Journal of Women's Health Issues (ISSN 2995-6331)

Search For Models Under The Machine Learning Philosophy Of Predictors Of High-Risk HPV Infection In Healthy Women.

Costa Santos, L. Eduardo[1], González Romero, Francisco Javier[2], Del Valle-Mendoza, Juana Mercedes[3], Ponce-Benavente, Luis[3], Rojas-Pinelo, Patricia[3], Aguilar-Luis, Miguel Angel[3], Palomares-Reyes, Carlos[3], Becerra-Goicochea, Lorena[4], Pinillos-Vilca, Luis[4], Silva-Caso, Wilmer[3], Weig, Pablo[3], Alvitrez-Arana, Juan[3], Bazán-Mayra, Jorge[5].

Affiliation:

*[1] Universidad Viña del Mar, Escuela de Ciencias, Viña del Mar, Chile. **Email:** lcosta@uvm.cl*

*[2] Universidad Viña del Mar, Escuela de Ciencias, Viña del Mar, Chile. **Email:** f.gonzalez@uvm.cl*

*[3] Universidad Peruana de Ciencias Aplicadas. Escuela de Medicina. Lima, Perú **Email:** juana.delvalle@upc.pe*

[3] Universidad Peruana de Ciencias Aplicadas. Escuela de Medicina. Lima, Perú Wilmer Silva-Caso

[3] Universidad Peruana de Ciencias Aplicadas. Escuela de Medicina. Lima, Perú Pablo Weilg

[3] Universidad Peruana de Ciencias Aplicadas. Escuela de Medicina. Lima, Perú Juan Alvitrez-Arana

[4] Hospital Regional Docente de Cajamarca, Cajamarca, Perú. Lorena Becerra-Goicochea

[4] Hospital Regional Docente de Cajamarca, Cajamarca, Perú. Luis Pinillos-Vilca

[5] Dirección Regional de Salud de Cajamarca (DIRESA), Cajamarca, Perú. Jorge Bazán-Mayra

***Corresponding author**
Luis Eduardo Costa-Santos ,
Universidad Viña del Mar, Escuela de Ciencias, Viña del Mar, Chile.
Email : lcosta@uvm.cl

Received Date : December 18, 2024
Accepted Date : December 19, 2024
Published Date : February 01, 2025

Keywords : *Risk predictors; HPV infection; sexual health; public health; primary care..*

INTRODUCTION

Cervical cancer is a public health problem that accounts for 36% of all cancers in women. Of these, 99.7% show the presence of HPV (human papillomavirus) DNA, its main aetiological agent, with viral genotypes 16 and 18 being the most common. Although cervical can-cers originate from cells with precancerous changes (precancers), only some women with precancerous cells will develop the disease (American Cancer Society 2014). Progression from precancer to cancer usually takes several years, although it can sometimes happen in less than a year. In some women, precancerous cells may remain unchanged or even disap-pear without treatment. However, in some cases, precancerous cells develop into full-blown (invasive) cancers. However, in some cases, precancerous lesions become full-blown (invasive) cancers. According to the Coordinator of the Cancer Prevention and Control Directorate of the Peruvian Ministry of Health, "75% of cancer cases in Peru are diagnosed at an advanced stage, which leads to a higher risk of death from this disease. For this reason, it is necessary to reduce the mortality from cancer by promoting that people themselves take care of early detection, with preventive evaluations and access to timely care for specialised The aim of this work is to develop predictive models by correlating additional risk factors with the presence of high-risk human papillomavirus in healthy women.

METHODOLOGY

A prospective study of cross-sectional exploratory design was conducted in a group of 692 healthy women with cervical cancer in the region of Cajamarca, Peru, who were sexually active and had given informed consent. A sample of cervical epithelium was collected in liquid medium and subjected to molecular biology techniques to detect the presence of HPV genetic material of the types known to date (International Committee on Taxonomy, ICTV, 2018).

These results will be used to determine the incidence of HPV infection in the study popula-tion.

In addition, a risk factor survey was administered to women enrolled in the study to correla-te the presence of high-risk viral genetic material with specific risk factors.

All data were processed using R-3.5.2 language (R Foundation 2018). Statistical analysis was performed as a multivariate analysis to determine whether there was an association (statistical significance) between the presence of HPV infection and selected risk factors, using the chi-squared statistic of independence (X2). To estimate the relative risk of developing the disease associated with the risk factors, the cross-product ratio (CVR) or odds ratio (OR) and its 95% confidence interval were calculated using a multiple-entry contingency table.

The data to be collected will be direct, structured, participatory and individual.

The instruments to obtain the data will be of two types:

- Exfoliative cytology samples of the cervix (endo and exto) in liquid medium, which will be used for:
- Observation under normal light microscopy (PAP technique).
- Analysis of the presence of viral genetic material.
- Survey of risk factors associated with each patient by means of a structured, clo-sed, multiple-choice and individual questionnaire.
- Samples of cervical cells from the ectocervix and endocervix were collected from each woman using a cytobrush by the staff of the hospital centres of the DIRESA of Cajamarca and sent to the Molecular Biology Laboratory of the Peruvian University of Applied Sciences in Lima, Peru, together with a risk questionnaire for each pa-tient.
- A tube containing phosphate buffered saline (pH 8.6) was used for preservation, and the samples were stored at -4°C and sent to the Molecular Biology Laboratory of the Peruvian University of Applied Sciences. Once the samples arrived at the laboratory, the cytobrushes were discarded and the tubes were vortexed and centrifuged to se-diment.
- Cells were resuspended in 1 mL phosphate buffered saline and separated into three aliquots from each fresh sample. They were then stored at -20°C until HPV testing for DNA extraction, amplification and genotype sequencing. Viral genomic DNA was ex-tracted from a total volume of 200 µL of the sample using guanidinium thiocyanate (Campos M., et al. 2008) and the purified material was resuspended in a final volume of 30 µL of deionised water. Samples were subjected to electrophoresis on a 1% aga-rose gel to check DNA quality. Polymerase chain reaction (PCR) amplification of hu-man papillomavirus was performed using primers and conditions described by Lur-chachaiwong (Lurchachaiwong. et al 2011). Primers were produced by MACROGEN, South Korea. PCR products were analysed on 2% agarose gel stained with ethidium bromide and bands were detected by UV transillumination (Kodack Logic 1500, USA).

PCR products were analysed on 2% agarose gel stained with ethidium bromide and bands were detected by UV transillumination (Kodack Logic 1500, USA). Positive samples were confirmed by direct genetic sequencing (sent to Macrogen-South Korea and analysed ac-cording to International Agency for Research on Cancer (IARC, 2018) protocols). HPV geno-types were categorised into three groups: high risk, probable oncogenic and low risk, based on the IARC classification (Bouvard V et al 2009).

RESULTS

A total of 692 cases were included in the study, ranging in age from 14 to 64 years. The age distribution is shown in **Table 1**. A frequency study was performed using the chi-squared test and the group discrimination in the association between AGE and HPV-PCR was highly significant (p-value = 0.0007866). The group of women aged 21-30 years was determined by predictive modelling to be the representative group for HPV-PCR positive results.

Table N° 1. Of the sample by age.

Label/Age	Range	Frequency	%
14 – 20	[14,20]	16	2,31
21 – 30	[21,30]	174	25,14
31 - 40	[31,40]	238	34,39
41 - 50	[41,50]	206	29,77
+51	[51, +]	58	8,38
	Total	692	

A frequency study was performed using the chi-squared test and the group discrimination in the association between AGE and HPV-PCR was highly significant (p-value = 0.0007866). The group of women aged 21-30 years was determined by predictive modelling to be the representative group for HPV-PCR positive results.

Of the sample, 168 cases tested positive for HPV, representing 23.5% of the sample. The majority of HPV-positive cases are concentrated in the 20-50 age group (151 cases out of 168, or 89.88% of the total). 18 cases had two HPV strains (10.7%) and 6 cases had more than 3 HPV strains (3.5%).

Table N° 2 shows the strains found according to the International Taxonomy Committee (ICTV, 2018) and classified as high risk (HR), low risk (LR) or indeterminate risk (IR) according to the National Cancer Institute of the United States of America (NIH). **Table 3** shows the cases where more than one strain was found.

Table 2. HPV genotyping results and risk of cervical cancer. HPV AR, high-risk human papi-llomavirus. HPV BR, low risk human papillomavirus. HPV RI, human papillomavirus of inde-terminate risk.

HPV AR	N	%	HPV BR	N	%	HPV RI	N	%
16	1	0,61	6	8	4,85	34	7	4,24
18	1	0,61	30	1	0,61	44	1	0,61
31	27	16,37	35	7	4,24	67	3	1,82
33	1	0.61	39	3	1,82	73	2	1,21
45	3	1,82	42	2	1,21	74	2	1,21
53	2	1,21	51	7	4,24	90	7	4,24
58	2	1,21	52	10	6,06	91	1	0,61
59	1	0,61	54	1	1,21	101	1	0,61
			56	4	2,42			
			67	3	1,82			
			68	6	3,64			
			69	4	2,42			
			70	9	5,4			

Table 3. Genotyping results of more than 2 HPV strains, ranked according to by associated risk.

with 2 strains	N	%	with 3 strains	N	%
6 y 11	1	0,61	9,40 y 43	3	1,82
16 y 9	1	0,61	14, 20 y 21	1	0,61
16 y 52	1	0,61	39, 45 y 68	2	1,21
31 y 90	1	0,61			
31 y 53	1	0,61			
33 y 52	1	0,61			
33 y 90	1	0,61			
56 y 66	1	0,61			
66 y 90	1	0,61			
69 y 26	2	1,21			

From the bivariate analysis, the most significant data are highlighted in **Table N° 4**

Table 4. Bivariate analysis of the most significant variables:

Variables with statistical significance	Chi-squared test, p-value
Edadtcon estadocivil	0.000000347800000000
Edad y numero de parejasen los ultimos 6 meses (21a50 anos)	0,000594800000000000
Uso de preservativo y edad	0,026080000000000000
No existencia de relaciones extramaritalesy edad	0,026080000000000000
Abuso sexualy edad primera relacion sexual	0.000009299000000000
Ultimo PAP y edad	0.000000002305000000
Historia ITS y ultimo PAP	0,002134000000000000
Historia ITS y numero de abortos	0,022040000000000000
HPV .PCR y edad	0,000786600000000000
HPV.PCR y nemro de parejas en el ultimo ano	0,009154000000000000
HPV.PCR y numero de partos	0,001399000000000000

In order to look for relationships between the variables studied, a modelling exercise was carried out with the following characteristics.

Decision tree methodology is used, which are classification models that divide data into subsets based on categories of input variables. This is helpful in making decisions along the pathway (early health intervention funnel). Decision trees are in the form of a tree, where each branch represents a choice among a number of alternatives, and each leaf represents a ranking or decision. It is a model that searches the data to find the variable that allows the data set to be divided into logical groups that are most different from each other. It allows good control of missing values and is useful for pre-selecting variables.

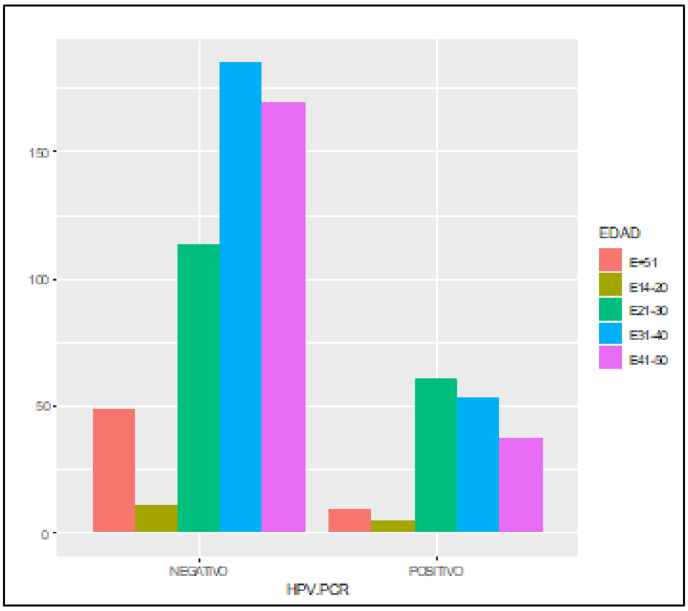
The model reflects 692 observations in its first split, of which 527, 76.16% are negative and 165, 23.84% are positive (p-value < 2.2e-16, phase 1). The graph below shows the first fitted model. It can be seen that the first (most important) split variable is AGE. This is consistent with the exploratory analysis (p-value = 0.0007866: phase 1) of the present study. The first group (HPV.PCR negative results) corresponds to the age groups 31 years and older, while the younger group of women show positive results (age group 14-30 years).

For the purposes of this study, the end nodes framed in red in the table above and shown graphically in Figure 1 are the focus of the analysis. PCR positive:

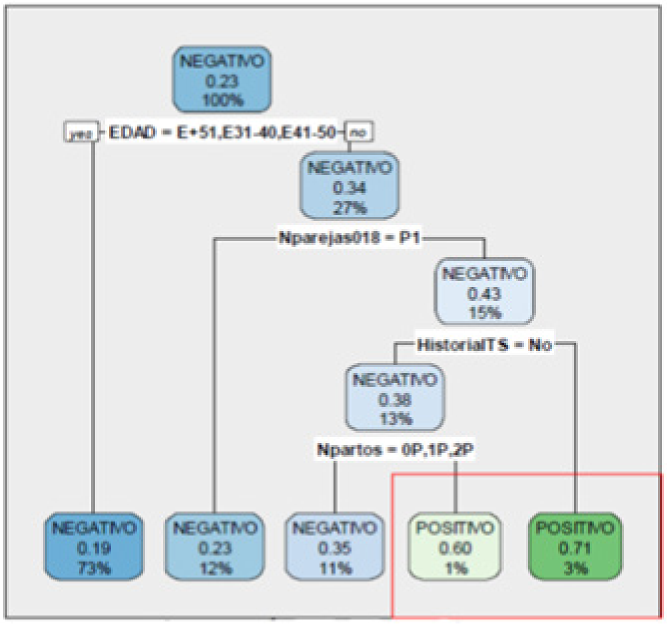
Group 1: with a probability of 0.71 (about 71%), women aged 14-30 years with 2 or more cu-rrent partners (P2, P3 and P4>) and STI history.

Group 2: with a probability of 0.60 (approx. 60%), women aged 14-30 years with 2 or more partners, but in this group there is no history of STIs and instead the number of births greater than or equal to 3 (P3, P4 and P5>) is presented as a characteristic of division.

Figure 1: HPV-PCR and AGE.



Decision Tree #1: Complete Data.



Model optimisation

Under the machine learning philosophy (supervised learning), the data is randomly divided into two subsets. The training set with 70% of the observations and the test set with the re-maining 30%. In this way, the behaviour of the model can be verified.

Machine learning is a form of AI (artificial intelligence) that allows a system to learn from data rather than through explicit programming. As the algorithm ingests training data, it is possible to produce more accurate data-driven models. A machine learning model is the output generated when you train your machine learning algorithm with data. After training, providing a model with an input will give you an output. For example, a predictive algorithm will produce a predictive model. If you then feed the predictive model data, you will get a prediction based on the data that trained the model.

In phase 1, a frequency study was performed using the chi-squared test and the group dis-crimination in the association between AGE and HPV-PCR was highly significant (p-value = 0.0007866). Predictive modelling identified the group of women aged 21-30 years as the representative group for HPV-PCR positive results.

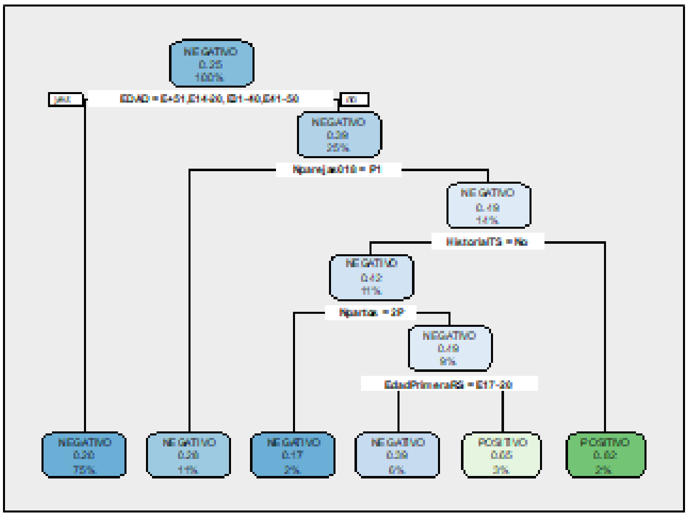
Optimised decision tree

Again for the purpose of generating risk predictors for healthy women, the two nodes of in-terest are highlighted in red in Tree N°1, distinguishing two groups (terminal nodes) sensitive to HPV results. PCR positive (decision tree 2)

Group 1: with a probability of 0.82 (approx. 82%) consists of women aged 14-20 years with 2 or more current partners (P2, P3 and P4>) and no history of STIs.

Group 2: with a probability of 0.65 (about 65%). This group includes the variables number of births (NBirths) and age at first sexual intercourse (AgeFirstRS).

Tree N° 2. train set.



FOREST GENETIC METHOD

This method is a hybrid between Random Forest (RF) and Genetic Algorithms (GA) in 3 pha-ses: normalisation, modelling and optimisation. The first phase corresponds to the prior preparation of the data set by means of normalisation functions. In the modelling phase, the objective function is determined using RF-based strategies to predict the value of the res-ponse for a given set of parameters. Finally, in the optimisation phase, the optimal combina-tion of parameter levels is obtained by integrating the properties given by our modelling scheme in the construction of the corresponding GA. This methodology allows us to focus on the most important variables resulting from the RF modelling process, which in turn allows us to more efficiently develop and control the new model generations in the optimisation phase, thus achieving significant improvements in terms of the quality objective conside-red.

The Forest-Genetic method allows the optimisation of the normalisation function due to the sparsity of the data, as in this case, and the overall performance of the algorithms (Villa-Murillo A et al. 2016).

Optimisation of forest parameters

1. Mtry: optimal number of predictors to evaluate in each partition: 1

The number of predictors used (variables versus error rate) is plotted, showing that as more variables are included, the error rate increases, demonstrating the inconvenience of mode-lling.

2. Node size: Minimum number of observations in the terminal nodes: It is possible to find 4 cases in the terminal branch of the tree when looking for predictive relationships, indicating low predictive power.

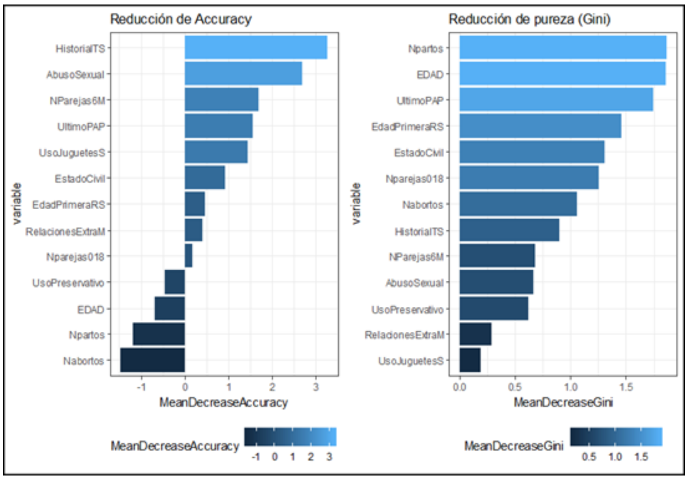
3. Optimal number of trees in the forest: At 250 splits, trees, the prediction result is stabili-sed.

Optimal model

After optimisation and supervised learning of the applied technique, Forest Genetic Method, a final model is obtained with a prediction error of 31.58% (0.3158) calculated from the test set and 34.19% (0.3419) through estimated error in these out of bag samples, known as out of bag error (OOB error). To fit the model, the values in the confusion matrix should be close to 0.

Most influential predictors: According to the behaviour of the relationships and the balance of the sample, the importance of the variables considered in this study is established. In Graph N° 1, Anova for HPV strains, where the history of sexually transmitted infections is the most relevant variable in accuracy and the number of births in the purity of the sample (concentration of the data).

Graph N° 1; Importance of the variables studied for this population.



CONCLUSIONS

Two predictive models were generated: The first model with a predictive capacity of 77.46%, where the most predictive variables are: Age, number of partners in the last year, history of sexually transmitted infections and number of births. Under the machine learning philosophy, the second model is built, this time with a predictive capacity of 74.74% evalua-ted in the test set. This model adds age at first sexual intercourse as a predictor variable in addition to those established in the first model.

Future challenges will be to focus on validating the models to find a more accurate predicti-ve tool for HPV-infected patients, and with associated relative risk factors in uninfected pa-tients, in different populations.

REFERENCES

1. American Cancer Society. Cancer Facts & Figures 2014. Atlanta, Ga: American Cancer Society; 2014.

2. Bouvard V, Baan R, Straif K, Grosse Y, Secretan B, El Ghissassi F, et al. A. review of hu-man carcinogens–Part B: biological agents. Lancet Oncol. 2009;10(4):321–2.

3. Campos M, Vilella A, Marcos MA, Letang E, Muñoz J, Salvadó E, et al. Incidence of respira-tory viruses among travelers with a febrile syndrome re turning from tropical and subtropical areas. J Med Virol. 2008;80(4):711– 5.

4. Fundación R 2018. Link: https://www.r-project.org/ International Agency for Research on Cancer, IARC, 2018 en link.: https://www.iarc.fr/media-centre/

5. International Committee on Taxonomy of Viruses (ICTV). Master Species Versión 2023, MSL #39).Revisado 26 noviembre 2024 en link: https://talk.ictvonline.org/files/master-species-lists/m/msl/7992

6. Lurchachaiwong W, Junyangdikul P, Payungporn S, Sampatanukul P, Chansaenroj J, Tresu-kosol D, et al. Human papillomavirus genotypes among infected Thai women with different cytological findings by analysis of E1 genes. New Microbiol. 2011;34(2):147–56.

7. Palacios V. Coordinador Prevención y Manejo Cáncer Cuello Uterino MINSA, Perú, 2018.

8. Villa-Murillo A, Carrión A, Sozzi A. Optimización del diseño de parámetros: Método Forest-Genetic univariante. Ciencias y Tecnología Vol. 10 No 1, Ene-Jun (2016) 12–24.