DETECTION OF CERVICAL CANCER RISK FACTORS IN VENEZUELA USING DECISION TREE ALGORITHM

*[1]Oladapo M. Oladoja & [2]Taiwo M. Adegoke

•

*[1,2]Department of Mathematics and Statistics, First Technical University, Ibadan, Nigeria
*[1]oladapo.oladoja@tech-u.edu.ng, [2]taiwo.adegoke@tech-u.edu.ng,

Abstract

Cervical cancer, a threat to female existence is one of major cancer affecting women in the developing countries of the world. Several factors are responsible which humans didn™t take cognizance of. These factors are numerous and can at times be difficult to explain using linear regression because it can™t handle many dummy variables that are not necessary to create qualitative predictors. This study uses decision trees to classify and identify the major risk factors causing cervical cancer in women depending on their age since it closely mirrors human decision making than the classical regression approach. A regression tree was constructed from the training data using recursive binary splitting. There was a minimum number of observations required for each terminal node before it stopped. Then cost complexity pruning to the large tree in order to obtain a sequence of best sub trees was applied. By using decision trees as building blocks, we can construct more powerful predictions for decision trees, bagging, random forests, and boosting. 858 cervical cancer patients were observed using 34 risk factor attributes from University Hospital of Caracas, Venezuela. Using classification trees, 14.22% of errors are produced during training. Based on the test data set, 91.5% of the predictions are correct. Based on the data set's pruned data, 91.75% of the observations can be classified correctly. Test predictions generated by this model are within 67 years of the true median age of patients, based on regression trees. Bagging and Random forest show improvement on the regression trees by setting a reduced mean square error. There are four most significant variables among all trees examined by the random forest, including age at first sexual intercourse, number of pregnancies, number of sexual partners, and hormonal contraceptives. The same goes for boosting, as a result of the relative influence statistics.

Keywords: Classification, Pruning, Bagging, Random Forest, Boosting.

1. Introduction

While routine cervical screenings, additional preventive treatments, such as HPV vaccination and safe sex practices, can often help women avoid developing cervical cancer, it is still a severe health risk for them. It's critical for women to monitor their cervical health and to consult with their doctor if they have any worries or queries. The bottom portion of the uterus that links to the vagina, known as the cervix, is typically referred to as the "cervical" in reference to females. The cervix is crucial to the health of female reproductive organs because it supports the baby throughout pregnancy and makes menstruation and birthing easier. A cervical screening test, sometimes referred to as a Pap smear or Pap test, can be used to look at the cells that line the cervix. A small sample of cervix cells is taken for this test in order to look for any abnormal cells that could point to the presence of cervical cancer or other illnesses. Although frequent cervical screenings and other preventive measures like HPV vaccination and safe sex practices can typically help avoid cervical cancer, it remains a severe health problem for women. Women should keep track of their cervical health often and consult with their doctor if they have any worries or inquiries. Hull *et al.* (2020) [1] discussed the difficulties that LMIC healthcare systems encounter in delivering cervical cancer screening and treatment. These difficulties include a lack of funding, poor infrastructure, and a paucity of qualified healthcare professionals. Moreover, barriers related to culture and society may restrict women's access to programs for cervical cancer prevention. The authors also emphasize how collaborations with international organizations help LMICs' efforts to prevent cervical cancer. They contend that in order to address the difficulties faced by women in these contexts and to lessen the incidence of cervical cancer worldwide, a concerted, international effort is required. Although the study describes the difficulties experienced in LMICs, it offers little advice on how to overcome these difficulties and enhance cervical cancer outcomes in these environments.

According to the cancer statistics, 311,000 women died of cervical cancer in 2018, which was an increase from 570,000 cases in 2017 (Arbyn *et al.*, 2020 [2]). In order to reduce the burden of cervical cancer and achieve global cervical cancer elimination goals, Arbyn *et al.* (2020) [2] argue that investments in preventative and control programs should be increased, especially in low- and middle-income countries. Despite its importance in reducing cervical cancer mortality, this study provides no information on cervical cancer screening in different regions.

Human papillomavirus infection with high-risk strains that continues over time is the main risk factor for cervical cancer (HPV). Smoking, immunosuppression, having several sexual partners, and beginning sexual activity at a young age are other risk factors. A limited resource and infrastructure make cervical cancer screening programs difficult in low- and middle-income countries, according to Zhang *et al.* (2020) [3]. Women's willingness to participate in screening programs may also be affected by social and cultural factors. There is no discussion of the advantages and disadvantages of different cervical cancer screening methods in the article, which provides a general overview of cervical cancer screening methods. Additionally, no detailed guidance is provided on how to implement the screening program. Despite the fact that a substantial share of cervical cancer cases and deaths occur in low- and middle-income countries, the article does not provide a comprehensive analysis of the unique challenges that must be overcome to prevent and treat cervical cancer in these countries.

Human papillomavirus infection with high-risk strains that continues over time is the main risk factor for cervical cancer (HPV). Smoking, immunosuppression, having several sexual partners, and beginning sexual activity at a young age are other risk factors. A limited resource and infrastructure make cervical cancer screening programs difficult in low- and middle-income countries, according to Zhang *et al.* (2020) [3]. Women's willingness to participate in screening programs may also be affected by social and cultural factors. There is no discussion of the advantages and disadvantages of different cervical cancer screening methods in the article, which provides a general overview of cervical cancer screening methods. Additionally, no detailed guidance is provided on how to implement the screening program. Despite the fact that a substantial share of cervical cancer cases and deaths occur in low- and middle-income countries, the article does not provide a comprehensive analysis of the unique challenges that must be overcome to prevent and treat cervical cancer in these countries.

Cervical cancer rates are six times higher among HIV-positive women than among HIV-negative women, and this group accounts for a substantial portion of cervical cancer cases around the world. Based on systematic reviews and meta-analyses, Stelzle *et al.* (2021) [4] assessed the robustness of their findings by conducting sensitivity analyses. They address the significance of their findings for efforts to prevent and control cervical cancer, including the requirement for women who are HIV-positive to have more access to cervical cancer screening and treatment. The study primarily uses information from studies that provide adjusted risk estimates for HIV-associated cervical cancer, which could lead to bias if confounding factor adjustments are insufficient or wrong.

The most frequent cancer that claims the lives of women in developing countries is cervical cancer. New technologies have been developed to enable cervical cancer screening and treatment that is speedier, more reasonably priced, and more sensitive. The use of self-sampling and home-based testing to increase screening uptake as well as the incorporation of biomarkers and personalized screening approaches to increase screening accuracy and decrease the number of pointless procedures are just a few of the potential future directions for cervical cancer screening discussed by Bedell *et al.* (2020) [5]. The incidence and death differences associated with cervical cancer by race and ethnicity are briefly mentioned in the article, but there is no in-depth discussion of the social and economic variables that contribute to these differences.

A major public health concern worldwide, cervical cancer is the ninth most prevalent disease in terms of new cases. Age-specific incidence and mortality trends presented by Sayo *et al.* (2022) [6] indicate that the decline in cervical cancer incidence and mortality has been greatest among women in their 20s and 30s. The Japanese Cancer Registry, one of the most trustworthy and complete cancer registration systems in the world, provided the data used in this study. The report cites regional and socioeconomic differences in cervical cancer incidence and mortality, but it doesn't go into great detail about how these differences are caused by social and economic variables. The COVID-19 pandemic, which may have altered incidence and mortality rates in recent years, may have had an impact on cervical cancer screening and treatment in Japan, but this is not taken into consideration in the study.

According to research by Rim *et al.* (2022) [7], cervical cancer is more prevalent in Uzbekistan while breast cancer is more prevalent in Korea. Also, compared to Korea, Uzbekistan has a higher mortality rate for cervical and breast cancer. The study stresses the need for enhanced screening programs in Uzbekistan and emphasizes the value of cervical cancer screening in lowering mortality rates. The study only looks at two nations, therefore its conclusions might not apply to other nations with diverse healthcare systems, risk factors, and cultural backgrounds.

Zhao *et al.* (2022) [8] identify several risk factors for cervical cancer in ethnic minorities in Yunnan Province, including advanced age, low education level, young age at first sexual experience, multiple sexual partners, smoking, misunderstanding of cervical cancer prevention, and non-participation in screening for cervical cancer. The study underlines the need for improved screening programs to raise awareness and participation, and emphasizes the significance of cervical cancer screening in lowering the incidence and mortality rates among ethnic minorities in Yunnan Province. The study uses a small sample size, which could restrict how broadly the results can be applied to other populations. The relationships between risk factors and cervical cancer may not be accurate because the study did not account for all relevant confounding variables.

Cervical cancer in women has been the subject of numerous studies, but only a small number of them used machine learning techniques. This study target at determining cervical risk factors of patients at a Venezuelan University Hospital of Caracas. 34 separate risk variables were used to monitor 858 patients.

2. Methodology

A good method for predicting a response from a single predictor variable is simple linear regression. In reality, we frequently use multiple predictors. Giving each predictor in a single model a unique slope coefficient enables us to achieve this. Assume that there are p different predictors in general. The multiple linear regression model then adopts the following form.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{1}$$

where ϵ is independently and identically distributed with mean zero and variance σ^2 The regression technique may result in accurate predictions on the training set, but it is likely to over-fit the data and perform poorly on the test set. This is because a tree that could be too complex. A smaller tree with fewer splits can lead to lower variance and better interpretation at the cost of some bias. An alternative to the method mentioned above is to only build the tree if the decrease in RSS brought on by each split is greater than a predetermined (high) cutoff point. This strategy will result in smaller trees, however it is too naive because a split that at first glance seems pointless in the tree may be followed by a really good split.

2.0.1 Algorithm 1: Building a Regression Tree

1. Using the training data, construct a huge tree using recursive binary splitting. Only stop when each terminal node has fewer than a predetermined number of observations.

2. Create a list of the top subtrees that are functions of $alpha$ by applying cost complexity pruning to the massive tree.

3. Choose alpha using K-fold cross-validation. Namely, create K folds from the training observations. When $k = 1, ..., K$:

    - Repeat steps 1 and 2 for every fold of the training data other than the kth fold.
    - Analyze the mean squared prediction error as a function of α for the data in the kth fold that were excluded.

    Calculate the results for each value of alpha, then choose the one with the lowest average error.

4. The Step 2 subtree that corresponds to the chosen alpha value should be returned.

2.1. Bagging, Random Forests and Boosting

2.1.1 Bagging

Bagging can enhance predictions for various regression techniques, but decision trees benefit the most from its use. We merely create B regression trees using B bootstrapped training sets, average the outcomes, and apply bagging to the regression trees. These trees are not manicured and are grown deeply. As a result, each tree has a high variation but a low bias. These B trees are averaged to lessen volatility. Bagging has been proven to dramatically improve accuracy by combining hundreds or even thousands of trees into a single step. We create B different bootstrapped training data sets, train our algorithm on the bth bootstrapped training set, and then use that data to generate B additional bootstrapped training data sets in order to obtain and eventually average all the predictions.

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \tag{2}$$

It's referred to as bagging.

2.1.2 Random Forests

Random forests provide an improvement over bagged trees by randomly making minor adjustments to the trees' decorations. We build a number of bagging-like decision tree forests using bootstrapped training samples. However, each time a split in a tree is taken into account when building these decision trees, a random sample of m predictors is chosen as split candidates from the whole collection of p predictors. One of those million dollar forecasters is restricted to the split. Even the majority of the potential predictors at each branch in the tree cannot be considered by the algorithm when building a random forest.

2.1.3 Boosting

Similar to other tree-growing techniques, boosting grows the trees in a certain order utilizing data from earlier trees. With boosting, each tree is fitted to a modified version of the original data set rather than using bootstrap sampling.

2.1.4 Algorithm 2: Boosting for Regression Trees

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.
2. For $b = 1, 2, ..., B$, repeat:
   - Fit a tree $\hat{f}^b$ with d splits (d+1 terminal nodes) to the training data (X, r).
   - Update $\hat{f}$ by adding in a shrunken version of the new tree:
   $$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x) \tag{3}$$
   - Update the residuals,
   $$r_i \leftarrow r_i - \lambda \hat{f}^b(x) \tag{4}$$
3. Output the boosted model,
$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \tag{5}$$

3. Results and Discussion

Both regression and classification problems can be solved using decision trees. We first evaluate the data set using classification trees. Age is a continuous variable in these data, therefore we start by re-coding it as a binary variable that has a value of Yes if the age variable is more than 20, and a value of No otherwise. 14.22% of training errors are made. A tree that offers a good fit to the (training) data will have a minimal deviation. Just divide the deviation by $n - |T_0|$ to get the reported residual mean deviance, which in this instance is $858 - 9 = 849$.

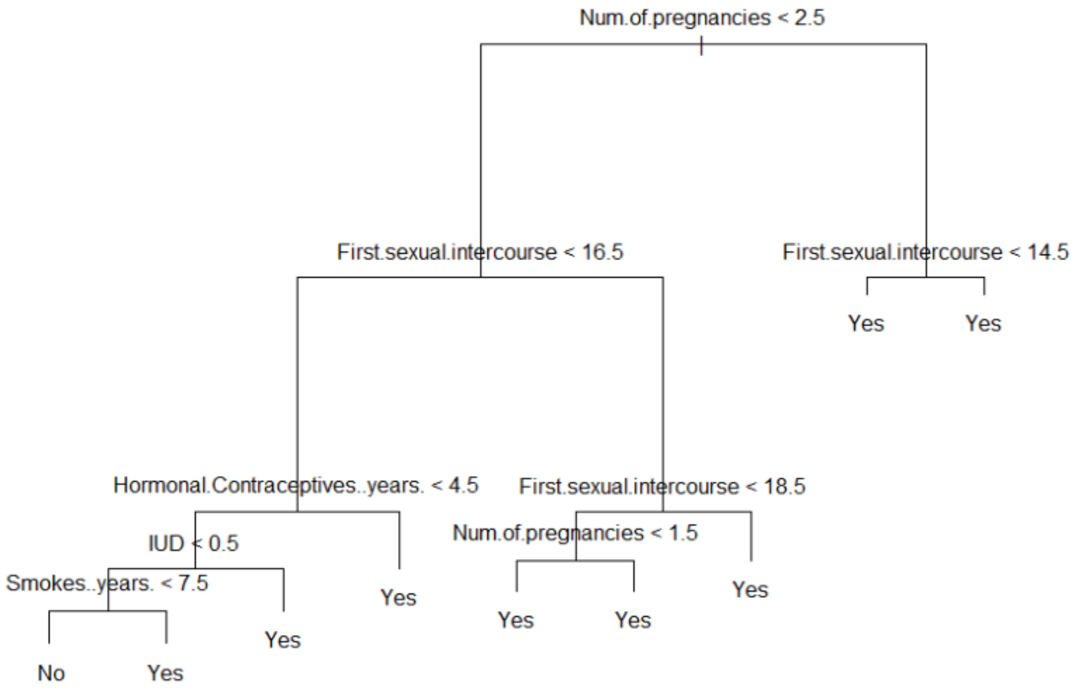
Num.of.pregnancies < 2.5
First.sexual.intercourse < 16.5
First.sexual.intercourse < 14.5
Yes
Yes
Hormonal.Contraceptives..years. < 4.5
First.sexual.intercourse < 18.5
IUD < 0.5
Num.of.pregnancies < 1.5
Smokes..years. < 7.5
Yes
Yes
Yes
Yes
Yes
No
Yes

Figure 1: *Regression Trees for Cervical Cancer in Venezuela*

According to Figure 1, which illustrates that the number of pregnancies is a significant risk factor component in determining the patients having cervical cancer depending on their age. The number of pregnancies appears to be the most significant indication of the cervical cancer of patients. Asterisks are used to denote branches that lead to terminal nodes. Instead of just estimating the training error, we must estimate the test error in order to assess a classification tree's performance on this data. The observations are divided into a training set and a test set, the tree is built using the training set, and its performance is assessed using the test set. In the test data set, this method yields accurate predictions for about 90.5% of the locations.

Table 1: *Machine Learning Accuracy Classification*

Tree.Pred	Age	High Test
	No	Yes
No	82	57
Yes	34	285

Now that the pruning procedure has been completed, 91.75% of the test observations are correctly classified, increasing both the interpretability of the tree and the classification accuracy. We get a larger pruned tree with lower classification accuracy if we increase the value of best.

A regression tree is used to fit the data set in this case. We initially create a training set before adjusting the tree to the training data. The findings demonstrate that the tree was constructed using only five of the factors. The total squared errors for the tree constitute the deviation in the context of a regression tree. In accordance with the cross-validation results, we make predictions on the test set using the unpruned tree. As shown in Table 2, we apply bagging and random forests on the data. Random forests beat bagging and regression trees in this case, according to the test set MSE, in terms of outcomes.

Table 2: *Mean Square Error (MSE) for the Test Set*

Algorithm	MSE
Regression Tree	46.05472
Bagging	38.3288
Random Forest	36.65896

There are two reported measurements of varied importance. The %IncMSE is based on the average decline in prediction accuracy on out-of-bag samples when a particular variable is left out of the model. The IncNodePurity is an average over all trees measurement of the total reduction in node impurity caused by splits over that variable (this was plotted in Figure 2).

Table 3: *Importance of Variables*

Variables	%IncMSE	IncNodePurity
Number.of.sexual.partners	13.935063364	1979.451415
First.sexual.intercourse	36.935258441	6489.721033
Num.of.pregnancies	40.324446650	8720.821614
Smokes	1.658543003	188.868735
Smokes..years.	7.417078338	1725.845382
Smokes..packs.year.	3.645987343	630.046003
Hormonal.Contraceptives	5.145181475	448.446348
Hormonal.Contraceptives..years.	19.550078277	3031.607582
IUD	10.366088887	1300.537849
IUD..years.	12.001987891	1696.846107
STDs	1.008508253	148.194058
STDs..number.	2.751653121	171.491002

The training RSS is used to assess node impurity in regression trees, whereas the deviation is used to measure it in classification trees. The findings show that the number of sexual partners, the first sexual encounter, and the number of pregnancies are by far the three most crucial variables across all of the trees taken into account in the random forest.

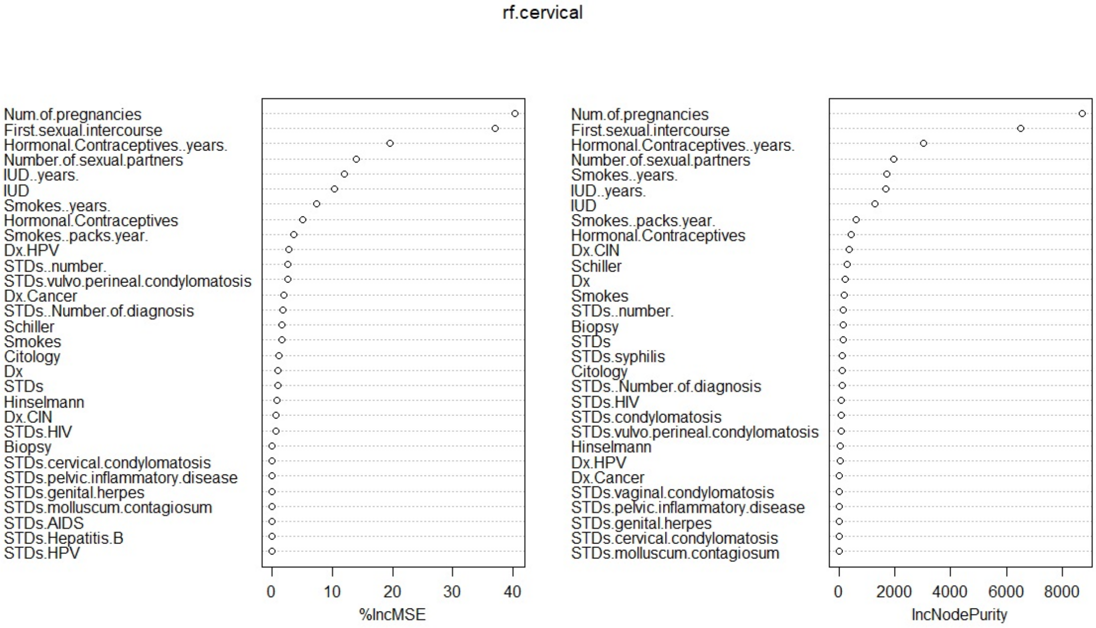

Figure 2: *Plot of Importance Measures*

Relative influence statistics are provided by boosted regression trees along with a relative influence plot as shown in Table 4. We can observe that the most significant factors are clearly the first sexual encounter, number of pregnancies, use of hormonal contraceptives throughout time, and number of sexual partners. Plots of the partial dependence between these four variables are also possible as shown in Figure 3. After integrating out the other factors, these charts show the marginal impact of the chosen variables on the response. Now, we employ the boosted model to forecast the test set's age at risk for developing cervical cancer. The test MSE obtained is 36.99571, which is higher than that for bagging and comparable to the test MSE for random forests. In boosting, smaller trees are frequently sufficient since the growth of a specific tree takes into account the other trees that have already been developed, as opposed to random forests. Interpretability can also be improved by utilizing smaller trees; for example, using stumps results

Table 4: *Relative Influence Statistics*

Variables	Rel.inf
First.sexual.intercourse	31.900870295
Num.of.pregnancies	23.007299808
Hormonal.Contraceptives..years.	15.871948215
Number.of.sexual.partners	10.889051795
Smokes..years.	3.356270114
IUD..years.	3.217962490
Hormonal.Contraceptives	2.848841347
Smokes..packs.year.	2.728422340
Schiller	2.150597021
IUD	1.786724939
STDs	0.552829213
Smokes	0.509029318
Biopsy	0.391805326
Citology	0.241696894
Dx	0.159287734
STDs..number.	0.147483560
STDs..Number.of.diagnosis	0.136255508
STDs.condylomatosis	0.076823947
Hinselmann	0.018564423
STDs.vulvo.perineal.condylomatosis	0.008235713

in an additive model.

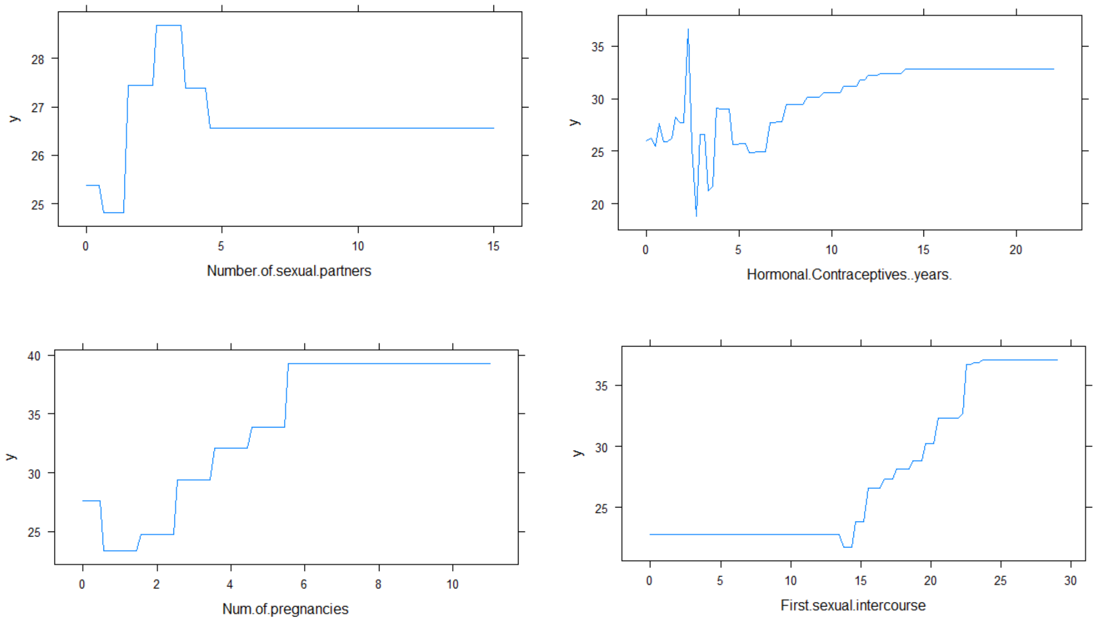

Figure 3: *Plot of Partial Dependence*

4. Conclusion

Regression and classification problems can be solved using decision trees. However, tree-based methods cannot compete with the most appropriate supervised learning approaches in terms of their ease and effectiveness for analysis. In this study, four machine learning algorithms were used namely, regression trees, bagging, random forests and boosting in determining the risk factors of cervical cancer patients due to their age. Random forest and boosting outperformed bagging and regression trees due to their mean square errors. It was discovered that number of pregnancies, first sexual intercourse, number of sexual partners and hormonal contraceptives are the most important risk factors to determine cervical cancer in women depending on their age. Concerned authorities in Venezuela need to take cognizance of the four variables in order to curb cervical cancer.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Hull, R., Mbele, M., Makhafola, T. Hicks, C., Wang, S., Reis, R.M., Mehrotra, R., Mkhize Kwitshana, Z. Kibiki, G., Bates, D.O.and Dlamini, Z. (2020). Cervical Cancer in Low and Middle Income Countries (Review). *Oncology of Letters* 20(3); 2058-2074. https://doi.org/10.3892/ol.2020.11754.

[2] Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosre, S., Saraiya, M., Ferlay, J., and Bray, F. (2020). Estimates of Incidence and Mortality of Cervical Cancer in 2018: A Worldwide Analysis. *The Lancet Global Health*, 8(2), e191-e203. https://doi.org/10.1016/S2214-109X(19)30482-6

[3] Zhang, S., Xu, H., Zhang, L., and Qiao, Y. (2020). Cervical Cancer: Epidemiology, Risk Factors and Screening. *Chinese Journal of Cancer Research*, 32(6), 720-728. https://doi.org/10.21147/j.issn.1000-9604.2020.06.05

[4] Stelzle, D., Tanaka, L. F., Lee, K. K., Ibrahim Khalil, A., Baussano, I., Shah, A. S. V., McAllister, D. A., Gottlieb, S. L., Klug, S. J., Winkler, A. S., Bray, F., Baggaley, R., Clifford, G. M., Broutet, N., and Dalal, S. (2021). Estimates of the Global Burden of Cervical Cancer Associated with HIV. *The Lancet Global Health*, 9(2), e161-e169. https://doi.org/10.1016/S2214-109X(20)30459-9

[5] Bedell, S.L, Goldstein, L.S., Goldstein, A.R. and Goldstein, A.T.(2020). Cervical Cancer Screening: Past, Present, and Future. *Sexual Medicine Reviews*, 8(1), 28–37, https://doi.org/10.1016/j.sxmr.2019.09.005

[6] Sayo, T., Matthew, P. and Kota, K. (2022). Trends in Cervical Cancer Incidence and Mortality of Young and Middle Adults in Japan. *Cancer Science*, 113(5),1801-1807. DOI:10.1111/cas.15320

[7] Rim, C.H., Lee, W.J., Musaev, B., Volichevich, T.Y., Pazlitdinovich, Z.Y., Lee, H.Y., Nigmatovich, T.M. and Rim, J.S. (2022). Comparison of Breast Cancer and Cervical Cancer in Uzbekistan and Korea: The First Report of The Uzbekistan–Korea Oncology Consortium. *Medicina*, 58, 1428. https://doi.org/10.3390/medicina58101428

[8] Zhao, M., Luo, L., Jia, Y., and Chen, B. (2022). Risk factors of Cervical Cancer Among Ethnic Minorities in Yunnan Province, China: A case–control study. *European Journal of Cancer Prevention*, 31(3), 287-292. https://doi.org/10.1097/CEJ.0000000000000704

[9] Obisesan, K.O. and Oladoja, O.M. (2022). On Normal Process of Diffusion Equation in Monitoring Carbon Monoxide Concentrations in Nigeria. *International Journal of Statistical Distributions and Applications*. 8(2)