Exploring Multi-Task Learning for Fairness in Machine Learning Regression

Bruno Pires[1], **Luiz Leduino**[1], **Lilian Berton**[1]

[1]Instituto de Ciência e Tecnologia – Universidade Federal de São Paulo (UNIFESP)

{bruno.pires22,lberton}@unifesp.br

Abstract. *Ensuring fairness in machine learning is a critical concern for high-stakes domains, yet most fairness-aware Multi-Task Learning (MTL) frameworks overlook regression problems in favor of classification. This work extends these techniques to regression, proposing a novel MTL framework that optimizes for equitable continuous outcomes across demographic subgroups. Our method dynamically reweights task-specific gradients during training to reduce disparities without compromising predictive accuracy. Evaluated on two real-world datasets, our approach, in the best scenarios, reduces subgroup disparity by up to 94.9% while also improving overall regression performance by up to 32.6%. These findings highlight the significant potential of fairness-aware MTL for creating more inclusive and responsible machine learning applications in sensitive domains.*

1. Introduction

Machine learning (ML) systems have become integral to decision-making pipelines across a wide range of domains, including healthcare, education, finance, and criminal justice. However, concerns regarding algorithmic *fairness* have grown with their widespread deployment. It is now well-understood that ML models can reflect and amplify existing biases present in the training data, resulting in disparate impacts for underrepresented or protected groups [Barocas et al. 2023, Mehrabi et al. 2021].

Bias in ML models may arise not only from skewed data distributions but also from algorithmic design choices, especially in tasks where sensitive attributes, such as race, gender, or age, are correlated with the prediction target. Traditional fairness-aware approaches often involve pre-processing data to mitigate bias, imposing fairness constraints during training, or post-processing model predictions [Hardt et al. 2016]. However, these strategies may fall short when applied to more complex learning paradigms, such as *multi-task learning*.

MTL is a framework in which multiple learning tasks are solved jointly, enabling models to leverage commonalities and differences across tasks [Caruana 1997]. While MTL has demonstrated improvements in generalization and data efficiency, it has recently gained attention in fairness research. Several studies have proposed fairness-aware MTL frameworks to mitigate bias across sensitive subgroups or tasks. For example, [Oneto et al. 2019] introduced fairness constraints into MTL to train group-specific classifiers, while maintaining independence from protected attributes at test time. [Li et al. 2023] proposed dynamic gradient reweighting during backpropagation to balance group fairness without sacrificing performance.

Despite these advances, existing fairness-aware MTL approaches have primarily focused on classification problems. In many real-world applications, such as predicting continuous outcomes in clinical or economic contexts, regression tasks are of paramount importance. However, fairness in MTL-based regression remains underexplored in the literature.

The main contribution of this work is to fill this gap by employing Multi-Task Learning in fairness-aware *regression tasks*. We demonstrate that our approach effectively mitigates bias across continuous outputs while preserving predictive performance. Our contributions are summarized as follows:

- We propose a fairness-aware MTL framework designed specifically for regression tasks, which generalizes existing classification-focused strategies.
- We introduce a dynamic task reweighting mechanism that reduces subgroup disparities without degrading accuracy.
- We evaluate our method on real-world datasets and show that it outperforms traditional fairness baselines on both fairness and performance metrics.

2. Related work

Recent studies highlight that bias in machine learning arises not only from data but also from algorithmic design choices, and traditional fairness techniques often fail to address the nuanced influence of protected attributes. Some papers have been exploring multi-task learning (MTL) to address fairness.

One proposed solution introduces a framework that combines MTL with Monte Carlo (MC) Dropout to assess and mitigate model uncertainty associated with protected labels [Zanna and Sano 2024]. Another study analyzes the impact of bias present in datasets on the fairness of deep learning model predictions, focusing on skin lesion classification using ResNet-based convolutional neural networks [Raumanns et al. 2024]. The research considers variations in patient sex within the training data and compares three learning approaches: single-task model, multitask model with reinforcement, and adversarial learning. A linear programming technique was used to create datasets with different distributions of patient sex and lesion classes.

The approach of [Oneto et al. 2019] combines MTL which trains group-specific models, with fairness constraints. To ensure fairness even when group-specific modeling is not allowed, they suggest first predicting the sensitive feature and then using this predicted value to guide MTL training. [Li et al. 2023] introduce dynamic gradient reweighting method during neural network training. This strategy automatically adjusts the influence of each subgroup in backpropagation, promoting fairness among tasks (subgroups) without compromising model generalization. The method was tested in a real-world scenario for mortality risk prediction in sepsis patients and demonstrated up to a 98% reduction in subgroup disparity, with less than a 4% loss in predictive accuracy.

[Li et al. 2022] introduces FairSR, a deep learning-based model that integrates sequential recommendation (SR) with algorithmic fairness. SR predicts future user-item interactions based on temporal dynamics, while fairness-aware recommendation mitigates biases in preference learning. They used MTL to enhance both recommendation accuracy and fairness in sequential recommendation systems. [Roy and Ntoutsi 2022] introduces L2T-FMT algorithm, a teacher-student network trained collaboratively. The student

is tasked with solving the fair MTL problem, while the teacher dynamically guides the learning process, emphasizing either accuracy or fairness based on which aspect is more challenging for each task. This adaptive selection of learning objectives at each step optimizes the trade-off process, reducing the number of required trade-off weights from 2T to T, where T represents the number of tasks.

Traditional fairness optimization approaches may fail in multi-task scenarios, so [Wang et al. 2021] proposes new metrics to better capture fairness-accuracy trade-offs in this setting. Additionally, the authors introduce Multi-Task-Aware Fairness (MTA-F), a method designed to enhance fairness in MTL while maintaining predictive performance.

While several of the referenced studies primarily concentrate on classification tasks, the scope of MTL can be expanded to encompass regression-based approaches, which align with the core objective of this work.

3. Methodology

This work proposes a fairness-aware Multi-Task Learning approach based on a Multi-Layer Perceptron (MLP) architecture. We address a regression task with potential bias by introducing two prediction heads: one for estimating the protected attribute (e.g., gender, race), and another for the target variable of interest. The learning process encourages shared representations that are informative for the primary task while accounting for fairness across the protected attribute.

3.1. Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i, z_i)\}_{i=1}^{N}$ be a dataset where $x_i \in R^d$ are the input features, $y_i \in R$ is the continuous target for the main task, and $z_i \in \{0, 1\}$ is the protected attribute (e.g., gender). The goal is to predict y accurately while ensuring fairness with respect to z.

3.2. Datasets

To evaluate the effectiveness and generalizability of the proposed approach, we apply it to two distinct regression tasks with different domains and fairness concerns: the Boston Housing dataset and the Parkinson Telemonitoring dataset. In both cases, the objective is to accurately predict a continuous target while reducing disparities associated with a sensitive attribute.

1. **Boston Housing Dataset** - The Boston Housing dataset consists of housing-related features aimed at predicting the median house value (MEDV) in thousands of dollars for different neighborhoods. In this context, the protected attribute is the variable B, which represents the proportion of Black residents in a neighborhood, defined as:

$$B = 1000 \cdot (1 - \text{proportion of Black residents})^2$$

   This variable has been widely discussed in the fairness literature as a proxy for racial composition. The fairness concern here arises from the possibility that predictive errors may disproportionately affect neighborhoods based on their racial demographics. Our approach aims to mitigate this by balancing the prediction of MEDV with fairness constraints related to B.
   Overall, the dataset comprises 506 samples with 13 features describing housing and demographic characteristics of neighborhoods. The target variable, MEDV,

represents the median house value in thousands of dollars, ranging approximately from 5 to 50. Figure 1 illustrates the distribution of the protected attribute B, highlighting its skewness across the dataset. This uneven distribution underscores the importance of fairness-aware modeling, aiming to prevent prediction errors from disproportionately affecting certain demographic groups. Figure 2 summarizes

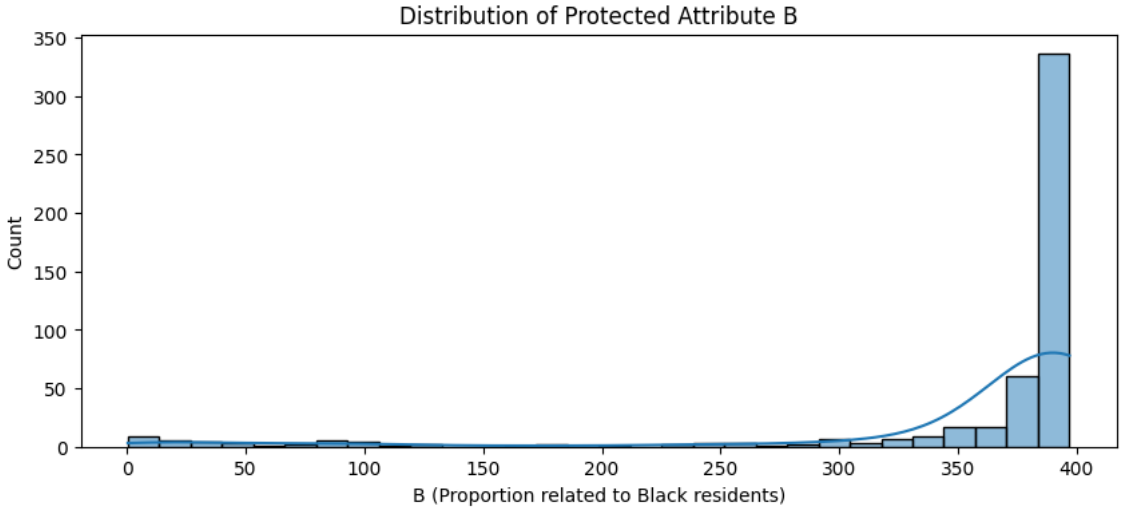
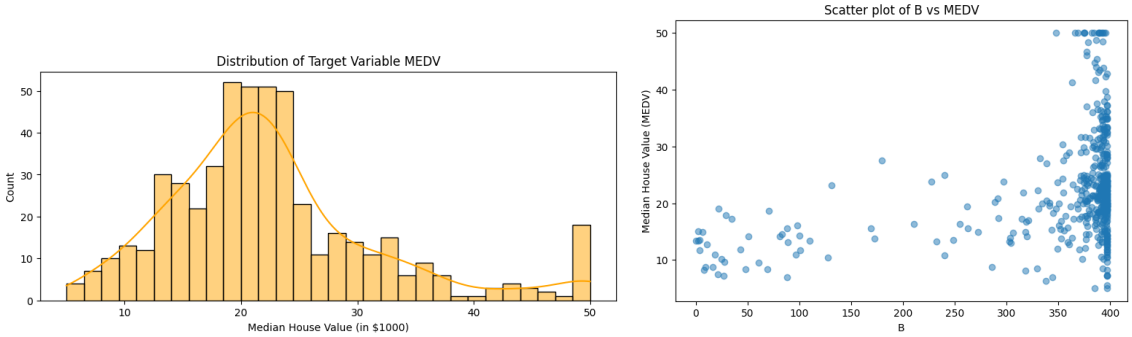


Figure 1. Distribution of the protected attribute B in the Boston Housing dataset.

the distribution of the target variable MEDV and the protected attribute B. The left panel shows that most house values are concentrated between $15,000 and $25,000. The right panel reveals a negative correlation between B and MEDV, indicating that neighborhoods with more Black residents tend to have lower median house values, highlighting a risk of unfairness if models reinforce this disparity.



(a) Distribution of the target variable MEDV. **(b) Scatter plot of B versus MEDV.**

Figure 2. Distribution of the target variable MEDV (left) and the relationship between the protected attribute B and MEDV (right) in the Boston Housing dataset.

2. **Parkinson Telemonitoring Dataset.**
The Parkinson Telemonitoring dataset includes 5,875 voice recordings from 42 patients, with the goal of predicting motor UPDRS scores. Gender is used as the protected attribute, with a noticeable imbalance (28 males, 14 females). Figure 3

shows a right-skewed distribution of UPDRS scores and highlights the gender imbalance, emphasizing the need for fairness-aware modeling to prevent biased prediction errors.

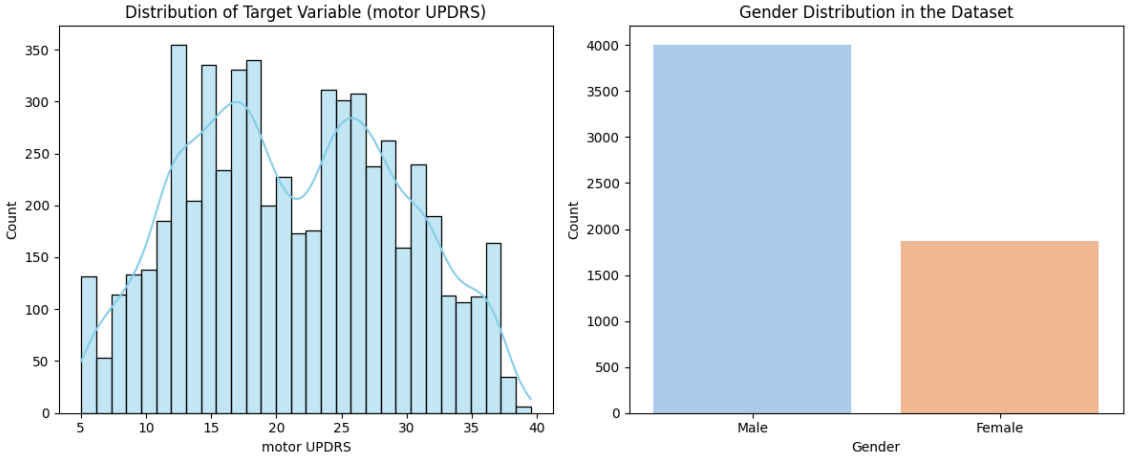


Figure 3. On the left, the histogram of the motor UPDRS target variable, which is moderately right-skewed. On the right, the gender distribution reveals an imbalance, with a larger proportion of male participants compared to female participants, for the Parkinson Telemonitoring Dataset.

In both datasets, the same modeling framework is employed: a predictive model trained to minimize the regression loss while incorporating a fairness regularization term. This setup allows us to investigate how the trade-off between accuracy and fairness behaves in different contexts, one socioeconomic and one biomedical.

3.3. MLP Multi-Task Architecture

The architecture consists of shared hidden layers followed by two task-specific heads:

- **Task 1**: Regression head for predicting the target y_i.
- **Task 2**: Head for predicting the protected attribute z_i. Classification in Parkinson Telemonitoring, regression in Boston Housing.

Given an input x_i, the shared layers produce a latent representation $h(x_i)$. The two outputs are then computed as:

$$\hat{y}_i = f_y(x_i) = W_y \cdot h(x_i) + b_y$$
$$\hat{z}_i = f_z(x_i) = \sigma(W_z \cdot h(x_i) + b_z)$$

where W_y, W_z, b_y, and b_z are the parameters of the respective task heads, and $\sigma(\cdot)$ is the sigmoid activation function used for binary classification.

3.4. Loss Function and Fairness-Aware Training

The total loss is a weighted combination of both task losses:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_y + (5 - \alpha) \cdot \mathcal{L}_z$$

- $\mathcal{L}_y$ is the Mean Squared Error (MSE) for regression tasks:

$$\mathcal{L}_y = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

- $\mathcal{L}_z$ is the Binary Cross-Entropy (BCE) loss for classification tasks:

$$\mathcal{L}_z = -\frac{1}{N}\sum_{i=1}^{N}[z_i \log(\hat{z}_i) + (1 - z_i)\log(1 - \hat{z}_i)]$$

The coefficient $\alpha \in [0, 5]$ balances between accuracy and fairness objectives. In practice, we perform hyperparameter tuning over α to optimize both fairness and predictive performance.

3.5. Pareto frontier

The Pareto frontier is the set of non-dominated solutions in a multi-objective optimization problem with conflicting objectives. A solution dominates another if it is no worse in all objectives and strictly better in at least one. Formally, for two objectives f_1 and f_2 we say that a solution dominates x_i dominates x_j if:

$$f_1(x_i) \leq f_1(x_j), \quad f_2(x_i) \leq f_2(x_j), \quad \text{and} \quad (f_1(x_i) < f_1(x_j) \text{ or } f_2(x_i) < f_2(x_j)).$$

The Pareto frontier consists of solutions for which no other solution dominates them. It represents the best possible trade-offs among the objectives, meaning that improving one objective necessarily worsens at least one other.

3.6. Training Procedure

We use the Adam optimizer with early stopping based on validation loss of the main task. Dropout and batch normalization are applied to reduce overfitting. Hyperparameters such as learning rate, hidden layer sizes, and α are selected using cross-validation. The training strategy for the Boston Housing dataset involves sequential training over many α values within blocks of 5, where the model is reset at the beginning of each block to avoid weight carryover across too many α configurations. This reset prevents excessive overfitting that might arise from long sequential fine-tuning on many different weighting parameters.

Moreover, the early stopping mechanism employed during training provides an additional safeguard against overfitting by monitoring the validation loss and terminating training once improvements become negligible.

In contrast, for Parkinson Telemonitoring, each α is trained independently from scratch for a fixed number of epochs without early stopping, ensuring a clean comparison across α values without implicit warm-start bias. Together, these practices help ensure fair and robust evaluation of the trade-off parameter α and the corresponding balance between fairness (auxiliary task) and regression performance (main task) in our multi-task framework.

3.7. Evaluation Metrics

We evaluate the model on both performance and fairness:

- **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** for the regression task.
- **Group Disparity**, defined as the difference in MAE across values of the protected attribute.
- **Fairness-Accuracy Trade-off**, which captures how the fairness-aware strategy affects overall prediction performance.

4. Results

In this section, we present the results obtained from the experiments conducted on the two datasets. The main objective is to evaluate disparities in predictions across sensitive groups, as well as their behavior as fairness-oriented approaches are applied, and finally to analyze the overall performance of the models, observing the trade-off that these methods entail. To this end, in all experiments we compare the original model, trained without any bias mitigation intervention, with the proposed approach, which incorporates mechanisms aimed at promoting fairness.

The evaluation was conducted considering two simultaneous tasks in a multi-task learning context, with the models using the modified loss function as previously described. The metrics used include mean absolute error and root mean squared error, both on the full dataset and stratified by sensitive groups defined in the dataset. This analysis allows quantifying not only the overall performance but also potential disparities between groups.

Additionally, after training models corresponding to 50 distinct values of the hyperparameter α, a Pareto frontier is constructed, which enables identifying the models that represent the best possible trade-offs between performance and fairness. This frontier serves as a basis to linearize the models on a continuous scale, ranging from exclusive prioritization of predictive performance to the maximization of equity among groups. In this way, it becomes possible to carefully select models that best meet the specific demands for balance between accuracy and fairness in each application context.

4.1. Boston Housing

The first dataset used was the Boston Housing dataset, with the multitask model performing two regression tasks, producing values for house prices (MEDV) and the proportion of Black people in the city (B). The model is composed of two shared dense layers with 64 and 32 neurons, respectively, both activated by ReLU functions. From this shared block, the model branches into two independent heads, each dedicated to one of the regression tasks, with no activation in the output layer.

Training was carried out using the Adam optimizer with a fixed learning rate of 0.001. To control overfitting, an early stopping strategy was implemented, with a patience limit of 10 epochs and a minimum improvement tolerance (δ) of 1×10^{-4}. The maximum number of epochs was set to 250, with a batch size of 32 samples.

In this first experiment, the model was trained in a standard way, that is, without any weighting in the loss function aimed at mitigating inequalities. The objective function

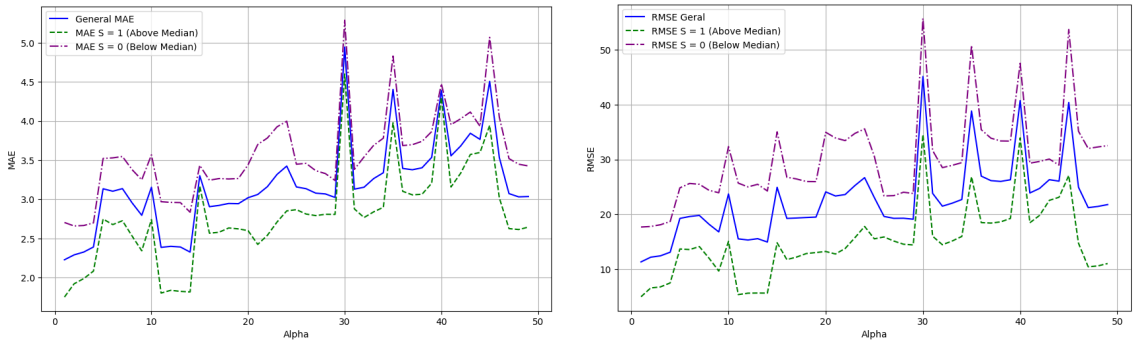
considered only the target variable (MEDV). This setup establishes a performance and inequality baseline that will later be compared with approaches that incorporate explicit fairness mechanisms.

For the purpose of fairness analysis, the test set was divided into two sensitive groups based on the median of variable B, which represents the proportion of Black residents in the localities. This approach was adopted to enable comparisons between distinct subgroups, avoiding the complexity of dealing directly with continuous sensitive variables. Thus, the model's performance was compared for localities with a proportion of Black population above the median versus those below it, in addition to evaluating the overall performance. The results obtained by this model, both in terms of overall performance and in metrics stratified by sensitive groups, are presented in Table 1 and will serve as a reference to analyze the impacts of the proposed interventions.

Metric	Value
General RMSE	11.45
General MAE	2.11
RMSE Disparity	13.76
MAE Disparity	0.96

Table 1. Boston Housing summary — no fairness interventions.

With the introduction of the fairness term in the loss function, the multitask model was trained for 150 epochs, allowing for an analysis of the impact of the hyperparameter α on the relationship between performance and fairness. The MAE and RMSE metrics were calculated for overall performance and also stratified by sensitive groups, defined based on the median of variable B. Figure 4 illustrates the errors across the models with different α values.



(a) MAE values across different values of α (fairness weighting).

(b) RMSE values across different values of α (fairness weighting).

Figure 4. Results for Boston Housing dataset.

At the end of the experiments, a Pareto frontier was constructed, gathering the best solutions in terms of the trade-off between predictive error and group disparity. This analysis makes it possible to visualize the potential trade-offs between fairness and performance, guiding the most appropriate choice of α according to the application's goals. Figure 5 and Table 2 present the corresponding results.

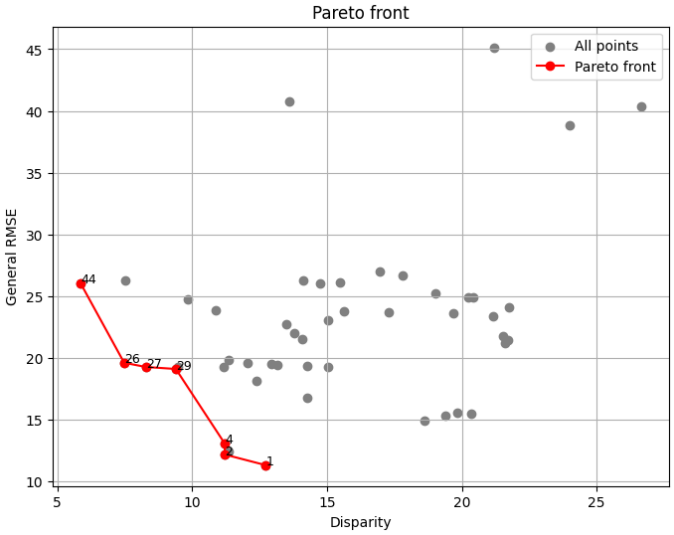


Figure 5. Pareto front with Boston Housing dataset.

α	RMSE Disparity	General RMSE
4.4	5.866693	26.037848
2.6	7.481788	19.600241
2.7	8.289066	19.255997
2.9	9.398005	19.096895
0.4	11.205297	13.074536
0.2	11.212414	12.165419
0.1	12.726402	11.306895

Table 2. Pareto front with Boston Housing dataset

Although the errors of the original model were not particularly high, a significant disparity between the defined groups was observed, greater than the overall RMSE itself, which was expected, given that this dataset is known to be biased. The implemented approach, however, revealed a useful and almost linear scale for mitigating this bias: by increasing the weight of fairness in the loss function (through α), the disparity between groups dropped to less than half of its original value, reaching a reduction of up to 68%. On the other hand, there was a considerable performance loss, with the overall error increasing by approximately 127%.

To minimize this impact, we identified a model with an α value very close to zero, in which the disparity decreased by only 8%, while performance slightly improved (approximately 1.3%). This behavior suggests that the fairness term introduced during training was beneficial in both aspects, as it enabled the discovery of solutions that outperform the baseline model. Overall, it can be observed that the hyperparameter α acts in an almost linear manner, and the construction of the Pareto frontier provides an adjustable scale for selecting the best values according to specific fairness and performance goals.

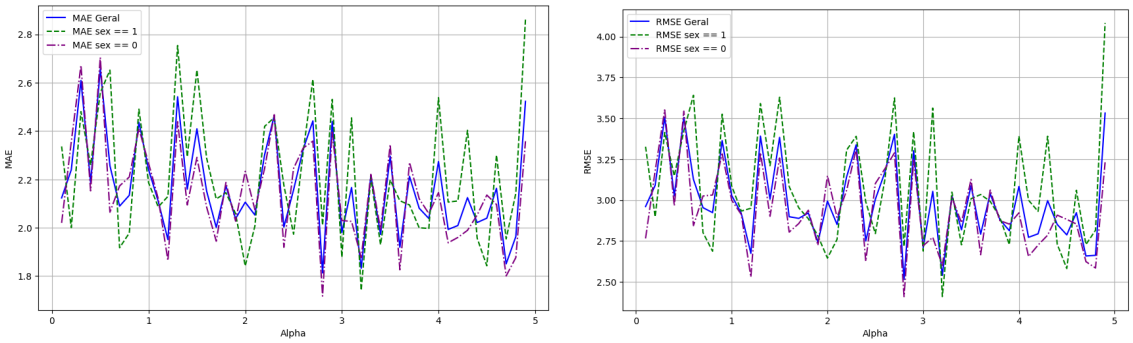
4.2. Parkinson Telemonitoring

In the second dataset, related to the problem of Parkinson's disease telemonitoring, the same procedure described earlier was adopted. A multitask model was used, consisting of two shared dense layers (with 64 and 32 neurons), followed by two outputs: one for regression and another for classification. Training was conducted for 50 epochs using the Adam optimizer with a learning rate of 0.01, and without applying early stopping. As in the first experiment, the tests included varying the hyperparameter α in order to analyze the impact of task weighting on performance and fairness of the predictions. The MAE and RMSE metrics were used to evaluate both overall performance and disparities between sensitive groups. Finally, the construction of the Pareto frontier allowed for the identification of optimal α values aimed at fairness and/or predictive performance. Table 3 shows the results without fairness modifications to the model.

Metric	Value
General RMSE	3.3341
General MAE	2.4486
RMSE Disparity	0.39
MAE Disparity	0.30

Table 3. Parkinson's Telemonitoring summary — no fairness interventions.

Figure 6 presents the results of applying fairness to the Parkinson Telemonitoring dataset. MAE and RMSE values were reported for overall performance as well as stratified by sensitive groups, across varying values of α.

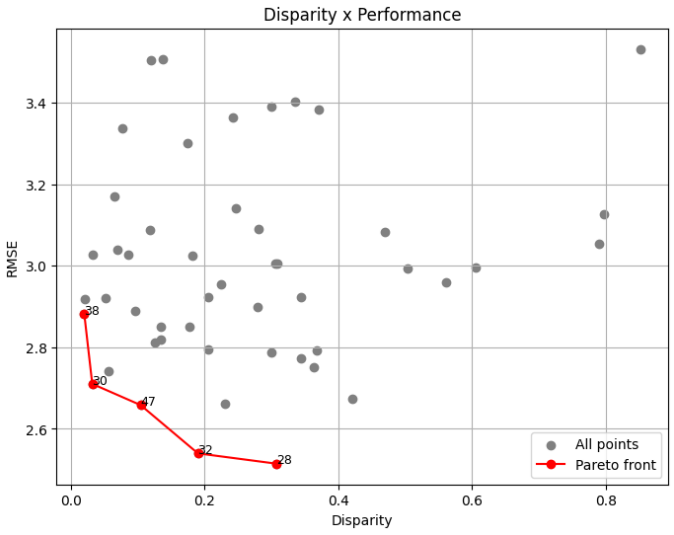


(a) MAE values across different values of α (fairness weighting).

(b) RMSE values across different values of α (fairness weighting).

Figure 6. Results for Parkinson Telemonitoring dataset.

Similarly to the previous experiment, the models generated were employed to build the Pareto frontier (see Figure 7 and Table 4), emphasizing the optimal trade-offs between predictive performance and fairness.



α	RMSE Disparity	General RMSE
3.8	0.020463	2.881697
3.0	0.032371	2.710295
4.7	0.104597	2.658332
3.2	0.190242	2.539749
2.8	0.307519	2.514439

Table 4. Pareto front with Parkinson Telemonitoring dataset

Figure 7. Pareto front with Parkinson Telemonitoring dataset.

In the second dataset, both the initial errors and the observed disparity were low, leading to a different scenario for evaluating the impact of fairness measures. Nevertheless, the results across different α values showed significant variation, indicating room for adjustment and the potential to use more robust models to fine-tune these parameters.

The most noteworthy point, however, lies in the fact that, in addition to a substantial reduction in group disparity, reaching up to 94.9%, there was also an improvement in the model's overall error, of 15.6%. This reveals a "double gain" scenario: the newly introduced fairness term in training not only promoted greater equity but also increased predictive accuracy.

This result reinforces an important premise in fairness studies: it is not enough to include a sensitive variable that seemingly improves model performance if it does so at the cost of increasing inequalities, perpetuating bias, and reproducing discriminatory behavior, as seen in the Boston Housing case. However, when such a variable is appropriately handled, with mitigation mechanisms and evidence that disparities are in fact being reduced, its use becomes justified, as it contributes to a model that is both fairer and more efficient.

As in the previous experiments, the hyperparameter α functioned as an almost linear scale to balance fairness and predictive performance. Although the lowest global errors were still associated with higher disparities, all outcomes obtained with interventions outperformed the initial results, in which no fairness measures were applied.

5. Conclusion

This work presented a multitask learning approach to incorporate fairness criteria into supervised models, with particular emphasis on regression problems, a domain still underexplored in the algorithmic fairness literature, which has traditionally focused on classification tasks. By treating the sensitive variable as an auxiliary task during training, we employed a hyperparameter α to control its influence on the loss function, enabling a continuous balance between predictive performance and equity.

Experiments were conducted on two datasets with complementary characteristics. In the Boston Housing dataset, known for its structural bias, the original model already exhibited significant group disparities despite strong overall performance. With the introduction of the sensitive task, we were able to reduce this disparity by up to 68%, even at some cost to performance. In more conservative cases, we achieved solutions that reduced inequality with minimal impact in overall error, demonstrating that the approach can produce models that are both fairer and more efficient.

In the Parkinson's Telemonitoring dataset, where initial disparity was low, tests still showed that the fairness term could contribute meaningfully. Disparity reduction reached over 94%, and in many α values we also observed performance gains. This suggests that, beyond regulating fairness, the new term may also incorporate useful information into the main regression task.

A central aspect of the results was the construction of Pareto frontiers that illustrated how α functions as a trade-off scale between equity and performance. This linear and adjustable behavior allows solutions to be adapted according to application-specific priorities, promoting more ethical models without major sacrifices in accuracy.

As future work, we propose testing the approach in other domains, particularly in sensitive areas such as public health, where bias can affect critical decisions. We also envision improvements to the loss function formulation, exploring strategies such as dynamic weighting or robust optimization, with the goal of increasing the effectiveness and applicability of the proposed method.

6. Ackowledgement

We thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant 2021/14725-3.

References

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28:41–75.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Li, C., Ding, S., Zou, N., Hu, X., Jiang, X., and Zhang, K. (2023). Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modeling. *Journal of biomedical informatics*, 143:104399.

Li, C.-T., Hsu, C., and Zhang, Y. (2022). Fairsr: Fairness-aware sequential recommendation through multi-task learning with preference graph embeddings. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(1):1–21.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Oneto, L., Doninini, M., Elders, A., and Pontil, M. (2019). Taking advantage of multitask learning for fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 227–237.

Raumanns, R., Schouten, G., Pluim, J. P., and Cheplygina, V. (2024). Dataset distribution impacts model fairness: Single vs. multi-task learning. In *MICCAI Workshop on Fairness of AI in Medical Imaging*, pages 14–23. Springer.

Roy, A. and Ntoutsi, E. (2022). Learning to teach fairness-aware deep multi-task learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 710–726. Springer.

Wang, Y., Wang, X., Beutel, A., Prost, F., Chen, J., and Chi, E. H. (2021). Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1748–1757.

Zanna, K. and Sano, A. (2024). Enhancing fairness and performance in machine learning models: A multi-task learning approach with monte-carlo dropout and pareto optimality. *arXiv preprint arXiv:2404.08230*.