POLYTECHNIC UNIVERSITY OF MADRID
E.T.S. DE INGENIERÍA AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS
E.T.S. DE INGENIEROS INFORMÁTICOS

APPLIED PROJECT MACHINE LEARNING

ACTIVITY SUPERVISED LEARNING

Computational Biology

IMPLEMENTED BY

Alberto González Calatayud

ACADEMIC COURSE: 2023-2024

Abstract

This study embarks on a computational journey to discern the most efficacious machine learning model for predicting treatment responses in colorectal cancer (CRC) based on genetic markers. In light of the significance of Single Nucleotide Polymorphisms (SNPs) in determining individual responses to CRC treatments, a dataset representing various SNPs was meticulously curated and prepared for analysis.

The research evaluated a suite of classifiers, namely Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), Random Forest, and Multilayer Perceptron, each subjected to a comprehensive cross-validation framework to fine-tune hyperparameters and enhance model performance.

The models were rigorously assessed on their accuracy and consistency across multiple iterations, aiming to identify a classifier that not only predicts with high precision but also exhibits stability across different subsets of data. The findings are presented in a comparative manner, highlighting the predictive prowess of each model and offering insights into their performance nuances. The culmination of this investigation provides a data-driven recommendation for a machine learning approach tailored to support clinical decisions in CRC treatment, marking a stride toward personalized medicine.

Keywords Colorectal Cancer, Machine Learning, SNP Analysis, Predictive Modeling, Cross-Validation, Personalized Medicine

1 INTRODUCTION

Colorectal cancer (CRC) stands as one of the leading causes of mortality worldwide, where early diagnosis and tailored treatment strategies can significantly improve patient survival rates. The complexity of CRC, influenced by individual genetic variations, presents a substantial challenge in predicting treatment responses.

This challenge underscores the necessity for robust analytical models capable of interpreting high-dimensional genetic data to predict therapeutic outcomes effectively.

The present work engages with this challenge by evaluating various machine learning classifiers to predict the response to treatment in individuals diagnosed with colorectal cancer.

Recognizing the critical role of Single Nucleotide Polymorphisms (SNPs) in shaping individual responses to medication, this study leverages SNP data transformed into a machine learning-compatible format to develop predictive models.

The **objective** is to assess and compare the performance of several classifiers, including Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), Random Forest, and Multilayer Perceptron, to ascertain the model that best generalizes to unseen data.

In pursuit of this goal, each classifier is subjected to a rigorous cross-validation process, tuning their hyperparameters to optimize predictive accuracy. This paper documents the methodological approach, from preprocessing the genetic data to training and validating the classifiers, followed by a critical analysis of the results. In doing so, the study aims to identify a model that not only exhibits high accuracy but also maintains consistency in performance, offering valuable insights for clinical decision-making in colorectal cancer treatment (Deulofeu et al., 2021).

2 METHODS

The computational analysis was performed using Python programming language with the support of several libraries known for their efficiency in data processing and machine learning tasks. The machine learning models were implemented using the `scikit-learn` library, which offers a wide range of tools for predictive data analysis. The specific classifiers from `scikit-learn` employed in this study include:

- `LogisticRegression` for logistic regression analysis.
- `DecisionTreeClassifier` for decision tree modeling.
- `KNeighborsClassifier` for the k-nearest neighbors algorithm.
- `RandomForestClassifier` for ensemble methods using random forests.
- `MLPClassifier` for neural network models with multiple layers.

Each model was fine-tuned and validated using `GridSearchCV` for hyperparameter optimization and `train_test_split` for data partitioning. Model performance metrics such as accuracy and F1 score were computed to assess the efficacy of each classifier, utilizing functions like `accuracy_score` and `f1_score`. The `confusion_matrix` function was used to evaluate the predictive capabilities of the best-performing model.

2.1 Library Versions

The versions of the primary libraries used in this study are listed below:

- numpy: `1.23.5`
- pandas: `1.5.3`
- matplotlib: `3.7.1`
- seaborn: `0.12.2`
- scikit-learn: `1.2.2`

3 MODEL EVALUATION

3.1 Dataset Composition and Preliminary Analysis

The study's dataset contains genetic data from 53 individuals related to colorectal cancer treatment responses. It includes 21 features representing different Single Nucleotide Polymorphisms (SNPs) and a target value indicating each individual's treatment response. For machine learning compatibility, SNP values were numerically encoded: 0 for 'MM', 1 for 'WW', 2 for 'WM', and 3 for 'MW'. This pre-processing step was crucial for applying various algorithms effectively. This initial analysis helped understand the data's structure, guiding informed decisions in model training and selection.

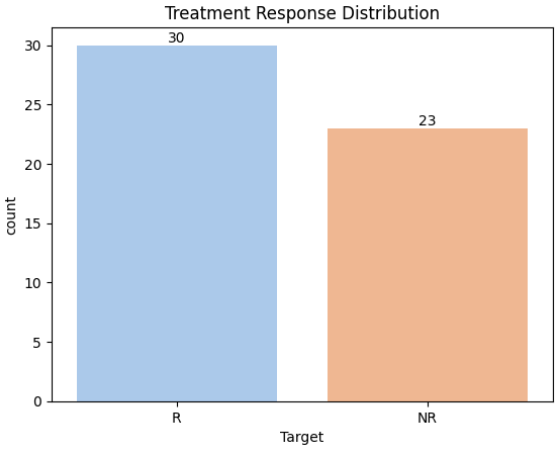

Figure 1: *Treatment Response Distribution*

An exploratory data analysis was conducted to gain insights into the structure and distribution of the dataset. Figure 1 presents the distribution of treatment responses among individuals, indicating the number of respondents with 'Response' (R) and 'No Response' (NR) to treatment.

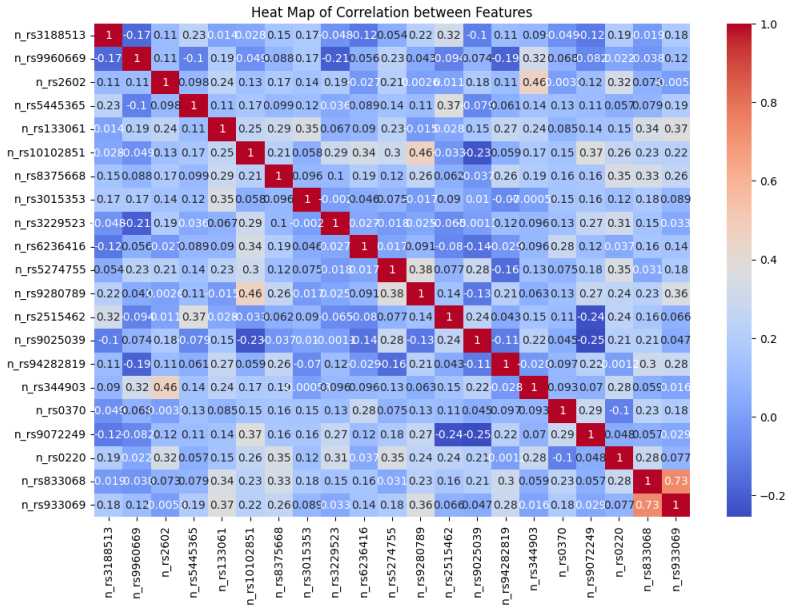

Figure 2: *Heat Map of Correlation between Features*

Additionally, to understand the relationships between different SNPs, a correlation heatmap was generated as shown in Figure 2. This heatmap visualizes the pairwise correlations between features, highlighting potential associations that may influence treatment response.

3.2 Model Evaluation and Selection

A systematic approach was employed to assess the performance of various classifiers and determine the optimal test size for each. The classifiers included Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), Random Forest, and Multilayer Perceptron, which were chosen for their diverse mechanisms of learning and suitability for the dataset at hand.

For each classifier, a range of test sizes [0.2, 0.3, 0.4] was evaluated to discern the proportion of the dataset that would yield the most reliable cross-validation results. The validation scores and their standard deviations were calculated across multiple subsets to ensure the robustness of the model against different data partitions. This iterative process aimed to mitigate overfitting and identify the test size that maximized the model's generalizability.

- Logistic Regression: Utilized a test size of 0.3, as determined to be optimal.
- Decision Tree: Employed a test size of 0.3, in line with the earlier findings.
- KNN: The best test size of 0.4 was used for this model.
- Random Forest: Conducted tuning using the test size of 0.2.
- MLP: Applied a test size of 0.4, as identified to be the most effective.

3.3 Hyperparameter Tuning and Model Validation

Following the determination of the optimal test size for each classifier, the next phase involved an intensive hyperparameter tuning process. This was executed through 20 iterations for each model, employing the test size previously identified as the most suitable for each respective classifier.
In each iteration, a Grid Search with cross-validation was used to find the best hyperparameters based on the F1 score. This search provided insights into model performance variability, measured by the standard deviation of F1 scores. The process highlighted each model's sensitivity to hyperparameters, enhancing their predictive accuracy and understanding of performance. Finally, models were evaluated on mean F1 score and its standard deviation, highlighting those with both high performance and consistency.

The performance of each model, post-tuning, was then visualized through bar charts. These charts depicted the mean F1 scores along with the standard deviations, providing a clear comparative view of the models' efficacies. The graphical representation aided in distilling complex numerical data into an interpretable format, facilitating a straightforward comparison of model performances.

This comprehensive approach to model evaluation and validation ensured that the final model selection was based on robust empirical evidence, laying a solid foundation for reliable predictions in the context of colorectal cancer treatment responses.

3.3.1 Logistic Regression

The optimal test size for the Logistic Regression model was determined to be 0.3. The best hyperparameters identified for this model were:

- Regularization Strength (C): 0.1
- Penalty: L2
- Solver: Newton-CG

Upon determining the optimal test size and hyperparameters for the Logistic Regression model, the model's performance was extensively evaluated. The key metrics considered were the validation score, train F1 score, test F1 score, train accuracy, and test accuracy. Each metric provides valuable insight into the model's ability to generalize and accurately predict outcomes on both seen and unseen data.

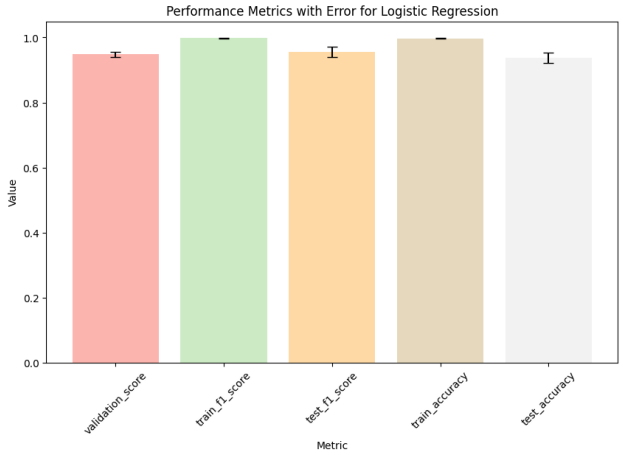

Figure 3: Bar chart illustrating the performance metrics with error for Logistic Regression. The metrics display the model's effectiveness across different evaluation criteria. Error bars represent the standard deviation, indicating the variability and consistency of the model's performance.

Figure 3 illustrates these performance metrics, with error bars signifying the standard deviation. Notably, the model exhibits a high validation score and F1 scores, which are critical measures of accuracy for binary classification tasks. The error bars are relatively small, indicating that the model's performance is consistent across different folds of the data. This consistency is a positive indicator of the model's reliability in practical applications.

3.3.2 Decision Tree

For the Decision Tree classifier, a test size of 0.3 was found to be most effective. The best hyperparameters for this model were:

- Maximum Depth: None (unlimited)
- Minimum Samples per Leaf: 1
- Minimum Samples Split: 5

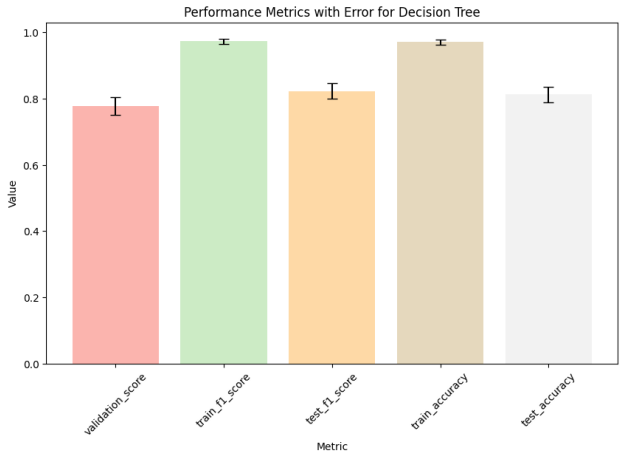

Figure 4: Performance metrics with error for the Decision Tree model.

As shown in Figure 4, the Decision Tree's performance metrics, particularly the validation and F1 scores, indicate a strong predictive capability.

3.3.3 KNN

Optimal test size for KNN was 0.4. The best hyperparameters were:

- Metric: Manhattan
- Number of Neighbors: 3
- Weights: Uniform

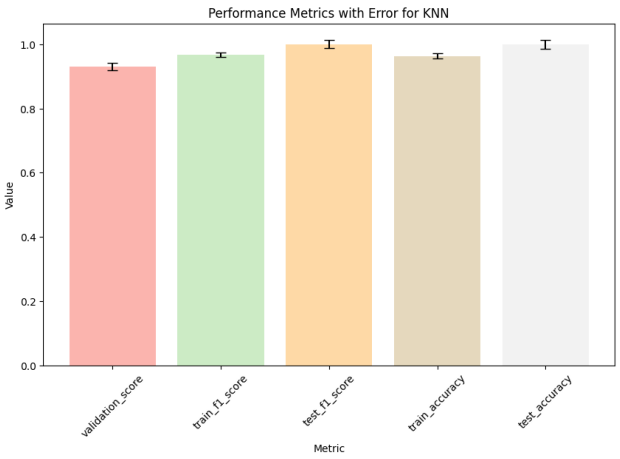

Figure 5: Performance metrics with error for the KNN model.

As depicted in Figure 5, the KNN model exhibits commendable validation and F1 scores, signifying its strong classification abilities.

3.3.4 Random Forest

The Random Forest model performed best with a test size of 0.2. The optimal hyperparameters identified were:

- Maximum Depth: None (unlimited)
- Minimum Samples per Leaf: 1
- Minimum Samples Split: 2
- Number of Estimators: 200

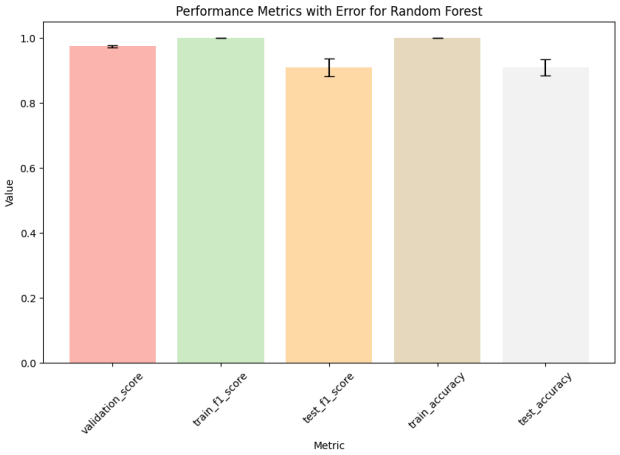

Figure 6: Bar chart illustrating the performance metrics with error for the Random Forest model.

Figure 6 presents these metrics with error bars, which reveal the Random Forest's solid validation and F1 scores. The relatively small error bars across all metrics underscore the model's stability and reliability, suggesting that the Random Forest's predictions are consistent and dependable.

3.3.5 MLP (Multilayer Perceptron)

For the MLP model, a test size of 0.4 was optimal. The best hyperparameters for this model included:

- Activation: ReLU
- Alpha: 0.0001
- Hidden Layer Sizes: (100, 100)
- Learning Rate: Adaptive
- Solver: Adam

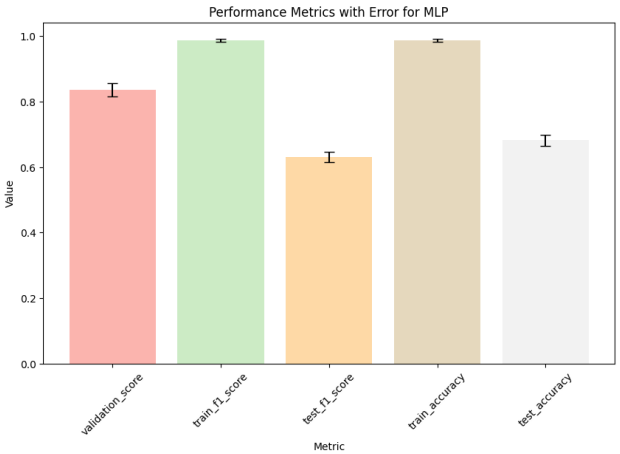

Figure 7: Performance metrics with error for the MLP model.

As depicted in Figure 7, this variability might affect its reliability in practical applications, pointing to a need for further tuning or consideration of model complexity to enhance stability and predictability.

4 FINAL RESULTS

4.1 Model Performance Comparison

The final evaluation of the machine learning models was conducted by comparing both the F1 score and accuracy, considering their respective errors, to determine the most reliable and robust classifier. The F1 score is particularly crucial in the context of binary classification as it conveys the balance between precision and recall. Accuracy, while a straightforward metric, is also essential as it represents the overall correctness of the model. The error bars in the figures indicate the standard deviation, thus reflecting the variability and consistency of the model's performance.

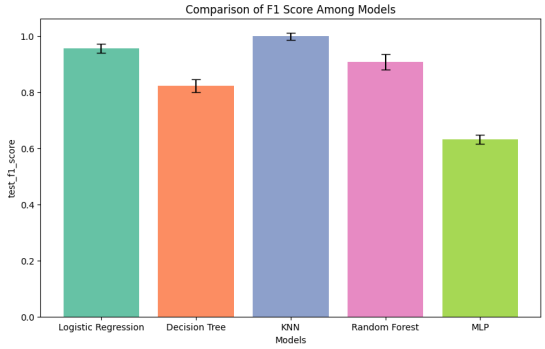

Figure 8: Comparison of F1 score among models

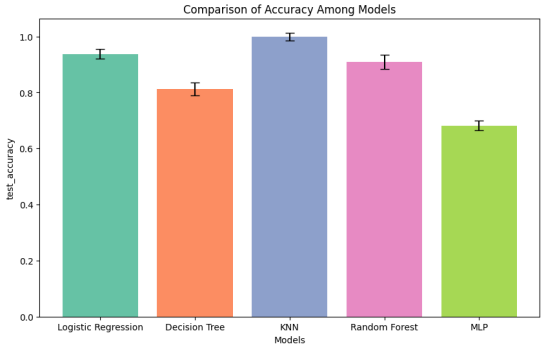

Figure 9: Comparison of accuracy among models

As shown in Figures 8 and 9, the KNN model achieved the highest F1 score and accuracy, indicating its strong predictive capabilities. However, the final selection favored the Logistic Regression model due to its lower variation in performance across training, testing, and validation. This lower variation signifies a model's robustness and suggests a higher likelihood of accurate predictions on new, unseen data. A model with less variability in its errors is preferred as it implies that the model is not overly fitted to the training data and can generalize well to new data, which is crucial for clinical applicability where the stakes are high and the cost of errors is significant.

4.2 Impact of Test Size on Model Performance

The evaluation of F1 scores across different test sizes provides a comprehensive picture of how model performance varies with respect to the proportion of data allocated for testing. The F1 score, a harmonic mean of precision and recall, is an essential metric for assessing the accuracy of a binary classifier, particularly when class distributions are imbalanced.

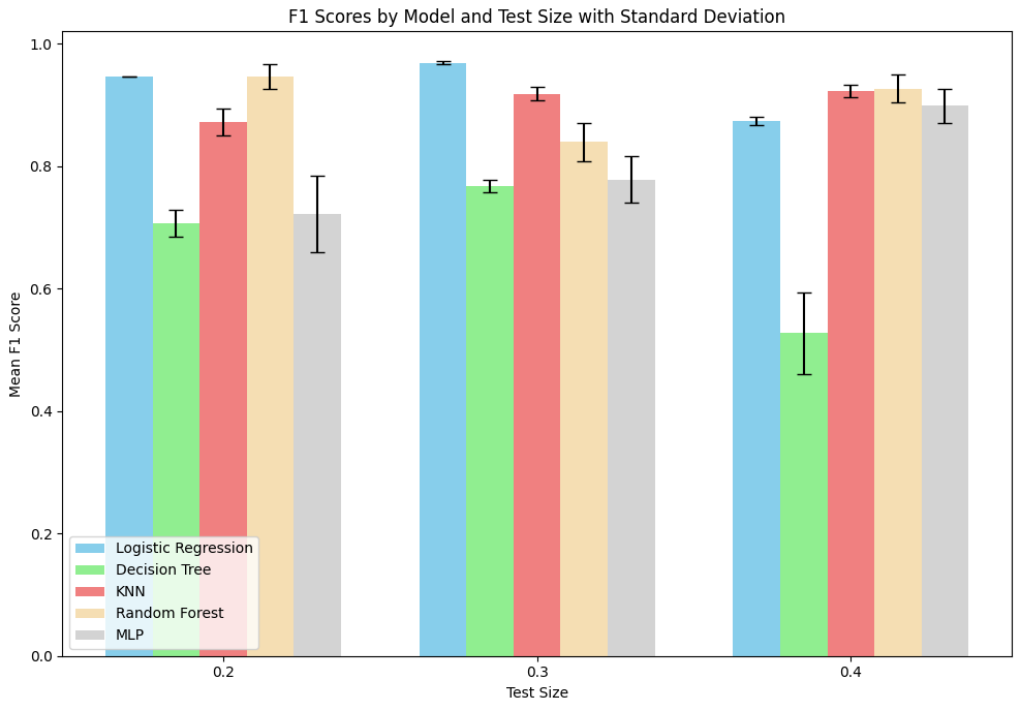

Figure 10: Mean F1 Scores by model and test size with standard deviation. This bar chart illustrates the dependency of model performance on the test size, with error bars representing the standard deviation of the mean F1 scores across different cross-validation folds.

Figure 10 demonstrates the mean F1 scores for each model at various test sizes, along with the associated standard deviation, which measures the spread of the scores. The chart reveals that certain models may perform better at specific test sizes, which could be indicative of their ability to generalize to new data. A smaller standard deviation is preferable, as it suggests that the model's performance is less sensitive to changes in the test data and is therefore more stable.

The analysis of these results led to the selection of the Logistic Regression model as the most suitable for our dataset, despite KNN showing slightly higher scores. The decision was informed by the observation that Logistic Regression had the least variation in performance across training, testing, and validation scores. This lower variation is indicative of a model's robustness, implying that it is not only able to fit well to the training data but can also generalize effectively to new, unseen data, thus increasing the likelihood of accurate predictions.

Table 1: Performance Metrics for Machine Learning Models Across Different Test Sizes

Model	Test Size	Validation Score (Mean ± SD)	Train F1 Score (Mean ± SD)	Test F1 Score (Mean ± SD)	Train Accuracy (Mean ± SD)	Test Accuracy (Mean ± SD)
Logistic Regression	0.3	0.9483 ± 0.0078	0.9978 ± 0.0015	0.9565 ± 0.0162	0.9973 ± 0.0018	0.9375 ± 0.0168
Decision Tree	0.3	0.7785 ± 0.0265	0.9736 ± 0.0075	0.8235 ± 0.0236	0.9716 ± 0.0080	0.8125 ± 0.0226
KNN	0.4	0.9293 ± 0.0113	0.9668 ± 0.0070	1.0000 ± 0.0127	0.9629 ± 0.0077	1.0000 ± 0.0136
Random Forest	0.2	0.9743 ± 0.0033	1.0000 ± 0.0000	0.9091 ± 0.0278	1.0000 ± 0.0000	0.9091 ± 0.0253
MLP	0.4	0.8360 ± 0.0201	0.9872 ± 0.0050	0.6316 ± 0.0158	0.9871 ± 0.0048	0.6818 ± 0.0168

As depicted in Table 1, we present a comprehensive overview of the performance metrics for each machine learning model, tailored to their optimal test size. This meticulous compilation showcases the models' mean validation scores and F1 scores, alongside the standard deviations, thus encapsulating both their accuracy and consistency. The table elucidates the models' predictive capabilities, with the highlighted test sizes being those that have yielded the most favorable balance between precision and reliability.

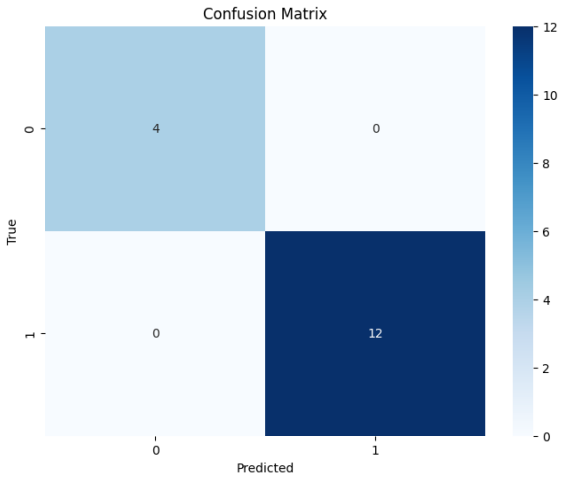

Figure 11: Confusion matrix of the Logistic Regression model demonstrating its predictive performance. The model has achieved 12 true positives with no false negatives or false positives, and 4 true negatives, indicating a high sensitivity and specificity. Such a perfect classification result, while encouraging, is atypical and suggests the need for further validation with a more extensive dataset.

The Logistic Regression model, selected for its consistent performance across various metrics, was further validated using a confusion matrix. The resulting matrix, presented below, indicates an exceptional performance with the model correctly predicting all instances of the positive class (1) with 12 true positives and no false negatives. This level of sensitivity is noteworthy. However, the absence of false positives and the presence of 4 true negatives may suggest an unusually perfect classification, which could be attributed to the limited size of the dataset. This anomaly calls for cautious interpretation and may warrant further investigation with a larger dataset to confirm the model's reliability.

5 CONCLUSIONS

- This study has undertaken a robust analytical approach to compare various machine learning models while contending with a limited dataset. Through meticulous cross-validation and hyperparameter tuning, each model was rigorously assessed to determine its performance stability and predictive power.
- The Logistic Regression model emerged as the most robust and reliable model, a finding that aligns with expectations given the nature of the problem at hand. Its inherent simplicity and effectiveness in binary classification tasks make it a suitable choice for predicting treatment responses in colorectal cancer cases, despite the challenges posed by the small sample size.

References

Deulofeu, M., García-Cuesta, E., Peña-Méndez, E. M., Conde, J. E., Jiménez-Romero, O., Verdú, E., Serrando, M. T., Salvadó, V., & Boadas-Vaello, P. (2021). Detection of sars-cov-2 infection in human nasopharyngeal samples by combining maldi-tof ms and artificial intelligence. *Frontiers in Medicine*, *8*. https://doi.org/10.3389/fmed.2021.661358 (cit. on p. 1).