Cervical Cancer Prediction Based on Machine Learning Algorithms

Jiangyue Mao LSA Data Science
maojy@umich.edu

STATISTICS
UNIVERSITY OF MICHIGAN

Background

- One of the main fatal diseases that threats women's health is Cervical cancer, which usually does not present any symptoms in early stages. When some symptoms appear, the patients' condition might have been worsened and the cancer may have become metastatic.

- Therefore, early diagnosis of Cervical cancer risk factors can stop its tracks and reduce the mortality rate and the associated complications.

Goal

To seek for more accurate earlier diagnosis, we proposed a study of Cervical cancer diagnosis based on machine learning classification.

Data

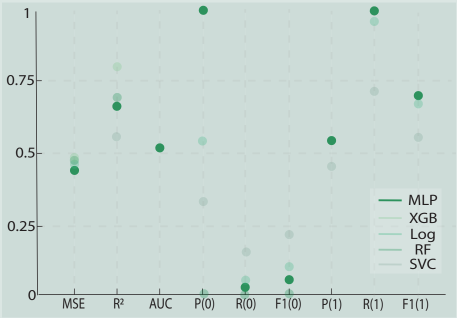
- The Cervical cancer (Risk Factors) Data Set has around 30 variables describing over 800 patients' demographic information, habits, and historic medical records.

- All the binary attributes are labeled as 0 or 1. Instead of having one Cervical cancer variable as the result, it includes four targets (Hinselmann, Schiller, Cytology, and Biopsy). The four targets are the widely used diagnostic test for cervical cancer, and will be used as the measurement for cancer predictions in this project.

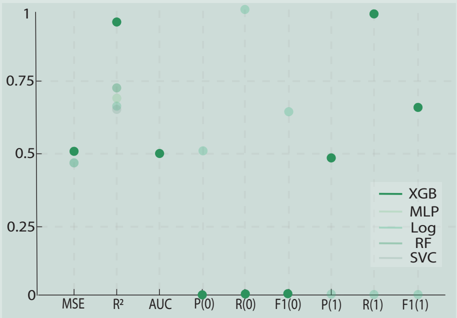
- In addition, there are some missing values due to the patients' unwillingness to disclose their information.

Methods

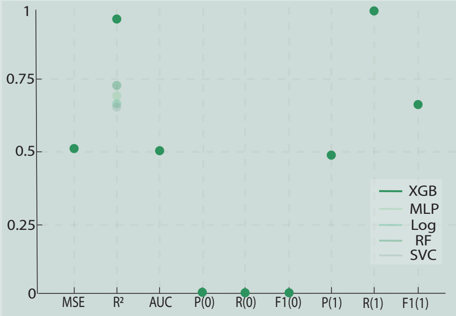
- Chi-squared test between the independent variables and the dependent one
- Data visualization to show feature correlations
- SMOTE method to balance the dataset
- Train-test split
- Support Vector Classifier, Kernel Ridge regression, Logistic Regression, Lasso Regression, Random Forest Claasifier, Gradient Boosting Classifier, and Multilayer Perceptron
- Forward and backward feature selections
- Grid Search on model parameters
- MSE, Cross Validation, R-squared, AUC, Precision, Recall, Accuracy, F1 Score



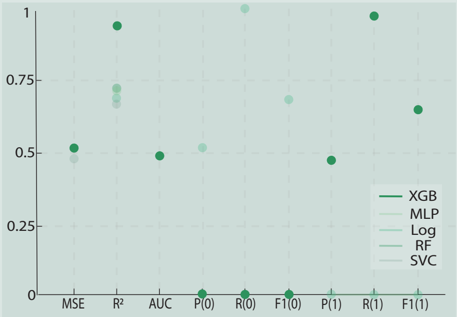
Hinselmann



Schiller



Citology



Biopsy

Results

- Model with best predictability (AUC, Precision, Recall):
Hinselmann: Multilayer Perceptron
Schiller, Citology, and Biopsy: Gradient Boosting

- Common key factors:
Was diagnosed as other cancer or HPV
Number of sexually transmitted diseases (STDs)
Years of using Intrauterine devices (IUDs)

Sometimes also related to age or a specific type of STD such as vulvo-perineal condylomatosis

Conclusions

- Take HPV vaccine

- Take good care of personal life, especially sex life to prevent getting and transmitting sexual diseases

- Prevent wearing IUDs for a long time (several years) and take good precautions during sex

- Disclose more medical information anonymously for data analysis use

Future work

- Further train and test on larger, balanced, and realdatasets, especially on hospital datasets

- Find key features that contribute to the above key factors by matrix minging and machine learning methods and suggest potential precautions

Cervical Cancer Prediction Based on Machine Learning Algorithms

Jiangyue Mao

Abstract

One of the main fatal diseases that threats women's health is Cervical Cancer, which usually does not present any symptoms in early stages. When some symptoms appear, the patients' condition might have been worsened and the cancer may have become metastatic. Therefore, early diagnosis of Cervical cancer risk factors can stop its tracks and reduce the mortality rate and the associated complications. To seek for more accurate earlier diagnosis, we proposed a study of Cervical cancer diagnosis based on machine learning classification using Support Vector Classifier, Kernel Ridge regression, Logistic Regression, Lasso Regression, Random Forest Classifier, Gradient Boosting Classifier, and Multilayer Perceptron. The dataset has around 30 variables describing over 800 patients' demographic information, habits, and historic medical records. It includes four targets (Hinselmann, Schiller, Cytology, and Biopsy). The four targets are used as the measurement for cancer predictions in this project. We also used SMOTE method to handle the imbalanced dataset. Evaluation metrics include Mean Squared Error, R squared, precision, recall, F1 score, and Area Under Curve. Our results show that the Multilayer Perceptron yields the best performance for Hinselmann, and the Gradient Boosting Classifier yields the best performance for the other three targets when compared with other machine learning methods. Future work includes the application of these machine learning methods to a more balanced and real large scale hospital dataset and the inclusion of more information-rich features.

1 Introduction (0.33 points)

Cervical cancer is frequently a fatal disease, common in females but also in males. Cervical cancer is both the fourth-most common type of cancer and the fourth-most common cause of death from cancer in women[1]. Around 70% of Cervical Cancers and 90% of deaths occur in developing countries[1][2]. Especially in low-income countries, Cervical Cancer is one of the most common causes of cancer death[3]. Although vaccines against the prime carcinogenic Human Papilloma Virus (HPV) types are available commercially and are publicized widely, the proportion of women receiving the vaccine is still low, especially in developing countries [4, 5]. In addition, in early stages of the cancer, no typical symptoms are seen[6], and thus the cancer is difficult to detect. Later symptoms may include abnormal vaginal bleeding, pelvic pain or pain during sexual intercourse[6]. Furthermore, despite the effective treatment of early Cervical Cancer with surgery and radiation therapy, late cervical cancer is usually hard to control [7, 8].

Since the survival of the cancer patients largely depends on the malignancy of the cancer cells, accurately forecasting of prognosis would be helpful for estimating the degree of malignancy and the time point of disease progression [9, 10]. Moreover, this can help patients and their families to set appropriate goals based on the accurate survival analysis. The advancement of machine learning makes it possible to predict the possibility of a patient suffering from Cervical Cancer based on patients' previous pathological conditions and not restrained by the lack of medical equipment or funding. We will apply machine learning algorithms to analyze the risk factors in the dataset, aiming to find the factors contributing most to Cervical Cancer and improve the early stage accuracy of diagnosis.

2 Problem Definition

In this project, we will employ seven machine learning algorithms to predict the probability of

Table 1: Model Equations

Model	Parameters	Objective Function
Kernelized SVM	α_i, y_i: label for sample i $\kappa(x_i, x_j) = \psi(x_i)\psi(x_j)$	$\max_{\alpha \in R^n} \left\{ \sum_{j=1}^{n} \alpha_j - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j) \right\}$ s.t. 1) $0 \leq \alpha_j \leq C$ 2) $\sum_{j=1}^{n} \alpha_j y_j = 0$
Kernel Ridge Regression	α, y: label vector i K : Kernel Matrix λ : Regularization Parameters	$\min_{\alpha \in R^n} \frac{1}{2} \Vert y - K\alpha \Vert_2^2 + \frac{\lambda}{2} \alpha^T K \alpha$
Logistic Regression	σ: logistic sigmoid function w: feature weights y_i: label for sample i	$\min_{w} \sum_{j=1}^{n} (\sigma(w_0 + w_1^T x_j) - y_j)^2$ $\sigma(u) = \frac{1}{1 + \exp{(-u)}}$
Lasso Linear Regression	w: feature weights λ: regularization parameters y_i: label of sample i	$\min_{w} \sum_{j=1}^{n} \left(y_i - w^T x_j \right)^2 + \lambda \Vert w \Vert$
Random Forest	pP: proportion of class 1 in parent node pL: proportion of class 1 in left child pR: proportion of class 1 in right child q: proportion of data in parent node that goes to left child H_p: $p \cdot log_2 p - (1-p) \cdot log_2(1-p)$	$IG = \ H(pP) \ - [qH(pL) + (1-q)H(pR)]$
Multilayer Perceptron	$(e_j)^2$: $y_j - y_j(v_j)$ $y_j(v_j)$: $\frac{1}{(1+exp(-v_j))}$ y_j: label of sample j	$\min_{vi} \frac{1}{2} \sum_{j=1} (e_j)^2$
Gradient Boosting Classifier	$L(y,t)$: e^{-yt} F_+: $\{\sum_{t=1}^{T} f_t \mid T \geq 1, f_t \in F\}$ y_j: label of sample x_j	$\min_{F \in F_+} \frac{1}{n} \sum_{i=1}^{n} L(y_i, F(x_i))$

the patients suffering from Cervical Cancer using sociodemographic factors and their pathological features and medical conditions. The conceptual and math notations are displayed in Table 1.

While relevant mathematical equations and variables for all models used are shown in Table

Figure 1: Dataset description

Feature type	Feature name	Hinselmann		Schiller		Cytology		Biopsy	
		Mean (μ) ± Std (σ)		Mean (μ) ± Std (σ)		Mean (μ) ± Std (σ)		Mean (μ) ± Std (σ)	
		Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Demographic	Age	26.7 ± 7.7	26.8 ± 8.5	29.6 ± 11	26.6 ± 8.2	26.2 ± 8.4	26.9 ± 8.5	28.6 ± 8.9	26.7 ± 8.5
	Smokes	0.2 ± 0.4	0.1 ± 0.4	0.2 ± 0.4	0.2 ± 0.3	0.1 ± 0.3	0.1 ± 0.4	0.2 ± 0.4	0.1 ± 0.3
	Smokes (years)	2.5 ± 7.2	1.1 ± 3.9	2.4 ± 6.2	1.1 ± 3.8	1.1 ± 3.4	1.2 ± 4.1	2.2 ± 6.2	1.1 ± 3.9
	Smokes (packs/year)	0.7 ± 2.6	0.4 ± 2.2	0.6 ± 1.9	0.4 ± 2.2	0.5 ± 2.3	0.4 ± 2.2	0.7 ± 2.3	0.4 ± 2.2
	Number of sexual partners	2.2 ± 0.9	2.5 ± 1.7	2.5 ± 1.2	2.5 ± 1.7	2.7 ± 1.3	2.5 ± 1.7	2.5 ± 1.3	2.5 ± 1.7
	First sexual intercourse (age)	16.8 ± 2.0	16.9 ± 2.8	17 ± 2.5	16.9 ± 2.8	16.9 ± 2.9	16.9 ± 2.8	17.1 ± 2.6	16.9 ± 2.8
	Number of pregnancies	2.4 ± 1.4	2.5 ± 1.7	2.6 ± 1.7	2.2 ± 1.4	2.1 ± 1.4	2.2 ± 1.4	2.3 ± 1.3	2.2 ± 1.4
	Hormonal contraceptives	0.7 ± 4.5	0.7 ± 0.5	0.6 ± 0.5	0.7 ± 0.5	0.7 ± 0.5	0.7 ± 0.5	0.7 ± 0.5	0.7 ± 0.5
Habit	Hormonal contraceptives (years)	2.9 ± 4.8	1.9 ± 3.5	3.2 ± 5.2	1.9 ± 3.4	3.3 ± 6.4	1.9 ± 3.4	3.3 ± 5.4	1.9 ± 3.4
	IUD	0.2 ± 0.4	0.1 ± 0.3	0.2 ± 0.4	0.1 ± 0.3	0.1 ± 0.3	0.1 ± 0.3	0.2 ± 0.4	0.1 ± 0.3
	IUD (years)	0.6 ± 1.5	0.4 ± 1.8	0.9 ± 2.9	0.4 ± 1.7	0.5 ± 1.7	0.4 ± 1.8	0.7 ± 2.0	0.4 ± 1.8
	STDs	0.2 ± 0.4	0.1 ± 0.3	0.2 ± 0.4	0.1 ± 0.3	0.2 ± 0.4	0.2 ± 0.3	0.2 ± 0.4	0.1 ± 0.3
	STDs (number)	0.3 ± 0.9	0.2 ± 0.5	0.4 ± 0.8	0.1 ± 0.5	0.3 ± 0.7	0.1 ± 0.5	0.3 ± 0.8	0.1 ± 0.5
	STDs: condylomatosis	0.1 ± 0.3	0.1 ± 0.2	0.1 ± 0.3	0.0 ± 0.2	0.1 ± 0.3	0.0 ± 0.2	0.1 ± 0.3	0 ± 0.2
	STDs: cervical condylomatosis	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	STDs: vaginal condylomatosis	0 ± 0	0.0 ± 0.1	0 ± 0	0 ± 0.1	0 ± 0	0 ± 0.1	0 ± 0	0 ± 0.1
History	STDs: vulvo-perineal condylomatosis	0.1 ± 0.3	0.0 ± 0.2	0.1 ± 0.3	0 ± 0.2	0.1 ± 0.3	0 ± 0.2	0.1 ± 0.4	0 ± 0.1
	STDs: syphilis	0.0 ± 0.2	0.0 ± 0.1	0.0 ± 0.2	0 ± 0.1	0 ± 0	0 ± 0.1	0 ± 0	0 ± 0.1
	STDs: pelvic inflammatory disease	0 ± 0	0.0 ± 0.0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	STDs: genital herpes	0 ± 0	0.0 ± 0.0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0.1	0 ± 0
	STDs: molluscum contagiosum	0 ± 0	0.0 ± 0.0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	STDs: AIDS	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	STDs: HIV	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1
	STDs: Hepatitis B	0 ± 0	0.0 ± 0.0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	STDs: HPV	0 ± 0	0.0 ± 0.0	0 ± 0	0 ± 0	0 ± 0	0 ± 0.1	0 ± 0	0 ± 0.1
Genomics	Dx: CIN	0 ± 0	0.0 ± 0.1	0.0 ± 0.1	0 ± 0.1	0 ± 0	0 ± 0.1	0.1 ± 0.2	0 ± 0.1
	Dx: HPV	0.1 ± 0.3	0.0 ± 0.1	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1
	Dx	0.1 ± 0.3	0.0 ± 0.2	0.1 ± 0.3	0 ± 0.2	0.1 ± 0.3	0 ± 0.2	0.1 ± 0.3	0 ± 0.1
	Dx: cancer	0.1 ± 0.3	0.0 ± 0.1	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1

1, the external library, sklearn, was imported and used for actual implementation due to its prior optimization with respect to the Python coding language in addition to the availability of built-in evaluation functions. In addition, we will measure the interpretability and predictability of the models and labels with or without Cervical Cancer by Mean Squared Error, R^2, Area Under Curve, precision, recall, and F1 score. This will be further elaborated in the Evaluation section.

3 Data

The Cervical Cancer (Risk Factors) Data Set comes from UCI machine learning repository, with around 30 variables describing over 800 patients' sociodemographic information, habits, and historic medical records. For example, it has age, smokes(years), number of sexual partners, type of diagnosed Sexually Transmitted Disease (STD)s, type of diagnosed other diseases such as other cancer or HPV(Dx:Cancer, Dx:HPV), and years of wearing Intrauterine devices (IUD(years)) etc. as variables, and all the binary attributes are labeled as 0 or 1. Instead of having one Cervical Cancer variable as the result, it includes four targets (Hinselmann, Schiller, Cytology, and Biopsy). The four targets are the widely used diagnostic tests for Cervical Cancer, and will be used as the dependent variables for cancer predictions in this project. In addition, there are some missing values due to the patients' unwillingness to disclose their information. The mean and standard deviations of the variables are illustrated in Figure 1 which comes from Irfan Ullah Khan and other authors' paper[10] that uses the same dataset.

4 Related Work

Similar or the same dataset has been employed in several papers, which have different focuses, ranging from the biomedical side of Cervical Cancer to machine learning outcomes and algorithm optimizations. The paper by Y. M. S. Al-Wesabi, Avishek Choudhury, and Daehan Won uses the same dataset[9] and applies Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN), Decision Trees (DT), Logistic Regression (LR), and Support Vector Machine (SVM) to predict patients' probability of suffering from Cervical Cancer and applies Wrapper methods such as Sequential Feature Selector to perform feature selection. The six features selected by this algorithm are Age, First Sexual Intercourse, Number of Pregnancies, Smokes, Hormonal Contraceptives and STDs:genital herpes. The methods that are used to measure classifiers' performance include computing accuracy, sensitivity, specificity, precision, and F statistics. Decision Tree yields the best performance measured by these metrics.

Another paper named Cervical Cancer Diagnosis Model by Irfan Ullah Khan and other authors also uses the same dataset, which applies Extreme Boosting Classifier in addition to the machine learning classifiers that were mentioned in the previous paper such as KNN, XGBoost, Random Forest (RF), etc., and the paper focuses more on the theoretical side of the classifiers such as the inferences of the parameters and formulas and the authors use Firefly Optimization to select variables. This paper uses the same evaluation metrics as the previous one, but it finds Extreme Gradient Boosting to be the most classifier with the best performance.

The third paper named Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods by Xiaoyu Deng, Yan Luo, and Cong Wang employing the same dataset uses similar machine learning techniques and evaluation methods but it further clarifies the SMOTE method to handle the imbalanced data[11].

The above three papers similarly processed the imbalanced data. However, with considerable more negative Cervical Cancer cases in the dataset than positive ones, the predictability of the models may not improve significantly as the positive Cervical Cancer patients' demographic and medical information is mostly randomly generated by making copies. And after balancing the dataset most independent variables still remain unbalanced. Furthermore, other machine learning methods such as Kernel Ridge Regression, Lasso Regression, and Multilayer Perceptron have yet to be thoroughly applied to this dataset to possibly find a model with better predictability and interpretability.

5 Methodology

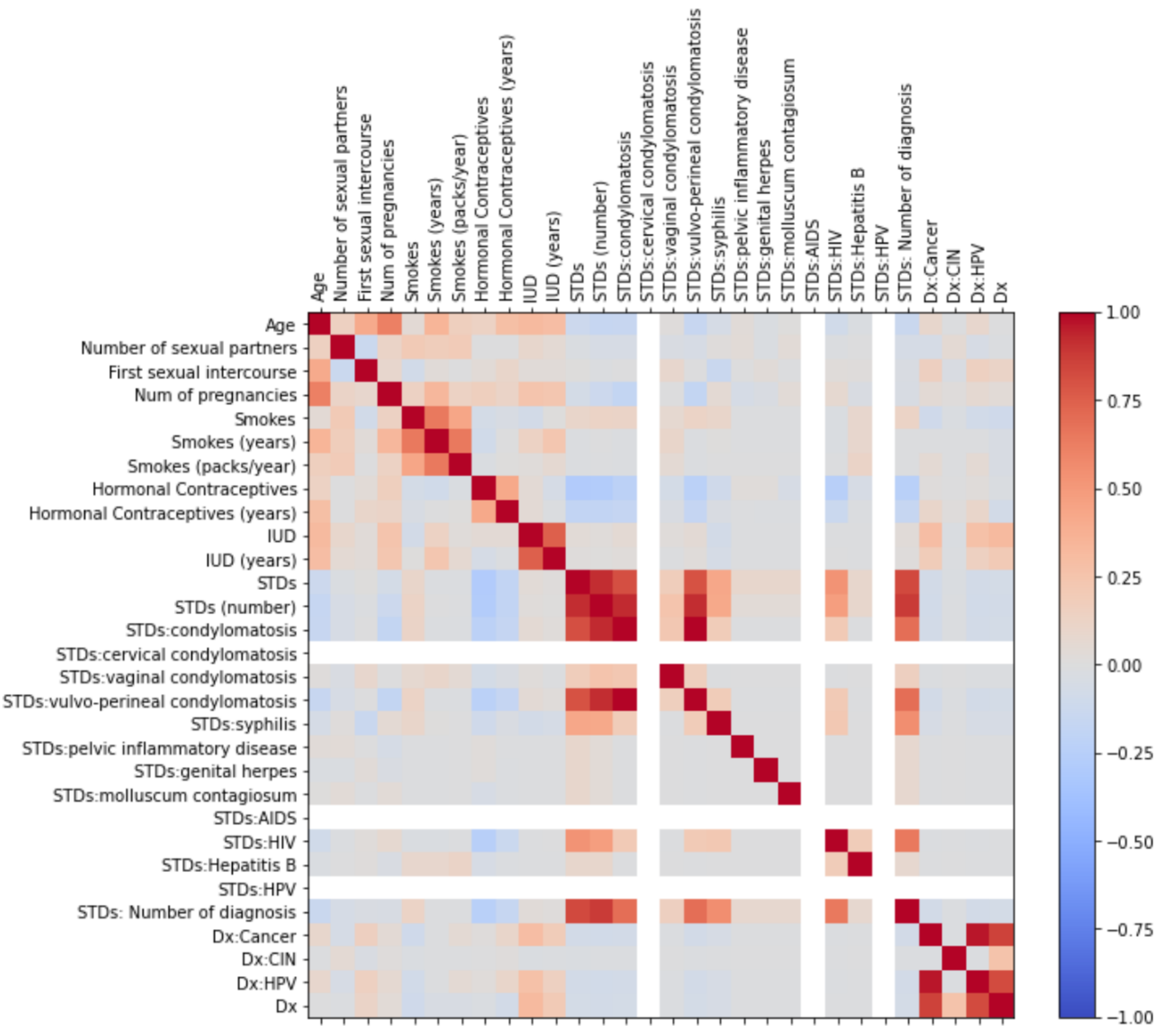
5.1 Data Preparation

Apart from dropping the missing values from the dataset and columns that have more than 20% of missing values, we perform Chi-squared Test to see the correlations among the variables. The correlations are visualized in Figure 2. From this figure we can see that Smokes, Smoke(years), and Smokes(packs/years), STDs, STDs(number), and STDs:condylomatosis, Dx:Cancer, Dx, and DX:HPV are highly correlated and we should try to avoid including all these variables in the same model or only include a few of them together if necessary.

Furthermore, by the Chi-squared Test between the features and the target variables we can primarily select the ones with the highest correlations with the target variables. For Hinselmann, the most corrected variables are IUD (years), STDs (number), STDs: Number of Diagnosis, Dx:Cancer, Dx:HPV, and Dx. For Schiller, the variables are Age, Num of Pregnancies, Hormonal Contraceptives (years), IUD, IUD (years), STDs, STDs (number), STDs:condylomatosis, STDs:vulvo-perineal condylomatosis, STDs:HIV, STDs: Number of Diagnosis, Dx:Cancer, and Dx:HPV. For Citology, variables are STDs:HIV, Dx:Cancer, Dx:HPV, and Dx. For Biopsy, the variables are Age, Hormonal Contraceptives (years), STDs, STDs (number), STDs:condylomatosis, STDs:vulvo-perineal condylomatosis, STDs: Number of Diagnosis, Dx:Cancer, Dx:HPV, and Dx. Since these variables are highly correlated with the targets and do not heavily overlap with other highly correlated variables, we will fit all these features into the model and perform feature selections during the model fitting process.

In addition, we need to figure out the ways to handle the imbalanced dataset. One way is to oversample the minority class. This can be achieved by simply making copies of the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model. A possible improvement on this is to synthesize new examples from the minority class. This can be very effective as this provides new information to the dataset. Synthetic Minority Oversampling Technique (SMOTE) fulfills this requirement by selecting examples that are close in the feature space, drawing a line between them and generating a new sample at a point along that line. By applying SMOTE, we can see the Mean Squared Error increased due to the increase of the posi-

Figure 2: Feature Correlation



tive sample size and the variance of patients' information, but this may help to produce a better prediction and help our model to be more generalizable. We also employed several other methods to deal with the imbalanced dataset such as Random Over-Sampling and Random over-sampling with imblearn(a Python package), but they do not produce a lower Mean Squared Error so we decide to keep using SMOTE as cited in other papers. The number of true positive samples of the four target variables each increase from around 30 to 638 after applying the SMOTE method.

5.2 Classification Methods

The classification task was accomplished by fitting each model with all selected variables for each target variable. Since there are four target variables and seven models for each target, the results of the models used to perform classification are listed in Table 1. All models used except the Random Forest are kernelized. This means model weights do not correspond to specific features, rather they correspond to the features resulting from the kernel function.

In order to perform model selection, both classification and regression models were used, in addition to the DT. For each model, forward and backward feature selections are applied. However, the MSE even becomes higher after fitting the models with the selected variables by forward and backward feature selections. Therefore, we continue to use the variables based on the Chi-squared results. Furthermore, a grid search using cross validation was applied to each model to select the hyperparameters that yield the lowest Mean Squared Error.

All dataset processing and machine learning algorithms were implemented in Python. Dataset preprocessing was implemented using the pandas and numpy libraries. Machine learning models and grid search using cross validation were implemented using the sklearn library. The sklearn library simply implements machine learning models by providing a model initialization function to produce a model object. After the cross validation process, the best Regularization parameter is selected to 10, and Gaussian Kernel was selected for the SVM and Kernel Ridge Regression models for all target variables. The best Regularization Parameters for the Logistic Regression models range from 0.1 to 1, and L2 penalty has the best performance for all target variables. In terms of MLP for Hinselmann and Schiller, the selected alphas are around 10^{-3}, and for the rest two targets they are around 10^{-6}, and the best hidden layer size is the default one (100). Furthermore, the maximum depth of the Random Forest Classifier is around 10 for all target variables, and the number of estimators is 10 for Hinselmann and Citology, and 50 for Schiller, and 100 for Biopsy. In addition, for Lasso Regression, the selected alpha is 0.1 and the maximum iteration is 100 for all targets except for Biopsy which is 1000.

After initialization, train and predict member functions for each model object can be called to apply data to the training process and generate classification probabilities for all test samples.

6 Evaluation and Results

6.1 Evaluation Metrics

Since our target is to predict patients' positivity and negativity of Hinselmann, Schiller, Cytology, and Biopsy, we apply evaluation metrics to each model based on their prediction for the positive (labeled 1) and negative (labeled as 0) target variables to observe the interpretability and predictability of each model. Evaluation metrics include Mean Squared Error (MSE), Area Under Curve(AUC), R^2, precision, recall, and F1 score. All these factors indicate how well the model can predict the positivities, and they are the metrics of predictability.

MSE measures the mean squared difference between the predicted values and the actual value. $MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. The lower the MSE is, the closer the predicted labels are to the true labels.

The True Positive Rate (TPR) represents the diseased patients who received positive results on the test and the False Positive Rate (FPR) represents the proportion of true negatives which are misclassified as positives. The receiver operating characteristic (ROC) curve is the TPR plotted against the FPR. If the model is perfect at detecting true positives, then FPR will always be 0 and TPR will always be 1. This will produce a curve that is upper triangular and the area under the curve (AUC) will be equal to 1, which is the ideal case. Typically the AUC is less than one, so if the AUC is close to 1, the model is very good at discrimination.

R-squared (R^2) is a statistical measure that represents the proportion of the variance for the dependent variable that is explained by the independent variables in the model. If the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs. $R^2 = 1 - \frac{UnexplainedVariation}{TotalVariation}$.

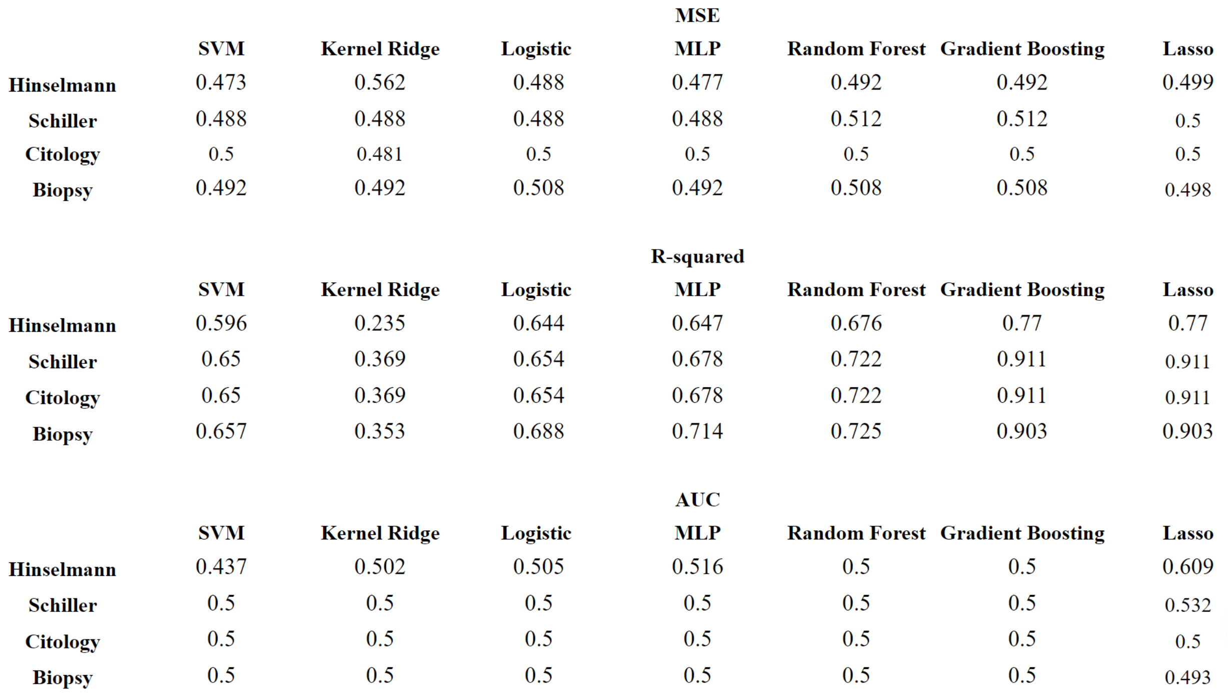
The precision is calculated as $\frac{tp}{tp+fp}$ where tp is the number of true positives and fp is the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative[12].

The recall is the ratio $\frac{tp}{tp+fn}$ where tp is the number of true positives and fn is the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples[12].

The F1 score can be interpreted as a weighted harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0 [12]. $F1 = 2 * \frac{precision*recall}{precision+recall}$.

In addition, the probability of randomly choosing the sample to be positive or negative is 0.5, and the corresponding MSE, AUC, R^2, precision, and recall are also 0.5, and F1 score is 0.25. If we always pick the majority class in the original dataset, then we will always classify the patients as negative on all tests since the dataset is extremely imbalanced. Although this has high accuracy, this does not make sense since we will classify all positive patients as negative and our predictions will be meaningless. But if we pick the majority class after using SMOTE to balance the dataset, this will have the same effect as randomly determining the patients as positive or negative on the tests since the number of people diagnosed as positive or negative on the tests is equal. Thus, the probability of choosing the sample to be test positive or negative is also 0.5, and other metrics except F1 score will also be 0.5, with F1 score being 0.25. Therefore, we can set 0.5 as the baseline for MSE, AUC,

Figure 3: MSE, R-squared and AUC results



				MSE			
	SVM	Kernel Ridge	Logistic	MLP	Random Forest	Gradient Boosting	Lasso
Hinselmann	0.473	0.562	0.488	0.477	0.492	0.492	0.499
Schiller	0.488	0.488	0.488	0.488	0.512	0.512	0.5
Citology	0.5	0.481	0.5	0.5	0.5	0.5	0.5
Biopsy	0.492	0.492	0.508	0.492	0.508	0.508	0.498

				R-squared			
	SVM	Kernel Ridge	Logistic	MLP	Random Forest	Gradient Boosting	Lasso
Hinselmann	0.596	0.235	0.644	0.647	0.676	0.77	0.77
Schiller	0.65	0.369	0.654	0.678	0.722	0.911	0.911
Citology	0.65	0.369	0.654	0.678	0.722	0.911	0.911
Biopsy	0.657	0.353	0.688	0.714	0.725	0.903	0.903

				AUC			
	SVM	Kernel Ridge	Logistic	MLP	Random Forest	Gradient Boosting	Lasso
Hinselmann	0.437	0.502	0.505	0.516	0.5	0.5	0.609
Schiller	0.5	0.5	0.5	0.5	0.5	0.5	0.532
Citology	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Biopsy	0.5	0.5	0.5	0.5	0.5	0.5	0.493

R^2, precision, and recall, and 0.25 as the baseline for F1 score, and we will compare our model predictability with these baselines.

6.2 Results

We use the above six metrics to measure the performance of the models, the results are in Figure 3. MLP has a much better (around 0.2 higher) precision and recall for the positive Hinselmann test compared with other models, and MLP also has a relatively low MSE and high AUC. Therefore, we choose MLP to be the model with the best performance for Hinselmann. We do not include precision, recall, and F1 score in the table because their results for both positive and negative cases of the other three target variables are very close and are nearly undifferentiable. The precision, recall, and F1 score for the rest three positive tests are around 0.5, 1, and 0.7 respectively, and the precision, recall, and F1 score for the rest three negative tests are close to 0.

From Figure 3 and the above precision, recall and F1 score, we conclude that the model yield best performance is Gradient Boosting for Schiller, Citology, and Biopsy because its MSE and AUC are close to that of other models and it has a much higher R^2 (approximately 0.2 to 0.3) compared with other models. The reason why we do not choose Lasso to be the final model for Schiller is that although it is around 0.3 higher in terms of AUC compared with that of Gradient Boosting, and it has exactly the same R^2 compared with that of Gradient Boosting, it is around 0.4 higher in terms of MSE. Although the MSE and AUC of the selected models are close to the baselines, their R^2, recall, and F1 score for the positive tests are significantly higher than that of the baseline. so we consider the models as better predictions than randomly choosing the test results to be positive. Despite the fact that the models are worse in terms of predicting negative results compared with the baselines, misclassifying negative outcomes as positive ones is better than its inverse since it will not misguide the patients to miss the best treatment period. Thus, we should focus more on the predictability of the positive tests (ie. the precision, recall, and F1 score for label 1). The key factors that we find in the variable selection process (Chi-squared Test) that are common among all targets are: Diagnosed as other cancer or HPV or not, Number of Sexually Transmitted Diseases (STDs), Years of Using Intrauterine devices (IUDs). Some targets are also related to age or a specific type of STD such as Vulvo-Perineal Condylomatosis.

7 Discussion

From this study, we suggest people, especially women, take HPV vaccine, take good care of personal life, especially sex life to prevent suffering from and transmitting sexual diseases, prevent wearing IUDs for a long time (several years) and take good precautions during sexual intercourse, and disclose more medical and sociodemographic information anonymously for data analysis use.

Although our model results show relatively low test error, high R^2 and AUC for all test cases, and high precision, recall, and F1 score for positive test cases, there are other factors we did not consider such as educational level and annual household income, etc. Besides, our sample size is relatively small and imbalanced, further training and testing on larger, balanced, and real datasets, especially on hospital datasets will be ideal. After applying the selected models to the proposed dataset, it is possible that this study to be used in future Cervical Cancer prediction analysis and improve the accuracy of early diagnostics. In addition, if time permits, we could find the key features that contribute to the above key factors by matrix mining and machine learning methods and suggest potential precautions.

References

[1] *World Cancer Report 2014.* World Health Organization. 2014. pp. Chapter 5.12. ISBN 978-9283204299.

[2] "Cervical cancer prevention and control saves lives in the Republic of Korea". *World Health Organization.*

[3] World Health Organization (February 2014). "Fact sheet No. 297: Cancer".

[4] Lei J, Ploner A, Elfström KM, Wang J, Roth A, Fang F, et al. *HPV vaccination and the risk of invasive cervical cancer.* N Engl J Med. 2020;383:1340–8.

[5] Williams EA, Newberg J, Williams KJ, Montesion M, Alexander BM, Lin DI, et al. *Prevalence of High-Risk nonvaccine human papillomavirus types in advanced squamous cell carcinoma among individuals of african vs Non-African ancestry.* JAMA Netw Open. 2021;4:e216481.

[6] "Cervical Cancer Treatment (PDQ®)". *NCI.* 14 March 2014.

[7] Huang H, Feng YL, Wan T, Zhang YN, Cao XP, Huang YW, et al. *Effectiveness of sequential chemoradiation vs concur-rent chemoradiation or radiation alone in adjuvant treatment after hysterectomy for cervical cancer: the STARS phase 3 randomized clinical trial.* JAMA Oncol. 2021;7:361–9.

[8] Arbyn M, Weiderpass E, Bruni L, de Sanjosé S, Saraiya M, Ferlay J, et al. *Estimates of incidence and mortality of cervi-cal cancer in 2018: a worldwide analysis.* Lancet Glob Health. 2020;8:e191-203.

[9] Al-Wesabi Y.M.S., Choudhury A: Classification of Cervical Cancer Dataset. In: Proceedings of the 2018 IISE Annual Conference. Edited by K. Barker, D. Berry, C. Rainwater: 2018; Orlando: IISE; 2018:1456-1461.

[10] Khan, I.U., Aslam, N., Alshehri, R., Alzahrani, S., Alghamdi, M., Almalki, A., Balabeed, M., 2021. *Cervical Cancer Diagnosis Model Using Extreme Gradient Boosting and Bioinspired Firefly Optimization.* Scientific Programming 2021, 1–10. doi:10.1155/2021/5540024.

[11] Deng, X., Luo, Y., Wang, C., 2018. *Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods.* doi:10.1109/ccis.2018.8691126.

[12] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.