UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

DATA AND TEXT MINING

FINAL PROJECT, JUNE 2020

Cervical cancer risk exploratory study

Author:
Antonio Vivace - 793509
a.vivace1@campus.unimib.it

***Abstract*—Cervical Cancer is one of the most treatable cancers when diagnosed at early stages. Yet, it killed 311000 women in 2018 and it's still the fourth cause of death from cancers in women, the second most common in developing areas, mainly because of the economic cost and the difficulties in implementing effectives screening programmes. Data Mining provides robust tools to verify the known causal relations and assess risk factors from medical datasets. Classification models can then exploit this scenario to help identifying groups of population at higer risk to improve planning of screening programmes.**

I. INTRODUCTION

This year in the US, there will be an estimated 13800 new cases of Cervical cancer. 4290 will be deadly [1]. Worldwide, this type of cancer causes makes up 8% of the total cancer cases and deaths. It's the fourth (second in developing countries [2]) cause of death from cancer in women. A lot of progress has been made, and nowadays it is one of the most curable: when diagnosed and treated at the earliest stages, the five-year survival rate can reach 95% [3].

In this work, we investigate the possibility of classifying high-risk patients from demographic and other medical data, but excluding the results of other related exams.

Known causal relations: HPV causal relation with the cervical cancer has been documented beyond reasonable doubt [4]. Additionally, Human papilloma virus (HPV) infection is necessary for the development of CIN, the abnormal growth of cells on the surface of the cervix, indicating a potentially precancerous transformation of cells of the cervix [5].

Risk factors include smoking, [6] early age at the first sexual intercourse and early pregnancies.

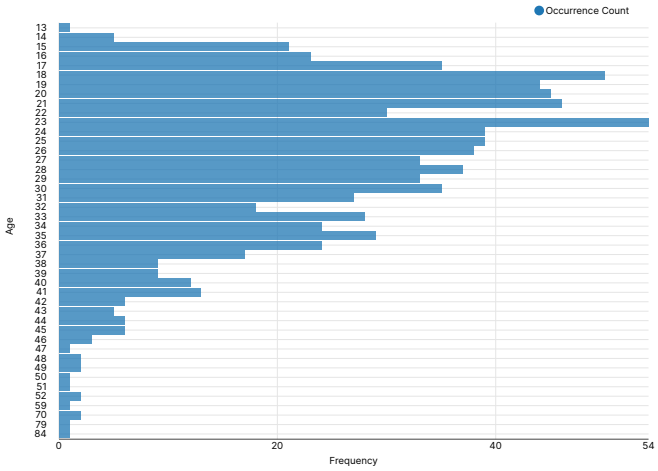

Figure 1. Age distribution

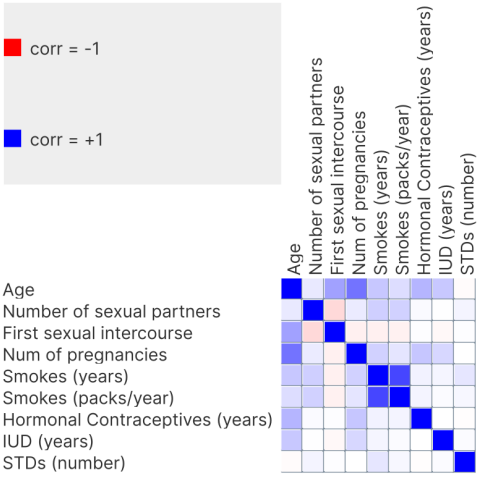

Figure 2. Correlation Matrix on the numeric attributes of the starting dataset

II. DATASET

The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela.

36 attributes describe demographic informations, sexual activity and frequency, pregnancies, age, smoke addiction, usage of contraceptives and diagnosed Sexualy Transmitted Diseases of 858 women patients.

The missing values are due to the decision of the patients to not disclose those details because of privacy concerns.

We'll introduce some context and clarifications on the attributes significance to outline the relationship with the disease:

- **Hormonal Contraceptives** indicates if the patient uses *hormonal* contraceptives. **IUD** indicates the usage of the Intra Uterine contraceptive Device.
- **STDs** indicates if the patient had at least one Sexually Transmitted Disease. It's positive when at least one of the STDs:condylomatosis, STDs:cervical condylomatosis, STDs:vaginal condylomatosis, STDs:vulvo-perineal condylomatosis, STDs:syphilis, STDs:pelvic inflammatory disease, STDs:genital herpes, STDs:molluscum contagiosum, STDs:AIDS, STDs:HIV, STDs:Hepatitis, STDs:HPV attributes is positive.
- **STDs:HPV** indicates if the patient is infected by the *Human Papilloma Virus*.
- **Dx:HPV** reveals if the patient was already diagnosed with HPV in the past.
- **Dx:Cancer** is positive if the patient already had cancer. It is not clear if this refers to Breast Cancer, as diagnosable by the Oncotype DX test or any type of cancer. The original paper presenting this dataset has no additional information on this [7].
- **Dx:CIN** indicates if the patient was affected by Cervical intraepithelial neoplasia (CIN).
- **Dx** is positive if at least one of the Dx variables is positive.
- **Hinselmann** and **Citology** are two (complementary [8]) tests to detect Cervical Cancer. **Schiller** is another test able to identify abnormal areas to be biopsied and examined.
- **Biopsy** is a medical procedure providing a confirmation of the diagnosis through a *colposcopy*, a magnified visual inspection of the cervix. This attribute has been chosen as our target.

III. The Methodological Approach

A. Exploration

The target class is highly unbalanced (Figure 4), with only 6.4% of positive values.

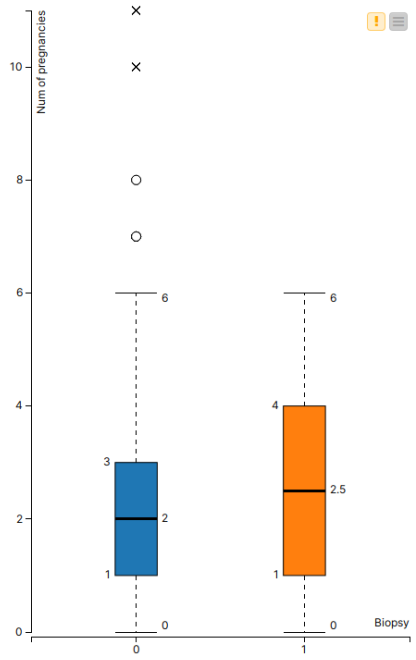

Figure 3. Biopsy VS number of pregnancies conditional box plot

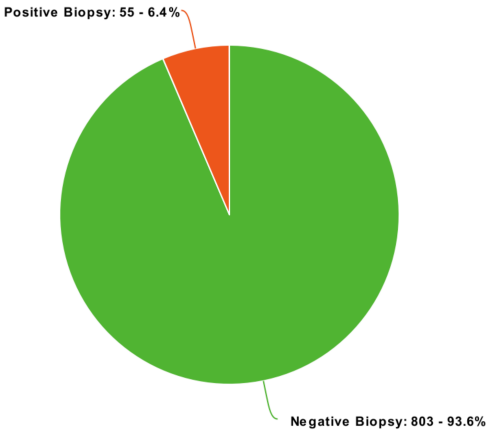

Figure 4. Unbalanced biopsy distribution

The main statistical descriptors have been computed and studied. Age distribution is shown in Figure 1: the majority of the instances refers to young women (younger than 15 and older than 38 years old can be considered outliers). Almost all of them had at least one pregnancy (98,1%).

B. Feature Selection

We decided to exclude the other tests (**Hinselmann**, **Citology**, **Schiller**) since, by conception, they share a similar purpose with the target **Biopsy**.

STDs:Time since first diagnosis and **STDs:Time since last diagnosis** are missing

in more than 90% of the records and were filtered out.

STDs:cervical condylomatosis and **STDs:AIDS**, **STDs:pelvic inflammatory disease**, **STDs: molluscum contagiosum**, **STDs:Hepatitis B** and **STDs:HPV** were removed because they had more than 99.995% negative (or missing) values.

Finally, **Smokes**, **Hormonal Contraceptives** and **IUD** were removed because the correspondings "Years" attributes already provide these informations.

C. Missing Values

Missing values have been imputed with the conditioned average value per Age class. 5 Age classes with same frequency have been created, using the *Equal frequency unsupervised discretization method*: 13-19, 19-23, 23,28, 28-34, 34-84.

Finally, to avoid inv the Z-Score Normalization has been applied.

After this steps, the dataset has 20 normalized attributes and no missing values:

- Age
- Number of sexual partners
- First sexual intercourse
- Num of pregnancies
- Smokes (years)
- Smokes (packs/year)
- Hormonal Contraceptives (years)
- IUD (years)
- STDs (number)
- STDs:condylomatosis
- STDs:vaginal condylomatosis
- STDs:vulvo-perineal condylomatosis
- STDs:syphilis
- STDs:genital herpes
- STDs:HIV
- STDs: Number of diagnosis
- Dx:Cancer
- Dx:CIN
- Dx:HPV
- Biopsy

D. Correlation

Some mentioned causal relations can be seen through this step: early age at first sexual intercourse, number of pregnancies.

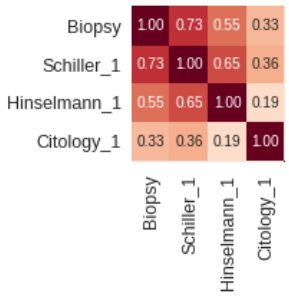

Figure 5. Correlation matrix of Cervical Cancer exams

		Ground Truth	
		Positive	Negative
Prediction	Positive	1.0	1.0
	Negative	10.0	1.0

Table I
PROPOSED COST MATRIX

Dx is not clearly correlated with Biopsy, probably because the type of cancer it refers to it's unrelated (Breast).

The other tests (**Hinselmann**, **Citology**, **Schiller**) are obviously very correlated with Biopsy (Figure 2).

Even if it's been shown that smoking it's a risk factor for CIN, the potential initial step of a Cervical Cancer, this is not visibile in this dataset.

E. Class imbalance

Two approaches have been experimented to handle this issue: oversampling with SMOTE ([9]) and the usage of cost-sensitive Random Forest classificator.

To avoid inflating metrics, SMOTE oversampling has been applied only on the training sets.

The cost-sensitive model makes use of the cost matrix shown in Table III, to try to minimize the False Negative Rate. The cost given to False Negatives (ignoring when a cancer is present) is 10 times the cost of False Positives (flagging a cancer even when it's not true).

F. Classification

The following models have been tested, using a 5-Fold stratified cross validation learning: Random

Attribute	Importance
First sexual intercourse	2.5964373464373462
Hormonal Contraceptives (years)	2.126984126984127
Number of sexual partners	1.593911719939117
Age	1.196969696969697
Num of pregnancies	1.0419717887154862
STDs (number)	0.8303292827102351
Dx:HPV	0.8073291050035236
Dx:Cancer	0.720280437756498
STDs: Number of diagnosis	0.4703452178834182
Smokes (years)	0.38181818181818183
Smokes (packs/year)	0.37965816755393667
IUD (years)	0.3663967611336032
STDs:HIV	0.35227272727272724
STDs:vulvo-perineal condylomatosis	0.28707870787078704
STDs:condylomatosis	0.2577920377160817
STDs:syphilis	0.152191894127378
Dx:CIN	0.14195402298850573
STDs:vaginal condylomatosis	0.08712121212121213
STDs:genital herpes	0.05130718954248366

Table II
FEATURE IMPORTANCE, ACCORDING TO THE RANDOM FOREST (SMOTE) MODEL (ALL LEVELS).

Forest and Logistic Regression. The features used are the selected one, while the split dataset (1/3, 2/3) and oversampled ones are being (separately) used (SMOTE and cost matrix).

Overview of the different configuration tried:

1) Logistic Regression with non-oversampled dataset (split 2/3 and 1/3)
2) Logistic Regression with SMOTE
3) Logistic Regression with Cost Matrix
4) Random Forest with non-oversampled dataset (split 2/3 and 1/3)
5) Random Forest with SMOTE
6) Random Forest with Cost Matrix

Higher number of folds are infeasible since the dataset is highly unbalanced. These metrics are computed and compared: Precision, Recall, Sensitivity, Specifity and F1 Score.

The trained Random Forest model is used to extract the importance (see Figure II) of the attributes. This step is very informative and can be used as a feature selector for other models and can highlight risk factors.

Importance from the RandomForest model is computed from the Splits and Candidates values:

$$\text{Importance}_i = \frac{\text{splits}_i}{\text{candidates}_i}$$

Where i is the level. The KNIME implementations gives 3 levels and we sum them all.

		Ground Truth	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

Table III
CONFUSION MATRIX

IV. RESULTS

A. Metrics

Table III explains the following quantities:

- TP True Positives
- FP False Positives
- FN False Negatives
- TN True Negatives

Receiver Operating Characteristics (ROC) diagrams show a classifier performance plotting the False Positive Rates (FPR) against the True Positive Rates (TPR) of the classifier for a different thresholds. The idea is maximise the area under the ROC curve.

Intuitively, precision is the ability of the classifier not to label as positive a sample that is negative while recall is a measure of the ability of the classifier to find all the positive samples. [**?**]

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

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

$$Precision = p = \frac{TP}{TP + FP} \tag{3}$$

Another metric is the F-measure, defined as follows:

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{p} \cdot \mathrm{r}}{(\beta^2 \cdot \mathrm{p}) + \mathrm{r}} \tag{4}$$

Where β is a positive real chosen such that recall is considered β times as important as precision. In the case $\beta = 1$, we have the F_1 score:

$$F_1 = 2 \cdot \frac{\mathrm{p} \cdot \mathrm{r}}{\mathrm{p} + \mathrm{r}} \tag{5}$$

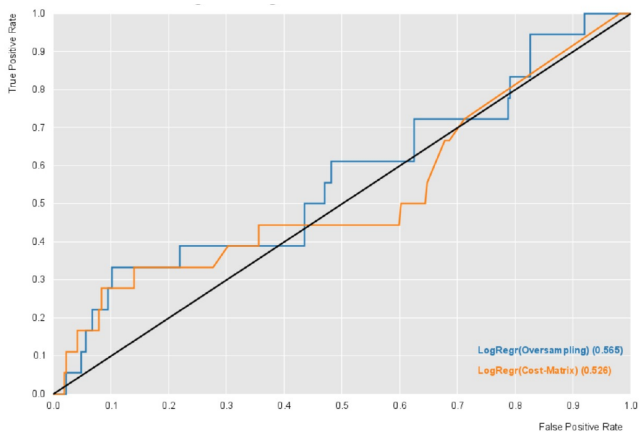

Figure 6. ROC comparison for two configuration of the Logistic Regression model

B. Comparisons

Metrics are reported in Table IV. Here we focus on the Recall and F-measure: we want the best model with an high cost on False Negatives. F-measure can also be used when a middleground balance between Precision and Recall is needed and the target class is unbalanced.

Random Forest models gave the highest Accuracy and Specifity, two metrics we not relevant here, especially with a so low Recall.

Oversampling generally never affected in a completely positive way the models, but helped improving Recall of the basic Logistic Regression model. However, the best value in this sense was obtained by the version sensitive to the Cost Matrix.

Overall, Logistic Regression resulted the best model, reaching 0.389 of Recall and 0.275 of F-Measure (highest values), while keeping a decent value of Accuracy.

Being able to set costs was actually a better strategy than oversampling, probably because of the (too) synthetic instances generated.

Exploring medical datasets on can help validate and investigate causal relations, but predicting a diagnosis without any targeted test or exam, having only partial information on behaviours related to risk factors is definitely a challenging task and we didn't get excellent results.

V. Conclusions

Perfomance of the classificators can appear unsatisfying, but this work provides a baseline where additional data and more complex ML tools can exploit additional (and less unbalanced) datasets. Exploiting trained models to rank risk factors can help build more relevant datasets, providing guidelines on which values and aspects about patients should be investigated and mined.

On the other hand, classificators results improve drastically when tests targeted to the same disease are available. This is expected, as each of exam have large medical literature describing their accuracy and diagnosing power.

Oversampling medical data with SMOTE also didn't show any particular enhancement, as the instances are syntethic and tend to over-represent particular patterns, thus confusing the classificators or not getting any improvement at all. The dataset also had a really low and unbalanced number of positive cases, on a already largely biased sample (Age distribution of patients, pregnancies).

Correctly evaluating the bias and the sample distribution from which the instances are sampled can also help in understanding how to handle the missing values. Here, our missing-imputation strategy keeps (and actually empowers) the bias, effectively supposing that if someone didn't answer, the truth should be similar to their peers (similar age, similar geographic provenience, ...). Since missing values weren't a minor issue, and they are due to privacy concerns, this is mostly a cultural consideration. A possible additional experimentation could be to change this approach.

However, the Correlation Analysis and the Feature importance extraction shows how the attributes related to known risk factors are popping out as the most important (Table II), and reflect the known causal relations we mentioned. Having richer datasets, providing more complete details about the medical history of the patient could lead to more accurate results and could enable *proper* clustering of patient population, suggesting the ones subject to higher risks.

Furthermore, cultural details plays a fundamental role in sexual habits and STDs related issues, but these elements are non-trivial to tackle and proper factor in medical machine learning models. E.g., it's been shown that early age at the first sexual intercourse and early pregnancies are strongly *interrelated* in most developing countries [10].

Bigger, richer and more complete datasets, including socio-economical and ethnic data on the patients could lead to intersting investigations and

improve this strategy to highlight and cluster high-risk groups.

REFERENCES

[1] American Cancer Society, "Cancer statistics center."

[2] R. Catarino, P. Petignat, G. Dongui, and P. Vassilakos, "Cervical cancer screening in developing countries at a crossroad: Emerging technologies and policy choices," pp. 281–290, dec 2015. [Online]. Available: /pmc/articles/PMC4675913/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4675913/

[3] Cancer Research UK, "Cervical cancer survival statistics."

[4] F. X. Bosch, A. Lorincz, N. Muñoz, C. J. Meijer, and K. V. Shah, "The causal relation between human papillomavirus and cervical cancer," pp. 244–265, 2002.

[5] V. Kumar, A. K. Abbas, and J. C. Aster, *Robbins basic pathology e-book*. Elsevier Health Sciences, 2017.

[6] S. Collins, T. P. Rollason, L. S. Young, and C. B. Woodman, "Cigarette smoking is an independent risk factor for cervical intraepithelial neoplasia in young women: A longitudinal study," *European Journal of Cancer*, vol. 46, no. 2, pp. 405–411, jan 2010.

[7] M. Ünlerşen, K. Sabanci, and M. Ozcan, "Determining cervical cancer possibility by using machine learning methods," *International Journal of Latest Research in Engineering and Technology*, vol. 3, pp. 65–71, 12 2017.

[8] I. D. Duncan, "Cervical screening," *The Obstetrician and Gynaecologist*, vol. 6, no. 2, pp. 93–97, apr 2004. [Online]. Available: http://doi.wiley.com/10.1576/toag.6.2.93.26984

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, jun 2011. [Online]. Available: http://arxiv.org/abs/1106.1813 http://dx.doi.org/10.1613/jair.953

[10] K. S. Louie, S. De Sanjose, M. Diaz, X. Castellsagué, R. Herrero, C. J. Meijer, K. Shah, S. Franceschi, N. Mũoz, and F. X. Bosch, "Early age at first sexual intercourse and early pregnancy are risk factors for cervical cancer in developing countries," *British Journal of Cancer*, vol. 100, no. 7, pp. 1191–1197, apr 2009. [Online]. Available: https://www.nature.com/articles/6604974

Classifier	Recall	Precision	Sensitivity	Specifity	F-Measure	Accuracy
Logistic Regression, non-oversampled	0.056	0.5	0.056	0.996	0.1	0.937
Logistic Regression, SMOTE oversampling	0.333	0.125	0.333	0.844	0.182	0.819
Logistic Regression, Cost Matrix	0.389	0.212	0.389	0.903	0.275	0.871
Random Forest, non-oversampled	0.056	0.167	0.056	0.981	0.083	0.923
Random Forest, SMOTE oversampling	0.056	0.143	0.056	0.978	0.08	0.92
Random Forest, Cost Matrix	0.111	0.133	0.111	0.952	0.121	0.899

Table IV

CLASSIFICATOR METRICS IN THE MAIN CONFIGURATIONS. 5 K-FOLD CROSS VALIDATION.