Task A
Supervised Learning Model Evaluation Metrics

Diego De Pablo

depablodiego@uma.es
Health Engineering. Málaga University.

This work investigates the performance of various classification methods in a supervised learning context, focusing on how certain techniques can yield misleadingly high metrics. Specifically, some methods that classify all samples into a single category may achieve better performance metrics compared to methods that accurately differentiate between positive and negative samples. The analysis highlights the importance of robust metrics like the Jaccard index and F-measure, which provide valuable insights into model performance. Ultimately, it underscores the necessity of understanding the specific goals of the analysis to determine which metrics are most relevant for evaluating a model's effectiveness.

1 Introduction

Artificial intelligence (AI) has emerged as a transformative solution to numerous challenges in a wide range of domains, often being viewed as a key to simple and efficient problem-solving. However, the performance of AI models must be critically assessed to understand their real-world applicability and limitations. In particular, supervised learning algorithms are often used in classification tasks, where the performance of these models can be evaluated through specific metrics.

In this work, we explore the key performance metrics derived from the *confusion matrix* to assess and compare supervised learning models. These metrics include **Precision**, **Recall**, **Specificity**, **False Positive Rate**, **False Negative Rate**, **Accuracy**, **Spatial Accuracy**, **Jaccard Index**, and **F-measure**. These metrics are commonly used to provide a detailed view of a model's performance across different dimensions. By understanding and comparing these metrics, we can gain insights into the strengths and weaknesses of the classification methods being evaluated.[1]

In computational learning, applying classification algorithms to predict disease progression is critical for deriving meaningful insights from complex biomedical data. The primary focus of this project is to evaluate and compare the performance of several classification methods using a dataset relevant to disease classification. The analysis will focus on various aspects, including the dataset's class distribution, balance, and overall characteristics, and will use a range of well-established performance metrics.

The metrics used in this analysis provide a quantitative basis to assess each model's strengths and limitations, allowing us to determine which algorithm performs best for predicting outcomes in the dataset. Specifically, we will implement an algorithm to calculate several well-known metrics based on the number of **True Positives (TP)**, **True Negatives (TN)**, **False Positives (FP)**, and **False Negatives (FN)**.

These metrics will enable a comprehensive evaluation of each method's effectiveness in predicting for example disease progression. Furthermore, graphical representations, such as heatmaps and radar charts, will be used to provide a visual comparison of the methods. This will facilitate better understanding of each model's advantages and limitations, allowing for more informed decision-making when selecting the appropriate classification model.

2 Dataset Description

The aim of this work is to highlight the validation methods and explore examples ranging from realistic to exaggerated cases that, in certain scenarios, might be considered good results if certain metrics are ignored.

Even though these results may seem favorable, they can actually be misleading. To demonstrate this, we will base our analysis on the following dataset (observe the figure 1).

Method	TP	FP	FN	TN
A	100	900	0	0
B	80	125	20	775
C	25	25	75	875
D	50	50	50	850
E	0	0	100	900

Fig. 1. The methods studied in this work

The five hypothetical methods in this study are evaluated on a dataset of **1000 samples**, divided into **two classes** (positive or negative). This dataset is notably **imbalanced**, with 100 positive and 900 negative samples, which can lead to biased models and misleading metrics, as seen with methods A and E that perform poorly on specific measures.

Imbalanced datasets skew results, making some models appear more effective than they are. Techniques such as **oversampling**, **undersampling**, and **weight adjustment** are used to mitigate these effects, ensuring more reliable evaluations.

Figure 2 illustrates the percentage confusion matrix for each method, alongside a sixth matrix representing a perfect model that correctly identifies all positive and negative samples.

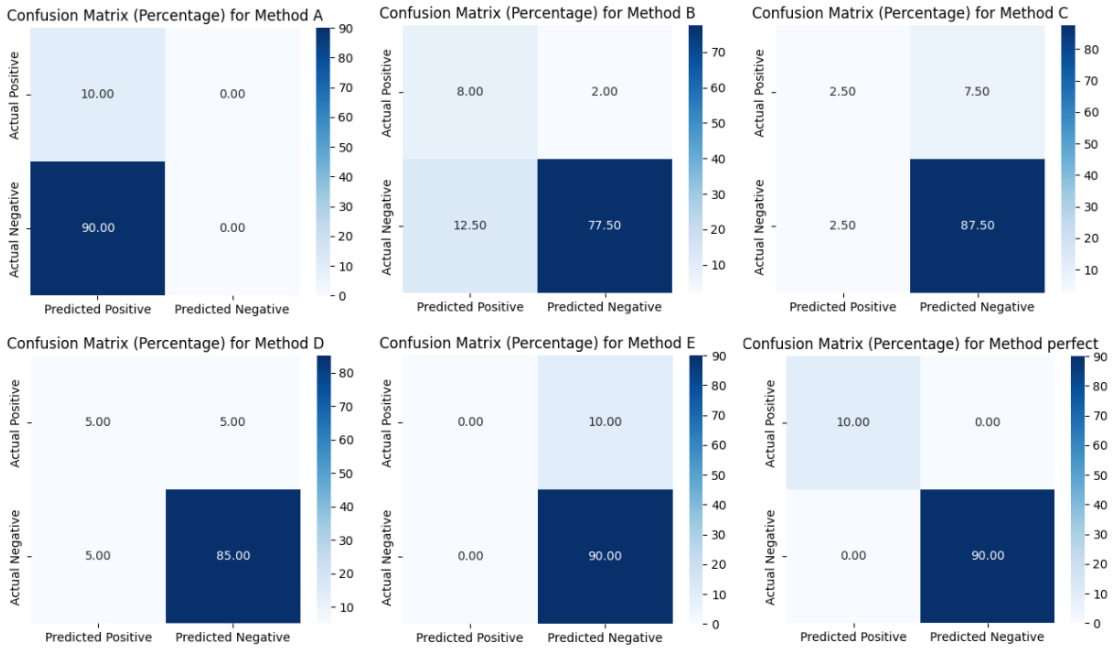

Fig. 2. 6 confusion matrices, from A to E and the case of the perfect confusion matrix for this dataset

Examining the confusion matrix for each method in relation to the ideal confusion matrix for the dataset reveals significant shortcomings in methods A and E, which classify all samples as either positive or negative. In contrast, methods B, C, and D demonstrate more realistic classification patterns, aligning more closely with the expectations of an effective machine learning approach. Although these methods approach the performance of a perfect model, they still exhibit a small percentage of false positives and false negatives. To determine which of these three methods is superior, it is essential to consider additional factors, such as validation metrics and the specific objectives for which the model is intended.

3 Metrics Overview

In this work, we implemented a function to compute various validation metrics using the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) yielded by each method. The results are visualized in a heatmap (see the Figure 3) to clearly present the performance of each method.

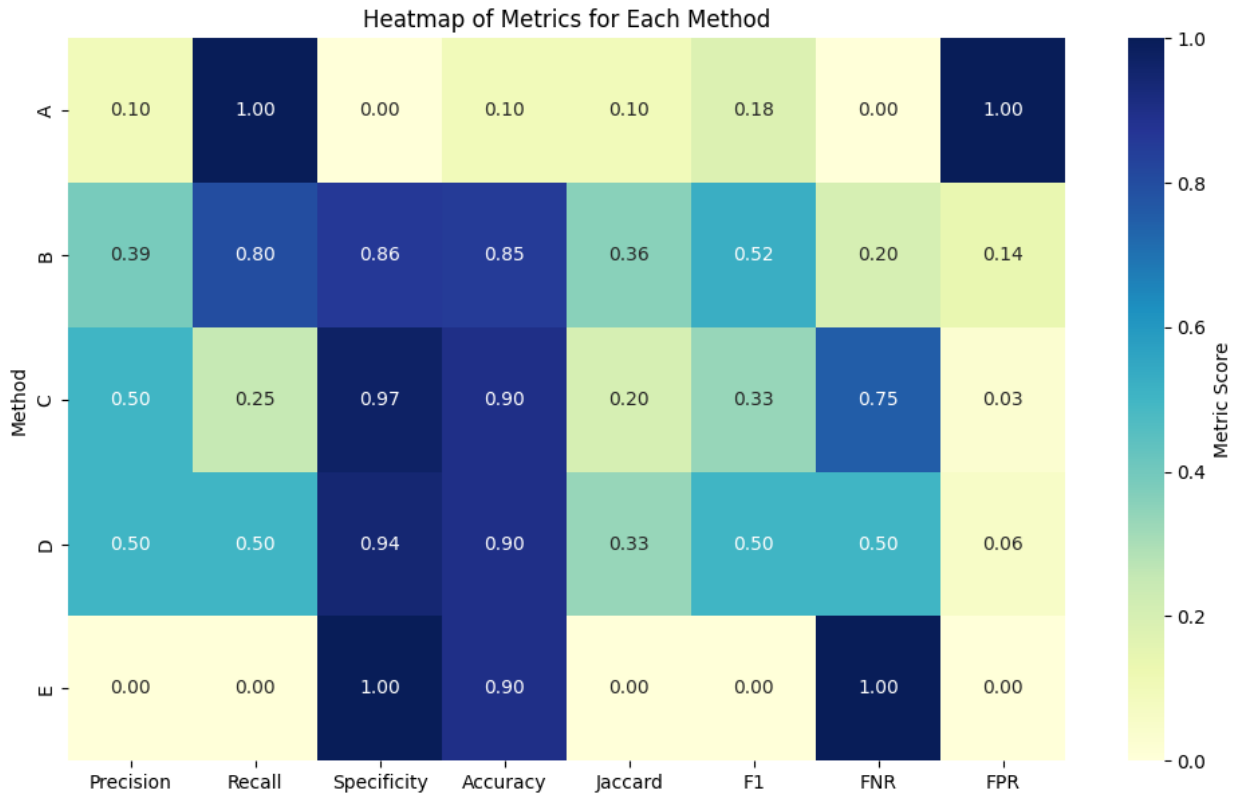

Fig. 3. The methods studied in this work

3.1 Summary of Expected Ranges

- Precision, Recall, Specificity, Accuracy, Jaccard, F1: Values range from 0 to 1, with higher values indicating better performance.
- FNR, FPR: Values range from 0 to 1, with lower values indicating better performance.

3.2 Evaluation Metrics

In machine learning, evaluation metrics are crucial for assessing model performance, particularly in classification tasks. Each metric provides unique insights into various aspects of the model's predictions. Below is a breakdown of the key metrics used in this study:

Precision (PR) Also known as positive predictive value, measures the proportion of true positive predictions out of all positive predictions made by the model. It provides insight into the model's ability to avoid false positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Indicates how many of the predicted positives were actually correct. It is especially useful when false positives are costly, such as in spam detection.

Looking at Table 1, where each method is accompanied by its precision metric:

Metric	**A**	**B**	**C**	**D**	**E**
Precision	0.10	0.39	**0.50**	**0.50**	0.00

Table 1. Precision metrics for each method.

It can be observed how methods C and D obtain the highest values, being 0.50 respectively, while the lowest is E with a precision of 0.00.

Recall (RC) Also known as sensitivity or true positive rate (TPR), tells us how many of the actual positives were correctly identified by the model. It answers: "Out of all the real positive cases, how many were captured by the model?"

It is crucial in cases where missing a positive case is costly (false negatives are highly undesirable). In medical diagnoses, high recall means fewer missed cases of a disease.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Looking at Table 2, where each method is accompanied by its recall metric:

Metric	**A**	**B**	**C**	**D**	**E**
Recall	**1**	0.80	0.25	0.50	0.00

Table 2. Recall metrics for each method.

Method A shows the highest recall, achieving a perfect score of 1 by capturing all positive cases. However, this metric does not consider negatives, and further analysis reveals that this method performs poorly overall. In contrast, Method E fails to identify any positives.

Specificity (SP) or true negative rate (TNR), measures the proportion of actual negatives that were correctly identified. It answers: "Of all the real negative cases, how many did the model correctly classify as negative?"

It is valuable when it's important to correctly identify negatives, such as in fraud detection, where you want to minimize false positives.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Looking at Table 3, where each method is accompanied by its specificity metric:

In this case, Method E achieves perfect specificity (Similar to the previous case), indicating it did not classify any negatives incorrectly, while Method A performs poorly.

Metric	A	B	C	D	E
Specificity	0.00	0.86	0.97	0.94	**1**

Table 3. Specificity metrics for each method.

Accuracy (ACC) Measures the overall proportion of correct predictions, including both positives and negatives. It answers: "Out of all the samples, how many did the model classify correctly?"

It is often used as a basic measure of model performance, but it can be misleading on imbalanced datasets.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Looking at Table 4, where each method is accompanied by its accuracy metric:

Metric	A	B	C	D	E
Accuracy	0.10	0.86	**0.9**	**0.9**	**0.9**

Table 4. Accuracy metrics for each method.

From the table, Methods C, D, and E exhibit high accuracy, reflecting their performance despite the imbalanced dataset.

Jaccard Index (J) Also known as Intersection over Union (IoU), measures the overlap between predicted positives and actual positives. It considers both false positives and false negatives.

$$\text{Jaccard} = \frac{TP}{TP + FN + FP}$$

Looking at Table 5, where each method is accompanied by its Jaccard index:

Metric	A	B	C	D	E
Jaccard	0.10	**0.36**	0.20	0.33	0.00

Table 5. Jaccard metrics for each method.

In this analysis, Method B performs best in terms of Jaccard index, showing a reasonable overlap between predicted and actual positives.

F-measure (F1) Is the harmonic mean of precision and recall. It balances the trade-off between precision and recall by giving more weight to lower values, which makes it especially useful when you need a balance between precision and recall.

$$F1 = \frac{2 \cdot PR \cdot RC}{PR + RC}$$

Looking at Table 6, where each method is accompanied by its F-measure:

As shown in the table, Method B has the highest F1 score, balancing the trade-offs between precision and recall.

False Negative Rate (FNR) Represents the proportion of actual positives that the model incorrectly predicted as negatives. It answers: "Of all the current positives, how many did the model miss?"

$$FNR = \frac{FN}{TP + FN}$$

Metric	A	B	C	D	E
F1	0.18	**0.52**	0.33	0.50	0.00

Table 6. F-measure metrics for each method.

Metric	A	B	C	D	E
FNR	**0**	0.20	0.75	0.50	1.00

Table 7. False Negative Rate metrics for each method.

Looking at Table 7, where each method is accompanied by its false negative rate:

Just as method A benefits from classifying all samples as positive, method B also stands out compared to the other methods, Method B has a significant false negative rate, which is concerning in scenarios where identifying positives is crucial.

False Positive Rate (FPR) Is the proportion of actual negatives that were incorrectly classified as positives. It answers: "Of all the real negative cases, how many did the model falsely classify as positive?"

$$FPR = \frac{FP}{FP + TN}$$

Looking at Table 8, where each method is accompanied by its false positive rate:

Metric	A	B	C	D	E
FPR	1.00	0.14	0.03	0.06	**0**

Table 8. False Positive Rate metrics for each method.

Here, similar to other cases, method E only classifies all samples as false, obtaining very good results, but methods B, C and D obtain great results.

3.3 Radar plot

Radar plots provide numerous advantages for visualizing data, particularly when comparing multiple variables. In this analysis (see the Figure 4) , the FNR and FPR metrics were excluded because they are assessed in reverse: values closer to 1 indicate poorer performance, while values nearer to 0 represent better results. To enhance clarity, only metrics where higher values indicate better performance will be included in the radar plots.

The radar plot (Figure 4) allows for several key observations:

- **Method A** shows excellent *recall*, achieving the highest score, but its performance is poor in all other metrics, such as precision, accuracy, and specificity. This suggests that while Method A captures all positive cases, it struggles to correctly identify negative ones.
- **Method B** presents a more balanced performance across all metrics. It performs well in *recall* and has moderate precision and specificity, making it a more reliable option compared to Method A.
- **Methods C and D** stand out with a precision of 0.5, and they show strong overall performance in various metrics like accuracy and specificity, indicating these models maintain a reasonable balance between correctly identifying both positive and negative cases.
- **Method E**, having no true positive cases, naturally scores low in most metrics, with zero precision and recall.

The radar plot highlights the strengths and weaknesses of each method, visually illustrating the trade-offs between metrics. Methods B and D offer the best balance, while Method A, despite its high recall, performs poorly in other key metrics, demonstrating its limitations.

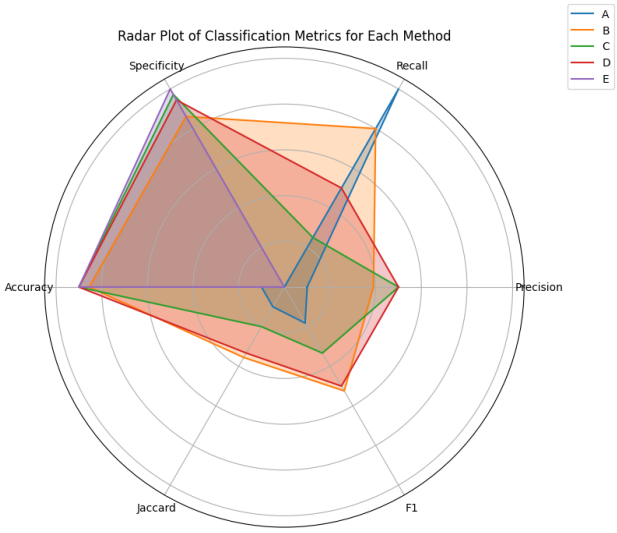

Fig. 4. The Radar plot

4 Metrics Comparison

In addition to analyzing all metrics at once, similar to bivariate analysis, we can highlight specific metrics for a more focused evaluation. By plotting methods on a 2-dimensional plane, where the x-axis represents one metric and the y-axis another, we can carefully assess the direct relationships between these values.

4.1 False Negatives (FN) against False Positives (FP)

It has already been observed that certain methods yield a high number of false negatives and false positives. Plotting both values on a 2-dimensional plane allows for a more intuitive visualization of which methods are performing better (see the figure 5). The closer the methods are to the origin (0,0) — having fewer false positives and false negatives — the more ideal they are.

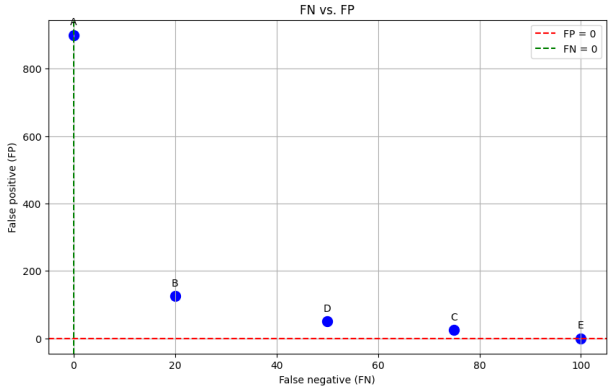

Fig. 5. FN vs. FP for each method.

As it has been a constant in the work methods A and E gave terrible results for the reasons already explained. And the methods B and D stand out positively as expected.

This visualization is important because false positives (FP) and false negatives (FN) directly impact the overall performance of a classifier. A method that minimizes both types of errors is considered optimal, especially in critical applications like medical diagnosis, where both kinds of mistakes can lead to significant consequences.

4.2 Precision (PR) against Recall (RC)

A Precision vs. Recall plot provides insight into the trade-off between these two metrics (see Figure 6), improving precision leads to lower recall, and vice versa. This visualization is particularly valuable when dealing with imbalanced classes. By observing the balance between precision and recall, one can determine the optimal threshold that best balances both metrics, allowing for model performance adjustments based on the specific needs of the task.

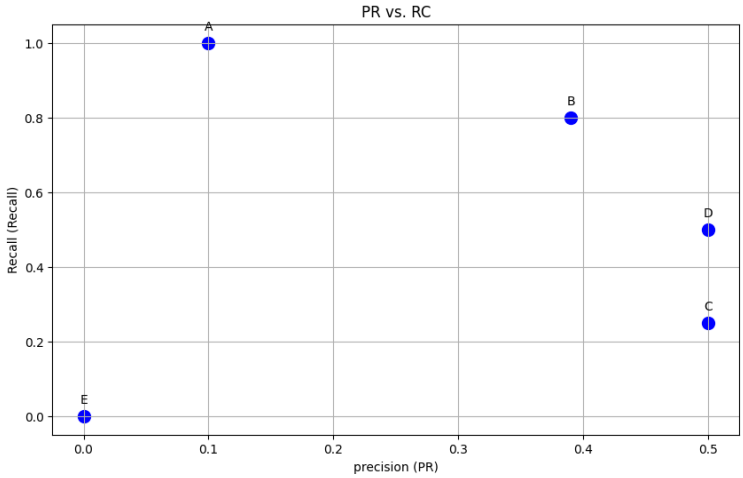

Fig. 6. PR vs. RC for each method.

The most ideal position for a method is close to the point (1,1), indicating high precision and high recall. Method B stands out as the best performing, achieving a balanced high value for both metrics, while method E performs poorly, as expected from previous observations.

References

1. S. Liu, F. Roemer, Y. Ge, et al. Comparison of evaluation metrics of deep learning for imbalanced imaging data in osteoarthritis studies. *Osteoarthritis and Cartilage*, 31(9):1242–1248, 2023.