Titre: Title:	Pruning Strategies in Random Forests: The Interplay Between Model Compression and Fairness
Auteur: Author:	Elaheh Rahmati
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Rahmati, E. (2025). Pruning Strategies in Random Forests: The Interplay Between Model Compression and Fairness [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/65863/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/65863/
Directeurs de recherche: Advisors:	Thibaut Vidal, & Golnoosh Farnadi
Programme: Program:	Maîtrise recherche génie industriel

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Pruning Strategies in Random Forests: The Interplay Between Model Compression and Fairness

ELAHEH RAHMATI
Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie industriel

Avril 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Pruning Strategies in Random Forests: The Interplay Between Model Compression and Fairness

présenté par **Elaheh RAHMATI**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Thibaut VIDAL, membre et directeur de recherche
Golnoosh FARNADI, membre et codirectrice de recherche
Louis-Martin ROUSSEAU, président
Ulrich AÏVODJI, membre

DEDICATION

To my supportive father, my nurturing mother, my lovely sister, and my sister-like friend,
Though distance separates us, your love and support have always made me feel as if you were right beside me.
Your unwavering encouragement and belief in me have been my constant source of strength.
I love you more than words can express. . . .

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Prof. Thibaut Vidal, for his unwavering support and guidance throughout my studies. Over the past two and a half years, I have faced numerous challenges, and he has always been there to provide encouragement, insightful advice, and invaluable mentorship. His patience and expertise have been instrumental in shaping my research and academic growth.

I am also immensely grateful to my co-supervisor, Prof. Golnoosh Farnadi, for her significant contributions to this project. Her guidance, constructive feedback, and constant support have played a crucial role in my progress. I truly appreciate the time and effort she dedicated to helping me navigate through the complexities of this research.

Beyond academia, I owe a heartfelt thank you to my family. Despite being thousands of kilometers away, their unwavering love, encouragement, and belief in me have been my greatest source of strength. Their support has been fundamental in allowing me to pursue my dreams, and for that, I am forever grateful.

Finally, I extend my appreciation to my friends, colleagues, and everyone who has supported me in this journey, whether through thought-provoking discussions, moments of laughter, or simply being there when I needed it most. This achievement would not have been possible without all of you.

RÉSUMÉ

Ce mémoire présente une étude approfondie de l'impact des méthodes d'élagage appliquées aux forêts aléatoires sur le compromis entre la performance prédictive et les mesures d'équité. Dans un contexte où l'équité algorithmique devient une préoccupation majeure, il est essentiel de comprendre comment les techniques de compression des modèles influencent ces métriques de manière multidimensionnelle. Cette recherche vise à explorer systématiquement ces effets selon plusieurs dimensions – les caractéristiques des jeux de données, les méthodes d'élagage utilisées et les niveaux d'élagage appliqués – dans le but de fournir des connaissances fondamentales et des recommandations exploitables pour la communauté scientifique.

L'objectif principal de cette étude est de déterminer dans quelle mesure les méthodes structurées d'élagage interagissent avec les biais initiaux des jeux de données et modifient les relations entre la précision des modèles et trois métriques d'équité clés : la différence de parité démographique, la différence d'égalité des chances et la différence de parité prédictive. L'étude examine également si certaines configurations d'élagage peuvent offrir des points d'équilibre stables permettant de préserver à la fois la performance prédictive et l'équité.

Pour répondre à ces objectifs, nous avons mené des expériences sur deux jeux de données distincts — le *jeu de données sur les revenus* et le *jeu de données sur le temps de déplacement* — chacun présentant des niveaux de stabilité et des structures de biais différents. Quatre méthodes d'élagage structurées (élagage par classement, élagage par regroupement, élagage MIQP et élagage glouton) ont été étudiées, en comparaison avec une méthode d'élagage aléatoire servant de base de référence. Chaque méthode a été appliquée à différents pourcentages d'élagage et les modèles résultants ont été évalués en termes d'exactitude et d'équité.

Notre méthodologie expérimentale suit une approche multidimensionnelle :

- Une première analyse globale regroupe les résultats à travers tous les niveaux d'élagage et les jeux de données afin d'identifier les tendances générales.
- Une analyse plus détaillée est ensuite réalisée en intégrant les dimensions du jeu de données et du niveau d'élagage, à l'aide de visualisations telles que des diagrammes en boîte, des nuages de points, des diagrammes radar et des matrices de corrélation.
- Enfin, chaque méthode est examinée individuellement afin d'évaluer sa sensibilité à l'élagage, son impact sur les relations entre métriques d'équité et sa robustesse dans différents scénarios.

Les résultats montrent que l'effet de l'élagage sur l'équité et l'exactitude n'est pas universel, mais fortement dépendant des propriétés du jeu de données et de la méthode utilisée. L'élagage par classement apparaît comme le plus sensible, provoquant de fortes variations métriques, notamment sur le jeu de données *travel time*. À l'inverse, l'élagage par regroupement démontre la plus grande stabilité et prévisibilité. Bien que les méthodes d'élagage glouton et MIQP montrent une certaine stabilité, elles peuvent, à des niveaux d'élagage forts, modifier de manière imprévisible les relations entre les métriques d'équité, en particulier sur les jeux de données sensibles. La magnitude de ces changements reste toutefois contenue dans des limites acceptables.

Un constat clé de ce travail est que la plupart des méthodes d'élagage ne changent pas la direction des biais initiaux présents dans les jeux de données, mais tendent plutôt à les amplifier ou à les atténuer. De plus, cette étude confirme que le compromis entre l'exactitude et l'équité n'est pas universel et varie selon la méthode, le jeu de données et le niveau d'élagage.

Ce mémoire se termine par des recommandations pratiques à l'intention des chercheurs et des praticiens. Le choix de la méthode d'élagage doit être guidé par la sensibilité du jeu de données aux perturbations d'équité, la stabilité des relations métriques sous l'élagage, ainsi que les objectifs d'équité propres à la tâche considérée.un élagage trop fort est déconseillé pour les jeux de données instables, sauf si des méthodes stables sont rigoureusement sélectionnées. Il est également recommandé de mener des analyses diagnostiques spécifiques avant toute application pratique.

Bien que cette étude porte sur deux jeux de données et trois métriques d'équité, elle constitue un socle d'analyse structuré et ouvre la voie à des travaux futurs plus étendus. Des recherches complémentaires pourraient intégrer d'autres jeux de données, des modèles d'ensemble alternatifs et des mesures de complexité computationnelle pour identifier des configurations optimales. Ce travail contribue ainsi à une meilleure compréhension des impacts de l'élagage sur l'équité dans les forêts aléatoires et fournit un cadre de référence pour les recherches futures.

ABSTRACT

This thesis presents a comprehensive study on the impact of random forest pruning methods on the fairness–accuracy trade-offs in predictive models. With growing concerns about algorithmic fairness, understanding how post-training model compression techniques such as pruning affect fairness metrics has become increasingly critical. This work aims to systematically explore these effects across multiple dimensions —dataset characteristics, pruning methods, and pruning levels— and to provide foundational insights that can guide future research and practical applications.

The core objective of this research is to determine how structured pruning methods interact with inherent dataset biases and influence the relationships between accuracy and three key fairness metrics: Demographic Parity Difference, Equalized Odds Difference, and Predictive Parity Difference. The study also investigates whether certain pruning configurations can serve as stable trade-off points, preserving fairness and predictive performance.

To address these objectives, experiments are conducted using two distinct datasets — the *income dataset* and the *travel time dataset* — which differ in terms of data variability and underlying bias characteristics. Four structured pruning methods (Rank Pruning, Cluster Pruning, MIQP Pruning, and Greedy Pruning) are analyzed alongside a random pruning baseline. Each method is applied across a range of pruning percentages, and the resulting models are evaluated with respect to both predictive accuracy and fairness metrics.

The experimental methodology consists of a multi-dimensional approach:

- First, high-level analyses aggregate results across datasets and pruning levels to capture overall trends.
- Subsequently, a more detailed analysis incorporates dataset and pruning-level dependencies, with visual explorations through boxplots, scatter plots, spider charts, and correlation heatmaps.
- Finally, each method is evaluated individually to assess its sensitivity to pruning, its effect on fairness metric interrelationships, and its robustness in different scenarios.

Our results demonstrate that the impact of pruning on fairness and accuracy is not universal; it is highly dependent on the dataset properties and the chosen pruning method. Rank Pruning is shown to be the most sensitive, with large metric shifts, particularly in unstable datasets such as the *travel time dataset.* In contrast, Cluster Pruning demonstrates the greatest overall stability and predictability across different datasets and pruning levels. Although Greedy Pruning and MIQP Pruning show relative stability, at high pruning levels, they can unpredictably alter the interrelationships between fairness metrics, especially in more sensitive datasets. While the magnitude of these shifts remains within acceptable ranges, their direction can vary depending on dataset characteristics, making these methods less predictable under extreme conditions.

A key finding of this thesis is that while pruning methods do not fundamentally change the direction of the bias inherent in the dataset, they tend to either amplify or attenuate its magnitude, depending on the method and the degree of pruning applied. Additionally, we confirm that the fairness–accuracy trade-off is not universal. While predictive parity can sometimes improve with increased pruning, other fairness metrics may degrade, especially at higher pruning levels.

The thesis concludes with practical recommendations for researchers and practitioners. The selection of pruning methods should be guided by the dataset's sensitivity to fairness shifts, the stability of metric relationships under pruning, and the specific fairness objectives of the task. Extreme pruning is discouraged for highly sensitive datasets unless methods with proven stability are used. Moreover, practitioners are encouraged to conduct thorough scenario-specific diagnostics before applying pruning strategies.

Although this study focuses on two datasets and three fairness metrics, it lays the groundwork for broader investigations. Future research can expand this multi-dimensional framework to include additional datasets, alternative ensemble models, and computational complexity measures to identify practical sweet spots. Overall, this thesis contributes essential reference insights for understanding and managing the fairness implications of pruning in random forests.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

DPD	Demographic Parity Difference
EOD	Equalized Odds Difference
PPD	Predictive Parity Difference

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

1.1 Context and Motivation

Ensemble methods, particularly random forests, have become essential tools in modern machine learning due to their high predictive accuracy, robustness, and ability to handle complex data structures. However, their large and complex architectures can lead to inefficiencies, prompting the development of ensemble pruning techniques. The goal of pruning is to reduce model complexity while maintaining — or potentially improving — predictive accuracy.

In recent years, research has extended beyond accuracy preservation toward fairness-aware pruning methods, reflecting growing concerns about bias and equity in algorithmic decision-making. While some pruning techniques focus purely on accuracy, others seek to preserve or even improve fairness metrics, and a few attempt to balance both objectives. However, despite these advancements, a significant gap persists in the literature: there is no systematic, comprehensive analysis comparing different pruning methods in terms of their trade-offs between fairness and accuracy, particularly for traditional machine learning models such as random forests.

Existing studies often address accuracy and fairness separately or in narrowly defined scenarios. Moreover, fairness-aware pruning methods are frequently designed for specific applications or datasets, with little focus on generalizability or broader methodological comparison. This lack of a unified understanding limits the ability of researchers and practitioners to make informed decisions about which pruning methods to use in fairness-critical applications.

1.2 Research Questions and Objectives

This thesis aims to fill this gap by addressing the following research questions:

- How do different random forest pruning methods impact the trade-off between predictive accuracy and multiple fairness metrics?
- To what extent are these trade-offs influenced by dataset characteristics, such as inherent bias and stability?
- How do these effects vary with different levels of pruning intensity?
- Can stable and generalizable recommendations or best practices be derived for applying pruning methods in fairness-sensitive scenarios?

To answer these questions, the main objectives of this research are:

- To conduct a systematic, multi-dimensional comparative analysis of structured random forest pruning techniques.
- To evaluate the behavior of these methods across multiple datasets and fairness metrics.
- To investigate how pruning intensity influences fairness–accuracy dynamics.
- To identify patterns and propose practical guidelines for selecting pruning strategies based on dataset sensitivity and fairness priorities.

1.3 Contributions

This thesis provides the first thorough comparative study of a selection of structured pruning methods for random forests, analyzing their effects on accuracy as well as three key fairness metrics: Demographic Parity Difference, Equalized Odds Difference, and Predictive Parity Difference. The analysis is conducted across two datasets with distinct characteristics, namely the income dataset and the travel time dataset, enabling an exploration of how dataset stability and bias influence pruning behavior. The study incorporates multiple experimental layers, examining results from an aggregated perspective down to method- and dataset-specific interactions, and includes a detailed examination of pruning-level effects. Through extensive visualization and correlation analysis, this thesis identifies critical patterns, trade-offs, and stability profiles associated with each pruning method. Finally, it offers actionable recommendations and best practices for practitioners aiming to balance accuracy and fairness in fairness-sensitive machine learning applications.

1.4 Thesis Outline

The remainder of this thesis is structured as follows:

Chapter 2. presents a comprehensive literature review, beginning with ensemble models and random forests, followed by an overview of pruning techniques in machine learning, fairness in machine learning, and finally the intersection of pruning and fairness, concluding with a synthesis of current gaps in the literature.

Chapter 3. outlines the methodology, introducing the pruning methods studied and providing detailed mathematical formulations for each.

Chapter 4. explains the experimental design, covering dataset selection, metric definitions, experimental setup, and execution strategy.

Chapter 5. is dedicated to dataset analysis, providing an in-depth examination of the datasets' properties, inherent biases, and stability characteristics to offer context for interpreting the experimental results.

Chapter 6. presents the experimental analysis and results, structured into key insights derived from multi-dimensional analyses. It also includes robustness assessments across various scenarios and a final comparative discussion of the pruning methods.

Chapter 7. concludes the thesis by summarizing contributions, limitations, and directions for future research.

CHAPTER 2 LITERATURE REVIEW

This literature review is structured to establish a clear understanding of key concepts and their interconnections in relation to this study. Section 2.1 begins by introducing ensemble models, outlining their benefits, limitations, and the main types of ensemble learning. Building on this foundation, the discussion then focuses on random forests as a widely used ensemble method, explaining their underlying mechanism, advantages, and limitations. From there, in Section 2.2, we narrow the focus to pruning techniques, first discussing their necessity in random forests and then reviewing methods applied in ensemble models more broadly.

Having established this technical foundation, the review moves to the second key area of this study, fairness in machine learning. In Section 2.3, we define bias and its sources in data and algorithms, define fairness (outlining group and individual fairness), and review existing studies at the intersection of pruning and fairness. Finally, in Section 2.4, we highlight some remaining gaps and define the objectives and contributions of this study

2.1 Ensemble Models

Ensemble methods enhance the predictive performance beyond that of individual models by training multiple models and aggregating their outputs [1] [2]. According to Sagi and Rokach [1], “Ensemble learning is considered the state-of-the-art solution for many machine learning challenges”. It has demonstrated superior performance over single models in applications such as fraud detection, medical diagnosis, and autonomous systems [2], [3].

Consider a labeled dataset $\{(x_i, y_i)\}_{i=1}^n$, where each feature vector $x_i \in \mathcal{X} \subseteq \mathbb{R}^p$ represents the i-th observation, and the associated class label y_i is an element of $[\![1, C]\!]$.

A classification ensemble consists of a collection of classifiers $\{(h_m, \alpha_m)\}_{m=1}^M$, where each classifier $h_m : \mathcal{X} \to [0,1]^C$ outputs a score vector for any input x, and α_m denotes its associated weight. The ensemble prediction function can be viewed as a mapping:

$$H : \mathcal{X} \times \mathbb{R}_{\geq 0}^M \to [\![1, C]\!]$$

which determines the predicted class for a given input x by employing a voting mechanism, defined by:

$$H(x; \alpha) = \arg \max_{c \in [\![1,C]\!]} \sum_{m=1}^{M} \alpha_m h_m^{(c)}(x)$$

Here, $h_m^{(c)}(x)$ represents the score assigned by classifier m for class c. If multiple classes share the maximum score, a deterministic tie-breaking rule can be applied to resolve the ambiguity.

Benefits of Ensemble Models. To provide a more detailed perspective on the advantages of ensemble models, Pintelas and Livieris [4] argue that combining multiple learning models results in significantly better performance than single learners. Similarly, Rane et al. [2] emphasize that ensemble methods enhance predictive accuracy, reliability, and overall generalization ability by leveraging the combined strengths of multiple models.

Regarding overfitting, Evans [5] explains that ensemble models help mitigate this issue by balancing complexity with regularization techniques. Additionally, Lim [6] underscores that ensembles reduce model dependency on specific data points, further preventing overfitting. Another notable advantage of ensemble models is their effectiveness in handling high-dimensional data. Capitaine et al. [7] state that random forests perform exceptionally well when dealing with datasets containing many features.

Limitations of Ensemble Models. Despite their advantages, ensemble models face several limitations and challenges:

- **Computational complexity:** Mienye and Sun [8] state that ensemble models require more computational power as multiple models must be trained and stored simultaneously. Shah et al. [9] reinforce this claim by analyzing boosting and stacking algorithms' increased memory and processing requirements.

- **Overfitting risk:** Although ensemble models have the potential to mitigate overfitting, they can also contribute to it under certain conditions. Kedziora [10] explains that ensemble learning may sometimes amplify biases present in the training data, ultimately leading to overfitting . Similarly, Pintelas and Livieris [4] argue that if not properly designed, an ensemble model can heighten the risk of overfitting by overemphasizing specific features.

- **Interpretability and decision-making transparency:** Besides the mentioned challenges, ensemble models are often considered "black boxes." Pardo [11] discusses the interpretability issue, noting that ensemble models often lack transparency, making it difficult to understand individual predictions. Evans [5] further argues that although ensemble methods enhance performance, they obscure decision-making processes. Dhanwanth et al. [12] highlights how black-box ensemble models raise concerns about fairness, particularly in criminal justice and healthcare applications. Similarly, Khalid et al. [3] suggest ensemble techniques must incorporate explainable AI (XAI) frameworks to improve transparency.

Types of Ensemble Learning. Ensemble learning methods include bagging, boosting, and stacking. Shah et al. [9] provide an overview of these techniques, comparing their effectiveness in different scenarios. Additionally, Mienye and Sun [8] offer mathematical and algorithmic representations of these methods, detailing their advantages and disadvantages.

One of the most common ensemble techniques is bagging, which aims to reduce variance by training multiple models on different bootstrap samples [13]. Since bagging is a form of ensemble learning, it inherits many advantages and disadvantages discussed earlier, such as reduced variance, improved generalization, computational challenges, and lack of interpretability [14].

Soloff et al. [15] state that bagging enhances stability in machine learning models while not necessitating assumptions regarding data distribution, base algorithms, or feature dimensionality. It shows that bagging successfully stabilizes even highly unstable base models, making it a valuable approach compared to other ensemble methods that may be sensitive to model-specific assumptions. Liu et al. [13] state that bagging inherently incorporates randomness via bootstrap sampling, rendering it naturally differentially private without requiring extra noise injection. Compared to other models that require external noise to achieve privacy, bagging naturally protects sensitive data while maintaining high predictive accuracy. A well-known implementation of bagging is the Random Forest algorithm [16].

2.1.1 Random Forests

Random Forests are a widely used type of ensemble learning method, build upon the principles of bagging (bootstrap aggregation) framework by introducing an additional level of randomness in feature selection, which provides robust predictive performance by averaging multiple trees, reducing variance, and mitigating the effects of noisy data [17]. This method has been particularly effective in classification tasks such as medical diagnostics and fraud detection [18], [19]. Partopour et al. [20] further state that Random Forests are effective in maintaining high accuracy while reducing computational overhead compared to other machine learning models.

Mechanism of Random Forests. Unlike single-decision trees prone to overfitting, Random Forests introduce randomness through bootstrapping and feature selection. Chi et al. [21] highlight that the key difference lies in how trees are trained independently on random subsets of data, increasing model robustness. Similarly, Daghero et al. [22] emphasize that ensemble aggregation enables better generalization in diverse datasets. As a result, random forests are highly adaptable to various data structures, especially high-dimensional tabular datasets, reducing the need for extensive feature engineering [23]. To provide a clearer understanding of how Random Forests are constructed, it is helpful to briefly recall the principle of bagging, on which they are based. In bagging, M base learners are trained on different bootstrap samples of the training dataset, and their predictions are combined, typically through majority voting [24].

Random Forest extends this framework by introducing additional randomness during the training of each tree. While each decision tree is built from a bootstrap sample, at each node split, a random subset of features of size m_{try} is selected, and the best split is determined from this subset rather than from the entire feature set.

Formally, a Random Forest consists of a collection of decision trees $\{T_m\}_{m=1}^{M}$, each trained on a bootstrapped version of the data. For a given input $\mathbf{x}$, each tree assigns a predicted class. The final prediction of the Random Forest is obtained by aggregating these predictions through majority voting:

$$H(\mathbf{x}) = \arg \max_{c \in \{1,\dots,C\}} \sum_{m=1}^{M} h_m^{(c)}(\mathbf{x})$$

where $h_m^{(c)}(\mathbf{x})$ is an indicator function equal to 1 if tree m assigns class c to $\mathbf{x}$, and 0 otherwise. In practice, each tree has an equal vote, with $\alpha_m = \frac{1}{M}$.

Additionally, Random Forests leverage out-of-bag (OOB) samples, which are observations not included in the bootstrap sample for a given tree. These OOB samples are used as an internal validation set to estimate the model's generalization error and to compute feature importance without requiring a separate validation dataset.

Advantages and Limitations of Random Forests. As mentioned, Random Forests are widely recognized for their high classification accuracy, especially in large datasets. Saadoon and Abdulamir [25] demonstrated that Random Forests achieved a very high accuracy when applied to big data, highlighting their ability to process vast amounts of information with minimal performance degradation. This robustness makes them suitable for complex real-world applications like IoT and big data analytics.

Another significant advantage of Random Forests is their ability to handle missing data and high-dimensional datasets. Unlike many other machine learning algorithms that require extensive pre-processing, Random Forests can make accurate predictions even when certain features are missing. Zhu [26] explains that Random Forests are immune to statistical assumptions and can efficiently process large datasets without excessive feature selection or transformation. Additionally, the paper supports that Random forests perform well in high-dimensional spaces, making them a preferred choice for applications in medical diagnosis, genetics, and finance.

Random Forests are also less sensitive to noisy data due to their ensemble nature [27]. Schmalohr et al. [19] provide empirical evidence that Random Forests outperform other methods in detecting epistatic interactions in noisy genetic data. Xiao et al. [28] reinforce this by showing that Random Forests achieve high accuracy in medical diagnosis despite noisy datasets. These benefits further establish Random Forests as a powerful, adaptable, and reliable machine-learning model for various domains.

Despite their advantages, Random Forests have their challenges and limitations. The black-box nature of Random Forests, inherent in bagging ensemble models, makes them challenging to interpret. Benard [29] note that the high number of computations per prediction obscures model decision-making. Furthermore, large ensembles require significant computational resources. Daghero et al. [22] discuss how inference latency and energy consumption increase with larger forests. Hyperparameter tuning is also essential for optimizing Random Forests. Kuyoro et al. [30] show that increasing tree depth improves accuracy but can lead to overfitting. Rhodes et al. [31] argue that feature selection strategies significantly impact model performance, emphasizing the need for careful hyperparameter adjustments.

2.2 Pruning in Machine Learning

As previously mentioned, Random Forests are among the most effective machine learning models, although they tend to generate large decision trees, leading to increased computational overhead and less interpretability [8]. This issue has led researchers to explore methods for improving their efficiency while preserving their predictive power. Studies such as those by Bernard et al. [32] investigate whether finding a subset of trees is possible while improving its accuracy. Beyond accuracy improvements, research on the interpretability of Random Forests suggests that pruning can reduce complexity, making models more transparent and easier to understand. Dorador et al. [33] highlight that, in some instances, the forest can be reduced so drastically that the remaining trees seamlessly merge into a single model, significantly enhancing interpretability compared to the original regression forest, which remains a black box. Research on the computational effects of pruning indicates that reducing the number of trees in a Random Forest decreases memory and processing requirements. Surjanovic et al. [34] show that their pruning technique can optimize energy consumption in resource-constrained environments.

Given the growing recognition of pruning's benefits, researchers have extensively explored various techniques to refine ensemble and Random Forest models. Kulkarni and Sinha [35] state that "For effective learning and classification of Random Forest, there is a need for reducing the number of trees (pruning) in Random Forest" . Similarly, Dorador et al. [33] emphasize that forest pruning plays a crucial role in balancing performance and interpretability.

Rose and Hassen [36], in their comprehensive review of Random Forest pruning techniques, highlight that many researchers have developed methods to select the optimal subset of trees. These pruning strategies focus on improving classifier performance by eliminating unnecessary trees.

Given these challenges, researchers have explored various strategies to enhance the efficiency and interpretability of Random Forests. One such approach is pruning, a technique that optimizes models by removing redundant or less important components, reducing their size and computational complexity while maintaining or even improving accuracy.

2.2.1 Pruning Methods for Ensemble Models

To better understand pruning techniques in ensemble models and Random Forests, researchers have classified these methods based on their implementation strategies and selection criteria. In a recent survey, Kulkarni and Sinha [35] classified ensemble pruning techniques into static and dynamic approaches, with most research focusing on static methods [37–40]. Static pruning involves selecting an optimal subset of trees before deployment to enhance efficiency from the outset. In contrast, dynamic pruning continuously adjusts the ensemble during runtime based on real-time performance metrics and criteria [41].

Although dynamic pruning is an emerging area of study with increasing research interest, it has been less frequently categorized in the literature. Since our focus is on static pruning techniques, we will primarily discuss methods falling under this category. Expanding on these classifier selection techniques, Tsoumakas et al. [42] Further classified ensemble pruning methods into ranking-based, clustering-based, and optimization-based approaches, each offering distinct strategies for refining ensemble models.

Several ensemble pruning approaches have been proposed in the literature and can be broadly categorized as follows:

- **Ranking-Based Method**: This approach evaluates the contribution of each tree to the overall accuracy and removes those with minimal impact. Classifiers are ranked based on evaluation metrics such as orientation and kappa pruning [42]. Studies such as [39], [40] have employed this method. Guo et al. [39] introduced the Margin and Diversity Measure (MDM), which ranks classifiers based on low margin, high diversity, and high accuracy. The final number of classifiers can be fixed or dynamically determined.

- **Clustering-Based Method**: This technique identifies redundant trees within the ensemble and eliminates similar ones to preserve diversity. The method clusters similar classifiers before selecting a diverse subset to enhance ensemble performance [42]. Popular clustering algorithms, such as k-means [43] and deterministic annealing [41], are frequently used in this process.

- **Optimization-Based Methods**: This approach leverages advanced optimization techniques such as genetic algorithms, hill climbing, and semi-definite programming to refine ensemble selection and identify the most effective subset of trees [41].

Since this research will focus on various methods across different categories of pruning approaches, each of these methods and their mathematical formulations will be presented in detail in Section 3.

Many Random Forest pruning methods align with these three main categories. Research in Random Forest optimization generally focuses on:

- **Base Classifier Construction**: Modifying the number of features used at each node or applying different evaluation criteria to determine the best split at every node [44].

- **Base Classifier Selection**: Identifying the most practical combination of classifiers from a given pool [44].

For example, Robnik-Sikonja [45] investigated both aspects to enhance the overall performance of Random Forests. Similarly, Tsymbal et al. [46] proposed an improved method by replacing the simple majority voting approach with a more advanced dynamic integration technique to enhance model performance. Given these advantages and studies, pruning techniques are widely applied across various domains to enhance efficiency and interpretability.

While pruning techniques offer significant advantages across various domains, they also present challenges, particularly in balancing accuracy, fairness, and efficiency. However, careful tuning of pruning parameters is essential to maintain accuracy while optimizing efficiency. Tarchoune et al. [47] demonstrate that fine-tuning pruning parameters helps balance accuracy and model simplicity by applying a pre-pruning technique that optimizes node splits based on performance thresholds, effectively reducing unnecessary tree growth. Their approach, validated on multiple medical datasets, achieved prediction accuracies above 80% across all cases. Nieth et al. [48] examine the effects of dataset pruning in adversarial training on model robustness. While pruning reduces data size, an improper implementation may remove essential features, making the model more vulnerable to adversarial attacks.

Beyond efficiency and robustness, pruning techniques also raise ethical concerns, particularly in relation to fairness. The process of removing trees or features can inadvertently amplify biases, affecting model decisions and outcomes [49]. As fairness in AI continues to gain attention, researchers emphasize the need to balance model optimization with ethical responsibility. This leads to a broader discussion on fairness in machine learning.

2.3 Fairness in Machine Learning

Bias in machine learning refers to systematic errors that lead to unfair outcomes, favoring or disadvantaging certain individuals or groups [50]. These biases can cause real-world harm, resulting in discriminatory decisions in critical areas such as healthcare, hiring, and criminal justice [51]. Sources of bias are generally categorized into two types: data bias and algorithmic bias. Data bias includes historical bias, representation bias, sampling bias, and label bias, all arising from unrepresentative datasets or data reflecting past societal prejudices [52]. Algorithmic bias, on the other hand, is introduced by model structure, feature selection, optimization methods, or the design of the algorithm itself [53], and can amplify existing disparities even if the data is corrected [54]. These concerns are particularly relevant in pruning algorithms, where the selection of features or patterns may unintentionally exacerbate bias and impact fairness.

Fairness in machine learning aims to reduce these biases and prevent discrimination. However, fairness is not only a technical challenge but also a socio-technical problem influenced by societal norms, legal frameworks, and context-specific ethical considerations [55]. Because fairness is context-dependent, multiple definitions have emerged to address different scenarios. At a high level, two primary concepts are commonly discussed: group fairness and individual fairness. Group fairness requires that different demographic or protected groups receive equal treatment. Metrics such as demographic parity and equal opportunity are frequently used. For example, demographic parity ensures that the probability of a positive outcome, such as loan approval, is the same across groups regardless of protected attributes [56]. Equal opportunity ensures equal probability of a correct positive prediction across groups [57], [58], while equalized odds balance both true positive and false positive rates across demographic groups [59], [60]. Another relevant measure, disparate impact, assesses whether certain groups are disproportionately disadvantaged [61], [59]. Individual fairness, by contrast, states that similar individuals should receive similar outcomes. For instance, two applicants with nearly identical qualifications should receive the same decision regardless of demographic characteristics [55].

Sensitive attributes — characteristics such as race, gender, age, and religion — are central to fairness-aware machine learning. These attributes, if not properly controlled for, can lead to discrimination through direct use or through proxy features that encode these attributes implicitly [62], [63]. The socio-technical nature of this challenge means fairness interventions must balance accuracy, societal context, and ethical responsibility, especially in models undergoing pruning or optimization changes that could alter feature importance and group impacts.

Fairness plays a vital role in ensuring that AI models do not discriminate against specific demographic groups or individuals and supports equitable access to opportunities across various domains [64]. Wachter et al. [65] highlight its significance in AI ethics, noting that biased models can lead to systemic discrimination. Similarly, Beutel et al. [66] stress the importance of fairness in algorithmic decision-making to prevent unintended social consequences. Zhou et al. [61] emphasize fairness in promoting ethical AI deployment, particularly in high-stakes applications. However, despite ongoing efforts, AI models often inherit biases from the data they are trained on, resulting in unfair outcomes. Pagano et al. [59] discuss how embedded biases in datasets, algorithms, and user interactions complicate fairness challenges, making it difficult to ensure unbiased decisions. Gohar and Cheng [58] further highlight that biased training data frequently results in discriminatory outcomes in real-world AI applications.

These biased models can reinforce existing inequalities, systematically disadvantaging certain groups. Shome et al. [60] illustrate how biased credit-scoring models have contributed to financial exclusion, while Pagano et al. [59] address legal challenges associated with bias in hiring and criminal justice models. Wan et al. [57] examine how AI-driven hiring models have historically disadvantaged women and minority candidates. Gohar and Cheng [58] provide additional examples of bias in policing and healthcare settings. The consequences of such biases extend beyond technology, influencing public policy and social structures, with fairness playing a critical role in domains like education, mental health, criminal justice, and housing safety [67]. Bird et al. [68] explore fairness considerations in consumer healthcare AI and highlight ongoing research to address bias in these systems.

Numerous studies underscore the necessity of fairness to prevent systemic discrimination in AI decision-making. Wachter et al. [65] investigate fairness in automated credit scoring, highlighting the challenges in ensuring equitable outcomes. Oneto and Chiappa [69] propose fairness-aware AI frameworks aimed at mitigating bias in algorithmic decisions. As fairness continues to gain attention, researchers have developed various methods to assess and address bias. Measuring fairness is critical for identifying disparities and evaluating model behavior. However, detection alone is insufficient. Mitigation strategies, including methods like statistical parity, equal opportunity, equalized odds, demographic parity, and disparate impact analysis, must be implemented to reduce discrimination [57–61, 70].

This socio-technical challenge demands ongoing attention to ensure AI models function responsibly in real-world settings. In this study, fairness is particularly examined in the context of pruning techniques and ensemble models, where changes in feature selection and model structure may influence fairness outcomes.

Challenges in Ensuring Fairness. While fairness metrics and mitigation strategies help address bias in AI models, they also introduce new complexities. Ensuring fairness is not just about applying the right techniques but also about navigating trade-offs between accuracy, interpretability, and ethical considerations. These ongoing challenges continue to shape fairness research in machine learning.

The balance between fairness and accuracy in machine learning remains a subject of ongoing debate. Rodolfa et al. [67] argue that fairness-aware models can achieve both high accuracy and improved fairness without significant trade-offs. Similarly, Beutel et al. [66] propose fairness-aware training techniques that aim to minimize the trade-offs between fairness and predictive performance.

Beyond the fairness-accuracy trade-off, another key challenge lies in the distinction between group fairness and individual fairness. A central question in fairness research is whether fairness should be assessed at the group level or the individual level. Lee et al. [71] argue that fairness metrics should balance these two perspectives to promote equitable outcomes. Kirat et al. [72] examine how legal frameworks shape fairness considerations at both the individual and group levels.

While fairness-aware techniques seek to mitigate bias in AI models, an emerging research direction explores how model compression techniques, such as pruning, influence fairness. Since pruning modifies model structures to improve efficiency, its impact on fairness must be carefully examined.

2.3.1 Intersection of Pruning and Fairness in Machine Learning

Recent studies have explored the intersection of pruning and fairness in machine learning models, aiming to balance efficiency gains with ethical considerations. Some research focuses on jointly optimizing pruning and fairness constraints to ensure that model compression does not disproportionately impact different demographic groups. Dai et al. [73] propose a bi-level optimization framework that simultaneously optimizes pruning and fairness, ensuring that compressed models maintain equitable performance. Similarly, Lin et al. introduce FairGRAPE, a gradient-based pruning method that minimizes fairness degradation during model compression. By assessing the per-group importance of model weights, FairGRAPE ensures that pruning does not disproportionately affect underrepresented groups, particularly in face attribute classification tasks [74].

Other studies apply fairness-aware pruning in specific applications, demonstrating its impact in real-world domains. Wu et al. [75] present FairPrune, a method that leverages pruning to enhance fairness in dermatological disease diagnosis models. Their findings reveal that fairness can be significantly improved while maintaining diagnostic accuracy by strategically removing parameters that contribute to bias. Addressing fairness concerns in structured pruning, Meyer and Wong [76] propose a fair loss function tailored for pruning, mitigating biased performance degradation during the pruning process. Their approach enables models to retain fairness while optimizing efficiency. Zayed et al. [77] investigate structured pruning in transformer-based language models, highlighting that pruning specific attention heads can inadvertently amplify bias, further emphasizing the importance of fairness-aware pruning strategies.

A growing body of research focuses on bias-aware pruning techniques for decision trees, where pruning is specifically designed to mitigate biased decision-making. Ranzato et al. [78] propose adversarial pruning, which removes biased nodes in decision trees to reduce unfair decision-making. Similarly, Kashyap et al. [79] demonstrate that fairness-aware pruning enhances model robustness across different demographic groups. In some cases, fairness-aware pruning can even outperform traditional pruning. Zhang et al. [80] introduce FairRepair, a method that selectively prunes biased branches while preserving accuracy, demonstrating its potential to maintain high performance while addressing bias. Fitzsimons et al. [81] argue that fairness-aware pruning can be effectively integrated into standard tree-based models with minimal performance loss, making it a viable strategy for improving fairness in decision tree ensembles.

Beyond static pruning approaches, researchers have also explored dynamic pruning techniques to maintain fairness over time. Zhang et al. [82] introduce dynamic pruning techniques to maintain fairness in neural networks.

Additionally, fairness-aware pruning has been extended to ensemble models, where the focus is on balancing predictive accuracy and bias mitigation. Bian and Zhang [83] present discriminative risk as a fairness metric that incorporates both individual and group fairness. They also introduce oracle bounds to provide theoretical learning guarantees for fairness enhancement and propose an ensemble pruning technique utilizing this measure. Lastly, Paganini [84] analyzes 100,000 image classification models, concluding that naive pruning strategies can exacerbate bias and disproportionately affect underrepresented classes, underscoring the importance of transparency and fairness considerations in pruning.

2.4 Gaps and Research Objectives

Through this literature review, we presented ensemble models with a particular focus on Random Forests, discussed their mechanism and pruning techniques, and examined fairness challenges and mitigation strategies in machine learning. We also reviewed studies exploring the intersection of pruning and fairness, noting contributions in areas such as gradient-based fairness-aware pruning, adversarial pruning, and loss function modifications. Existing work has demonstrated the potential for pruning methods to influence fairness outcomes, especially in deep learning models and specific applications like healthcare, finance, and image classification.

However, several research gaps remain. First, most studies address fairness-aware pruning within the context of deep learning or highly specific domains. There is a lack of systematic, high-level analysis on how different pruning techniques affect fairness in Random Forest models, despite their widespread use in real-world, tabular, and interpretable settings. Second, prior research typically focuses on a single pruning method applied to a particular dataset or industry use case. There is no comparative analysis that evaluates multiple pruning strategies across different scenarios, modeling goals, or objective fairness and accuracy metrics. This makes it difficult to generalize findings or understand the broader implications of pruning on fairness.

This study addresses these gaps by conducting a structured comparison of multiple Random Forest pruning methods. We evaluate each approach across diverse datasets and modeling objectives, assessing their impact on fairness and predictive performance using established fairness metrics. In doing so, we provide insights into how different pruning methods perform across varying scenarios, moving beyond application-specific solutions to offer a broader, foundational understanding. This contribution advances fairness-aware machine learning research by systematically analyzing an underexplored area and identifying patterns in the trade-offs between accuracy and fairness. By filling this gap, our work offers a knowledge base that can help researchers and practitioners better understand the impact of pruning methods on Random Forest models, enabling them to design their studies and settings with greater awareness and informed decision-making.

CHAPTER 3 Methodology

In this work, we perform a thorough evaluation of four foundational ensemble pruning methods — Greedy Pruning, Cluster-based Pruning, Rank-based Pruning, and Mixed-Integer Quadratic Programming (MIQP)-based Pruning — applied to random forests, aiming to assess their impact on various fairness metrics. Each of these methods follows different strategies and optimization goals, and we present their formal definitions, configurations, and settings in a self-contained manner. Our objective is not to propose a new pruning technique, but rather to experimentally compare these established methods and analyze their effects on model performance and fairness across different sensitive attributes. This section presents a comprehensive overview of the pruning techniques employed in this study, outlining their theoretical foundations, selection criteria, and algorithmic implementation.

3.1 Cluster Pruning

In cluster-based Pruning methods, the fundamental idea is to identify groups (clusters) of similar classifiers within the ensemble and then select a representative subset of these clusters to form a smaller, more efficient ensemble [43], [85]. The underlying assumption is that classifiers within the same cluster make similar errors and are thus somewhat redundant [43].

Methodological Framework and Mathematical Formulation. The general cluster pruning approach consists of generating an initial ensemble, measuring similarity (or dissimilarity) between classifiers, clustering them, selecting representatives, and combining their outputs. First, a diverse ensemble of classifiers $\{h_m\}_{m=1}^{M}$ is created using methods like bagging, boosting, or varying architectures and data [86], [43]. To group similar classifiers, a similarity or distance metric is defined based on either model outputs or, less commonly, model parameters [41]. For outputs, classifiers are compared using their predictions on a dataset S_{data}. The similarity between classifiers h_i and h_j can be computed as the correlation of their prediction vectors Y_i and Y_j, where $Y_i = [y(h_i; \mathbf{x}_1), \ldots, y(h_i; \mathbf{x}_m)]$ [43]. Alternatively, distance can be measured using the compound error probability [43], [86]:

$$d(h_i, h_j) = 1 - \text{prob}(h_i \text{ fails on } \mathbf{x}, h_j \text{ fails on } \mathbf{x})$$

These similarity values form a similarity or distance matrix S of size $M \times M$, with entries $S_{ij} = \textit{similarity}(h_i, h_j)$ or $D_{ij} = \textit{distance}(h_i, h_j)$. Using this matrix, classifiers are clustered into k clusters, $Cl_1, Cl_2, \ldots, Cl_k$, with algorithms such as k-means [43], hierarchical clustering [86], or spectral clustering [85]. For instance, in k-means clustering, classifiers represented by their prediction vectors are partitioned by minimizing the sum of squared distances to cluster centroids $\{M_j\}_{j=1}^k$:

$$\operatorname*{argmin}_{\{Cl_i\}_{i=1}^k} \sum_{i=1}^{k} \sum_{Y_j \in Cl_i} d^2(Y_j, M_i)$$

Once clusters are formed, a representative classifier $h_i^* \in Cl_i$ is selected from each cluster, either by choosing the most accurate classifier on a validation set, the one closest to the cluster centroid, selecting a small number of diverse and accurate classifiers, or distributing the voting weight of pruned classifiers among the selected ones based on their similarity [43]. The pruned ensemble then consists of $\{h_1^*, \ldots, h_k^*\}$. Finally, predictions for new instances $\mathbf{x}$ are obtained by combining the outputs of the selected classifiers using an aggregation rule (e.g., majority voting or weighted averaging) [86], [43]:

$$\hat{y}(\mathbf{x}) = aggregate(\{y(h_1^*; \mathbf{x}), \ldots, y(h_k^*; \mathbf{x})\})$$

Key Variations in Literature. The literature on cluster pruning of classifier ensembles explores several variations and refinements of this general framework.

Choice of clustering algorithm is one important variation. Methods include k-means clustering, where classifiers are grouped based on their prediction vectors and the number of clusters k is determined by monitoring diversity between clusters or performance on a validation set [43]. Hierarchical agglomerative clustering is also used, employing compound error probability as a distance measure to identify subsets of classifiers [86]. Additionally, spectral clustering has been proposed to prune ensembles by analyzing pairwise diversity within the set of classifiers [85].

Another key variation lies in the **similarity and distance measures** used. Some approaches use correlation coefficients between prediction vectors to quantify classifier disagreement [43], [85]. Others employ distance measures based on compound error probability to group classifiers that make coincident errors. Spectral clustering approaches often utilize Q statistics to measure pairwise diversity [85].

Representative selection strategies also differ. One common method selects the most accurate classifier within each cluster based on validation performance and discards the rest. Another approach chooses the classifier closest to the cluster centroid in the output space. Furthermore, some techniques implement a strategy where the voting weights of removed classifiers are distributed among the remaining unpruned classifiers according to their similarity. This aims to maintain ensemble authenticity, particularly at high pruning rates.

Determining the optimal number of clusters is another challenge. Some approaches involve **gradually increasing the number of clusters** and stopping when diversity between cluster representatives deteriorates. Others monitor the prediction accuracy of the pruned ensemble on a validation set, stopping when further pruning causes significant performance degradation [43].

Finally, certain pruning techniques are adapted for **specific ensemble types**. While the general framework is broadly applicable, some methods are designed for ensembles like Random Forests. An example is CLUB-DRF, a clustering-based random forest pruning method [85].

In summary, cluster pruning offers a principled way to reduce the size and complexity of classifier ensembles by identifying and removing redundant classifiers based on their similarity. The effectiveness of different cluster pruning methods depends on the choice of similarity measure, clustering algorithm, representative selection strategy, and the method used to determine the optimal level of pruning.

The practical implementation of cluster pruning used in this work is described in Algorithm 1.

Algorithm 1 Cluster Pruning Algorithm

```
Require: Number of classifiers ($M$), clustering method ($ClusterClassifiers$), selection
    method ($SelectClassifiers$), clustering mode ($ClusterMode$)
Compute classifier representations based on $ClusterMode$
Cluster classifiers into $k$ clusters using $ClusterClassifiers$
for each cluster $Cl_i$ do
    Select a representative classifier using $SelectClassifiers$
end for
Return pruned ensemble $\{h_1^*, \ldots, h_k^*\}$
```

3.2 Greedy Pruning

Greedy pruning methods for ensemble learning aim to **reduce the size of an ensemble of classifiers by iteratively selecting a subset of the original ensemble based on a specific evaluation criterion**. The goal is to improve storage and computational efficiency as well as potentially enhance generalization performance. Instead of directly selecting an optimal or near-optimal subset, greedy methods explore the space of possible classifier subsets by making locally optimal choices at each step [87].

Methodological Framework and Mathematical Formulation. The general greedy pruning approach can be performed in two directions: **forward** or **backward**. In **greedy forward pruning**, the process starts with an empty sub-ensemble $S_0 = \emptyset$. At each iteration

k, each candidate classifier $h \in \{h_m\}_{m=1}^{M} \setminus S_k$ is evaluated by computing the evaluation criterion $J(S_k \cup \{h\})$, considering performance metrics and/or diversity. The next classifier h_{next} to be added is the one that optimizes this criterion:

$$h_{\text{next}} = \arg \max_{h \in \{h_m\}_{m=1}^{M} \setminus S_k} J(S_k \cup \{h\})$$

The selected classifier is then added to the sub-ensemble, $S_{k+1} = S_k \cup \{h_{\text{next}}\}$, and the process continues until a stopping condition is met, such as achieving a desired ensemble size or when adding further classifiers no longer improves performance [87].

In **greedy backward pruning**, the process starts with the full ensemble $\{h_m\}_{m=1}^{M}$. At each iteration k, the classifier h_{remove} to be eliminated from the current sub-ensemble S_k is chosen to minimize the degradation or maximize the improvement of the evaluation criterion:

$$h_{\text{remove}} = \arg \min_{h \in S_k} J(S_k \setminus \{h\})$$

if the criterion measures performance (higher is better), or equivalently:

$$h_{\text{remove}} = \arg \max_{h \in S_k} L(S_k \setminus \{h\})$$

if the criterion measures loss or error. This process repeats until the ensemble reaches the desired pruned size or removing classifiers no longer benefits the criterion. The evaluation criterion $J(S)$ or $L(S)$ can be based on factors such as the **accuracy** of the sub-ensemble on a validation set and the **diversity** among classifiers [87].

Greedy pruning methods offer a computationally efficient way to reduce ensemble size while making locally optimal decisions based on the chosen evaluation criterion. The effectiveness of these methods strongly depends on selecting an appropriate evaluation criterion suited to the task and dataset [87].

Key Variations in Literature. The primary variations in greedy pruning methods lie in the choice of the **evaluation criterion** used to guide the selection or removal of classifiers. Several criteria have been proposed in the literature. One common approach is **accuracy-based** selection, where classifiers are chosen based on their individual or collective ability to improve the ensemble's accuracy on a validation set. Techniques such as "nice Bagging" and "trimmed Bagging" fall under this category, excluding classifiers with poor accuracy [85].

Another approach is **diversity-based**, where classifiers are selected to ensure they differ from one another, reducing redundancy and improving generalization. Measures like the Kappa statistic, complementarity, orientation, and margin distance are often used to quantify diversity [87]. Some methods combine both **accuracy and diversity**, using criteria that simultaneously consider predictive strength and diversity among classifiers. The Diversity Regularized Ensemble Pruning (DREP) method explicitly integrates diversity regularization into the selection process [85], [87].

Margin-based criteria are also used, focusing on the distribution of margins within the ensemble. Quadratic Margin Maximization is one technique aimed at optimizing margin distribution and increasing diversity [85]. Additionally, **error reduction**-based approaches, such as a simplified version of reduce-error pruning, iteratively add classifiers that lower the ensemble's error on a validation set [88].

Other variations include the use of a **complementariness measure**, where classifiers are ordered according to how complementary their predictions are to the ensemble. **Margin distance minimization** is another criterion that ranks classifiers based on minimizing margin distance [88]. Finally, methods like **Information Exchange Glowworm Swarm Optimization (IEGSO)** combined with the **Complementarity Measure (COM)** have been proposed, leveraging swarm optimization to select a diverse and accurate subset of classifiers based on complementarity [85].

The direction of the greedy search (forward or backward) is another variation, although some studies suggest that the direction does not significantly affect the final performance, with forward selection often producing smaller ensembles.

The practical implementation of cluster pruning used in this work is described in Algorithm 2.

Algorithm 2 Greedy Pruning Algorithm

```
Require: Number of classifiers ($M$), scoring metric ($Metric$)
Initialize an empty set $S$ of selected classifiers
for each iteration until $M$ classifiers are selected do
    Compute the contribution of each candidate classifier using $Metric$
    Select the classifier maximizing improvement in $Metric$
    Add the selected classifier to $S$
end for
Return pruned ensemble $S$
```

3.3 MIQP-Based Pruning

Zhang et al. [89] formulates the pruning problem as a quadratic integer program and solves it using SDP relaxation, the general approach, mathematical formulation, and related concepts of Mixed-Integer Quadratic Programming (MIQP)-based pruning for ensemble learning are presented in the following content. Finding the optimal subset of classifiers is a combinatorial optimization problem, which becomes computationally intractable for even moderately sized ensembles [89], [90]. Many heuristic methods, often based on greedy search, have been proposed to find approximate solutions. However, these heuristics often lack theoretical or empirical quality guarantees [89]. This motivates using more sophisticated optimization techniques, such as formulating the pruning problem as a mathematical program. While the primary source utilizes quadratic integer programming (QIP) and solves it with Semidefinite Programming (SDP), the core idea of formulating the selection with integer variables and a quadratic objective relates to the broader concept of MIQP, especially if mixed-integer variables were to be involved in variations.

Methodological Framework and Mathematical Formulation. The MIQP-based pruning method, the goal is to select a subset of classifiers that simultaneously optimizes the ensemble's accuracy and diversity. Let there be M classifiers in the original ensemble, denoted by $\{h_m\}_{m=1}^{M}$. A binary variable $x_m \in \{0, 1\}$ is associated with each classifier h_m, where $x_m = 1$ if classifier h_m is included in the pruned ensemble, and $x_m = 0$ otherwise. The objective is to select exactly k classifiers to minimize a quadratic cost function.

The foundation of this method lies in constructing an error matrix P, where each entry $P_{im} = 1$ if classifier h_m misclassifies the i-th data point, and 0 otherwise. From this, an error correlation matrix $G = P^T P$ is computed. The diagonal elements G_{mm} represent the number of errors made by classifier h_m, and the off-diagonal elements $G_{m_1 m_2}$ indicate the number of common errors between classifiers h_{m_1} and h_{m_2}, capturing error correlation.

To standardize the scales, the matrix G is normalized into $\tilde{G}$. The diagonal elements $\tilde{G}_{mm}$ reflect the error rate of classifier h_m, while the off-diagonal elements are defined as:

$$\tilde{G}_{m_1 m_2} = \frac{1}{2}\left(\frac{G_{m_1 m_2}}{G_{m_1 m_1}} + \frac{G_{m_1 m_2}}{G_{m_2 m_2}}\right),$$

which represents the average conditional probability of simultaneous misclassification between two classifiers. Lower values in $\tilde{G}$ indicate lower overlap and better diversity.

The quadratic objective function to be minimized is:

$$\min_{x} x^T \tilde{G} x$$

subject to the constraints:

$$\sum_{m=1}^{M} x_m = k$$

ensuring the pruned ensemble size is k, and

$$x_m \in \{0, 1\} \text{ for all } m = 1, \ldots, M$$

enforcing binary selection decisions. Expanding the objective function:

$$x^T \tilde{G} x = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} x_{m_1} \tilde{G}_{m_1 m_2} x_{m_2}$$

shows that the goal is to select classifiers with both low individual error rates and minimal error overlap [89].

Since solving this QIP problem is NP-hard, the approach uses a semi-definite programming (SDP) relaxation. By relaxing the binary constraints and reformulating the problem in terms of a matrix variable $V = vv^T$, approximate solutions can be obtained through convex optimization. Finally, randomized rounding, inspired by methods for solving the Max-Cut problem, is applied to the SDP solution to produce a binary vector x, indicating the chosen subset of classifiers.

Key Variations in Literature. In their paper, Zhang et al. [89] highlight various strategies for ensemble pruning that could potentially be formulated using MIQP or related integer programming techniques. One important variation involves **diversity maximization**, where certain methods aim to maximize diversity among selected classifiers. This objective can be incorporated into the optimization function or constraints of an integer programming formulation. Cavalcanti et al. [90] also emphasize the role of diversity and propose combining multiple diversity measures to improve ensemble performance.

Another key variation is the **accuracy and diversity trade-off**. Many pruning methods attempt to balance individual classifier accuracy with ensemble diversity [89]. The QIP formulation directly captures this balance by considering both the diagonal elements (reflecting accuracy) and off-diagonal elements (representing diversity) of the $\tilde{G}$ matrix.

Weight-based optimization is also explored in the literature. Some approaches assign weights to classifiers and prune those with low or zero weights [90], [89]. While the formulation discussed by Zhang et al. focuses on binary weights for subset selection, extending this approach to continuous or integer weights within specified bounds could lead to more flexible MIQP formulations.

Another variation includes **clustering-based pruning**, where classifiers are grouped, and representatives are selected from each cluster. This approach may involve integer variables to indicate cluster membership and the selection of a single classifier per cluster, with the objective of improving overall ensemble performance [85].

Lastly, the **performance on a validation set** can guide pruning decisions. The genetic algorithm proposed by Cavalcanti et al. [90] uses a fitness function based on validation set error rates, indicating that MIQP formulations could also integrate validation performance into their objectives.

In essence, the MIQP-based pruning method encodes the ensemble's strength (accuracy) and diversity into a quadratic optimization framework. By using error structures and advanced optimization techniques like SDP relaxation and randomized rounding, it aims to identify an optimal or near-optimal subset of classifiers that collectively minimize individual errors and redundant error overlap [89].

The practical implementation of MIQP pruning used in this work is described in Algorithm 3.

Algorithm 3 MIQP Pruning Algorithm

```
Require: Number of classifiers ($M$), individual metric ($SingleMetric$), pairwise metric ($PairwiseMetric$), trade-off parameter ($\alpha$)
Compute individual classifier performance scores as vector $q$
Compute pairwise interaction terms as matrix $P$
Solve the optimization problem:
    $$\arg\min_{w} \Big( (1-\alpha) q^T w + \alpha w^T P w \Big)$$
Return the optimized selection vector $w$
```

3.4 Rank-Based Pruning

The concept of **Rank pruning** aligns with **ordering-based ensemble pruning methods**, where each classifier in the ensemble is evaluated according to a defined criterion, **ranked** based on these evaluation scores, and the top-ranked classifiers are selected to form a pruned sub-ensemble [91], [92].

Methodological Framework and Mathematical Formulation. The general procedure consists of three main steps: (1) defining an evaluation metric that measures the importance or contribution of each classifier (based on accuracy, diversity, margin, or combined metrics); (2) ranking classifiers based on these scores; and (3) selecting the top-ranked classifiers according to a target ensemble size or resource constraints. Mathematically, let $H = \{h_1, h_2, \ldots, h_M\}$ be the original ensemble of M classifiers. An evaluation function $E(h_m)$ assigns a score to each classifier $h_m \in H$. The scores $s_m = E(h_m)$ are calculated for all $m = 1, \ldots, M$, and classifiers are ranked so that if $s_i > s_j$, then h_i precedes h_j in the ordered list OL. The top p percent or the top T classifiers from OL are selected to form the pruned ensemble H_{pruned} [91], [92].

Rank pruning is based on the observation that not all ensemble members contribute equally to the overall performance [91–93]. Some classifiers may be redundant, poorly performing, or detrimental to accuracy [85]. By evaluating and **ranking** classifiers, it becomes possible to eliminate less valuable ones, forming a smaller ensemble that can maintain or improve performance with reduced computational cost [91], [92].

The following methods highlight different implementations of rank pruning.

Ensemble Pruning via Individual Contribution Ordering (EPIC): EPIC is based on the insight that classifiers which are more accurate and correctly predict samples from the minority group are more valuable, while incorrect predictions on minority group samples are considered less harmful [92]. The procedure begins with the definition of a heuristic metric for evaluating individual contribution (IC_i), which combines accuracy and diversity by accounting for predictions on both majority and minority classes. The IC_i score for each classifier h_i is computed on a pruning set D_{pr} using the following formula:

$$IC_i = \sum_{j=1}^{N_{pr}} (\alpha_{ij}(2v_{max}^{(j)} - v_{h_i(\mathbf{x}_j)}^{(j)}) + \beta_{ij} v_{sec}^{(j)} + \theta_{ij}(v_{correct}^{(j)} - v_{h_i(\mathbf{x}_j)}^{(j)} - v_{max}^{(j)}))$$

In this formulation, α_{ij} is equal to 1 if classifier h_i correctly predicts the j-th data point and this prediction belongs to the minority group, and 0 otherwise. Similarly, β_{ij} is set to 1 if the prediction is correct in the majority group, and 0 otherwise. The parameter θ_{ij} is equal to 1 if the classifier's prediction is incorrect, and 0 otherwise. The term $v_k^{(j)}$ represents the number of votes for label k on the j-th data point in D_{pr}. Furthermore, $v_{max}^{(j)}$ denotes the count of the most common prediction for that data point, $v_{sec}^{(j)}$ indicates the second-highest prediction count, and $v_{correct}^{(j)}$ corresponds to the count for the true label. Once all IC_i scores are computed, the classifiers are ranked in decreasing order based on these scores, and the top classifiers are selected to form the pruned ensemble [91], [92].

Margin & Diversity based ordering Ensemble Pruning (MDEP): The MDEP method measures the importance of classifiers by combining both the margin of examples and ensemble diversity, with particular focus on examples that have small margins and on classifiers that correctly classify these challenging examples [91]. The procedure begins by defining the

Margin & Diversity based Measure (MDM) for each classifier as:

$$MDM(h,H) = \sum_{\mathbf{x}_i \in D_{pr}} \Big[I(h(\mathbf{x}_i) = y_i) \cdot \Big(\alpha f_m(\mathbf{x}_i) + (1-\alpha) f_d(h,\mathbf{x}_i) \Big) \Big]$$

In this formulation, the parameter α controls the balance between margin and diversity. The term $f_m(\mathbf{x}_i)$ is defined as $\log(|margin(\mathbf{x}_i)|)$, where $margin(\mathbf{x}_i)$ is calculated as $\frac{v_{y_i}^{(i)} - v_{\tilde{y}_i}^{(i)}}{M}$, with $v_{y_i}^{(i)}$ representing the number of votes for the correct class and $v_{\tilde{y}_i}^{(i)}$ denoting the number of votes for the most popular incorrect class. The diversity contribution is captured by $f_d(h, \mathbf{x}_i)$, defined as $\log\left(\frac{v_{y_i}^{(i)}}{M}\right)$. After assigning MDM scores to each classifier, they are ranked based on these scores, and the top T classifiers are selected to form the pruned ensemble [91].

Other Ordering-Based Methods: Several additional ordering-based methods have been proposed in the literature. Orientation Ordering (OO) ranks classifiers based on the angle between their signature vectors and a reference vector [91], [92]. The Unsupervised Margin based Ordering Ensemble Pruning (UMEP) method orders classifiers according to an unsupervised ensemble margin criterion, with a particular emphasis on low-margin examples [91]. Another approach, known as Kappa Pruning, orders pairs of classifiers based on their diversity as measured by the kappa statistic and selects the most diverse subset [91], [93]. Finally, Reduce-Error Pruning ranks classifier subsets based on their voted performance on a pruning set and selects the best-performing subset for the final ensemble [93].

Key Variations in Literature. The primary variations in rank pruning methods lie in the **choice of the evaluation metric** $E(h_m)$ used to rank classifiers [91], [92]. One commonly used approach involves **accuracy-based metrics**, where classifiers are ranked based on their individual accuracy on training, validation, or pruning sets. Another important variation involves **diversity-based metrics**, which rank classifiers according to their diversity relative to other ensemble members. Measures such as the kappa statistic or disagreement metrics are often employed for this purpose [91], [92].

Additionally, **margin-based metrics** are used to rank classifiers based on their contribution to the ensemble margin, placing greater emphasis on examples with small margins. There are also **contribution-based metrics**, where classifiers are ranked using heuristic measures that consider both accuracy and diversity to assess each member's overall contribution [91]. Finally, some methods use **combined metrics** that simultaneously integrate accuracy, diversity, and margin considerations to produce a more comprehensive ranking of classifiers [91], [92].

In summary, rank pruning is a general ensemble pruning framework that evaluates and **ranks** classifiers based on various metrics to identify the most valuable members. The effectiveness of this method heavily depends on the chosen evaluation metric and its ability to capture the contribution of each classifier [91], [92]. The practical implementation of rank pruning used in this work is presented in Algorithm 4.

Algorithm 4 Rank Pruning Algorithm

```
Require: Number of classifiers ($M$), evaluation metric ($EvaluateEstimator$), selection fraction or number ($TopK$)
Compute evaluation scores for each classifier using $EvaluateEstimator$
Rank classifiers based on their evaluation scores in descending order
Select the top $TopK$ classifiers from the ranked list
Return pruned ensemble of selected classifiers
```

3.5 Statistical Testing and Visualization Tools

In this study, statistical analyses are performed to assess the significance and magnitude of the observed effects and to better understand how various factors influence fairness metrics and accuracy. These analyses are conducted using a combination of hypothesis testing, permutation-based ANOVA, and visualization tools.

For pairwise comparisons of models under identical conditions, we use the Wilcoxon signed-rank test when the metric distributions deviate from normality, and paired t-tests when normality assumptions are satisfied. These tests are implemented using the `scipy.stats` library in Python, with the functions `ttest_rel` and `wilcoxon`. To quantify the effect size in non-parametric comparisons, we calculate the commonly used effect size "r", using the formula:

$$r = \frac{Z}{\sqrt{N}}$$

where Z is the Wilcoxon test statistic and N is the total number of observations, providing an interpretable measure of the strength of observed effects.

For multi-factor analyses, permutation-based ANOVA is applied to examine the impact of factors such as pruning level, dataset, and pruning method on fairness metrics and accuracy. This approach is chosen for its robustness against violations of normality and homoscedasticity assumptions, relying on resampling to obtain reliable significance testing [94]. The analyses are performed using the `statsmodels` library in Python, specifically the modules `statsmodels.formula.api`, `statsmodels.api`, and the `anova_lm` function. The pruning percentage is treated as a categorical factor to capture potential non-linear effects and complex interaction patterns. The general factorial ANOVA model structure is given by:

$$Metric\_Value \sim C(Factor_1) \times C(Factor_2) \times \cdots \times C(Factor_n)$$

where each factor is modeled as categorical, and main effects, two-way interactions, and higher-order interactions are considered as needed.

Finally, for visualization, Tableau is used for interactive exploration and graphical presentation of results, while Python is employed for implementing statistical analyses and generating custom plots.

CHAPTER 4 Experimental Design and Analytical Framework

The experimental design in this study systematically examines the impact of various pruning techniques on the fairness and predictive performance of Random Forest classifiers. Using datasets from the Folktables repository, controlled experiments are conducted to assess how pruning influences fairness metrics and model accuracy. Each experiment follows a standardized workflow, ensuring consistent dataset splits, model training, and evaluation procedures. To isolate the effects of pruning, all hyperparameters remain fixed while pruning strategies and levels are systematically varied. The following sections describe the research questions, datasets, fairness metrics, pruning methods, and evaluation framework applied in this study.

4.1 Research Questions

This study investigates the impact of pruning on fairness metrics and predictive performance. Two central research questions guide the analysis:

- **Question 1:** How do different pruning methods impact fairness metrics and model performance compared to both the baseline (non-pruned) model and the random pruning model, under the same dataset and experimental setup across varying pruning levels?
- **Question 2:** Do different pruning methods demonstrate varying fairness and performance outcomes across different pruning levels within the same dataset and experimental configuration?

Each research question is evaluated across multiple datasets using the experimental configuration outlined in the following sections. Pruned models are tested at varying levels to assess how pruning influences fairness and predictive performance. Statistical hypothesis testing is applied to ensure that observed differences are significant and not due to random variation.

4.2 Datasets

The datasets used in this work include the *ACS Income* dataset, which focuses on income prediction, and the *ACS Travel Time* dataset, which focuses on commute time prediction. The dataset selection for this study is based on three main criteria. First, each dataset has a tabular structure compatible with Random Forest classifiers. Second, the datasets have a sufficient sample size to support robust training and statistical evaluation. Third, they exhibit clear evidence of systematic bias against defined groups, which allows for the assessment of fairness interventions.

To address the last criterion, we use the Fair Fairness Benchmark (FFB), a PyTorch-based framework for evaluating fairness in machine learning [95]. It provides standardized metrics and algorithms with flexibility for customization. The FFB highlights that many widely used fairness datasets lack consistent, measurable bias, making careful dataset selection essential for reliable and meaningful results. In their work, the authors analyze various datasets, assessing which ones meet these critical conditions.

Based on their analysis and the criteria defined for this study, we selected the ACS Income and ACS Travel Time datasets. These datasets offer large sample sizes, broader demographic coverage, and improved data quality compared to other common datasets like UCI Adult or KDD Census. Additionally, they provide diversity in application domains and have demonstrated potential to exhibit fairness-related bias, enabling a comprehensive assessment of fairness interventions across distinct tasks.

Following the dataset selection process and criteria, we describe the specific tasks and configurations used in this study:

ACS Income. The task involves predicting whether a working adult's annual income exceeds $50,000. The target variable is PINCP (Total personal income), which is labeled as 1 if PINCP is greater than 50,000, and 0 otherwise. Additionally, the income threshold is configurable through the provided experimental framework, allowing for alternative prediction tasks.

ACS Travel Time. The task involves predicting whether a working adult's commute time exceeds 20 minutes. The target variable is JWMNP (Travel time to work), which is labeled as 1 if JWMNP is greater than 20, and 0 otherwise.

Both datasets span multiple years and cover all U.S. states and Puerto Rico, from 2014 to 2018. This results in 255 distinct datasets per task, allowing the study of distribution shifts across time and geography, and enabling robust fairness evaluation under varying conditions.To introduce variability and ensure that findings are not limited to a single geographic or temporal setting, different configurations of states and years were selected for each task. These configurations are shown in Table 4.1.

These selections allow for controlled comparisons of fairness and performance under different temporal and regional conditions.

Choice of Sensitive Attributes. The sensitive attribute selected for fairness evaluation in this study is "Sex" which is a binary attribute with values 1 (male) and 2 (female). This

Table 4.1 Dataset configurations for each task

Task	Year	Selected States
ACS Income	2018	New York (NY), California (CA)
ACS Travel Time	2014	Washington (WA), Oregon (OR)

attribute is explicitly defined in the Folktables dataset and is commonly used in fairness research.

The choice of sex as the primary sensitive attribute is motivated by its widespread use in fairness studies and its documented role in historical inequalities across socio-economic outcomes. Furthermore, the consistent availability and quality of this attribute in the Folktables dataset ensure reliability and comparability in fairness assessments.

Primary Dataset Characteristics. Table 4.2 presents the initial distributions of the sensitive attribute (sex) for each dataset, showing the proportions of male and female instances prior to any pruning or modification, as well as their distribution across class labels. These baseline distributions serve as reference points for subsequent fairness assessments, offering insight into the relationship between gender and predicted outcomes and establishing the initial fairness conditions against which pruning interventions are evaluated.

Table 4.2 Initial distribution of the sensitive attribute (sex) and distribution by class label for each Dataset

Dataset	Sex	Initial Ditribution(%)	Class = False (%)	Class = True (%)
ACS Income	Male	52.06	47.15	59.06
	Female	47.94	52.85	40.94
ACS Travel Time	Male	52.35	49.40	56.68
	Female	47.65	50.60	43.32

An in-depth analysis of the datasets and their statistical characteristics is provided in Section 5.

4.3 Configurations of Pruning Methods and Implementation Details

To make fair and clear comparisons, this study uses key methods from each type of ensemble pruning described in the methodology. We focus on well-established and widely adopted approaches that serve as standard baselines in the field. This selection facilitates a systematic evaluation of commonly used pruning strategies and establishes a solid foundation for future research into more advanced or adaptive pruning methods. A consistent baseline configuration is used to enable fair comparisons among pruning strategies. This ensures that no model gains an inherent advantage and prevents bias toward accuracy at the expense of diversity. Each pruning method is applied under standardized conditions, allowing objective evaluation of its impact on ensemble composition, predictive performance, and fairness.The methodology section (Section 3) provides a detailed explanation of each method. In this section, we focus only on the specific configurations used in our experiments.

Cluster Pruning We employ a probability-based representation to ensure a neutral view of classifier behavior, avoiding bias toward accuracy. K-means clustering is used for its scalability and efficiency in grouping similar classifiers. From each cluster, we select the most accurate representative to maintain predictive strength while preserving diversity, ensuring that pruning decisions are not solely accuracy-driven.

Rank Pruning For rank pruning, we use negative AUC as the ranking criterion, as it captures both predictive strength and variation among classifiers. While alternatives such as the Kappa statistic were considered, they proved less interpretable and less effective in preserving diversity. Ranking by individual error was deliberately avoided to prevent excessive bias toward accuracy. By relying on negative AUC, the ranking process balances predictive performance with ensemble complementarity.

Greedy Pruning For greedy pruning, we use the complementariness metric for classifier selection to ensure that each chosen classifier contributes unique predictive patterns rather than simply minimizing error. Alternative metrics, such as reduced error, were considered but found to introduce excessive bias toward accuracy, which conflicts with the fairness objectives of this study. By using complementariness, we support the construction of an ensemble that balances predictive strength with diversity.

Mixed-Integer Quadratic Programming (MIQP) Pruning For MIQP pruning, we use a pairwise error metric that captures both individual classifier performance and the degree of error overlap among models. This approach prevents the selection of classifiers that make similar mistakes, thereby promoting ensemble diversity. The weighting parameter α is set to 1 to avoid placing excessive emphasis on individual error, as it is already reflected in the combined error measure. This configuration ensures that MIQP pruning maintains a balanced selection process, reducing collective error while preserving diversity.

In addition to these methods, this study includes **Random Pruning**, which serves as a baseline approach. In this method, trees are removed at random, providing a control reference for evaluating the effectiveness of structured pruning techniques.

All pruning methods are implemented using an open-source GitHub repository to ensure reproducibility and transparency.

4.4 Evaluation Metrics

Fairness Metrics. Fairness in machine learning is inherently complex, with multiple definitions that often conflict depending on the context and goals of the analysis. The Fairness Impossibility Theorem states that certain fairness criteria cannot be satisfied simultaneously, making trade-offs inevitable. In light of this, our study adopts a set of fairness metrics that reflect distinct, sometimes competing, fairness perspectives. This approach allows for a comprehensive evaluation of how pruning affects different dimensions of fairness.

The selection of fairness metrics in this study is guided by two key considerations. First, dataset-specific bias patterns play an important role, as highlighted in FFB [95], where it is noted that datasets can exhibit different levels of bias depending on the metric chosen, thereby directly influencing fairness assessments. Second, the Fairness Impossibility Theorem [96] demonstrates that demographic parity, equalized odds, and predictive parity cannot all be satisfied simultaneously by a well-calibrated classifier. This highlights the need to acknowledge inherent trade-offs when evaluating fairness.

Based on the principles of fairness, three fairness metrics are selected for this study. In the following formulations, we use these notations. We denote $\hat{Y}$ as the predicted outcome, with $\hat{Y} = 1$ indicating a positive prediction and $\hat{Y} = 0$ indicating a negative prediction. The true outcome is represented by Y, where $Y = 1$ denotes a positive instance and $Y = 0$ a negative one. The sensitive attribute is denoted by A, with $A = 1$ referring to the privileged group and $A = 0$ to the unprivileged group. All probabilities are computed over the evaluation dataset.

- **Demographic Parity:** Measures whether different demographic groups receive positive predictions at the same rate, regardless of the actual outcomes. This metric captures the influence of group membership on predicted outcomes. A value close to zero indicates that outcomes are evenly distributed across groups.

$$DP = P(\hat{Y} = 1 \mid A = 1) - P(\hat{Y} = 1 \mid A = 0) \tag{4.1}$$

- **Predictive Parity:** Assesses whether the probability of a correct positive prediction is consistent across groups. This reflects dependence on predicted outcomes and the reliability of positive predictions for different demographic groups.A value close to zero suggests parity in predictive reliability across demographic groups.

$$PP = P(Y = 1 \mid \hat{Y} = 1, A = 1) - P(Y = 1 \mid \hat{Y} = 1, A = 0) \tag{4.2}$$

- **Equalized Odds:** Evaluates whether false positive and false negative rates are equally distributed across groups, capturing the dependence on true outcomes and ensuring balanced error rates.It is computed as the sum of absolute differences in these rates.Lower values indicate more balanced error rates across groups.

$$EO = |P(\hat{Y} = 1 \mid Y = 1, A = 1) - P(\hat{Y} = 1 \mid Y = 1, A = 0)| \tag{4.3}$$

$$+|P(\hat{Y} = 1 \mid Y = 0, A = 1) - P(\hat{Y} = 1 \mid Y = 0, A = 0)| \tag{4.4}$$

Together, these metrics represent three key dimensions of fairness: dependence on predicted outcomes (Predictive Parity), dependence on true outcomes (Equalized Odds), and dependence on group membership (Demographic Parity). This selection ensures a comprehensive and multidimensional assessment of fairness across models and pruning configurations. By including these metrics, the study ensures a balanced and robust assessment of fairness, reflecting the inherent trade-offs and complexities present in real-world decision-making models.

Performance Metrics. Model performance is evaluated using the Accuracy metric which is The proportion of correctly classified instances. Accuracy serves as the primary measure of predictive performance in this study, providing a clear and interpretable assessment of model effectiveness across different pruning levels and datasets.

4.5 Experiment Setup

The baseline model used in this study is a Random Forest classifier, implemented using the Scikit-learn library. Bootstrap sampling is enabled in the model, which generates multiple random subsets of the training data to train different decision trees. This process improves robustness and helps mitigate overfitting by ensuring that each tree is trained on a different subset of data. Additionally, we set the random seed (`random_state=42`) to ensure that the model's random processes, such as data splitting and random initialization, are consistent across runs, making the training process reproducible.

The model configuration consists of 100 estimators. Each internal node requires a minimum of 10 samples to split, and bootstrap sampling is enabled. All other parameters are set to their default values. By default, the Scikit-learn Random Forest classifier (`RandomForestClassifier`) uses the Gini impurity for node splitting (criterion="gini") and considers all features when splitting each node (`max_features="auto"`).

This configuration strikes a balance between model complexity and generalization, ensuring stable and interpretable baseline performance. The same setup is applied consistently across all datasets and experimental runs to isolate the effects of pruning and other experimental factors. The model is first trained on the designated training set, after which pruning techniques are applied. Evaluation is then conducted on the test set to assess the impact of pruning on both predictive performance and fairness metrics.

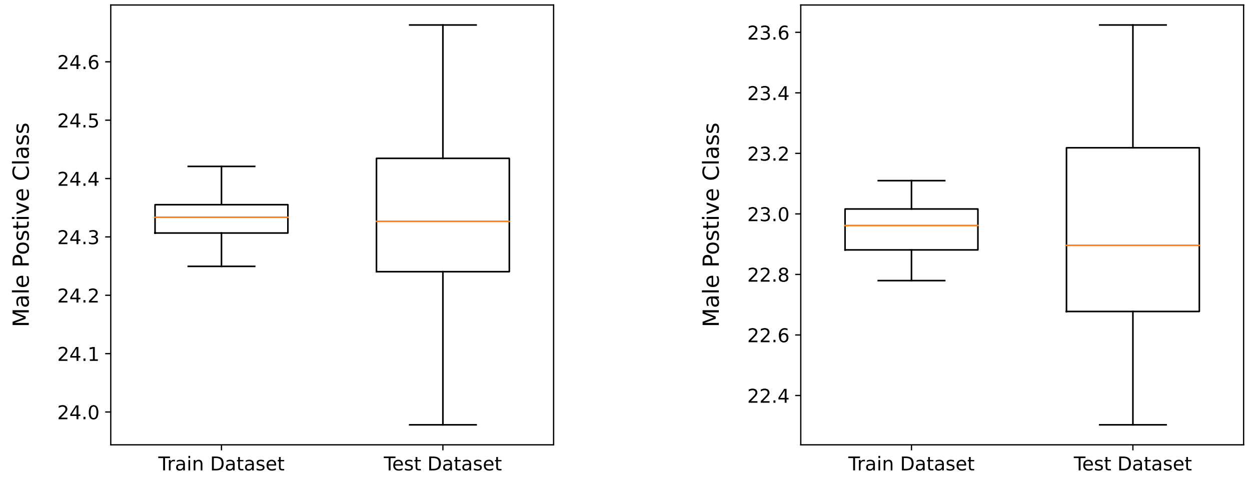
Train-Test Splits, Cross-Validation and Repetition Setup. A 5-fold cross-validation with 10 repetitions is employed to ensure statistical robustness and result stability. Each dataset is divided into five folds, with each fold consisting of a 75% training set and a 25% test set. This entire process is repeated 10 times using different random seeds, resulting in 50 unique train-test splits for each dataset. In each split, the distribution of sensitive attributes and target labels is carefully preserved to ensure that both the training and test sets remain representative of the initial dataset.

The choice of 50 experimental runs is designed to balance statistical reliability with computational feasibility. This decision is based on several considerations. First, conducting 50 independent experiments ensures statistical robustness by reducing the likelihood that results are influenced by any single random partition; performance and fairness metrics are averaged across diverse splits, capturing consistent trends. Second, this number of repetitions achieves an effective trade-off between computational cost and reliability, as increasing the number of runs would only marginally reduce variance while significantly increasing resource demands. Finally, empirical observations confirm that with 50 experiments, patterns in fairness and performance metrics are stable and not artifacts of specific dataset divisions.

This experimental setup provides a statistically sound and reproducible foundation for evaluating the impact of pruning techniques on model fairness and predictive accuracy.

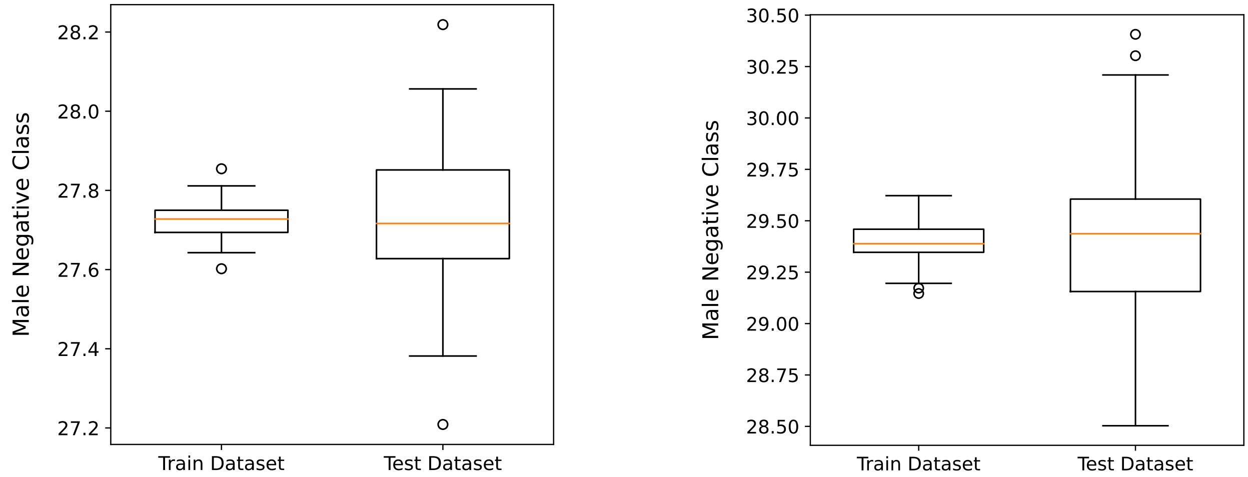
Distribution Preservation in Splits. For each train-test batch used in our 50-run experiment, we carefully maintained the ratios and percentages stated in Table 4.2. This ensures that without altering the initial characteristics of the datasets, we can analyze the pure effect of pruning models on fairness. We observed some outlier batches in both test and train datasets, but they were retained to avoid manipulating batch randomness.

The boxplots below illustrate the distribution of the sensitive attribute over 50 runs of train-test batches for each dataset.



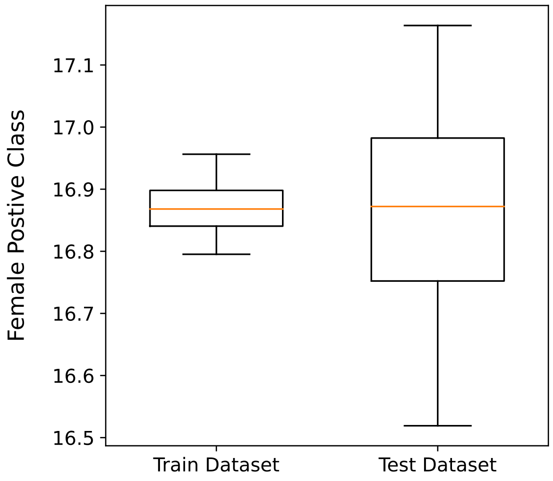
(a) Male Positive - Income Dataset

(b) Male Positive - Travel Time Dataset

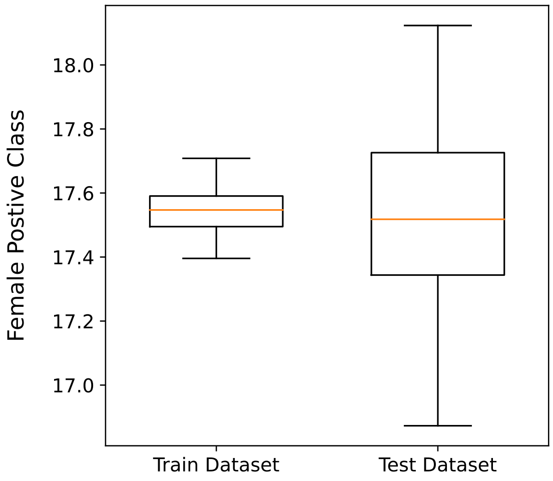


(a) Male Negative - Income Dataset

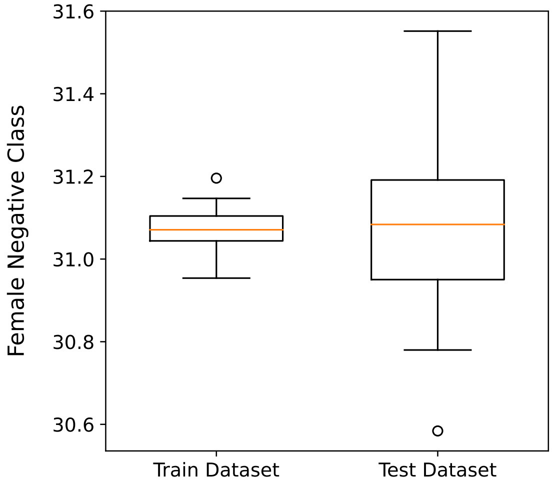
(b) Male Negative - Travel Time Dataset



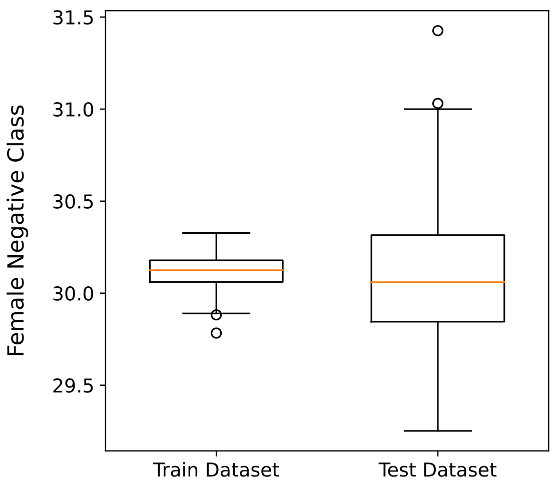
(a) Female Positive - Income Dataset



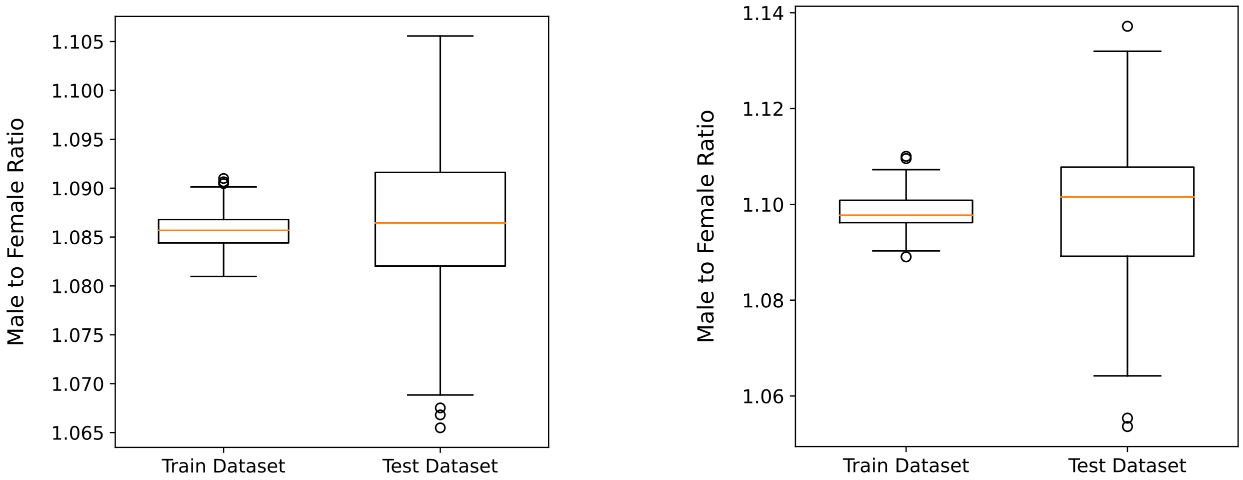
(b) Female Positive - Travel Time Dataset



(a) Female Negative - Income Dataset



(b) Female Negative - Travel Time Dataset



(a) Male-to-Female Ratio - Income Dataset

(b) Male-to-Female Ratio - Travel Time Dataset

Figure 4.5 Boxplots illustrating the distribution of sensitive attributes across 50 runs for the Income and Travel datasets.

Ensuring Experimental Consistency. To ensure fair and controlled comparisons across all experiments, several measures are implemented. All models, regardless of pruning method or configuration, are evaluated on the same 50 dataset splits. This guarantees direct comparability and eliminates variability introduced by differing data partitions. The same baseline Random Forest model configuration is applied across all pruning techniques, ensuring that observed differences are attributable solely to the pruning process. Furthermore, pairwise statistical analyses are conducted on results from identical splits to confirm that variations in fairness and performance metrics arise directly from pruning interventions rather than random fluctuations.

4.6 Statistical Analysis

Statistical hypothesis testing is employed to evaluate the significance of differences between pruning methods, with test selection guided by data properties and comparison structure. The primary test used is the Wilcoxon Signed-Rank Test, chosen for its robustness and applicability to both normally and non-normally distributed paired data. Although pairwise t-tests were also conducted for normally distributed results, the Wilcoxon test was selected as the main comparison method to maintain consistency and comparability across all scenarios.

A comparison between the t-test and the Wilcoxon test showed five mismatches across all scenarios, where the Wilcoxon test detected significant differences while the t-test did not.

All tests are conducted at a significance level of $p < 0.05$.

Effect Size Calculation. To assess the practical relevance of statistically significant differences, the commonly used effect size r is calculated as the standardized test statistic z divided by the square root of the sample size n, i.e.,

$$r = \frac{z}{\sqrt{n}}$$

The interpretation thresholds for r are:

r Value	Interpretation
0.1 to 0.3	Small effect
0.3 to 0.5	Medium effect
> 0.5	Large effect

Table 4.3 Interpretation guidelines for Wilcoxon effect size

The Wilcoxon Signed-Rank Test, coupled with effect size measurement, provides a robust, non-parametric framework for evaluating the statistical and practical significance of pruning's impact on fairness metrics. These methods ensure that reported differences are not only statistically valid but also practically meaningful, reflecting real shifts in fairness beyond random variation or experimental noise.

CHAPTER 5 Characterization and Initial Analysis

5.1 Exploratory Dataset Analysis

Before analyzing the results, we need a solid understanding of our datasets and their inherent characteristics. Below is a detailed analysis of the two datasets and the tasks we worked on in this model.

Income Dataset. This dataset consists of 11 columns and 298,686 rows. All features are recorded as float types, and the class column is a boolean indicating whether a person's annual income exceeds $50,000. The features are described as follows:

- **AGEP (Age):** Integer values from 0 to 99, where 0 indicates less than one year old.
- **COW (Class of Worker):** Employment type, including private (for-profit or non-profit), various government positions, self-employed (incorporated or not), unpaid family workers, and unemployed individuals.
- **SCHL (Educational Attainment):** Ranges from 1 (no schooling) to 24 (doctorate degree). Categories: 1–15 for no schooling through 12th grade without diploma, 16 for high school diploma, 17 for GED or equivalent, 18–19 for some college without a degree, and 20–24 for associate's, bachelor's, master's, professional, or doctorate degrees.
- **MAR (Marital Status):** Includes married, widowed, divorced, separated, or never married/under 15 years old.
- **OCCP (Occupation)** and **POBP (Place of Birth):** Follow standard coding as per ACS PUMS documentation.
- **RELP (Relationship):** Defines relationships to the reference person, including spouse, child (biological or adopted), stepchild, sibling, parent, in-law, grandchild, roommate, or nonrelative.
- **WKHP (Usual Hours Worked Per Week):** Records weekly work hours from 1 to 98 (with 99 indicating 99+ hours). Not applicable for individuals under 16 or those who did not work.
- **SEX (Sex):** Binary variable where 1 indicates male and 2 indicates female.

- **RAC1P (Race):** Categorized into nine groups: White alone, Black or African American alone, American Indian, Alaska Native, Asian, Pacific Islander, Other Race, and Two or More Races.

The correlation matrix heatmap 5.1 reveals that the features most associated with the class outcome are **SCHL (education level)** and **WKHP (weekly working hours)**, both showing a moderate positive correlation of ~0.34. **AGEP (age)** also show moderate positive correlation(0.26). In contrast, **OCCP (occupation)** and **MAR (marital status)** demonstrate negative correlations with the class outcome (-0.32 and -0.26, respectively). The **SEX** feature shows a weak negative correlation of -0.12, indicating a negligible relationship with the class. Furthermore, notable correlations between features include a strong negative correlation between age and marital status (-0.49), and moderate associations between marital status and relationship status (0.38), as well as occupation and education level (-0.39). These findings suggest that educational attainment, working hours, and demographic attributes are key factors linked to the target class.

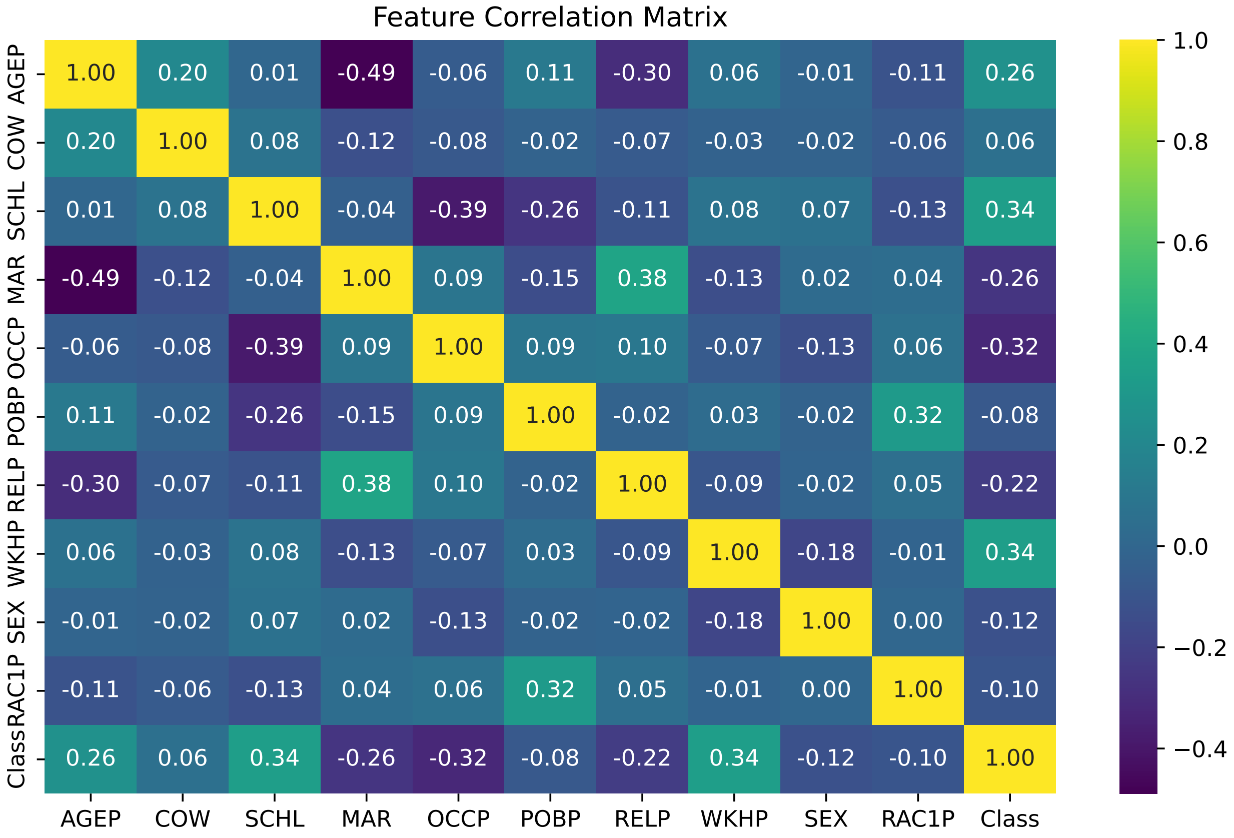


Figure 5.1 Correlation matrix for the Income dataset, illustrating correlations between dataset features

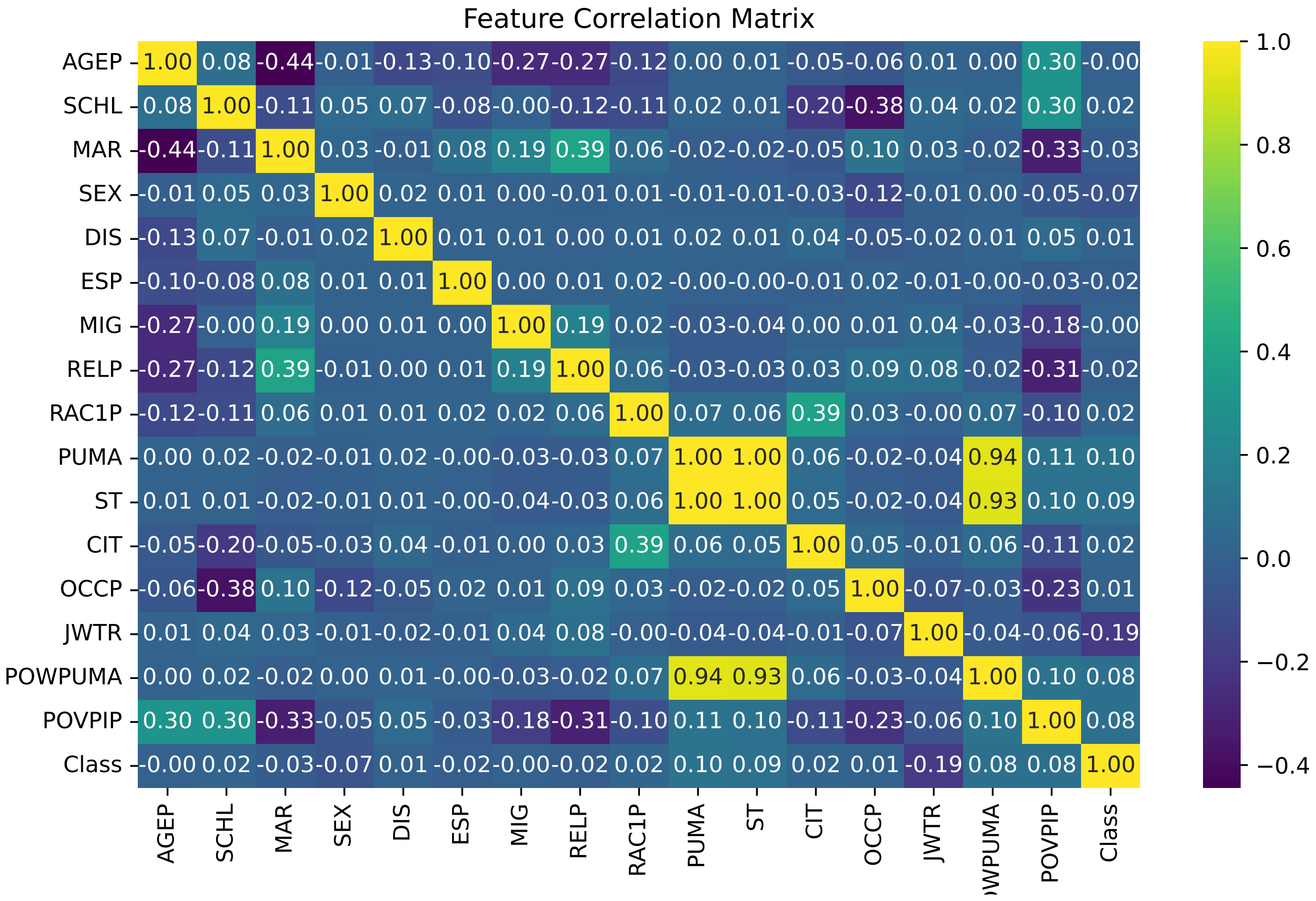
Travel Time Dataset. This dataset contains 17 features, all recorded as float types, with a class label indicating whether a person's travel time from home to their destination exceeds 20 minutes.

The features included in the travel time dataset are as follows:

- **AGEP (Age):** Integer values from 0 to 99, where 0 indicates less than one year old.
- **SCHL (Educational Attainment):** Follows the same classification as in the income dataset, ranging from no schooling to a doctorate degree.
- **MAR (Marital Status):** Categories include married, widowed, divorced, separated, and never married individuals.

- **SEX (Sex):** A binary feature indicating male or female.
- **DIS (Disability Recode):** Indicates the presence or absence of a disability.
- **ESP (Employment Status of Parents):** Captures parental employment status, ranging from both parents in the labor force to only one or neither parent working.
- **MIG (Mobility Status):** Indicates whether individuals lived in the same house one year ago, moved within the U.S./Puerto Rico, or moved from outside the U.S.
- **RELP (Relationship):** Describes the individual's relationship to the reference person, including categories for family members, in-laws, roommates, and nonrelatives.
- **RAC1P (Race):** Categorizes individuals into groups such as White, Black, American Indian, Asian, Pacific Islander, and multiracial individuals, following ACS PUMS standards.
- **PUMA (Public Use Microdata Area)** and **ST (State Code):** Geographic codes based on Census definitions.
- **CIT (Citizenship Status):** Distinguishes between U.S.-born citizens, citizens born in U.S. territories, naturalized citizens, and non-citizens.
- **OCCP (Occupation):** Employment categories defined by ACS documentation.
- **JWTR (Means of Transportation to Work):** Includes modes such as car, bus, subway, ferry, motorcycle, bicycle, walking, working from home, and other methods.
- **POWPUMA (Place of Work PUMA):** Coded by Census region definitions.
- **POVPIP (Income-to-Poverty Ratio Recode):** Records ratios from 0 to 500, with a value of 501 indicating 501% or higher.

The correlation analysis based on heatmap 5.2 shows that most features have weak correlations with the class outcome. The feature with the strongest positive relationship to the class is **PUMA (Public Use Microdata Area)** with ~0.10. Other features such as **REL (relationship status)** and **SCHL (education level)** show very weak positive correlations of around 0.03 and 0.02, respectively. **SEX** exhibits an almost negligible negative correlation of 0.07 with the class outcome. Significant correlations among features include **AGEP and MAR** with a strong negative correlation of -0.44, **SCHL and OCCP** with -0.38, and **REL and MAR** with a positive correlation of 0.39. Overall, relationships between features and the class remain weak, but strong inter-feature correlations suggest multicollinearity considerations in modeling.



	AGEP	SCHL	MAR	SEX	DIS	ESP	MIG	RELP	RAC1P	PUMA	ST	CIT	OCCP	JWTR	POWPUMA	POVPIP	Class
AGEP	1.00	0.08	-0.44	-0.01	-0.13	-0.10	-0.27	-0.27	-0.12	0.00	0.01	-0.05	-0.06	0.01	0.00	0.30	-0.00
SCHL	0.08	1.00	-0.11	0.05	0.07	-0.08	-0.00	-0.12	-0.11	0.02	0.01	-0.20	-0.38	0.04	0.02	0.30	0.02
MAR	-0.44	-0.11	1.00	0.03	-0.01	0.08	0.19	0.39	0.06	-0.02	-0.02	-0.05	0.10	0.03	-0.02	-0.33	-0.03
SEX	-0.01	0.05	0.03	1.00	0.02	0.01	0.00	-0.01	0.01	-0.01	-0.01	-0.03	-0.12	-0.01	0.00	-0.05	-0.07
DIS	-0.13	0.07	-0.01	0.02	1.00	0.01	0.01	0.00	0.01	0.02	0.01	0.04	-0.05	-0.02	0.01	0.05	0.01
ESP	-0.10	-0.08	0.08	0.01	0.01	1.00	0.00	0.01	0.02	-0.00	-0.00	-0.01	0.02	-0.01	-0.00	-0.03	-0.02
MIG	-0.27	-0.00	0.19	0.00	0.01	0.00	1.00	0.19	0.02	-0.03	-0.04	0.00	0.01	0.04	-0.03	-0.18	-0.00
RELP	-0.27	-0.12	0.39	-0.01	0.00	0.01	0.19	1.00	0.06	-0.03	-0.03	0.03	0.09	0.08	-0.02	-0.31	-0.02
RAC1P	-0.12	-0.11	0.06	0.01	0.01	0.02	0.02	0.06	1.00	0.07	0.06	0.39	0.03	-0.00	0.07	-0.10	0.02
PUMA	0.00	0.02	-0.02	-0.01	0.02	-0.00	-0.03	-0.03	0.07	1.00	1.00	0.06	-0.02	-0.04	0.94	0.11	0.10
ST	0.01	0.01	-0.02	-0.01	0.01	-0.00	-0.04	-0.03	0.06	1.00	1.00	0.05	-0.02	-0.04	0.93	0.10	0.09
CIT	-0.05	-0.20	-0.05	-0.03	0.04	-0.01	0.00	0.03	0.39	0.06	0.05	1.00	0.05	-0.01	0.06	-0.11	0.02
OCCP	-0.06	-0.38	0.10	-0.12	-0.05	0.02	0.01	0.09	0.03	-0.02	-0.02	0.05	1.00	-0.07	-0.03	-0.23	0.01
JWTR	0.01	0.04	0.03	-0.01	-0.02	-0.01	0.04	0.08	-0.00	-0.04	-0.04	-0.01	-0.07	1.00	-0.04	-0.06	-0.19
POWPUMA	0.00	0.02	-0.02	0.00	0.01	-0.00	-0.03	-0.02	0.07	0.94	0.93	0.06	-0.03	-0.04	1.00	0.10	0.08
POVPIP	0.30	0.30	-0.33	-0.05	0.05	-0.03	-0.18	-0.31	-0.10	0.11	0.10	-0.11	-0.23	-0.06	0.10	1.00	0.08
Class	-0.00	0.02	-0.03	-0.07	0.01	-0.02	-0.00	-0.02	0.02	0.10	0.09	0.02	0.01	-0.19	0.08	0.08	1.00

Figure 5.2 Correlation matrix for the Travel Time dataset, illustrating correlation between dataset features

5.1.1 Comparison of Feature Correlation Analysis Between Two Datasets

The correlation matrices from the two datasets present notable differences. In the first dataset, the most significant positive correlations with the class were observed for **SCHL (education level)** and **WKHP (weekly working hours)** at approximately 0.34, with **AGEP (age)** also contributing moderately. In contrast, the second dataset shows very weak correlations between features and the class. The strongest relationship observed is with **JWTR (Means of Transportation to work)** at only -0.19, followed by **PUMA** at 0.10, indicating a weaker predictive relationship.

Additionally, the first dataset displays clearer strong inter-feature correlations, such as **AGEP and MAR** (-0.49) and **REL and MAR** (0.38). The second dataset also exhibits strong feature-to-feature relationships, highlighting potential multicollinearity.

Overall, the first dataset presents more meaningful and stronger feature relationships with the target class, while the second dataset is characterized by weaker associations and stronger inter-feature redundancy. Models trained on the income dataset would be simpler, more interpretable, and possibly more robust. We consider the income dataset more stable, allowing for generalized insights. The travel time dataset is more biased, sensitive, and vulnerable, amplifying the fairness aspect of our experiment in unstable scenarios. We will leverage these initial differences and characteristics in our insights.

5.1.2 Initial Evaluation of Fairness and Accuracy — Unpruned Base Model

Spider chart 5.3 illustrate the relative comparison of the computed fairness metrics for each dataset, based on the results of 50 runs of the Base Model on the test set prior to pruning. The base random forest model with 100 trees exhibits greater bias in predictive parity for the income dataset. In contrast, the travel time dataset displays more pronounced bias in demographic parity and equalized odds.

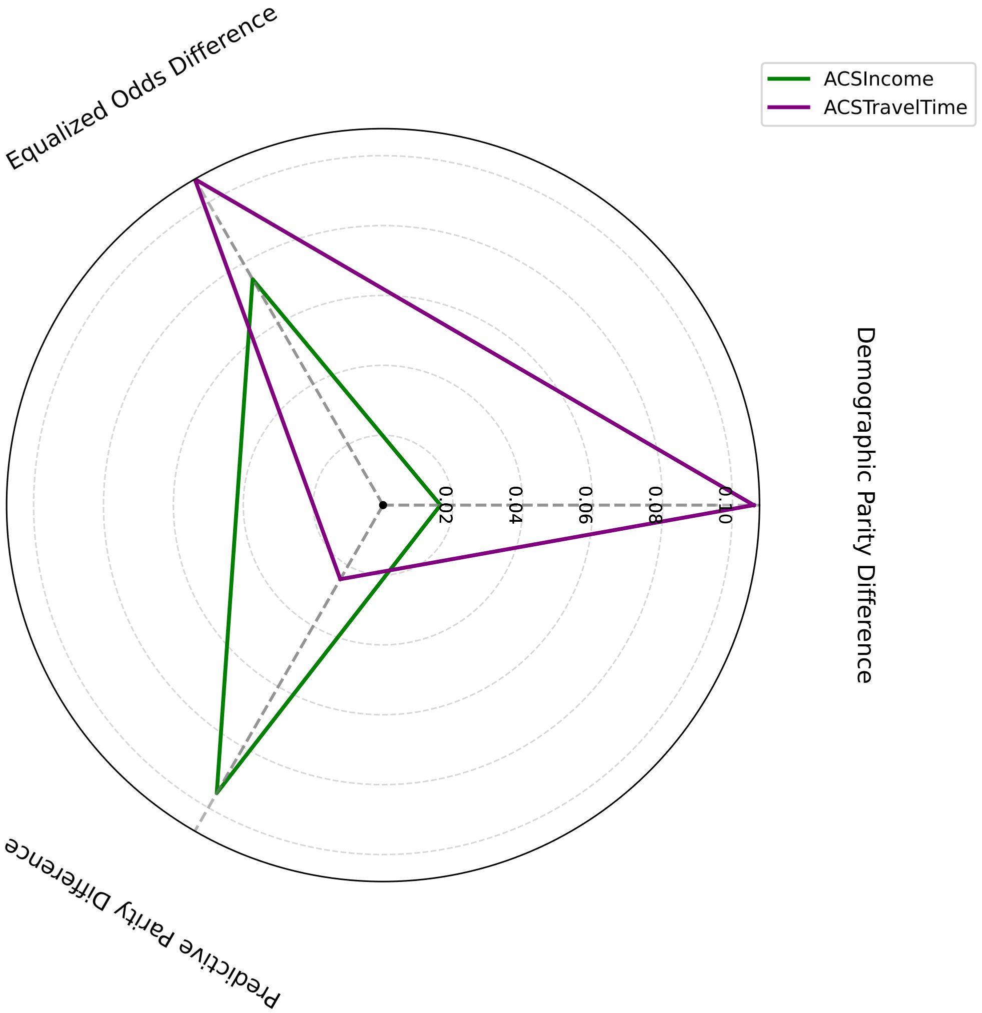


Figure 5.3 Spider Chart illustrating the relative relationships between fairness metrics for the baseline (non-pruned) model. Each color represents one dataset (Income and Travel Time), highlighting the inherent fairness metric profiles before applying any pruning.

Also, the boxplots 5.4 and 5.5 for each metric for the Base model show that during 50 different runs with 50 different train-test validations, the distribution of each metric is more even and less varied in the income dataset. This indicates that for the Income Dataset, there is a lower median and tighter spread in Equalized Odds Difference and Demographic Parity Difference, making it more stable in these metrics. In contrast, the Travel Time Dataset shows a higher median and greater variance, indicating higher bias in terms of demographic parity and equalized odds. For predictive parity, the Income Dataset has a higher median value, while the Travel Time Dataset has a lower median with wider variance, suggesting inconsistency across different samples that could affect pruning models' performance regarding fairness. Additionally, the Travel Time Dataset exhibits lower accuracy (approximately

0.65), higher fairness disparities, and greater variability in fairness metrics, all pointing to significant prediction variation and potential inconsistency due to demographic imbalance and skewed data distribution. The Income Dataset, on the other hand, maintains higher accuracy (approximately 0.75) with a tight distribution. In comparison, the Travel Time Dataset has a lower median accuracy (approximately 0.65) with broader variance, making it more challenging for models to balance accuracy and fairness.

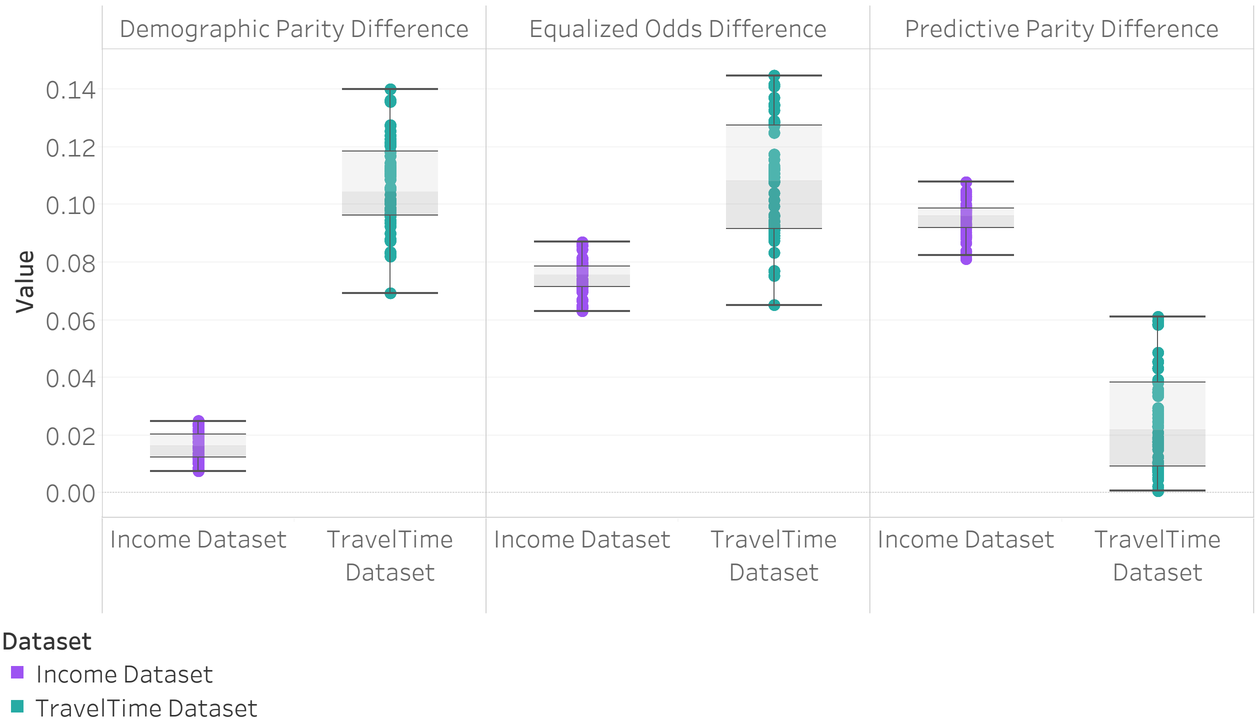


Figure 5.4 Box plot showing the distribution of fairness metrics for the baseline (non-pruned) model across both datasets. Each panel represents a specific fairness metric, and each color corresponds to one dataset (Income or Travel Time), highlighting the variability and differences in fairness metrics between the two datasets.

In summary, these observations collectively highlight that the Travel Time Dataset exhibits worse fairness metrics (higher Equalized Odds Difference and Demographic Parity differences), indicating more bias. Predictive Parity Difference is lower for the Travel Time Dataset but accompanied by high variance, reducing its fairness reliability. The Travel Time Dataset's

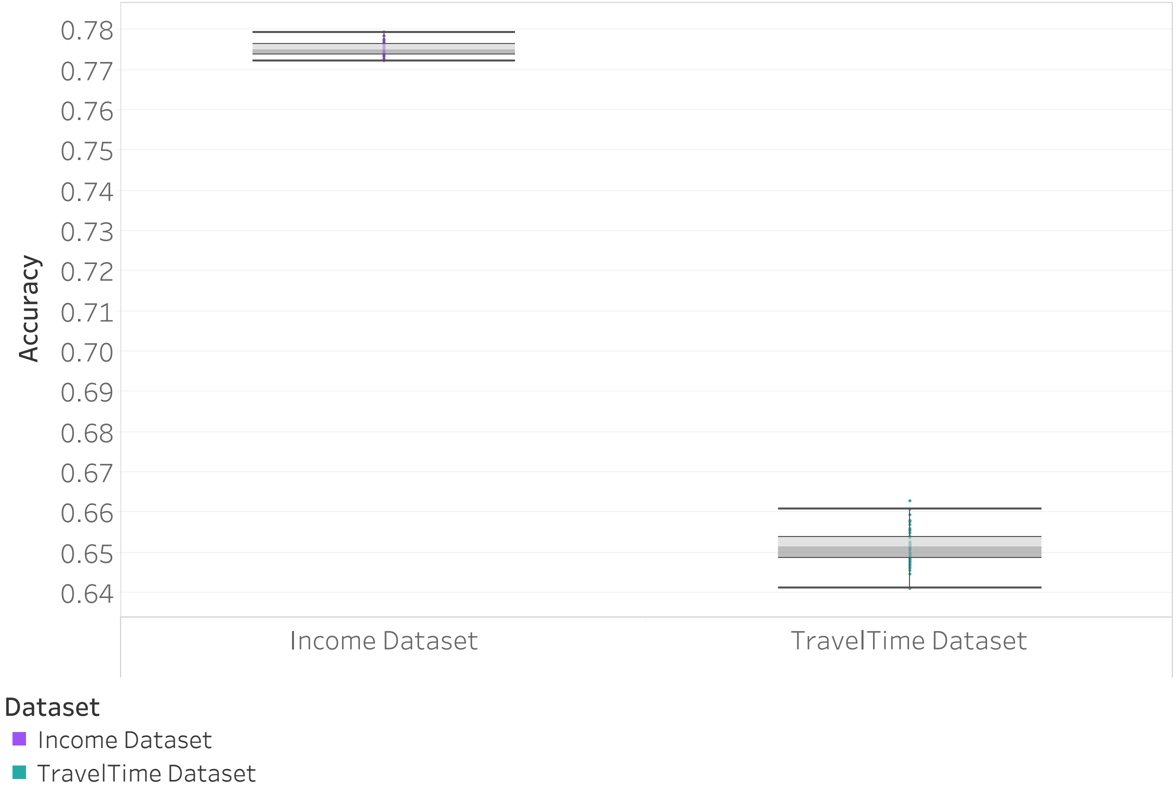


Figure 5.5 Box plot showing the distribution of Accuracy for the baseline (non-pruned) model across both datasets. Each color corresponds to one dataset (Income or Travel Time), highlighting the variability and differences in Accuracy between the two datasets.

lower accuracy shows that the base model performs less reliably on it. The Income Dataset maintains better fairness scores and higher accuracy, indicating cleaner data and stronger generalizability. Lower fairness disparities with smaller variance in the Income Dataset suggest stable, balanced predictions across demographic groups. Overall, the Income Dataset demonstrates better fairness and performance. In contrast, the Travel Time Dataset shows stronger fairness disparities and lower accuracy, highlighting the need for fairness corrections and additional preprocessing.

Based on these observations, the Income dataset — characterized by its stability — provides a solid foundation for deriving general insights in this study. In contrast, the Travel Time dataset, with greater variance and weaker correlations, serves as a more challenging benchmark to evaluate how initial dataset characteristics influence pruning model performance.

Now that we established a clear understanding of our datasets and the behavior of the Base model across key metrics, we can proceed to analyze the experimental results of our pruning techniques and their impact on the accuracy–fairness trade-off.

CHAPTER 6 Experimental Analysis and Results

This section is structured to progressively build a comprehensive understanding of pruning methods and their impact on fairness and performance metrics. The analysis begins by establishing a baseline comparison between random pruning and structured pruning methods, aiming to assess whether structured pruning methods provide added value and in what ways. Next, the inherent biases present within the datasets are examined to determine whether different pruning methods and pruning levels alter these underlying biases. To explore this relationship further and in greater depth, the fairness-relative relationships between metrics are investigated. This is achieved through the application of three-way factorial ANOVA and three-way permuted factorial ANOVA tests, enabling the identification of the effects of key factors —dataset, method, and pruning level— on these relative relationships. The analysis then moves beyond relative relationships to examine the individual behavior of each metric (both fairness and accuracy) independently, once again utilizing three-way factorial ANOVA and permuted factorial ANOVA tests to quantify the effect of each factor on metrics behavior. Subsequently, the focus shifts to understanding how these factors influence the trade-off between fairness and accuracy. Having established these foundations, the final stage of the analysis delves into the direct relationship between pruning levels and each of the metrics, analyzing how increasing pruning impacts both performance and fairness. This leads to an overall synthesis, where all factors and scenarios are considered collectively to identify overarching trends, patterns, similarities, and differences. The results culminate in a robustness analysis across scenarios, followed by the presentation of best practices derived from the findings of this study.

6.1 Comparison of Structured Pruning Methods and Random Pruning

For the analysis in this section, Random Pruning was used as the baseline for comparison with structured pruning methods. Comparisons between Random Pruning and structured methods were performed for each dataset, metric, and method by aggregating results across all pruning levels. This approach was chosen for statistical robustness. Analyzing each pruning level separately would result in instability and reduced statistical power, as each pruning percentage has limited data points. Aggregation captures the overall behavior of each method relative to Random Pruning, avoiding conclusions based on isolated fluctuations.

Accuracy. In a first step, we compare the accuracy of the methods. As shown in Figure 6.1 and 6.2, which presents the boxplot for accuracy across different pruning levels and methods for both datasets, we observe that—with the exception of Rank Pruning, which underperforms compared to Random Pruning in some scenarios—all other structured methods outperform Random Pruning at most pruning levels in terms of accuracy for the Income dataset. This is expected, as these structured pruning methods are designed to preserve accuracy.

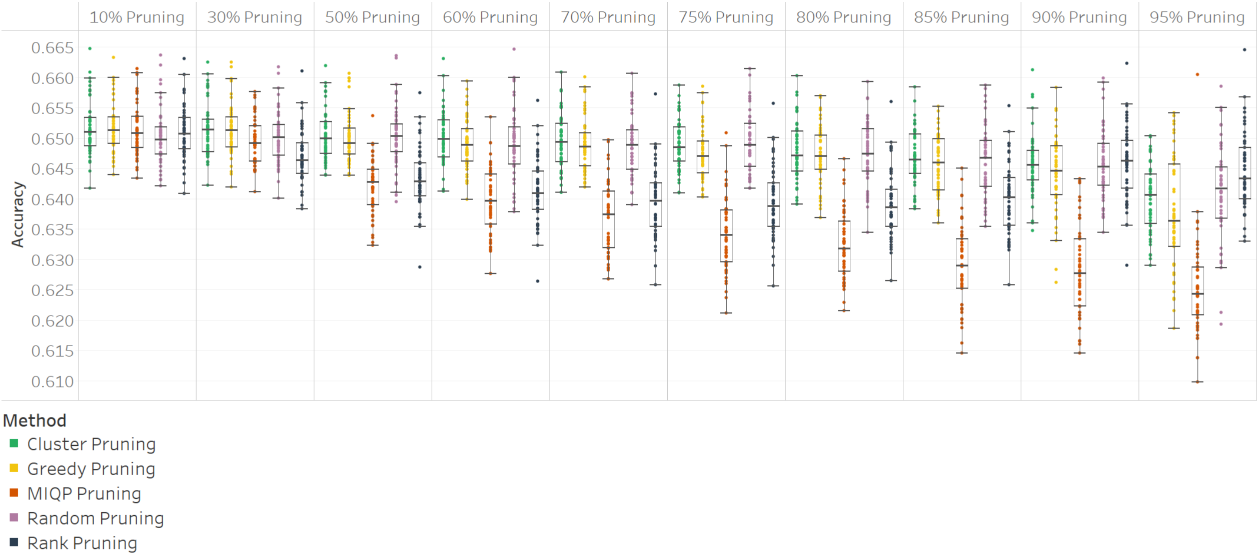


Figure 6.1 Box plot illustrating the accuracy of different pruning methods across varying pruning levels for the Travel Time dataset. The x-axis represents pruning percentages, and the y-axis shows accuracy values. Each color corresponds to a distinct pruning method. This plot highlights how accuracy changes as pruning intensity increases and allows comparison of method stability and performance at different pruning levels.

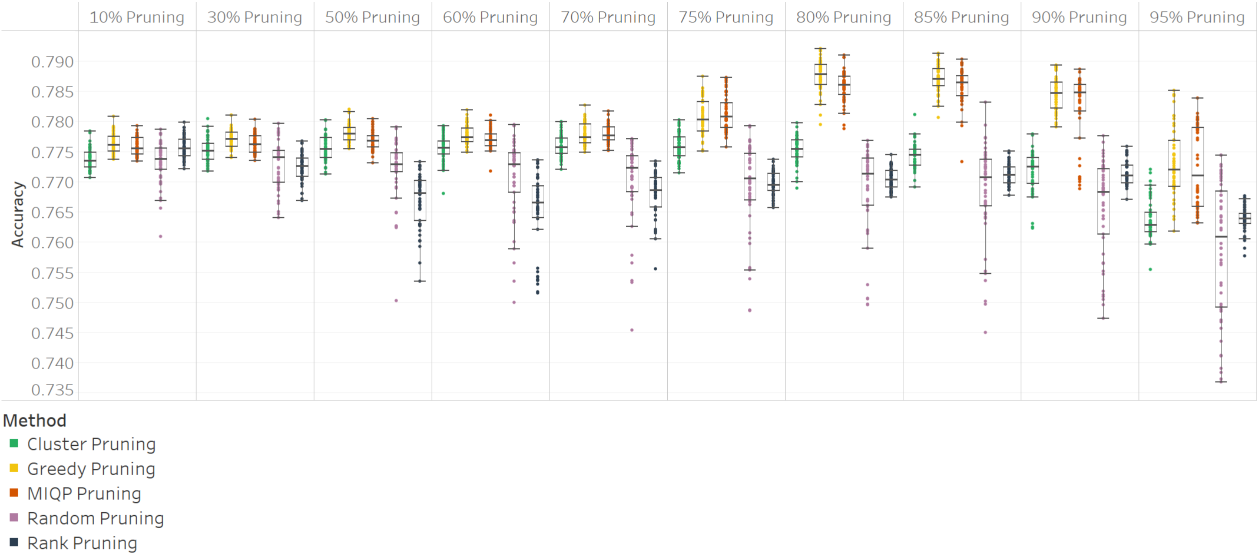


Figure 6.2 Box plot illustrating the accuracy of different pruning methods across varying pruning levels for the Income dataset. The x-axis represents pruning percentages, and the y-axis shows accuracy values. Each color corresponds to a distinct pruning method. This plot highlights how accuracy changes as pruning intensity increases and allows comparison of method stability and performance at different pruning levels.

Cluster Pruning and Greedy Pruning are the most effective at preserving accuracy, consistently surpassing Random Pruning and even improving accuracy in some cases. Overall, structured pruning methods outperform Random Pruning in terms of accuracy. Therefore, if accuracy is the main objective, structured pruning is the preferred approach. However, Rank Pruning remains the notable exception, performing worse than Random Pruning under our study's configuration. MIQP exhibits unstable behavior across different datasets, performing particularly poorly on the Travel Time dataset. While Random Pruning performs competitively with MIQP in terms of accuracy on this dataset, it generally falls behind other structured pruning methods.

For the comparison of each method's performance in terms of fairness metrics against random pruning, the Wilcoxon statistical test was applied to assess significance. Effect size measurements were also used to quantify differences across various metrics and scenarios, defined by combinations of method, pruning level, and dataset. All statistical analyses were conducted on the raw experimental results. However, given the large number of plots and variations, only key visualizations are presented. The interpretation of Cohen's d effect sizes follows standard guidelines, where values around $d = 0.2$ are considered small effects, $d = 0.5$ medium effects, and $d = 0.8$ large effects. Values below 0.2 are deemed negligible, while those above 0.8 indicate strong practical significance.

To facilitate clear comparisons, we computed win counts and win percentages for each combination of dataset, metric, and pruning method. These counts represent only the scenarios where one method significantly outperformed the other, based on the Wilcoxon test results (p-value < 0.05). Scenarios without significant differences were excluded. Each "Random Wins" or "Method Wins" value reflects the number of scenarios where either Random Pruning or the respective structured method showed a statistically significant advantage. The full results of these statistical tests, including the winning scenarios for each method and the corresponding average effect sizes, are presented in Tables 6.1 to 6.4.

Table 6.1 Summary of Wilcoxon test Results for Accuracy

Dataset	Pruning Method	Wins (Random)	Wins (Other)	Avg Effect Size (Random)	Avg Effect Size (Other)
Income	Cluster Pruning	0	9	–	0.1269
Income	Greedy Pruning	0	10	–	0.0102
Income	MIQP Pruning	0	10	–	0.0262
Income	Rank Pruning	3	3	0.1916	0.2139
Travel Time	Cluster Pruning	0	2	–	0.2116
Travel Time	Greedy Pruning	2	2	0.3153	0.2024
Travel Time	MIQP Pruning	8	1	0.0025	0.2290
Travel Time	Rank Pruning	7	1	0.0246	0.3012

Table 6.2 Summary of Wilcoxon test Results for Demographic Parity Difference

Dataset	Pruning Method	Wins (Random)	Wins (Other)	Avg Effect Size (Random)	Avg Effect Size (Other)
Income	Cluster Pruning	2	4	0.1447	0.2531
Income	Greedy Pruning	0	8	-	0.1150
Income	MIQP Pruning	1	6	0.3380	0.0344
Income	Rank Pruning	4	3	0.2600	0.2186
Travel Time	Cluster Pruning	0	9	-	0.1410
Travel Time	Greedy Pruning	0	7	-	0.1440
Travel Time	MIQP Pruning	3	5	0.2259	0.1857
Travel Time	Rank Pruning	10	0	0.0261	-

Table 6.3 Summary of Wilcoxon test Results for Equalized Odds Difference

Dataset	Pruning Method	Wins (Random)	Wins (Other)	Avg Effect Size (Random)	Avg Effect Size (Other)
Income	Cluster Pruning	2	1	0.2753	0.3365
Income	Greedy Pruning	5	1	0.1490	0.2925
Income	MIQP Pruning	5	3	0.1468	0.2727
Income	Rank Pruning	3	3	0.2607	0.2648
Travel Time	Cluster Pruning	0	9	-	0.1526
Travel Time	Greedy Pruning	0	7	-	0.1227
Travel Time	MIQP Pruning	2	5	0.2486	0.1035
Travel Time	Rank Pruning	10	0	0.0271	-

Table 6.4 Summary of Wilcoxon test Results for Predictive Parity Difference

Dataset	Pruning Method	Wins (Random)	Wins (Other)	Avg Effect Size (Random)	Avg Effect Size (Other)
Income	Cluster Pruning	6	1	0.1876	0.3004
Income	Greedy Pruning	10	0	0.1060	-
Income	MIQP Pruning	8	0	0.0890	-
Income	Rank Pruning	10	0	0.2168	-
Travel Time	Cluster Pruning	1	0	0.2525	-
Travel Time	Greedy Pruning	2	1	0.2639	0.3051
Travel Time	MIQP Pruning	1	8	0.3224	0.0731
Travel Time	Rank Pruning	0	8	-	0.0812

Demographic Parity Difference. Cluster and Greedy pruning generally outperform Random Pruning in demographic parity, but the effect sizes are small, highlighting statistical but limited practical significance. In the Income dataset, Greedy and MIQP methods outperform Random Pruning in most scenarios, but again with small effect sizes. Rank pruning is in close competition with Random Pruning, showing medium effect sizes in the Income dataset but consistently losing in the Travel Time dataset — although with negligible effect sizes. MIQP pruning shows inconsistent performance; while sometimes close to Random Pruning, Random even shows larger effect sizes against MIQP in certain metrics, suggesting meaningful practical differences. Rank pruning rarely outperforms Random Pruning in this fairness metrics and often performs worse, making it the weakest method in this regard.

Equalized Odds Difference. The results for equalized odds vary by dataset. In the Income dataset, there is no clear dominant method, with Cluster, MIQP, and Rank pruning in close competition with Random Pruning, and Greedy pruning losing in most scenarios. In the Travel Time dataset, all methods except Rank pruning (which loses consistently but with negligible effect sizes) statistically outperform Random Pruning, though the effect sizes remain small. Random Pruning remains competitive, particularly with MIQP, and often matches or outperforms methods like Rank pruning. Overall, there is no consistent superiority of structured pruning methods over Random Pruning for equalized odds for both datasets.

Predictive Parity Difference. The results for predictive parity differ from the other fairness metrics and show the opposite pattern across datasets. In the Income dataset, Random Pruning outperforms all other methods in most scenarios, though with small effect sizes. It is clearly better than MIQP, Greedy, and Rank pruning for predictive parity. In the Travel Time dataset, the results shift: there is almost no difference between Random and Cluster or between Random and Greedy pruning. MIQP and Rank pruning, in contrast to previous metrics, outperform Random Pruning in most scenarios, but with negligible effect sizes. Greedy pruning shows little to no advantage over Random Pruning and is often significantly worse. Overall, Random Pruning emerges as the strongest performer for predictive parity in the Income dataset.

Based on these results, Rank Pruning is generally less effective than Random Pruning in most scenarios, with the exception of Predictive Parity Difference in the Travel Time dataset, where it shows comparatively better performance. Also What can be observed is that the effectiveness of pruning methods depends on the initial fairness bias present in the dataset. In the Travel Time dataset, when there is a noticeable bias toward a specific fairness metric, structured pruning methods generally perform better than Random Pruning. However, the opposite is true for the Income dataset—when it exhibits bias toward a particular metric, Random Pruning tends to outperform structured methods on that metric. The extent of this performance difference varies by method: Cluster and Greedy Pruning follow this pattern most strongly, while MIQP shows more balanced performance and is more competitive with Random Pruning.

The Effect of Pruning Levels. Taking the analysis one step further, we introduce the pruning level dimension into our study, allowing us to examine how these patterns vary not only across datasets and tasks but also across different levels of pruning intensity. To explore the impact of pruning levels on fairness and performance, we divided pruning into three categories.

Figures 6.3 - 6.5 present the comparison between Random Pruning and each structured method across all metrics at three pruning intensity levels: low pruning (below 50), medium pruning (between 50 and 75), and extreme pruning (above 75).The plots reveal the following results.

Predictive Parity Difference Across Pruning Levels. At all pruning levels, Random Pruning outperforms or matches structured pruning on the Income dataset. Predictive Parity does not show improvement with structured pruning in this dataset. However, in the Travel Time dataset, at medium and extreme pruning levels, structured pruning—particularly the MIQP and Rank Pruning methods—outperforms Random Pruning in most scenarios. In general, though, Random Pruning performs better than structured methods at low pruning levels.

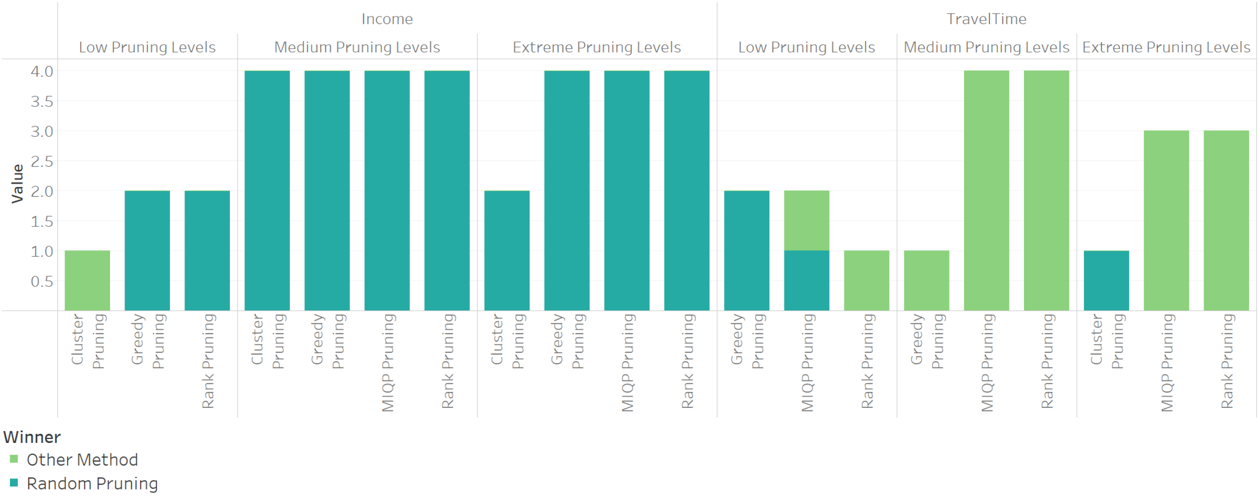


Figure 6.3 Stacked bar plots illustrating the comparison between random pruning and structured pruning methods for the metric of Predictive Parity Difference, across different pruning levels (low, medium, and extreme) for each dataset (Income and Travel Time). Each bar represents the number of scenarios in which either random pruning or the structured pruning method outperformed the other, based on pairwise statistical tests. The colored segments show the frequency of random pruning wins versus structured method wins, providing a visual understanding of performance dominance across pruning intensities.

Equalized Odds Difference Across Pruning Levels. Across all pruning levels, structured pruning performs similarly to—or only slightly better than—Random Pruning, particularly in the Travel Time dataset. In some extreme pruning levels, it even underperforms compared to Random Pruning; for instance, MIQP and Greedy in the Income dataset, and MIQP and Rank in the Travel Time dataset. Furthermore, structured pruning does not demonstrate consistent superiority in terms of Equalized Odds.

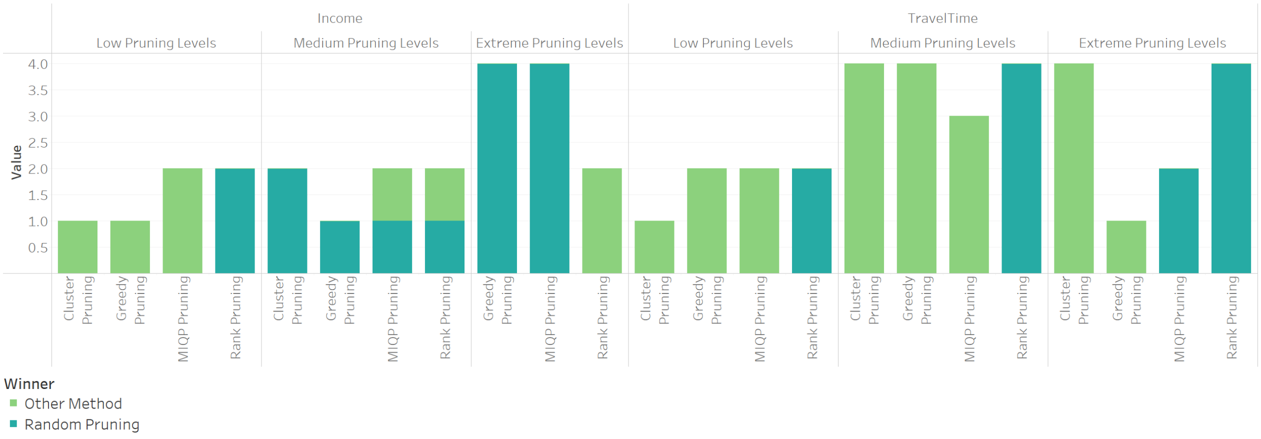


Figure 6.4 Stacked bar plots illustrating the comparison between random pruning and structured pruning methods for the metric of Equalizd Odds Difference, across different pruning levels (low, medium, and extreme) for each dataset (Income and Travel Time). Each bar represents the number of scenarios in which either random pruning or the structured pruning method outperformed the other, based on pairwise statistical tests. The colored segments show the frequency of random pruning wins versus structured method wins, providing a visual understanding of performance dominance across pruning intensities.

Demographic Parity Difference Across Pruning Levels. The results for this metric vary across datasets, but structured pruning methods generally outperform Random Pruning at medium pruning levels in both datasets. At extreme pruning levels, however, Random Pruning sometimes matches or even surpasses structured methods in certain cases. Moreover, even for fairness-sensitive metrics such as Demographic Parity, structured pruning is not always the superior choice. Depending on the pruning level and dataset, Random Pruning can show better performance, highlighting that there is no consistent winner across all scenarios.

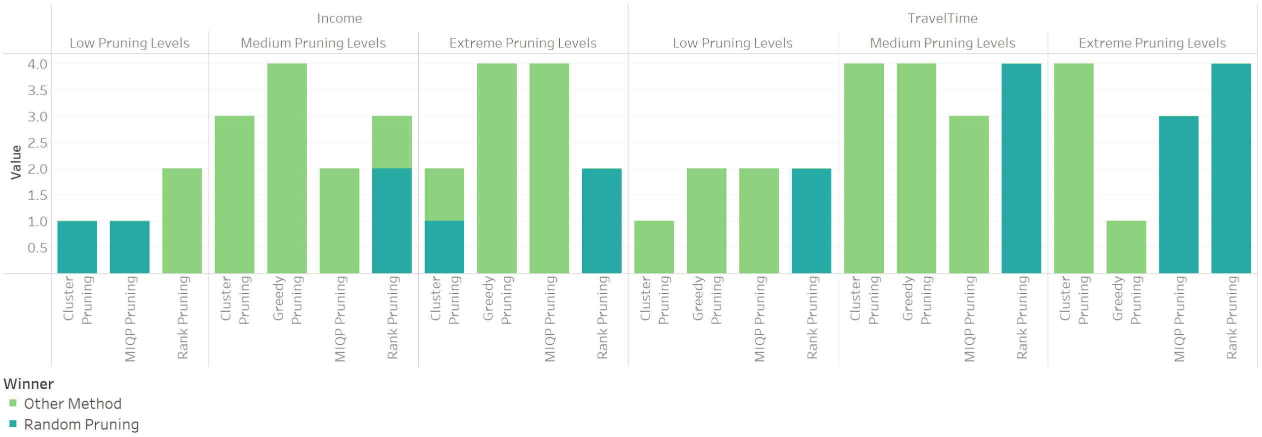


Figure 6.5 Stacked bar plots illustrating the comparison between random pruning and structured pruning methods for the metric of Demographic Parity Difference, across different pruning levels (low, medium, and extreme) for each dataset (Income and Travel Time). Each bar represents the number of scenarios in which either random pruning or the structured pruning method outperformed the other, based on pairwise statistical tests. The colored segments show the frequency of random pruning wins versus structured method wins, providing a visual understanding of performance dominance across pruning intensities.

Case Study: Random Pruning vs. Specific Structured Pruning Methods

We can perform a detailed pairwise comparison between Random Pruning and other pruning methods at a very granular level.

Random vs. Greedy Pruning According to Figures 6.3 - 6.5, In the Income dataset, at medium and extreme pruning levels, Greedy Pruning outperforms Random Pruning in terms of Demographic Parity. However, for Predictive Parity, Random Pruning consistently outperforms Greedy Pruning. Similarly, for Equalized Odds at medium and extreme pruning levels, Random Pruning shows better performance than Greedy. Overall, while Greedy Pruning is favored for Demographic Parity, it tends to underperform compared to Random Pruning in the other two fairness metrics—especially at medium and extreme pruning levels. In the Travel Time dataset, Greedy Pruning demonstrates better performance in Demographic Parity in 80% of cases. While Random Pruning outperforms Greedy Pruning in some scenarios for Predictive Parity, there is no consistent superiority between the two. For Equalized Odds, Greedy Pruning outperforms Random Pruning in several scenarios across all pruning levels. Overall, Greedy Pruning proves to be more effective for certain fairness metrics—particularly Equalized Odds Difference and Demographic Parity Difference—across varying pruning levels. Nonetheless, it tends to underperform in Predictive Parity at most pruning levels.

Random vs. Cluster Pruning Figures 6.3 - 6.5 reveal that Cluster Pruning consistently outperforms Random Pruning in terms of accuracy. However, at extreme pruning levels, Random Pruning competes closely with Cluster Pruning on fairness metrics. For Predictive Parity, Random Pruning often matches or even outperforms Cluster Pruning across all levels of pruning and all datasets. While Cluster Pruning effectively maintains accuracy, it does not always lead to improvements in fairness, particularly under conditions of high pruning.

Random vs. MIQP Pruning According to Figures 6.3 - 6.5 Fairness metrics vary across datasets and pruning levels, and neither method consistently emerges as superior. At extreme pruning levels, Random Pruning competes closely with MIQP. Additionally, MIQP demonstrates inconsistency and does not reliably outperform Random Pruning in terms of fairness in any of the metrics or levels.

Random vs. Rank Pruning According to Figures 6.3 - 6.5 Rank Pruning emerges as the weakest performer in both accuracy and fairness. Random Pruning consistently outperforms Rank Pruning across these metrics, except for Predictive Parity in the Travel Time dataset, where Rank Pruning consistently demonstrates superior performance compared to Random Pruning.

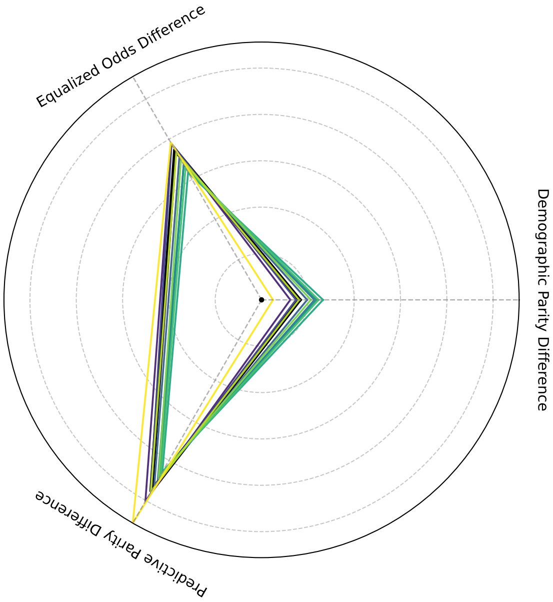
Key Insights The key insights from the case study highlight several important observations. In the Income dataset, Random Pruning significantly outperforms structured pruning in terms of Predictive Parity and Equalized Odds across almost all pruning levels. Although structured pruning improves Demographic Parity and Equalized Odds in certain scenarios specially in the travel time dataset, it tends to lose this advantage at extreme pruning levels. The effectiveness of structured pruning in comparison to the random pruning is highly dependent on the dataset, the pruning level, and the specific fairness metric being targeted. At higher levels of pruning, Random Pruning often competes with or even surpasses structured methods, suggesting that structured approaches become increasingly focused on accuracy rather than fairness. Interestingly, Random Pruning can sometimes offer fairness improvements without causing major accuracy sacrifices, presenting a practical trade-off for certain situations. Rank Pruning, however, emerges as the least effective method and should only be used with caution unless the ranking function is carefully designed. While structured pruning can contribute to fairness improvements, the choice of method must take into account the dataset's characteristics, the degree of pruning, and the fairness goals. Additionally, computational cost should not be overlooked when selecting a pruning approach.

6.2 Most Pruning Techniques Do Not Alter Underlying Fairness Metric Relationships but Amplify or Deamplify Existing Bias Patterns

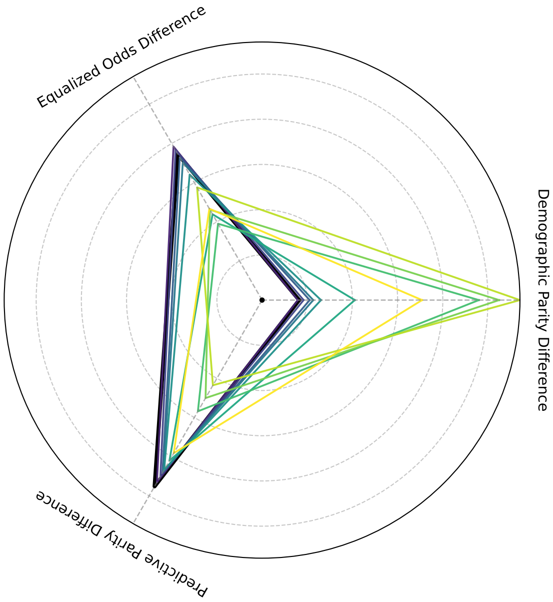
This finding suggests that pruning does not fundamentally restructure the bias present in the ensemble but instead modulates its strength. Consequently, understanding the base model's bias profile is critical, as in most cases, pruning will preserve these dynamics.

A key observation from the experimental analysis is that most pruning techniques, especially at low to medium pruning levels, do not fundamentally change the relationships between fairness metrics. Instead, they amplify or deamplify bias patterns already present in the base model.

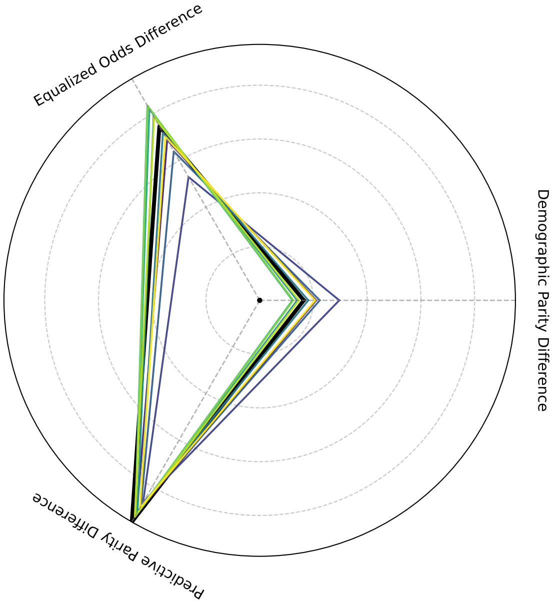
Based on the spider charts 6.8, which show the relative relationships between three fairness metrics for each dataset and pruning method, the following patterns are observed. The relative relationships between fairness metrics remain largely consistent across different datasets, methods, and pruning levels. While the magnitude of these metrics shifts with pruning, this reflects amplification or suppression effects rather than structural changes. Some methods, such as greedy and MIQP pruning, exhibit changes in these relationships at extreme pruning levels or on more sensitive datasets. Specifically, in the income dataset, MIQP and greedy pruning begin to alter these relationships around the 75% pruning mark, while in the travel time dataset, greedy pruning demonstrates similar behavior starting at 75% pruning. In contrast, cluster pruning and rank pruning maintain these relative relationships between metrics across nearly all levels of pruning, thereby preserving the initial fairness dynamics.



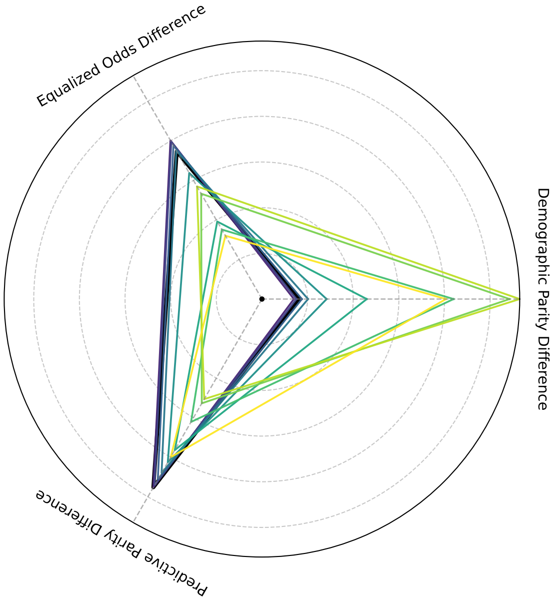
Income Dataset — Cluster Pruning Method



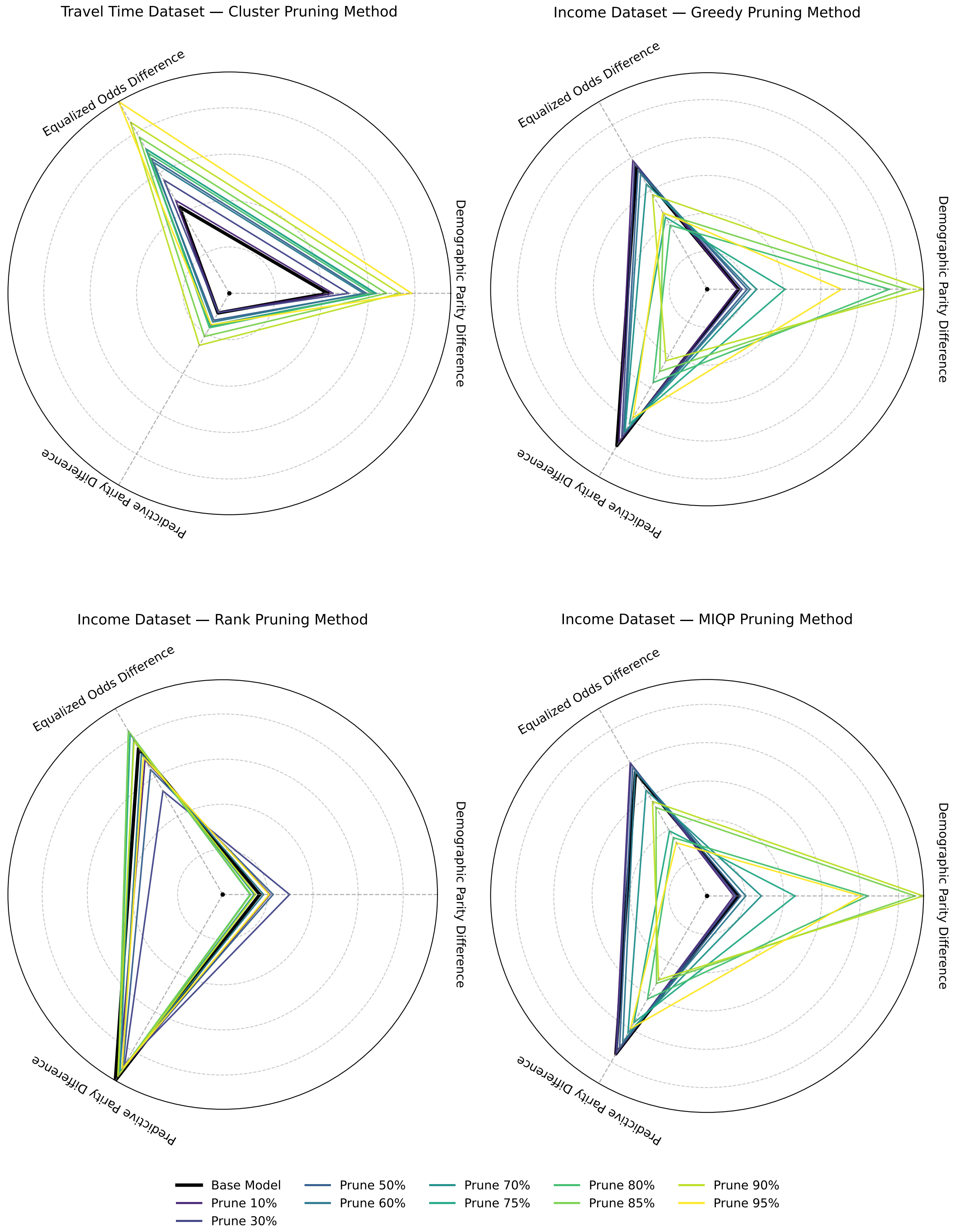
Income Dataset — Greedy Pruning Method



Income Dataset — Rank Pruning Method



Income Dataset — MIQP Pruning Method



Travel Time Dataset — Cluster Pruning Method

Income Dataset — Greedy Pruning Method

Income Dataset — Rank Pruning Method

Income Dataset — MIQP Pruning Method

Figure 6.8 Spider charts demonstrating the relationship between various fairness metrics across different pruning methods (Cluster, Greedy, Rank, MIQP) and datasets (Income, Travel Time) at different pruning levels. Each color line corresponds to one pruning level, with the base random forest model shown in black as a reference line.

Additional scatter plots 6.9 and 6.10, filtered by pruning method and dataset across all pruning levels and including the base model for reference, further reinforce these findings.

Trend lines, represented by colored lines of best fit across scatter plots, demonstrate almost stable and consistent relationships between fairness metrics across varying pruning intensities. The data points cluster along predictable linear or smooth curves, indicating that the relationships between metrics remain persistent and do not become erratic. Additionally, different pruning percentages shift the metrics along the axes while continuing to follow the same correlation structures, except for some scenarios mentioned previously, confirming that pruning scales these biases rather than changing their direction.

Since our analysis includes multiple dimensions — models, datasets, metrics, and pruning levels — it generates numerous visualizations. In this section, we present only the scatter plots for the most informative and significant scenarios.

Specifically, we include the scatter plots for the cluster pruning method on the Travel Time dataset, as well as the scatter plots for the greedy pruning method on the Income dataset.These selections allow us to cover two different datasets, namely Income and Travel Time, and two distinct pruning methods from different categories, both of which exhibit varying behaviors at extreme pruning levels. For a comprehensive view, the complete set of scatter plots is provided in Appendix A.

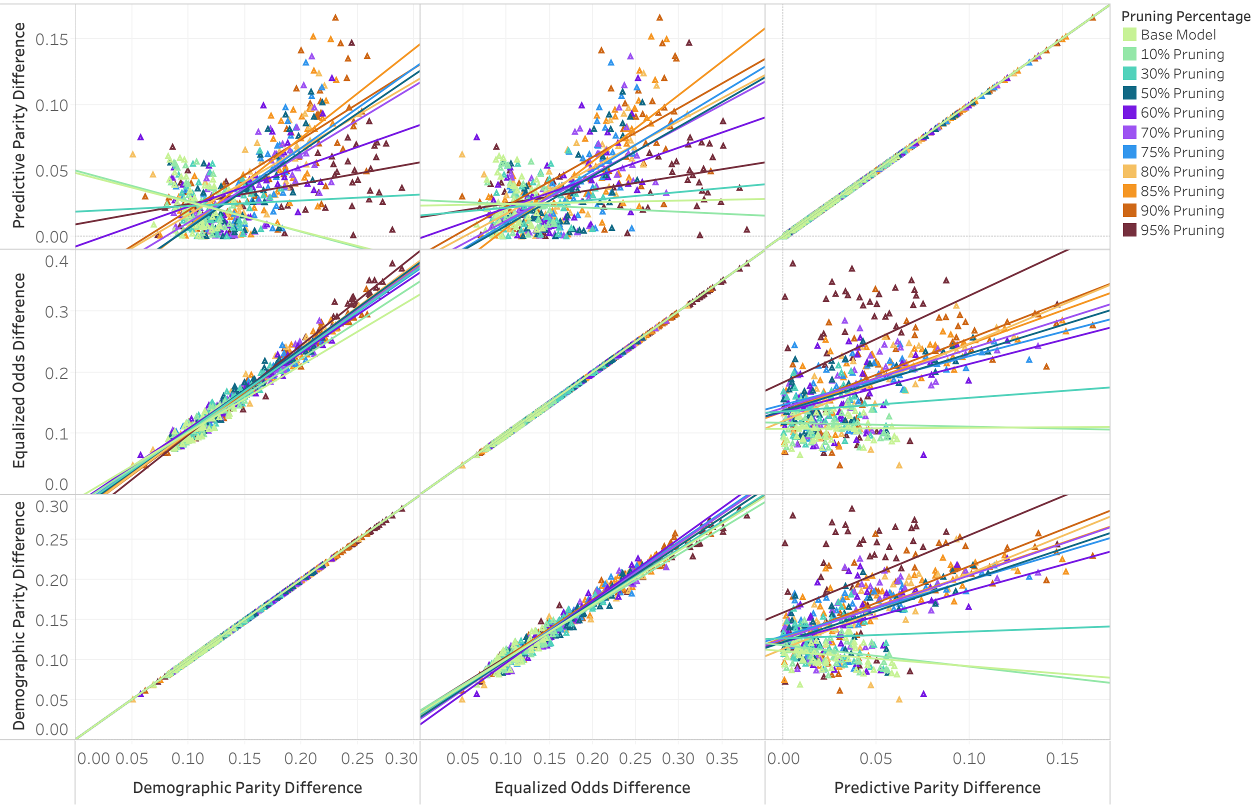


Figure 6.9 Scatter Plot of 2x2 Combinations of Fairness Metrics for Cluster Pruning on the Travel Time Dataset This scatter plot illustrates the relationship between all possible 2x2 combinations of fairness metrics at different pruning levels for the cluster pruning method applied to the travel time dataset. Each color represents a different pruning level.

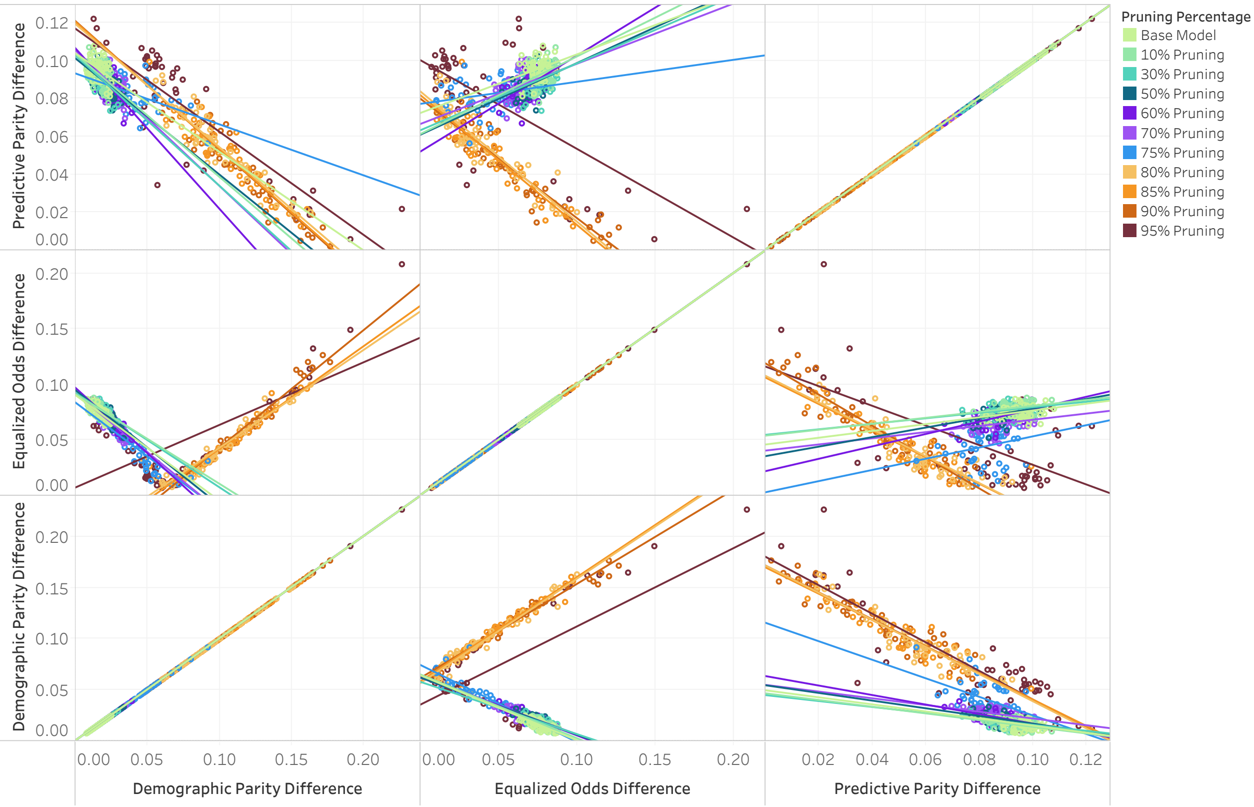


Figure 6.10 Scatter Plot of 2x2 Combinations of Fairness Metrics for Greedy Pruning on the Income Dataset This scatter plot shows the relationship between all possible 2x2 combinations of fairness metrics at different pruning levels for the greedy pruning method applied to the income dataset. Each color represents a different pruning level.

In examining the relationships between fairness metrics and their variations across pruning levels, we observe that structural patterns present in the base (unpruned) models remain consistent. Rather than altering these relationships, pruning either amplifies or suppresses them. This behavior indicates that these specific pruning methods operate on top of existing bias structures rather than fundamentally changing the fairness dynamics of the model. Thus, pruning reinforces patterns already embedded in the ensemble's architecture, either exacerbating or mitigating them as pruning levels increase. These results emphasize the importance of carefully assessing the bias profile of the base model prior to pruning, as the pruning process will inherently propagate these tendencies rather than correct them.

It is important to note, however, that while most pruning techniques preserve these relationships at moderate levels, the structure and direction of fairness metric correlations are not universal and can vary between datasets. Furthermore, under certain methods or extreme pruning intensities, these relationships may be altered. We investigate these nuances in detail in the following section.

6.3 Fairness Metric Relationships Are Dataset- and Method-Dependent, Revealing Unique Correlation Patterns

This result highlights that different fairness metrics do not always reflect the same biases and that their relationships can be influenced by dataset structures and pruning levels. Understanding whether fairness metrics inter-relate differently based on experimental conditions is critical for assessing the stability and predictability of fairness behaviors across pruning methods and datasets.

We analyzed this relationship from the broader perspective of general trends in Result 2; now, we dig deeper into the correlation between metrics in each scenario — each unique combination of dataset, pruning level, and pruning method. We now examine how fairness metrics relate to each other under different pruning methods, levels, and datasets, supported by statistical tests.

In the previous section, we demonstrated that most pruning methods do not fundamentally alter the inherent bias structures present in the dataset and do not significantly change the relationships between fairness metrics; rather, they tend to amplify or deamplify existing patterns. However, we also noted that under certain rare conditions — particularly for methods such as MIQP and Greedy pruning, and especially at extreme pruning levels — deviations from these general patterns can occur. This observation naturally leads to a deeper question: to what extent do the dataset characteristics, pruning methods, and pruning levels statistically influence these fairness metric relationships?

To assess the impact of various factors on fairness metrics and accuracy, we conduct a structured analysis using permutation-based factorial ANOVA, supported by visual inspection. This approach confirms the overall stability of metric relationships while revealing that certain conditions — such as extreme pruning levels or specific methods — can cause notable shifts. In our analyses, we constructed permutation factorial ANOVA models based on the defined dependent variable and relevant independent factors. The general model structure is:

$$Metric\_Value \sim C(Factor_1) \times C(Factor_2) \times \cdots \times C(Factor_n)$$

where:$C(\cdot)$ denotes that the factor is treated as categorical. The model includes main effects of each factor, two-way interactions, and higher-order interactions between factors as applicable. The number and combination of factors are determined based on the specific research question and scenario under analysis.

Interpretation of Permuted factorial ANOVA Metrics. The *F-value* measures the strength of the effect of each factor by comparing explained and unexplained variance; higher values indicate stronger effects. The *permutation p-value* assesses statistical significance by comparing the observed result to a distribution generated through random permutations, with lower values (typically below 0.05) indicating significance. The *Eta-squared* (η^2) represents the proportion of variance explained by the factor, indicating effect size, where values below 0.06 are considered small, between 0.06 and 0.14 moderate, and above 0.14 large. *Interaction effects* show whether the influence of one factor depends on the level of another, and *three-way interactions* indicate more complex dependencies between multiple factors.

The results of the permutation factorial ANOVA test are presented in Tables 6.5- 6.7.

Table 6.5 Permutation ANOVA results for the correlation between Demographic Parity Difference and Predictive Parity Difference.

Factor	**Observed F**	**Permutation p-value**	**Eta squared**
Dataset	576.98	0	0.6719
Method	1.55	0.2118	0.0054
Dataset:Method	47.59	0	0.1663
Prune percentage	11.06	0.0018	0.0129
Dataset:Prune percentage	43.66	0	0.0508
Method:Prune percentage	3.05	0.0316	0.0107
Dataset:Method:Prune percentage	2.13	0.1012	0.0074
Residual		0	0.0745

Table 6.6 Permutation factorial ANOVA results for the correlation between Demographic Parity Difference and Equalized Odds Difference.

Factor	**Observed F**	**Permutation p-value**	**Eta squared**
Dataset	283.47	0	0.6333
Method	10.99	0	0.0737
Dataset:Method	5.01	0.004	0.0336
Prune percentage	11.14	0.001	0.0249
Dataset:Prune percentage	15.88	0	0.0355
Method:Prune percentage	5.56	0.001	0.0373
Dataset:Method:Prune percentage	2.81	0.0446	0.0188
Residual		0	0.1430

Table 6.7 Permutation ANOVA results for the correlation between Predictive Parity Difference and Equalized Odds Difference.

Factor	**Observed F**	**Permutation p-value**	**Eta squared**
Dataset	12.44	0.0008	0.0709
Method	8.33	0.0002	0.1424
Dataset:Method	7.50	0.0002	0.1282
Prune percentage	1.40	0.2368	0.0080
Dataset:Prune percentage	30.60	0	0.1743
Method:Prune percentage	2.63	0.0632	0.0449
Dataset:Method:Prune percentage	3.90	0.0138	0.0667
Residual		0	0.3647

Dataset Dependency The analysis reveals that the dataset factor has a very strong influence on the relationships between the different fairness metrics, consistently showing large effects across all examined metric correlations. Moreover, there is a strong interaction between the dataset and the method, indicating that the effect of different pruning methods varies significantly across datasets. The interaction between the dataset and the pruning percentage is moderate, suggesting that the impact of pruning levels also changes depending on the dataset, but to a lesser degree.

Method Dependency The method factor does not show a significant influence on the correlation between demographic parity difference and predictive parity difference. However, for the correlations between demographic parity difference and equalized odds difference, and between predictive parity difference and equalized odds difference, the method factor has a noticeable impact — particularly strong in the latter case. Additionally, small but significant interactions between method and pruning percentage are observed, indicating that pruning levels slightly influence the fairness outcomes of different methods, except for the Predictive Parity Difference and the Equalized Odds Difference.

Pruning Level Dependency The pruning percentage demonstrates a small but statistically significant effect on all correlations between the fairness metrics, except for the Predictive Parity Difference and the Equalized Odds Difference, indicating that pruning levels do affect fairness relationships, though modestly. Furthermore, small to medium in some cases, but significant three-way interactions among dataset, method, and pruning percentage are present, suggesting that the combined influence of these factors subtly shapes the relationships between the fairness metrics.

To further clarify the permutation factorial ANOVA test results, the visual results of the permutation factorial ANOVA can be observed in the line plots 6.11 - 6.13, which illustrate the correlation between each pairwise combination of fairness metrics across different pruning levels. These plots demonstrate how the relationships between metrics vary by pruning method, dataset, and prune percentage, providing a clear visual representation that aligns with the statistical findings. The distinct patterns for each pruning Method and the varying trends between datasets visually confirm the systematic differences captured by the permutation factorial ANOVA results.

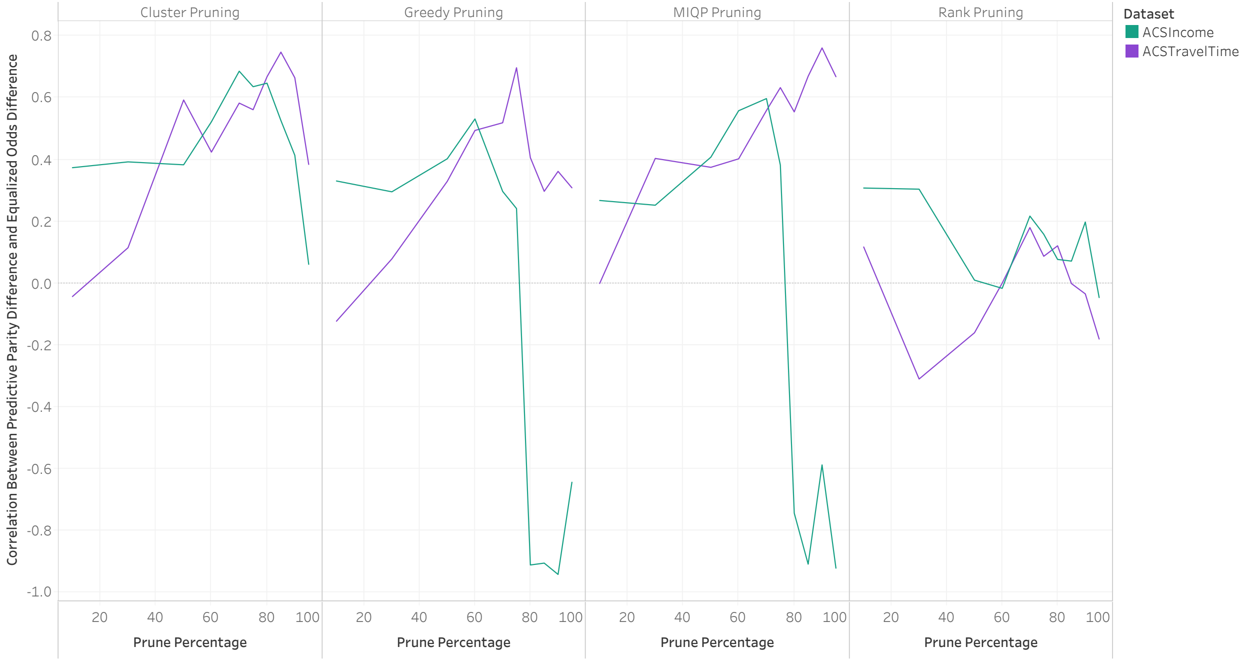


Figure 6.11 Line plot showing the correlation between predictive parity difference and equalized odds difference across different pruning levels and methods. Each subplot represents a pruning method, and lines represent different datasets.

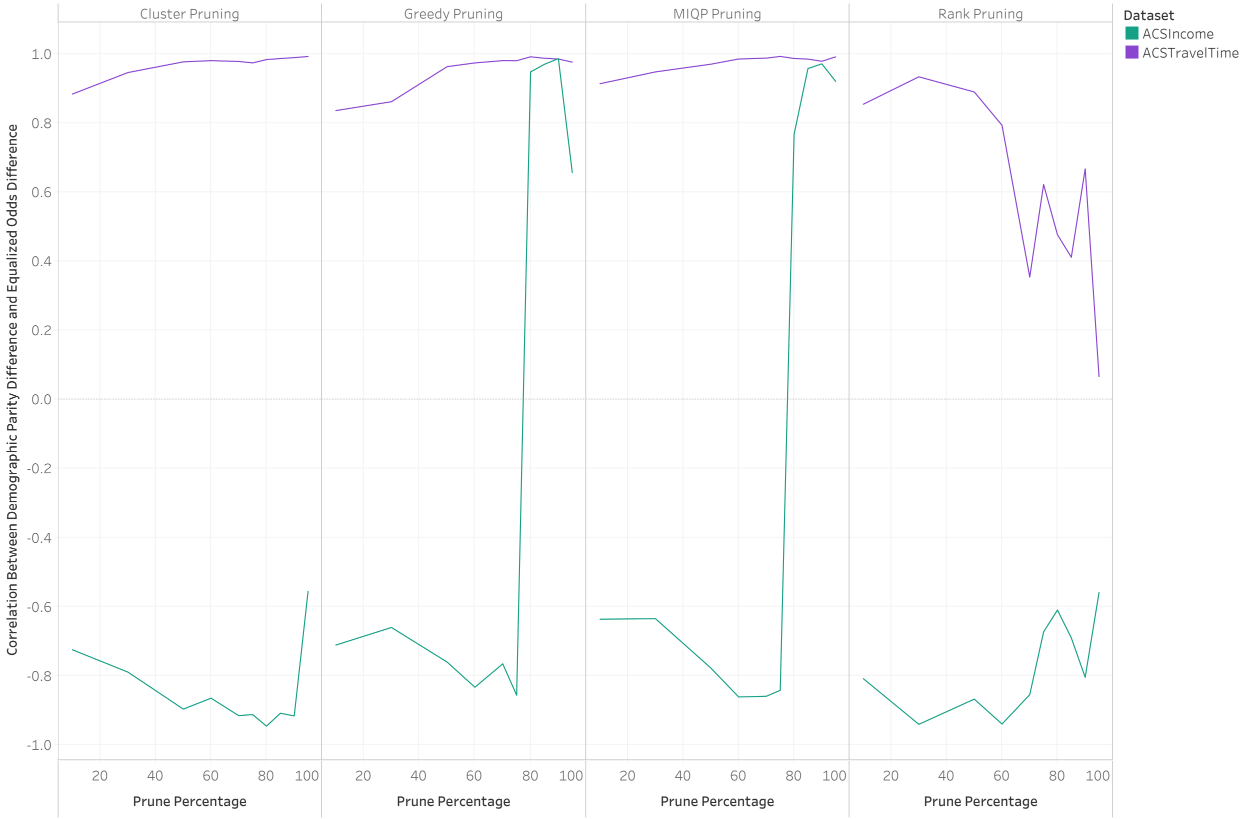


Figure 6.12 Line plot showing the correlation between demographic parity difference and equalized odds difference across different pruning levels and methods. This visualization demonstrates how the relationships between these fairness metrics evolve with pruning and differ across datasets and pruning methods.

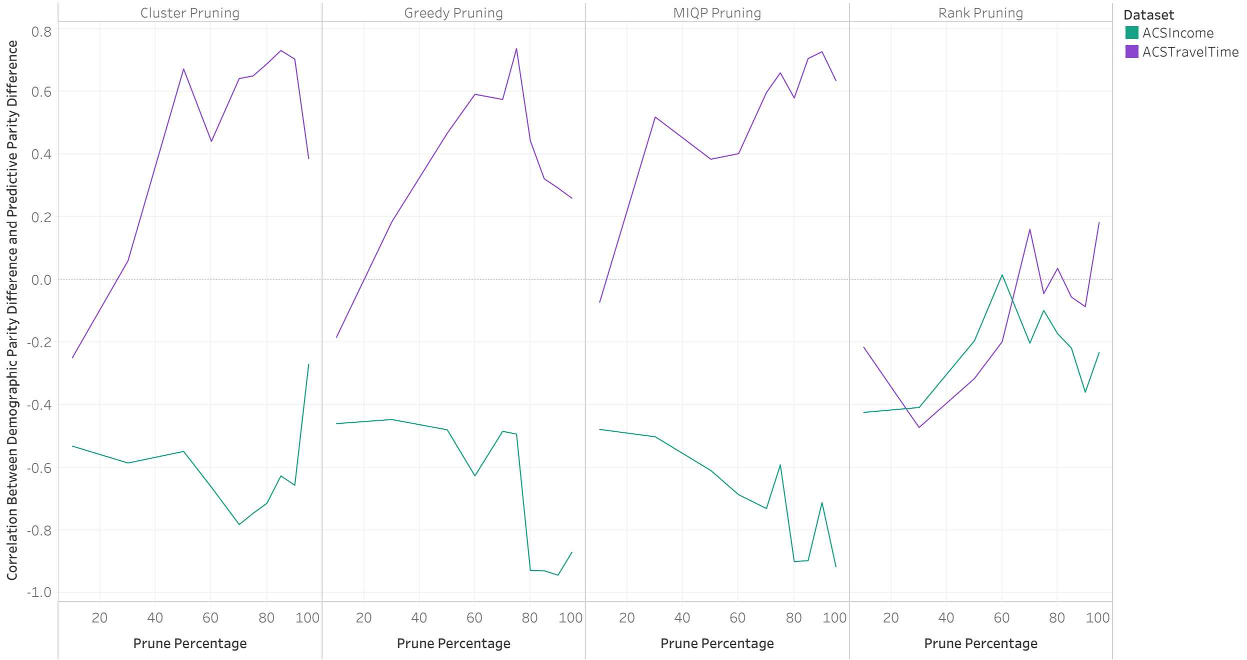


Figure 6.13 Line plot showing the correlation between demographic parity difference and predictive parity difference across various pruning levels and methods. These plots illustrate the stability and shifts in fairness metric relationships under different pruning conditions and datasets.

To visually analyze these relationships, we also created scatter plots for different scenarios. Given the large number of scenarios, we present only a limited and representative selection here, with the complete set of plots available in the appendix B. Specifically, we present results for three representative pruning levels: 30% (low pruning), 60% (medium pruning), and 85% (extreme pruning). For each of these levels, we analyze scatter plots across different methods and datasets to further support and interpret the findings discussed above.

According to the scatter plots 6.14 - 6.16 for the different pruning measures and for different datasets, for three different levels of pruning — 30%, 60%, and 85% — we can see that for each pairwise comparison of the methods and datasets at a specific pruning level, the trend lines and the scatter plots show significant differences. According to these plots, different pruning methods are indicated by different colors, and each dataset is represented by a different shape. In each grid of the plot, which shows the relationship between two fairness metrics, the scatter shape for the travel time dataset and the income dataset shows significant differences. For instance, at a 30% pruning level, in the relationship between predictive parity difference and equalized odds difference for the income dataset, the points are more densely concentrated in one specific area of the grid. In contrast, for the travel time dataset, the points and results are more scattered throughout the grid. The same behavior is observed in the relationship between predictive parity difference and demographic parity difference, the density of the data points is focused in one part of the grid for the income dataset, while for the travel time dataset, the points are more scattered.

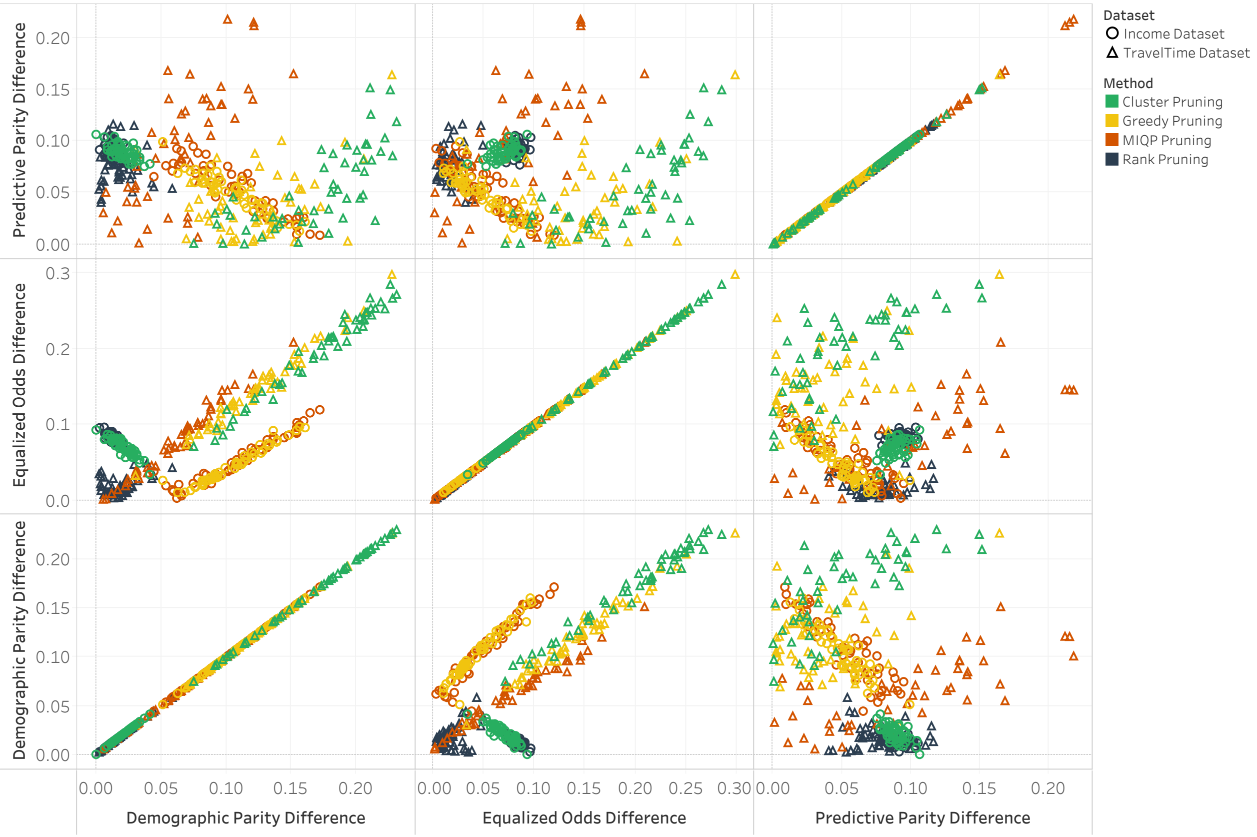


Figure 6.14 Scatter plots illustrating the relationships between demographic parity difference, predictive parity difference, and equalized odds difference for pruning level 85% for each dataset. Different colors represent pruning methods, and point shapes distinguish between datasets.

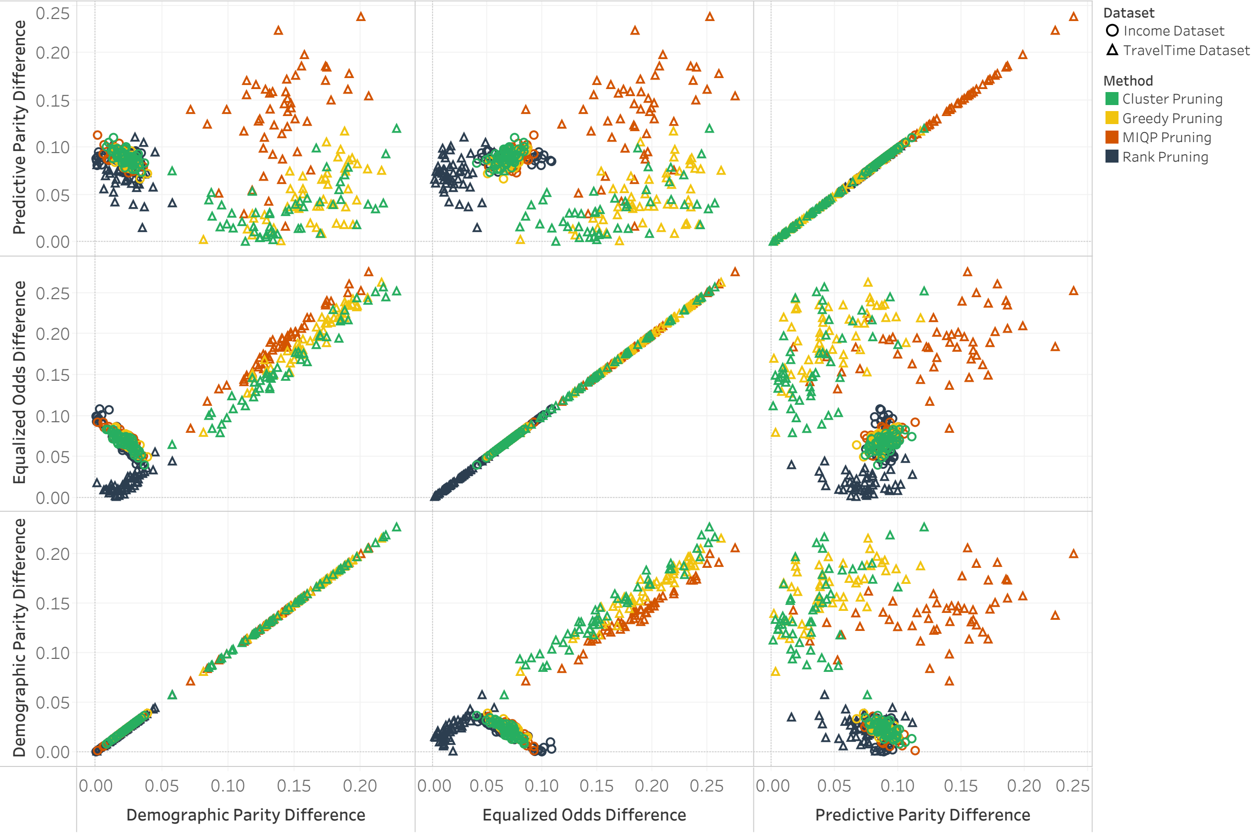


Figure 6.15 Scatter plots illustrating the relationships between demographic parity difference, predictive parity difference, and equalized odds difference for pruning level 60% for each dataset. Different colors represent pruning methods, and point shapes distinguish between datasets.

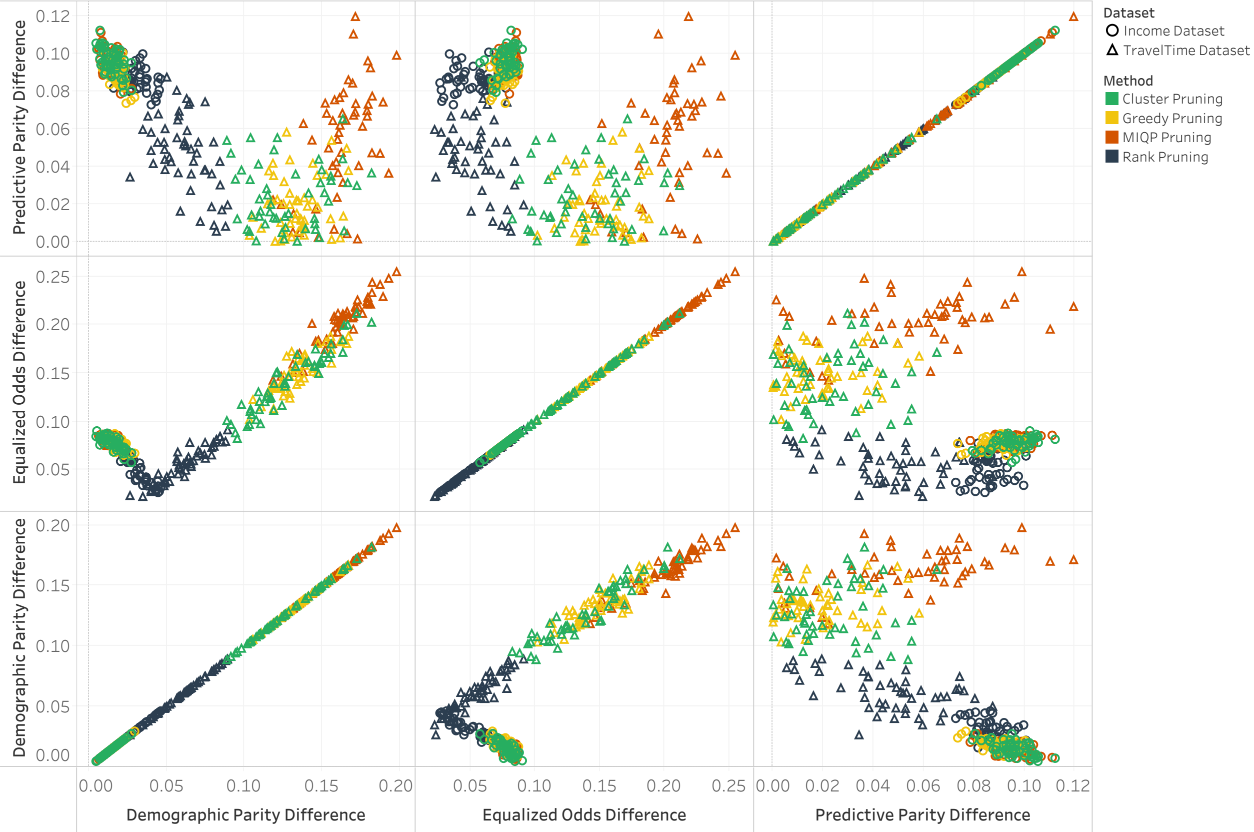


Figure 6.16 Scatter plots illustrating the relationships between demographic parity difference, predictive parity difference, and equalized odds difference for pruning level 30% for each dataset. Different colors represent pruning methods, and point shapes distinguish between datasets.

This specific pattern — more scatteredness for the travel time dataset and more density in one place for the income dataset — can be seen in all different 2x2 combinations of the metrics. Moving to another level of pruning, this behavior is even more amplified. For the travel time dataset, the scatteredness of the results increases, while for the income dataset, the points remain denser in one specific area. At the next level of pruning, both the travel time and income datasets show more scatteredness and less density in one specific area. However, a very different behavior emerges as we reach extreme pruning levels. For example, for the relationship between demographic parity difference and equalized odds difference in the travel time dataset, the results of different methods become similar. This means that with an increasing demographic parity difference, the equalized odds difference also increases. However, for the income dataset, the behavior is different: cluster pruning and rank pruning show a trend where increasing demographic parity difference causes a decrease in equalized odds difference. In contrast, MIQP exhibits the opposite trend, where an increase in demographic parity difference also leads to an increase in equalized odds difference. This behavior of MIQP matches the behavior of all other methods on the travel time dataset. From the scatter plots, we can understand that the relationship between fairness metrics is highly dependent on the dataset and method as well as the pruning level.

Overall, the statistical analysis reveals clear and consistent patterns in the inter-correlation between fairness metrics. The trade-offs are primarily driven by the dataset factor, with large effect sizes across all metrics — showing that dataset characteristics play the biggest role in shaping fairness relationships. Method choice also matters, though its impact varies depending on which fairness metrics are being compared, ranging from negligible to strong. Pruning levels show no effect or only minor effects individually, but their impact becomes more apparent when interacting with other factors. The strongest interactions appear between dataset and method, highlighting that these trade-offs are highly context-dependent. In short, dataset and method choices have the largest influence on fairness metric trade-offs, while pruning levels add moderate, yet meaningful, variations — especially through their interactions. These findings highlight that while the relationships between fairness metrics are generally preserved and tend to be amplified or deamplified by pruning, certain combinations of datasets, methods, and extreme pruning levels can lead to deviations or shifts in these relationships. This underlines the importance of understanding both systematic bias patterns and the influence of pruning strategies when assessing fairness behaviors in practice.

6.4 Metric Behavior Depends on Dataset and Shows Significant Variation

In this section, we analyze the effect of each of the factors — dataset, method, and pruning level — on each of the metrics. Here, we do not consider any trade-offs or correlations; we only aim to understand the effect of each factor on each metric.

To analyze the influence of dataset, method, and pruning percentage on individual fairness metrics, we conducted permutation-based three-factor ANOVA tests. For each fairness metric (Demographic Parity Difference, Equalized Odds Difference, Predictive Parity Difference, and Accuracy), we built a factorial model with the formula:

$$\text{Metric} \sim \text{C}(Dataset) \times \text{C}(Method) \times \text{C}(PrunePercentage)$$

The permutation factorial ANOVA test results are shown in Table 6.8.

Factor	Sum Sq	Df	F	Permutation p-value	Eta squared
Demographic Parity Difference					
C(Dataset)	17.3437	1	631812.68	0	0.9865
C(Method)	0.0324	3	392.88	0	0.0018
C(prune_percentage)	0.0160	9	21.05	0	0.0009
C(Dataset):C(Method)	0.0454	3	550.91	0	0.0026
C(Dataset):C(prune_percentage)	0.0177	9	214.72	0	0.0010
C(Method):C(prune_percentage)	0.0034	27	16.51	0	0.0002
C(Dataset):C(Method):C(prune_percentage)	0.0140	27	67.55	0	0.0008
Residual	0.1094	3984			0.0062
Equalized Odds Difference					
C(Dataset)	17.9102	1	665558.36	0	0.9843
C(Method)	0.0363	3	450.50	0	0.0020
C(prune_percentage)	0.0169	9	20.93	0	0.0009
C(Dataset):C(Method)	0.0484	3	601.59	0	0.0027
C(Dataset):C(prune_percentage)	0.0194	9	240.95	0	0.0011
C(Method):C(prune_percentage)	0.0038	27	17.60	0	0.0002
C(Dataset):C(Method):C(prune_percentage)	0.0153	27	71.02	0	0.0008
Residual	0.1073	3984			0.0059
Predictive Parity Difference					
C(Dataset)	6.7410	1	657184.98	0	0.9839
C(Method)	0.0139	3	452.39	0	0.0020
C(prune_percentage)	0.0065	9	20.88	0	0.0010
C(Dataset):C(Method)	0.0191	3	622.63	0	0.0028
C(Dataset):C(prune_percentage)	0.0077	9	251.13	0	0.0011
C(Method):C(prune_percentage)	0.0015	27	18.21	0	0.0002
C(Dataset):C(Method):C(prune_percentage)	0.0060	27	72.60	0	0.0009
Residual	0.0399	3984			0.0058
Accuracy					
C(Dataset)	18.6542	1	749233.27	0	0.9849
C(Method)	0.0396	3	475.50	0	0.0021
C(prune_percentage)	0.0179	9	21.53	0	0.0009
C(Dataset):C(Method)	0.0512	3	613.95	0	0.0027
C(Dataset):C(prune_percentage)	0.0204	9	248.71	0	0.0011
C(Method):C(prune_percentage)	0.0040	27	17.93	0	0.0002
C(Dataset):C(Method):C(prune_percentage)	0.0160	27	70.86	0	0.0008
Residual	0.1124	3984			0.0059

Table 6.8 permutation factorial ANOVA results with permutation p-values and eta squared for each metric (Demographic Parity Difference, Equalized Odds Difference, Predictive Parity Difference, and Accuracy.

Permutation Factorial ANOVA Results Analysis for Each Fairness Metric

Demographic Parity Difference. The analysis shows that the dataset factor has an overwhelmingly dominant effect on the variation in Demographic Parity Difference, explaining the vast majority of the variance. The method factor also contributes significantly but on a much smaller scale. The pruning percentage factor has a small but statistically significant influence. While all interaction effects between dataset, method, and pruning percentage are statistically significant, their impact is minor compared to the dominant role of the dataset.

Equalized Odds Difference. For Equalized Odds Difference, the dataset factor again plays the most dominant role, explaining nearly all of the variance. The method factor has a strong, though much smaller influence, while the pruning percentage factor has a small but significant effect. Interactions between dataset, method, and pruning percentage are present and significant but account for only a very minimal portion of the variance compared to the dataset's dominant effect.

Predictive Parity Difference. In the case of Predictive Parity Difference, the dataset factor remains the most significant contributor to variation. The method factor shows a considerable effect, though much smaller than that of the dataset, and the pruning percentage has a small but notable influence. All interaction terms are significant but play a relatively minor role in explaining variance when compared to the dataset factor.

Accuracy. For Accuracy, we observe a similar pattern to previous metrics: the dataset factor exhibits a large and significant effect, explaining the majority of the variance. The method factor also has a significant but comparatively smaller impact — still greater than that of the pruning level. Meanwhile, the pruning percentage factor shows a small yet statistically significant effect.

These results demonstrate that the dataset factor is the dominant source of variance across all fairness metrics and accuracy. Among all metrics, the factor dataset has a strong, real, and meaningful effect. It's statistically and practically significant. The method factor and pruning level have much smaller but significant contributions. Meaning they are statistically significant but not practically important. Notably, interactions between factors are present and significant, but their contribution to variance is minimal compared to the dataset factor, meaning small practical effect. This finding highlights that the dataset-specific characteristics play a critical role in metric behaviors, overshadowing other factors. In subsequent analyses, we aim to isolate and examine the effects of pruning levels and methods without the dominating influence of dataset effects.

The results from the permutation factorial ANOVA analysis reveal extremely large F-values (for example, values exceeding 600,000) and near-zero permutation p-values across almost all factors. Most notably, the eta-squared values for the `Dataset` factor are exceptionally high, with values around 0.98, indicating that approximately 98% of the variance in fairness metrics is explained by the dataset itself. In other words, the `Dataset` factor almost entirely explains the variation in fairness metrics. This observation aligns with and statistically reinforces the findings presented earlier in 6.2 and 6.3, where we identified strong dataset-specific relationships in fairness metric behavior.

Based on these findings, we conclude that fairness metric behaviors are largely dataset-specific and scenario-driven. Therefore, to gain further insights, we move forward by isolating and examining the effects of pruning levels and methodological strategies independently. This approach enables us to assess whether these factors contribute meaningfully to fairness outcomes within each dataset context, without being masked by the overwhelming influence of the dataset factor.

6.4.1 The Metric Behavior Varies by Method and Pruning Level

In this section, we perform an isolated analysis of method and pruning level effects on the metrics. In the previous analysis, we observed that the **Dataset** factor exhibited an overwhelmingly dominant effect on fairness metrics, accounting for the majority of variance. To ensure that this dominance does not overshadow the contributions of other factors, we proceed by isolating and analyzing the impact of **method** and **pruning level** independently. Following the observation of an overwhelming influence of the dataset factor on fairness metrics, it became necessary to isolate and analyze the effects of pruning levels and methods without the dominating presence of the dataset factor. The purpose of this analysis is to understand whether pruning strategies and method choices influence fairness and accuracy metrics across scenarios when the dataset-specific characteristics are removed from the model.

A permutation-based Analysis of Variance (ANOVA) test is performed for each metric, based on the model

$$\text{Metric} \sim \text{C}(Method) \times \text{C}(prune\_percentage)$$

This approach allows us to assess the pure effects of method design choices and pruning intensity levels. The goal of this analysis is to evaluate whether different pruning strategies and methodological approaches introduce significant variations in fairness metrics when considered without the confounding effect of dataset-specific characteristics.

The key question we aim to answer is: *To what extent do pruning levels and methods influence fairness metrics, once the dominating influence of the dataset is removed?* This step allows us to better understand the practical impact of pruning techniques and methodological strategies on fairness behavior in isolation. The results of the Analysis of Variance (ANOVA) test are presented in Tables 6.9 - 6.12.

Factor	sum_sq	F	Permutation p-value	Eta_squared
C(Method)	2.9115	320.2011	0	0.1866
C(prune_percentage)	0.1216	4.4595	0	0.0078
C(Method):C(prune_percentage)	0.5671	6.9304	0	0.0363
Residual	12.0023	NaN	0	0.7693

Table 6.9 permutation factorial ANOVA results for Demographic Parity Difference (Method and Pruning Level effects isolated)

Factor	sum_sq	F	Permutation p-value	Eta_squared
C(Method)	2.8716	303.4322	0	0.1682
C(prune_percentage)	0.3610	12.7140	0	0.0211
C(Method):C(prune_percentage)	1.3444	15.7836	0	0.0788
Residual	12.4921	NaN	0	0.7319

Table 6.10 permutation factorial ANOVA results for Equalized Odds Difference (Method and Pruning Level effects isolated)

Factor	sum_sq	F	Permutation p-value	Eta_squared
C(Method)	0.4883	126.0838	0	0.0810
C(prune_percentage)	0.1704	14.6677	0	0.0283
C(Method):C(prune_percentage)	0.2589	7.4261	0	0.0429
Residual	5.1126	NaN	0	0.8478

Table 6.11 permutation factorial ANOVA results for Predictive Parity Difference (Method and Pruning Level effects isolated)

Factor	sum_sq	F	Permutation p-value	Eta_squared
C(Method)	0.0324	2.4398	0.0580	0.0018
C(prune_percentage)	0.0301	0.7577	0.6320	0.0017
C(Method):C(prune_percentage)	0.0147	0.1229	1.0000	0.0008
Residual	17.5048	–	–	0.9956

Table 6.12 ANOVA results for Accuracy (Method and Pruning Level effects isolated)

Demographic Parity Difference. The analysis indicates that the choice of pruning method has a strong and significant impact on Demographic Parity Difference, making it the primary factor influencing this metric. The pruning percentage also contributes a small but consistent effect. Additionally, the interaction between the method and pruning intensity plays a moderate role, showing that both factors together can shape variations in demographic parity. Overall, while the method is the dominant driver, pruning intensity introduces subtle yet systematic changes.

Equalized Odds Difference. For Equalized Odds Difference, the pruning method has a strong and notable influence, with pruning intensity also contributing a small but significant effect. The interaction between method and pruning percentage is moderate to strong, indicating that their combination can either amplify or mitigate disparities in equalized odds. This highlights that both the choice of method and pruning level collectively shape outcomes in this metric.

Predictive Parity Difference. The Predictive Parity Difference metric is moderately influenced by the pruning method, while pruning intensity has a small but noticeable effect. The interaction between method and pruning level is also moderate, suggesting that both factors jointly impact this metric, though to a slightly lesser extent than observed for Equalized Odds Difference.

Accuracy. In contrast to the fairness metrics, Accuracy is largely unaffected by variations in pruning methods or intensities. The method factor shows only a marginal effect, while the pruning percentage and their interaction are not significant. This indicates that accuracy remains stable regardless of pruning choices, and decisions around pruning design have minimal impact on accuracy compared to their influence on fairness outcomes.

Overall, the dataset emerges as the primary driver of metric behavior. Nevertheless, fairness metrics are also significantly influenced by the pruning method and pruning level, albeit to a much smaller extent. This aligns with previous findings in this study, where we showed that the relationship between fairness metrics remains generally stable within each dataset and is not substantially altered by the method, although differences can appear at extreme pruning levels and with certain methods. here in this section we statistically proved that while dataset characteristics strongly influence each metric, particularly fairness metrics, the interaction between pruning level and method can also have a significant impact. In specific scenarios, this interaction can shift underlying biases. Overall, these results confirm that fairness behaviors are primarily driven by dataset characteristics, but pruning intensity and pruning method also play meaningful roles. Even after removing the overwhelming dataset influence, the method and pruning level still show systematic contributions to fairness variations.

6.5 The Trade-Off Between Accuracy and Fairness Metrics Varies Significantly Across Datasets and Methods

In this section, we investigate whether the trade-off between accuracy and fairness metrics is influenced by three key factors: dataset, method, and pruning level. The trade-off is quantified using the ratio of accuracy difference to fairness metric difference:

$$TradeOff\ Ratio = \frac{Accuracy\ Difference}{Fairness\ Metric\ Difference}$$

In this calculation, both the accuracy difference and fairness metric difference are computed relative to the base model. Specifically, for each scenario in the raw dataset, consisting of 50 repetitions, each data point was paired with its corresponding base model result under that specific combination of dataset and method. The accuracy difference and fairness metric difference were then calculated between each of these 50 data points and their respective base model values. This approach ensures that all changes are measured relative to a stable and consistent baseline, making the interpretation clear and focused on the impact of pruning and tuning rather than arbitrary fluctuations.

This ratio represents how much accuracy changes for each unit change in a fairness metric in comparison to the base model. It does not assume that accuracy is always decreasing or that fairness is always improving. Instead, it reflects the proportional relationship between changes in accuracy and fairness metrics, regardless of direction.

We selected this metric because simple correlations do not adequately convey proportional changes. The trade-off ratio allows us to directly assess how strongly fairness adjustments are related to accuracy changes, providing with a clearer understanding of the compromise or benefit associated with modifying pruning levels or methods.

Permutation-based three-factor ANOVAs were applied to analyze trade-off metrics for demographic parity difference, predictive parity difference, and equalized odds difference. The statistical model used for this analysis is defined as:

$$\text{TradeOff\_Metric} \sim \text{C}(Dataset) \times \text{C}(Method) \times \text{C}(Prune\_Percentage)$$

The ANOVA test results are shown in the tables 6.13 - 6.15.

Table 6.13 Result for Accuracy Trade-Off with Demographic Parity Difference

Factor	Sum of Squares	F-value	Permutation p-value	Eta-squared
C(Dataset)	10.6730	13.0028	0.0001	0.0037
C(Method)	1.2406	0.5038	0.6892	0.0004
C(Dataset):C(Method)	22.2872	9.0508	0.0000	0.0078
Prune Percentage	4.1911	5.1060	0.0232	0.0015
C(Dataset):Prune Percentage	0.2351	0.2864	0.5796	0.0001
C(Method):Prune Percentage	13.7288	5.5752	0.0008	0.0048
C(Dataset):C(Method):Prune Percentage	6.5786	2.6715	0.0465	0.0023
Residual	2807.2055	–	–	0.9794

Table 6.14 Result for Accuracy Trade-Off with Predictive Parity Difference

Factor	Sum of Squares	F-value	Permutation p-value	Eta-squared
C(Dataset)	7.1410	6.5649	0.0101	0.0018
C(Method)	84.8719	26.0084	0.0000	0.0210
C(Dataset):C(Method)	210.8758	64.6214	0.0000	0.0521
Prune Percentage	0.7861	0.7227	0.3989	0.0002
C(Dataset):Prune Percentage	5.1821	4.7641	0.0294	0.0013
C(Method):Prune Percentage	10.3877	3.1832	0.0244	0.0026
C(Dataset):C(Method):Prune Percentage	8.9710	2.7491	0.0446	0.0022
Residual	3720.1046	–	–	0.9189

Table 6.15 Result for Accuracy Trade-Off with Equalized Odds Difference

Factor	Sum of Squares	F-value	Permutation p-value	Eta-squared
C(Dataset)	13.9835	13.4057	0.0002	0.0039
C(Method)	22.6375	7.2340	0.0000	0.0062
C(Dataset):C(Method)	4.4787	1.4312	0.2289	0.0012
Prune Percentage	0.0039	0.0037	0.9488	0.0000
C(Dataset):Prune Percentage	2.0163	1.9330	0.1674	0.0006
C(Method):Prune Percentage	8.1931	2.6182	0.0543	0.0023
C(Dataset):C(Method):Prune Percentage	6.3659	2.0343	0.1055	0.0018
Residual	3567.4098	–	–	0.9841

Dataset Dependency. For the trade-off between accuracy and demographic parity difference, the dataset has a significant influence, and the interaction between dataset and method also plays an important role, indicating that method effects vary depending on the dataset. However, the effect of pruning level does not seem to depend on the dataset for this trade-off.

For the trade-off between accuracy and predictive parity difference, the dataset again has a significant effect, and both the dataset–method and dataset–pruning interactions are significant. This suggests that both methods and pruning levels impact this trade-off differently across datasets.

In the case of the trade-off between accuracy and equalized odds difference, while the dataset factor itself is significant, neither the dataset–method nor the dataset–pruning interactions show significant effects. This indicates that the dataset influences the trade-off, but this influence is not dependent on method or pruning variations.

Method Dependency. For the trade-off between accuracy and demographic parity difference, the method factor does not show a significant effect, indicating that method choice does not strongly influence this trade-off. In contrast, methods significantly impact the trade-offs with predictive parity difference and equalized odds difference, demonstrating that different methods can meaningfully alter these fairness–accuracy relationships.

Pruning Level Dependency. For the trade-off between accuracy and demographic parity difference, the pruning level has a small but significant effect, and its impact varies across methods. Additionally, a significant three-way interaction suggests that the combined influence of dataset, method, and pruning level contributes to shaping this trade-off.

For the trade-off between accuracy and predictive parity difference, the pruning level alone is not significant. However, the effects of pruning do depend on the combination of method and dataset, as indicated by significant interactions. Finally, for the trade-off between accuracy and equalized odds difference, neither the pruning level nor its interactions show substantial effects, suggesting that pruning does not meaningfully affect this particular trade-off.

To help visualize the ANOVA test results, boxplots of the trade-off metrics are provided, offering a clearer understanding of the findings discussed.

Box plots 6.17 - 6.19 display the distribution of trade-off ratios across datasets and methods. Each box illustrates the spread of trade-off values for a given combination of dataset and method. Clear differences in distribution shapes, spreads, and medians are observed. For example, the MIQP method demonstrates a wider spread in predictive parity difference trade-off ratios, indicating greater variability in accuracy–fairness relationships. The clustering of values for certain methods suggests that some techniques result in more stable trade-offs. These visual patterns support the statistical significance of dataset and method effects identified in the tests.

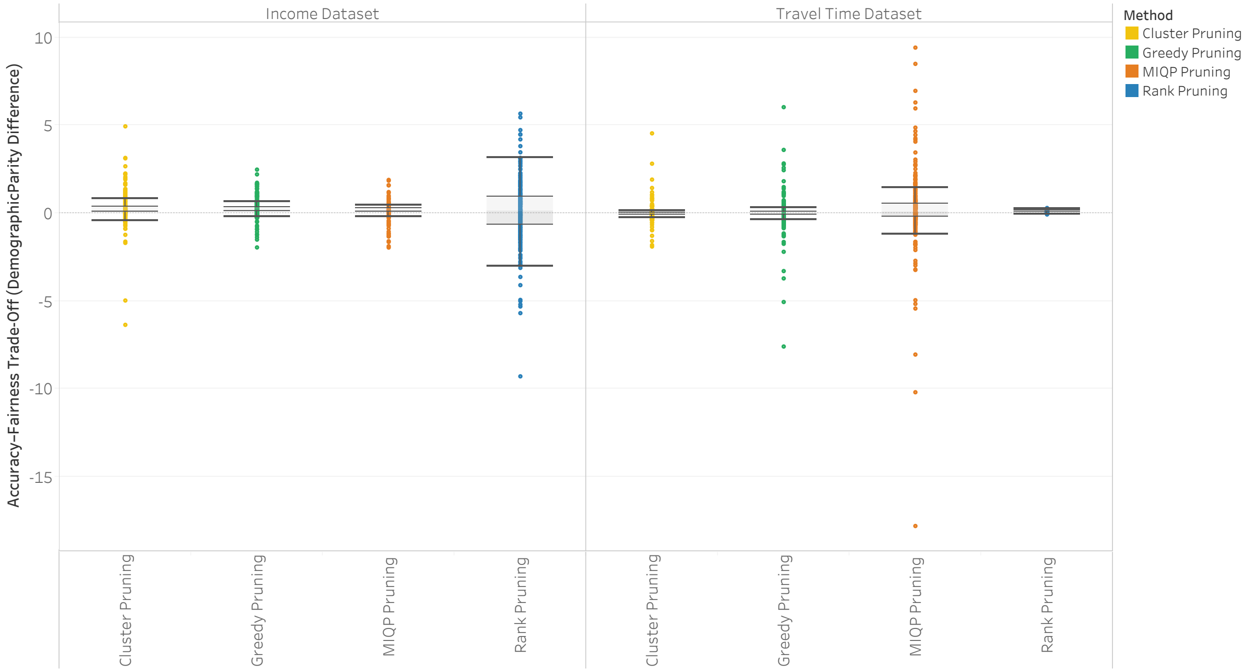


Figure 6.17 Box Plot of Accuracy–Fairness Trade-Off Ratios for Demographic Parity Difference This figure presents the distribution of the accuracy–fairness trade-off ratios for demographic parity difference across different pruning methods and both datasets (Income and Travel Time). Each color represents a specific pruning method, and separate grids correspond to each dataset. The box plots illustrate the defined trade-off measure used in the permutation factorial ANOVA analysis.

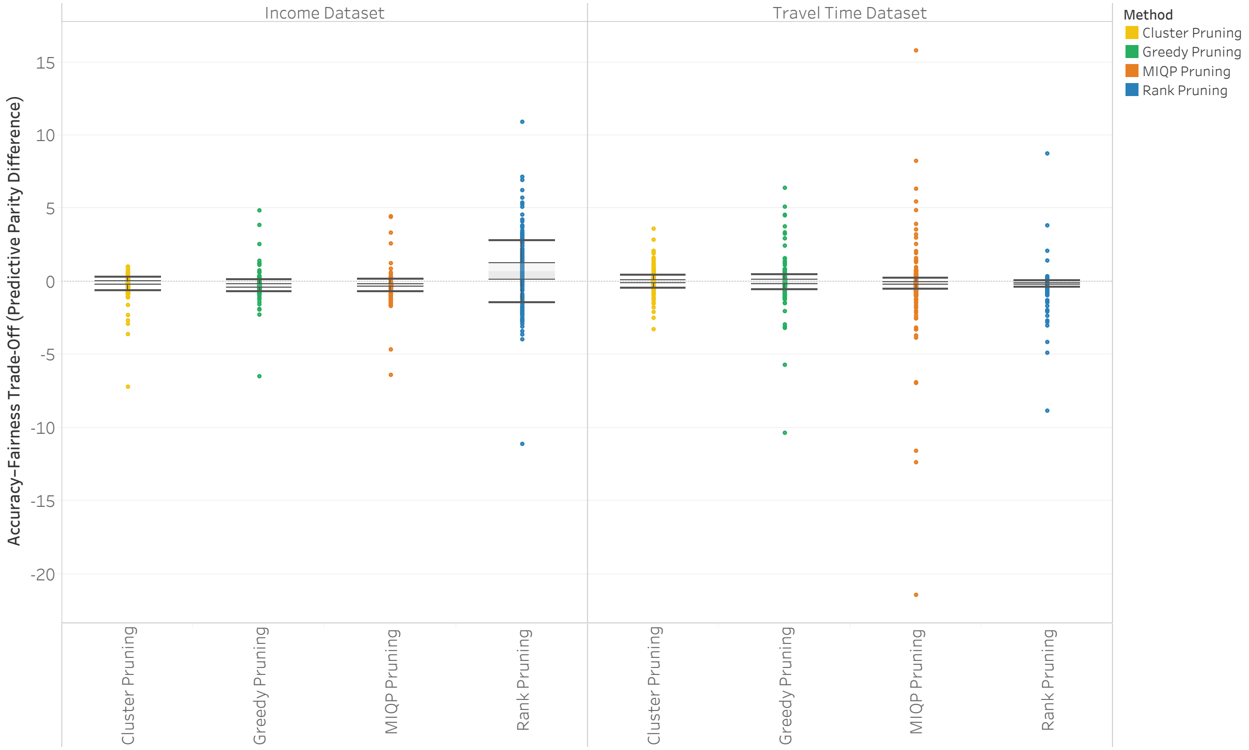


Figure 6.18 Box Plot of Accuracy–Fairness Trade-Off Ratios for Predictive Parity Difference This figure presents the distribution of the accuracy–fairness trade-off ratios for predictive parity across different pruning methods and both datasets (Income and Travel Time). Each color represents a specific pruning method, and separate grids correspond to each dataset. The box plots illustrate the defined trade-off measure used in the permutation factorial ANOVA analysis.

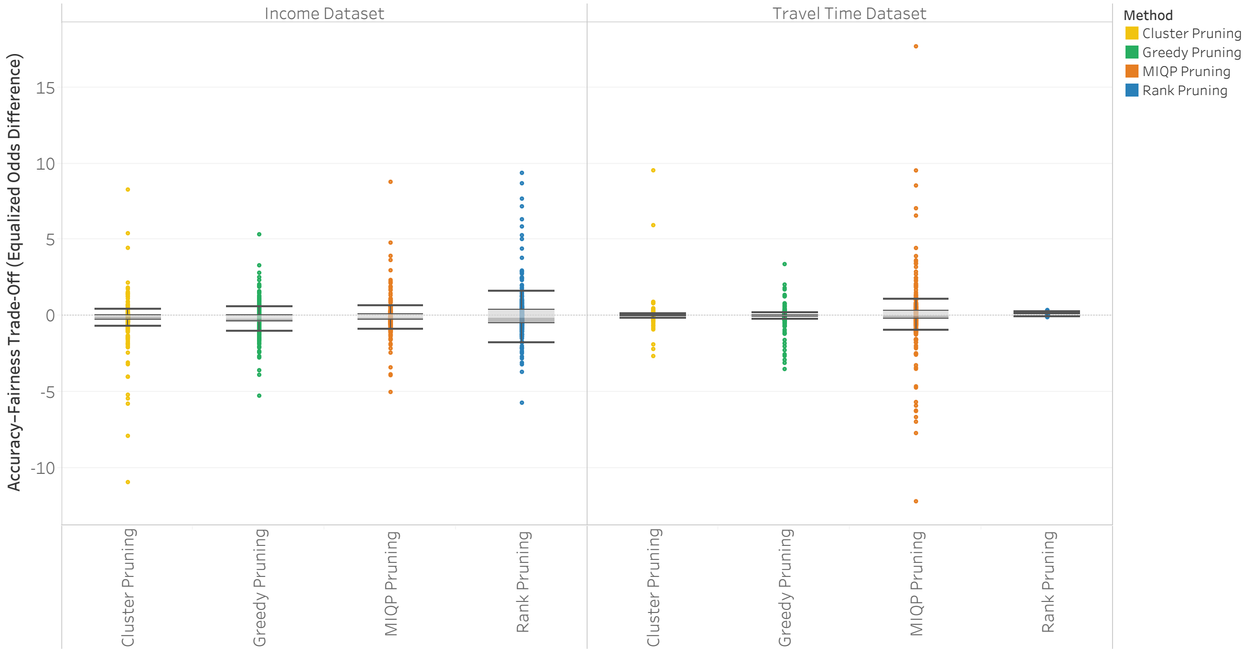


Figure 6.19 Box Plot of Accuracy–Fairness Trade-Off Ratios for Equalized Odds Difference This figure presents the distribution of the accuracy–fairness trade-off ratios for equalized odds difference across different pruning methods and both datasets (Income and Travel Time). Each color represents a specific pruning method, and separate grids correspond to each dataset. The box plots illustrate the defined trade-off measure used in the permutation factorial ANOVA analysis.

The analysis highlights that dataset characteristics consistently influence accuracy–fairness trade-offs, though the complexity of this impact varies by metric. Method choice strongly affects trade-offs involving predictive parity difference and equalized odds difference but has no significant influence on demographic parity difference. Pruning level has a modest direct effect for demographic parity difference, but becomes more important through interactions with method and dataset. Overall, predictive parity difference emerges as the most sensitive metric, showing the strongest and most complex interaction effects, while Equalized Odds Difference demonstrates simpler, less complex dependencies, showing no statistically or practically significant effects for the pruning percentage factor and different interactions of factors. Thus, the balance between accuracy and fairness is shaped by a combination of dataset, method, and pruning level, with their importance differing across fairness metrics.

No Universal Behavior: Dataset Characteristics Govern Pruning Trade-Offs and Metric Behavior

In the previous sections, we established that the dataset, pruning method, and pruning level each contribute to variations in fairness metrics and accuracy, with the dataset factor demonstrating the most dominant influence. While the method and pruning level also have significant impacts, their effects are more subtle and scenario-dependent. These findings suggest that pruning does not affect models in a simple manner. Instead, certain thresholds or tipping points emerge, where fairness metrics and accuracy can rapidly shift, and the magnitude of these changes is closely tied to the characteristics and inherent bias of the dataset.

While the previous parts of our analysis were primarily statistical, focusing on hypothesis testing and variance analysis, this section relies on different visualizations of the raw results to provide an intuitive and comprehensive understanding of model behavior. Here, we consolidate all previous insights and extend the analysis by adopting a scenario-based perspective. In this context, each scenario is defined as the combination of pruning method, dataset, and pruning level. We assess robustness in terms of both the stability and the direction of changes in accuracy and fairness metrics across different pruning intensities, as well as their relative interactions and trade-offs.

To structure this analysis, we first present a detailed exploration of the trade-off behavior between accuracy and fairness metrics through line plot visualizations, identifying common patterns and distinctive behaviors across scenarios. Following this, we complement these observations with other visualizations — including boxplots and correlation heatmaps — to quantify relationships, assess variability, and evaluate robustness across different pruning methods, datasets, and pruning levels.

Trends in Accuracy and Fairness Metrics Trade-off Across different scenarios

In the presented plots 6.20, we illustrate the relationship between accuracy and each fairness metric across varying pruning percentages for both datasets. Each line in the plot corresponds to one pruning method, with the line thickness representing the level of pruning applied — thicker lines indicate higher pruning percentages, while thinner lines represent lower pruning percentages. The initial state for each method is marked on the plot, and progression along each line corresponds to increasing pruning levels.

By following the trajectory along each line, we can observe how increasing the pruning level affects the trade-off between accuracy and fairness metrics. These visualizations are provided for both datasets and across all pruning methods, offering a comprehensive overview of their respective behaviors.

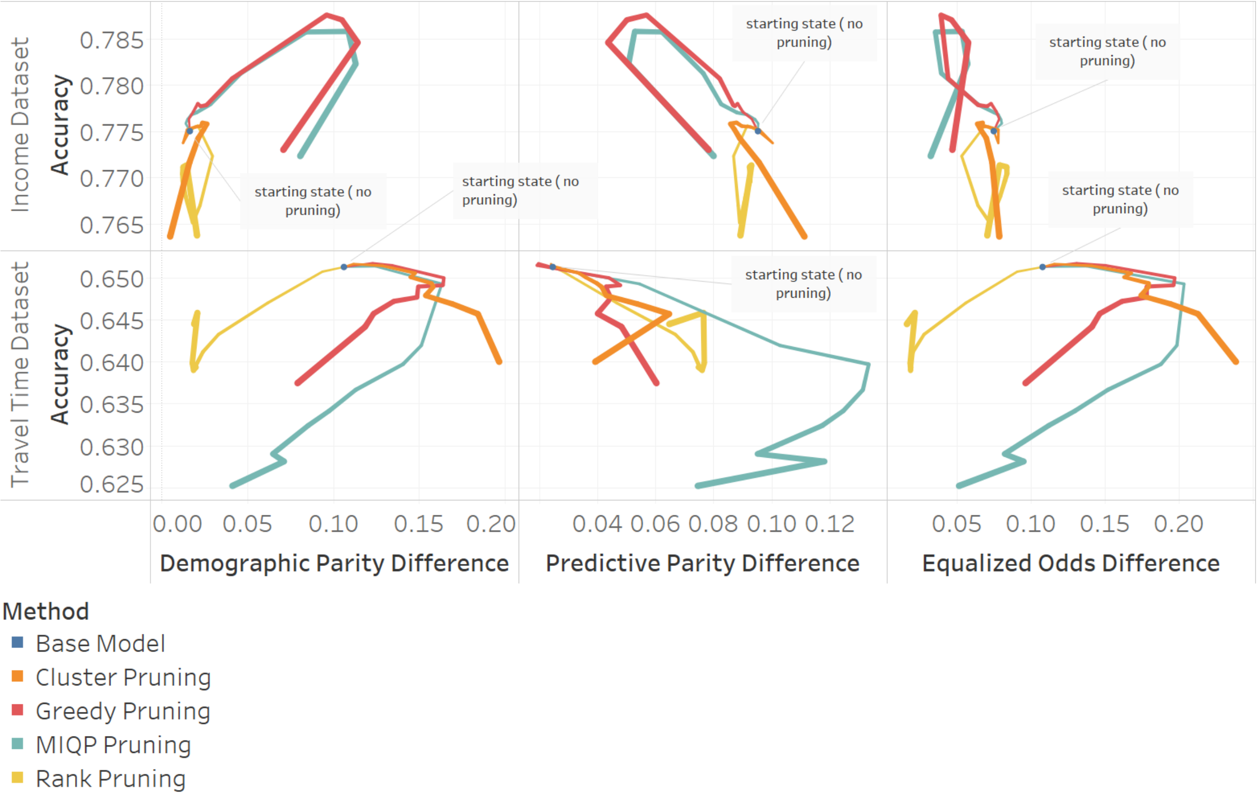


Figure 6.20 Line plots illustrating the trade-off between accuracy and each fairness metric across different datasets and pruning methods. Each color corresponds to a specific pruning method, while the thickness of each line represents the pruning level — thinner lines indicate lower pruning levels and thicker lines represent higher pruning levels. All methods originate from a common baseline point (the initial unpruned model) and progress along the lines as pruning intensity increases. This visualization allows comparison of how each method's accuracy–fairness trade-off evolves with increasing pruning.

This plot is particularly helpful for identifying similarities and differences in trade-off behavior between the pruning methods. For instance, on the Income dataset, MIQP and Greedy pruning methods demonstrate notably similar behavior patterns. Similarly, Cluster pruning and Rank pruning also exhibit comparable tendencies on this dataset. Overall, the methods applied to the Income dataset tend to display more stable and aligned behaviors.

In contrast, for the Travel Time dataset, each pruning method demonstrates distinct and unique patterns. There is little to no overlap in tarde-off behavior, suggesting that the methods respond differently to pruning on this dataset, highlighting dataset-dependent variability in fairness-accuracy trade-offs.

Another key observation from the plots is that, at more extreme pruning levels, the accuracy generally decreases for most methods. However, for MIQP and Greedy pruning on the Income dataset, accuracy initially increases up to around 80% pruning and only begins to decrease beyond that point. This observation is consistent with the earlier findings, and the plot clearly illustrates this behavior.

Additionally, for the Income dataset, it can be seen that the equalized odds difference remains relatively stable across different pruning levels. Even as the accuracy decreases at higher pruning percentages, the equalized odds difference stays within a narrow range, showing almost no significant changes.

For the Travel Time dataset, another notable trend emerges: as pruning levels increase, the demographic parity difference decreases for Rank pruning, Greedy, and MIQP methods, while for Cluster pruning, it increases. This same contrasting behavior is observed in the equalized odds difference, where Rank pruning, Greedy, and MIQP show decreasing trends, while Cluster pruning shows an increasing trend. Thus, Cluster pruning displays a different behavior compared to the other three methods.

Moreover, for the predictive parity difference in the Travel Time dataset, we observe that it generally increases as the pruning level increases. However, for Cluster pruning, the direction of change differs from the other methods, indicating a distinct pattern in how predictive parity responds to pruning in this case.

From the perspective of the line plots, these trends become even more evident for each metric as the pruning level increases. In terms of accuracy, both the Income and Travel Time datasets exhibit some similar patterns. Notably, Rank pruning consistently performs the worst in both datasets. However, the overall performance of all methods on the accuracy of the Travel Time dataset is lower compared to the Income dataset. Additionally, for the Income dataset, the optimization-based pruning methods (MIQP and Greedy) show some improvement in accuracy at moderate pruning levels before declining at higher pruning levels.

When analyzing the demographic parity difference, it is evident that for the Travel Time dataset, Cluster pruning exhibits significant shifts and large variances across different pruning levels. In contrast, on the Income dataset, the variations are smaller and more stable. Furthermore, Greedy and MIQP pruning demonstrate similar behaviors on the Income dataset, with both methods showing relatively controlled trends until reaching more extreme pruning levels.

As we move toward more extreme pruning levels, the predictive parity difference often decreases initially and then increases. For both Greedy and MIQP pruning, there is a notable sharp decrease in demographic parity difference and equalized odds difference at around the 90% pruning level, accompanied by a large increase in predictive parity difference at that same point.

For the Travel Time dataset, Rank pruning shows strong performance in reducing both equalized odds difference and demographic parity difference, but it performs poorly with respect to predictive parity difference. In almost all cases for the Income dataset, predictive parity difference worsens at extreme pruning levels, particularly under MIQP pruning. The same trend can be observed for equalized odds difference: except for Rank pruning, most methods worsen this metric up to a certain point. Beyond that point — typically between medium to extreme pruning levels — MIQP and Greedy pruning begin to reduce the equalized odds difference, showing a shift in direction. Interestingly, both MIQP and Greedy pruning methods display similar drop points where this change occurs, and this observation is also supported by insights presented earlier in the analysis (Insight 2 and Insight 3). These shifts highlight how these methods can behave differently at extreme pruning levels.

A similar pattern can be seen in demographic parity difference. Cluster pruning continues to worsen the metric at extreme pruning levels, while MIQP and Greedy pruning methods, after a certain threshold, begin to decrease the unfairness, indicating that medium to high pruning percentages are needed to achieve better fairness outcomes with these methods.

Among all methods, Cluster pruning and Rank pruning display more predictable and stable behaviors. Their trends generally follow a steady pace, either consistently increasing or decreasing, depending on the dataset characteristics and the metric in question. In contrast, MIQP and Greedy pruning show less stable patterns; they often follow one direction (increasing or decreasing) up to a medium or high pruning threshold, after which they sharply shift direction and move in the opposite trend. This characteristic non-monotonic behavior distinguishes MIQP and Greedy pruning from the more steady patterns observed in Cluster and Rank pruning methods.

While the above analysis provides an intuitive understanding of metric trends through line plots, we now extend this analysis by incorporating statistical relationships, distributional patterns, and robustness assessments using additional visualization techniques.

6.5.1 Trends in Accuracy and Fairness Metrics behavior Across different scenarios

This analysis is conducted using complementary visualizations, including correlation heatmaps, boxplots, and line charts. Each visualization provides a unique perspective: correlation heatmaps quantify monotonic relationships between pruning levels and metrics; boxplots reveal distributional characteristics and variance across scenarios; and line charts illustrate average trends and overall trajectories. By integrating these visual tools, we aim to identify stable and unstable regimes and assess the robustness of metrics under varying pruning intensities.

The boxplots and line charts, presented in Figures 6.21–6.29, reveal several consistent patterns across both datasets. For low pruning levels, methods display stable and compact distributions for both fairness metrics and accuracy, indicating consistency and minimal variability. However, as pruning levels increase, the variance of all metrics grows. This widening of distributions and the appearance of more extreme outliers highlight increasing instability, particularly in the Travel Time dataset. The divergence between methods also becomes more pronounced at higher pruning levels, with some methods demonstrating marked upward shifts in fairness differences and declines in accuracy, while others maintain stability. These patterns are less severe in the Income dataset, where models show better balance and more predictable trade-offs. Outlier behavior becomes increasingly prominent under extreme pruning, reflecting inconsistency and volatility in certain methods. In the Travel Time dataset, higher pruning levels lead to significant instability in fairness metrics and varying declines in accuracy. Some methods demonstrate resilience by preserving stability or even improving fairness while maintaining performance, whereas others exhibit clear trade-offs. Conversely, in the Income dataset, several methods maintain balanced stability or show improvements across both fairness and accuracy metrics, indicating robustness even at high pruning levels.

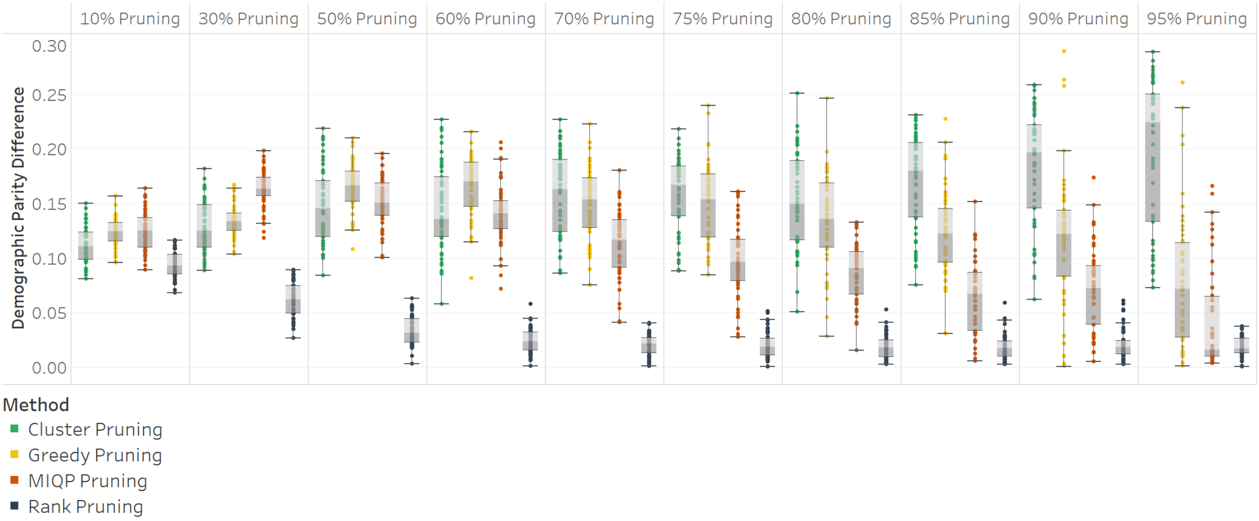


Figure 6.21 Box plots illustrating the distribution of Demographic Parity Difference across varying pruning levels for the Travel Time dataset. Each color represents a specific pruning method, and each panel corresponds to a distinct pruning level. This visualization highlights how each metric evolves with increasing pruning intensity for each method.

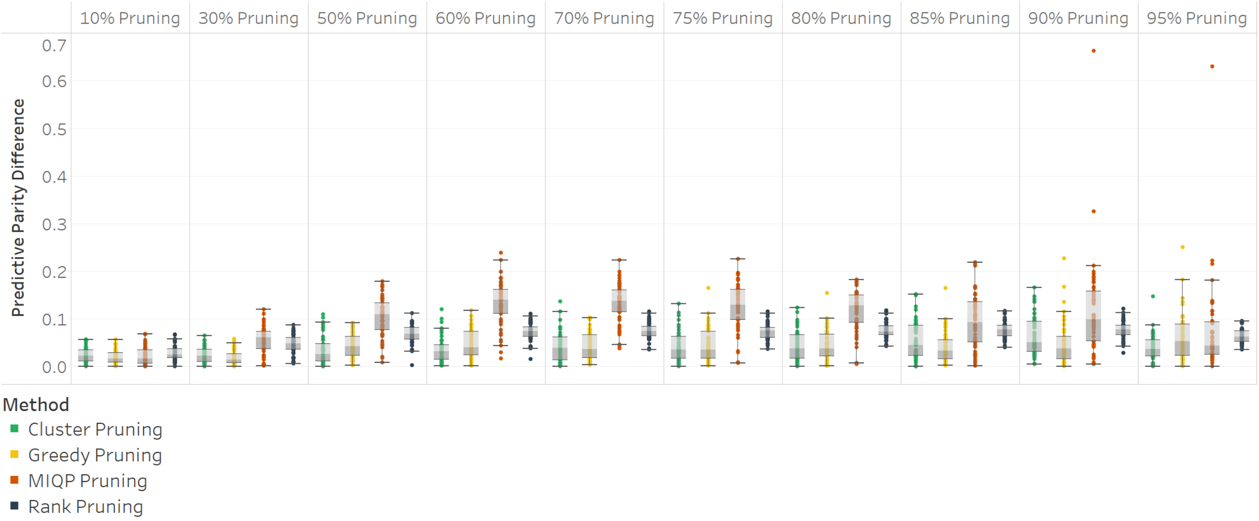


Figure 6.22 Box plots illustrating the distribution of Predictive Parity Difference across varying pruning levels for the Travel Time dataset. Each color represents a specific pruning method, and each panel corresponds to a distinct pruning level. This visualization highlights how each metric evolves with increasing pruning intensity for each method.

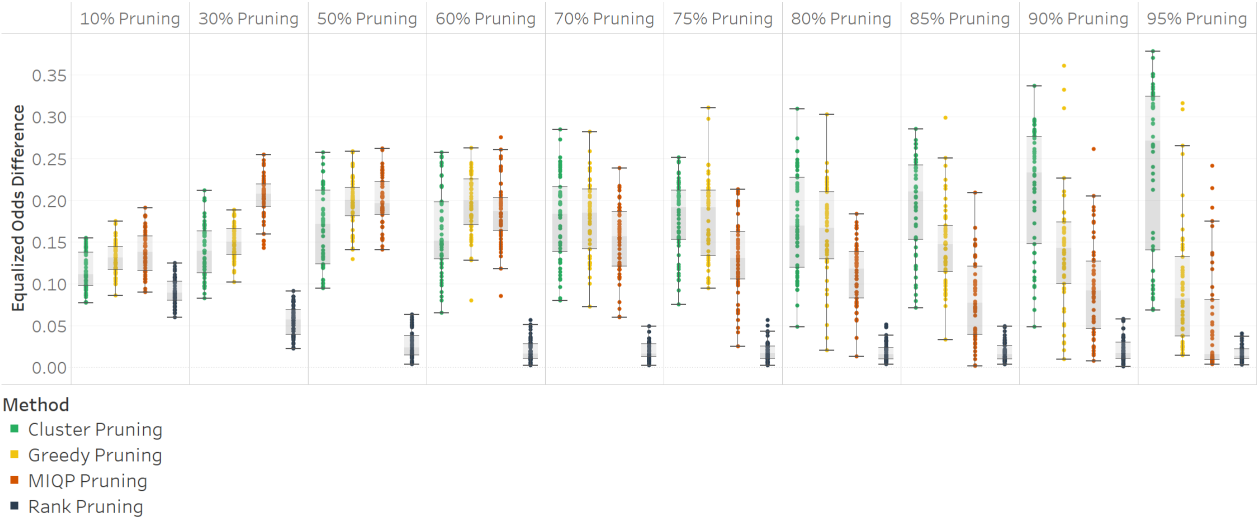


Figure 6.23 Box plots illustrating the distribution of Equalized Odds Difference across varying pruning levels for the Travel Time dataset. Each color represents a specific pruning method, and each panel corresponds to a distinct pruning level. This visualization highlights how each metric evolves with increasing pruning intensity for each method.

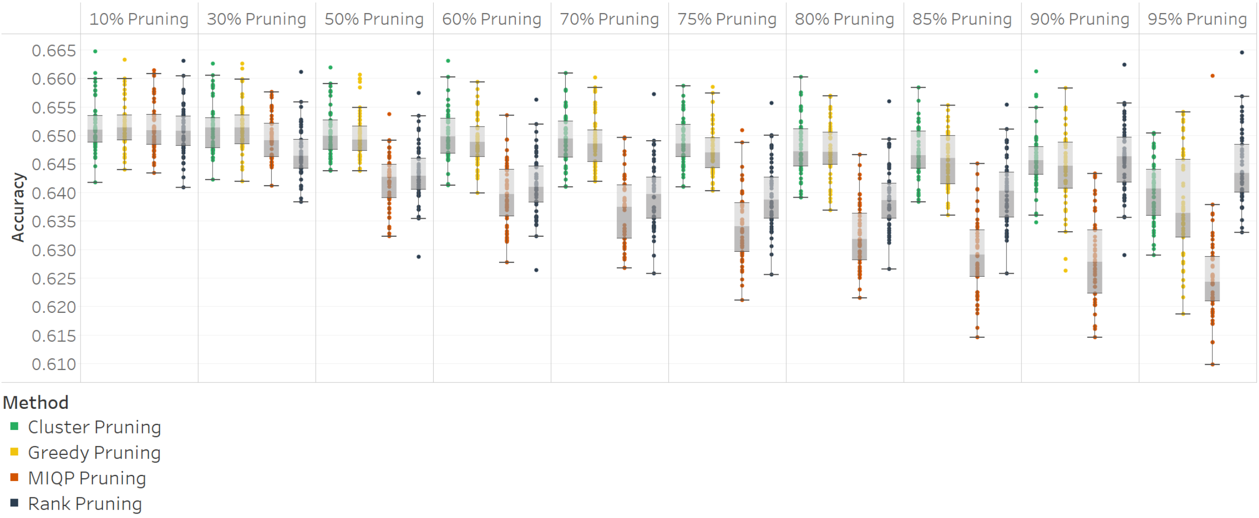


Figure 6.24 Box plots illustrating the distribution of Accuracy across varying pruning levels for the Travel Time dataset. Each color represents a specific pruning method, and each panel corresponds to a distinct pruning level. This visualization highlights how each metric evolves with increasing pruning intensity for each method.

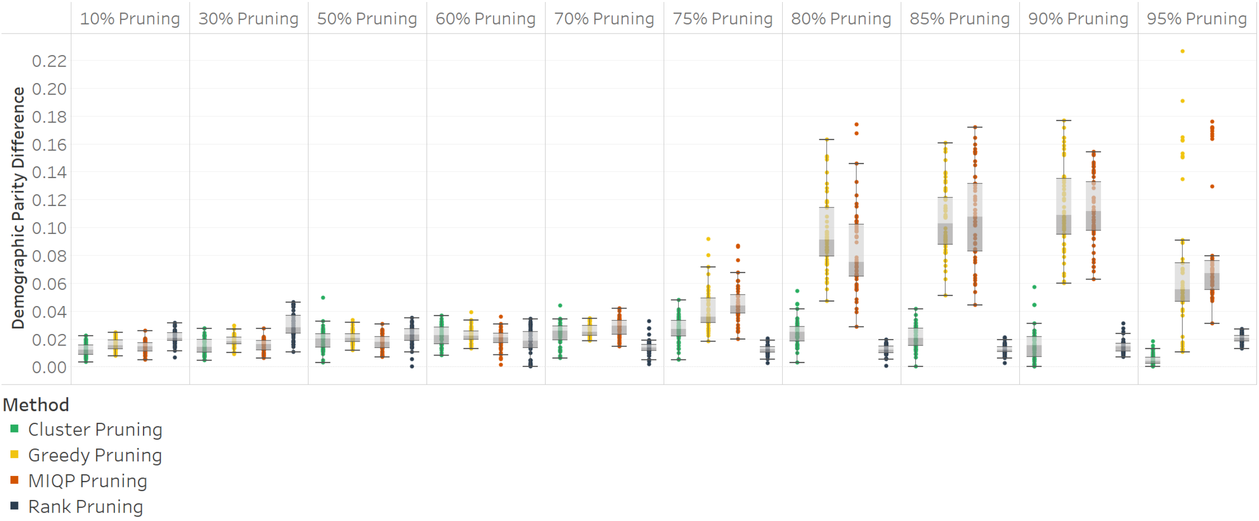


Figure 6.25 Box plots illustrating the distribution of Demographic Parity Difference across varying pruning levels for the Income dataset. Each color represents a specific pruning method, and each panel corresponds to a distinct pruning level. This visualization highlights how each metric evolves with increasing pruning intensity for each method.

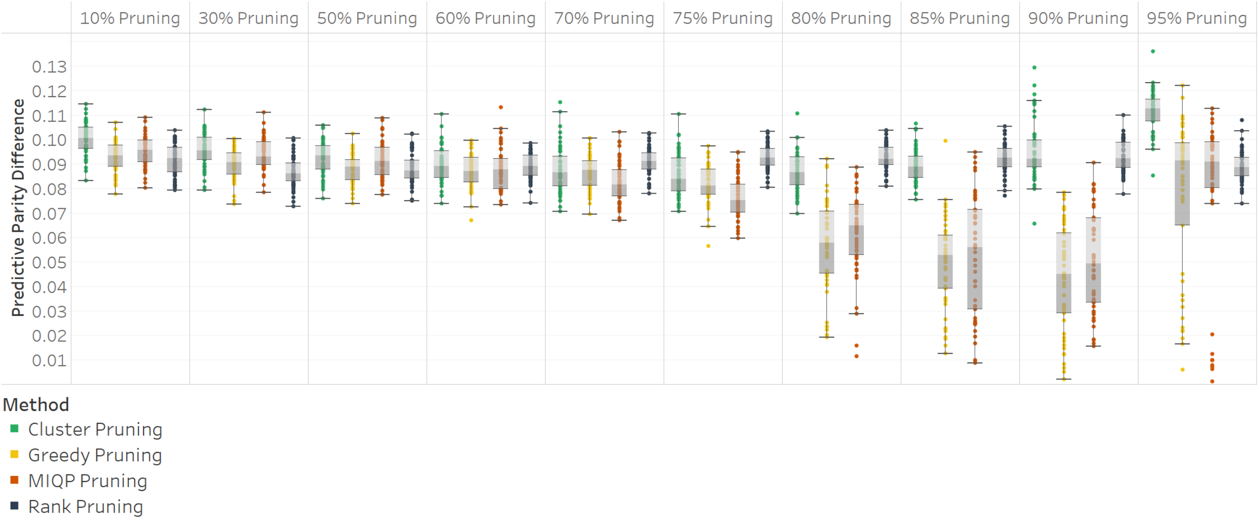


Figure 6.26 Box plots illustrating the distribution of Predictive Parity Difference across varying pruning levels for the Income dataset. Each color represents a specific pruning method, and each panel corresponds to a distinct pruning level. This visualization highlights how each metric evolves with increasing pruning intensity for each method.

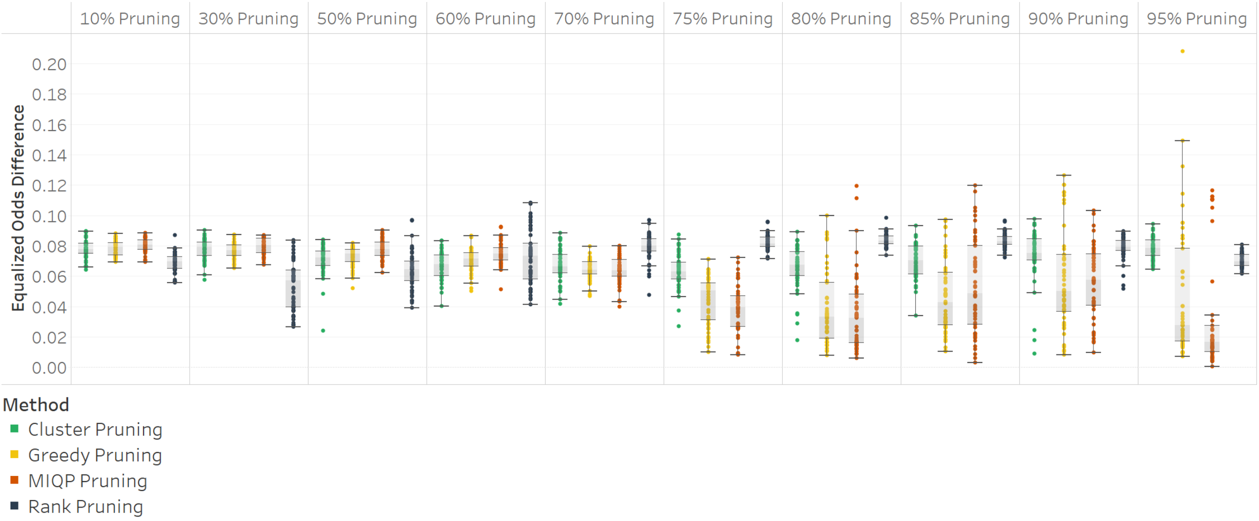


Figure 6.27 Box plots illustrating the distribution of Equalized Odds Difference across varying pruning levels for the Income dataset. Each color represents a specific pruning method, and each panel corresponds to a distinct pruning level. This visualization highlights how each metric evolves with increasing pruning intensity for each method.

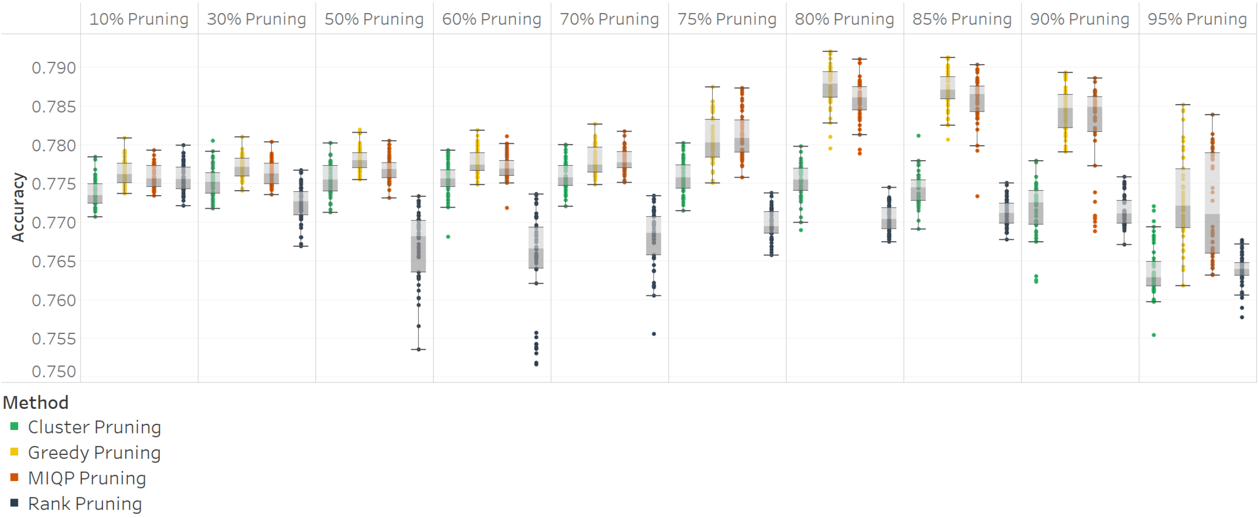


Figure 6.28 Box plots illustrating the distribution of Accuracy across varying pruning levels for the Income dataset. Each color represents a specific pruning method, and each panel corresponds to a distinct pruning level. This visualization highlights how each metric evolves with increasing pruning intensity for each method.

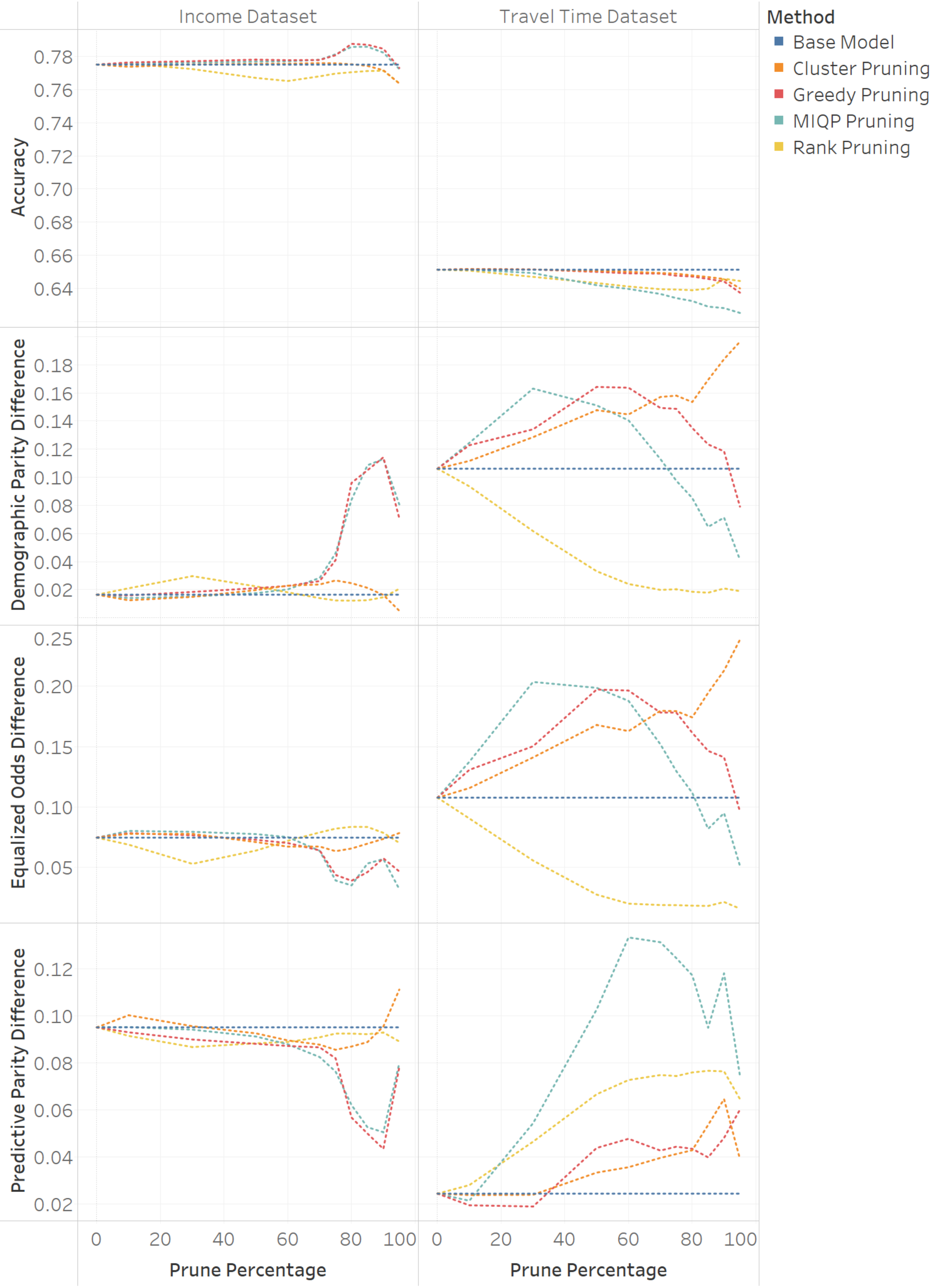


Figure 6.29 Line plots illustrating the behavior of each fairness metric and accuracy across different pruning percentages. Each color represents a specific pruning method, and each panel corresponds to a different dataset. The plots show how each metric evolves as pruning intensity increases, allowing for comparisons between methods and datasets in terms of metric stability and sensitivity to pruning.

The correlation heatmap analysis further supports these observations by quantifying the strength and direction of relationships between pruning percentage and metrics for each pruning method and dataset. While pruning levels are represented by discrete values (e.g., 10%, 20%, . . . , 90%), they are inherently ordinal and exhibit a natural ordering. On the other hand, fairness metrics such as Demographic Parity Difference, Equalized Odds Difference, and Predictive Parity Difference are continuous variables. To appropriately assess the relationship between an ordinal predictor and continuous response variables without assuming linearity, Spearman's rank correlation coefficient is employed. Spearman correlation measures the strength and direction of monotonic relationships by converting both variables into their respective ranks before computing the correlation. This allows for capturing trends where metrics increase or decrease in a consistent manner as pruning levels change, even if the relationship is not strictly linear. By using Spearman correlation, we ensure that the ordinal nature of pruning levels is respected and that non-linear but monotonic patterns are effectively detected, providing robust and interpretable insights into how pruning intensities influence fairness metrics.

For the Rank Pruning method, correlation heatmaps 6.30 show a moderate negative correlation between accuracy and pruning percentage for both datasets, confirming that accuracy declines with increased pruning. Accuracy consistently declines with higher pruning levels across both datasets. It also signals that aggressive pruning sacrifices predictive performance. In the Travel Time dataset, fairness metrics fluctuate sharply, with demographic parity difference decreasing and predictive parity difference increasing alongside higher pruning levels. Equalized odds difference also decreases, indicating mixed impacts on fairness. In the Income dataset, these correlations are moderate and more stable, suggesting less sensitivity to pruning. Demographic parity difference decreases moderately with increased pruning, while equalized odds difference increases. Predictive parity difference, however, appears relatively independent of pruning levels, indicating stability in this fairness dimension despite pruning.

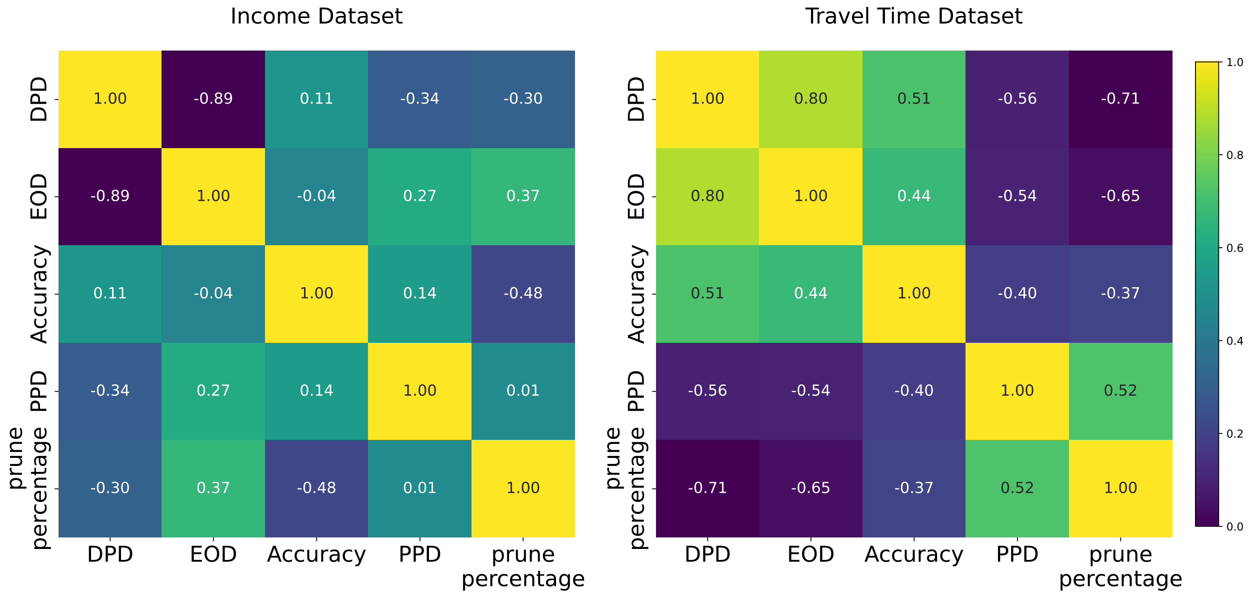


Figure 6.30 Correlation heatmap between accuracy and pruning percentage for the Rank Pruning method.

Across all datasets and pruning methods, accuracy and pruning percentage generally exhibit moderate to strong negative correlations. This indicates that higher pruning percentages tend to reduce the classifier's predictive performance, aligning with expected behavior. However, exceptions exist. According to the correlation heatmap 6.31, the MIQP pruning method displays an interesting pattern with moderate positive correlation between accuracy and pruning percentage for the Income dataset, contrasting with a negative relationship for the Travel Time dataset. This suggests that higher pruning percentages may be associated with slight improvements or stability in accuracy, potentially due to the removal of unstable or noisy branches. Fairness metrics in the Income dataset show moderate negative correlations with pruning percentage for equalized odds difference and predictive parity difference, while demographic parity difference is strongly positively correlated, illustrating complex trade-offs. On the other hand, the Travel Time dataset shows moderate correlations between fairness metrics and pruning percentage. Predictive parity difference increases as pruning intensifies, while demographic parity difference and equalized odds difference decrease.

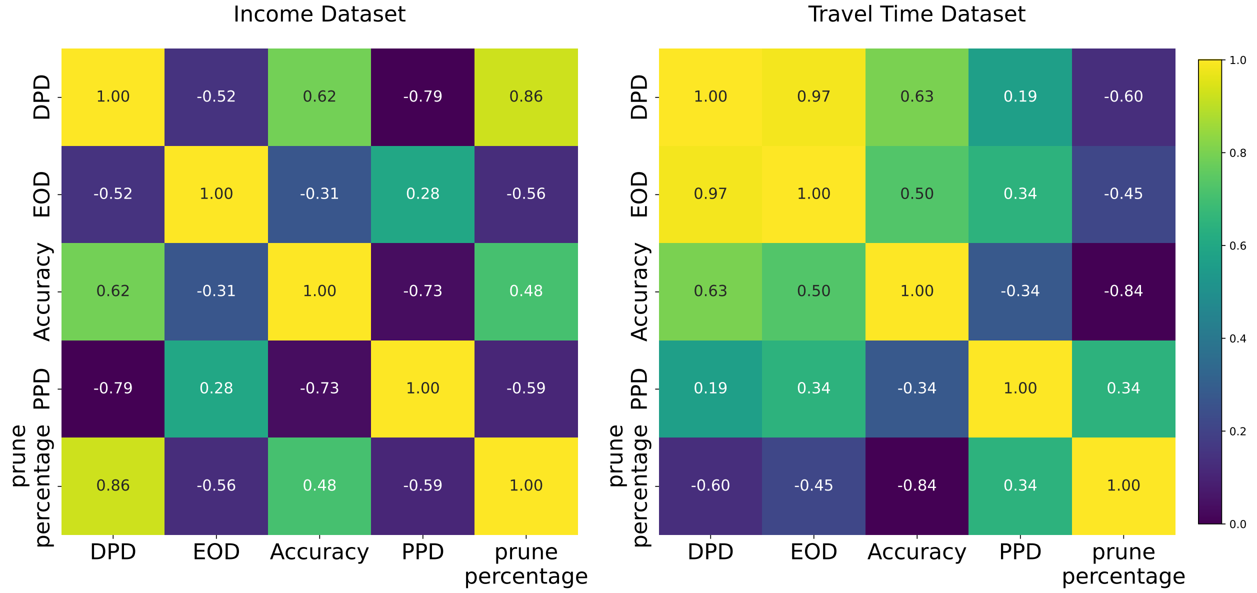


Figure 6.31 Correlation heatmap between accuracy and pruning percentage for the MIQP Pruning method.

Cluster pruning shows mild negative correlations between accuracy and pruning levels across both datasets Figure 6.32, with stronger impacts in the Travel Time dataset. This reduction in accuracy remains within a reasonable and acceptable range, indicating controlled degradation. In the Income dataset, fairness metrics are largely unaffected by pruning percentage, while in the Travel Time dataset, moderate positive correlations indicate that increasing pruning can lead to fairness degradation.

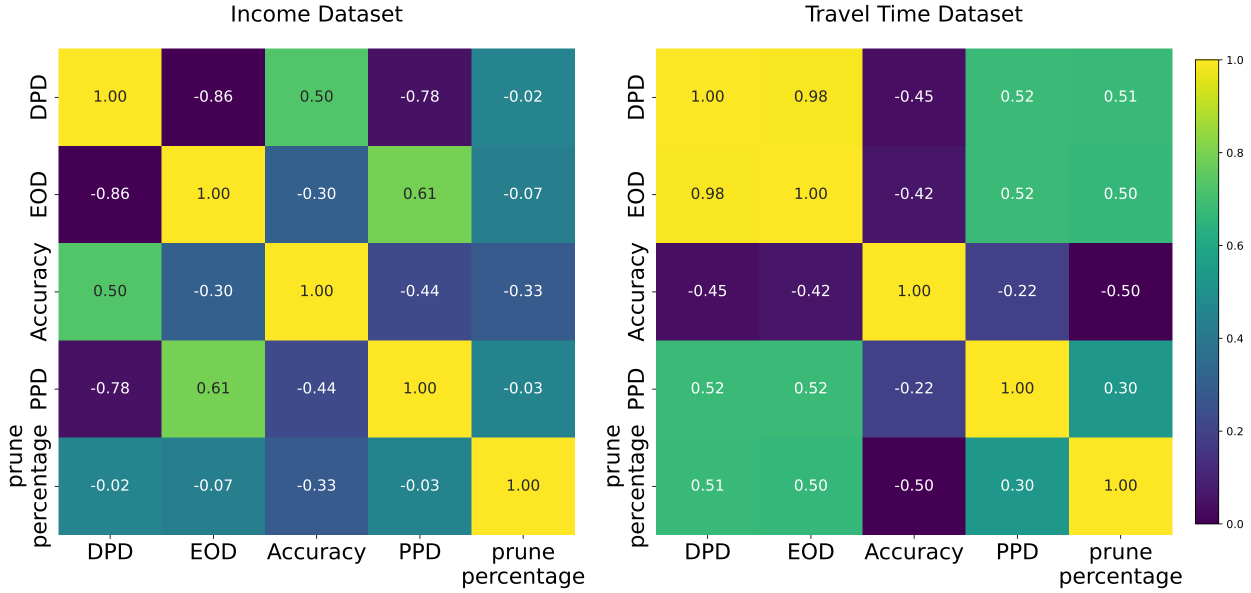


Figure 6.32 Correlation heatmap between accuracy and pruning percentage for the Cluster Pruning method.

For the Greedy pruning method 6.33, accuracy shows a moderate positive correlation with pruning percentage in the Income dataset and a negative correlation in the Travel Time dataset. This is similar to MIQP behavior. Notably, they are the only two methods that show accuracy improvements in certain scenarios, and both belong to the same category of pruning methods (optimization-based). In terms of fairness, the Income dataset displays moderate to strong correlations: demographic parity difference increases with pruning, while equalized odds difference and predictive parity difference decrease. In the Travel Time dataset, fairness metrics show low-magnitude relationships, indicating relative stability.

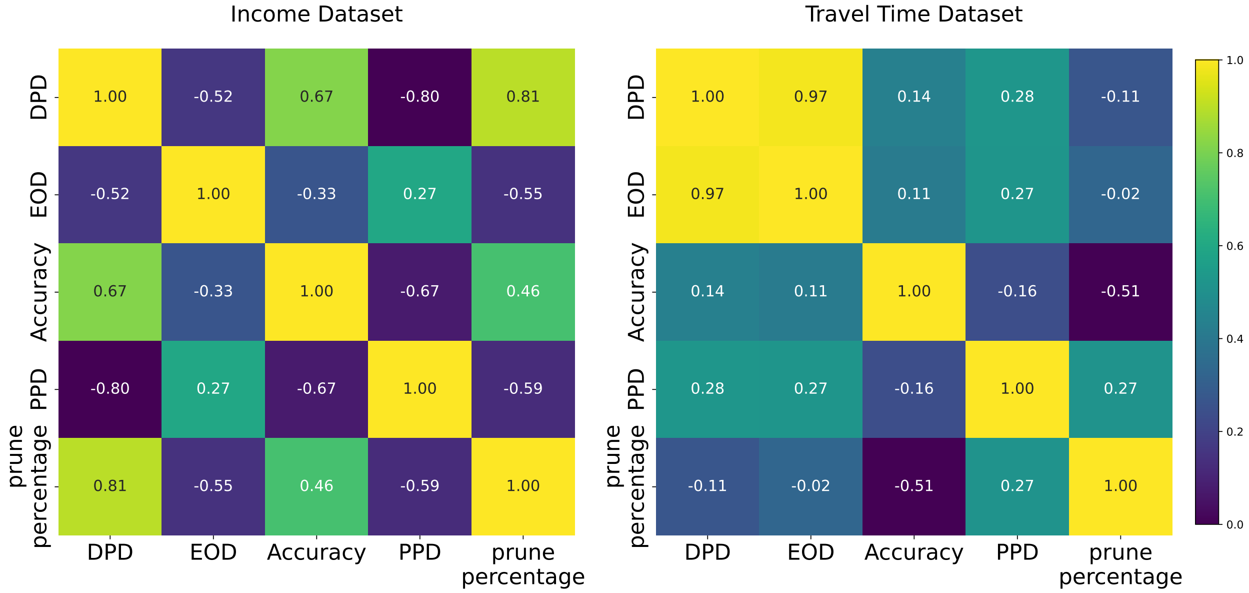


Figure 6.33 Correlation heatmap between accuracy and pruning percentage for the Greedy Pruning method.

Fairness Metrics vs. Pruning Levels

Overall, the combined analysis of boxplots, line charts, and correlation heatmaps reveals that pruning level has a complex and scenario-dependent impact on both accuracy and fairness metrics. The Income dataset consistently demonstrates greater stability and predictability across metrics and pruning methods. In contrast, the Travel Time dataset exhibits more volatility, larger spreads, and more pronounced trade-offs between fairness and accuracy. In short, pruning impacts are more controlled and predictable in the Income dataset, with less volatility between methods, while in the Travel Time dataset, pruning leads to more divergent, less stable outcomes. Some methods, particularly optimization-based methods like MIQP and Greedy, show potential for maintaining or even improving accuracy at moderate to extreme pruning levels, but their fairness behaviors become less predictable at higher pruning intensities based on the sensitivity of the dataset, in our case Travel Time dataset.

Based on the analysis, we can also identify some outstanding behavior for each method according to our case study. Rank pruning interacts differently with fairness metrics across datasets, indicating that this method must account for pruning levels and dataset-inherent biases carefully. Very strong shifts in fairness metrics under rank pruning make it risky at extreme pruning levels. Sharp fluctuations highlight that if fairness is critical, applying rank pruning at high levels leads to unpredictable behavior. Among all methods, the Greedy pruning method emerges as one of the most stable and reliable options for preserving the initial fairness state of the Travel Time dataset, whereas other methods display higher variability and shifts in metrics on this sensitive dataset. Compared to other methods, cluster pruning demonstrates more stability in terms of metric shifts, particularly for the Income dataset, and can be a good choice for extreme pruning levels if needed. However, extreme pruning should still be applied with caution, as moderate correlations between pruning percentage and all metrics persist.

These findings emphasize the importance of considering dataset characteristics, initial biases, and specific objectives when applying pruning methods. As along the all sections of our analysis shown, the dataset itself, along with its inherent bias, plays the most significant role in determining metric behavior and outcomes. No single method consistently outperforms others across all scenarios. Instead, the choice of pruning method and pruning intensity must be carefully aligned with the intended balance between fairness and accuracy, as well as the stability requirements of the application context.

6.6 Multi-Metric Trade-Offs: Final Observations and Recommendations from Our Case Study

This study has focused on evaluating four pruning methods — *ClusterPruningClassifier*, *GreedyPruningClassifier*, *MIOPruningClassifier*, and *RankPruningClassifier* — on two distinct datasets: in our study and based on our initial dataset analysis one considered less stable (the *Travel time* dataset) and another more stable (the *Income* dataset). The analysis was conducted across varying pruning percentages and multiple evaluation metrics: *Accuracy*, *Demographic Parity Difference*, *Equalized Odds Difference*, *Predictive Parity Difference*. In this final section, we extend our comparison by incorporating additional metrics — *Elapsed Time*, *Maximum RSS* (memory consumption), and *Total CPU Usage* — to evaluate the trade-offs between model performance, computational cost, and time efficiency.

All experiments were conducted on the MILA cluster, using nodes with NVIDIA A100 (40GB/80GB), V100 (32GB), or RTX8000 (48GB) GPUs. Jobs were submitted with a SLURM configuration of 16 CPUs per task, 32 GB memory per CPU, and a maximum runtime of 10 hours. Resource and time metrics were collected from SLURM output logs.

In this final wrap-up, the relationships between these metrics are visualized using **parallel coordinate plots** in Figures 6.34 and 6.35. In our case, each line in the parallel coordinate plot represents a specific **scenario** — defined as the combination of a given dataset, pruning method, and pruning level. For each scenario, all metrics were **normalized** to a common scale before plotting. This normalization allows for a fair comparison between metrics that are measured on different scales (for example, accuracy vs. elapsed time), and emphasizes relative changes rather than absolute values. This visualization approach provides an intuitive and holistic view of the complex trade-offs between accuracy, fairness, and computational resource consumption.

Context-Aware Observations and Recommendations. While it is important to emphasize that these findings are derived from our case study of two datasets and four methods — and therefore should not be considered general rules — a few cautious observations can be highlighted. These observations are made in the context of the extensive analyses performed in earlier sections, including variance analysis, stability exploration, and fairness stability assessment.

Accuracy Trends. Across both datasets, accuracy remained relatively stable at low to mid-range pruning percentages (0–30%) but started to decline at higher pruning levels. This decline was more pronounced in the less stable Travel dataset and more muted in the more stable Income dataset. Based on these findings, practitioners are advised to maintain pruning percentages within this stable range to preserve model accuracy, especially when working with datasets that exhibit variability or instability. It is important to note that this observation holds for general and basic pruning methods; if the pruning method is specifically configured to preserve accuracy, these trends might differ.

Fairness Metrics (Demographic Parity, Equalized Odds, Predictive Parity). Mid-level pruning (typically between 30% and 60%) does not result in significant practical shifts in fairness metrics compared to the initial state across both datasets and all methods. Therefore, this range of pruning can be regarded as a cautiously reliable zone. In some cases, mid-level pruning even led to fairness improvements, which were smoother and more predictable in the Income dataset, while the Travel dataset exhibited more fluctuations. However, practitioners should be prepared for more volatility in less stable datasets and apply incremental adjustments while closely monitoring changes.

Balancing accuracy and fairness. For those aiming to achieve a balance between accuracy and fairness, moderate pruning levels (approximately 30–60%) are recommended. This balance appears robust across both stable and less stable datasets but should always be verified within the specific dataset context. Practitioners are encouraged to regularly assess both fairness and performance throughout the pruning process to avoid unintended degradation in predictive capabilities.

Dataset stability consideration. The stable Income dataset demonstrated more predictable and steady metric trends under pruning, suggesting that stable datasets may tolerate more aggressive pruning with fewer risks of metric shifts. In contrast, the Travel dataset, being less stable, required more conservative pruning strategies and careful observation of metric behavior. This highlights the importance of evaluating dataset characteristics before deciding on pruning intensities. Furthermore, across all tests and analyses, the dataset itself showed a significant impact on metric behavior. This indicates that practitioners should prioritize understanding the dataset's inherent bias and characteristics above all when aiming to prune while preserving fairness.

Choosing pruning methods. The choice of pruning method is equally critical. Cluster and Greedy pruning methods tended to preserve accuracy for longer, although they sometimes offered smaller improvements in fairness or even caused fairness deterioration. On the other hand, MIQP and Rank pruning methods demonstrated higher potential for fairness improvements but introduced more variability in outcomes. Among all methods, cluster pruning showed the most stability across scenarios, while rank pruning exhibited greater variability, making it less reliable for fairness-aware pruning. MIQP and greedy pruning methods displayed differing behaviors under extreme pruning levels, suggesting that these methods may require more careful calibration and sensitivity analysis, particularly for less stable datasets or higher pruning intensities.

Resource efficiency. High pruning levels (60–85%) can lead to significant reductions in computational resource usage, especially in methods like cluster-based pruning or backward greedy pruning. Nevertheless, caution is advised, and such levels should only be applied with continuous monitoring to safeguard model stability and fairness integrity.

Final Note

All observations presented are derived from controlled experiments conducted on two widely used datasets in fairness research and applied to four common structured ensemble pruning methods. While these findings offer valuable guidance for practitioners — helping to raise awareness of potential trade-offs and key focus areas such as fairness monitoring, computational efficiency, and model stability — they should not be viewed as universally applicable rules. It is strongly advised that each deployment scenario undergo its own thorough testing and validation process, supported by a dedicated fairness and stability monitoring framework.

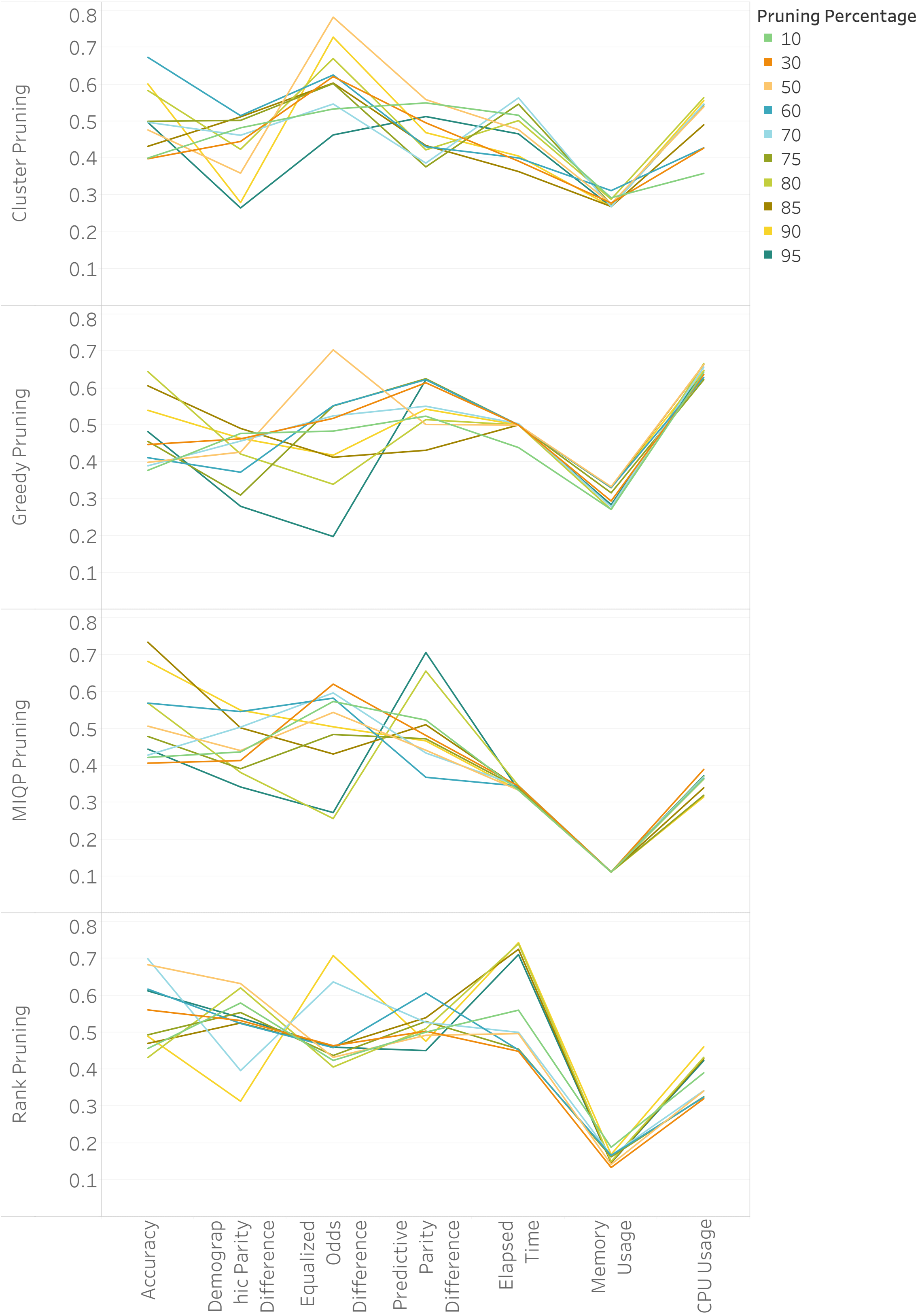


Figure 6.34 Parallel coordinate plots illustrating accuracy, fairness metrics, and resource usage across varying pruning levels for the Income dataset. Each color represents a distinct pruning intensity, and each pane corresponds to one pruning method. The plots highlight the trade-offs between predictive performance, fairness, and resource efficiency, offering visual insight into how different pruning levels and methods impact these dimensions.

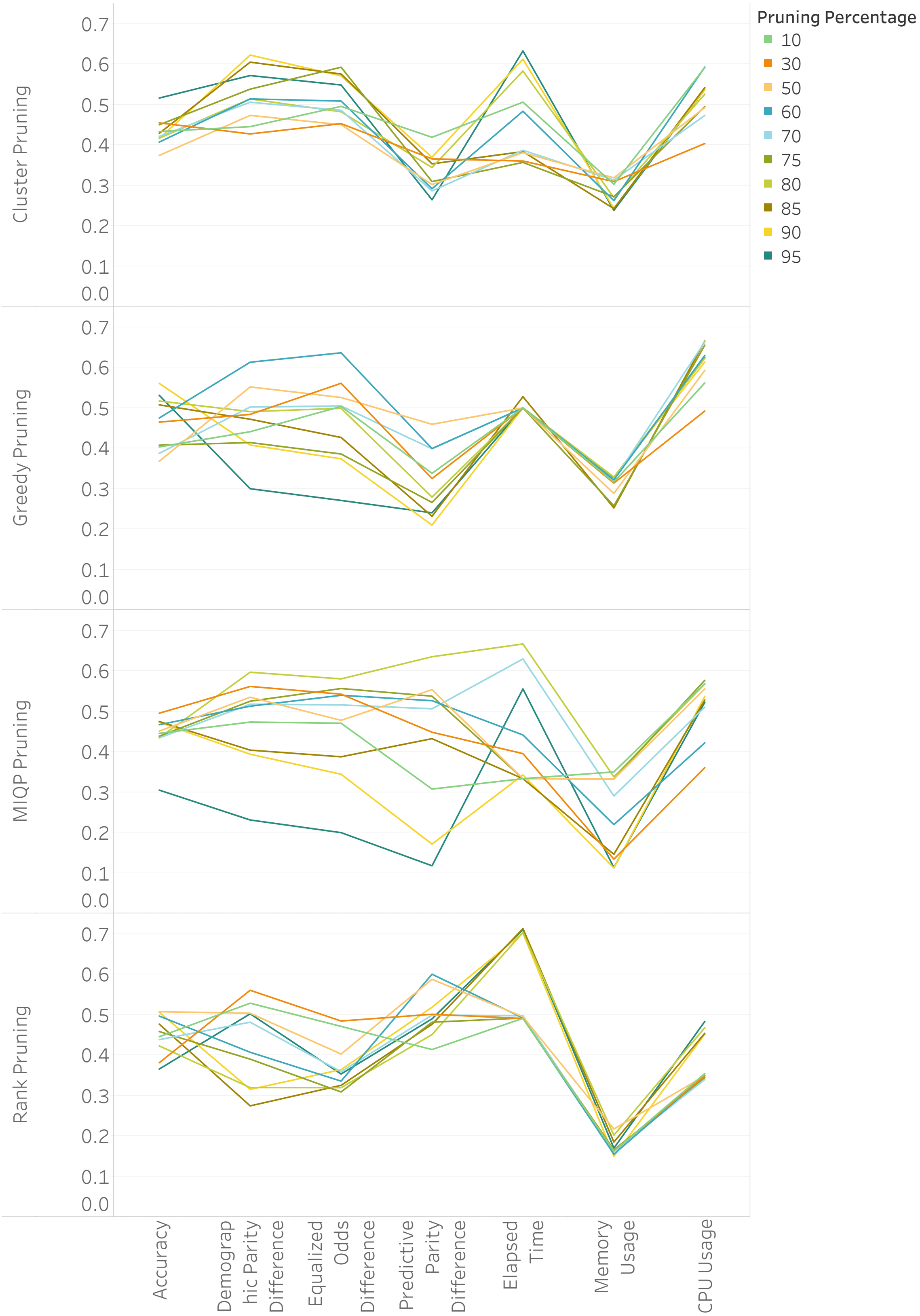


Figure 6.35 Parallel coordinate plots illustrating accuracy, fairness metrics, and resource usage across varying pruning levels for the Travel Time dataset. Each color represents a distinct pruning intensity, and each pane corresponds to one pruning method. The plots highlight the trade-offs between predictive performance, fairness, and resource efficiency, offering visual insight into how different pruning levels and methods impact these dimensions.

CHAPTER 7 CONCLUSION

In this study, we conducted a comprehensive and multi-dimensional analysis of the impact of pruning techniques on the fairness and predictive performance of random forest models. Through systematic experimentation across two distinct datasets (Income andTravelTime), multiple structured pruning methods (Rank, Cluster, MIQP, and Greedy pruning), and varying pruning levels, we examined how different aspects interact and shape the metrics behavior and fairness–accuracy trade-offs.

Our analysis confirms that pruning does not operate in isolation. Instead, its effects on fairness and accuracy are intricately tied to the characteristics of the dataset, the nature of the pruning method, and the degree of pruning applied. While some pruning methods, such as Rank Pruning, were shown to be highly sensitive and capable of causing drastic shifts in fairness metrics, others, like Greedy and Cluster Pruning, demonstrated more stable behavior, particularly for less volatile datasets. Importantly, our findings highlight that pruning often amplifies or deamplifies existing biases in the dataset, rather than altering their direction.

We further observed that the trade-off between accuracy and fairness metrics is neither universal nor predictable. Instead, it varies across methods, datasets, and pruning levels, with the dataset itself playing a significant role in shaping these outcomes. In some scenarios, predictive parity difference improved with pruning, while demographic parity difference and equalized odds difference worsened, illustrating the complex interplay between metrics and underlying biases. This complexity underscores the necessity of tailored approaches when applying pruning methods in fairness-sensitive tasks.

Throughout this study, we provided detailed visualizations — including correlation heatmaps, scatter plots, spider charts, and parallel coordinate plots — to aid in understanding the interactions between metrics, methods,pruning intensities and datasets. These visual tools, coupled with statistical analysis and correlation studies, allowed us to identify patterns and stability trends that can inform best practices.

While this work focused on two datasets, four widely used structured ensemble pruning methods, and three core fairness metrics, it provides foundational insights that researchers and practitioners can use to guide pruning decisions in fairness-aware machine learning. Our findings highlight that the choice of pruning method, understanding of dataset bias, and careful calibration of pruning levels are all critical factors that cannot be generalized and must instead be assessed on a case-by-case basis.

This study opens several avenues for future research. Expanding this analysis to a broader range of datasets and sensitive attributes could help verify the generalizability of the findings. Additionally, incorporating computational complexity or inference time into the analysis, alongside conducting Pareto Analysis, may offer more practical and actionable guidance for real-world applications.

REFERENCES

[1] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 8, no. 4, p. e1249, 2018.

[2] N. Rane, S. P. Choudhary, and J. Rane, “Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions,” *Studies in Medical and Health Sciences*, vol. 1, no. 2, pp. 18–41, 2024.

[3] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, and J. Adejoh, “Enhancing credit card fraud detection: an ensemble machine learning approach,” *Big Data and Cognitive Computing*, vol. 8, no. 1, p. 6, 2024.

[4] P. Pintelas and I. E. Livieris, “Special issue on ensemble learning and applications,” p. 140, 2020.

[5] B. Evans, “Population-based ensemble learning with tree structures for classification,” Ph.D. dissertation, Open Access Te Herenga Waka-Victoria University of Wellington, 2019.

[6] J. S. Lim, “Ensemble learning of high dimension datasets,” Ph.D. dissertation, The University of Waikato, 2020.

[7] L. Capitaine, R. Genuer, and R. Thiébaut, “Random forests for high-dimensional longitudinal data,” *Statistical methods in medical research*, vol. 30, no. 1, pp. 166–184, 2021.

[8] I. D. Mienye and Y. Sun, “A survey of ensemble learning: Concepts, algorithms, applications, and prospects,” *Ieee Access*, vol. 10, pp. 99 129–99 149, 2022.

[9] M. Shah, K. Gandhi, K. A. Patel, H. Kantawala, R. Patel, and A. Kothari, “Theoretical evaluation of ensemble machine learning techniques,” in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, 2023, pp. 829–837.

[10] J. Kedziora, “Prediction instability in machine learning ensembles,” *arXiv preprint arXiv:2407.03194*, 2024.

[11] B. Seijo Pardo, “Information fusion and ensembles in machine learning,” 2019.

[12] B. Dhanwanth, R. A. Roshan, C. Bhargavi, G. V. Shri, and S. Raja, “Ensemble machine learning for better crime detection and prevention,” in *2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE, 2023, pp. 701–707.

[13] H. Liu, J. Jia, and N. Z. Gong, “On the intrinsic differential privacy of bagging,” *arXiv preprint arXiv:2008.09845*, 2020.

[14] C. Zhao, R. Peng, and D. Wu, “Bagging and boosting fine-tuning for ensemble learning,” *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 4, pp. 1728–1742, 2023.

[15] J. A. Soloff, R. F. Barber, and R. Willett, “Bagging provides assumption-free stability,” *Journal of Machine Learning Research*, vol. 25, no. 131, pp. 1–35, 2024.

[16] J. Deshmukh, M. Jangid, S. Gupte, S. Ghosh, and S. Ingle, “Ensemble method combination: bagging and boosting,” in *Advanced Computing Technologies and Applications: Proceedings of 2nd International Conference on Advanced Computing Technologies and Applications—ICACTA 2020*. Springer, 2020, pp. 399–409.

[17] B. Liu and R. Mazumder, “Randomization can reduce both bias and variance: A case study in random forests,” *arXiv preprint arXiv:2402.12668*, 2024.

[18] Z. Li, X. Du, T. Wu, and Y. Cao, “Explaining random forests as single decision trees through distance functional optimization,” in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[19] C. Lewis Schmalohr, J. Grossbach, M. Clément-Ziza, and A. Beyer, “Detection of epistatic interactions with random forest,” *BioRxiv*, p. 353193, 2018.

[20] B. Partopour, R. C. Paffenroth, and A. G. Dixon, “Random forests for mapping and analysis of microkinetics models,” *Computers & Chemical Engineering*, vol. 115, pp. 286–294, 2018.

[21] C.-M. Chi, P. Vossler, Y. Fan, and J. Lv, “Asymptotic properties of high-dimensional random forests,” *The Annals of Statistics*, vol. 50, no. 6, pp. 3415–3438, 2022.

[22] F. Daghero, A. Burrello, C. Xie, L. Benini, A. Calimera, E. Macii, M. Poncino, and D. J. Pagliari, “Adaptive random forests for energy-efficient inference on microcontrollers,” in *2021 IFIP/IEEE 29th International Conference on Very Large Scale Integration (VLSI-SoC)*. IEEE, 2021, pp. 1–6.

[23] A. Ziegler and I. R. König, “Mining data with random forests: current options for real-world applications,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 1, pp. 55–63, 2014.

[24] S. Suthaharan and S. Suthaharan, “Random forest learning,” *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, pp. 273–288, 2016.

[25] Y. A. Saadoon and R. H. Abdulamir, “Improved random forest algorithm performance for big data,” in *Journal of Physics: Conference Series*, vol. 1897, no. 1. IOP Publishing, 2021, p. 012071.

[26] T. Zhu, “Analysis on the applicability of the random forest,” in *Journal of Physics: Conference Series*, vol. 1607, no. 1. IOP Publishing, 2020, p. 012123.

[27] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, “Random forests for land cover classification,” *Pattern recognition letters*, vol. 27, no. 4, pp. 294–300, 2006.

[28] M. Xiao, C. Yan, B. Fu, S. Yang, S. Zhu, D. Yang, B. Lei, R. Huang, and J. Lei, “Risk prediction for postpartum depression based on random forest.” *Zhong nan da xue xue bao. Yi xue ban= Journal of Central South University. Medical Sciences*, vol. 45, no. 10, pp. 1215–1222, 2020.

[29] C. Bénard, “Random forests: A sensitivity analysis perspective,” in *Proceedings of the MascotNum Annual Conference, Aussois, France*, 2021, pp. 28–30.

[30] A. O. Kuyoro, O. A. Ogunyolu, T. G. Ayanwola, and F. Y. Ayankoya, “Dynamic effectiveness of random forest algorithm in financial credit risk management for improving output accuracy and loan classification prediction,” *Ingénierie des systèmes d’information*, vol. 27, no. 5, p. 815, 2022.

[31] J. S. Rhodes, A. Cutler, and K. R. Moon, “Geometry-and accuracy-preserving random forest proximities,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 947–10 959, 2023.

[32] S. Bernard, L. Heutte, and S. Adam, “On the selection of decision trees in random forests,” in *2009 International joint conference on neural networks.* IEEE, 2009, pp. 302–307.

[33] A. Dorador, “Improving the accuracy and interpretability of random forests via forest pruning,” *arXiv e-prints*, pp. arXiv–2401, 2024.

[34] N. Surjanovic, A. Henrey, and T. M. Loughin, "Alpha-trimming: Locally adaptive tree pruning for random forests," *arXiv preprint arXiv:2408.07151*, 2024.

[35] V. Y. Kulkarni and P. K. Sinha, "Pruning of random forest classifiers: A survey and future directions," in *2012 International Conference on Data Science & Engineering (ICDSE).* IEEE, 2012, pp. 64–68.

[36] M. Rose, H. R. Hassen *et al.*, "Asurvey of random forest pruning techniques," in *CS & IT Conference Proceedings*, vol. 9, no. 18. CS & IT Conference Proceedings, 2019.

[37] K. Fawagreh, M. M. Gaber, and E. Elyan, "Club-drf: A clustering approach to extreme pruning of random forests," in *Research and Development in Intelligent Systems XXXII: Incorporating Applications and Innovations in Intelligent Systems XXIII 32.* Springer, 2015, pp. 59–73.

[38] H. Zhang and L. Cao, "A spectral clustering based ensemble pruning approach," *Neurocomputing*, vol. 139, pp. 289–297, 2014.

[39] H. Guo, H. Liu, R. Li, C. Wu, Y. Guo, and M. Xu, "Margin & diversity based ordering ensemble pruning," *Neurocomputing*, vol. 275, pp. 237–246, 2018.

[40] K. Fawagreh, M. M. Gaber, and E. Elyan, "An outlier detection-based tree selection approach to extreme pruning of random forests," *arXiv preprint arXiv:1503.05187*, 2015.

[41] B. Bakker and T. Heskes, "Clustering ensembles of neural network models," *Neural networks*, vol. 16, no. 2, pp. 261–269, 2003.

[42] G. Tsoumakas, I. Partalas, and I. Vlahavas, "An ensemble pruning primer," *Applications of supervised and unsupervised ensemble methods*, pp. 1–13, 2009.

[43] A. Lazarevic and Z. Obradovic, "Effective pruning of neural network classifier ensembles," in *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, vol. 2. IEEE, 2001, pp. 796–801.

[44] E. E. Tripoliti, D. I. Fotiadis, and G. Manis, "Modifications of the construction and voting mechanisms of the random forests algorithm," *Data & Knowledge Engineering*, vol. 87, pp. 41–65, 2013.

[45] M. Robnik-Šikonja, "Improving random forests," in *European conference on machine learning.* Springer, 2004, pp. 359–370.

[46] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, “Dynamic integration with random forests,” in *Machine Learning: ECML 2006: 17th European Conference on Machine Learning Berlin, Germany, September 18-22, 2006 Proceedings 17*. Springer, 2006, pp. 801–808.

[47] I. Tarchoune, A. Djebbar, and H. Merouani, “Improving random forest with pre-pruning technique for binary classification,” 2023, all Sciences Abstracts, Accessed: Mar. 26, 2025.

[48] B. Nieth, T. Altstidl, L. Schwinn, and B. Eskofier, “Large-scale dataset pruning in adversarial training through data importance extrapolation,” *arXiv preprint arXiv:2406.13283*, 2024.

[49] E. Iofinova, A. Peste, and D. Alistarh, “Bias in pruned vision models: In-depth analysis and countermeasures,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 364–24 373.

[50] K. Mavrogiorgos, A. Kiourtis, A. Mavrogiorgou, A. Menychtas, and D. Kyriazis, “Bias in machine learning: A literature review,” *Applied Sciences*, vol. 14, no. 19, p. 8860, 2024.

[51] I. Sharma and B. Rathodiya, “Bias in machine learning algorithms,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 2019.

[52] S. Richardson, “Exposing the many biases in machine learning,” *Business Information Review*, vol. 39, no. 3, pp. 82–89, 2022.

[53] W. Blanzeisky and P. Cunningham, “Algorithmic factors influencing bias in machine learning,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2021, pp. 559–574.

[54] L. Bald, “Identifying and mitigating bias in machine learning applications,” 2019.

[55] L. Bothmann, K. Peters, and B. Bischl, “What is fairness? implications for fairml,” *arXiv preprint*, vol. arXiv:2205.09622, 2022.

[56] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, “Differential fairness,” in *NeurIPS 2019 Workshop on Machine Learning with Guarantees*, 2019.

[57] M. Wan, D. Zha, N. Liu, and N. Zou, “In-processing modeling techniques for machine learning fairness: A survey,” *ACM Transactions on Knowledge Discovery from Data*, vol. 17, no. 3, pp. 1–27, 2023.

[58] U. Gohar and L. Cheng, “A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges,” *arXiv preprint arXiv:2305.06969*, 2023.

[59] T. P. Pagano, R. B. Loureiro, F. V. Lisboa, R. M. Peixoto, G. A. Guimarães, G. O. Cruz, M. M. Araujo, L. L. Santos, M. A. Cruz, E. L. Oliveira *et al.*, “Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods,” *Big data and cognitive computing*, vol. 7, no. 1, p. 15, 2023.

[60] A. Shome, L. Cruz, and A. Van Deursen, “Data vs. model machine learning fairness testing: An empirical study,” in *Proceedings of the 5th IEEE/ACM International Workshop on Deep Learning for Testing and Testing for Deep Learning*, 2024, pp. 1–8.

[61] N. Zhou, Z. Zhang, V. N. Nair, H. Singhal, and J. Chen, “Bias, fairness and accountability with artificial intelligence and machine learning algorithms,” *International Statistical Review*, vol. 90, no. 3, pp. 468–480, 2022.

[62] S. Rathore, *Model Agnostic Feature Selection for Fairness.* University of Rhode Island, 2022.

[63] Z. Zhu, Y. Yao, J. Sun, Y. Liu, and H. Li, “Evaluating fairness without sensitive attributes: A framework using only auxiliary models,” 2022.

[64] R. Luo, T. Tang, F. Xia, J. Liu, C. Xu, L. Y. Zhang, W. Xiang, and C. Zhang, “Algorithmic fairness: A tolerance perspective,” *arXiv preprint arXiv:2405.09543*, 2024.

[65] S. Wachter, B. Mittelstadt, and C. Russell, “Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai,” *Computer Law & Security Review*, vol. 41, p. 105567, 2021.

[66] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi, “Putting fairness principles into practice: Challenges, metrics, and improvements,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 453–459.

[67] K. T. Rodolfa, H. Lamba, and R. Ghani, “Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy,” *Nature Machine Intelligence*, vol. 3, no. 10, pp. 896–904, 2021.

[68] S. Bird, K. Kenthapadi, E. Kiciman, and M. Mitchell, “Fairness-aware machine learning: Practical challenges and lessons learned,” in *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 834–835.

[69] L. Oneto and S. Chiappa, "Fairness in machine learning," in *Recent trends in learning from data: Tutorials from the inns big data and deep learning conference (innsbddl2019)*. Springer, 2020, pp. 155–196.

[70] J. S. Franklin, K. Bhanot, M. Ghalwash, K. P. Bennett, J. McCusker, and D. L. McGuinness, "An ontology for fairness metrics," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 265–275.

[71] M. S. A. Lee, L. Floridi, and J. Singh, "From fairness metrics to key ethics indicators (keis): a context-aware approach to algorithmic ethics in an unequal society," *Available at SSRN*, 2020.

[72] T. Kirat, O. Tambou, V. Do, and A. Tsoukiàs, "Fairness and explainability in automatic decision-making systems. a challenge for computer science and law," *EURO journal on decision processes*, vol. 11, p. 100036, 2023.

[73] Y. Dai, G. Li, F. Luo, X. Ma, and Y. Wu, "Coupling fairness and pruning in a single run: a bi-level optimization perspective," *arXiv preprint arXiv:2312.10181*, 2023.

[74] X. Lin, S. Kim, and J. Joo, "Fairgrape: Fairness-aware gradient pruning method for face attribute classification," in *European Conference on Computer Vision*. Springer, 2022, pp. 414–432.

[75] Y. Wu, D. Zeng, X. Xu, Y. Shi, and J. Hu, "Fairprune: Achieving fairness through pruning for dermatological disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 743–753.

[76] R. Meyer and A. Wong, "A fair loss function for network pruning," *arXiv preprint arXiv:2211.10285*, 2022.

[77] A. Zayed, G. Mordido, S. Shabanian, I. Baldini, and S. Chandar, "Fairness-aware structured pruning in transformers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 20, 2024, pp. 22 484–22 492.

[78] F. Ranzato, C. Urban, and M. Zanella, "Fairness-aware training of decision trees by abstract interpretation," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 1508–1517.

[79] S. Kashyap, S. Mehta *et al.*, "A comparative study of fairness and bias in decisiontree and random forest machine learning models," in *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)*. IEEE, 2024, pp. 491–495.

[80] J. Zhang, I. Beschastnikh, S. Mechtaev, and A. Roychoudhury, “Fairness-guided smt-based rectification of decision trees and random forests,” *arXiv preprint arXiv:2011.11001*, 2020.

[81] J. Fitzsimons, A. Ali, M. Osborne, and S. Roberts, “Equality constrained decision trees: For the algorithmic enforcement of group fairness,” *arXiv preprint arXiv:1810.05041*, 2018.

[82] X. Fan, H. Xu, J. Wu, C. Tong, K. Wang, J. Song, and Z. Huang, “Dynamic connected neural decision classifier and regressor with dynamic softing pruning,” in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 1014–1021.

[83] Y. Bian and K. Zhang, “Increasing fairness in compromise on accuracy via weighted vote with learning guarantees,” *CoRR*, 2023.

[84] M. Paganini, “Prune responsibly,” *arXiv preprint arXiv:2009.09936*, 2020.

[85] Y. Manzali and M. Elfar, “Random forest pruning techniques: a recent review,” in *Operations research forum*, vol. 4, no. 2. Springer, 2023, p. 43.

[86] G. Giacinto, F. Roli, and G. Fumera, “Design of effective multiple classifier systems by clustering of classifiers,” in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 2. IEEE, 2000, pp. 160–163.

[87] D. Koutra, C. Plant, M. G. Rodriguez, E. Baralis, and F. Bonchi, *Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part V*. Springer Nature, 2023, vol. 14173.

[88] G. Martınez-Munoz and A. Suárez, “Aggregation ordering in bagging,” in *Proc. of the IASTED International Conference on Artificial Intelligence and Applications*. Citeseer, 2004, pp. 258–263.

[89] Y. Zhang, S. Burer, W. Nick Street, K. P. Bennett, and E. Parrado-Hernández, “Ensemble pruning via semi-definite programming.” *Journal of machine learning research*, vol. 7, no. 7, 2006.

[90] G. D. Cavalcanti, L. S. Oliveira, T. J. Moura, and G. V. Carvalho, “Combining diversity measures for ensemble pruning,” *Pattern Recognition Letters*, vol. 74, pp. 38–45, 2016.

[91] H. Guo, H. Liu, R. Li, C. Wu, Y. Guo, and M. Xu, “Margin & diversity based ordering ensemble pruning,” *Neurocomputing*, vol. 275, pp. 237–246, 2018.

[92] Z. Lu, X. Wu, X. Zhu, and J. Bongard, “Ensemble pruning via individual contribution ordering,” in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 871–880.

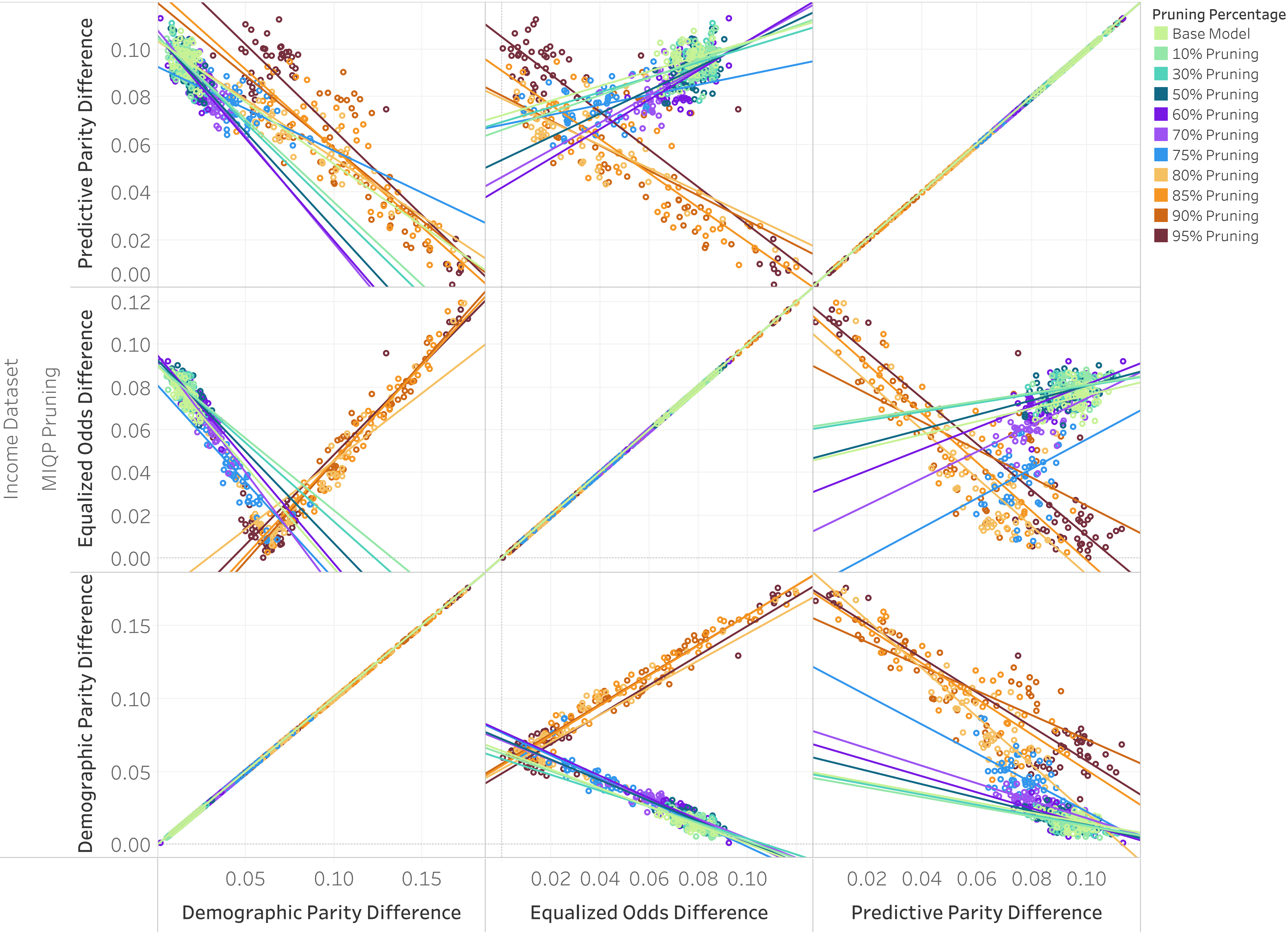
[93] J. Zhu, H. Zou, S. Rosset, T. Hastie *et al.*, “Multi-class adaboost,” *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

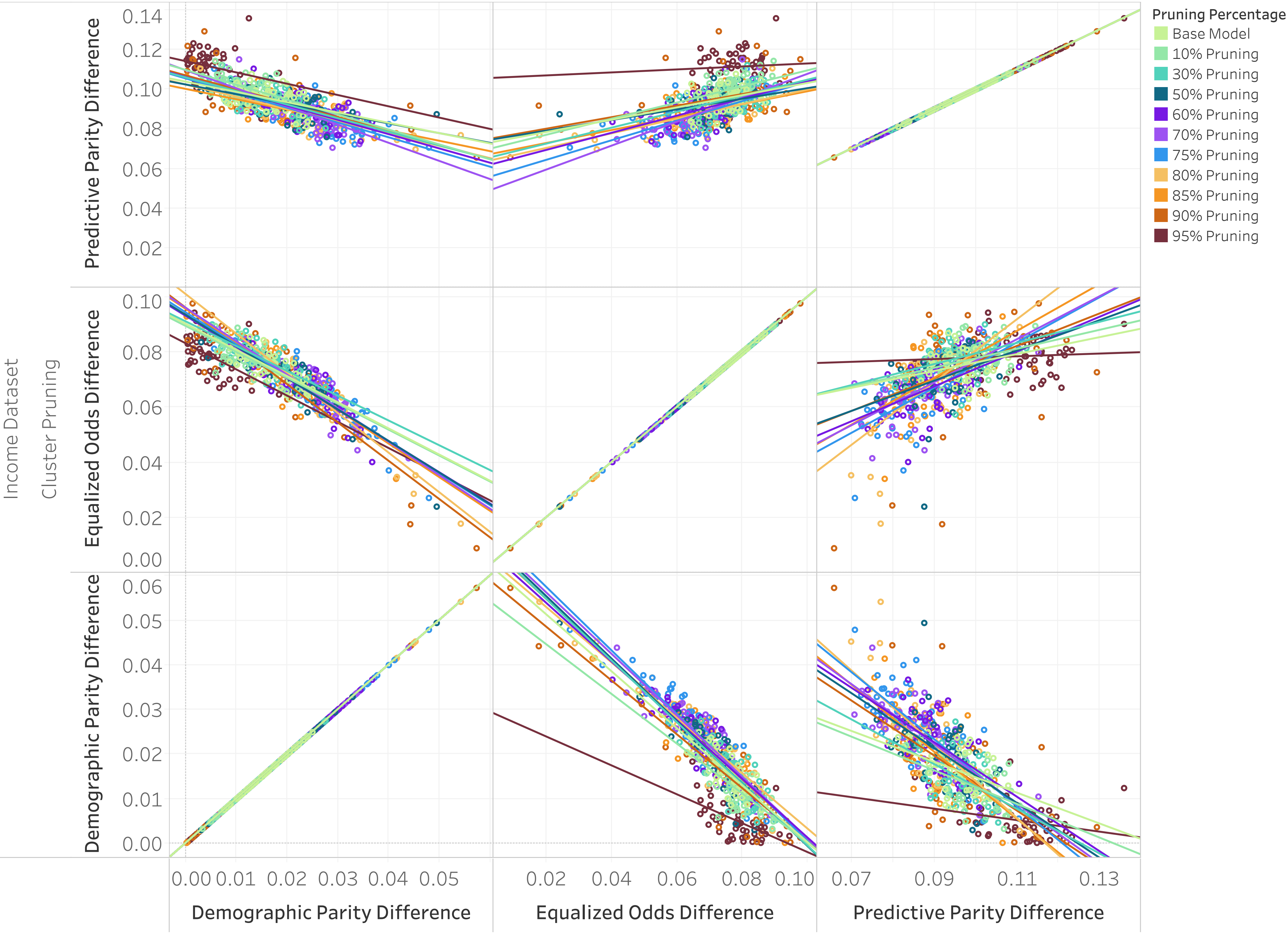
[94] D. H. Annis, “Permutation, parametric, and bootstrap tests of hypotheses,” 2005.

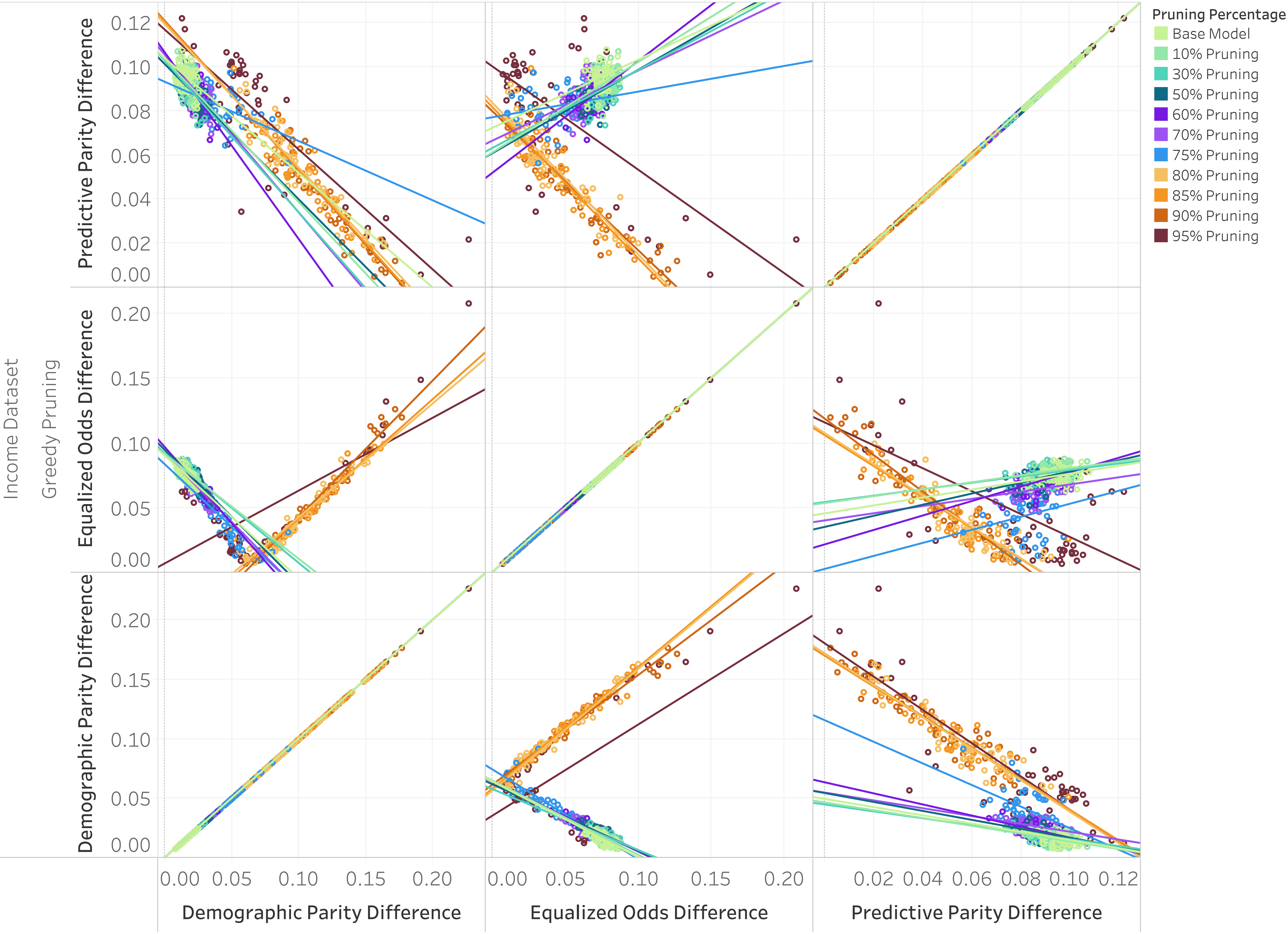
[95] X. Han, J. Chi, Y. Chen, Q. Wang, H. Zhao, N. Zou, and X. Hu, “Ffb: A fair fairness benchmark for in-processing group fairness methods,” *arXiv preprint arXiv:2306.09468*, 2023.

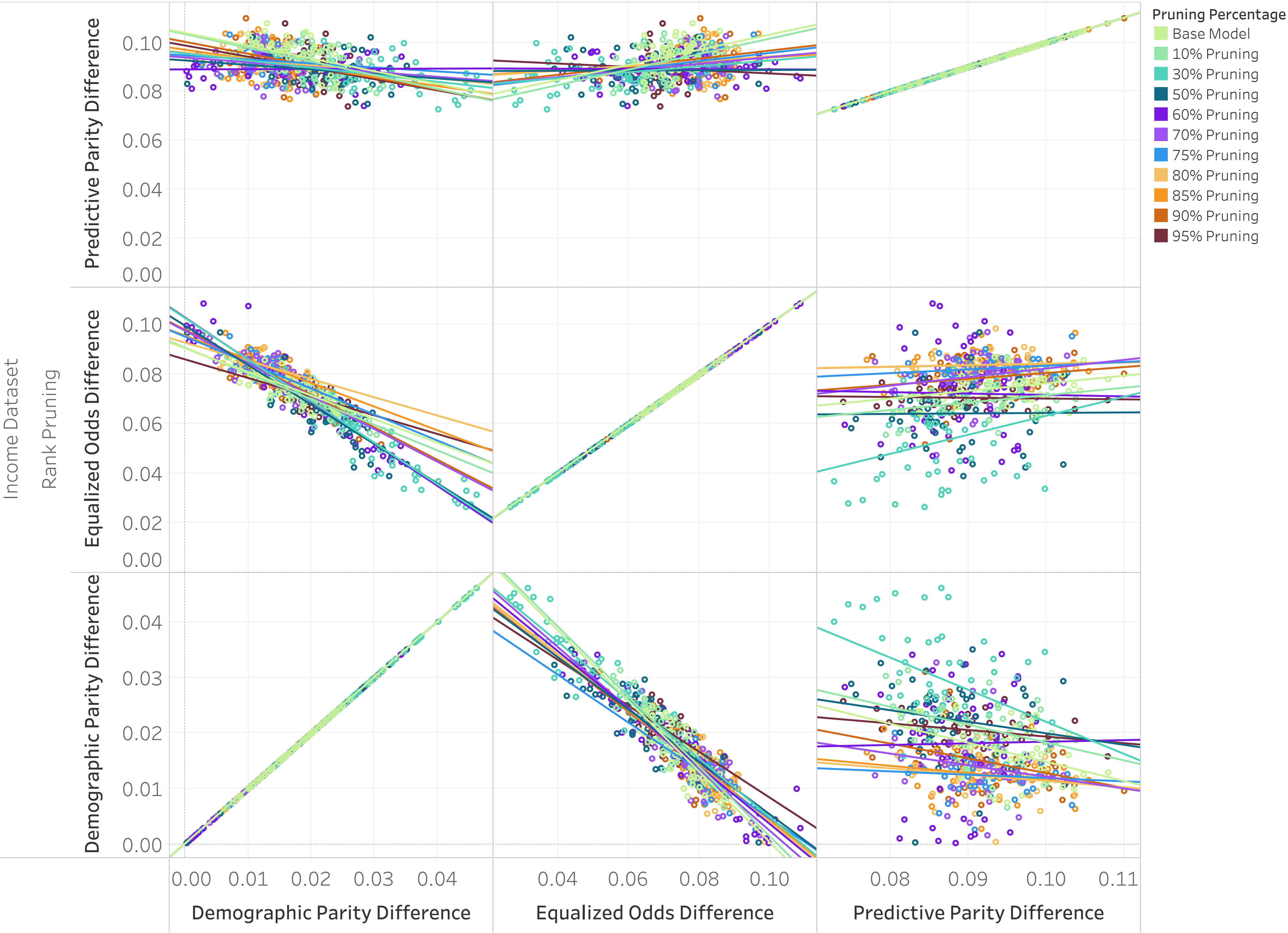
[96] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” *arXiv preprint arXiv:1609.05807*, 2016.

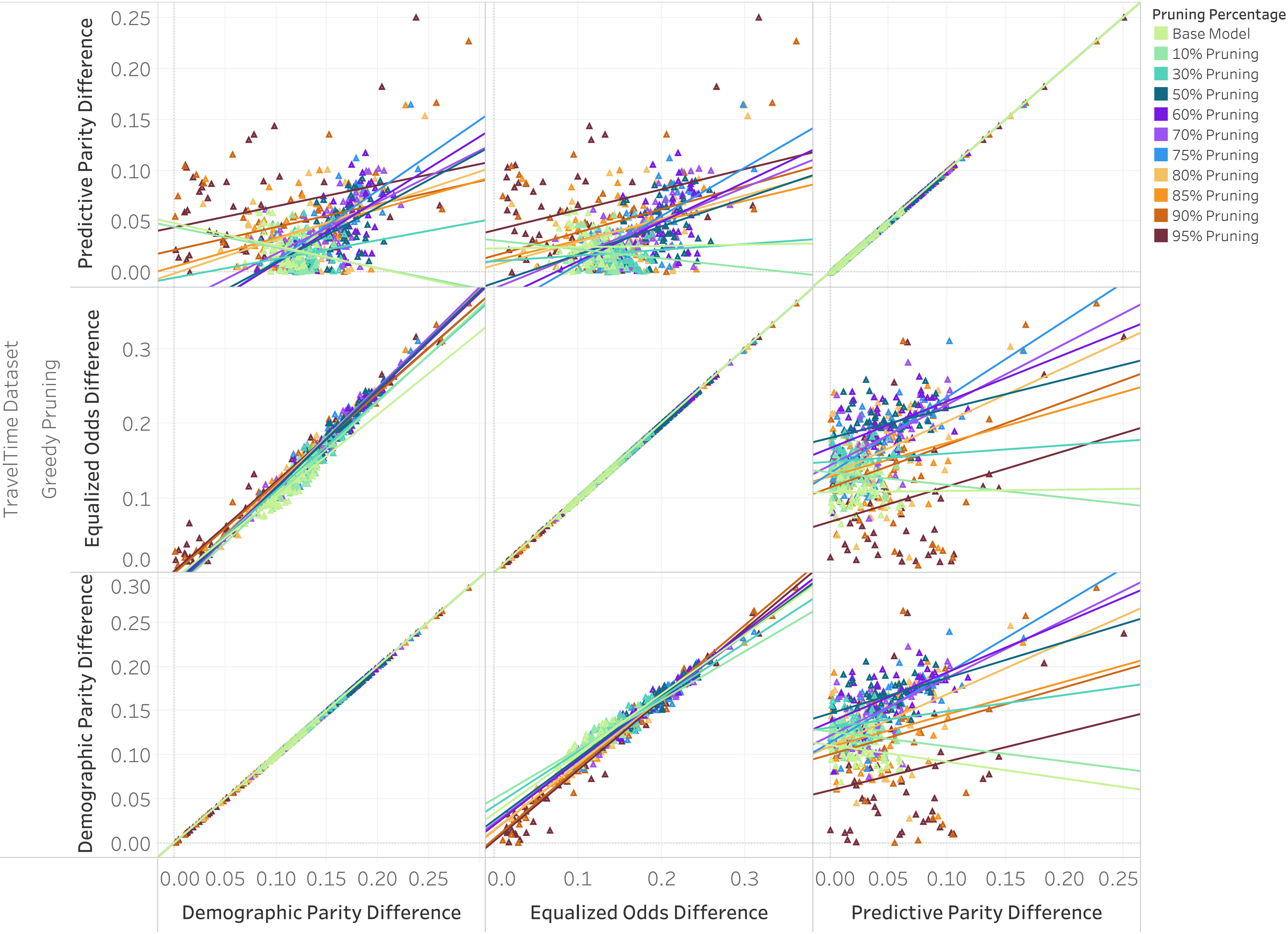
APPENDIX A SCATTER PLOTS SHOWING THE RELATIONSHIP BETWEEN EACH PAIRWISE COMBINATION OF FAIRNESS METRICS ACROSS DIFFERENT PRUNING INTENSITY LEVELS FOR EACH PRUNING METHOD ON BOTH DATASETS.

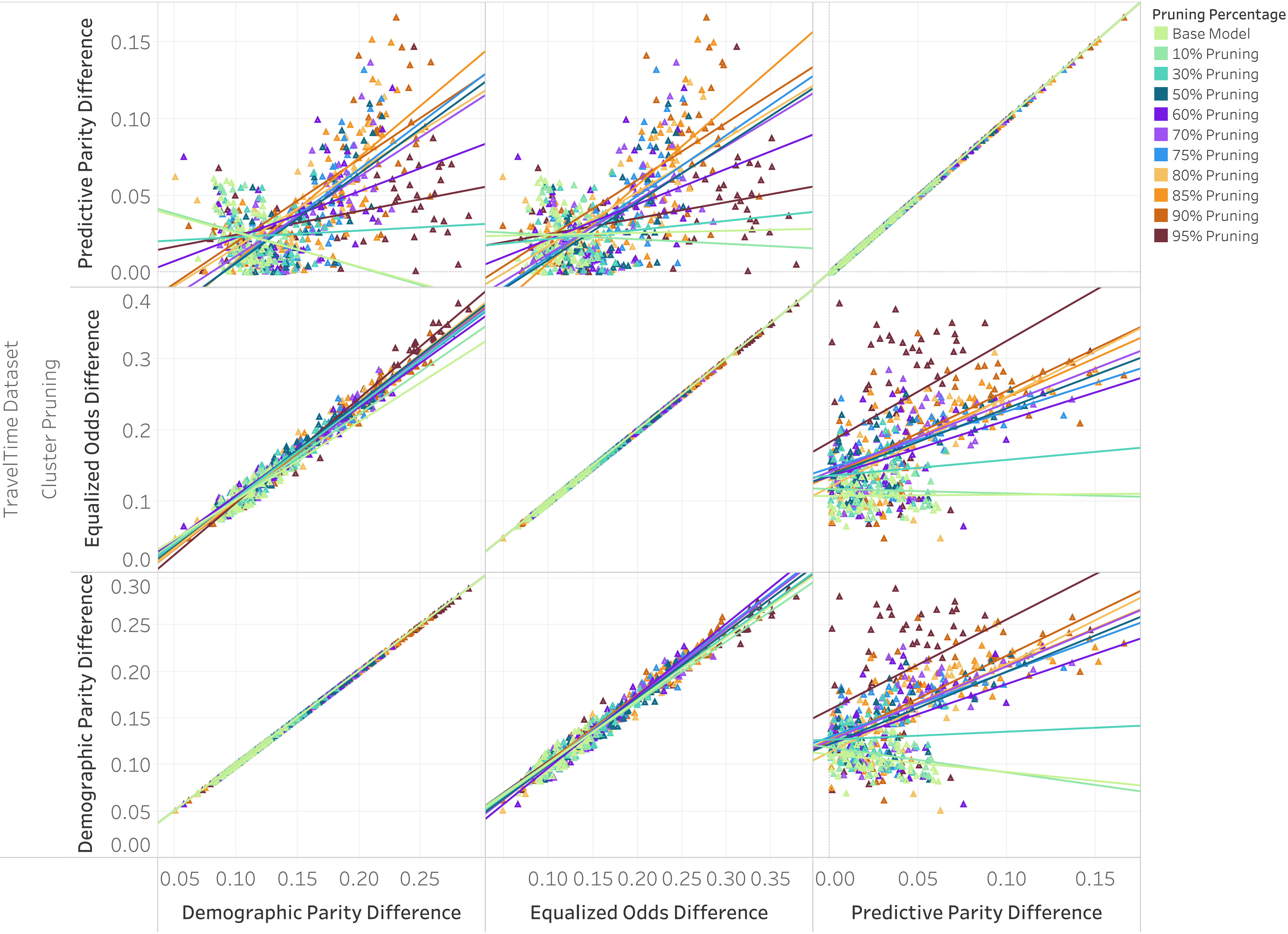

Income Dataset
MIQP Pruning
Predictive Parity Difference
Equalized Odds Difference
Demographic Parity Difference
0.10
0.08
0.06
0.04
0.02
0.00
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.15
0.10
0.05
0.00
0.05
0.10
0.15
Demographic Parity Difference
0.02
0.04
0.06
0.08
0.10
Equalized Odds Difference
0.02
0.04
0.06
0.08
0.10
Predictive Parity Difference
Pruning Percentage
Base Model
10% Pruning
30% Pruning
50% Pruning
60% Pruning
70% Pruning
75% Pruning
80% Pruning
85% Pruning
90% Pruning
95% Pruning


Income Dataset
Cluster Pruning
Predictive Parity Difference
Equalized Odds Difference
Demographic Parity Difference
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.10
0.08
0.06
0.04
0.02
0.00
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0.00 0.01 0.02 0.03 0.04 0.05
0.02 0.04 0.06 0.08 0.10
0.07 0.09 0.11 0.13
Demographic Parity Difference
Equalized Odds Difference
Predictive Parity Difference
Pruning Percentage
Base Model
10% Pruning
30% Pruning
50% Pruning
60% Pruning
70% Pruning
75% Pruning
80% Pruning
85% Pruning
90% Pruning
95% Pruning

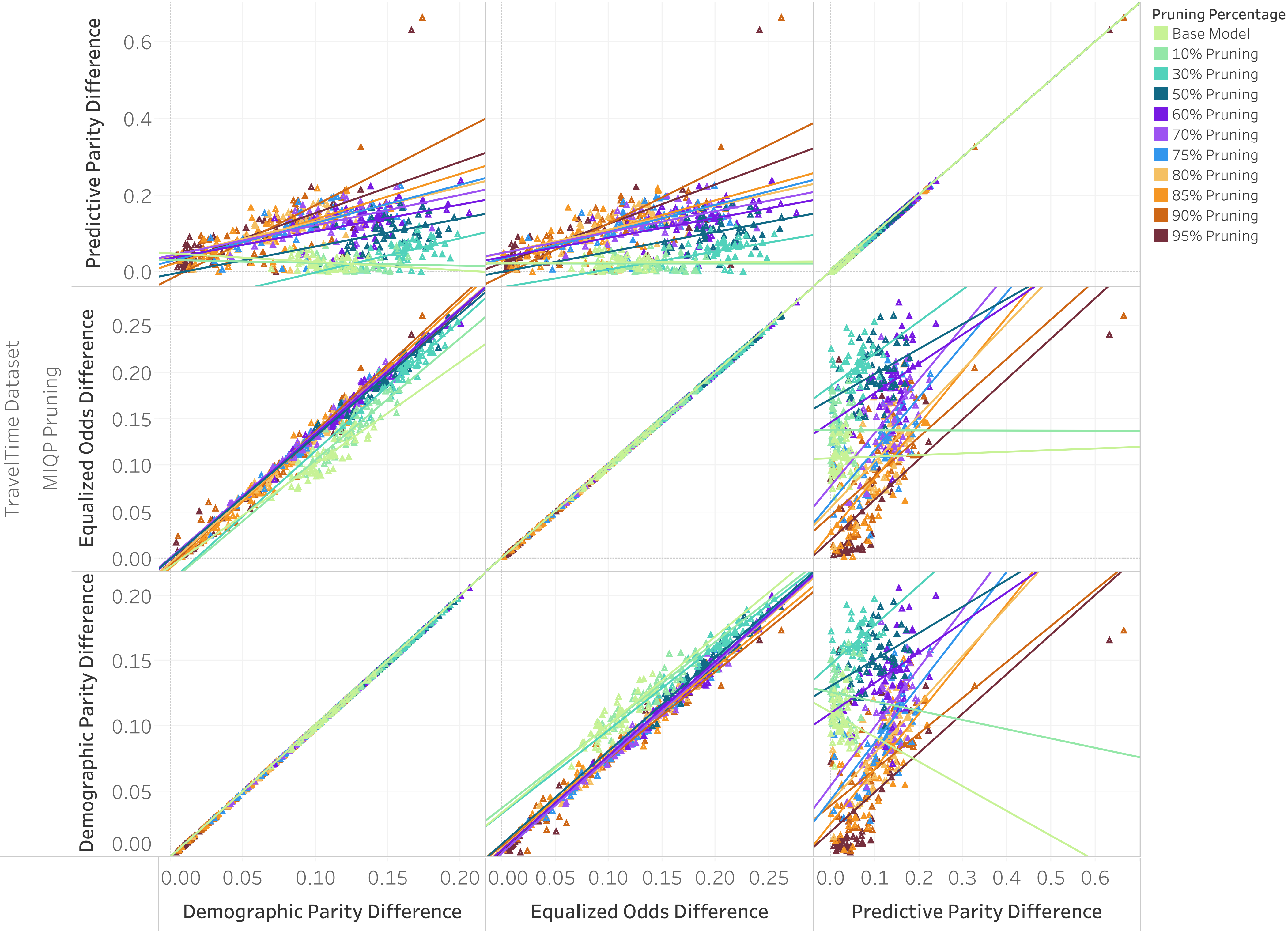

Income Dataset
Greedy Pruning
Predictive Parity Difference
Equalized Odds Difference
Demographic Parity Difference
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.20
0.15
0.10
0.05
0.00
0.20
0.15
0.10
0.05
0.00
0.00 0.05 0.10 0.15 0.20
Demographic Parity Difference
0.00 0.05 0.10 0.15 0.20
Equalized Odds Difference
0.02 0.04 0.06 0.08 0.10 0.12
Predictive Parity Difference
Pruning Percentage
Base Model
10% Pruning
30% Pruning
50% Pruning
60% Pruning
70% Pruning
75% Pruning
80% Pruning
85% Pruning
90% Pruning
95% Pruning

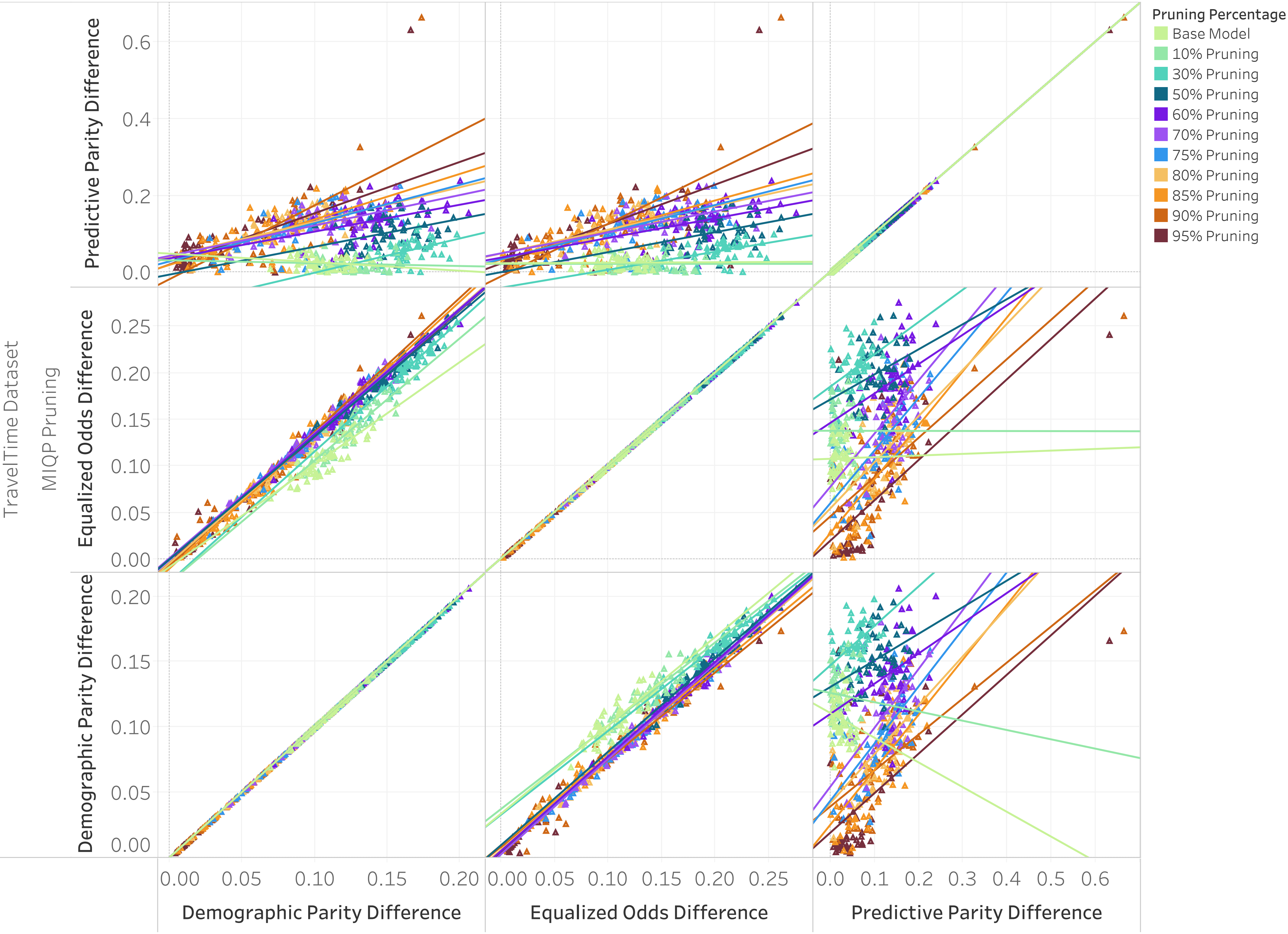

Income Dataset
Rank Pruning
Predictive Parity Difference
0.10
0.08
0.06
0.04
0.02
0.00
Equalized Odds Difference
0.10
0.08
0.06
0.04
0.02
0.00
Demographic Parity Difference
0.04
0.03
0.02
0.01
0.00
0.00 0.01 0.02 0.03 0.04
Demographic Parity Difference
0.04 0.06 0.08 0.10
Equalized Odds Difference
0.08 0.09 0.10 0.11
Predictive Parity Difference
Pruning Percentage
Base Model
10% Pruning
30% Pruning
50% Pruning
60% Pruning
70% Pruning
75% Pruning
80% Pruning
85% Pruning
90% Pruning
95% Pruning


TravelTime Dataset
Greedy Pruning
Predictive Parity Difference
Equalized Odds Difference
Demographic Parity Difference
0.25
0.20
0.15
0.10
0.05
0.00
0.3
0.2
0.1
0.0
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.00 0.05 0.10 0.15 0.20 0.25
Demographic Parity Difference
0.0 0.1 0.2 0.3
Equalized Odds Difference
0.00 0.05 0.10 0.15 0.20 0.25
Predictive Parity Difference
Pruning Percentage
Base Model
10% Pruning
30% Pruning
50% Pruning
60% Pruning
70% Pruning
75% Pruning
80% Pruning
85% Pruning
90% Pruning
95% Pruning

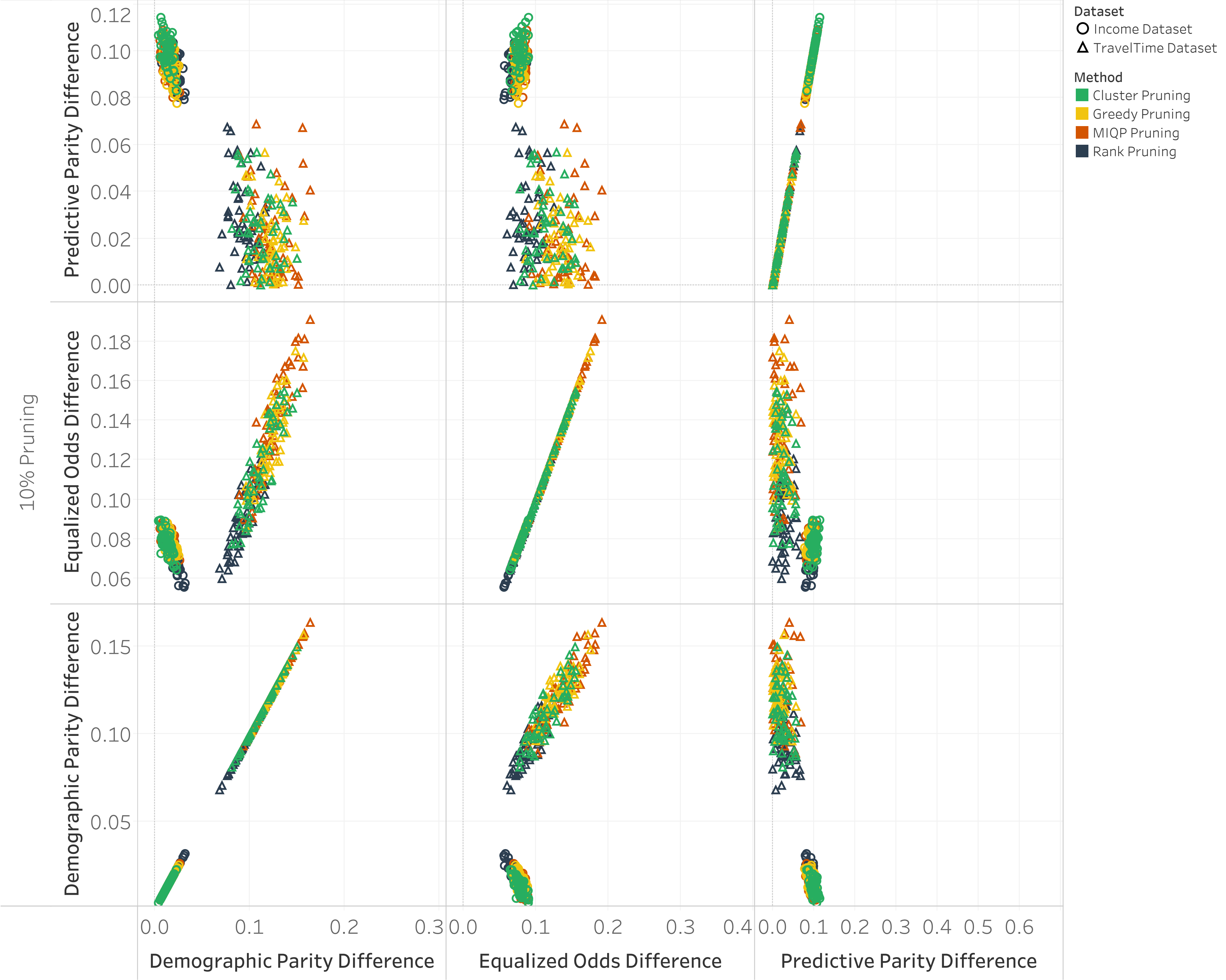


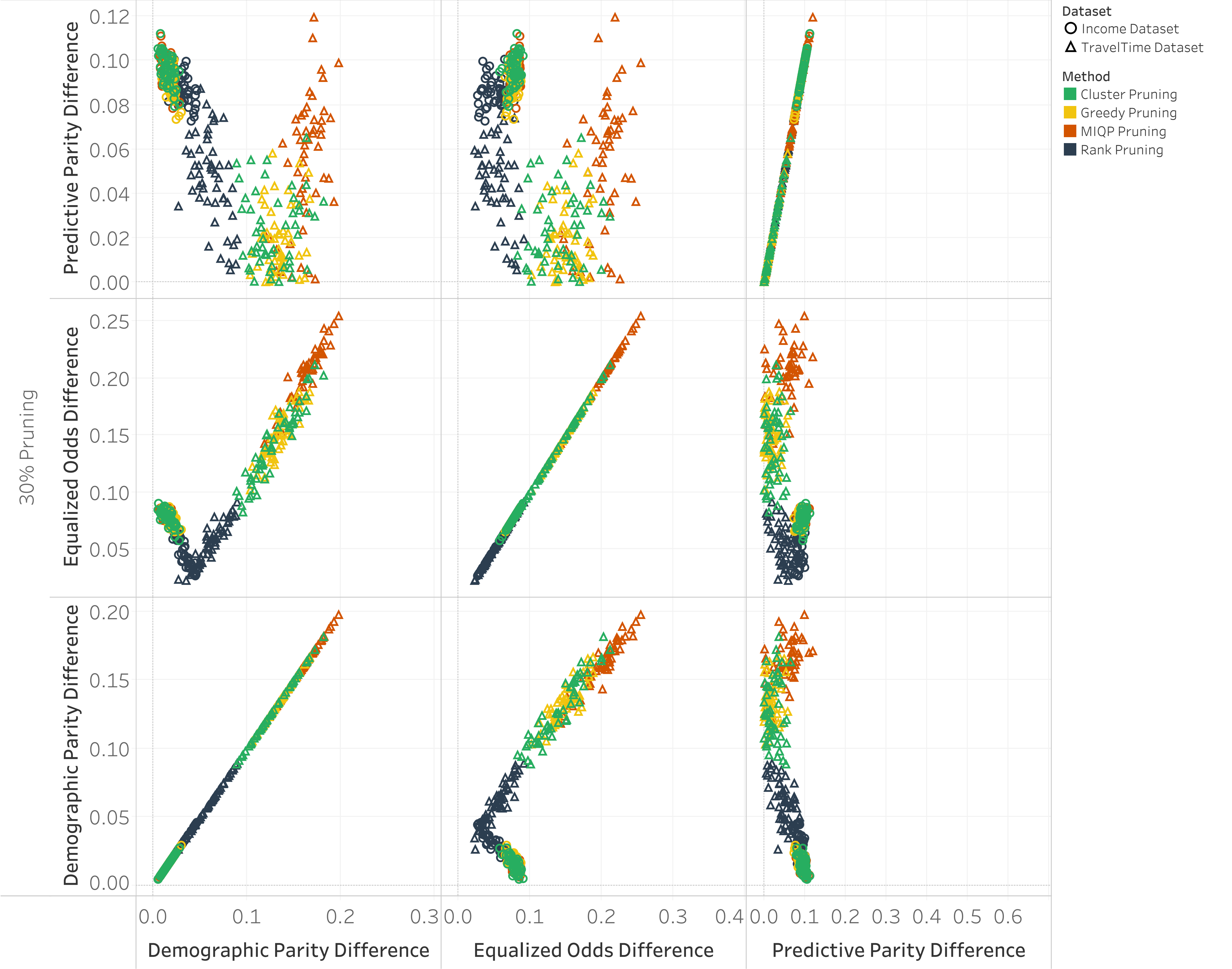
TravelTime Dataset
Cluster Pruning
Predictive Parity Difference
Equalized Odds Difference
Demographic Parity Difference
0.15
0.10
0.05
0.00
0.4
0.3
0.2
0.1
0.0
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.05 0.10 0.15 0.20 0.25
0.10 0.15 0.20 0.25 0.30 0.35
0.00 0.05 0.10 0.15
Demographic Parity Difference
Equalized Odds Difference
Predictive Parity Difference
Pruning Percentage
Base Model
10% Pruning
30% Pruning
50% Pruning
60% Pruning
70% Pruning
75% Pruning
80% Pruning
85% Pruning
90% Pruning
95% Pruning

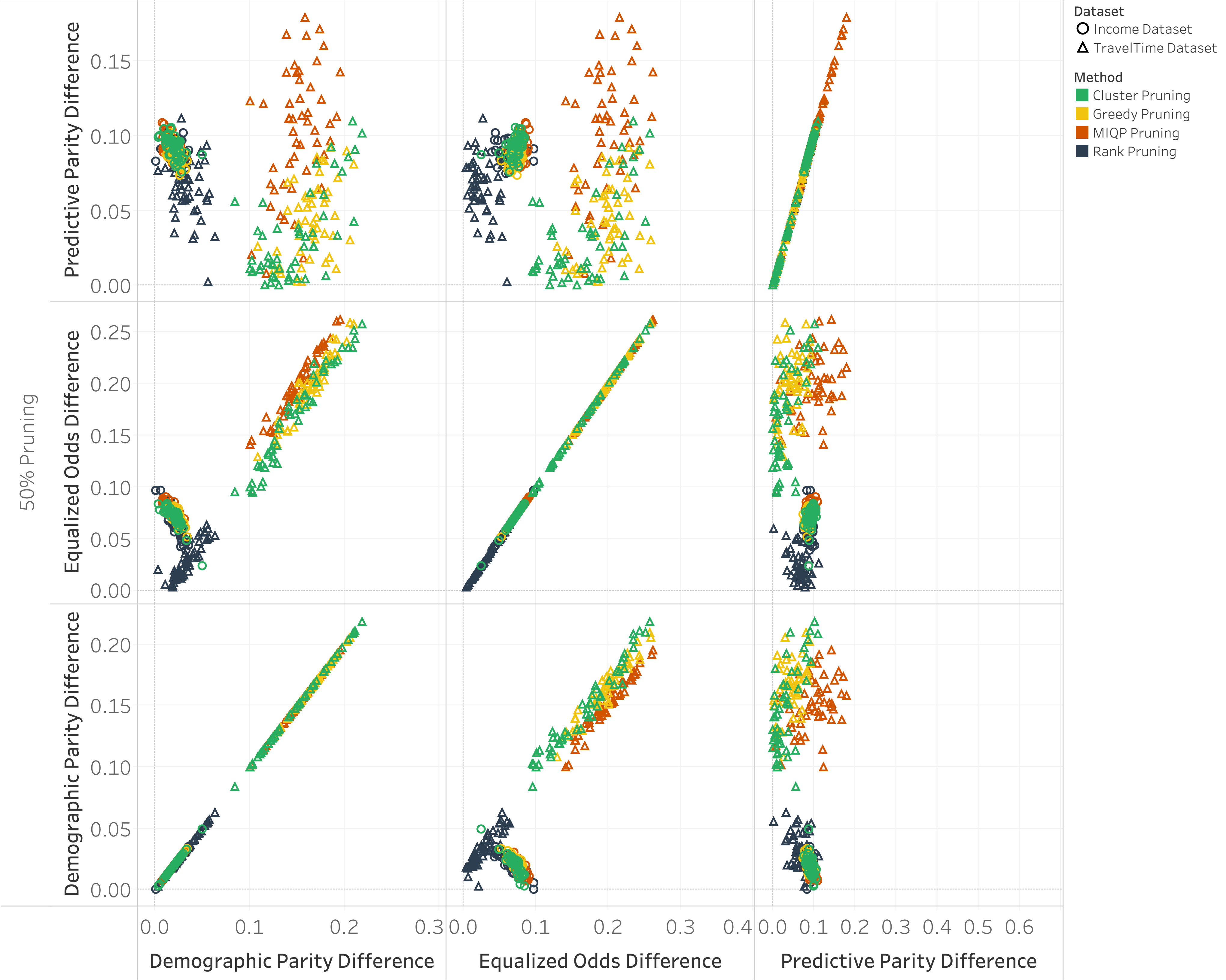

Pruning Percentage
Base Model
10% Pruning
30% Pruning
50% Pruning
60% Pruning
70% Pruning
75% Pruning
80% Pruning
85% Pruning
90% Pruning
95% Pruning
TravelTime Dataset
MIQP Pruning
Predictive Parity Difference
0.6
0.4
0.2
0.0
Equalized Odds Difference
0.25
0.20
0.15
0.10
0.05
0.00
Demographic Parity Difference
0.20
0.15
0.10
0.05
0.00
0.00
0.05
0.10
0.15
0.20
Demographic Parity Difference
0.00
0.05
0.10
0.15
0.20
0.25
Equalized Odds Difference
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Predictive Parity Difference

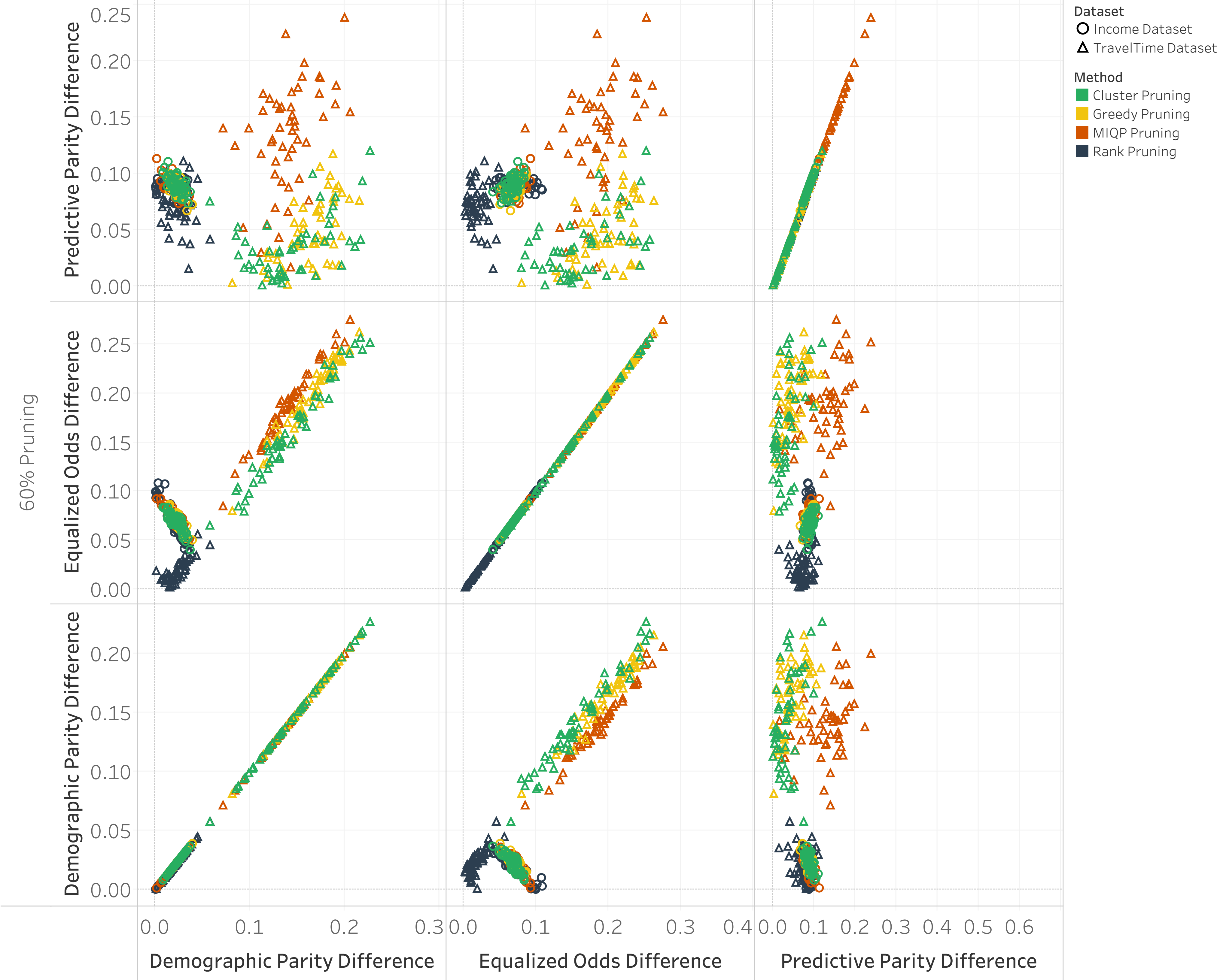

TravelTime Dataset
MIQP Pruning
Pruning Percentage
Base Model
10% Pruning
30% Pruning
50% Pruning
60% Pruning
70% Pruning
75% Pruning
80% Pruning
85% Pruning
90% Pruning
95% Pruning
Predictive Parity Difference
0.0
0.2
0.4
0.6
Equalized Odds Difference
0.00
0.05
0.10
0.15
0.20
0.25
Demographic Parity Difference
0.00
0.05
0.10
0.15
0.20
0.00
0.05
0.10
0.15
0.20
Demographic Parity Difference
0.00
0.05
0.10
0.15
0.20
0.25
Equalized Odds Difference
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Predictive Parity Difference

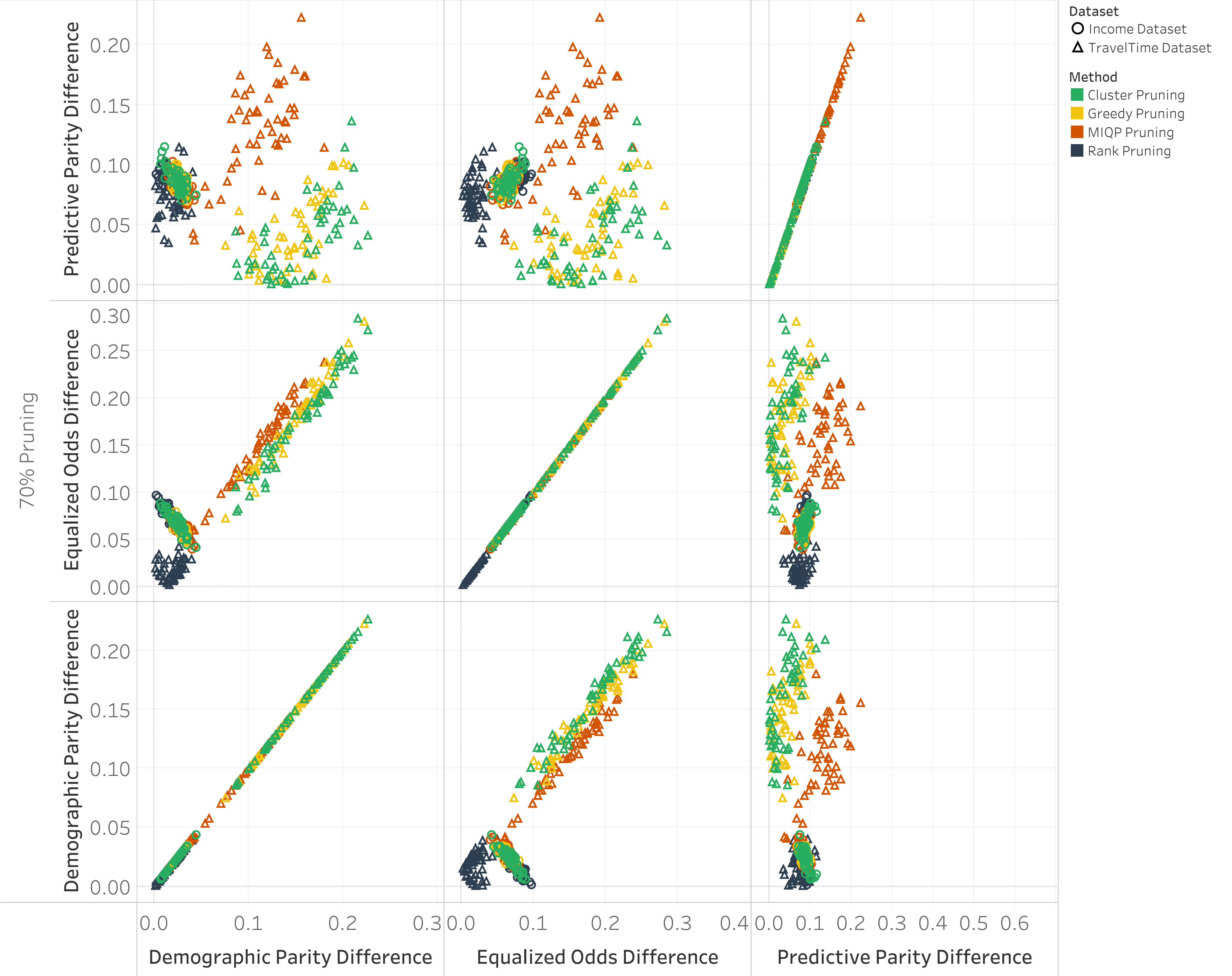
APPENDIX B SCATTER PLOTS SHOWING THE RELATIONSHIP BETWEEN EACH PAIRWISE COMBINATION OF FAIRNESS METRICS ACROSS DIFFERENT SCENARIOS.

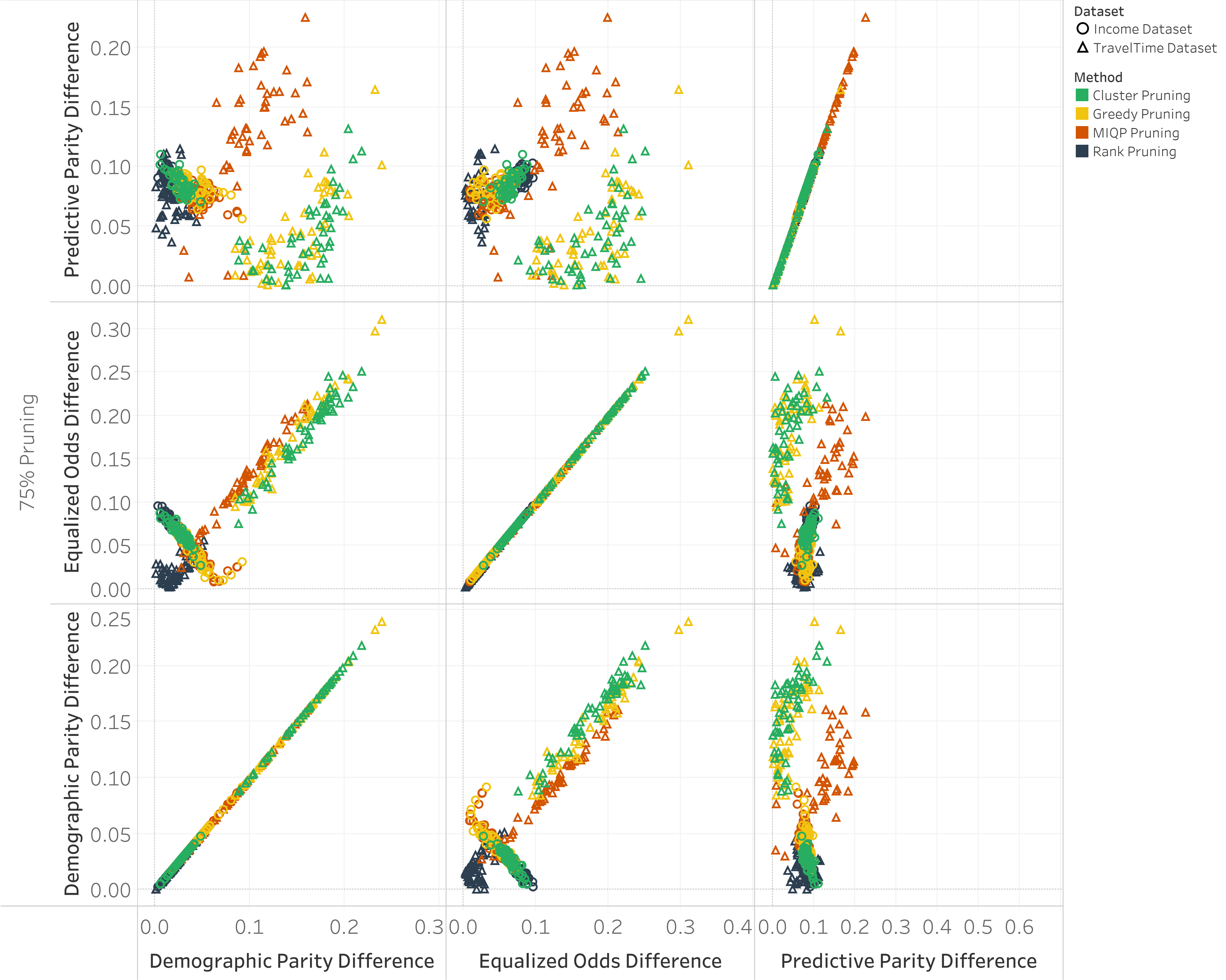

Dataset
○ Income Dataset
△ TravelTime Dataset
Method
🟩 Cluster Pruning
🟨 Greedy Pruning
🟧 MIQP Pruning
⬛ Rank Pruning
10% Pruning
Predictive Parity Difference
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Equalized Odds Difference
0.18
0.16
0.14
0.12
0.10
0.08
0.06
Demographic Parity Difference
0.15
0.10
0.05
0.0
0.1
0.2
0.3
0.0
0.1
0.2
0.3
0.4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Demographic Parity Difference
Equalized Odds Difference
Predictive Parity Difference

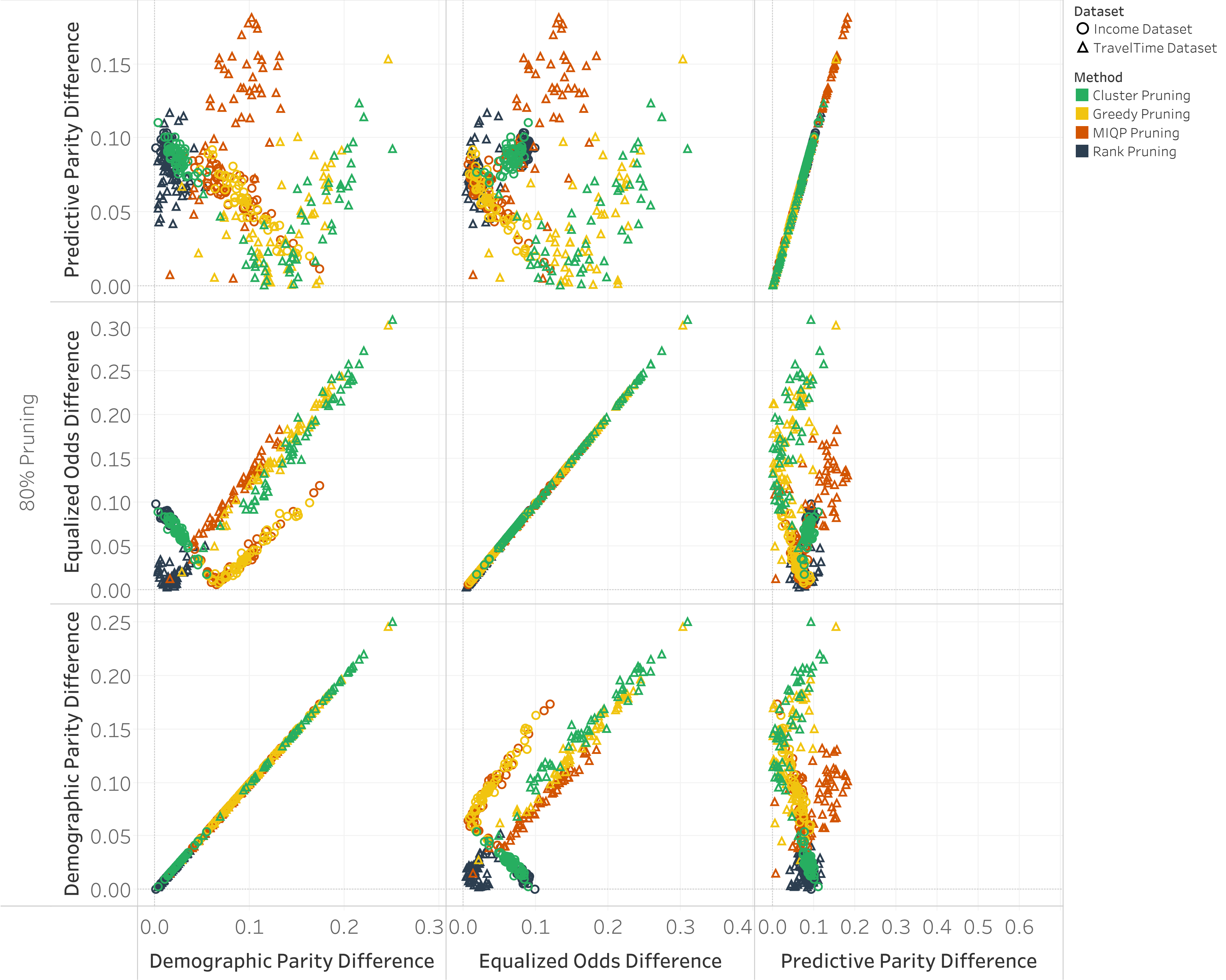

Dataset
○ Income Dataset
△ TravelTime Dataset
Method
🟩 Cluster Pruning
🟨 Greedy Pruning
🟥 MIQP Pruning
■ Rank Pruning
30% Pruning
Predictive Parity Difference
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Equalized Odds Difference
0.05
0.10
0.15
0.20
0.25
Demographic Parity Difference
0.00
0.05
0.10
0.15
0.20
0.0
0.1
0.2
0.3
Demographic Parity Difference
0.0
0.1
0.2
0.3
0.4
Equalized Odds Difference
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Predictive Parity Difference

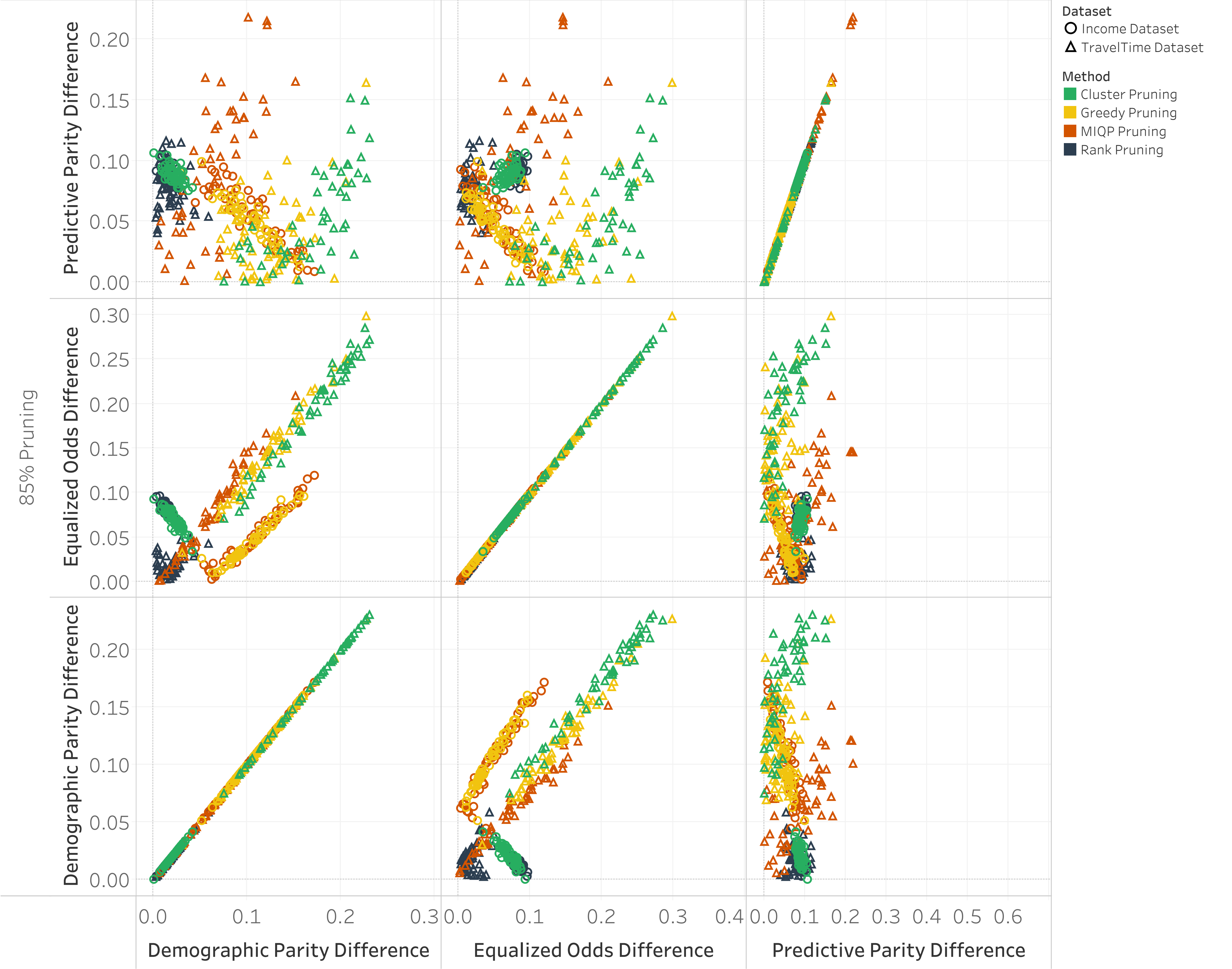

Dataset
○ Income Dataset
△ TravelTime Dataset
Method
Cluster Pruning
Greedy Pruning
MIQP Pruning
Rank Pruning
50% Pruning
Predictive Parity Difference
Equalized Odds Difference
Demographic Parity Difference
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Demographic Parity Difference
Equalized Odds Difference
Predictive Parity Difference

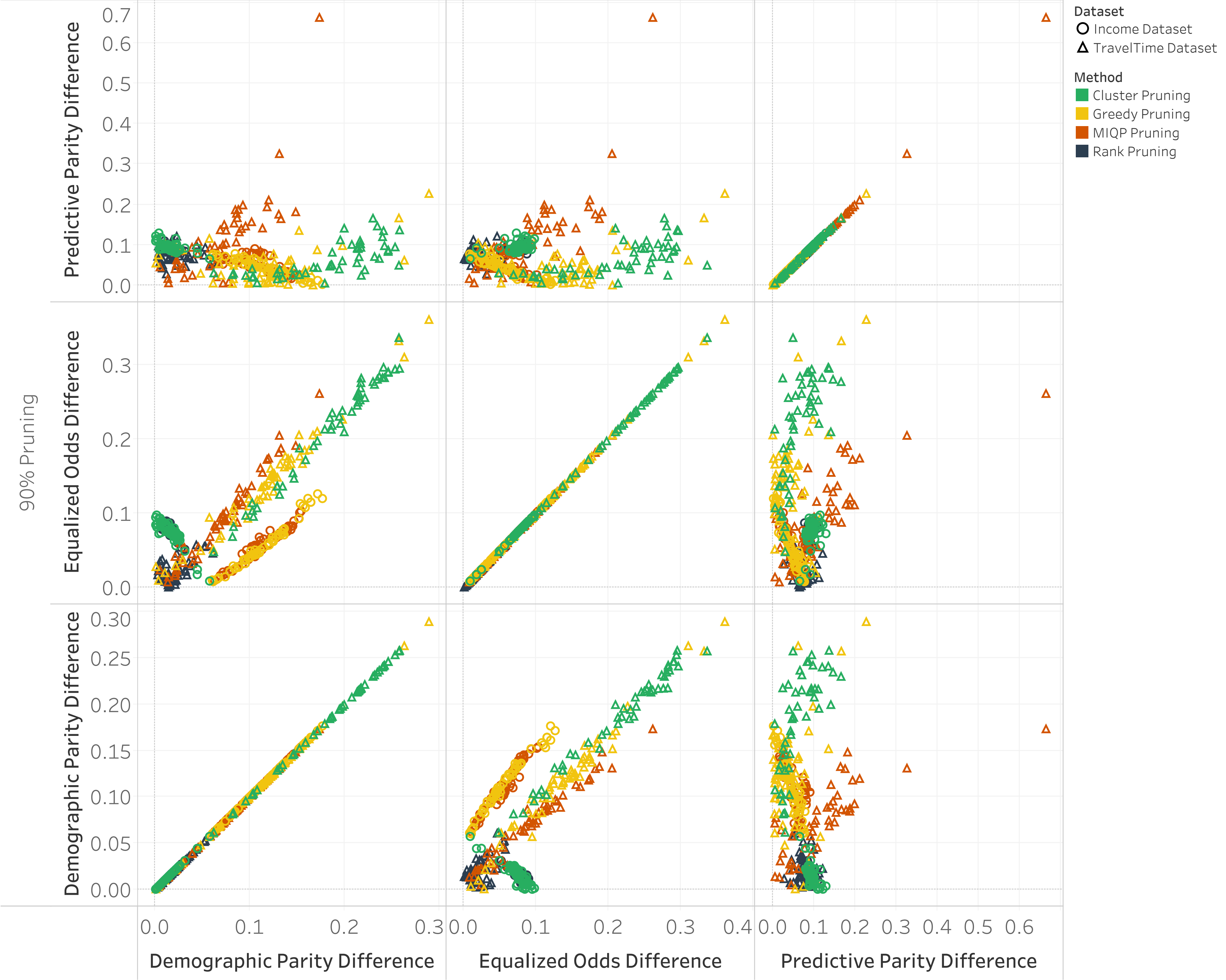

Dataset
○ Income Dataset
△ TravelTime Dataset
Method
🟩 Cluster Pruning
🟨 Greedy Pruning
🟧 MIQP Pruning
⬛ Rank Pruning
60% Pruning
Predictive Parity Difference
Equalized Odds Difference
Demographic Parity Difference
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Demographic Parity Difference
Equalized Odds Difference
Predictive Parity Difference

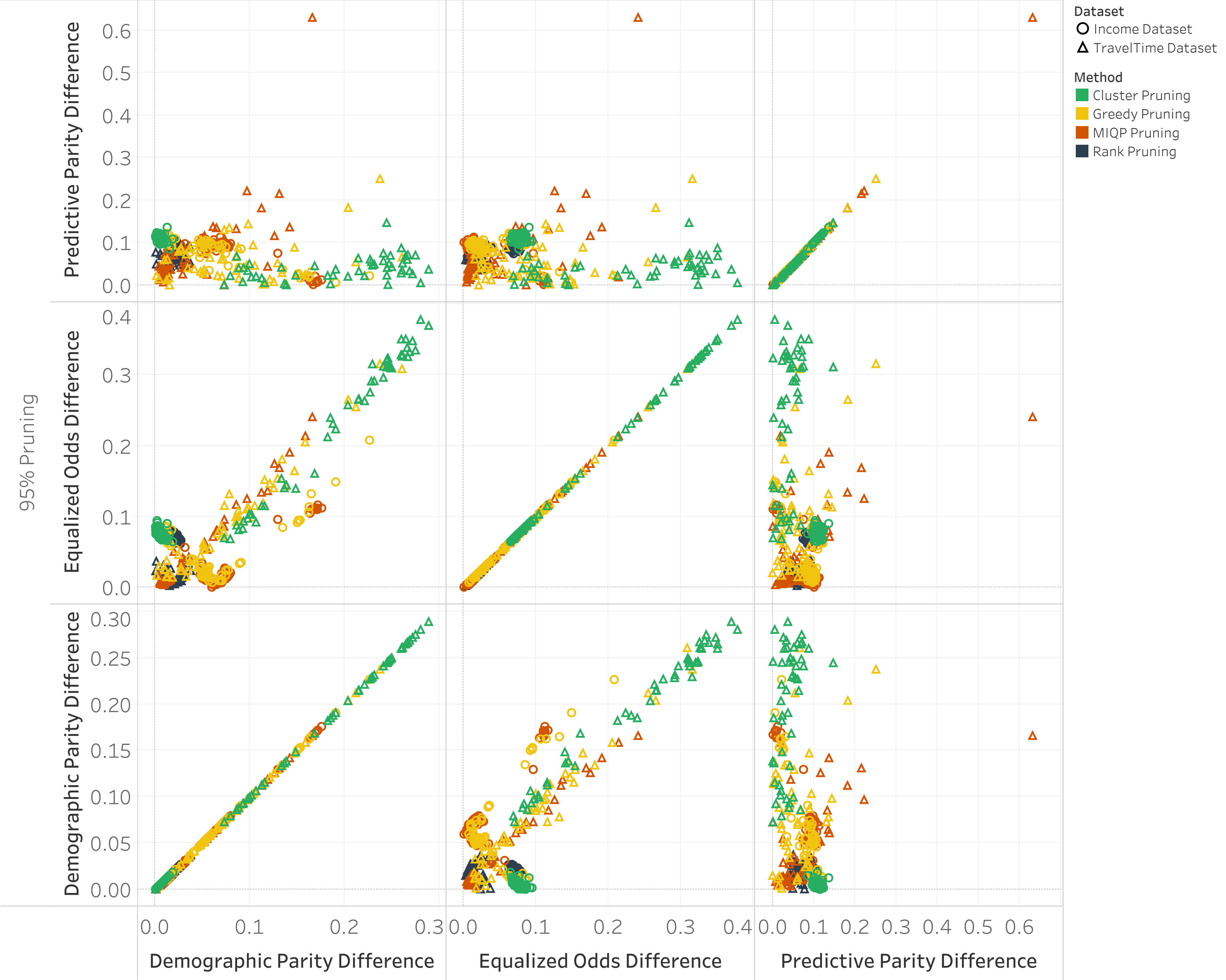

Dataset
○ Income Dataset
△ TravelTime Dataset
Method
🟩 Cluster Pruning
🟨 Greedy Pruning
🟧 MIQP Pruning
⬛ Rank Pruning
70% Pruning
Predictive Parity Difference
Equalized Odds Difference
Demographic Parity Difference
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Demographic Parity Difference
Equalized Odds Difference
Predictive Parity Difference


Dataset
○ Income Dataset
△ TravelTime Dataset
Method
🟩 Cluster Pruning
🟨 Greedy Pruning
🟧 MIQP Pruning
⬛ Rank Pruning
75% Pruning
Predictive Parity Difference
0.00
0.05
0.10
0.15
0.20
Equalized Odds Difference
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Demographic Parity Difference
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.1
0.2
0.3
Demographic Parity Difference
0.0
0.1
0.2
0.3
0.4
Equalized Odds Difference
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Predictive Parity Difference


Predictive Parity Difference
0.15
0.10
0.05
0.00
80% Pruning
Equalized Odds Difference
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Demographic Parity Difference
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.1
0.2
0.3
0.0
0.1
0.2
0.3
0.4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Demographic Parity Difference
Equalized Odds Difference
Predictive Parity Difference
Dataset
○ Income Dataset
△ TravelTime Dataset
Method
■ Cluster Pruning
■ Greedy Pruning
■ MIQP Pruning
■ Rank Pruning


85% Pruning
Predictive Parity Difference
Equalized Odds Difference
Demographic Parity Difference
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Demographic Parity Difference
Equalized Odds Difference
Predictive Parity Difference
Dataset
○ Income Dataset
△ TravelTime Dataset
Method
🟩 Cluster Pruning
🟨 Greedy Pruning
🟧 MIQP Pruning
⬛ Rank Pruning


Dataset
○ Income Dataset
△ TravelTime Dataset
Method
🟩 Cluster Pruning
🟨 Greedy Pruning
🟧 MIQP Pruning
⬛ Rank Pruning
90% Pruning
Predictive Parity Difference
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Equalized Odds Difference
0.3
0.2
0.1
0.0
Demographic Parity Difference
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.1
0.2
0.3
Demographic Parity Difference
0.0
0.1
0.2
0.3
0.4
Equalized Odds Difference
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Predictive Parity Difference


Dataset
○ Income Dataset
△ TravelTime Dataset
Method
🟩 Cluster Pruning
🟨 Greedy Pruning
🟧 MIQP Pruning
⬛ Rank Pruning
95% Pruning
Predictive Parity Difference
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Equalized Odds Difference
0.4
0.3
0.2
0.1
0.0
Demographic Parity Difference
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.1
0.2
0.3
Demographic Parity Difference
0.0
0.1
0.2
0.3
0.4
Equalized Odds Difference
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Predictive Parity Difference