# Task B

## 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.

## 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

## 3 Conclusion

## 4 Repository Github

Further information, including the source code and full project documentation, can be accessed through the GitHub repository. Click here to go to the repository.

## 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.