Cervical Cancer Risk: A Case Study

Authors: Giulia Saresini (matr. 864967), Nicola Perani (matr. 864755), Sara Nava (matr. 870885)
University of Milano-Bicocca – Master's Degree Program in Data Science

Abstract

Cervical cancer poses a significant global health threat, with 11,000 new cases in the U.S. annually. This project, based on a Kaggle dataset, aims to develop a predictive model for identifying high-risk cervical cancer cases requiring biopsy. The dataset, with 36 variables, underwent meticulous pre-processing, including handling of imbalanced data. The study explores model robustness, variable selection, and confidence interval analysis, with a focus on balanced training data. Results show high accuracy, with J48 standing out for robustness and high Recall. Future research involves exploring ensemble methods and fine-tuning models for improved cervical cancer risk assessment.

Index

1 – Introduction 1

2 – Pre-processing 2

2.a – Categorical Variable Processing and Recoding..... 2

2.b –Missing Data Handling 2

2.c – Class Imbalance Problem..... 2

2.e – Training Set Balancing..... 3

2.f – Data Scaling..... 3

3 – Model Training..... 3

3.a – Classification Models 3

3.b – Feature Selection and Data Scaling..... 3

3.c – Models Robustness Evaluation..... 3

3.d – Confidence Interval Evaluation..... 4

3.e – Sensitivity and Specificity Evaluatiom 5

3.f – ROC curve Evaluation..... 5

4 – Test Results 6

5 – Conclusions 6

6 – References..... 7

1 – Introduction

Cervical cancer, responsible for approximately 11,000 new cases of invasive cervical cancer diagnosed each year in the United States, poses a significant threat to women's health globally. Despite a steady decrease in new cases over the past decades, this form of cancer continues to claim lives of about **4,000 women in the United States** and **over 300,000 women worldwide**.

The current project aims to **explore the risk factors** associated with cervical cancer **to identify critical cases in need of a biopsy.** A dataset obtained from Kaggle [1], sourced from the UCI Repository, has been utilized.

The primary challenge lies in the accurate classification of cervical cancer risks, using the **biopsy** outcomes **as the target variable**. This project seeks to develop a predictive model capable of precisely identifying high - risk cases to implement targeted preventive interventions. Understanding the underlying risk factors is crucial for optimizing screening strategies and further reducing the incidence of this disease.

The dataset, collected from a cohort of female patients, provides a wide range of information, including 36 variables:

- *Age.*
- *Number of sexual partners.*
- *First sexual intercourse*: age at first sexual intercourse.
- *Num of pregnancies:* Number of pregnancies.
- *Smokes, Smokes (years), Smokes (packs/year):* Information on smoking habits.
- *Hormonal Contraceptives, Hormonal Contraceptives (years):* Use of oral contraceptives and duration of use.
- *IUD, IUD (years)*: Use of intrauterine devices and duration of use.

- *STDs, STDs (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, STDs:Hepatitis B, STDs:HPV:* Information on sexually transmitted diseases.
- *STDs: Number of diagnoses, STDs: Time since first diagnosis, STDs: Time since last diagnosis*: Number of diagnoses and time passed since diagnoses of sexually transmitted diseases.
- *Dx:Cancer, Dx:CIN, Dx:HPV, Dx*: Results of cervical cancer diagnoses and related conditions.
- *Hinselmann, Schiller, Citology*: Outcomes of specific examinations related to cervical cancer.
- *Biopsy*: Undergoing a biopsy, used as the target variable.

2 – Pre-processing

During this phase, several transformations were implemented to ensure **data consistency and quality**.

2.a – Categorical Variable Processing and Recoding

A crucial step involved the **treatment of categorical variables**, initially imported as double variables from the original dataset. To avoid the presence of decimal places at the end of characters, resulting from their initial representation as double, we first converted these variables to integers and subsequently to strings.

Binary categorical variables were represented using numerical values 0 and 1, following common practices in statistical analyses. However, despite being binary, we recoded for clarity the response variable ***Biopsy***, assigning *No* to the value 0 and *Yes* to the value 1. This modification aims to clearly differentiate the ***Biopsy*** variable from others in the dataset, thereby enhancing the understanding of the results.

2.b –Missing Data Handling

We decided to systematically **exclude** attributes exhibiting **more than 30% missing data** to mitigate potential distortions stemming from imputation methods such as mean, median, or mode, which might lead to an unwarranted homogenization of attribute distributions.

Regarding other instances of missing data, we implemented a **conditional imputation strategy**, contingent on the response variable ***Biopsy***. For numerical variables, we performed the imputation utilizing the **median** instead of the mean, given its resistance to the influence of outliers. Conversely, for string variables, imputation was conducted based on the **mode**, representing the most frequently occurring value.

2.c – Class Imbalance Problem

Exploratory data analysis provided an overview of the class distribution in the dataset. We observed a **significant imbalance**, with **53 positive cases** and **700 negative cases**, highlighting the need to address this aspect in the subsequent stages of the project. Awareness of this imbalance is crucial for an accurate assessment of model performance and for making informed decisions during project development.

This is particularly important as, in this case, the variable ***Biopsy*** indicates whether a biopsy procedure is recommended or not. ***Biopsy*** involves the removal of tissue from the cervix to test for abnormal or precancerous conditions, or cervical cancer. The primary objective was to avoid false negatives, ensuring that the model does not overlook cases where a biopsy is strongly recommended, thus maintaining a focus on effective cancer detection.

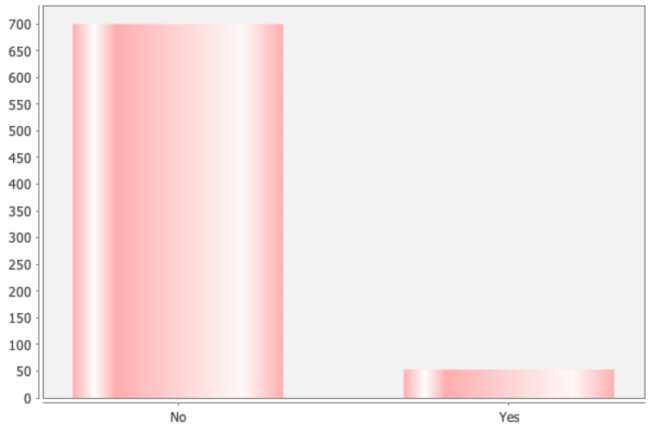

Figure 1: Display data distribution with the respect to the class variable to check the presence of class imbalance problem.

2.d – Dataset Partitioning

After completing the data exploration phase, we proceeded with **partitioning the dataset**. The adopted methodology was that of **stratified sampling** concerning the ***Biopsy*** target variable, with an allocation of **80%** of the data to the **training set** and **20%** to the **test set**.

The choice to use stratified sampling was motivated by the need to maintain a proportional distribution of classes within the target variable in both the training and test sets. This approach ensures a more accurate representation of the original dataset's characteristics in both sets, avoiding imbalances that could adversely affect the model's performance.

2.e – Training Set Balancing

Once the training dataset was obtained, we first addressed the issue of imbalanced data mentioned in the previous section [2.c]. In this context, we chose to pursue two distinct approaches by **testing the models on both** properly **balanced and imbalanced data**. This decision was made after consulting various sources and scientific references that presented contrasting yet valid opinions on the matter [2][3][4].

In both scenarios, the **validation dataset** was generated through an additional partition of our training set (using the same partitioning used to split data into training and test set [2.d]), **maintaining the imbalanced nature of the data**. This approach aimed to obtain realistic results on data reflecting real-world conditions, enabling us to make more informed choices to optimize performance on the test data.

To **balance the training data**, we opted for **oversampling** using the **SMOTE method** [5]. In this context, under sampling would not have been sensible due to the limited number of instances with the ***Biopsy*** class set to *Yes*. Training on such a small number of observations would not have ensured a proper representation of the characteristics of the class of interest.

It's worth noting that our dataset includes not only continuous variables but also integer variables. Since **SMOTE** inherently **operates with continuous numerical values**, we addressed this challenge by **rounding** the **results** obtained for integer variables.

2.f – Data Scaling

As for data scaling, we have decided to assess this aspect only during the algorithm selection phase [3.a]. This decision is influenced by the fact that **not all algorithms require data normalization or standardization** [6],[7],[8]. Therefore, data scaling must be carefully evaluated based on the specific needs of each model adopted in the analysis. It is important to note that there are no rigid rules regarding these approaches, and various solutions can be tested to find the one that best yields optimal results.

3 – Model Training

The techniques described subsequently were **applied to both balanced and imbalanced training data**.

3.a – Classification Models

We identified six distinct classification models to train on our data, each requiring a specific data scaling approach applied to the training set [3.b]. The identified models include **J48, Random Forest** (RF)**, Logistic Regression** (Logistic)**, Multilayer Perceptron** (MLP)**, Naive Bayes Classifier** (Naïve Bayes)**, Support Vector Machine** (SVM).

3.b – Feature Selection and Data Scaling

After identifying the classification models to be trained to address the classification problem, we chose to perform **variable selection** to reduce the dataset's dimensionality, with the aim of avoiding potential overfitting during the model testing phase. The variable selection technique adopted was a **Multivariate Filter Method** (evaluator: *CfsSubsetEval*, search: *BestFirst*), implemented simultaneously with a **10-fold cross-validation**. This approach yielded two distinct sets of results for the two training sets:

- **Variables selected for balanced data**: Firstsexualintercourse, Numofpregnancies, HormonalContraceptives(years), IUDyears, STDs(number), Dx, Schiller.
- **Variables selected for unbalanced data**: HormonalContraceptives(years), STDs:condylomatosis, STDs:syphilis, STDs:genital herpes, STDs:HIV, Dx:Cancer, Schiller.

Simultaneously with feature selection, we determinate the **appropriate data scaling method**. We not only considered the sources cited in paragraph 2.f but, also, tested different data scaling approaches on the training set. Ultimately, we decided to retain the original data for all models, except **for the Support Vector Machine** (SVM), for which we opted to **standardize the training set**.

3.c – Models Robustness Evaluation

The implementation of a 10-fold cross-validation also allowed us to **assess the robustness of the selected classification algorithms** by evaluating the variability of accuracies obtained across different models in each iteration of the cross-validation through boxplot analysis.

Balanced Data

As clearly highlighted in the boxplots depicted in **Figure 2,** each model exhibits a **high accuracy value**, exceeding **90%**. These results imply an effective initial performance of the estimated classification models. However, for a more in-depth assessment of algorithm robustness, it is beneficial to consider the range of accuracy variation for each model: **broad ranges may indicate fewer algorithm robustness**.

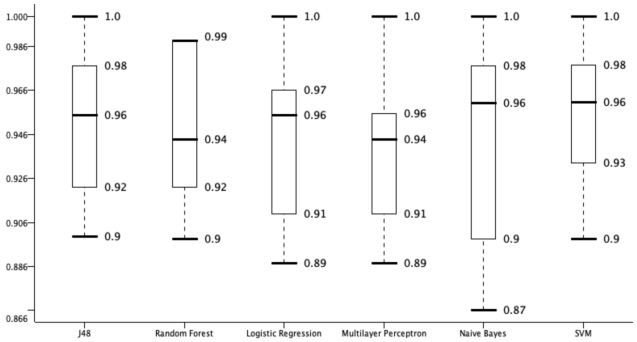

Figure 2: Accuracy boxplots of each model obtained with the 10-fold cross validation on Balanced Data.

Imbalanced Data

As depicted in **Figure 3**, **accuracy values surpassing 88%** are observed in the context of imbalanced data, representing a substantial outcome. As in the case of balanced data (**Figure 2**), for models trained with unbalanced data, we observe fairly wide ranges of variation for accuracy. This suggests that different training and test data can significantly alter the model's performance. This leads us to not have strong preferences between balanced and unbalanced data after an initial analysis of robustness.

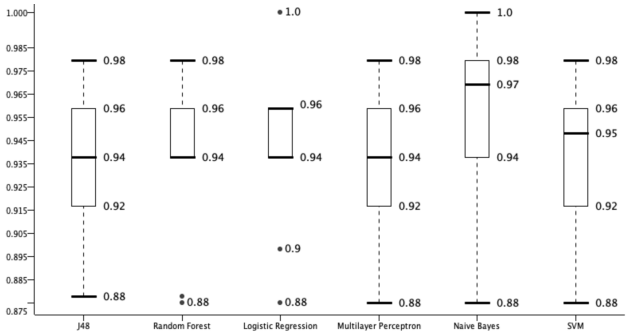

Figure 3: Accuracy boxplots of each model obtained with the 10-fold cross validation on Imbalanced Data.

3.d – Confidence Interval Evaluation

After conducting an initial analysis of algorithm robustness in the previous paragraph [3.c], we delved into a detailed examination of our models. We utilized the **same variables selected** through the attribute selector in the preceding step, proceeding with the training of the models on the entire training set, applying the **Holdout Method**.

Subsequently, we performed a check for potential overfitting by **comparing the accuracy's confidence intervals of the models** (obtained on training set) with the **accuracy obtained in the validation test**.

Balanced Data

As highlighted in **Figure 4**, the **confidence intervals (CI) for accuracy** obtained from the training set have a **range of variation between 5% and 7%.** The accuracies in the validation set all fall within the created confidence intervals, except for Naive Bayes, whose estimated interval ranges from the lower extreme 0.94 to the upper extreme at 0.99, an accuracy of 0.97 on training set but achieved an accuracy in the validation set of 0.93.

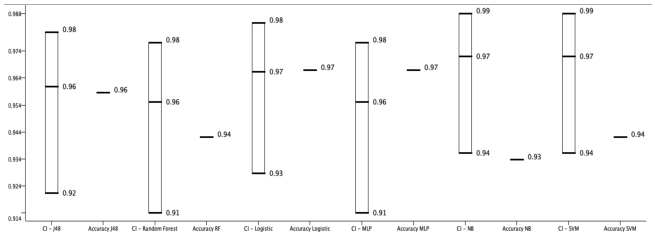

Figure 4: Confidence Interval of Accuracy computed for each model on balanced training set, against validation set accuracy.

It is noteworthy that the **accuracies in the validation set are all very close to the accuracy of the training set**, except for Naïve Bayes Classifier and SVM**.**

Such proximity between the two values indicates that the accuracy estimates obtained during training could be representative of the actual model performance when applied to new data in the validation set. In other words, the **consistency between the training and validation sets** may indicate **good model generalization**, where the predicted performances closely align with those observed in real-world scenarios.

Imbalanced Data

In **Figure 5**, it is evident that, in the case of imbalanced data, the **confidence intervals** all have a **width of approximately 10%**, which is higher than that shown in **Figure 4**.

All accuracies obtained during validation tests fall within their respective confidence intervals, and, similar to the case of balanced data, the accuracy values obtained on validation test are very similar to those of the training test.

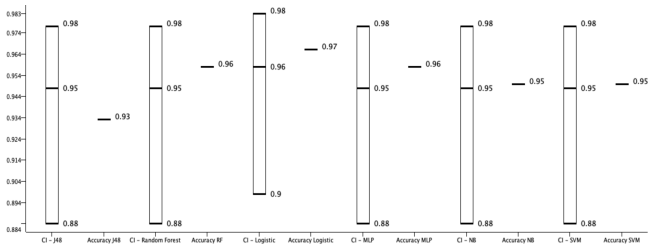

Figure 5: Confidence Interval of Accuracy computed for each model on imbalanced training set, against validation set accuracy.

3.e – Sensitivity and Specificity Evaluation

In a second check, still based on the results from the validation phase, we examined **specificity and sensitivity**. This evaluation was crucial to ensure that the models could achieve **high Recall/Sensitivity**, thus ensuring the accurate prediction of the ***Biopsy*** variable with the *Yes* mode. This approach was adopted because relying solely on accuracy would risk depending on misleading results. This risk is particularly pronounced in the case of unbalanced data, as models tend to perform better or create bias toward the majority class.

Balanced Data

By analyzing the graph in **Figure 6**, it is observed that models such as J48, Logistic, and MLP exhibit a very high sensitivity, suggesting that they are excellent models to consider in the final testing phase.

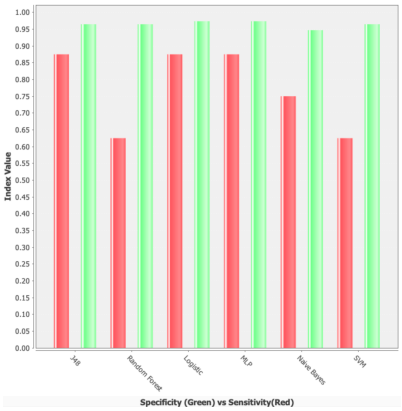

Figure 6: Specificity (Green) against Sensitivity (Red) computed for each model trained on balanced data.

Imbalanced Data

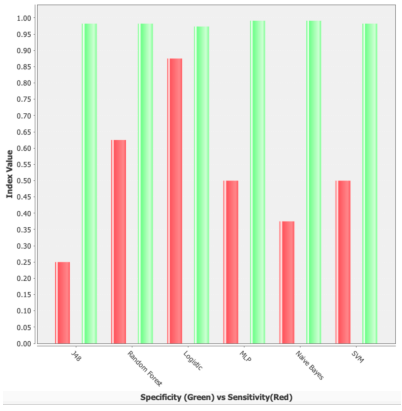

Figure 7: Specificity (Green) against Sensitivity (Red) computed for each model trained on imbalanced data.

Contrastingly, examining the graph in **Figure 7**, it is noticeable that, except for logistic regression, which maintains results similar to the previous scenario, all other models exhibit **significantly reduced sensitivity**, averaging **around 40%**. This phenomenon can be attributed to various reasons, particularly when training on imbalanced datasets. The low sensitivity may stem from an excessive focus on the majority class, inadequate representation of the target class, and the intrinsic complexity of the latter. In such contexts, models tend to adapt to predicting the majority class, neglecting the correct identification of positive cases.

3.f – ROC curve Evaluation

Before delving into the comparison of **ROC curves**, it's essential to highlight the significance of this analysis in assessing the performance of classification models. The **Receiver Operating Characteristic** (ROC) **curve** provides a visual representation of a model's ability to discriminate between classes at various classification thresholds. This aids in understanding the trade-off between sensitivity and specificity.

The **area under the ROC curve** (**AUC**) serves as a quantitative measure of the model's overall performance.

Balanced Data

Figure 8 illustrates the ROC curves for models applied to balanced data, showcasing their performance in predicting ***'Biopsy=Yes'***. The corresponding Area Under the Curve (AUC) values for each model are displayed in the bottom-right section of the graph.

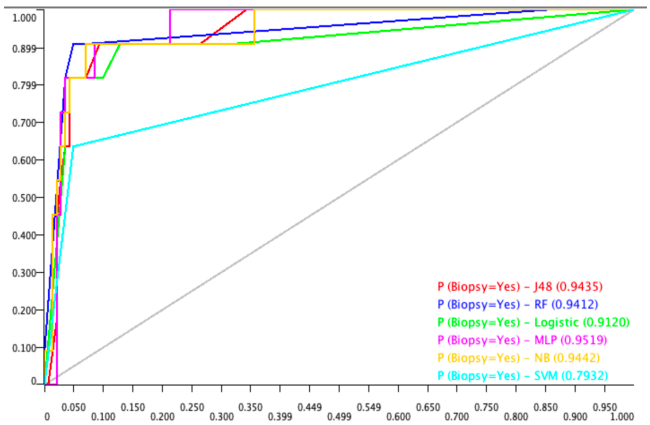

Figure 8: ROC curves for models applied to balanced data, showcasing their performance in predicting *'Biopsy=Yes'*.

Notably, Multi-Layer Perceptron (MLP) stands out with the highest AUC of 0.9519, indicating superior performance in balanced conditions. J48, Random Forest (RF), Logistic Regression, and Naive Bayes (NB) also exhibit robust performances, with AUC values exceeding 0.90. However, Support Vector Machine (SVM) lags behind

with a comparatively lower AUC of 0.7932. This suggests that SVM may face challenges in achieving high predictive accuracy in balanced datasets.

Imbalanced Data

The Multilayer Perceptron experiences a significant variation in its AUC when transitioning from balanced to imbalanced data (**Figure 9**). In contrast, despite the class imbalance, Random Forest (RF) maintains a remarkable AUC of 0.9364, emphasizing its resilience to varying class distributions. The Support Vector Machine (SVM), on the other hand, continues to demonstrate its limited ability to classify data optimally, as indicated by its AUC of 0.7166.

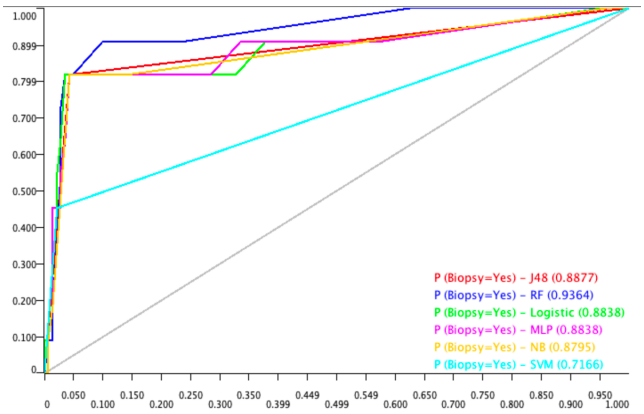

Figure 9: ROC curves for models applied to imbalanced data, showcasing their performance in predicting *'Biopsy=Yes'*.

Models trained on Imbalanced Data

Model	Accuracy	Recall
J48	0.947	0.818
Random Forest	0.940	0.545
Logistic	0.954	0.818
MLP	0.940	0.545
Naive Bayes	0.947	0.818
SVM	0.940	0.455

Table 2: Accuracy and Recall results on imbalanced data.

Observing the results, it is noted that, in the case of models trained **on balanced data**, the **average Recall is significantly higher**, while the **Accuracy is satisfactory for both balanced and imbalanced data**. It is also highlighted that, in the case of unbalanced data, Accuracy is misleading as it does not indicate the model's classification capability but rather its tendency to classify most observations into the majority class, as observed in SVM, MLP, and Random Forest.

Based on these considerations, the model selection was oriented towards **those trained on balanced data**, and among these, the choice fell on the **J48 model**, given its **high recall and overall performance**.

4 – Test Results

In the analysis of the test results, focused on the classification of the "Biopsy" class as *Yes* or *No* for the assessment of cervical cancer risk, the Recall metric emerges as crucial, considering the priority of correctly identifying positive cases. Below are the results of the models.

Models trained on Balanced Data

Model	Accuracy	Recall
J48	0.940	0.909
Random Forest	0.934	0.818
Logistic	0.934	0.818
MLP	0.927	0.818
Naive Bayes	0.927	0.818
SVM	0.927	0.636

Table 1: Accuracy and Recall results on balanced data.

5 – Conclusions

In our examination of cervical cancer risk classification based on biopsy outcomes, the pivotal metric of Recall takes precedence, emphasizing the imperative need for accurately identifying positive cases.

The primary challenge lay in precisely classifying cervical cancer risks, with biopsy outcomes serving as the target variable. Our objective was to develop a predictive model capable of identifying high-risk cases, facilitating targeted preventive interventions, and optimizing screening strategies. The dataset, featuring 36 variables collected from a cohort of female patients, underwent meticulous pre-processing to ensure data consistency and quality.

Categorical variables were processed, and missing data were handled with a strategic imputation strategy. The class imbalance problem was addressed through oversampling using the SMOTE method. The dataset was then partitioned, and a balanced training set was obtained,

considering the proportional distribution of classes within the target variable.

Variable selection, robustness evaluation, and confidence interval analysis were performed on both balanced and imbalanced data.

Moving forward, several avenues for improvement and exploration present themselves. Future endeavors could delve into ensemble methods [9], fine-tuning model [10] and collaborating with healthcare professionals for domain-specific insights.

After conducting an in-depth critical analysis, it became evident that caution was necessary when interpreting results. While accuracy remained satisfactory across both scenarios (balanced and imbalanced data), basing the model selection solely on this metric would not have been the correct choice in our case study, as the error of failing to predict a *Yes* was more severe than misclassifying a *No*. Therefore, it was decided to select the model considering its performance in predicting the minority class, thus taking into account the recall. The final test results emphasized the superiority of models trained on balanced data. In fact, these models exhibited significantly higher average recall, highlighting their effectiveness. The model selection favored those trained on balanced data, with J48 standing out for its robustness and high Recall.

To enhance the study's impact, ongoing refinement and dataset expansion are crucial for advancing cervical cancer risk assessment and improving patient outcomes. The study lays a solid foundation, and future research should focus on addressing the identified limitations for a more comprehensive understanding and effective application of predictive models in real-world healthcare scenarios.

6 – References

1. Data source: https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification
2. Introduction to Balanced and Imbalanced datasets: https://encord.com/blog/an-introduction-to-balanced-and-imbalanced-datasets-in-machine-learning/
3. Unbalanced data is a problem? No, balanced data is worse: https://matloff.wordpress.com/2015/09/29/unbalanced-data-is-a-problem-no-balanced-data-is-worse/
4. Imbalance class problem: https://stats.stackexchange.com/questions/227088/when-should-i-balance-classes-in-a-training-data-set
5. Smote technique: https://www.blog.trainindata.com/overcoming-class-imbalance-with-smote/
6. Does SVM need feature scaling? https://forecastegy.com/posts/does-svm-need-feature-scaling-or-normalization/
7. Does Random Forest Need Feature Scaling or Normalization? https://forecastegy.com/posts/does-random-forest-need-feature-scaling-or-normalization/
8. Which models require normalized data? https://www.yourdatateacher.com/2022/06/13/which-models-require-normalized-data/
9. Ensemble Methods: https://corporatefinanceinstitute.com/resources/data-science/ensemble-methods/
10. Fine tuning: https://encord.com/blog/training-vs-fine-tuning/#h2