Biased logistic models applied to cervical cancer risk

José Antonio Cárdenas Garro
Universidad de Lima
Lima, Perú
jacarden@ulima.edu.pe

José Antonio Taquía Gutiérrez
Universidad de Lima
Instituto de Investigación Científica
Lima, Perú
jtaquia@ulima.edu.pe

Abstract

In the present work, the CRISP-DM methodology was proposed to develop a set of machine learning models applied to evaluate cervical cancer risk suffer. For this research, a sample of 858 patients was taken, who were asked a series of questions regarding this pathology. The database has an unbalanced dependent variable, since this is a health study, the balancing technique will not be used to identify the variables that will enter the model, the Boruta library was used for variable selection. For the development, five algorithms will be used: Support Vector Machine (SVM), decision trees using the CHAID and CART algorithms, logistic regression and "asymmetric link" models. The models proposed in this work were refined by means of the Auc, Gini, Log loss and KS (Kolmogorov-Smirnov) indicators, as a result using the proposed models, AUC values of 98% were obtained.

Keywords
Cervical cancer, Machine learning, Boruta library, Asymmetric linkage models, Predictive models

1. Introduction
Cervical cancer occurs in the lower part of the uterus, this type of pathology is one of the most common and with high mortality rate if not detected in time. Various strains of human papillomavirus (HPV), or a sexually transmitted infection, play an important role in causing most types of cervical cancer. Most used test to detect are Hinselmann, Schiller, citology and biopsy (Moldovan, 2020), which are very methods to detect this oncology pathology. Because sexually transmitted infections play an important role in the risk of having cervical cancer, variables with information on sexually transmitted infections were included in the study (Islami et al. 2017; Sindiani et al. 2020). In this sense, our work presents a predictive model using machine learning techniques to classify patients which might be more likely to have cervical cancer and thus be able to detect them and prevent future deaths.

This paper has been structured as follows: Sections 2 and 3 describe the literature reviewed and our research methodology. Section 4 discusses data treatment and modeling. Section 5 provides results and the improvements proposed. Finally, Section 6 presents our conclusions and recommendations for future lines of research on the subject matter.

1.1 Objectives
For these purposes of this work, the following objectives were defined

- Design a predictive classification model to determine who is more likely to be at risk of cervical cancer.
- Compare models using unbalanced data

2. Literature Review
Machine learning applied to solve health problems is a big trend on different phases of medical research. On the diagnostic phase different ensemble models can obtain good results as reported in Hsu, et al. (2021). The main aspect

in this field is where obtain data to create models. In relation to this, open access image databases (Weegar and Sundström, 2020), and tabular data as in the UCI database, are helpful to evaluate new models (Sindiani et al. 2020; Adem et al. 2018; Richter and Khoshgoftaar, 2018; Dua and Graf, 2019; Priya and Karthikeyan, 2020).

The study of the detection of this type of cancer had increased in techniques such as convolutional neural networks and image processing in combination with KNN or SVM filtering methods. (Jia et al. 2020; Arora et al. 2021; Lee, 2021; Bhuvaneshwari, 2019). Using nuclei cellular images by cytopathologists to identify features that train hierarchical Bayesian and stacked models to detect the presence of anomalies (Diniz et al. 2021). The use on bio markers for cancer detection are part on these methods, this increase precision metrics (Hajjo et al. 2021). In relation of applying image classification techniques, some studies describe better performance using "transferred learning" techniques (Chandran et al. 2021). In the work of Meng (2020) supervised algorithms, ensemble techniques with random forest and boosting methods perform well using stack model with AUC values in range 77-78%. In the same sense cluster classification as a function of survival time is also applied as a prediction of cervical cancer as a variable (Ding et al. 2021). There are works related to the application of analytical models in unbalanced data which explain this approach (Jiang et al. 2021; Tay, 2016). An important fact related to data manipulation in this type of study is presented by Ijaz et al. (2020), using synthetic methods (SMOTE) to increase number of cases to obtain balance data.

3. Methods

In the case of a data mining projects, CRISP-DM methodology is a standard for experimental purposes using databases. It is structured in six phases, some bidirectional, which means that in some of them, would be iterative review of data. Completion of phases closes a complete implementation cycle of a datamining project (Meng, 2020; Yaacob et al. 2019).

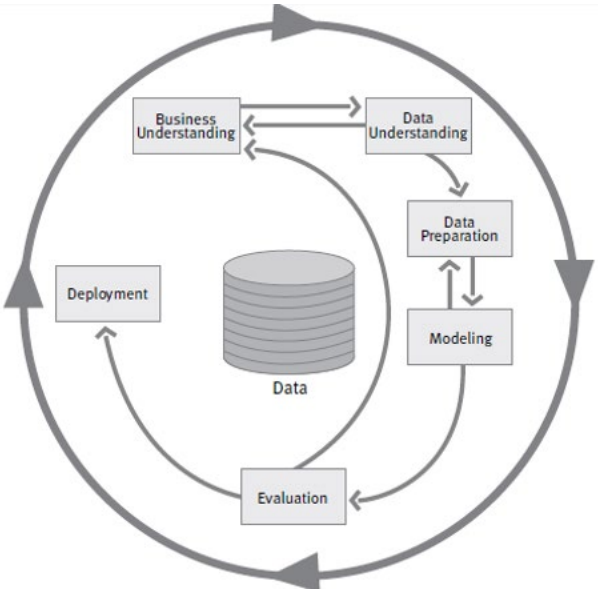

Figure 1: CRISP-DM Methodology Phases

Figure 1 shows the incremental nature of the methodology emphasizing in our study the application of supervised models that have a dichotomous qualitative variable as target. The method suggests successive iterations on data understanding and modeling phases to increases accuracy metrics in successive iterations (Aggarwal, 2015).

4. Data Collection

Data was obtained from UCI Irvine database, collecting 858 cases from the “Hospital Universitario de Caracas-Venezuela” (Dua and Grafff, 2019). In relation to the treatment of missing values, all variables were imputed with median and mode, so as not to lose variables that could contribute to the model. (Figure 2)

The sample was divided into two parts: train and test for the construction and validation of the model. The train represents 70% and the test 30%.

In the exploratory analysis of the data, the original set of records 859 and variables 34 was used.

- 10 variables have no missing values in their records.
- 5 variables have less than 5% missing values (Smokes, Smokes_years, Smokes_packs_year, FSI, PA).
- 1 variable of 6% of predicted values (pregnancies).
- 17 variables with missing values between 12% and 14% (HC, HCA, IUD, IUD_years, STDs, STD_number, 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 and STDs_Hepatitis.B)

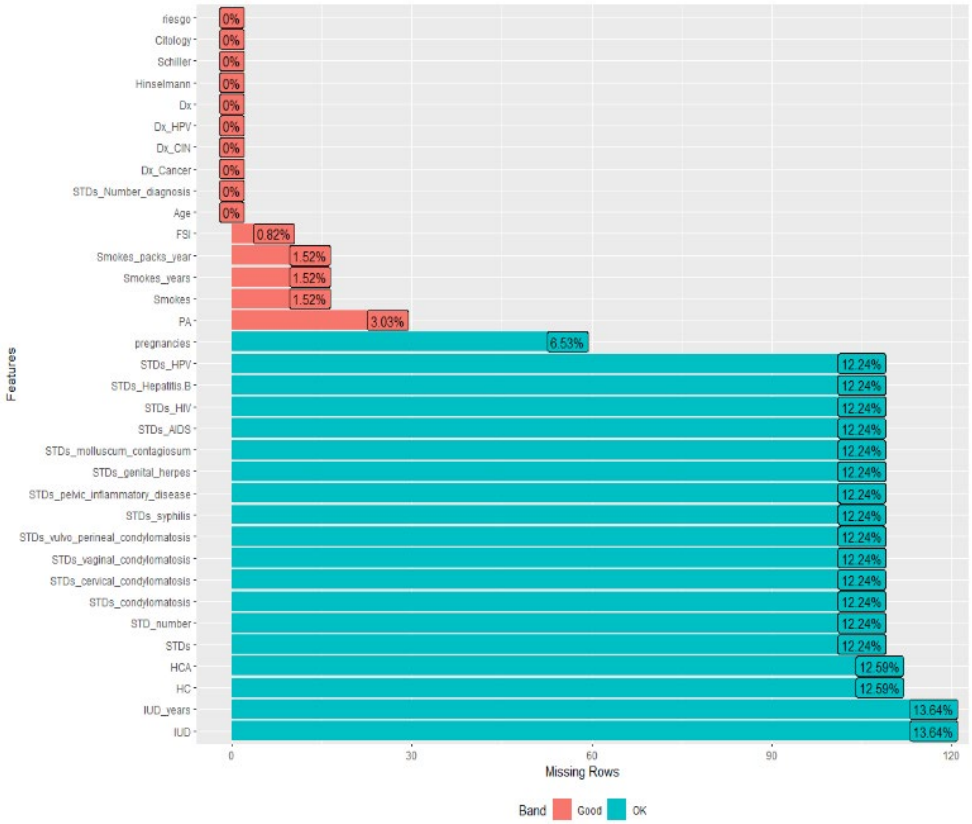

Figure 2: Percentage of missing values by variable

The variables that exceed a cutoff point of 0.6 (recommended, but not influential) are the variables STD_number, Smokes years, which will be removed in the model since the variables have a high relationship with the target variable. (Figure 3)

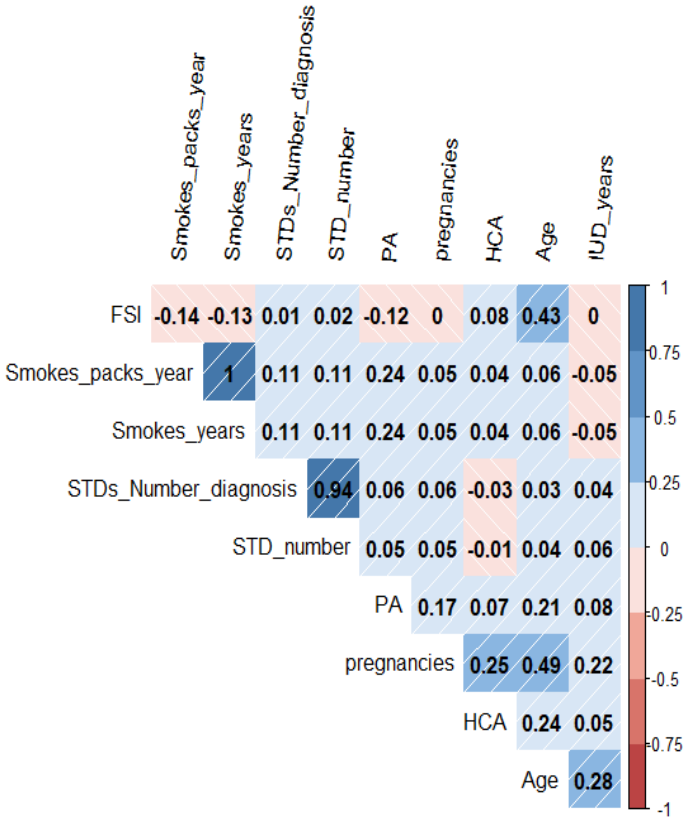

Figure 3: "Spearman" correlation matrix

4.1 Balancing of the target variables

As the study is a case of health, having or not having the risk of having cervical cancer, the baseline is unbalanced. This has been decided so as not to manipulate the nature of the data (Buja, et al. 2005). For this reason, it makes sense to use a skewed logistic model (Tay, 2016). (Table 1)

Table1: Distribution of the target variable

Dependent variable: risk	Amount of data	Percentage of total
0	803	93.59%
1	55	6.41%
Total	858	100.0%

5. Results and Discussion

5.1 Numerical Results

For the variable selection the Boruta library was implemented. It is a selection method subsets data obtaining the best possible fit. This method uses Random Forest as the underlying algorithm, consisting of entering all variables into the equation and then excluded one after the other. At each iteration the variable with less importance is taken out quantify contrast. Applying this method, we have five selected variables for the proposed models. Those marked in green are the most important, followed by those marked in yellow and red. The important variables for the model are Age, Dx, Cytology, Shiller, Hinselmann.

selection will be applied only to the logistic regression algorithm. The variables that were selected are: STDs_condylomatosis, STDs_HIV, Dx_Cancer, Hinselmann, Schiller.

The models obtained in the modeling stage were evaluated with the test data, obtaining the following indicators shown in Table 2.

Table2: Comparison of models. Validation indicators

Model	AUC	GINI	LOG-LOSS	K-S
Logistics-BK	0.877	0.754	0.123	0.774
Logistics	0.917	0.833	0.124	0.774
CHAID Tree	0.872	0.744	0.128	0.711
CART tree	0.872	0.744	0.128	0.711
Radial SVM	0.910	0.821	0.131	0.774
Linear SVM	0.919	0.838	0.123	0.824
Cloglog	0.982	0.964	0.098	0.726
Scobit	0.974	0.947	0.096	0.724

5.2 Graphical Results

Variable importance are obtains using Boruta library using R programming language. Figure 4 shows box plots on importance for the model of each variable to detect the ones to be used on the model.

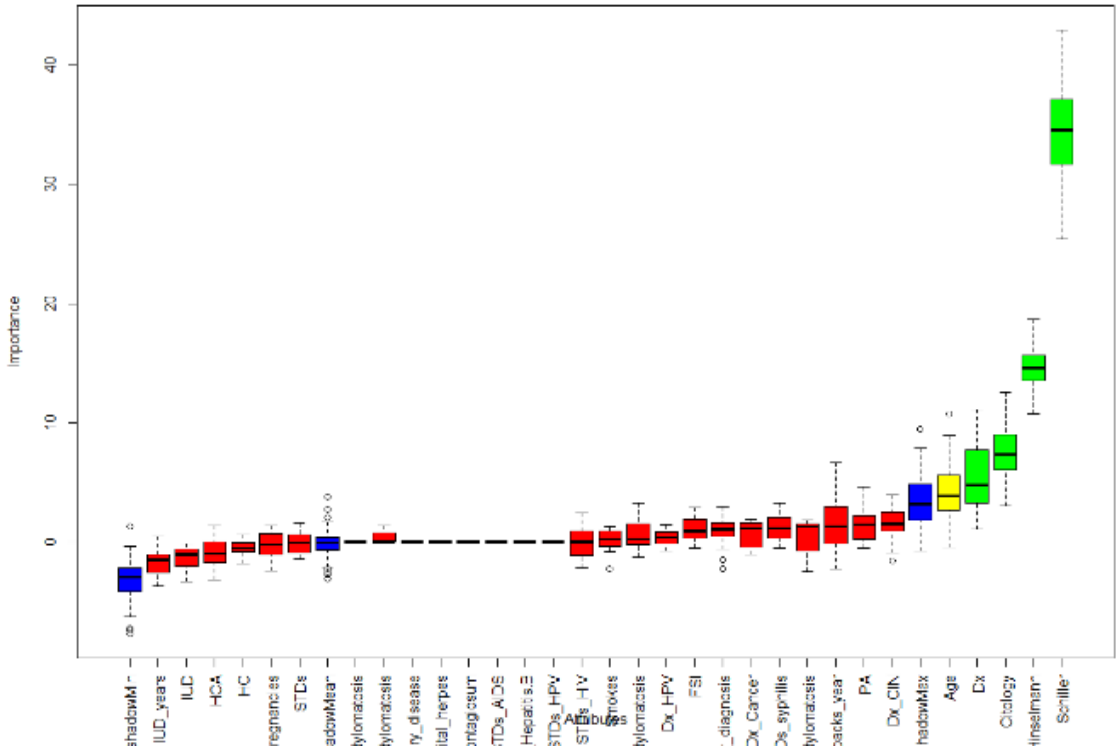

Figure 4: Result of Boruta's method

5.3 Proposed Improvements

It will be interesting to try other variable selection techniques to determine the best variables before creating the models, for example: Stepwise, Forward, and penalized models such as Lasso regularization. In addition, algorithms such as Random Forest, neural networks, etc. could be applied to evaluate performance

6. Conclusion

In our study Cloglog algorithm is slightly superior to Scobit, both algorithms perform better analyzing different performance metrics in comparison to other traditional methods. These algorithms have better predictive application on obtaining cervical cancer risk due to the fact dataset is unbalanced.

References

Adem, K., Kilicarslan, S., & Comert, O. Classification and diagnosis of cervical cancer with softmax classificationwith stacked autoencoder. *Expert Systems with Applications*. (2019). DOI: https://doi.org/10.1016/j.eswa.2018.08.050

Aggarwal, C. C. Data mining: the textbook. Springer. (2015).

Arora M., Dhawan S., Singh K. *Deep Learning in Health Care: Automatic Cervix Image Classification Using Convolutional Neural Network.* In: Marriwala N., Tripathi C.C., Kumar D., Jain S. (eds) Mobile Radio Communications and 5G Networks. Lecture Notes in Networks and Systems, vol 140. (2021) Springer, Singapore. https://doi.org/10.1007/978-981-15-7130-5_10

Bhuvaneshwari, K. V., & Poornima, B. Cervical Cancer Cell Identification & Detection Using Fuzzy C Mean and K nearest Neighbor Techniques. *International Journal of Innovative Technology and Exploring Engineering*. 2019. DOI: 10.35940/ijitee.I.7892.0881019

Buja, A., Stuetzle, W., & Shen, Y. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November 3*. (2005)

Chandran, V., Sumithra, M. G., Karthick, A., George, T., Deivakani, M., Elakkiya, B., ... & Manoharan, S. Diagnosis of Cervical Cancer based on Ensemble Deep Learning Network using Colposcopy Images. *BioMed Research International*, 2021. DOI: 10.1155/2021/5584004.

Ding, D., Lang, T., Zou, D. et al. Machine learning-based prediction of survival prognosis in cervical cancer. *BMC Bioinformatics* 22, 331 (2021). https://doi.org/10.1186/s12859-021-04261-x

Diniz, D. N., Rezende, M. T., Bianchi, A. G. C., Carneiro, C. M., Ushizima, D. M., de Medeiros, F. N. S., & Souza, M. J. F. A Hierarchical Feature-Based Methodology to Perform Cervical Cancer Classification. *Applied Sciences*, 11(9), 4091. (2021). DOI: http://dx.doi.org/10.3390/app11094091

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Hajjo, R., Sabbah, D. A., Bardaweel, S. K., & Tropsha, A. Identification of Tumor-Specific MRI Biomarkers Using Machine Learning (ML). *Diagnostics* (Basel, Switzerland), 11(5), 742. (2021) https://doi.org/10.3390/diagnostics11050742

Hsu, C. H., Chen, X., Lin, W., Jiang, C., Zhang, Y., Hao, Z., & Chung, Y. C. Effective multiple cancer disease diagnosis frameworks for improved healthcare using machine learning. *Measurement, 175, 109145. (2021)*

Ijaz, M. F., Attique, M., & Son, Y. Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods. *Sensors,* 20(10), 2809. (2020) DOI: http://dx.doi.org/10.3390/s20102809

Islami, F., Torre, L. A., Drope, J. M., Ward, E. M., & Jemal, A. Global cancer in women: cancer control priorities. Cancer Epidemiology and Prevention Biomarkers, 26(4), 458-470. (2017) DOI: 10.1158/1055-9965.EPI-16-0871

Jia, A. D., Li, B. Z., & Zhang, C. C. Detection of cervical cancer cells based on strong feature CNN-SVM network. *Neurocomputing*, 411, 112-127. (2020). Doi: https://doi.org/10.1016/j.neucom.2020.06.006

Jiang, C., Tay, R., & Lu, L. A skewed logistic model of two-unit bicycle-vehicle hit-and-run crashes. *Traffic injury prevention*, 22(2), 158-161. (2021)

Lee, D., & Yoon, S. N. Application of Artificial Intelligence-Based Technologies in the Healthcare Industry: Opportunities and Challenges. *International Journal of Environmental Research and Public Health*, 18(1), 271. (2021) DOI: http://dx.doi.org/10.3390/ijerph18010271

Meng, Q. Machine learning to predict local recurrence and distant metastasis of cervical cancer after definitive radiotherapy. *International Journal of Radiation Oncology, Biology, Physics, 108(3), e767*. (2020)

Moldovan, D. Cervical Cancer Diagnosis Using a Chicken Swarm Optimization Based Machine Learning Method. *International Conference on e-Health and Bioengineering (EHB)* (pp. 1-4). IEEE. (2020, October).

Priya, S., & Karthikeyan, N. K. A Heuristic and ANN based Classification Model for Early Screening of Cervical Cancer. *International Journal of Computational Intelligence Systems*, 13(1), 1092-1100. (2020) DOI: 10.2991/ijcis.d.200730.003

Richter, A. N., & Khoshgoftaar, T. M. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artificial intelligence in medicine*, 90, 1-14. (2018) DOI: 10.1016/j.artmed.2018.06.002

Sindiani, A. M., Alshdaifat, E. H., & Alkhatib, A. J. Investigating Cervical Risk Factors that Lead to Cytological and Biopsy Examination. *Medical archives (Sarajevo, Bosnia and Herzegovina)*, 74(4), 294-297. (2020) https://doi.org/10.5455/medarh.2020.74.294-297.

Tay, R. Comparison of the binary logistic and skewed logistic (Scobit) models of injury severity in motor vehicle collisions. *Accident Analysis & Prevention*, 88, 52-55. (2016).

Weegar R, Sundström K Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations. *PLoS ONE 15(8): e0237911*. (2020) https://doi.org/10.1371/journal.pone.0237911.

Yaacob, W. F. W., Nasir, S. A. M., Yaacob, W. F. W., & Sobri, N. M. Supervised data mining approach for predicting student performance. *Indonesian Journal of Electrical Engineering and Computer Science*, 16(3), 1584-1592. (2019)

Biographies

José Antonio Cárdenas Garro holds a master's in data science from Universidad Ricardo Palma. Licensed in Statistics from Universidad Nacional Mayor de San Marcos. Currently playing the role of Head of Models and Risk Strategies in Scotiabank main offices at Lima, applying different statistical techniques and machine learning methods to discriminate credit risk in Peru and Canada. Professor at different universities teaching courses in Data Science (Universidad de Lima - Universidad Ricardo Palma - Universidad Nacional Mayor de San Marcos). Data science internships in Mexico and Colombia such as: Machine learning at the Instituto Tecnológico Autónomo de México (ITAM-México), Centro de Investigación Matemática (CIMAT - México) and Business Analytics at Universidad Católica de Colombia.

José Antonio Taquía is a Doctoral Researcher from Universidad Nacional Mayor de San Marcos at Lima-Peru and holds a Master of Science degree in Industrial Engineering from University de Lima, where he is a member of the School of Engineering and Architecture teaching courses on quantitative methods, predictive analytics, and research methodology. He has a vast experience on applied technology related to machine learning and industry 4.0 disrupting applications. In the private sector he was part of several implementations of technical projects including roles as an expert user and in the leading deployment side. He worked as a senior corporate demand planner with emphasis on the statistical field for a multinational Peruvian company in the beauty and personal care industry with operations in Europe and Latin America. Mr. Taquía has a strong background in supply chain analytics and operations modeling applied at different sectors of the industry. He is also a member of the Scientific Research Institute at the Universidad de Lima being part of the exponential technology and circular economy groups. His main research interests are on statistical learning, predictive analytics, and industry 4.0.