which improves the accuracy of the model [5].

Additionally, when utilizing a random forest there are a number of parameters that can be tuned. Parameter tuning results in faster learning and improved accuracy of the model [6]. In this study the number of estimators (n_estimators) and the max depth of the tree (max_depth) were experimented with to find the parameters that yielded the highest testing set accuracy. In order to find the optimal parameters, a for loop was employed to test a range of values for both n_estimators and max_depth.

2.6 Visualization of the Random Forest Classification Trees

The individual random forest classification trees were visualized by importing plot_tree from sklearn.tree. This package allowed us to visualize a single decision tree at a time and analyze the variations between each tree.

3 Results and Discussion

The study aimed to optimize the random forest classifier to obtain the highest possible accuracy for the model. The for loop and function implemented in the study to determine the optimal number of estimators and the max depth of the tree concluded that 64 estimators with a max depth of 2 would give the best results. These two parameters yielded a training set accuracy of 94.83% and a testing set accuracy of 92.86%. After running a random forest classifier with the optimal parameters found above, the study was able to take a new set of behavioral feature scores and input them to predict whether a patient will develop cervical cancer.

One of the most important elements of the study was determining which features are the most influential when determining the onset of cervical cancer. The random forest classification algorithm was able to detect the importance of each feature included in the study and these results are seen below in Table 1 and Figure 2. Table 1 shows the importance for each behavioral feature and concludes that empowerment_abilities, motivation_strength, and behavior_personalHygine are the most important features included in the study. Additionally, it shows that norm_significantPerson, socialSupport_emotionality, and attitude_spontaneity are the features that are the least important for determining the onset of cervical cancer. Figure 2 is a graphical representation of the results reported in Table 1.

Feature	Importance
empowerment_abilities	0.145
motivation_strength	0.112
behavior_personalHygine	0.109
empowerment_knowledge	0.102
perception_vulnerability	0.090
motivation_willingness	0.084
norm_fulfillment	0.069
perception_severity	0.064
behavior_sexualRisk	0.058
empowerment_desires	0.050
intention_commitment	0.032
intention_aggregation	0.023
socialSupport_instrumental	0.016
socialSupport_appreciation	0.014
behavior_eating	0.011
attitude_consistency	0.010
attitude_spontaneity	0.007
socialSupport_emotionality	0.003
norm_significantPerson	0.000

Table 1: Feature Importance
The table above give the individual features and their respective importance in the model created in this study. The higher the importance, the more significant that behavioral feature is at determining whether or not a person will develop cervical cancer.

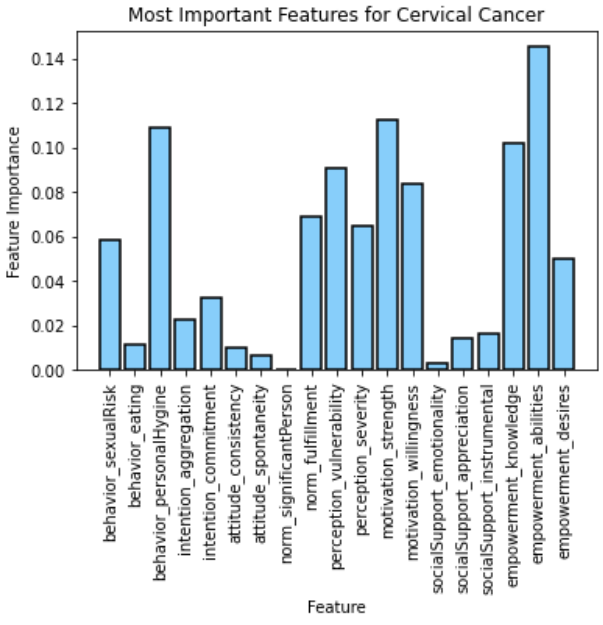

Figure 2: Visualization of Feature Importance
This graph is a visualization of the feature importances shown in Table 1.

The study concluded by creating a method to visualize each of the individual decision trees that were incorporated into the random forest. Comparisons can be made between the 64 different trees and analysis can be performed to see which features were included the most frequently. One of the random forest classification trees is included below in Figure 3.

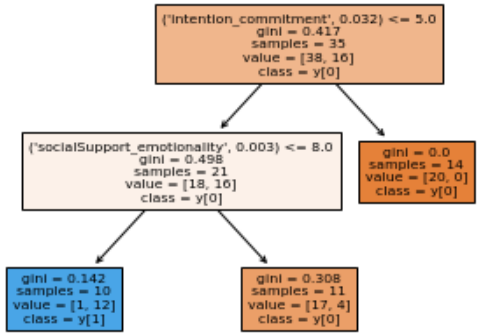

Figure 3: Visualization of Random Forest Classification Tree
This figure is a visualization of a single classification tree created by the random forest model.

The methodology used in this study was able to create a supervised machine learning algorithm that can predict a patient's likelihood of developing cervical cancer with great accuracy. Although this study was very successful and was able to satisfy its objectives, there was a limitation. The data set that was utilized included a fairly small subset of patients. This model could potentially become more accurate if it were to be trained with larger group of patient data. If there is a clear outline that defines what constitutes each score in every behavioral category, many different physicians will be able to contribute to this data set over time, and there will be significantly more data to include in future studies.

4 Conclusion

This study was able to successfully develop an algorithm to predict whether or not a patient will develop cervical cancer with 92.86% accuracy, and determine the most important behavioral features that contribute to the onset of cervical cancer. It was found that the following behavioral features are the three most important at predicting the development of cervical cancer: empowerment_abilities, motivation_strength,

and behavior_personalHygine. Additionally, it was concluded that norm_significantPerson, socialSupport_emotionality, and attitude_spontaneity play very little or no role in determining whether or not a patient will develop cervical cancer.

As mentioned above, more resources can be dedicated to increasing the amount of patient data included in data sets such as the one used in this study. The larger the training data set, the more accurate the machine learning model can become. In future studies this model can continue to be refined in attempts to improve its classification accuracy. One way in which this can potentially be accomplished is by reducing the number of features that are analyzed in the model. Taking away features that are not as indicative could potentially increase the speed and accuracy of the algorithm. Additionally, physicians can devote time to determining other behavioral features that might play a role in the development of cervical cancer. Overall, these types of algorithms can be applied to other diseases that are preventable or benefit from early detection.

In conclusion, machine learning techniques, such as the one created and implemented above, have the potential to make a significant impact on the medical field. Diagnostic, prognostic, and risk prediction methods can notice trends in large data sets that are complex and multifactorial, which are often not apparent to physicians. The ability for machines to detect these trends can improve the speed and accuracy of patient diagnosis. The results from machine learning studies can also advise physicians as to what medical advice and therapeutics are appropriate for a specific patient. As technology continues to improve, there will be a pronounced expansion in the presence of machine learning techniques utilized in the medical field.

5 Acknowledgements

I would like to thank Professor Anibal and Christina for their valuable feedback throughout this project.

References

[1] "Ascent of machine learning in medicine," Nat Mater, vol. 18, no. 5, p. 407, May 2019, doi: 10.1038/s41563-019-0360-1.

[2] R. C. Deo, "Machine Learning in Medicine," Circulation, vol. 132, no. 20, pp. 1920-30, Nov 17 2015, doi: 10.1161/CIRCULATIONAHA.115.001593.

[3] A. F, S. C, and A. L, "Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer," J Biomed Phys Eng, vol. 10, no. 4, pp. 513-522, Aug 2020, doi: 10.31661/jbpe.v0i0.1912-1027.

[4] G. S. Handelman, H. K. Kok, R. V. Chandra, A. H. Razavi, M. J. Lee, and H. Asadi, "eDoctor: machine learning and the future of medicine," J Intern Med, vol. 284, no. 6, pp. 603-619, Dec 2018, doi: 10.1111/joim.12822.

[5] D. Petkovic, R. Altman, M. Wong, and A. Vigil, "Improving the explainability of Random Forest classifier - user centered approach," Pac Symp Biocomput, vol. 23, pp. 204-215, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29218882.

[6] Z. Salod and Y. Singh, "Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol," J Public Health Res, vol. 8, no. 3, p. 1677, Dec 4 2019, doi: 10.4081/jphr.2019.1677.

[7] Cervical Cancer Behavior Risk Data Sheet, UCI, 2016. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Cervical+Cancer+Behavior+Risk

[8] J. Wiens and E. S. Shenoy, "Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology," Clin Infect Dis, vol. 66, no. 1, pp. 149-153, Jan 6 2018, doi: 10.1093/cid/cix731.