Machine Learning project

Davide Ghilardi, Lorenzo Lobosco, Fabio Salerno

Abstract

Each year cervical cancer kills about 4,000 women in the United States and about 300,000 women worldwide, even though being the most preventable type of cancer. The purpose of the project is to investigate the risk factors that lead to cervical cancer and evaluate the effectiveness of screening tests used to diagnose it. The study consists of a supervised classification analysis through several machine learning models, to predict the outcome of a biopsy test. Furthermore, evaluation criteria were used to choose the best performing model. Finally, possible future developments are proposed to improve the accuracy and robustness of the predictions.

Keywords: Machine Learning; Classification; Cervical; Cancer

Contents

1. Introduction

In 2020, an estimated 604.000 women were diagnosed with cervical cancer worldwide and about 342.000 women died from the disease [5]. Each year in the United States, about 13.000 new cases of cervical cancer are diagnosed and about 4.000 women die of this cancer [7]. Cervical cancer is cancer that starts in the cells of the cervix. The cervix is the lower, narrow end of the uterus (womb). The cervix connects the uterus to the vagina (birth canal). Cervical cancer usually develops slowly over time. Before cancer appears in the cervix, the cells of the cervix go through changes known as dysplasia, in which abnormal cells begin to appear in the cervical tissue. Over time, if not destroyed or removed, the abnormal cells may become cancer cells and start to grow and spread more deeply into the cervix and to surrounding areas [2]. Cervical cancer was once one of the most common causes of cancer death for American women. The cervical cancer death rate dropped significantly with the increased use of the Pap test (also called Cytology test). This screening procedure can find changes in the cervix before cancer develops. It can also find cervical cancer at early stages, so it is easier to be cured [10]. In the United States, cervical cancer mortality rates abated by 74% from 1955 - 1992 thanks to increased screening and early detection [10]. The main cause of cervical cancer is persistent infection with high-risk types of human papillomavirus (HPV), an extremely common family of viruses that are transmitted through sexual contact; women most at risk for cervical cancer are those with a history of multiple sexual partners, sexual intercourse at age 17 years or younger, or both [10, 5]. The age of the women needs also to be taken into account, about 50% of cervical cancer diagnoses occur in women ages 35 - 54, and about 20% occur in women over 65 years of age. There are many other factors that can impact on the cervical cancer diagnosis such as: family history, use of oral contraceptives, the number of children birthed, smoking and also socio and economic factors, numerous studies report that high poverty levels are linked with low screening rates [10, 9]. This project was born with the goal of investigating the potential risk factors that can impact on the cervical cancer diagnosis and also to investigate how commonly used screening tests such as the Hinselmann test, the Schiller test and the PAP test perform on cancer diagnosis. The goal is trying to predict the result of a biopsy test employing different classification techniques specific to Machine Learning, the results of which were then compared in order to identify the best performing model that can potentially be used by the medical industry.

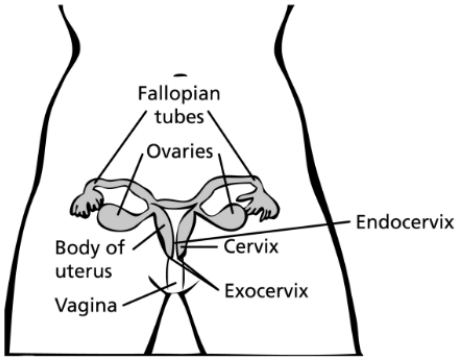

Figure 1 Cervical cancer related body parts.

2. Dataset

The dataset used for our analyzes contains a list of risk factors for cervical cancer leading to a biopsy examination. It is composed of 858 records, each referring to a different patient, and 36 variables, including the response variable. the dataset's variables are the following:

- **Age**: the age of the patient
- **Number** of sexual partners: the number of sexual partners of the patient
- **First sexual intercourse**: age of first sexual intercourse of the subject
- **Num of pregnancies**: number of pregnancies of the subject
- **Smokes**: dichotomous variable that tells us whether the subject smokes: 1 yes, 0 no
- **Smokes(years)**: how many years the subject has been smoking
- **Smokes(packs/year)**: is used to describe how many cigarettes the subject has smoked in his lifetime *"It is calculated by multiplying the number of packs of cigarettes smoked per day by the number of years the person has smoked. For example, 1 pack-year is equal to smoking 1 pack per day for 1 year, or 2 packs per day for half a year, and so on"*.
- **Hormonal Contraceptives**: dichotomous variable that tells us whether the subject takes hormonal contraceptives: 1 yes, 0 no.
- **Hormonal Contraceptives(years)**: how many years the subject has been taking hormonal contraceptives.
- **IUD**: dichotomous variable that tells us whether the subject uses The intrauterine device: 1 yes, 0 no. *"The intrauterine device, IUD, or coil, is a small, birth-control device that is inserted into the uterus to prevent unintended pregnancy for up to 5 years"*.
- **IUD (years)**: how many years the subject has been using The intrauterine device.
- **STDs**: the subject has had sexually transmitted diseases, 1 yes 0 no. *"Sexually transmitted diseases (STDs), also known as sexually transmitted infections (STIs)"*.
- **STDs(number)**: number of STDs experienced by the patient.
- **STDs condylomatosis**: whether the subject has been infected by condylomatosis.
- **STDs cervical condylomatosis**: whether the subject has been infected by cervical condylomatosis.
- **STDs vaginal condylomatosis**: whether the subject has been infected by vaginal condylomatosis.
- **STDs vulvo-perineal condylomatosis**: whether the subject has been infected by vulvo-perineal condylomatosis.
- **STDs syphilis**: whether the subject has been infected by syphilis condylomatosis.
- **STDs pelvic inflammatory disease**: whether the subject has been infected by pelvic inflammatory disease.
- **STDs genital herpes**: whether the subject has been infected by genital herpes.
- **STDs molluscum contagiosum**: whether the subject has been infected by molluscum contagiosum.
- **STDs AIDS**: whether the subject has been infected by AIDS.
- **STDs HIV**: whether the subject has been infected by HIV.
- **STDs Hepatitis B**: whether the subject has been infected by Hepatitis B.
- **STDs HPV**: whether the subject has been infected by HPV.
- **STDs Number of diagnosis**: number of diagnoses made for sexually transmitted diseases
- **STDs Time since first diagnosis**: years since the first diagnosis
- **STDs Time since last diagnosis**: years since the last diagnosis
- **Dx Cancer**: past cancer diagnosis,1 yes, 0 no.
- **Dx CIN**: diagnosis of CIN, 1 yes 0 no. *"Cervical intraepithelial neoplasia (CIN), also known as cervical dysplasia, is the abnormal growth of cells on the surface of the cervix that could potentially lead to cervical cancer"*
- **Dx HPV**: diagnosis on the presence of HPV(The human papillomavirus). 1 yes, 0 no. *"A positive test result means that you have a type of high-risk HPV that's linked to cervical cancer. It doesn't mean that you have cervical cancer now, but it's a warning sign that cervical cancer could develop in the future [9]."*
- **Dx**: dx is an abbreviation for analysis, if the client has been diagnosed.
- **Hinselmann**: Hinselmann is a dichotomous variable that refers to the test colposcopy. 1 if positive, 0 if negative. A colposcopy is a type of cervical cancer test. It lets your doctor or nurse get a close-up look at your cervix — the opening to your uterus. It's used to find abnormal cells in your cervix [6].
- **Schiller**: Schiller is a dichotomous variable that refers to the cervical cancer test in which iodine is applied to the cervix. The iodine colors healthy cells brown; abnormal cells remain unstained, usually appearing white or yellow. 1 if positive, 0 if negative [1].
- **Cytology**: Cytology test also called PAP test is the exam of a single cell type, as often found in fluid specimens. 1 if tested positive, 0 if negative [4].
- **Biopsy**: A cervical biopsy is a procedure to remove tissue from the cervix to test for abnormal or precancerous conditions, or cervical cancer. This is the target variable of our classification. Through the biopsy it is possible to diagnose with certainty whether the subject has cervical cancer or not, thus we use the other variables of the dataset to obtain a model that is able to predict the outcome of the biopsy [3].

A. imbalanced class

Through exploratory data analysis, it was noticed that the number of patients affected by cancer is lower than the number of healthy ones (7% Biopsy=1, 93% Biopsy=0), so the class attribute is imbalanced. This aspect is very important to keep in mind when doing classification, so the methods where adapted in order to deal with the imbalanced class.

3. Preprocessing

The preprocessing stage is used to process the raw input data and transform it with the purpose of simplifying the analysis. Were implemented the following steps:

- Preliminary feature selection
- Handling of missing value
- Train-test split

A. Preliminary feature selection

Variables were analyzed to find redundant or useless ones for the analysis. It was found out that some attributes have a high amount of missing values (92%) as very sensitive data to which most subjects did not respond, such as "STDs: Time since first diagnosis" and "STDs: Time since last diagnosis" and so it was decided to discard them to leaning the dataset. Features, such as "Smokes" contained redundant information already contained in others ones, such as "Smokes (years)", so it was decided to discard some of them to avoid multicollinearity. The variables excluded from the analysis at this stage were the following: “Smokes”; “Hormonal Contraceptives"; “IUD“; “STDs“; “STDs (number)”; “STDs: Number of diagnosis“; “STDs: Time since first diagnosis“; “STDs: Time since last diagnosis“; “Dx“.

B. Handling missing values

The dataset has some missing values due to the fact that several patients decided not to answer some of the questions because of privacy concerns [9]. About 12% of the patients interviewed decided to not disclose information about their sexually transmitted diseases those information are stored in 14 columns of the dataset; considering the loss of those information and the fact that of 105 patients which not disclosed STDs information, just two had positive value of the response, it was decided to delete those records in order to avoid distortions in the dataset due to wrong imputation. After this operation the remaining dataset was made of 753 records and the features with still NAs were: Number of sexual partners (1.8%), First sexual intercourse (0.8%), Number of pregnancies (6.2%), Smokes (year) (1.3%), Smokes (packs/year) (1.3%), Hormonal Contraceptives (years) (1.7%) and IUD (years) (2.1%). Taking out the variable “Number of pregnancies”, there was a relatively low number of NAs, so it was decided to handle the remaining missing values which were all numerical features, by replacing them with the mean value. Taking a look at the distribution of those features there wasn’t a significant presence of outliers so it was decided to use the mean instead of the median.

C. Train test split

To assure that predictions weren’t biased by training data, causing overfitting, it was decided to split the dataset into train and test partitions. In particular, the first included 80% of the data, while the second the remaining 20%. Then, the classifiers were trained with specific methods only on the train set. Finally, the best models for each research question were chosen to be evaluated on the test partition.

4. Models

A. Naïve Bayes

Naïve Bayes is a supervised learning algorithm that is particularly suited for classification tasks, and falls in the category of Bayesian models. It tries to estimate the conditional probability P(Y=1 | X=x) of the output given the inputs exploiting the Bayes theorem, which says that:

$$P(y=1|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y=1)P(y=1)}{P(\boldsymbol{x}|y=1)P(y=1) + P(\boldsymbol{x}|y=0)P(y=0)}.$$

Naïve Bayes model is a simple yet effective model, it scales fast, and it performs well on imbalanced datasets.

B. Random Forest

Random Forest (RF) is a supervised algorithm that is used for classification and regression tasks. It’s an ensemble model, which means that the final prediction is made by aggregating, by majority voting, the ones of individual trees. RF model is able to reduce overfitting and improve robustness, since every tree is trained on a random subset of data. It can handle large and high-dimensional datasets, it is relatively fast to train, and it is also relatively easy to interpret, as the importance of each feature can be determined by examining the individual trees in the model. RF model has many hyperparameters that can be adjusted to improve its performance; in particular, we have to define:

- The number of trees; a larger number of trees will usually result in a better model, but at the cost of increased runtime and memory usage.
- The maximum depth of the individual decision trees in the forest. A deeper tree can model more complex relationships between the features and the target variable, but it is also more prone to overfitting.
- The minimum number of samples required to split an internal node in a decision tree. A larger value will result in fewer splits, which can make the model more robust but also less able to capture fine-grained patterns in the data.
- The minimum number of samples required to be at a leaf node in a decision tree. A larger value will result in fewer leaf nodes, which, as before, can make the model more robust but at the same time unable to recognize fine-grained patterns in the data.
- The number of features that are considered when splitting a node in a decision tree. A larger value will allow the tree to consider more features, which can make the model more powerful but also more prone to overfitting.

Other parameters exists to fine-tune the learning algorithm.

C. Logistic Regression

Logistic regression is a linear algorithm that tries to learn a function that maps input features $\boldsymbol{x}$ to the probability of a binary outcome y. Mathematically, the function is parameterized by a weight vector $\boldsymbol{w}$ and a bias term b:

$$P(y=1|\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{w}, b)$$

The function f is called the logistic function, and it is defined as:

$$f(\boldsymbol{x}; \boldsymbol{w}, b) = \frac{e^{\boldsymbol{w}\boldsymbol{x}+b}}{1+e^{\boldsymbol{w}\boldsymbol{x}+b}}$$

It’s well-suited for modeling probabilities since it compress any real-valued number to unit interval. Training a logistic regression model involves maximizing the likelihood function, which is defined as the probability of the parameters given the data:

$$\mathcal{L}(\boldsymbol{w}, b) = P(y|\boldsymbol{x}, \boldsymbol{w}, b)$$

In practice, this is done by applying the IRLS optimization algorithm to the negative log-likelihood function. The predictions of a LR model can be interpreted as the probability that the example belongs to the positive class. Although logistic regression is very simple, it widely used for its explainability. In fact, the linearity makes it easy to understand the influence of each feature on the prediction.

D. SVMs

Support Vector Machines (SVMs) are a supervised learning model that, in classification tasks, tries to find an hyperplane $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$ that maximally separates the positive and negative classes. Often, observations aren't perfectly divided, so the optimization problem can be relaxed to allow some misclassified instances. This solution is called "soft margin" and involves the solution of

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b))$$

In the sum above, the first term tries to maximize the margin between the classes. The second term tries to minimize the number of misclassified examples by penalizing any training examples that are on the wrong side of the decision boundary. The regularization parameter C determines the trade-off between these two goals. Predictions are given by the sign of the projection of the point on the hyperplane. SVMs can handle high-dimensional data efficiently, thanks to the use of the kernel trick, a technique that allows to find nonlinear decision boundaries by projecting the data into a infinite-dimensional space using a kernel function. Common kernel functions include the linear kernel, the polynomial kernel, and the radial basis function (RBF) kernel.

E. ANNs

An artificial neural network (ANN) is a popular machine learning model inspired by the structure and function of the biological neural networks that make up the brain. ANNs are composed of layers of interconnected "neurons," which process and transmit information. Each neuron, also called perceprton, receives input (x_i) from some number of other neurons, and weights them to output (y) a signal to other neurons in the next layer. In particular the perceprton computes the following formula:

$$y_i = f(\sum_{i=1}^{n} w_i x_i - \theta)$$

There, w_i is the weight associated with the i-th input, θ is a bias term, and f is the activation function which scales the sum in a specific interval. Popular activation functions are the sigmoid, the hyperbolic tangent, and the ReLU. ANNs are trained with backpropagation, an algorithm that exploits gradient descent to adjust the weights to minimize the error between the output of the ANN and the desired one. The process starts from the connections between the last layer and the output node and proceeds backwards toward the input layer.

F. AdaBoost

AdaBoost (short for Adaptive Boosting) is a meta-algorithm (can improve the performance of a base learning algorithm) that falls in the category of ensemble methods. AdaBoost works by weighting the observations in the training data such that the weak learners focus more on the observations that were misclassified by the previous weak learner. The weak learners are then combined using a majority vote weighted on the accuracy they reached. To implement the Ada Boosting algorithm it was used the RealAdaBoost node applied to a Decision Stump, an algorithm which exploits decision trees with one split level.

G. Cross validation

Each model was evaluated using a 10 fold cross validation method on the training partition. The dataset was splitted in ten subsets where each fold was disjoint, exhaustive and with a constant number of records. To account for imbalanced class, each fold was created using stratified sampling. The models were then trained ten times, using, at each iteration, nine folds as training set and the remaining fold as test set. This was done in order to have a more robust estimation of the performance measures. This procedure was developed through the X-partitioner and X-aggregator nodes, and for comparison purposes it was important to be sure that each model uses the same partitions and in order to guarantee that a random seed was selected and used in every model.

H. Feature selection

To better understand the value added by the screening tests (Hinselmann, Schiller, and Cytology) to the performances of risk factors in order to predict the biopsy result, it was decided to implement a feature selection. In particular, by using the CfsSubsetEval method of the Weka AttributeSelectedClassifier node, the dimensionality of the training set was reduced by evaluating the individual attributes before submitting them to the classifier. Through the forward BestFirst search method, which searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility, the features that most influence the response variable were selected without neglecting the correlation between them.

I. Performance metrics

For the analysis, several criteria were used to assess performance. Since the response class variable suffered from an imbalanced class, the computation of the accuracy as a performance measure wasn't suitable. This is due to the fact that it considers every class equally important; however, in an imbalanced dataset, the rare class is considered more interesting than the majority one. It therefore becomes appropriate to use other evaluation measures that are better suited to take into account the imbalanced class problem. In particular the following were used: Recall, Precision, F_1-score. Recall, also called True Positive Rate, is defined as

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

and represents the portion of positive records (Biopsy=1) correctly classified by the model (TP, true positive). A high Recall value indicates a low portion of incorrectly classified positive records (FN, false negatives).

Precision, on the other hand, describes the fraction of records that are actually positive among all those predicted as such by the classifier. It's defined as

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

and a high value indicates a low portion of incorrectly classified negative records (FP, false positives). These two measures compete with each other, the more recall you get the less precision you have and vice versa. It's important to assess which

error has the priority: the number of false positives or the number of false negatives. In this case, preventing the mistake of a cancer patient classified as healthy is more important than mistaking a healthy person as a cancer patient, so the priority must be to minimize the amount of false negatives produced by the classifier. The F_1-score is a measure that summarize Precision and Recall by computing their harmonic mean

$$F_1 = \frac{2 * Recall * Precision}{Recall + Precision}. \quad (3)$$

It takes value ([0,1]) and a high value of F_1 ensures that both recall and precision are reasonably high. A further method for evaluating the classification model without taking the class distribution and error costs into account is the representation of the ROC curve, which relates the percentage of false positives (FPR, false positives rate) with the percentage of true positives (TPR, true positives rate). Once the ROC curve is defined, the Area Under the Curve (AUC) can be computed; it's the measure of the ability of a classifier to distinguish between classes. When AUC = 1, then the classifier is able to perfectly distinguish between all the positive and the negative class points correctly. When 0.5<AUC<1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values. When AUC=0.5, then the classifier is not able to distinguish between positive and negative class points.

J. Cost-sensitivity

To improve the Recall score were implemented cost-sensitive classifiers. They are trained by assuming that errors made are unequal. In this case, more importance was given to false negatives than to false positives. That choice was made because of the assumption that the consequences of predicting a patient as healthy, when in reality it is not, are far more worse than the opposite. To implement cost-based classification, WEKA's CostSensitive Classifier node was used and set to minimize the expected misclassification cost (EMC), a metric that measures the cost of a misclassification. EMC is computed as follows

$$EMC(CM, C) = C_{TP} \cdot P(TP) + C_{FP} \cdot P(FP) + \\ + C_{FN} \cdot P(FN) + C_{TN} \cdot P(TN) \quad (4)$$

In the equation above, CM is the confusion matrix, while C is the cost matrix, which gives the cost of all the 4 possible outcomes: TP, FP, TN, FN. C was defined as

Biopsy/prediction	**0**	**1**
0	0	1
1	20	-1

So the importance of producing a FN is considered 20 times greater than the one of producing a FP. Moreover, since it's fundamental that the algorithm produces TPs (even more than TNs), they were made able to reduce the cost by 1.

5. Results

A. Risk factors

The following table summarizes the results of the model trained using an equally weighted cost-matrix, giving the same importance to false positives and false negatives.

Models	Recall	Precision	F_1	AUC
AdaBoost	0 (0)	0 (0)	0 (0)	0.647 (0.07)
Random Forest	0.097 (0.104)	0.6 (0.548)	0.164 (0.169)	0.681 (0.053)
Logistic	0.047 (0.065)	0.267 (0.435)	0.076 (0.105)	0.579 (0.076)
SVM	0 (0)	0 (0)	0 (0)	0.5 (0)
ANN	0.094 (0.13)	0.267 (0.435)	0.133 (0.189)	0.557 (0.159)
Naïve Bayes	0.186 (0.145)	0.162 (0.084)	0.17 (0.108)	0.6 (0.077)

Table 1 Performance indicators for the chosen models and (standard deviations) with an equally weighted cost-matrix, risk factors.

At a first glance, we can see that the results aren't satisfactory. In fact, all recall scores don't exceed 0.20, which means that among all patients with a positive biopsy test, less than 20% of them are classified as at risk. Moreover metrics' standard deviations are really high for certain models, so predictions aren't stable. On the other hand, AdaBoost and Random Forest AUCs arrive nearly to 70%, a fairly good value, and also their standard deviations are pretty low. Overall, since Naïve Bayes has the best recall score and fair F_1 score and AUC, it has been taken as the best model for this group. The second set of models has been trained with the cost-matrix defined before. The results are shown in the table below

Models	Recall	Precision	F_1	AUC
AdaBoost	0.74 (0.129)	0.103 (0.017)	0.18 (0.028)	0.625 (0.066)
Random Forest	0.22 (0.219)	0.357 (0.356)	0.257 (0.241)	0.682 (0.153)
Logistic	0.29 (0.188)	0.16 (0.107)	0.202 (0.129)	0.573 (0.132)
SVM	0.29 (0.188)	0.217 (0.197)	0.238 (0.175)	0.599 (0.103)
ANN	0.33 (0.274)	0.248 (0.178)	0.257 (0.17)	0.594 (0.184)
Naïve Bayes	0.355 (0.224)	0.148 (0.108)	0.205 (0.139)	0.594 (0.12)

Table 2 Performance indicators for the chosen models and (standard deviations) with the defined cost-matrix, risk factors.

We can observe that recall scores grow significantly, with AdaBoost reaching a value of 74%; however, standard deviations remain high. F_1 scores improve for all models while AUC values remain roughly the same (except for SVM). AdaBoost and Naïve Bayes were chosen as the best models to examine on the test set.

Models	Recall	Precision	F_1	AUC
Naïve Bayes	0.182	0.25	0.211	0.692
AdaBoost(cost)	0.545	0.125	0.203	0.623
Naïve Bayes(cost)	0.364	0.167	0.229	0.67

Table 3 Performance indicators for the chosen models on the test partition, risk factors.

The table above collects the results of the selected models on the test set. We can see that Naïve Bayes and Naïve Bayes (cost) maintained the same performances in the test set. On the other hand, AdaBoost seemed to overfit, in particular its recall value decreased consistently from train to test set.

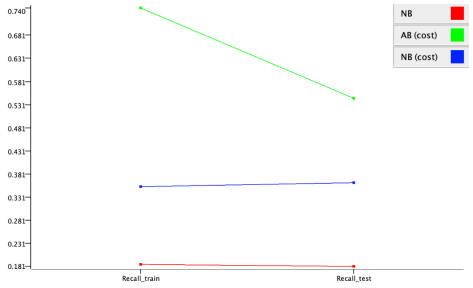

Figure 2 Line plot of the recall obtained in the training and test partitions, risk factors.

Nevertheless, AdaBoost reached the best recall score (54.5%), but the worst precision (12.5%), so it generates a low number of false negatives at the price of producing a big one of false positives. As we could imagine, Naïve Bayes trained with a symmetric cost-matrix gives the best precision (25%), the best accuracy (90.1%), and also the best AUC (69.2%), while the worst recall (18.2%). F_1 score is similar for all models. Even though Naïve Bayes (non-cost) exceeds by 10 to 20 percentage points all the other models in precision, accuracy, and AUC scores, we consider AdaBoost the best model as it reaches the best recall score, and a fairly good AUC (62.3%).

B. Risk factors and screening tests

After analyzing the risk factors alone, it was tried to jointly analyze the risk factors in combination with the results of the three screening tests present in the dataset: Hinselmann, Schiller and Cytology to assess how their introduction in the models would affect the prediction of a biopsy test. The inspiration that led to conducting this type of in-depth study on screening tests was an article which analysed whether the use of Schiller's test and Pap smear could increase detection rate of cervical dysplasias [8]. The following table summarizes the results of the model trained using an equally weighted cost-matrix.

Models	Recall	Precision	F_1	AUC
AdaBoost	0.5 (0.268)	0.678 (0.335)	0.532 (0.241)	0.962 (0.039)
Random Forest	0.625 (0.247)	0.728 (0.164)	0.641 (0.177)	0.948 (0.066)
Logistic	0.6 (0.181)	0.648 (0.177)	0.591 (0.127)	0.934 (0.093)
SVM	0.81 (0.182)	0.637 (0.124)	0.701 (0.123)	0.886 (0.09)
ANN	0.55 (0.235)	0.638 (0.246)	0.553 (0.166)	0.916 (0.085)
Naïve Bayes	0.645 (0.203)	0.457 (0.158)	0.527 (0.164)	0.865 (0.107)

Table 4 Performance indicators for the chosen models and (standard deviations) with an equally weighted cost-matrix, risk factors+screening tests.

The model with the best Recall value is the SVM reaching a value of 81% followed by the Naïve Bayes reaching a value of 64.5%. Taking into account the F_1 score, the SVM (70.1%) has the best score and the second best is achieved by the Random Forest (64.1%), so they reached a fair balance between Recall and Precision. The highest AUC is achieved by the Random Forest (94.8%). However, metrics' standard deviations are really high for almost all the models, so predictions aren't stable. Considering the remaining models, all of them have a relatively high value of AUC. Finally, it was decided to take the SVM as the best for this group and to try it on the test set. The results of the second set of models, the ones trained with the cost matrix, are shown in the table below.

Models	Recall	Precision	F_1	AUC
AdaBoost	0.825 (0.237)	0.585 (0.153)	0.676 (0.177)	0.962 (0.039)
Random Forest	0.735 (0.268)	0.683 (0.166)	0.665 (0.159)	0.949 (0.065)
Logistic	0.855 (0.172)	0.563 (0.113)	0.672 (0.118)	0.931 (0.091)
SVM	0.925 (0.169)	0.632 (0.143)	0.744 (0.142)	0.941 (0.086)
ANN	0.765 (0.193)	0.544 (0.128)	0.62 (0.107)	0.943 (0.066)
Naïve Bayes	0.76 (0.205)	0.377 (0.113)	0.497 (0.134)	0.83 (0.103)

Table 5 Performance indicators for the chosen models and (standard deviations) with the defined cost-matrix, risk factors+screening tests.

With the cost-sensitive approach, the mean Recall scores of all models have improved. In particular, the best score was still the SVM's (92.5%). All the F_1 scores have also improved except for the Naïve Bayes. However, other than AUC, metrics' standard deviations remained still high for almost all the models. AUC remained roughly the same for almost all models, except for

SVM which significantly improved it. Finally, ADA and the SVM were chosen as the best model to examine on the test set.

Models	Recall	Precision	F_1	AUC
SVM	0.909	0.714	0.8	0.94
AdaBoost(cost)	0.909	0.714	0.8	0.941
SVM(cost)	0.909	0.667	0.769	0.937

Table 6 Performance indicators for the chosen models on the test partition, risk factors+screening tests.

The table above collects the results of the selected models on the test set. We can see that all models achieved exactly the same recall score (90.9%), which is modest. Also, other metrics almost converge to the same values.

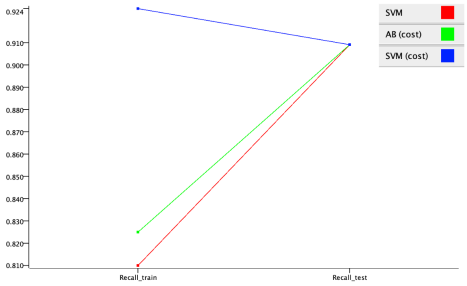

Figure 3 Line plot of the recall obtained in the training and test partitions, risk factors+screening tests.

Moreover, it can be said that models don't overfit since all the metrics don't decrease from the train to the test set.

Feature selection chose the following attributes: STDssyphilis, STDsgenitalherpes, STDsHIV, DxCIN, DxHPV, Schiller.

Models	Recall	Precision	F_1	AUC
SVM(filter)	0.909	0.714	0.8	0.94
AdaBoost(cost+filter)	0.909	0.667	0.769	0.939
SVM(cost+filter)	0.909	0.667	0.769	0.937

Table 7 Performance indicators for the chosen models with the subset of feature chosen on the test partition, risk factors+screening test.

The models with the subset of features chosen by the filter, reached performances close to the ones of the models trained with all attributes.

6. Conclusions

The aim of the first research question was to understand how well risk factors are able to predict the outcome of a biopsy test. Initially the results were not very satisfactory, the performance of each model was rather low and specifically low Recall values were obtained, so the model assigned as "healthy" several patients who actually resulted positive to the biopsy test. Through cost-sensitive models, which introduced a cost matrix assigning a higher cost to the FN error, the performance improved but not enough to achieve a good result. Although it seemed to overfit, AdaBoost was the model which achieved the best combination of recall and AUC. The second research question aimed at understanding the level to which information on screening tests (Hinselmann, Schiller and Cytology), analysed in conjunction with risk factors, impacted in predicting the outcome of a biopsy test. There, the results were much more satisfying, all models were able to achieve similar and better levels of Recall than the previous ones with just the risk factors. Finally, comparing the results obtained on the training set with those obtained on the test set, no major differences were found. Furthermore, the feature selection models helped to better understand which were the most relevant features in determining the outcome of a biopsy test. In particular, their results seemed coherent with the domain analysis. In fact, the attributes related to Cervical Dysplasia (Dx:CIN) and to Human Papilloma Virus (HPV) were chosen. These are the main causes of cervical cancers and so it seems right that they should receive particular attention in the predictions. Schiller was the only test included by the filter. A possible reason for this is that the three tests may be correlated and considered redundant by the filter. Overall, the feature selection didn't negatively impact the model's performances and helped to make the results more interpretable. Finally, these results are suggesting the importance of screening tests in order to early detect cervical cancer. Further analysis should be done on the following topics: First, since the poor results obtained by using just the risk factors could be due to the fact that the dataset does not contain a sufficient number of observations, the access to a bigger dataset may improve the models' results and robustness. Second, domain experts should be interviewed during both the analysis, and the interpretations of the results. They would have allowed us to give answers to important assumptions such as the choice of the cost-matrix coefficients, to better understand the impact of classification errors, adjust, and optimize the models' parameters, and the significance of the subset of features selected by the filter. In conclusion, it has to be said that machine learning models can assist experts in cancer detection, and thus become an important asset for the medical industry.

References

[1] National Cancer Insitute. *Schiller test*. 2023. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/schiller-test.

[2] National Cancer Insitute. *What Is Cervical Cancer?* 2022. URL: https://www.cancer.gov/types/cervical#:~:text=Cervical%20cancer%20is%20cancer%20that,usually%20develops%20slowly%20over%20time.

[3] Johns Hopkins Medicine. *Cervical Biopsy*. 2023. URL: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/cervical-biopsy#:~:text=What%20is%20a%20cervical%20biopsy,that%20opens%20into%20the%20vagina..

[4] Johns Hopkins Medicine. *Cytology*. 2023. URL: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/cytology#:~:text=Cytology%20is%20the%20exam%20of,other%20screening%20and%20diagnostic%20areas..

[5] World Health Organization. *Cervical Cancer Awareness Month 2022*. 2022. URL: https://www.iarc.who.int/infographics/cervical-cancer-awareness-month-2022/.

[6] Planned Paretnthood. *What is a colposcopy?* 2023. URL: https://www.plannedparenthood.org/learn/cancer/cervical-cancer/what-colposcopy.

[7] Centers for desease control prevention. *Cervical Cancer Statistics*. 2022. URL: https://www.cdc.gov/cancer/cervical/statistics/index.htm#:~:text=Each%20year%20in%20the%20United,of%20dying%20from%20cervical%20cancer.

[8] A. A. Khazi Ramaraju H. E. Nagaveni Y. C. "Use of Schiller's test versus Pap smear to increase detection rate of cervical dysplasias". In: (2016).

[9] American Cancer Society. *Cervical Cancer Risk Classification*. URL: https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification.

[10] American Cancer Society. *Key Statistics for Cervical Cancer*. 2023. URL: https://www.cancer.org/cancer/cervical-cancer/about/key-statistics.html#:~:text=The%20American%20Cancer%20Society%27s%20estimates,will%20die%20from%20cervical%20cancer.