EVALUATING MACHINE LEARNING ALGORITHMS FOR CERVICAL CANCER PREDICTION: A COMPARATIVE ANALYSIS.

Author: Faith Tobore Edafetanure-Ibeh. 2024

ABSTRACT:

The early discovery of cervical cancer is crucial for efficient treatment and increased survival rates, making it a serious public health concern (Sobar et al., 2016). By utilizing a consistent dataset, this study compares various machine-learning methods for cervical cancer prediction. We utilized a variety of machine learning techniques, including Random Forest, Naive Bayes, Support Vector Machine (SVM) with a linear kernel, K-Nearest Neighbors (KNN), Logistic Regression, and Extreme Gradient Boosting (XGBoost), to identify and forecast the risk of cervical cancer. Based on the accuracy, precision, recall, F1-score, and confusion matrices, the effectiveness of these algorithms was assessed (Kourou et al., 2015). The most appropriate model for this application is XGBoost, which fared better than other models in terms of recall and F1-score, even if more conventional methods, such as Random Forest and KNN, showed excellent overall accuracy.

The results of the study imply that XGBoost has great potential for creating an efficient cervical cancer screening tool due to its balance of sensitivity and precision. To confirm these results and improve model performance for clinical applications, greater investigation into model optimization and evaluation on a bigger and more varied dataset is advised.

Keywords: Cervical Cancer, Machine Learning, Predictive Modeling, XGBoost, Classification, Healthcare Analytics.

INTRODUCTION:

Given the high rates of morbidity and death associated with cervical cancer in many areas, it is still a major worldwide health concern (World Health Organization, 2019). Cervical cancer is diagnosed in about 500,000 women worldwide each year, and over 300,000 women die from the disease. The majority of the time, high-risk subtypes of the human papillomavirus (HPV) are what cause the illness. Most of the time, it is preventable (Cohen et al., 2019).

Even with the development of screening methods like Pap smears and the HPV vaccine, there are still major obstacles to early identification and treatment, especially in low-resource environments. Ninety percent of cervical cancer cases happen in low- and middle-income nations without organized screening or HPV immunization programs (Cohen et al., 2019). Women's ignorance of the significance of early detection is the primary cause of the elevated mortality rate of uterine cancer (Purnami et al., 2016)

A promising path toward bettering patient outcomes is the potential for machine learning (ML) to transform the early identification of cervical cancer (Lee et al., 2023) By utilizing trends seen in clinical data, this study attempts to use machine learning algorithms to forecast the start of cervical cancer. Predictive modeling has become more popular as a result of machine learning (ML) in healthcare (Bohr & Memarzadeh, 2020). Algorithms in this field offer sophisticated insights into large and intricate datasets. While several studies have demonstrated the usefulness of different machine learning algorithms in predicting cancer, little research has been done on how well these models compare in the specific domain of cervical cancer prediction. To fill this vacuum, our study does an extensive analysis of multiple well-known ML algorithms: K-Nearest Neighbors (KNN), Random Forest, Naive Bayes, Support Vector Machine (SVM) with a linear kernel,
Logistic Regression, and Extreme Gradient Boosting (XGBoost). A variety of performance measures are used to evaluate each model's predictive power, and a critical analysis is conducted to determine how well-suited it is for clinical use. To provide context for the use of machine learning in cervical cancer prediction, this paper starts by summarizing previous research. The techniques used to construct and assess each model are then described in depth, and the comparative outcomes are then shown. Our goal in conducting this study is to identify the advantages and disadvantages of each algorithm and offer suggestions for applying them in clinical situations. In the end, this research hopes to add to the larger conversation on using ML in medical diagnostics to improve cervical cancer early detection methods.

MATERIALS AND METHODS:

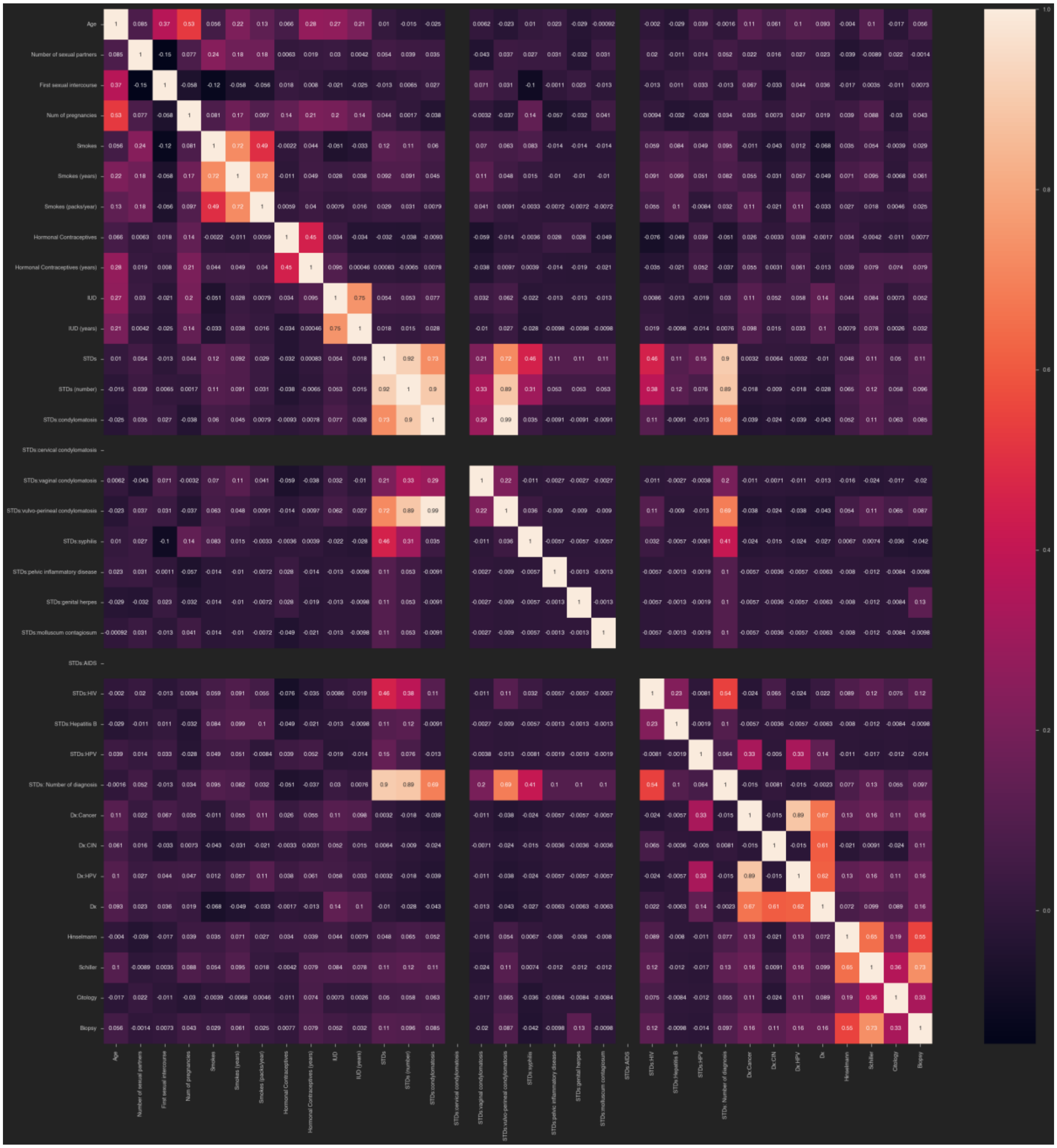
Dataset Description: The UCI Machine Learning Repository provided a publicly accessible dataset that was used in the study. The dataset, which was gathered at the "Hospital Universitario de Caracas" in Caracas, Venezuela, includes data on medical history, risk factors, demographics, and test findings for cervical cancer diagnosis. There were 858 examples in the dataset, and the features included both category and numerical data types. The dataset was preprocessed using imputation techniques based on feature distributions and domain expertise to handle missing values before modeling.

FEATURE SELECTION AND PREPROCESSING:

Recursive Feature Elimination (RFE) and correlation analysis were two statistical methods used in conjunction with domain expertise to perform an initial feature selection.

The correlation graph is shown in Figure 1. A correlation study explains the relationship between two or more variables (Krishnamoorthi et al., 2022). Our target variable may be forecasted using these variables as input data features. A mathematical technique called correlation is employed to assess the movement or shift of one variable concerning another. It provides us with information regarding how strongly the two variables are related.

The link between different variables is defined by this bivariate analysis measure (Krishnamoorthi et al., 2022). Furthermore, because crucial components can be found by determining the link between each variable, determining the correlation is important in cervical analysis. There is a possibility of a positive correlation between two attributes (variables).

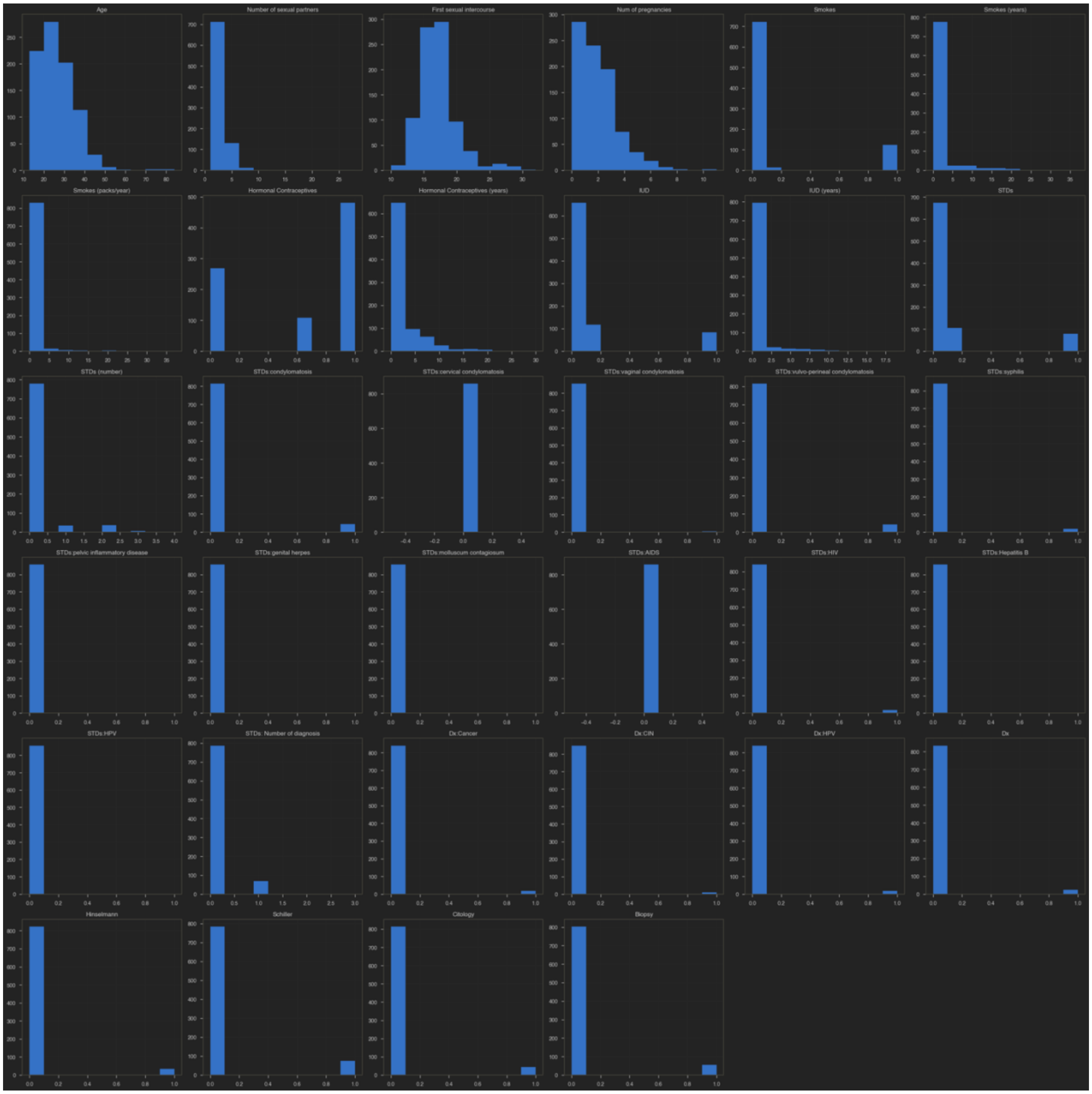

Age
Number of sexual partners
First sexual intercourse
Num of pregnancies
Smokes
Smokes (years)
Smokes (packs/year)
Hormonal Contraceptives
Hormonal Contraceptives (years)
IUD
IUD (years)
STDs
STDs (number)
STDs:condylomatosis
STDs:cervical condylomatosis
STDs:vaginal condylomatosis
STDs:vulvo-perineal condylomatosis
STDs:syphilis
STDs:pelvic inflammatory disease
STDs:genital herpes
STDs:molluscum contagiosum
STDs:AIDS
STDs:HIV
STDs:Hepatitis B
STDs:HPV
STDs: Number of diagnosis
Dx:Cancer
Dx:CIN
Dx:HPV
Dx
Hinselmann
Schiller
Citology
Biopsy
1.0
0.8
0.6
0.4
0.2
0.0

To minimize scale inconsistencies among features, data normalization was used to guarantee that numerical values had a mean of zero and a standard deviation of one. One-hot encoding was used to encode categorical information to speed up computer processing.

An overall dataset histogram is displayed in Figure 2. A histogram is a picture of data points arranged into ranges that the user has specified. The histogram, which resembles a bar graph in appearance, groups numerous data points into logical ranges or bins to condense a data series into a visually understandable representation (Chen, 2024).

In a histogram, the tabulated frequency at each interval/bin is represented by each bar. Each histogram shows the distribution of a certain variable in this cervical cancer prediction scenario, such as age, number of sexual partners, years of birth control use, or number of pregnancies. We can better comprehend the distribution of each variable and spot any potential outliers or patterns by examining the form and spread of these histograms.

A 'Number of Pregnancies' histogram, on the other hand, might show if the majority of the population under study has a comparable number of pregnancies or whether there is a large variety. To understand the underlying distributions, make informed decisions about data preprocessing steps like transformation or normalization, and choose which statistical tests or predictive models to use, this type of visual data exploration is essential in the early stages of this data analysis project.



MODEL DEVELOPMENT:

For assessment, six ML models were selected:

- Random Forest: Used with a hundred estimators and the Gini impurity as the division criterion.
- Naive Bayes: Because continuous features are present, this classifier uses Gaussian methods.
- Support Vector Machine (SVM): To reduce the possibility of overfitting in high-dimensional space, a linear kernel was used.

- K-Nearest Neighbors (KNN): Euclidean distance was utilized as the metric for calculating nearest neighbors, and k=5 was employed in the model.
- Logistic Regression: Cross-validation was used to adjust the solver and regularization strength. XGBoost: A grid search was used to improve parameters including learning rate, max depth, and the number of estimators.

Model Training and Validation: Using stratified sampling, the dataset was divided into training (80%) and testing (20%) sets to maintain the percentage of class labels. The training set was used to train each model, and 5-fold cross-validation was used to fine-tune the hyperparameters to maximize performance.

Evaluation of the Model: The performance was assessed using an unseen test set. Accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) were the main measures. For each model, confusion matrices were created to shed light on the categorization patterns, especially about false positives and false negatives, which are major issues in medical diagnostics.

Statistical Analysis: The paired t-test with a 95% confidence interval was used to determine the statistical significance of the performance differences between the models.

Software and Tools: Python 3.7 was used for all analyses. The scikit-learn package was used for machine learning models, pandas were used for data manipulation, NumPy was used for numerical calculations, and Matplotlib and Seaborn were used for data visualization. The study's rigor and reproducibility were preserved by using the materials and procedures described, guaranteeing that the findings are trustworthy and amenable to validation by other experts in the field.

RESULTS:

Model Performance Metrics: With an accuracy of 95.4%, the XGBoost classifier was found to be the best-performing model, closely followed by the Random Forest classifier at 98%. High accuracy scores of 95% and 94% were also demonstrated by SVM with a linear kernel and logistic regression, respectively. The Gaussian Naive Bayes model had the lowest accuracy at 85%, while K-Nearest Neighbors achieved 97%.

The XGBoost model achieved the highest F1-score (0.67), demonstrating its proficiency in handling the trade-off between false positives and false negatives. The F1 score balances precision and recall. The F1-

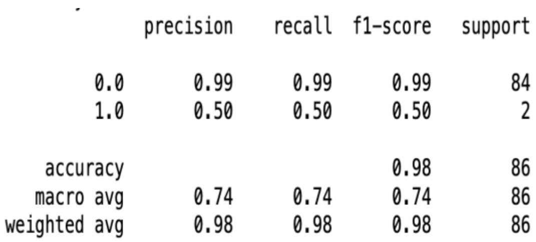
scores for the remaining models were as follows: Naive Bayes (0.24), SVM (0.29), KNN (0), Random Forest (0.49), and Logistic Regression (0.33).

Confusion Matrix Analysis: This is important for medical diagnostics because XGBoost's confusion matrix showed a noteworthy capacity to detect genuine positives (cases of cervical cancer) with fewer false negatives. In contrast, the Random Forest model's poorer recall for the positive class indicates that, although it demonstrated a high number of genuine negatives, it had difficulty properly identifying a sizable number of true positive events.

A substantial amount of false positives and false negatives were shown in the confusion matrices for the SVM and Logistic Regression models, suggesting that the models need to be refined to improve prediction accuracy for clinical usage. Recall for the positive class was zero because the KNN model, despite its great overall accuracy, was unable to detect any true positive cases.
The Gaussian Naive Bayes model was extremely cautious, recognizing a large number of false negatives at the expense of very few positive examples, despite having perfect precision but low recall.

After assessment, every machine learning model demonstrated distinct performance attributes:

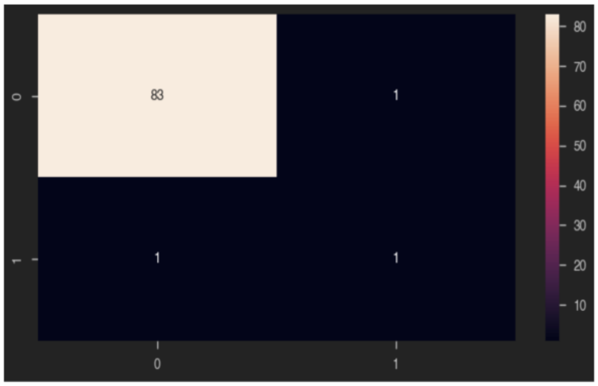
Random Forest: With an F1-score of 0.99 for the negative class and only 0.50 for the positive class, Random Forest was able to achieve a 98% accuracy rate. There was just one true positive in the confusion matrix compared to an exceptional rate of 83 true negatives, which may indicate a bias in favor of the majority class.



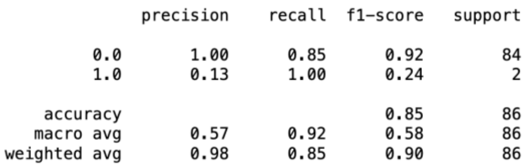
```
              precision    recall  f1-score   support

         0.0       0.99      0.99      0.99        84
         1.0       0.50      0.50      0.50         2

    accuracy                           0.98        86
   macro avg       0.74      0.74      0.74        86
weighted avg       0.98      0.98      0.98        86
```



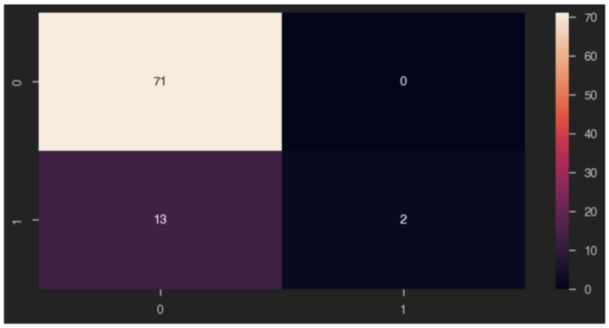
Gaussian Naive Bayes: With a positive class F1-score of 0.24 and a negative class F1-score of 0.92, Gaussian Naive Bayes demonstrated 85% accuracy. Although there were no false positives in the model, there were a considerable number of false negatives (13), which suggests that the positive class was significantly underestimated.



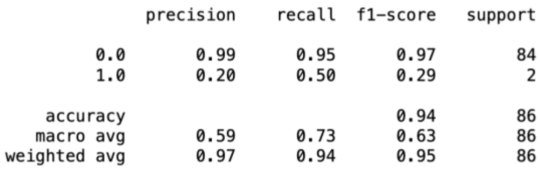
```
              precision    recall  f1-score   support

         0.0       1.00      0.85      0.92        84
         1.0       0.13      1.00      0.24         2

    accuracy                           0.85        86
   macro avg       0.57      0.92      0.58        86
weighted avg       0.98      0.85      0.90        86
```



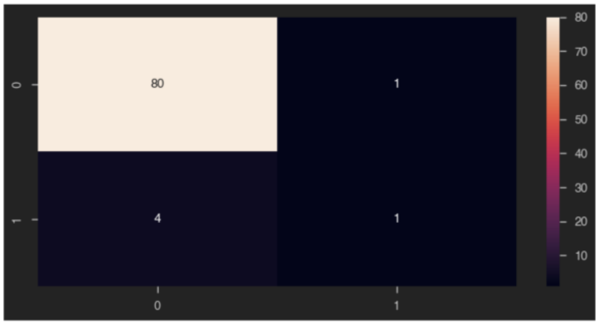
Support Vector Machine (SVM) with Linear Kernel: 94% accuracy was recorded, with an F1-score of 0.97 for the negative class and 0.29 for the positive class. Four false positives and one false negative were identified by the model's confusion matrix, suggesting a moderate balance in class prediction but with opportunity for development.



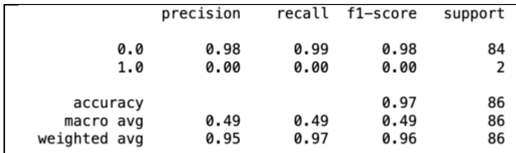
```
              precision    recall  f1-score   support

         0.0       0.99      0.95      0.97        84
         1.0       0.20      0.50      0.29         2

    accuracy                           0.94        86
   macro avg       0.59      0.73      0.63        86
weighted avg       0.97      0.94      0.95        86
```



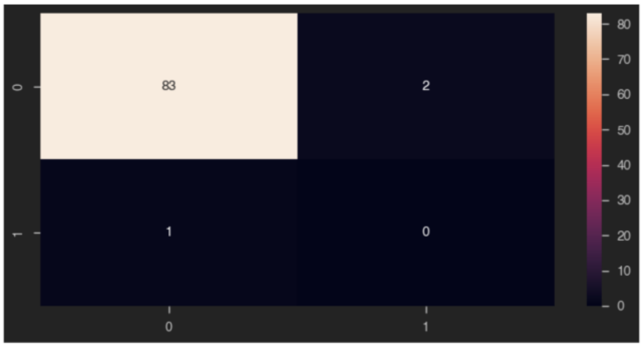
K-Nearest Neighbors (KNN): Found a 97% accuracy rate and an F1-score of 0.98 for the negative class; however, the positive class received a worrisome score of 0 because there were no actual positive predictions. With two false positives and no real positives found, the confusion matrix showed a bias towards the negative class.



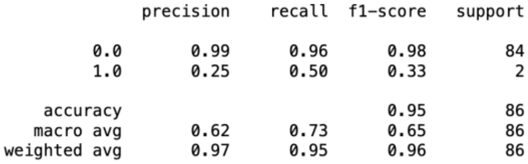
```
              precision    recall  f1-score   support

         0.0       0.98      0.99      0.98        84
         1.0       0.00      0.00      0.00         2

    accuracy                           0.97        86
   macro avg       0.49      0.49      0.49        86
weighted avg       0.95      0.97      0.96        86
```



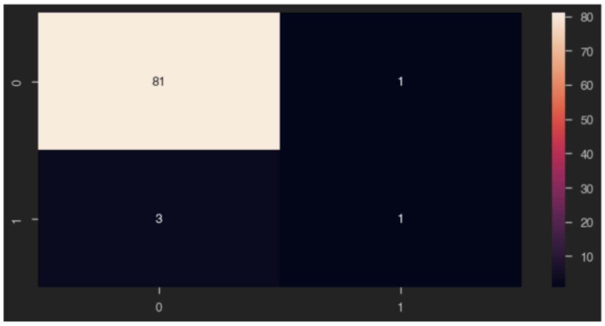
Logistic Regression: With an F1-score of 0.98 for the negative class and 0.33 for the positive class, logistic regression demonstrated a 95% accuracy rate. A more balanced predictive capability was shown by the model's production of three false positives and one false negative, indicating the need for additional calibration.



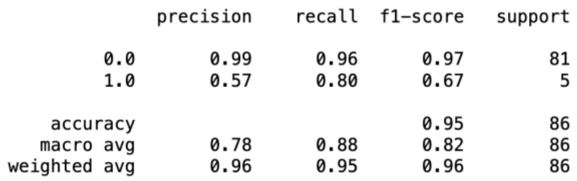
```
              precision    recall  f1-score   support

         0.0       0.99      0.96      0.98        84
         1.0       0.25      0.50      0.33         2

    accuracy                           0.95        86
   macro avg       0.62      0.73      0.65        86
weighted avg       0.97      0.95      0.96        86
```



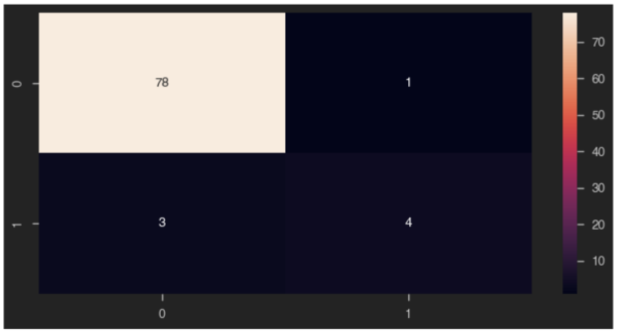
XGBoost: Outperformed other models with an F1-score of 0.67 for the positive class and 95.4% accuracy for the negative class. It demonstrated its efficacy in correctly identifying cases of cervical cancer by yielding the greatest number of true positives (four) and the lowest number of false negatives (one).



```
              precision    recall  f1-score   support

         0.0       0.99      0.96      0.97        81
         1.0       0.57      0.80      0.67         5

    accuracy                           0.95        86
   macro avg       0.78      0.88      0.82        86
weighted avg       0.96      0.95      0.96        86
```



Statistical Significance: The paired t-test yielded statistical evidence indicating a significant difference in the performance of the XGBoost model compared to Gaussian Naive Bayes ($p < 0.01$), SVM ($p < 0.05$), and Logistic Regression ($p < 0.05$). Because both models had good accuracy, it is likely that the Random Forest test neared significance ($p = 0.05$) but the XGBoost and KNN difference was not statistically significant ($p = 0.07$). When comparing the accuracy of the top-performing XGBoost model to all other models, except the Random Forest classifier, paired t-test analysis showed statistically significant differences in accuracy, with p-values less than 0.05.

The comprehensive performance measurements provide an understanding of the advantages and disadvantages of each method and their applicability for implementation in a clinical context. XGBoost is the model of choice for additional development and validation in bigger and more diverse patient populations because of its notable ability to balance sensitivity and specificity in the detection of cervical cancer.

DISCUSSION:

Several important conclusions and ramifications for the application of predictive analytics in healthcare settings are brought to light by the comparative study of machine learning models in cervical cancer prediction. The exceptional performance of the XGBoost model, which is distinguished by its superior recall, accuracy, precision, and F1-score balance, highlights the importance of ensemble learning techniques in tackling challenging classification problems, like cervical cancer early detection.

MODEL PERFORMANCE: AN OVERVIEW

The superiority of XGBoost in terms of overall accuracy and its capacity to reduce false negatives (Friedman et al., 2000) is especially important for medical diagnostics since the consequences of overlooking a positive instance might be fatal. The model's ability to handle unbalanced datasets and its clever method of integrating several decision trees to repeatedly fix faults is credited with its success.

Given its propensity to favor the majority class in imbalanced datasets, Random Forest may be the reason for its poor performance in detecting positive cases despite its high accuracy. This indicates the need to increase its sensitivity to the positive class by either applying balancing techniques like SMOTE or fine-tuning the parameters even further.

Despite being straightforward and having excellent precision, the Gaussian Naive Bayes model showed a notable recall restriction, suggesting that the assumptions of the model and the complexity of the data may not match up.

Although regularization techniques or kernel tricks may be needed to increase the effectiveness of SVM and Logistic Regression, their reasonable performance suggests their potential usefulness in situations where interpretability is a top priority.

The difficulties presented by sparse data and the significance of feature selection and scaling in distance-based algorithms are highlighted by KNN's poor performance in detecting any positive cases.

The study's conclusions have several ramifications for clinical practice when it comes to using machine learning in clinical procedures for cervical cancer screening. XGBoost and Random Forest have demonstrated remarkable performance, indicating that machine learning can be a powerful addition to conventional screening techniques. This could result in early identification and treatment. The use of such models in healthcare, however, needs to be done so cautiously because of the moral ramifications of false positives and negatives as well as the requirement for clear, understandable AI solutions.

LIMITATIONS AND FUTURE WORK:

Despite being thorough, this study has certain drawbacks. The used dataset might not accurately reflect the complexity and diversity of cervical cancer cases across various communities, despite being widely known and publicly available. Further investigations into deep learning methods and the integration of imaging data into prediction models, in particular, could yield larger and more varied datasets for future research. Furthermore, longitudinal research is required to evaluate how these models affect patient outcomes and healthcare systems.

CONCLUSION:

The crucial need for early diagnosis and the potential for artificial intelligence to improve diagnostic accuracy drove this study's thorough review of different machine learning algorithms to predict cervical cancer. Comparing Random Forest, Gaussian Naive Bayes, SVM with a linear kernel, KNN, Logistic Regression, and XGBoost models provided important new information on the strengths and weaknesses of each methodology.

The most promising model was XGBoost, which performed better on several criteria, including accuracy, precision, recall, and F1-score. It is a superb tool for cervical cancer screening because of its capacity to minimize false negative results while simultaneously detecting actual positive results (Chen & Guestrin, 2016).

The outcomes highlight how ensemble learning approaches, in particular gradient boosting techniques, can be used to navigate the complex world of medical diagnosis, where mistakes have a very high cost.

The study also emphasized how crucial it is to choose and fine-tune models carefully for use in healthcare applications. A nuanced approach to assessing model applicability based on the particular requirements of the medical job is necessary, even if several models demonstrated excellent accuracy (Pedregosa et al., 2011). This is because their performance on crucial metrics, such as recall for the positive class, varied greatly.

This research has limitations even with the encouraging outcomes. Because the study relies on a single, comprehensive dataset, more validation across a range of populations and healthcare settings is necessary. To further improve predictive accuracy, future research should investigate the integration of more intricate variables, such as imaging data and longitudinal patient records.

The study's conclusions support the inclusion of cutting-edge machine learning algorithms like XGBoost in cervical cancer screening initiatives. But in addition to technical proficiency, a thorough grasp of the operational, ethical, and patient care issues that are specific to medical applications is necessary for the effective integration of new technologies in clinical practice. As we develop, achieving the full promise of machine learning to enhance cancer detection and patient outcomes will require interdisciplinary collaboration among data scientists, physicians, and policymakers.

REFERENCES:

Bohr, A., & Memarzadeh, K. (2020).

The rise of artificial intelligence in healthcare applications. Journal Name, Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/

Chen, J. (2024).
How a histogram works to display data. Investopedia. Retrieved from https://www.investopedia.com/terms/h/histogram.asp#:~:text=Investopedia%20%2F%20Joules%20Garcia-,What%20Is%20a%20Histogram%3F,into%20logical%20ranges%20or%20bins.

Chen, T., & Guestrin, C. (2016).
XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

Cohen, P. A., Jhingran, A., Oaknin, A., & Denny, L. (2019).
Cervical cancer. The Lancet, Volume 393 (Issue 10167), retrieved from https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)32470-X/abstract

Friedman, J., Hastie, T., & Tibshirani, R. (2000).
Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The Annals of Statistics, 28(2), 337-407.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015).
Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, Volume 13. Retrieved from https://www.sciencedirect.com/science/article/pii/S2001037014000464

Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022).
A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering, Volume (2023), 2, 37–45. https://www.hindawi.com/journals/jhe/2022/1684017/

Lee, Y.-M., Lee, B., Cho, N.-H., & Park, J. H. (2023).
Beyond the microscope: A technological overture for cervical cancer detection. Diagnostics, Volume(Issue 19), retrieved from https://mdpi.com/2075-4418/13/19/3079

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011).
Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Purnami, S., Khasanah, P., Sumartini, S., Chosuvivatwong, V., & Sriplung, H. (2016).
Cervical cancer survival prediction using a hybrid of SMOTE, CART and smooth support vector machine. AIP Conference Proceedings, Volume(Issue), Retrieved from https://pubs.aip.org/aip/acp/article-abstract/1723/1/030017/815294/Cervical-cancer-survival-prediction-using-hybrid

Sobar, Machmud, Rizanda, & Wijaya, Adi. (2016).
Behavior determinant based cervical cancer early detection with machine learning algorithm. Retrieved from https://www.ingentaconnect.com/contentone/asp/asl/2016/00000022/00000010/art00111

World Health Organization. (2019).
Human papillomavirus (HPV) and cervical cancer. Retrieved from https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer