# Evaluating Machine Learning Models:Multi-Metric, Fairness & Multi-Class Perspective

Mikołaj Sujka

Applied Computer Science, Jagiellonian University

## Abstract

This poster examines how different evaluation metrics influence model selection for regression, binary and multi-class classification, as well as generative tasks. We apply RMSE, MAE, $R^2$, Accuracy, F1 (micro/macro), AUC, BLEU/ROUGE, and fairness metrics (Demographic Parity, Equal Opportunity) across multiple datasets. Automated tuning via GridSearchCV and experiment tracking with MLflow ensure reproducibility. Key findings highlight trade-offs between accuracy and fairness, limitations of standard metrics in generative scenarios, and challenges extending binary metrics to multi-class settings.

## Introduction

In real-world applications, no single metric captures all aspects of model performance. Different metrics emphasize different errors: RMSE penalizes large deviations in regression, while F1 and AUC focus on class imbalance in classification. Moreover, optimizing accuracy alone can mask systematic biases against protected groups. We also discuss why binary metrics do not trivially generalize to multi-class evaluation, and outline difficulties in assessing generative model outputs.

- Objectives: compare multiple metrics, analyze fairness vs. accuracy trade-offs, explore generative and multi-class evaluation techniques.

## Related Work

The comprehensive survey by Author et al. maps the evolution from accuracy and MSE to advanced metrics like F1 and AUC, and demonstrates conflicting model rankings under different criteria [**?** ]. Reviews on fairness outline data, algorithmic, and interaction biases, proposing metrics such as Demographic Parity and Equal Opportunity Difference [**?** ]. Emerging research highlights the need for robust evaluation of generative models (using BLEU, ROUGE, and human judgment) and identifies gaps in multi-class metric theory.

## Methods

**Datasets:** UCI Boston (regression), UCI Adult (binary classification with sensitive attributes), Iris and MNIST (multi-class).

**Models:** Linear Regression, Random Forest Regressor; Logistic Regression, XGBoost, Feed-forward Neural Network.

**Metrics:** RMSE, MAE, $R^2$; Accuracy, Precision, Recall, F1 (micro/macro), AUC-ROC; Demographic Parity Difference, Equal Opportunity Difference; BLEU, ROUGE, and subjective human evaluation for generative models.

**Validation & Tuning:** Stratified k-fold cross-validation; GridSearchCV for hyperparameter tuning; MLflow for experiment tracking and reproducibility.

## Results

- **Regression:** RMSE vs. MAE comparison shows that models with lower RMSE may have higher MAE, indicating distribution-dependent error trade-offs.
- **Binary Classification:** ROC curves and precision-recall analysis reveal that optimizing for AUC can reduce F1-score in imbalanced settings.
- **Fairness Analysis:** On the Adult dataset, enforcing Demographic Parity resulted in a 5% drop in overall accuracy while reducing group disparity by 40%.
- **Multi-Class:** Confusion matrices and macro/micro-F1 comparisons demonstrate that macro-F1 penalizes rare classes more heavily.
- **Generative Evaluation:** BLEU and ROUGE scores correlate poorly with human judgments of text diversity and relevance.

## Discussion

Our experiments confirm that the choice of metric fundamentally alters which model is deemed "best." Improving fairness metrics often comes at the cost of predictive accuracy, highlighting a stakeholder-driven trade-off. Standard metrics fail to capture nuances in generative tasks, requiring human-in-the-loop evaluation. Extending binary classification metrics to multi-class problems introduces additional complexity in weighting and aggregation.

## Conclusion Future Work

**Key Takeaways:** There is no universal evaluation metric; practitioners must align metric choice with domain-specific goals and fairness considerations.
**Future Directions:** Develop advanced metrics for generative models, integrate AutoML and MLflow pipelines for seamless evaluation, and extend fairness metrics to multi-class contexts.

## References

[1] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

## Source Tools

- **Literature:** Google Scholar, IEEE Xplore, ACM Digital Library, ArXiv.
- **Key Papers:** Survey on ML metrics (2021), Systematic review of bias (2022).
- **Libraries:** scikit-learn (GridSearchCV, cross$_v al_s core$), $TensorFlow$, $PyTorch$.**Platforms:**$Kaggle, UCIML$