Cervical Cancer detection: Classification models evaluation

Raffaele Anselmo[1], Lorenzo Pastore[2]

Abstract

Cancer is a major cause of morbidity and mortality affecting population all around the world. Although cervical cancer is one of the most preventable diseases, it still has very high incidence rates in medium to low HDI countries. The main reasons for the higher incidence and mortality in developing countries are attributed to lack of awareness of cervical cancer, absence or poor quality of screening programmes and limited access to health care services. As other research has shown, we believe that Machine Learning has the potential to provide great low-cost aid in the early detection of cervical cancer to many medical structures around the world. In this paper we propose a Machine Learning classification models comparison based on a real case of data availability.

Keywords: Machine learning, Cancer, Comparison, Classification.

[1] University of Milano-Bicocca

[2] University of Milano-Bicocca

Table of Contents

1. Introduction

According to the World Health Organization (WHO), the five most common types of cancer diagnosed among women in 2012 were breast cancer (25.2%), colorectum cancer (9.2%), lung cancer (8.7%), cervical cancer (7.9%), and stomach (4.8%) cancer.[1]

While from the global picture it might not seem immediately clear, prioritizing the cure of cervical cancer is of critical importance, considering its disproportionately high incidence rate in the developing world. Cervical cancer is indeed the second most prevalent cancer type in less developed areas of the world.

According to WHO, effective interventions against cervical cancer exist, including screening for and treatment of precancer and invasive cancer. However, the cure rate for invasive cervical cancer is closely related to the stage of disease at diagnosis and the availability of treatment.[2]

From this perspective, early detection is of utmost importance. The asymptomatic nature of cervical cancer is the major challenge faced in the diagnosis in the early stage.[3] As other research has shown, we believe that Machine Learning has the potential to provide great low-cost aid to many medical structures around the world in the early detection of cervical cancer.[4]

In this paper we propose a machine learning (ML) approach based on a real hospital data. The dataset was collected at the “Hospital Universitario de Caracas” in Venezuela.[5,6]

Our aim is to predict the class attribute Diagnosis. In particular we want to test and investigate the effectiveness of different ML algorithm approaches on a real dataset which we expect to be partially incomplete.

After this brief introduction, in Chapter 2 we first conduct a preliminary assessment of the dataset, trying to take care of missing values and all preprocessing operations needed for our analysis. In Chapter 3 we present the five classification algorithms we selected. In Chapter 4 all performance evaluation of our classifiers are compared. In Chapter 5 we finally attempt to draw conclusions to validate our analysis and identify areas of interest for further researches.

2. Data Exploration and Preprocessing

The dataset used for our analysis was obtained from the UCI[7] repository and made available through the Kaggle platform. The analysis was conducted using the software Knime and in this paper we will refer to nodes and the relative properties we used.

2.1 Data Exploration

The dataset comprises demographic information, habits, and historic medical records of 858 patients. The data is represented by 32 risk factors (Appendix 1). In addition, there are 4 target variables: Hinselmann, Schiller, Cytology and Biopsy. Some patients did not answer all questions for individual privacy reasons. Hence, the dataset had to be pre-treated to deal with the missing values.

Our initial data exploration highlighted the need to deal with numerical, boolean and nominal attributes. At the same time, the correlation matrix showed some very high correlations that need to be analyzed (Appendix 2).

During the data exploration we were driven by our concern about the dataset dimension and handling of missing values. We focused on understanding the meaning of the variables, trying to build a basis of domain knowledge while referring as much as possible to existing literature[8–10] to minimize error.

2.2 Preprocessing

In this section we will refer to the variables with an (id) which is the variable identifier for Appendix 1 where the number of missing values is shown.

First we focus on the target variables. Rather than choosing one of the four target variables already in the dataset, we decided to construct a new variable that takes into account all the information of the four variables together. The new variable, named

- *Diagnosis*, takes value 1 if at least three of the four boolean target variables above take value 1, it takes 0 in all other cases.

We use *Diagnosis* as unique identifier of the presence of cancer, so we proceed with deleting the previous target variables: Hinselmann (33), Schiller (34), Cytology (35) and Biopsy (36).

The dataset initially consisted of 858 records. During this phase we noticed that some variables were particularly affected by missing values. Below we list some operations we performed on the featured variables:

- *STDs: Time since first diagnosis* (27) and *STDs: Time since last diagnosis* (28) were removed as a consequence of lack of available values.
- *STDs: Cervical Condylomatosis* (15) and *STDs: AIDS* (22) were removed because characterized for taking no other value except 0.
- The variables *STDs (number)* (13) and *STDs: Number of diagnosis* (26) seemed redundant and have an high correlation (Appendix 2) with most of the attributes, probably because these two variables represent an aggregation of the other attributes' information; so we decided to remove them.
- We also noticed that some of the STDs attributes had the same number of missing values (105), so we decided to delete these rows because we did not have sufficient information to replace such a high number of missing values.
- For numerical attributes that have few missing values, we used the median to replace them, due to unbalanced classes.

Through these operations we deleted or replaced all missing data and drastically cleaned the dataset.

The very last step of the preprocessing phase consisted in the normalization of numerical attributes, due to different scale of values.

3. Classification Algorithms

Once we completed the preprocessing operations, the dataset was ready to be used for our machine learning analysis and we only needed to select a classification algorithm. It is important to highlight that our purpose is not only to perform a classification, but a classification which is optimal for our data. For this reason, according to the data exploration analysis and other research we consulted,[11,12] we decided to use five different classification algorithms:

- J48 Random Forest[13] (J48)
- Naivë Bayes (NB)
- Support Vector Machine[14] (SVM)
- Logistic Regression[15] (LR)
- Multi Layer Perceptron[16] (MLP)

Additionally, in this paper we also aimed to build a good practice evaluation models comparison procedure.

Before we let our classification algorithms work, we integrated a 10 folds Cross Validation with stratified sampling with respect of the target variable, *Diagnosis*, to avoid underfitting and overfitting problems.

The stratified sampling was needed due to the unbalance of Diagnosis class (instead LOOCV is not applicable).

The most significant classifiers parameters are outlined as follows. We decided not to use pruning for J48 Random Forest because it didn't work with our data distribution and the limited data available.

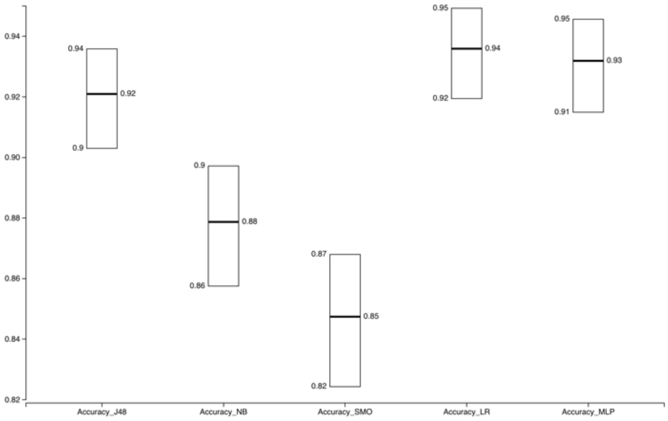
For the Multi Layer Perceptron we selected 3 hidden layers with 15 neurons for each layer. However we could use any other combination of hidden layers and neurons for layer. We leave the selection of the optimal parameters to further studies.

For the remaining classifiers we used the knime's node default parameters.

4. Performance Evaluation

To evaluate the classifiers performance we first have a look at the accuracy and its confidence interval.

Table 1 - Accuracy confidence interval



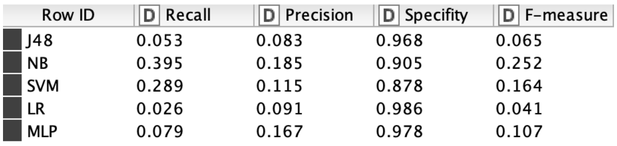
At first glance Naivë Bayes (0.88) and the Support Vector Machine (0.85), respectively second and third from the left of Table 1, seem to have lower performance than J48 (0.92), Logistic Regression (0.94) and Multi Layer Perceptron (0.93).

To have a better understanding of the performance and a more exhaustive comparison between the classifiers, we decided to compute their paired error difference with a level of significance of 0.90.

The error difference between pairs of classifiers, shown in Appendix 3, confirms that the errors of NB and SVM are significantly higher than the ones of all other classifiers, while the error difference between J48, MLP and LR is not significantly different from zero.[*]

To have a complete overview of the performance of each classifier, we also have to take into account four other evaluation measures: *Recall, Precision, Specifity* and *F-measure.*

Table 2 - Recall, Precision, Specifity and F-measure



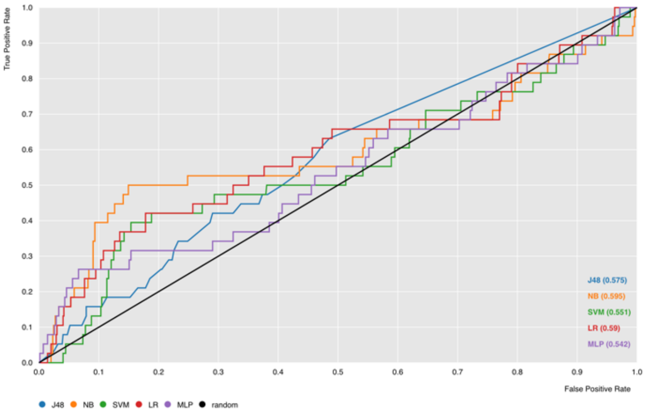
Row ID	D Recall	D Precision	D Specifity	D F-measure
■ J48	0.053	0.083	0.968	0.065
■ NB	0.395	0.185	0.905	0.252
■ SVM	0.289	0.115	0.878	0.164
■ LR	0.026	0.091	0.986	0.041
■ MLP	0.079	0.167	0.978	0.107

These measures allow us to evaluate the classifiers with respect to the context and the domain knowledge. In this case, considering that in medical environment the cost of a wrong diagnosis could be higher for a false negative (FN) than a false positive (FP), we focused our attention on the recall measure.

Although Naivë Bayes and Support Vector Machine had the worst Accuracy, they have the best values for Recall, so they could be the best predictors to minimize the FN. Also F-measure, that evaluate together Recall and Precision, have the highest value in correspondence of NB and SVM.

To have a visual cue of performance comparison, we look at the ROC Curve.

Table 3 - ROC Curve



The ROC Curve shows that if we accept a low rates of false positives, the predictor that ensures the highest True Positive Rate is the Naivë Bayes. Instead if we want to reach a higher values of True Positive Rate, we should use J48 predictor.

5. Conclusion and suggestion

With this analysis, our aim was to predict the class attribute Diagnosis and to build a classification models comparison that could help us detect the fittest model for our data. For this task we found the ROC curve to be a particularly efficient solution to the problem of comparing multiple classifiers in imprecise and changing environments.[17]

[*] assuming data is normally distributed at 90% confidence interval

This allowed us, not only to select the best model but also to take care of the costs' trade-off characterizing our very specific case.

In conclusion, based on the analysis and the model, this paper has shown the potential value added of using ML for early detection of cervical cancer in the context of small and incomplete data collection.

Below we share the main difficulties we've found during the project, hoping that these will be useful for future research.

- **Preprocessing**: Although Knime has its own dedicated nodes, we suggest to conduct preprocessing operations with other software (Python, R) that allows to better manipulate the data. Alternatively, it is possible to use Knime integration nodes (Rsnippet, PythonScript).
- **Dataset**: Our analysis focused on the ML algorithms effectiveness in the context of a real data collection. For a more accurate analysis on cervical cancer we suggest to refer to a larger dataset which should include records from different hospital around the world and allow for a more accurate selection of demographic variable.

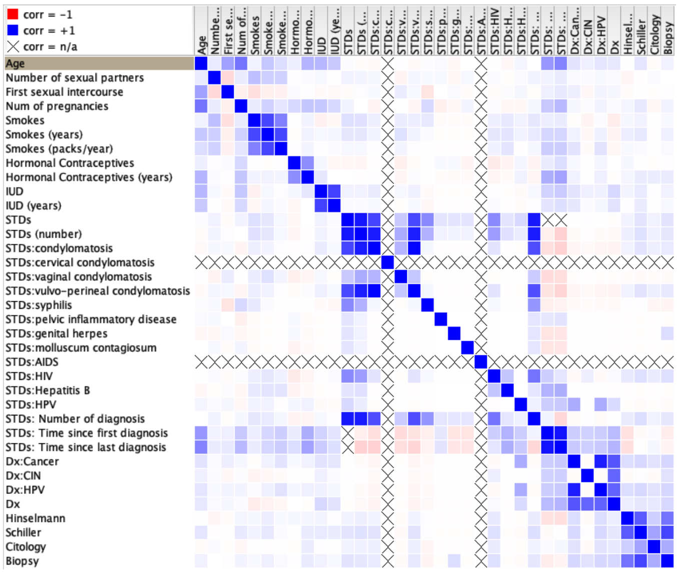
Finally, we wish to focus on the importance of clear data collection in the medical environment. We believe this would help not only to have better ML working models, but also to stimulate knowledge and awareness in-country where the lack of programmes hampers the detection of cervical cancer detection in early stage.

Appendix

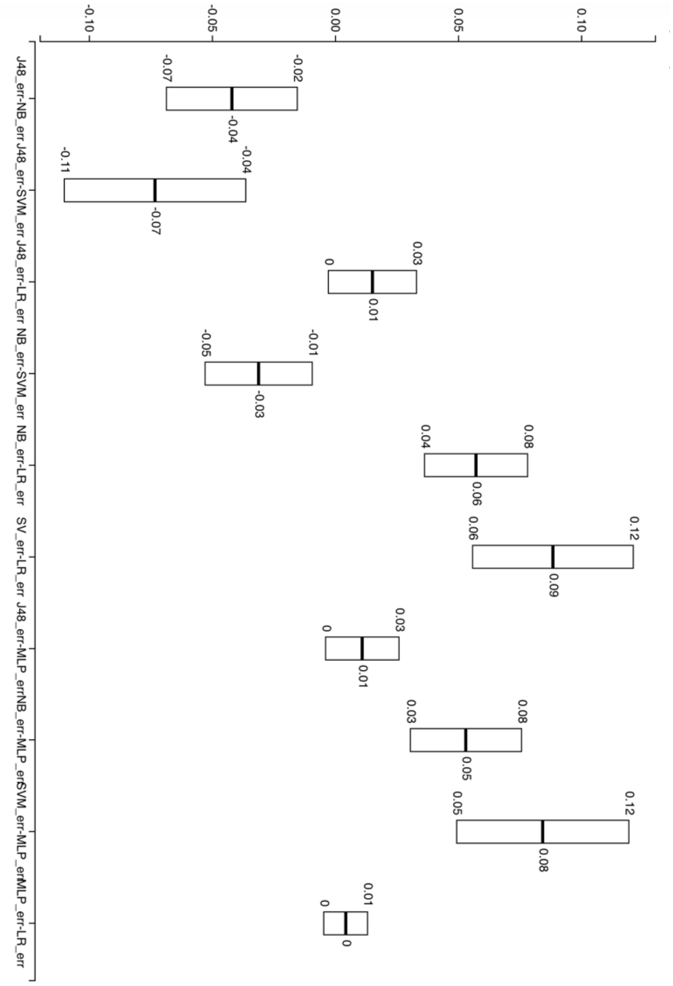
Appendix 1 – Variable description

	Variable	Type	Role	N. Missing value
1	Age	integer	feature	0
2	Number of sexual partners	integer	feature	26
3	Age of first sexual intercourse	integer	feature	7
4	Number of pregnancies	integer	feature	56
5	Smokes	boolean	feature	13
6	Smokes (years)	double	feature	13
7	Smokes (packs/years)	double	feature	13
8	Hormonal Contraceptives	boolean	feature	28
9	Hormonal Contraceptives (years)	integer	feature	108
10	IUD	boolean	feature	117
11	IUD (years)	integer	feature	117
12	STDs	boolean	feature	105
13	STDs (number)	integer	feature	105
14	STDs: Condylomatosis	boolean	feature	105
15	STDs: Cervical Condylomatosis	boolean	feature	105
16	STDs: Vaginal Condylomatosis	boolean	feature	105
17	STDs: Vulvo-perineal Condylomatosis	boolean	feature	105
18	STDs: Syphilis	boolean	feature	105
19	STDs: Pelvic inflammatory disease	boolean	feature	105
20	STDs: Genital herpes	boolean	feature	105
21	STDs: Molluscum contagiosum	boolean	feature	105
22	STDs: AIDS	boolean	feature	105
23	STDs: HIV	boolean	feature	105
24	STDs: Hepatitis B	boolean	feature	105
25	STDs: HPV	boolean	feature	105
26	STDs: Number of diagnosis	integer	feature	0
27	STDs: Time since first diagnosis	integer	feature	787
28	STDs: Time since last diagnosis	integer	feature	787
29	Dx: Cancer	boolean	feature	0
30	Dx: CIN	boolean	feature	0
31	Dx: HPV	boolean	feature	0
32	Dx	boolean	feature	0
33	Hinselmann	boolean	target	0
34	Schiller	boolean	target	0
35	Cytology	boolean	target	0
36	Biopsy	boolean	target	0

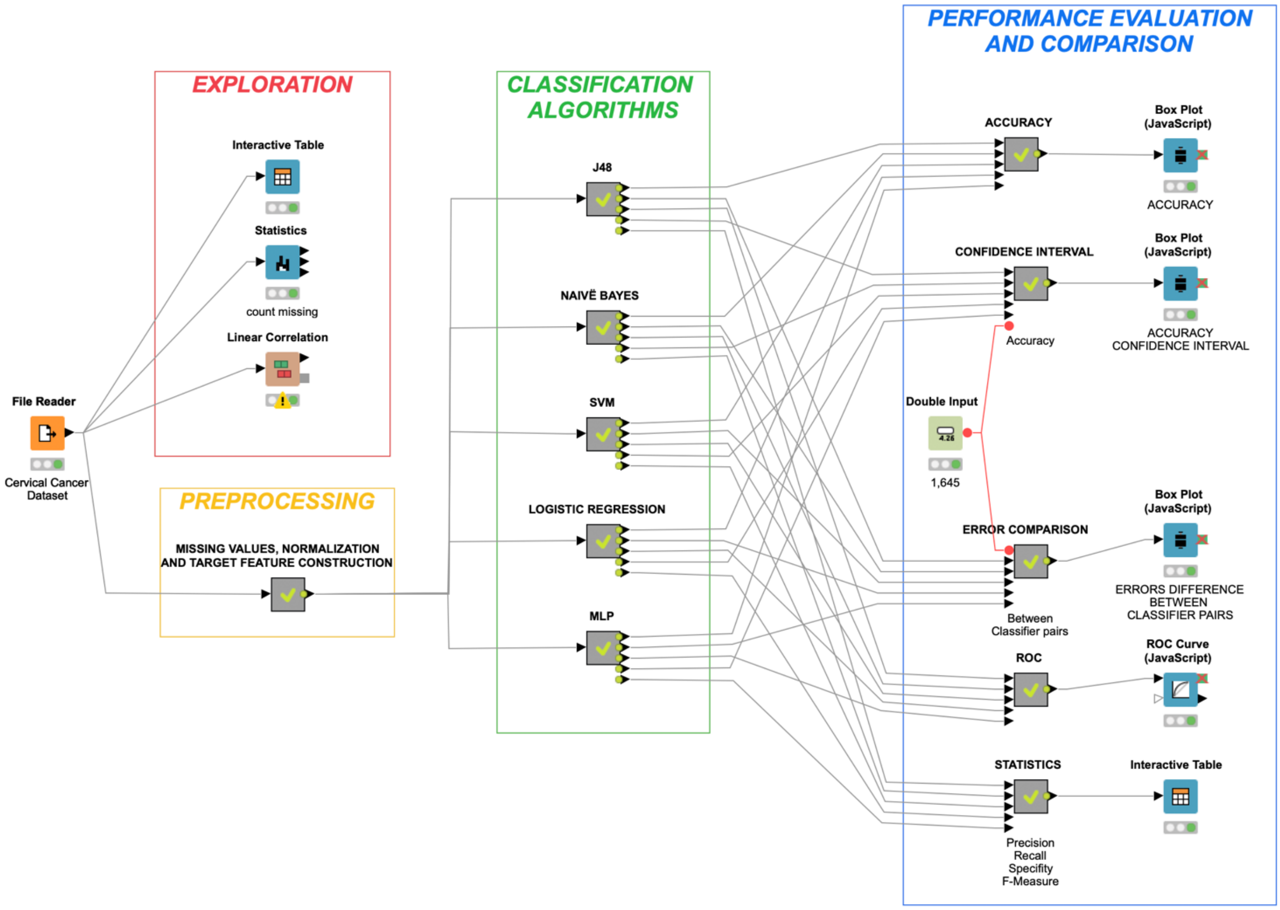
Appendix 2 – Correlation matrix



Appendix 3 – Error difference between pair of classifiers



Appendix 4 – Knime workflow



Bibliography

1. Stewart, B. W. & Wild, C. P. *World Cancer Report 2014.* (International Agency for Research on Cancer/World Health Organization, 2014).
2. *Comprehensive cervical cancer control: a guide to essential practice*. (2006).
3. Shetty, A. & Shah, V. Survey of Cervical Cancer Prediction Using Machine Learning: A Comparative Approach. in *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* 1–6 (IEEE, 2018). doi:10.1109/ICCCNT.2018.8494169
4. Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine* **23**, 89–109 (2001).
5. Fernandes, K., Cardoso, J. S. & Fernandes, J. Transfer Learning with Partial Observability Applied to Cervical Cancer Screening. in *Pattern Recognition and Image Analysis* (eds. Alexandre, L. A., Salvador Sánchez, J. & Rodrigues, J. M. F.) **10255**, 243–250 (Springer International Publishing, 2017).
6. *Pattern Recognition and Image Analysis*. **10255**, (Springer International Publishing, 2017).
7. UCI Machine Learning Repository: Cervical cancer (Risk Factors) Data Set. Available at: https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29. (Accessed: 12th February 2019)
8. Luhn, P. *et al.* The role of co-factors in the progression from human papillomavirus infection to cervical cancer. *Gynecologic Oncology* **128**, 265–270 (2013).
9. Gadducci, A., Barsotti, C., Cosio, S., Domenici, L. & Riccardo Genazzani, A. Smoking habit, immune suppression, oral contraceptive use, and hormone replacement therapy use and cervical carcinogenesis: a review of the literature. *Gynecological Endocrinology* **27**, 597–604 (2011).
10. Schiffman, M., Castle, P. E., Jeronimo, J., Rodriguez, A. C. & Wacholder, S. Human papillomavirus and cervical cancer. *The Lancet* **370**, 890–907 (2007).
11. Wu, W. & Zhou, H. Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches. *IEEE Access* **5**, 25189–25195 (2017).
12. keymasi, M., Mishra, V., Aslan, S. & Asem, M. M. Theoretical Assessment of Cervical Cancer Using Machine Learning Methods Based on Pap-Smear Test. in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* 1367–1373 (IEEE, 2018). doi:10.1109/IEMCON.2018.8615041
13. Breiman, L. *Classification And Regression Trees*. (Routledge, 2017). doi:10.1201/9781315139470
14. Sweilam, N. H., Tharwat, A. A. & Abdel Moniem, N. K. Support vector machine for diagnosis cancer disease: A comparative study. *Egyptian Informatics Journal* **11**, 81–92 (2010).
15. Tu, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology* **49**, 1225–1231 (1996).
16. West, D., Mangiameli, P., Rampal, R. & West, V. Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *European Journal of Operational Research* **162**, 532–551 (2005).
17. Provost, F. J. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions.