Prediction of Cervical Cancer Risk based on Machine Learning Approaches

Group - VioletGo

1. Problem

The objective of this study is to develop a predictive model for the likelihood of cervical cancer diagnosis based on a range of medical and sexual history variables, including age, number of sexual partners, age of first sexual intercourse, number of pregnancies, smoking status, use of hormonal contraceptives or IUDs, history of STDs, and results of diagnostic tests including Hinselmann, Schiller, Cytology, and Biopsy tests.

2. Motivations and background

Cervical cancer is a significant global health problem, particularly in low- and middle-income countries.In 2008, 275,000 deaths occurred due to cervical cancer. Of which, 88% occurred in developing countries[1]. It is the second most common type of cancer in women worldwide, and it seriously threatens women's health. The primary cause of cervical cancer is persistent infection with high-risk human papillomavirus (HPV)[2]. Early detection is crucial for preventing and treating cervical cancer, but it can be difficult in areas lacking resources, leading to high mortality rates. Developing a machine learning model to predict cervical cancer risk could be a promising solution. This model could identify high-risk individuals for early intervention and treatment, potentially saving lives and reducing the global burden of cervical cancer.

3. Data

This dataset [3] contains information about various risk factors associated with cervical cancer in women. The data was collected from 858 patients, and includes 36 features including demographic information, medical history, and lifestyle factors, as well as 4 target variables indicating the presence or absence of cervical intraepithelial neoplasia or cervical cancer based on different diagnostic criteria.We used frequency bar graphs for categorical variables, which showed an imbalance in the outcome variable with more 0s than 1s. To address this, we consolidated four variables into one. For integer variables, QQ plots revealed non-normality, so we standardized the data. A correlation matrix heatmap helped us identify highly correlated variables and mitigate multicollinearity by potentially removing one. E.g., Smoke package year vs. Smoke year.

Feature	Description	Feature	Description
Smokes	whether the individual smokes or not.	Age	The age of the individual.
Hormonal Contraceptives	whether to use hormonal contraceptives.	Sexual partners	The number of sexual partners the individual has had.
IUD	whether uses an intrauterine device.	First sexual intercourse (age)	The age at which the individual had their first sexual intercourse.
STDs	whether has had a sexually transmitted disease.	Pregnancies	The number of pregnancies the individual has had.
STDs:condylomatosis	whether the individual has had condylomatosis.	Hormonal Contraceptives	The number of years the individual has been using hormonal contraceptives.
STDs:cervical condylomatosis	whether the individual has had cervical condylomatosis.	IUD (years)	The number of years the individual has been using an IUD.
STDs:vaginal condylomatosis	whether the individual has had vaginal condylomatosis.	STDs (number)	The number of STDs the individual has had.
STDs:vulvo-perineal condylomatosis	whether the individual has had vulvo-perineal condylomatosis.	STDs: Number of diagnosis	The number of STD diagnoses the individual has received.
STDs:syphilis	whether the individual has had syphilis.	STDs: Time since first diagnosis	The time since the individual's first STD diagnosis.
STDs:pelvic inflammatory disease	whether the individual has had pelvic inflammatory disease.	STDs: Time since last diagnosis	The time since the individual's last STD diagnosis.
STDs:genital herpes	whether the individual has had genital herpes.	Dx:CIN	whether the individual has been diagnosed with CIN.
STDs:molluscum contagiosum	whether the individual has had molluscum contagiosum.	Dx:HPV	whether the individual has been diagnosed with HPV.
STDs:AIDS	whether the individual has AIDS.	Dx	whether the individual has been diagnosed with any disease.
STDs:HIV	whether the individual has HIV.	Hinselmann	whether diagnosed with cervical cancer by Hinselmann.
STDs:Hepatitis B	whether the individual has Hepatitis B.	Schiller	whether diagnosed with cervical cancer based on the Schiller.
STDs:HPV	whether the individual has HPV.	Cytology	whether diagnosed with cervical cancer based on cytology.
Dx:Cancer	whether diagnosed with cervical cancer.	Biopsy	whether diagnosed with cervical cancer based on biopsy.
Smokes (years)	How many years they have been smoking	Smokes (packs)	Indicates individual smoking intensity in packs per year.

4. Approach

4.1 Data Preprocessing

Data preprocessing is a critical step in the machine learning pipeline. First, we applied mean imputation to fill the missing values within continuous variables, such as age, number of sexual partners and number of pregnancies. Mean imputation replaces missing value with the mean value of the non-missing values in the same variable. For categorical variables, we randomly sample values from non-missing values to fill the missing value, which is similar to mode imputation where the most frequent category tends to be selected. After addressing the missing data, we standardized the features to ensure that they were on a comparable scale and contributed equally in subsequent modeling. We then calculated a risk score by summing up four test results (0 or 1), representing the likelihood of cervical cancer. Lastly, we partitioned the preprocessed data into training and testing sets, maintaining an 8:2 ratio, to enable model development and evaluation, respectively.

4.2 Exploratory Modeling

We conducted an exploratory analysis to better understand the underlying patterns and structure of the data, as well as to evaluate the performance of different modeling approaches. Our feature selection method involved the use of LASSO (Least Absolute Shrinkage and Selection Operator) regression, a well-established technique that can effectively handle high-dimensional data with many predictors. Next, we applied the LASSO-selected features to a set of baseline models, including Linear Regression, Random Forest, K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), and Gradient Boosting Machine (GBM), to assess their predictive performance. We observed that all models performed well in predicting whether an individual's risk score exceeded zero (equivalent to being classified as a positive case). However, when examining the predicted values, we found that all observations were classified as negative cases. Additionally, we noted a severe class imbalance issue, with negative cases representing 87.56% of the training set and 82.07% of the testing set. This imbalance can significantly affect the performance of the models and needs to be addressed in the subsequent modeling steps.

4.3 Downsampling

To manage the negative class imbalance, we chose downsampling to adjust the ratio of positive and negative cases, and thus improve the performance of predictive models and ensure that the model is not biased towards the majority class. It is interesting to note that the downsampled data can be viewed as a nested case-control study, with the positive cases comprising the cases and a random subset of negative cases serving as the controls from an epidemiological perspective.

Furthermore, a sensitivity analysis will be conducted to examine the stability of downsampling in order to ensure the robustness of the predictive models. The sensitivity analysis will involve testing the stability of the results, such as the mean squared error (MSE), by varying the degree of downsampling. Specifically, the downsampling ratios of 35%, 45%, 55%, and 65% reduction will be evaluated to assess the consistency of the results across different levels of downsampling. In this study, these downsampling ratios correspond to the positive to negative case ratios of 1:5, 1:4, 1:3, and 1:2, respectively. If the results are consistent across varying degrees of downsampling, then the downsampling process can be considered stable.

4.4 Final Modeling

In this study, we evaluated the impact of different degrees of downsampling on feature selection using LASSO, as well as on the performance of five prediction models: Linear Regression, Random Forest, K-NN, SVM, and GBM. Specifically, we employed four different degrees of downsampling, the positive to negative case ratios of 1:5, 1:4, 1:3, and 1:2, and subsequently evaluated the stability and performance of each model. Through this approach, we aim to identify the optimal degree of downsampling that balances model performance and stability, while mitigating the effects of class imbalance in the dataset.

5. Evaluation

In this section, we present the evaluation results of the machine learning models employed to predict the risk of cervical cancer in women. We used Linear Regression, Random Forest,knn, SVM, and GBM to compare their performance on different datasets with varying case-control ratios (1:2, 1:3, 1:4, and 1:5).

The evaluation metrics used for assessing the performance of the models are MSE, RMSE and MAE, cross-validated Mean Squared Error (CV_MSE), cross-validated Root Mean Squared Error (CV_RMSE), cross-validated Mean Absolute Error (CV_MAE), and Accuracy. Lower values for MSE, RMSE, MAE, CV_MSE, CV_RMSE, and CV_MAE indicate better model performance, while higher values for Accuracy indicate better classification performance.

Based on the line graph: MSE for Each Datasets, the bar chart: Accuracy for Each Model and the table below:

The results indicate that the GBM model consistently outperforms the other models in terms of accuracy, achieving its highest accuracy of 0.835 on the 1:4 case-control ratio dataset. This suggests that the Gradient Boosting Machine is particularly well-suited for this problem, as it can effectively capture complex relationships between features and the target variable while remaining relatively robust to overfitting.

The SVM model also performs reasonably well, with its highest accuracy of 0.804 on the 1:3 case-control ratio dataset. This indicates that the SVM model is able to create a relatively accurate decision boundary for classification.

On the other hand, the Linear Regression model exhibits lower accuracy across all case-control ratios, suggesting that it may not be the best choice for this problem, as the relationships between the features and the target variable may be too complex for a linear model to capture effectively.

The Random Forest and KNN models show moderate performance, with their highest accuracies at 0.691 and 0.742, respectively, on the 1:4 and 1:2 case-control ratio datasets. These models may benefit from further hyperparameter tuning or feature engineering to improve their performance.

The sensitivity analysis of the downsampling process demonstrates a relatively stable performance across varying case-control ratios, indicating that the models' performance is robust and reliable even when the dataset is downsampled.

In conclusion, the Gradient Boosting Machine model exhibits the best performance in predicting the risk of cervical cancer in women across different case-control ratio datasets. Future work may include further tuning of the model's hyperparameters, exploring additional feature engineering techniques, or investigating other advanced machine learning algorithms to further enhance the model's performance and reduce the global burden of cervical cancer.

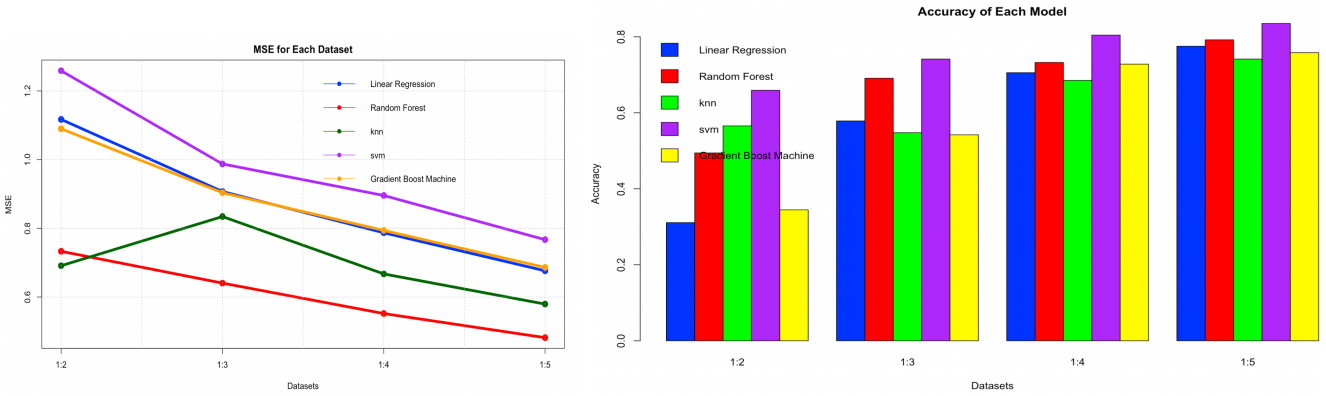

Table 1： Performance Comparison of Five Machine Learning Models using MSE, RMSE, and MAE Metrics

Model	Dataset	MSE	RMSE	MAE	CVS_MSE	CV_RMSE	CV_MAE
linear regression	1:2	1.117	1.057	0.843	0.296	0.15	0.171
	1:3	0.733	0.856	0.661	0.296	0.15	0.171
	1:4	0.691	0.831	0.564	0.296	0.15	0.171
	1:5	1.259	1.122	0.657	0.296	0.15	0.171
Random Forest	1:2	1.09	1.044	0.83	0.215	0.109	0.162
	1:3	0.907	0.952	0.696	0.215	0.109	0.162
	1:4	0.64	0.8	0.573	0.215	0.109	0.162
	1:5	0.834	0.913	0.63	0.215	0.109	0.162
KNN	1:2	0.987	0.994	0.523	0.234	0.123	0.151
	1:3	0.904	0.951	0.693	0.234	0.123	0.151
	1:4	0.787	0.887	0.608	0.234	0.123	0.151
	1:5	0.552	0.743	0.5	0.234	0.123	0.151
SVM	1:2	0.667	0.817	0.508	0.143	0.071	0.135
	1:3	0.896	0.946	0.454	0.143	0.071	0.135
	1:4	0.794	0.891	0.625	0.143	0.071	0.135
	1:5	0.676	0.822	0.525	0.143	0.071	0.135
Gradient Boost Machine	1:2	0.482	0.694	0.435	0.198	0.1	0.1333
	1:3	0.579	0.761	0.446	0.198	0.1	0.1333
	1:4	0.767	0.876	0.39	0.198	0.1	0.1333
	1:5	0.686	0.828	0.533	0.198	0.1	0.1333

6. Discussion and next step

1. Given the skewed distribution, using the mean value and random sampling to deal with missing data may not be ideal, as these methods may not accurately capture the true distribution of the data. In future work, more advanced methods can be explored for imputing missing values to better represent the true data distribution. Techniques such as k-Nearest Neighbors Imputation, Iterative Imputer, or Bayesian methods could be employed to address this issue.
2. Lasso regression may indeed struggle with collinearity problems and may inadvertently eliminate variables that are collinear with other variables. To improve the performance of Lasso for feature selection, you could consider using Elastic Net regularization, which combines the L1 penalty of Lasso with the L2 penalty of Ridge regression. This approach helps mitigate the limitations of Lasso while still encouraging sparsity in the selected features.
3. The use of neural networks and deep learning methods should be considered for future work, as these techniques have demonstrated success in capturing complex relationships and handling large, high-dimensional datasets. These advanced methods may provide even better predictive performance than traditional machine learning models.
4. Dividing the data into more groups based on the case-control ratio can provide a more comprehensive understanding of the models' performance and stability. By creating a curve that illustrates the trend of the data and the stability of the models across varying case-control ratios, we can gain valuable insights into how the models respond to changes in the data distribution and identify potential areas for improvement or further investigation.

Reference

[1] Kashyap N, Krishnan N, Kaur S, Ghai S. Risk Factors of Cervical Cancer: A Case-Control Study. *Asia Pac J Oncol Nurs*. 2019;6(3):308-314. doi:10.4103/apjon.apjon_73_18
[2] Zhang S, Xu H, Zhang L, Qiao Y. Cervical cancer: Epidemiology, risk factors and screening. *Chin J Cancer Res*. 2020;32(6):720-728. doi:10.21147/j.issn.1000-9604.2020.06.05
[3] Fernandes,Kelwin, Cardoso,Jaime & Fernandes,Jessica. (2017). Cervical cancer (Risk Factors). UCI Machine Learning Repository. https://doi.org/10.24432/C5Z310.

Group Work

Collaborate via google folder, zoom,github,overleaf and regular meeting.

Task:	Deadline:	Responsibility:
Data cleaning and preprocessing	March 25,2023	Xinyu,Yingyu
Subset selection and feature engineering	March 25,2023	Xinyu,Jingwei
Baseline model and advanced model	March 31,2023	Xinyu,Yingyu,Jingwei
Data visualization	April 8, 2023	Yingyu,Jingwei
Evaluation and limitation	April 15, 2023	Yingyu,Xinyu
Report writing & Video recording	April 28, 2023	Jingwei,Xinyu,YIngyu

Team evaluations

Jingwei Zhou: 5 points
Yinyu Peng: 5 points
Xinyu Wang: 5 points