Predicting the risk factors of Cervical Cancer using Data Mining Techniques

Pronob Barman

Cervical Cancer Data Set Analysis

Introduction

The data set used for this project is publicly available through the UCI Machine Learning Repository (Cervical Cancer Data Set). The dataset was collected at "Hospital Universitario de Caracas" in Caracas, Venezuela. The dataset comprises demographic information, habits, and historic medical records of 858 patients. It contains 32 variables and 4 target variables. Below are the variables that used in our analysis:

Attribute	Type	Attribute	Type	Attribute	Type
Age	Integer	STDs	Bool	STDs:HIV	Bool
Number of sexual partners	Integer	STDs (number)	Integer	STDs:Hepatitis B	Bool
First sexual intercourse (age)	Integer	STDs:condylomatosis	Bool	STDs:HPV	Bool
Number of pregnancies	Integer	STDs:cervical condylomatosis	Bool	STDs: Number of diagnosis	Integer
Smokes	Bool	STDs:vaginal condylomatosis	Bool	STDs: Time since first diagnosis	Integer
Smokes (years)	Bool	STDs:vulvo-perineal condylomatosis	Bool	STDs: Time since last diagnosis	Integer
Smokes (packs/year)	Bool	STDs:syphilis	Bool	Dx:Cancer	Bool
Hormonal Contraceptives	Bool	STDs:pelvic inflammatory disease	Bool	Dx:CIN	Bool
Hormonal Contraceptives (years)	Integer	STDs:genital herpes	Bool	Dx:HPV	Bool
IUD	Bool	STDs:molluscum contagiosum	Bool	Dx	Bool
IUD (years)	Integer	STDs:AIDS	Bool		

Figure 1: Attributes and their types

And the four target variables are: Hinselmann: bool, Schiller: bool, Cytology: bool, Biopsy: bool. These four target variables were used as screening test to detect the cervical cancer. To make our analysis more easier and understandable, thus, we combined these four target variables as

"Cancer". Here, Cancer=Hinselmann+Schiller+Cytology+Biopsy

$$Cancer = \begin{cases} 1 & \text{if any of the four tests is positive} \\ 0 & \text{if all of the four tests are negative} \end{cases}$$

Now, cancer is our new outcome variable. This data analysis is appropriate for data mining. Because we have lots of variables in our dataset and we will find our best model to accurately predict the outcome variable. And this the main purpose of data mining. Therefore, we believe that this dataset is completely accurate for data mining. If we could successfully predict the outcome variable, people would aware of the causes of cervical cancer. Also, government or any health policy organization could implement new policies to prevent cervical cancer death.

Exploration of the Data

At first, we looked at the number of cases of our outcome variable, cancer.

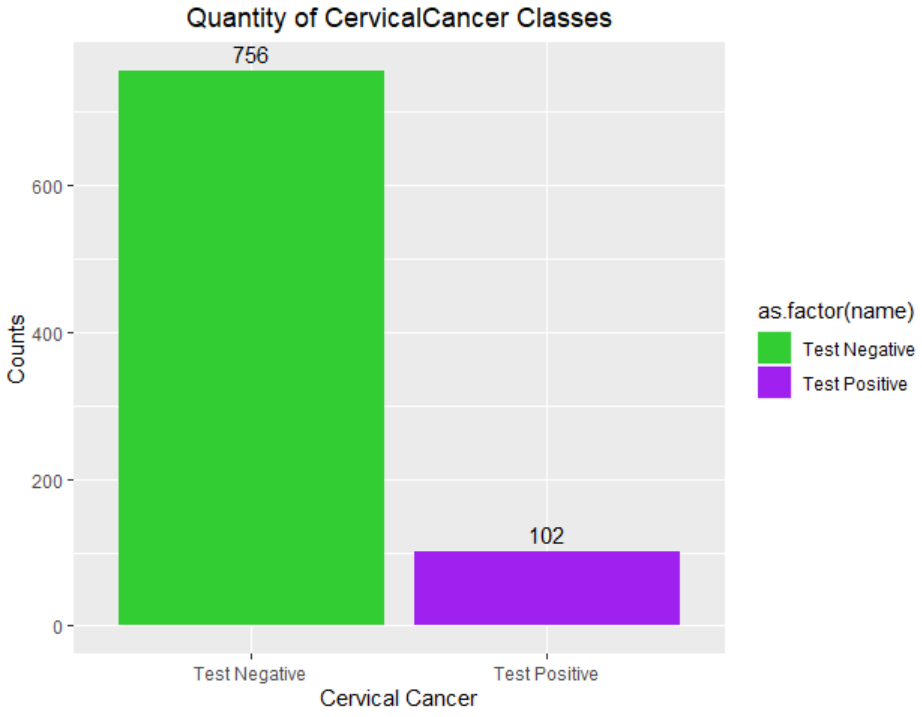

Figure 2: Cervical Cancer Cases

From the bar plot [1] we see that there were 102 patients who had cervical cancer out of 858 patients which is approximately 12% of the total observations. Before going any further step, we need to check whether data contains any missing observations. We found that data contains only

[1] used ggplot2 [8] package to draw the graph

59 complete cases. So, we count the missing observations and make a graphical representation of it.

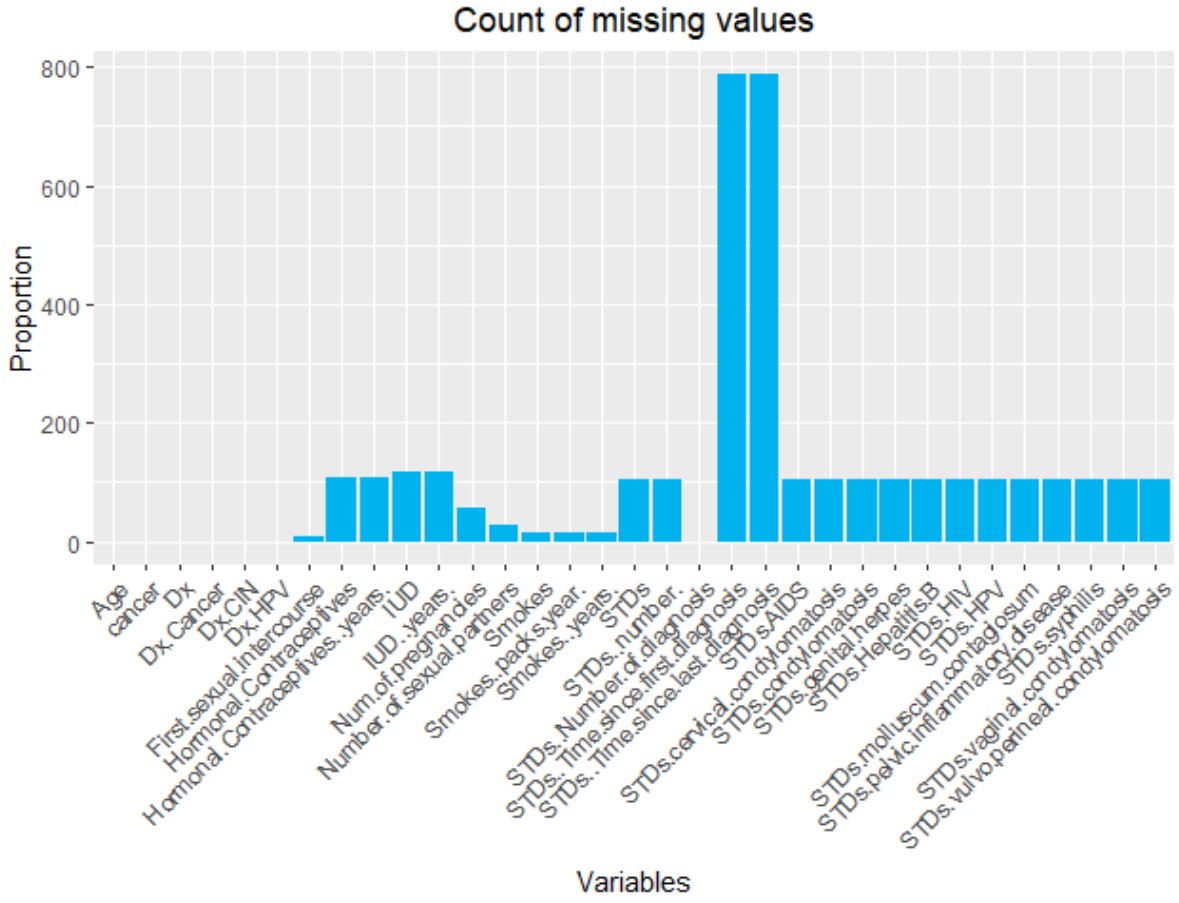

Figure 3: Missing Data Observation

From the graph we see that there were lots of missing observations in our dataset. Specially if we look at the middle of our bar chart, we see two bars are significantly higher than the rest of the bars. These two bars are STDs Time since first diagnosis and STDs Time since last diagnosis. Since, most of the observations of these two variables are missing (approximately 93%), therefore, we dropped these two variables without any consideration. Still some of the variables have around 12% missing observations. So, we imputed the missing observations. There are lots of ways we can impute the missing observations. In R, there is a package called "MICE" [6] is very popular for imputing the missing observations. MICE assume that the missing data are missing at random, which means that the probability that a value is missing depends only on observed value and can be predicted using them. It imputes data on a variable by variable basis by specifying an imputation model per variable. We used predictive mean matching as imputation method here. For checking the data structure [2] of each continuous explanatory variables it was found that all

[2]Hmisc [1] package were used

of the explanatory variables are rightly skewed except the first sexual intercourse.

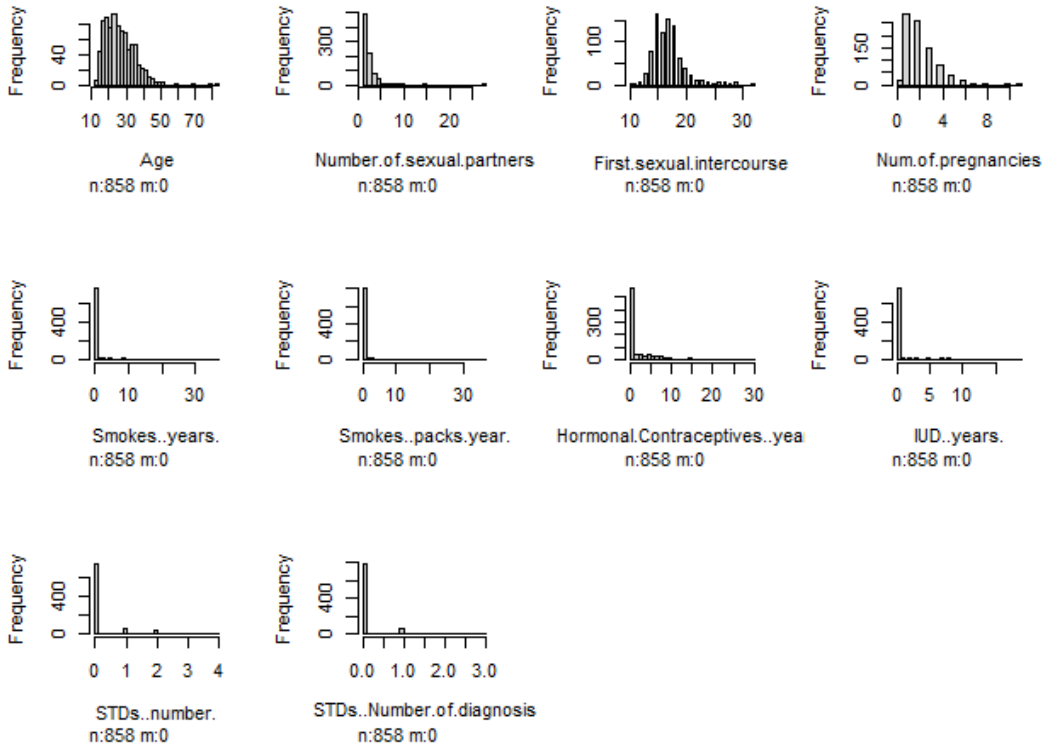

Figure 4: Distribution form of all continuous explanatory variables

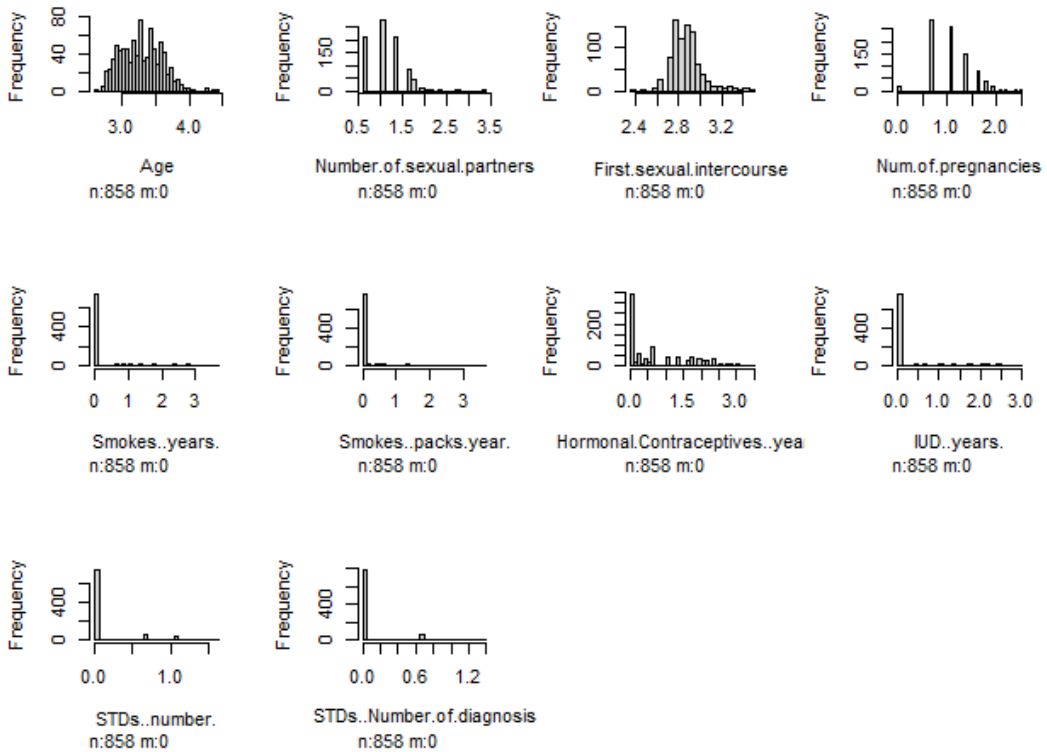

Figure 5: Distribution form of all continuous explanatory variables after log transformation

First sexual intercourse follows very closed to normal shape. Since, there were lots of 0 values present in the dataset. Therefore, a modified log transformation log(variable+1) was taken here. In the Figure 5, some of the variables are now closed to normal and some of them still rightly skewed.

The dataset contains 30 variables now. That implies a high-dimensional data, in terms of number of features. Therefore, a Boruta analysis [3] was taken to identify the most important features for cervical cancer among the 30 variables. Boruta analysis captures all the important features in the dataset with respect to an outcome variable. At first, it duplicates the original dataset by shuffling the predictors' values known as shadow features and make a random forest by combining the shadow features with the original predictors. After that, make a comparison between original variables with randomized variables to measure their importance. And choose the one which have higher importance ratio than that of randomized variables.

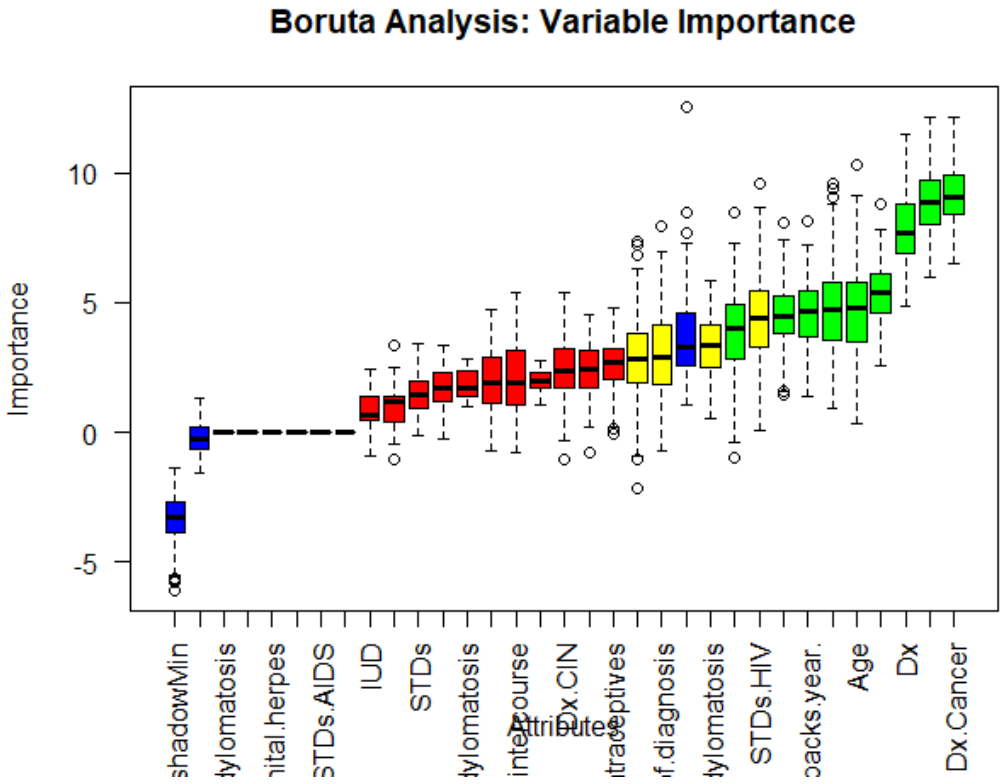

Figure 6: Boruta Analysis: Variable importance

After performing Boruta method we got 11 variables as the most important features for cervical cancer. Here is the list of selected variables: Age, Number of pregnancies, Smokes (years), Smokes (packs/year), Hormonal Contraceptives (years), STDs (number), STDs condylomatosis, STDs HIV, Dx Cancer, Dx HPV, Dx. Now, we checked the each selected variables' distribution form with cervical cancer. So, we fitted boxplot with cervical cancer.

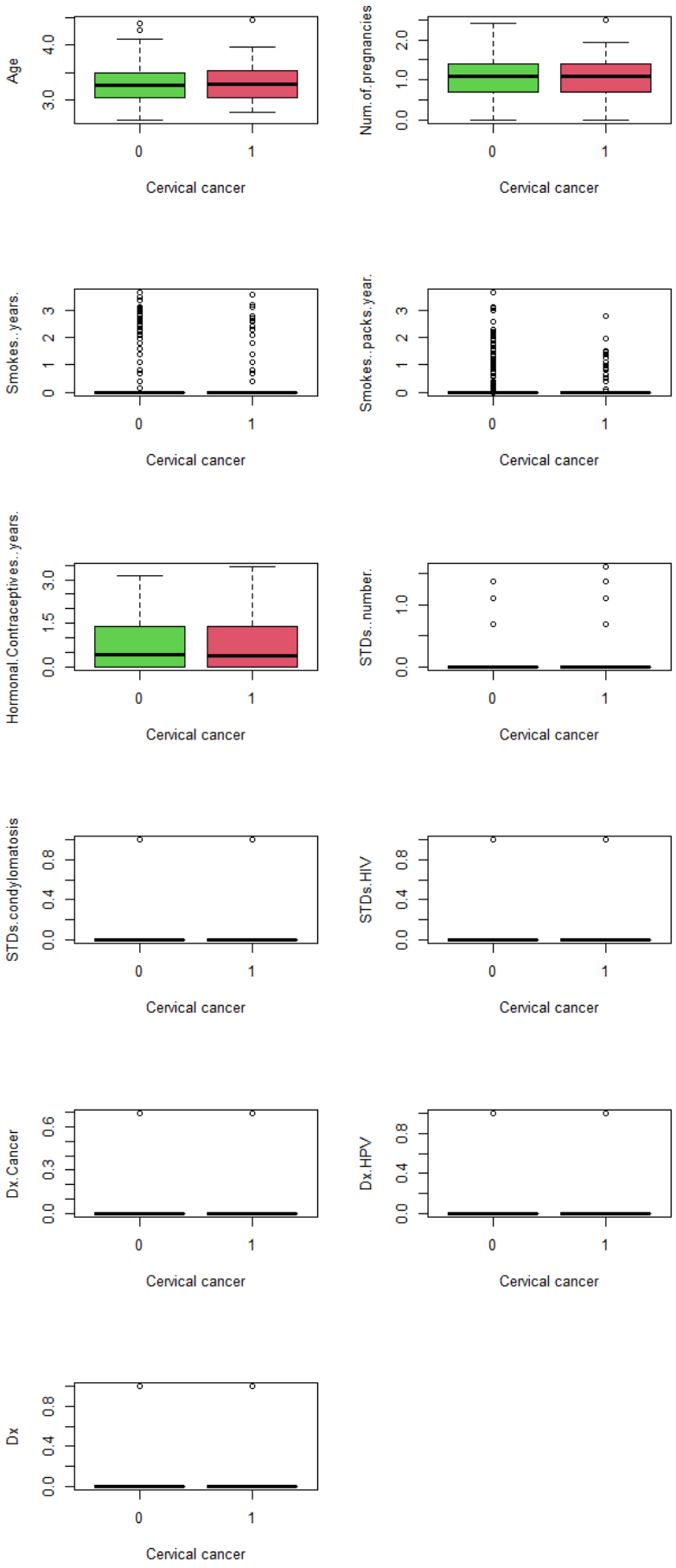

Figure 7: Boxplot of the selected variables with cervical cancer

From the Figure 7, we see that some of the variables follow very closed to normal shape. Also, there are some outliers present in our dataset. However, that does not hinder our future analysis. Therefore, we did not do anything with outliers.

Assessment of Predictions

Before we performed our data analysis we randomly partitioned [3] our dataset into two parts: Training and Validation. Data was randomly assigned to either group with probabilities of $\frac{3}{4}$ and $\frac{1}{4}$ respectively. In training dataset, 76 (11.82%) patients had cancer out of 643 patients. And in the validation dataset, 26 (12.09%) patients had cancer out of 215 patients.

For predicting the cervical cancer we perform three models.

1. Linear Discriminant Analysis (LDA)
2. Decision Tree
3. Random Forest

1. Linear Discriminant Analysis (LDA):

Before performing LDA [4], we needed to examine whether the eigenvalues of the sample variance/covariance matrix (consisting of all explanatory variables) are close to 0. And we found that no eigenvalues were closed to 0. Therefore, we did not need to drop any variables from our model. The results from our LDA analysis are found in Table 1.

	Validation
Correct Classification Rate	0.8605
Sensitivity	0.0000
Specificity	0.8768
Positive Predictive Value	0.0000
Negative Predictive Value	0.9788

Table 1: LDA model performance: Sensitivity, Specificity, PPV, and NPV

[3] tidyverse [9] and caret [2] packages were used
[4] Used MASS [7] package to run

2. Decision Tree Model:

We performed the decision tree [5] with all of the variables and found the results shown in Table 2.

	Validation
Correct Classification Rate	0.8651
Sensitivity	0.0000
Specificity	0.8774
Positive Predictive Value	0.0000
Negative Predictive Value	0.9841

Table 2: Decision tree model performance: Sensitivity, Specificity, PPV, and NPV

The results suggest that decision tree model perform better prediction than LDA.

3. Random Forest Model:

We performed the random forest [6] with all of the variables and found the results shown in Table 3.

	Validation
Correct Classification Rate	0.8791
Sensitivity	0.5000
Specificity	0.8826
Positive Predictive Value	0.0385
Negative Predictive Value	0.9947

Table 3: Random forest model performance: Sensitivity, Specificity, PPV, and NPV

The results suggest that random forest model perform better prediction than LDA and decision tree models.

Discussion:

From the Table 1, Table 2, and Table 3, we found that random forest model perform better than the LDA and decision tree models. Even though random forest model's correct classification rate and specificity rate are closed to LDA and decision tree models. Nonetheless, it has the highest sensitivity rate (50%). Therefore, random forest model can classify the presence of cervical cancer better than LDA and decision tree models. Random forest model can gain the high values of sensitivity and specificity when ensuring the accuracy.

[5] Used rpart [5] package

[6] Used randomForest [4] package

With this limited information the data mining analysis have been done successfully. We could have been got a better prediction by using a more advanced model. Additionally, if we would have had more information and more details about each variables' characters, we could have improved our data mining analysis better.

Several difficulties were encountered during the analysis. The major difficulty was data contains ample of missing observations. Even though we used "MICE" package to impute the missing observations. However, this procedure could increase the biased. Also, log transformation would not be helpful for following normal distribution. Other explanatory variables could have been more important for cervical cancer than our selected variables. Nonetheless, we overlooked this matter.

If the analysis were starting over, a more rigid data imputation method would be used. Moreover, a better way of selecting the best important features would be implemented. A more advanced algorithm would be applied for prediction in cervical cancer.

In the future research, it is recommended to use a better data imputation method. Also, it is recommended to use a more informative dataset rather than our dataset or gather more information about each variable of our dataset. Besides, other advanced model can be applied to enhance the clarity and quality of the results. Moreover, find a better way of diagnose cervical cancer rather than using the combination of the four screening test.

References

[1] Frank E Harrell Jr, with contributions from Charles Dupont, and many others. *Hmisc: Harrell Miscellaneous*. R package version 4.5-0. 2021. URL: https://CRAN.R-project.org/package=Hmisc.

[2] Max Kuhn. *caret: Classification and Regression Training*. R package version 6.0-86. 2020. URL: https://CRAN.R-project.org/package=caret.

[3] Miron B. Kursa and Witold R. Rudnicki. "Feature Selection with the Boruta Package". In: *Journal of Statistical Software* 36.11 (2010), pp. 1–13. URL: http://www.jstatsoft.org/v36/i11/.

[4] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22. URL: https://CRAN.R-project.org/doc/Rnews/.

[5] Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15. 2019. URL: https://CRAN.R-project.org/package=rpart.

[6] Stef van Buuren and Karin Groothuis-Oudshoorn. "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45.3 (2011), pp. 1–67. URL: https://www.jstatsoft.org/v45/i03/.

[7] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: http://www.stats.ox.ac.uk/pub/MASS4/.

[8] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[9] Hadley Wickham et al. "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43 (2019), p. 1686. DOI: 10.21105/joss.01686.