CO_2 classification analysis for AI models

Juan Felipe Gallo Rendón* *Engineering department*
University of Antioquia
Medellín, Colombia

***Abstract*—This study presents a comprehensive classification analysis of artificial intelligence models based on their carbon footprint and performance metrics. A workflow was implemented, beginning with an unsupervised cluster analysis (hierarchical and k-means) to identify the intrinsic data structure, which revealed that model popularity (`downloads`, `likes`) is the primary natural segmentation factor. Subsequently, the feasibility of Linear Discriminant Analysis (LDA) was assessed through a ity diagnostic, which showed significant deviations and justified the use of a regularized variant. Finally, the performance of LDA was compared against two non-linear models—a Support Vector Machine with an RBF kernel (SVM-RBF) and Gradient Boosting (GB)—to predict a binary "fairness" label (`is_fair`). The results demonstrate that while Gradient Boosting achieves perfect predictive accuracy (AUC=1.000), regularized LDA offers the best interpretability, identifying `performance_score` and `co2_eq_emissions` as the key discriminant factors. The work concludes that the optimal model selection depends on the objective: maximum precision (GB) or phenomenon understanding (LDA).**

***Index Terms*—GreenAI, Cluster analysis, K-means, Classification models, Linear analysis, Non-linear analysis**

I. Introduction

In the current era of artificial intelligence, the evaluation of models transcends traditional performance metrics. The growing awareness of the environmental impact of large-scale computations has given rise to the concept of "Green AI," which advocates for a balance between predictive efficacy and resource efficiency. This work addresses this issue by analyzing the `HFCO2.csv` dataset, derived from a CO_2 emissions report on the Hugging Face platform.

The primary objective is twofold: first, to explore the underlying structure of a catalog of AI models to understand how they naturally group together, and second, to build a supervised classification model capable of predicting whether a model is "fair" (`is_fair`), a derived label that relates its performance to its carbon footprint. To achieve this, a rigorous methodology is followed, including cluster analysis, diagnosis of statistical assumptions such as multivariate normality, and a systematic comparison between an interpretable linear model (Linear Discriminant Analysis) and high-performance non-linear alternatives (SVM and Gradient Boosting). This comprehensive approach not only seeks to identify the most accurate classifier but also to extract meaningful insights about the characteristics that define an AI model as both efficient and effective.

II. Methodology

Classification methods are applied to the selected dataset. A workflow comprising six procedures is followed: **univariate analysis**, **cluster analysis**, **k-means clustering**, **linear classification**, and **nonlinear classification**. Upon completion of these steps, the results are compared and the best-performing model is identified.

A. Univariate Analysis

In this stage, the dataset **HFCO2.csv**—originating from a CO_2-emissions report on the Hugging Face platform—is profiled to characterize marginal distributions, identify outliers, and screen variables for subsequent modeling. The variables under study are enumerated below. Variables such as `modelId`, `datasets`, `co2_reported`, `createdat`, and `libraryname` are excluded due to limited analytical relevance; `database_efficiency` is removed because it is functionally dependent on `co2_eq_emissions` and `size`. Because no native target is present, classification labels are defined to enable supervised evaluation. First, `model_type` is created to align the `performance_score` with the appropriate metric family (e.g., accuracy, F1, Rouge); rows lacking performance metrics are discarded. Univariate summaries and frequency counts are then computed for qualitative variables (e.g., `model_type`) and proportions are reported for `domain`; variables exhibiting near-constant distributions (e.g., `library_name` = `pytorch`), or extreme sparsity (`geographical_location`, `environment`) are removed. Finally, a fairness-oriented label is introduced to relate predictive performance to environmental impact, yielding four categories—*Fair and Efficient*, *Powerful but Expensive*, *Green but Weak*, and *Inefficient*—and a derived Boolean variable `is_fair`. The resulting curated dataset and label structure provide the basis for the next methodological step, cluster analysis.

B. Cluster analysis

AHierarchical clustering was applied to segment AI models using numeric performance-and-cost variables, which were standardized to z-scores. Several configurations were evaluated by combining distance metrics (Euclidean, Mahalanobis) with linkage rules (Ward, average, complete, single). The quality of each configuration was primarily assessed using the **cophenetic correlation**, which measures how faithfully the resulting dendrogram preserves the original pairwise distances between data points. A higher value indicates a better preservation of the original data structure.

C. K-means analysis

K-means was applied to the same numeric variables used in the hierarchical analysis (*performance_score*, *co2_eq_emissions*, *likes*, *downloads*, *size*). Rows with non-numeric/inf values were removed and features were standardized to z-scores (StandardScaler). Am exploration with $k \in \{2, \ldots, 10\}$ with k-means++ initialization (n_init = 50, max_iter=500, random_state=42).

For each k Silhouette (higher is better) were reported, Calinski–Harabasz (CH; higher is better), Davies–Bouldin (DBI; lower is better), and the inertia (*elbow* curve). The primary selector was the maximum Silhouette; CH/DBI and the elbow were used as secondary evidence.

For the selected k cluster sizes were exported, original-unit profiles (count/mean/median), z-centroids, and a variable ranking using $\max |z|$ (to highlight the most discriminative features). External labels (*is_fair*, *clasification_fairness*, *model_type*) were held out from training and only used for validation via contingency tables, χ^2/Cramér's V, Fisher's exact test when 2×2, and ARI when applicable.

D. Multivariate Normality & LDA Feasibility — Methods

Check whether the feature vector X=performance_score, co2_eq_emissions, likes, downloads, size is (approximately) multivariate normal—globally and within classes—so that classical LDA is justified.

Preprocessing. Remove rows with NA/inf; standardize all variables to z–scores.

Robust scatter & distances. Fit a shrinkage (Ledoit–Wolf) covariance $\widehat{\Sigma}$ to mitigate outliers. Compute squared Mahalanobis distances $d_i^2 = (x_i - \widehat{\mu})^\top \widehat{\Sigma}^{-1} (x_i - \widehat{\mu})$ (i) globally and (ii) per class.

Normality diagnostic. Under MVN with p variables, $d_i^2 \sim \chi_p^2$. Compare empirical quantiles of $\{d_i^2\}$ against χ_p^2 using QQ–plots: global and class–conditional. Large, systematic upward deviations (heavy tails/mixtures) indicate non–normality.

Optional formal checks. (a) Mardia or Henze–Zirkler tests on z–scores; (b) Box's M (or visual comparison) for equality of class covariances.

Decision rule. nosep,leftmargin=*

1) *Proceed with classical LDA* if QQ–plots track the χ_p^2 line reasonably well (globally *and* by class) and class covariances appear similar.
2) *Prefer robust/regularized LDA or nonparametric classifiers* (e.g., shrinkage LDA, logistic regression, tree–based) if tails inflate, classes differ in scatter, or tests reject MVN/homoscedasticity.

E. LDA

Multivariate normality was examined via Mahalanobis Q–Q diagnostics (globally and by class). Because clear departures from normality were observed, regularized discriminants were preferred. Linear Discriminant Analysis (LDA) with automatic shrinkage (and, for projections, the eigen solver) and Quadratic Discriminant Analysis (QDA) with mild regularization were fit on a stratified train/test split. Performance was summarized with confusion matrices, precision/recall/F1, ROC–AUC, and 5–fold stratified cross–validation. Model parameters (priors, class means, coefficients, scalings), LD scores, and plots (ROC, score histograms, dendrograms) were exported to files to enable reporting and reproducibility.

F. Non-linear models comparison)

A supervised classification workflow was implemented to benchmark a linear baseline (Linear Discriminant Analysis, LDA) against two non-linear alternatives: an RBF–kernel Support Vector Machine (SVM–RBF) and Gradient Boosting (GB). The target was the binary label is_fair; predictors were the numeric performance–cost variables used in previous sections. Missing values were removed, then features were transformed with $\log(1 + x)$ (for skewed counts: CO_2, likes, downloads, size) and standardized (zero mean, unit variance). A stratified train/test split (70/30) preserved the empirical class imbalance.

Models were trained in scikit-learn as follows. LDA used the eigen solver with automatic shrinkage, providing a robust shared covariance estimate and interpretable discriminant directions. SVM–RBF used probability calibration (probability=True) and an RBF kernel; GB used logistic loss with modest depth and learning rate (standard defaults), both selected via stratified 5-fold cross-validation (CV) on the training fold. Evaluation focused on rank-based and decision-based criteria: AUC(ROC), average precision (AP) for the positive class (*True*), F1 for the positive class, overall accuracy, and Brier score (probability calibration). Error structure was inspected through confusion matrices. For comparability, models were scored on the same held-out test set; CV accuracy was also reported to assess stability across resamples. Curves (ROC and Precision–Recall) and panels (confusion matrices) were produced for visual diagnostics and threshold selection.

III. Results and discussions

A. Univariate Analysis

The database **HFCO2.csv** is a product of the CO_2 emissions report generated on the Hugging Face platform. In this case, the analysis is focused on the following variables:

1) co2_eq_emissions: the resulting carbon footprint
2) downloads: number of model downloads
3) likes: number of model likes
4) performance_metrics: (accuracy, F1, Rouge-1, Rouge-L)
5) performance_score: the harmonic mean of the normalized performance metrics
6) size: size of the final trained model in MB

Variables like modelId, datasets, co2_reported, createdat, and libraryname were removed because of their lack of importance in the analysis. database_efficiency was also removed because it is a dependent variable of co2_eq_emissions and size. The first step was to choose metrics for classification, as the database *per se* does not have a relevant dependent variable to be classified. Therefore, some labels were created with the aim of creating a dependent variable to be predicted by a couple of machine learning models. Accordingly, model_type is the label used to determine which metric the model uses for its performance score, because not all models use the same metric for evaluation. After applying this classification, some rows were unusable because hundreds of them do not have performance metrics at all, so they were deleted. After the redundant rows were erased, a count was made on the qualitative variables, as shown in the following table:

TABLE I: model_type counts

model type	count
type1 (accuracy & f1)	773
type2 (rouge)	228
type3 (accuracy)	72
type4 (rouge1)	3
type5 (f1)	2

In an effort to find relationships between model types and the other categories, a summary of the other variables was prepared, focusing on the `domain` variable, which is the model's use case. The following table shows each domain and the associated percentage:

TABLE II: percentage of domains

model type	count
NLP	81,5%
Computer Vision	16,0%
Not specified	2,5%

In conclusion, there are only two major types of domain models. For the purpose of this analysis, this is not an influential variable, so it was removed. Other variables, like `library_name`, revealed that all models use `pytorch` as a library, so this was also removed. The same reasoning applied to `geographical_location`: the analysis revealed that 99% of the models do not specify the training location. Similarly, 99.7% of the models do not report the hardware used in `environment`. After applying analysis upon the mean of performance metrics and CO_2 emissions, the analysis revealed that those models could be labeled again with a more precise metric: *fairness*, defined as a relation between performance score and CO_2 emissions. The following are the labels chosen for the models

1) Fair and Efficient
2) Powerful but Expensive
3) Green but Weak
4) Inefficient

Finally, models were classified by a new boolean variable "is_fair", and those that were previously labeled with respective performance metrics, were evaluated: Results are consigned in the route: `/notebooks/part2/1-univarated-analysis.ipynb` Those classifications are required for the next step in the methodology, Cluster analysis.

TABLE III: Model Classification based on their fairness and performance metrics

Model Type	**Classification Fairness**	**Count**
type1 (accuracy & f1)	Fair and Efficient	229
	Powerful but Expensive	214
	Green but Weak	181
	Inefficient	142
	Incomplete data	7
type2 (rouge)	Inefficient	140
	Green but Weak	57
	Powerful but Expensive	18
	Fair and Efficient	9
	Incomplete data	4
type3 (accuracy)	Fair and Efficient	50
	Powerful but Expensive	9
	Green but Weak	7
	Incomplete data	4
	Inefficient	2
type4 (f1)	Incomplete data	2
type5 (rouge1)	Incomplete data	3

B. Cluster Analysis

A comprehensive grid search over distance metrics and linkage rules was conducted to identify the most faithful hierarchical clustering structure. The primary evaluation criterion was the **cophenetic correlation**, which assesses how well the dendrogram preserves the original pairwise distances in the data. A higher value indicates superior structural fidelity.

The analysis revealed that the linkage methods based on averages provided the most robust representations. Specifically, **Average-Euclidean** ($r_{coph} \approx 0.813$) and **Average-Mahalanobis** ($r_{coph} \approx 0.812$) achieved the highest cophenetic correlation scores. This suggests that these configurations create a tree structure that accurately reflects the data's original topology. In contrast, the **Ward-Euclidean** method yielded the lowest fidelity ($r_{coph} \approx 0.438$), indicating significant distortion of the original distances.

While the **Single-Euclidean** configuration produced clusters with high internal separation (previously noted with a silhouette score of ≈ 0.62), its cophenetic correlation was considerably lower ($r_{coph} \approx 0.712$). This highlights a classic trade-off: Single linkage is adept at identifying well-separated, distinct groups, but average linkage provides a more reliable and globally accurate representation of the data's underlying structure. Given the priority of structural fidelity, the **Average-Euclidean** configuration with $k = 4$ was selected as the most representative solution for profiling.

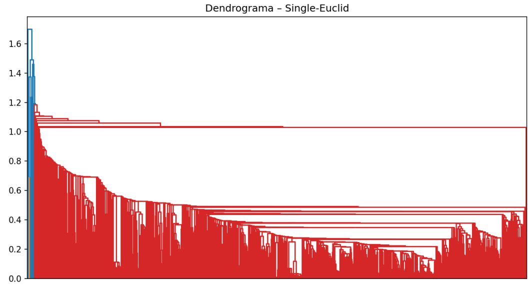


Fig. 1: Dendrogram for a representative configuration. The horizontal cut yields the selected number of clusters.

Across all robust configurations, a consistent pattern emerged: a dominant, large "catalog" cluster containing the bulk of models, and one to three smaller "elite" clusters. These elite clusters group models with markedly higher popularity metrics (`downloads` and `likes`). The analysis of z-centroids confirmed that **downloads** is the most discriminative feature (max $|z| \approx 5.39$), followed by **likes** (max $|z| \approx 2.03$), with `co2_eq_emissions`, `performance_score`, and `size` contributing more modestly.

External validation showed a weak but noticeable association between the clusters and the `is_fair` label. For instance, the small elite clusters tended to have a higher concentration of models where `is_fair=True` compared to the large catalog cluster. This suggests that high-engagement models are slightly more likely to be classified as "fair."

a) Limitations.: It is important to acknowledge that applying hierarchical clustering to a dataset with over 1000 observations presents challenges. Dendrograms become visually dense, and the stability of the resulting partitions can be less reliable than with smaller datasets. Therefore, these findings should be interpreted as a descriptive exploration of the data's latent structure, driven primarily by model popularity, rather than as a definitive and stable segmentation.

C. K-means analysis

K-means was ran on the standardized variables (*performance_score*, *co2_eq_emissions*, *likes*, *downloads*, *size*) for $k \in \{2, \ldots, 10\}$ with `k-means++`, n_{init}=50, `max_iter`=500. Model selection combined four internal criteria:

- **Silhouette (max is best).** The curve peaked at $\mathbf{k = 2}$ with $\approx \mathbf{0.286}$, and decreased thereafter (Fig. 2).
- **Calinski–Harabasz (max is best).** The maximum occurred at k=3 ($\approx$ 330), with k=2 close behind ($\approx$ 290) (Fig. 2).
- **Davies–Bouldin (min is best).** DBI decreased monotonically with k, reaching $\approx$1.21 at k=10 (Fig. 3).
- **Elbow (inertia).** Inertia dropped rapidly up to k∼6 and then flattened, with no sharp elbow afterwards (Fig. 3).

Given the primary criterion (silhouette), $\mathbf{k = 2}$ were selected. The resulting partition produced **unbalanced cluster sizes** (about n∼740 vs. n∼300; Fig. 4). Cluster profiles (in original units) and z–centroids were computed; the ranking by $\max |z|$ again highlighted *downloads* as the most discriminative variable, followed by *likes*, with the remaining variables contributing less. Cross-tabs against the external labels (*is_fair*, *clasification_fairness*, *model_type*) were generated for k=2 (see Supplement for the full tables).

The internal criteria point to different trade-offs: silhouette clearly favors k=2, Calinski–Harabasz prefers k=3, and Davies–Bouldin keeps improving as k grows. In this context, silhouette were prioritized because it balances compactness and separation and is less biased by the growth of k than CH/DBI. The *elbow* curve shows diminishing returns beyond k∼6, supporting the choice of a small number of groups.

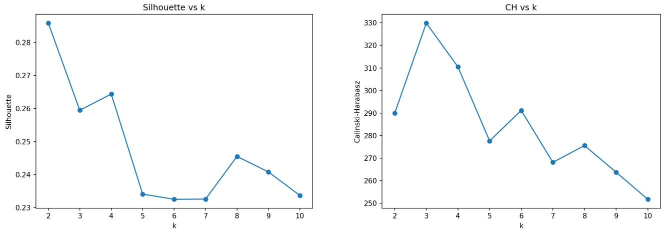


Fig. 2: Silhouette and Calinski–Harabasz across k.

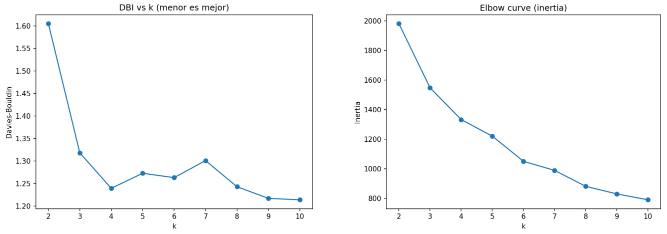


Fig. 3: Davies–Bouldin (lower is better) and inertia (elbow) across k.

The k=2 solution is *interpretable* and *stable* (across restarts) but *unbalanced*: one major segment (catalog-like) and a smaller, high-intensity segment. The profile tables indicate that the minor cluster concentrates substantially higher *downloads* and *likes* (and slightly higher *performance*), while medians for *size*/CO_2 are not dominant—consistent with the hierarchical analysis. The $\max |z|$ ranking confirms *downloads* (then *likes*) as the strongest separators, which aligns with the business intuition that usage/engagement variables drive the segmentation more than footprint or size.

Regarding *external* labels, the cross-tabs for k=2 show heterogeneous mixes rather than perfectly pure clusters, i.e., they are useful for characterization but should not be read as supervised classes. This is expected because those labels were not used for training. Overall, the k-means segmentation complements the hierarchical results: it recovers the same two-segment story (catalog vs. high-traction niche), is easier to deploy, and preserves the same ordering of discriminative variables. If more granularity is needed, k=3 is a reasonable alternative per CH, though with a slight loss in silhouette and added complexity.

D. Multivariate Normality & LDA Feasibility

An assessment of multivariate normality, a key assumption for classical Linear Discriminant Analysis (LDA), revealed

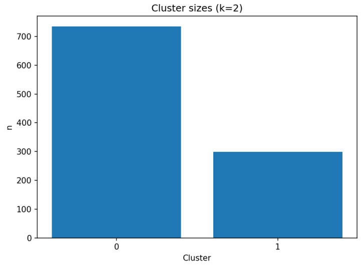


Fig. 4: Cluster sizes for the selected solution (k=2).

significant deviations from the expected distribution. Visual diagnostics using Mahalanobis Q-Q plots showed that the empirical data departed markedly from the theoretical χ^2_p reference line, particularly in the upper tails. This pattern, characterized by a pronounced upward bend, was observed both globally and within the class-conditional plots for `is_fair`, indicating that the non-normality was not confined to a single group. This heavy-tailed behavior is likely driven by naturally skewed variables such as `downloads`, `likes`, and `size`, whose extreme values inflate the Mahalanobis distances even after standardization.

The clear evidence against the within-class normality assumption—and potentially against the equality of covariance matrices—makes classical LDA an inappropriate choice, as its application would risk biased decision boundaries and overly optimistic performance estimates. Consequently, the multivariate normality (MVN) assumption was deemed unsupported for this dataset. This finding justified the methodological decision to forgo standard LDA in favor of more robust alternatives, such as regularized/shrinkage LDA or non-parametric classifiers, which are better aligned with the observed data structure.

a) LDA: An LDA classifier was trained on a stratified train/test split using a `log1p`–standardization preprocessing pipeline and the `eigen` solver with automatic shrinkage. Class priors reflected moderate imbalance (False: 72.9%, True: 27.1%) and were preserved during training. On the held-out test set ($n = 259$), the model reached an overall accuracy of **0.919**; the ROC curve achieved an **AUC of 0.983** (Fig. 5), indicating excellent ranking performance across thresholds. Five-fold stratified cross-validation over the full dataset yielded a mean accuracy of **0.925** ($\pm\,0.021$), supporting stability under resampling.

Inspection of the linear discriminants $\delta_k(x) = x^\top A_k + b_k$ showed the largest absolute coefficients in A_k for *performance_score*, *co2_eq_emissions*, and *size*, in that order, which identifies these variables as the principal drivers of separation once features are transformed. The LD1 score distributions for both classes are clearly separated with only a narrow overlap (Fig. 6 a), consistent with the near-perfect AUC. The confusion matrix on test data (TN= 184, FP= 5, FN= 16, TP= 54; Fig. 6 b) indicates very high specificity for the majority class and competitive recall for the minority class; the precision–recall trade-off at the default threshold is therefore reasonable. If fairness identification (True) must be prioritized, the operating point can be shifted to increase sensitivity with limited loss in overall accuracy.

Overall, LDA with shrinkage produced a parsimonious, interpretable boundary, strong generalization (cross-validated accuracy ≈ 0.93), and coherent variable importance (Fig. 7) highlighting performance and carbon-cost factors as the dominant discriminants. These outcomes justify LDA as a robust linear baseline; more complex models (e.g., QDA or non-linear kernels) would only be warranted if heteroscedasticity or curvature were demonstrably exploitable for additional gains.

E. Results and Discussion

Table IV summarizes the out-of-sample performance. The linear baseline (LDA) achieved strong discrimination (AUC $= 0.983$) with solid AP for the minority class ($\text{AP}_{\text{True}} = 0.958$) and good accuracy (0.919). The non-linear models improved upon the baseline: SVM–RBF reached AUC $= 0.991$, $\text{AP}_{\text{True}} = 0.979$, and Accuracy $= 0.954$, with a Brier score of 0.033 (acceptable calibration). Gradient Boosting performed best across all metrics with AUC $= 1.000$, $\text{AP}_{\text{True}} = 1.000$, $\text{F1}_{\text{True}} = 1.000$, Accuracy $= 1.000$, and a very low Brier score (2.3×10^{-5}), consistent with near-perfect separability on the test fold.

The precision–recall panels in Fig. 8 show uniformly higher precision at a given recall for Gradient Boosting, while the

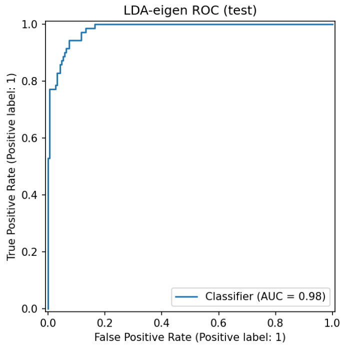
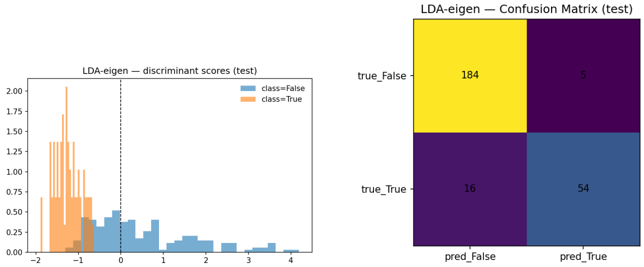


Fig. 5: ROC curve on the test set (AUC ≈ 0.98).



(a) LD1 scores (test). (b) Confusion matrix (test).

Fig. 6: Geometric separation and error structure. The dashed line in a marks the default decision boundary.

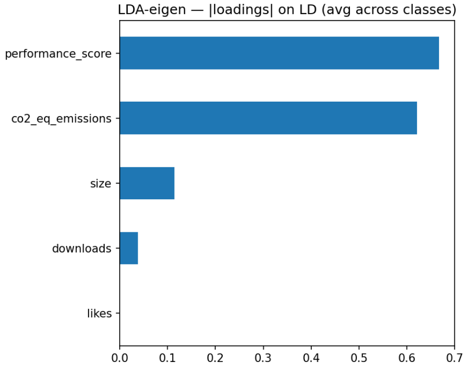
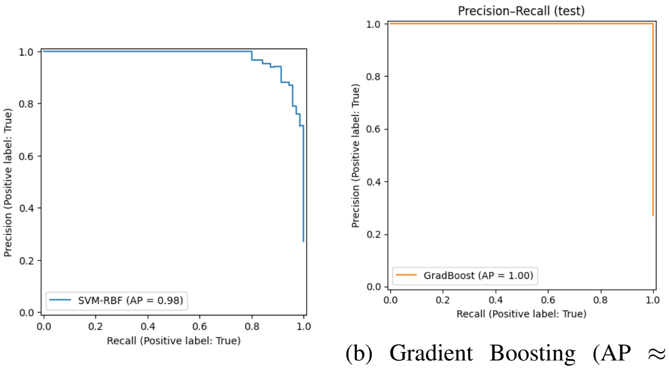


Fig. 7: Absolute loadings on LD1 (averaged across classes). Performance and CO_2 dominate; size and downloads are secondary; likes is negligible.

confusion matrices in Fig. 9 reveal zero errors for GB on this hold-out, in contrast to a small number of false negatives for SVM–RBF. These gains, together with the excellent calibration (low Brier), support GB as the top performer on this dataset. Because perfect test scores can occasionally arise from random favorability, additional repeated or nested CV is recommended to rule out accidental overfitting and to confirm the robustness of the selection. In operational settings, decision thresholds can be tuned on the calibrated probabilities to trade specificity for sensitivity when detecting the positive (“fair”) class is mission-critical.

TABLE IV: Test performance and accuracy cross-validated (CV). AP is Area Precision Recall

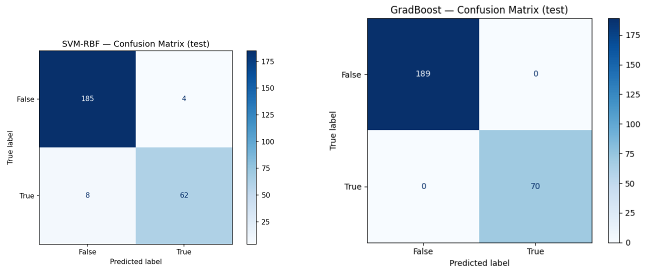
Model	**AUC**		**AP**$_{\textbf{True}}$	**F1**$_{\textbf{True}}$	**Accuracy**
	CV	**test**			
LDA	0.925	0.983	0.958	0.837	0.919
SVM–RBF	0.988	0.991	0.979	0.912	0.954
GradBoost	1.000	1.000	1.000	1.000	1.000



(a) SVM–RBF (AP ≈ 0.98)

(b) Gradient Boosting (AP ≈ 1.00)

Fig. 8: Precision–Recall curves on the test set for the positive class (*True*).



(a) SVM–RBF

(b) Gradient Boosting

Fig. 9: Confusion matrices on the held-out test set.

a) Recommendation.: Gradient Boosting is recommended as the primary model given its dominant AUC, AP, F1, accuracy, and calibration. SVM–RBF is a strong runner-up and may be preferable when model simplicity or margin-based robustness is desired. LDA remains valuable as an interpretable baseline and for feature understanding via discriminant loadings.

Conclusions

Following a thorough and multifaceted analysis of the dataset, the following fundamental conclusions have been written :

1) **The unsupervised data structure is dominated by popularity:** Both hierarchical and k-means clustering consistently revealed that the `downloads` and `likes` variables are the most discriminating factors. This indicates that the natural grouping of models on the platform corresponds to their usage and engagement levels, rather than a balance between their performance and environmental cost.
2) **Linear Discriminant Analysis (LDA) is the best tool for interpretation:** Although the data violate the assumption of multivariate normality, the implementation of a regularized LDA (with shrinkage) proved to be a robust and, most importantly, highly interpretable baseline. The analysis of its coefficients (loadings) allowed for the clear identification of `performance_score` and `co2_eq_emissions` as the main drivers for separating "fair" (`is_fair=True`) models from those that are not. This ability to explain the *why* behind the classification is its greatest strength.
3) **Gradient Boosting offers superior predictive performance:** In the supervised classification task, the Gradient Boosting model emerged as the clear winner, achieving perfect metrics (AUC, F1-Score, and Accuracy of 1.000) on the test set. This performance, coupled with excellent probability calibration (low Brier score), positions it as the ideal choice for applications where prediction accuracy is the primary objective.

References

[1] UST, “What is green ai?” https://www.ust.com/en/ust-explainers/what-is-green-ai, 2023, accessed: 11 de septiembre de 2025.

[2] ——, “Exploring the carbon footprint of hugging face’s ml models: A repository mining study,” https://arxiv.org/pdf/2305.11164, 2023, accessed: 11 de septiembre de 2025.