entropy

MDPI

Article

A General Framework for Fair Regression

Jack Fitzsimons [1,*], AbdulRahman Al Ali [2], Michael Osborne [1] and Stephen Roberts [1]

[1] Department of Engineering Science, University of Oxford, Oxford OX13PJ, UK
[2] Faculty of Business and Law, Northampton University, Northampton NN15PH, UK
* Correspondence: jack@robots.ox.ac.uk

Received: 29 April 2019; Accepted: 22 July 2019; Published: 29 July 2019

Abstract: Fairness, through its many forms and definitions, has become an important issue facing the machine learning community. In this work, we consider how to incorporate group fairness constraints into kernel regression methods, applicable to Gaussian processes, support vector machines, neural network regression and decision tree regression. Further, we focus on examining the effect of incorporating these constraints in decision tree regression, with direct applications to random forests and boosted trees amongst other widespread popular inference techniques. We show that the order of complexity of memory and computation is preserved for such models and tightly binds the expected perturbations to the model in terms of the number of leaves of the trees. Importantly, the approach works on trained models and hence can be easily applied to models in current use and group labels are only required on training data.

Keywords: machine learning; algorithmic fairness; kernel methods; constrained learning; Gaussian process; decision tree; neural network

1. Introduction

As the proliferation of machine learning and algorithmic decision making continues to grow throughout industry, the net societal impact of them has been studied with more scrutiny. In the USA under the Obama administration, a report on big data collection and analysis found that "big data technologies can cause societal harms beyond damages to privacy" [1]. The report feared that algorithmic decisions informed by big data may have harmful biases, further discriminating against disadvantaged groups. This along with other similar findings has led to a surge in research around algorithmic fairness and the removal of bias from big data.

The term *fairness*, with respect to some sensitive feature or set of features, has a range of potential definitions. In this work, impact parity is considered. In particular, this work is concerned with group fairness under the following definitions as taken from [2].

Group Fairness: *A predictor $\mathcal{H} : X \to Y$ achieves fairness with bias ϵ with respect to groups $A, B \subseteq X$ and $O \subseteq Y$ being any subset of outcomes iff,*

$$|\mathrm{P}\{\mathcal{H}(x_i) \in O | x_i \in A\} - \mathrm{P}\{\mathcal{H}(x_j) \in O | x_j \in B\}| \le \epsilon.$$

The above definition can also be described as statistical or demographical parity. Group fairness has found widespread application in India and the USA, where affirmative action has been used to address discrimination against caste, race and gender [3–5].

The above definition does not, unfortunately, have natural application to regression problems. One approach to get around this would be to alter the definition to bound the absolute difference between the respective marginal distributions over the output space. However, this is a strong requirement and may hinder the model's ability to model the function space appropriately. Rather, a weaker and potentially more desirable constraint would be to force the expectation of the marginal distributions over the output space to equate. Therefore, statements such as "the average expected outcome for population A and B is equal" would be valid. This can be seen as a constraint such that the total entropy of the partition is maximised.

The second issue encountered is that the generative distribution of groups A and B are generally unknown. In this work, it is assumed that the empirical distribution $p_A(x)$ and $p_B(x)$, as observed from the training set, is equal to or negligibly perturbed from the true generative distributions.

Combining these two caveats, we arrive at the definition:

Group Fairness in Expectation: *A regressor $f(\cdot) : X \to Y$ achieves fairness with respect to groups $A, B \subseteq X$ iff,*

$$\mathbb{E}[f(x_i)|x_i \in A] - \mathbb{E}[f(x_i)|x_j \in B] = 0$$

$$\int (p_A(x) - p_B(x))\, f(x)dx = 0.$$

There are many machine learning techniques with which *Group Fairness in Expectation* constraints (GFE constraints) may be incorporated. While constraining kernel regression is introduced in Section 3, the main focus of the paper is examining decision tree regression and respective ensemble methods which build on decision tree regression such as random forests, extra trees and boosted trees due to their widespread use in industry and hence their extensive impact on society [6]. The reason for this is to show that such an approach will not affect the order of computational or memory complexity of the model.

The main contributions of this paper are:

I We use quadrature approaches to enforce GFE constraints on kernel regression with applications to Gaussian processes, support vector machines, neural network regression and decision tree regression, as outlined in Section 3.

II We incorporate these constraints on decision tree regression without affecting the computational or memory requirements, as outlined in Sections 5 and 6.

III We derive a tight bound for the variance of the perturbations due to the incorporation of GFE constraints on decision tree regression in terms of the number of leaves of the tree, as outlined in Section 7.

IV We show that these *fair* trees can be combined into random forests, boosted trees and other ensemble approaches while maintaining fairness, as shown in Section 8.

2. Related Work

There are many ways in which the now huge volume of literature on algorithmic fairness may be split. One such approach is to break the proposed literature into three branches of research based upon the stage of the machine learning life cycle they belong. The first is the data alteration approach, which endeavours to modify the original dataset in order to prevent discrimination or bias due to the protected variable [7,8]. The second is an attempt to regularise such that the model is penalised for bias [9–13]. Finally, the third endeavours to use post-processing to re-calibrate and mitigate against bias [14,15].

The literature also differs dramatically as to what is the objective of the fairness algorithm. Recent work has made efforts towards grouping these into consistent objective formalisation [2,16]. Often, the focus of algorithmic fairness is on classification problem with regression receiving very little attention.

The approach applied to enforce fairness may be from a plethora of definitions, anti-classification [16], or fairness through unawareness as it is also referred to as [2], endeavour to treat data agnostic of protected variables and hence enforces fairness via treatment rather than outcome. The second popular method is classification parity, i.e., the error with respect to some given measure is equal across groups defined by the protected variable. Finally, calibration is the term used when outcomes are independent of protected group conditioned on risk.

Narrowing our focus to regression, two contradicting objectives once again arise, namely group level fairness and individual fairness. Individual fairness implies that small changes to a given characteristic of an individual leads to small changes in outcome. Group fairness on the other hand endeavours to make aggregate outcomes of protected groups similar. The latter is the focus of this work and an overview of where this fits into the broader litterature may be found in Table 1.

Table 1. This table is amended from [2], highlighting some of the major contribution currently in the domain of fairness in machine learning. Parity versus preference refers to whether fairness means achieving equality or satisfying the preferences. Treatment versus impact refers to whether fairness is to be maintained in treatment or process of the learning algorithm or resulting output of the system. To the best of the authors knowledge, this work is the first group fair framework for regression problems.

	Parity		**Preference**
Treatment	Unawareness [7,17]	Counterfactual [18–20]	Preferred Treatment [21]
Outcome	Group Fairness [22–24]	Individual Fairness [7,10,17]	Preferred Impact [21,25]
	Equal Opportunity [26–29]		

Our work endeavours to create group level parity of expected outcome, of Group Fairness in Expectation as introduced in this work, with application to all kernel based regression which minimise the L2 norm. This includes decision trees, Gaussian processes and multi-layer perceptrons.

Specifically to decision trees, discrimination aware decision trees have been introduced [30] for classification. They offer dependency aware tree construction and leaf relabelling approach. Later, fair forests [13] introduced a further tree induction algorithm to encourage fairness. They did this by introducing a new gain measure to encourage fairness. However, the issue with adding such regularisation is two-fold. Firstly, discouraging bias via a regularising term does not make any guarantee about the bias of the post trained model. Secondly, it is hard to make any theoretical guarantees about the underlying model or the effect the new regulariser has had on the model.

The approach offered in this work seeks to perform model inference in a constrained space, leveraging basic theory from Bayesian quadrature such that the predicted marginal distributions are guaranteed to have equal means. Such moment constraints have a natural relationship to maximum entropy methods. By utilising quadrature methods, it is also possible to derive bounds for the expected absolute perturbation induced by constraining the space. This is shown explicitly in Section 7. Ultimately, the paper develops a general framework to perform group-fair regression, an important open problem as pointed out in [23].

We emphasise to the reader that, as outlined in the next section, there are many definitions of fairness, each with reasonable motives but conflicting values. Group fairness, addressed in this work, inherently leads to individual unfairness, i.e., to create equal aggregate statistics between sub-population, individuals in each sub-population are treated inconsistently. The reverse is also true. As such, we should always think through the adverse effects of our approach before applying it in the real world. The experiments in this paper are aimed to explore and demonstrate the approach introduced, but are not meant to advocate using group fairness specifically for the task in hand.

3. Constrained Kernel Regression

We first show how one can create such linear constraints on kernel regression models. This work builds on the earlier contributions in [31], where the authors examined the incorporation of linear constraints on Gaussian processes (GPs). Gaussian processes are a Bayesian kernel method most popular for regression. For a detailed introduction to Gaussian processes, we refer the reader to [32]. However, for the reader unfamiliar with GPs specifically, they may simply think of a high dimensional Gaussian distribution parameterised by a kernel $K(\cdot,\cdot)$, with zero mean and unit variance without loss of generality. Given a set of inputs and respective outputs, $\{x_i, y_i\}_{i=1}^{N}$, split into training and testing sets, $\{\mathrm{x}_i, \mathrm{y}_i\}_{i=1}^{n}$ and $\{\bar{\mathrm{x}}_i, \bar{\mathrm{y}}_i\}_{i=1}^{N-n}$, inference is performed as,

$$\mathbb{E}[\bar{\mathrm{y}}] = K_{\bar{x},x} K_{x,x}^{-1} \mathrm{y},$$

$$\mathbb{V}[\bar{\mathrm{y}}] = K_{\bar{x},\bar{x}} - K_{\bar{x},x} K_{x,x}^{-1} K_{x,\bar{x}}$$

where $K_{x,x}$ denotes the kernel matrix between training examples, $K_{\bar{x},x}$ is the kernel matrix between the test and training examples and $K_{\bar{x},\bar{x}}$ is the prior variance on the prediction point defined by the kernel matrix. Gaussian processes differ from high dimensional Gaussian distributions as they can model the relationships between points in continuous space, via the kernel function, as opposed to being limited to a finite dimension.

An important note is that any combination of Gaussian distributions via addition and subtraction is a closed space, i.e., the sum of Gaussians is also Gaussian and so on. While this may at first appear trivial, it is, in fact, a very useful artefact. For example, let us assume there are two variables, a and b, drawn from Gaussian distributions with mean and variance $\mu_a, \mu_b, \sigma_a^2, \sigma_b^2$, respectively. Further, assume that the correlation coefficient ρ describes the interaction between the two variables. Then, a new variable c, which is equal to the difference a and b, is drawn from a Gaussian distribution with mean and variance,

$$\mu_c = \mu_a - \mu_b,$$

$$\sigma_c^2 = \sigma_a^2 + \sigma_b^2 - 2\rho\sigma_a\sigma_b.$$

We can thus write all three variables in terms of a single mean vector and covariance matrix,

$$\mu = \begin{bmatrix} \mu_a - \mu_b \\ \mu_a \\ \mu_b \end{bmatrix}$$

$$K = \begin{bmatrix} \sigma_a^2 + \sigma_b^2 - 2\rho\sigma_a\sigma_b & \sigma_a^2 - \rho\sigma_a\sigma_b & \sigma_b^2 - \rho\sigma_a\sigma_b \\ \sigma_a^2 - \rho\sigma_a\sigma_b & \sigma_a^2 & \rho\sigma_a\sigma_b \\ \sigma_b^2 - \rho\sigma_a\sigma_b & \rho\sigma_a\sigma_b & \sigma_b^2 \end{bmatrix}$$

Given any two of the above observations, the third can be inferred exactly. We refer to this as a degenerate distribution as K will naturally be low rank. If we observe that $\mu_a - \mu_b$ is equal to zero, we are thus constraining the distribution of a and b. This can easily be extended to the relationship between sums and differences of more variables.

Bayesian quadrature [33] is a technique used to incorporate integral observations into the Gaussian process framework. Essentially, quadrature can be derived through an infinite summation and the above relationship between these summations can be exploited [34]. An example covariance structure thus looks akin to,

$$K = \begin{bmatrix} \int\int p(x)K(x,x')p(x')\mathrm{d}x\mathrm{d}x' & \int p(x)K(x,x_0)\mathrm{d}x & \int p(x)K(x,x_1)\mathrm{d}x \\ \int p(x)K(x,x_0)\mathrm{d}x & K(x_0,x_0) & K(x_1,x_0) \\ \int p(x)K(x,x_1)\mathrm{d}x & K(x_1,x_0) & K(x_1,x_1) \end{bmatrix}$$

where $p(x)$ is some probability distribution over the domain of x, on which the Gaussian process is defined and against which the quadrature is performed against.

Reiterating the motivation of this work, given two generative distributions $p_A(x)$ and $p_B(x)$ which subpopulations A and B of the data are generated from, we wish to constrain the inferred function $f(\cdot)$ such that,

$$\int p_A(x)f(x)\mathrm{d}x = \int p_B(x)f(x)\mathrm{d}x.$$

This constraint can be rewritten as,

$$\int (p_A(x) - p_B(x)) f(x) \mathrm{d}x = 0,$$

which allows us to incorporate the constraint on $f(\cdot)$ as an observation in the above Gaussian process. Let $q_{A,B}(x) = p_A(x) - p_B(x)$ be the difference between the generative probability distributions of A and B; then, by setting the corresponding observation as zero, the covariance matrix becomes,

$$K = \begin{bmatrix} \int\int q_{A,B}(x)K(x,x')q_{A,B}(x')\mathrm{d}x\mathrm{d}x' & \int q_{A,B}(x)K(x,x_0)\mathrm{d}x & \int q_{A,B}(x)K(x,x_1)\mathrm{d}x \\ \int q_{A,B}(x)K(x,x_0)\mathrm{d}x & K(x_0,x_0) & K(x_1,x_0) \\ \int q_{A,B}(x)K(x,x_1)\mathrm{d}x & K(x_1,x_0) & K(x_1,x_1) \end{bmatrix}$$

We refer to these as equality constrained Gaussian processes. Let us now turn to incorporate these concepts into decision tree regression.

4. Trees as Kernel Regression

Decision tree regression (DTR) and related approaches offer a white box approach for practitioners who wish to use them. These methods are among the most popular methods in machine learning [6] in practice as they are generally intuitive even for those not from statistics, mathematics or computer science background. It is their proliferation, especially in businesses without machine learning researchers, that makes them of particular interest.

DTR regress data by sorting them down binary trees based partitions in the input domain. The trees are created by recursively partitioning the domain of input along axis aligned splits determined by a given metric of the data in each partition, such as information gain or variance reduction. In this work, we do not consider the many possible techniques for learning decision trees, but rather assume that the practitioner has a trained decision tree model. For a more complete description of decision trees, the authors refer the readers to [35].

For the purposes of this work, DTR can be described as a partitioning of space such that predictions are made by averaging the observations in the local partition, referred to as the leaves of the tree. As such, DTR has a very natural formulation as a degenerate kernel whereby,

$$K(x_i, x_j) = \begin{cases} 1, & \text{if } L(x_i) = L(x_j) \\ 0, & \text{otherwise} \end{cases}$$

where $L(\cdot)$ is the index of the leaf in which the argument belongs. The kernel hence becomes naturally block diagonal and the classifier/regressor written as,

$$f(\bar{\mathrm{x}}) = \mathbb{E}[\bar{\mathrm{y}}] = K_{\bar{x},x} K_{x,x}^{-1} \mathrm{y}$$

with $K_{\bar{x},x}$ denoting the vector of kernel values between $\bar{\mathrm{x}}$ and the observations, $K_{x,x}$ denoting the covariance matrix of the observations as defined by the implicit decision tree kernel and y denoting the values of the observations.

It is also worth noting how one can also write the decision tree as a two-stage model: first by averaging the observations of associated with each leaf and then by using a diagonal kernel matrix to perform inference. Trivially, the diagonal kernel matrix acts only as a lookup and outputs the leaf average that corresponds to the point being predicted. Let us refer to this compressed kernel matrix approach as the compressed kernel representation and the block diagonal variant as the explicit kernel representation.

5. Fairness Constrained Decision Trees

Borrowing concepts from the previous section on equality constrained Gaussian processes using Bayesian quadrature, decision trees may be constrained in a similar fashion. The first consideration to note is that we wish the constraint observation to act as a hard equality, i.e., noiseless. In contrast, we are willing for the observations to be perturbed in order to satisfy this hard equality constraint. To achieve this, let us add a constant noise term, σ^2_{noise}, to the diagonals of the decision tree kernel matrix. Similar to ordinary least squares regression, the regressor now minimises the L2-norm of the error induced on the observations, conditioned on the equality constraint, which is noise free. In the explicit kernel representation, this implies the minimum induced noise per observation, whereas in compressed kernel representation this implies the minimum induced noise per leaf.

An important note is that the constraint is applied to the kernel regressor equations, hence the method is exact for regression trees or when the practitioner is concerned with relative outcomes of various predictions. However, in the case that the observations range within $[0,1]$, as is the case in classification, then we must renormalise the output to $[0,1]$. This no longer guarantees a minimum L2-norm perturbation and while potentially still useful, is not the focus of this work.

The second consideration is how to determine the generative probability distributions $p_A(x)$ and $p_B(x)$. Given the frequentist nature of decision trees, it makes sense to consider $p_A(x)$ and $p_B(x)$ as the empirical distributions of subpopulations A and B, as described in Section 1. Thus, the integral of the empirical distribution on a given leaf, $\int_{L_i} p_A(x)\mathrm{d}x$, is defined as the proportion of population A observed in the partition associated with leaf L_i. We emphasise that how $p_A(x)$ and $p_B(x)$ are determined is not the core focus of this work and many approaches have merit. For example, a Gaussian mixture model could be used to model the input distribution, in which case $\int_{L_i} p_A(x)\mathrm{d}x$ would equal the cumulative distribution of the generative PDF over the bounds defined by the leaf. This is demonstrated in the Experimental Section. Many other such models would also be valid and determining which method to use to model the generative distribution is left to the practitioner with domain expertise.

6. Efficient Algorithm For Equality Constrained Decision Trees

At this point, an equality constrained variant of a decision tree has been described, in both explicit representation and compressed representation. In this section, we show that equality constraints on a decision tree do not change the computational or memory order of complexity. The motivation for considering the order of complexities is that decision trees are one of the more scalable machine learning models, whereas kernel methods such as Gaussian processes naively scale at $\mathcal{O}(n^3)$ in computation and $\mathcal{O}(n^2)$ in memory, where n is the number of observations. While the approach presented in this work utilises concepts from Bayesian quadrature and linearly constrained Gaussian processes, the model's usefulness would be drastically hindered if it no longer maintained the performance characteristics of the classic decision tree, namely computational cost, and memory requirements.

6.1. Efficiently Constrained Decision Trees in Compressed Kernel Representation

As Figure 1 shows, the compressed kernel representation of the constrained decision tree creates an arrowhead matrix. It is well known that the inverse of an arrowhead matrix is a diagonal matrix with a rank-1 update. Letting D represent the diagonal principal sub-matrix with diagonal elements equal to one, z being vector such that the ith element is equal to the relative difference in generative populations distributions for leaf i, $z_i = \int_{L_i}(p_A(x) - p_B(x))dx$, then the arrowhead inversion properties state that,

$$\begin{bmatrix} D & z \\ z^T & 0 \end{bmatrix}^{-1} = \begin{bmatrix} D^{-1} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix} + \rho u u^T,$$

with $\rho = -\frac{1}{z^T D^{-1} z}$ and $u = \begin{bmatrix} D^{-1}z \\ -1 \end{bmatrix}$. Note that the integral of the difference between the two generative distributions when evaluated over the entire domain is equal to zero, as both $p_A(x)$ and $p_B(x)$ must sum to one by definition and hence their differences to zero. Returning to the equation of interest, namely $f(\bar{\mathrm{x}}) = K_{\bar{x},x}K^{-1}_{x,x}\mathrm{y}$ with y as the average value of each leaf of the tree, and subbing in $K_{\bar{x},x}$ as a vector of zeros with a one indexing the jth leaf in which the predicted point belongs to and is equal to zero, as it does not contribute to the empirical distributions, we arrive at,

$$f(\bar{\mathrm{x}}) = \frac{1}{1+\sigma_n^2}\left(\mathrm{y}_j + z_j \frac{\sum_i z_i y_i}{\sum_i z_i^2}\right).$$

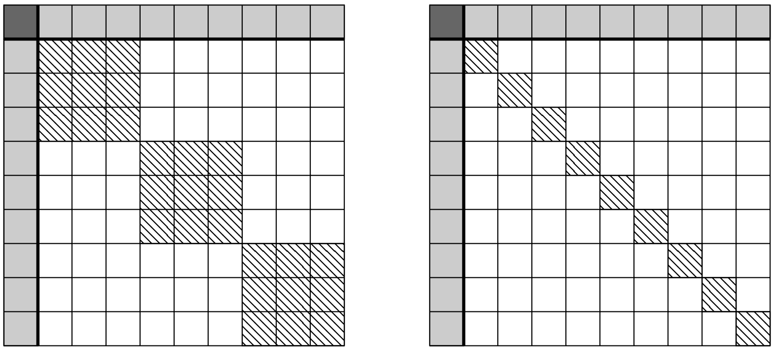

Figure 1. This is a visualisation of a decision tree kernel matrix with marginal constraint, left in explicit representation and right in compressed representation. The dark cell in the upper left of the matrix is the double integrated kernel function with respect to the difference of input distributions, which constrain the process. The solid grey row and column are single integrals of the kernel function. White cells have zero values and the dashed (block) diagonals are the kernel matrix between observations or leaves of the tree. We can note that the above, compressed representation kernel matrix is an arrowhead matrix, which we exploit to create an efficient algorithm.

The term $\frac{1}{1+\sigma_n^2}$ is the effect of the prior under the Gaussian process perspective; however, by post-multiplying by $(1+\sigma_n^2)$, this prior effect can be removed. While relatively simple to derive, the above equation shows that only an additive update to the predictions is required to ensure group fairness in decision trees. Further, if the same relative population is observed for Group A and Group B on a single leaf j, then $z_j = 0$ and no change is applied to the original inferred prediction before the constraint is applied other than the effect of the noise. In fact, the perturbation to a leaf's expectation grows linearly with the bias in the population of the leaf.

From an efficiency standpoint, only the difference in generative distributions, z, needs to be stored, which is an additional $\mathcal{O}(L)$ extra memory requirement and the update per leaf can be pre-computed in $\mathcal{O}(L)$. These additional memory and computational requirements are negligible compared to $\mathcal{O}(N)$ cost of the decision tree itself.

6.2. *Efficiently Constrained Decision Trees in Explicit Kernel Representation*

Let us now turn our attention to the explicit kernel representation case, where the D in the previous subsection is replaced with the block diagonal matrix equivalent. First, let us state the *bordering method*, a special case of the block diagonal inversion lemma,

$$\begin{bmatrix} D & z \\ z^T & 0 \end{bmatrix}^{-1} = \begin{bmatrix} D^{-1} - \rho D^{-1} z z^T D^{-1} & \rho D^{-1} z \\ \rho z^T D^{-1} & -\rho \end{bmatrix}.$$

with $\rho = -\frac{1}{z^T D^{-1} z}$ once again. Substituting this into the kernel regression equation once more, we find,

$$
\begin{aligned}
f(\bar{\mathrm{x}}) &= \begin{bmatrix} \mathbb{I}_j \\ 0 \end{bmatrix} \begin{bmatrix} D & z \\ z^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{y} \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \mathbb{I}_j \\ 0 \end{bmatrix} \begin{bmatrix} D^{-1} - \rho D^{-1} z z^T D^{-1} & \rho D^{-1} z \\ \rho z^T D^{-1} & -\rho \end{bmatrix} \begin{bmatrix} \mathrm{y} \\ 0 \end{bmatrix}
\end{aligned}
$$

where $\mathbb{I}_j$ denotes a vector of zeros with ones placed in all elements relating to observations in the same leaf. Expanding the above linear algebra,

$$f(\bar{\mathrm{x}}) = \mathbb{I}_j D^{-1} \mathrm{y} + \rho \mathbb{I}_j D^{-1} z z^T D^{-1} \mathrm{y}.$$

As D is a block-diagonal matrix, it is straight forward to show $\rho = -\left(\sum_{j \in L} \frac{m_j z_j^2}{m_j + \sigma_n^2} \right)^{-1}$ where j is iterating over the set of leaves. Note that, when $m_j = 1$ for all j, we arrive at the same value for ρ as we did in the previous subsection. We can continue to apply this result to the other terms of interest,

$$
\begin{aligned}
X_1 &= \mathbb{I}_j D^{-1} \mathrm{y} = \frac{m_j \mathrm{y}_j}{m_j + \sigma_n^2} \\
X_2 &= \mathbb{I}_j D^{-1} z = \frac{m_j z_j}{m_j + \sigma_j^2} \\
X_3 &= z^T D^{-1} \mathrm{y} = \sum_{j \in L} \frac{m_j z_j \mathrm{y}_j}{m_j + \sigma_n^2}
\end{aligned}
$$

where y_j is once again the average output observation over leaf j. The terms have been labelled X_1, X_2 and X_3 for shorthand. The computation time for the three terms, along with ρ, can be computed in linear time with respect to the size of the data, $\mathcal{O}(n)$, and can be pre-computed ahead of time, hence not affect the computational complexity of a standard decision tree. Once again, only z_j and m_j have to be stored for each leaf and hence the additional memory cost is only $\mathcal{O}(L)$. As such, we can simplify the full expression for the expected outcome as,

$$f(\bar{\mathrm{x}}) = X_1 + \rho X_2 X_3.$$

6.3. Expected Perturbation Bounds

In imposing equality constraints on the models, the inferred outputs become perturbed. In this section, the expected magnitude of the perturbation is analysed for the compressed kernel representation. We define the perturbation due to the equality constraint, not due to the incorporation of the noise, as,

$$\epsilon = z_j \frac{\sum_i z_i y_i}{\sum_i z_i^2}.$$

Theorem 1. *Given a decision tree with L leaves, with expected value of leaf observations denoted by the vector $y \in \mathbb{R}^L$ normalised to have zero mean and unit variance and leaf frequency imbalance denoted as $z \in \mathbb{R}^L$, the expected variance induced by the perturbation due to the incorporating a Group Fairness in Expectation constraint is bounded by,*

$$\mathbb{E}[\epsilon^2] \leq \frac{1}{L}$$

Proof. As the expectation of z_j is zero due to it being the difference of two probability distributions, the variance is equal to the expectation of ϵ^2,

$$\mathbb{E}[\epsilon^2] = \frac{1}{L}\sum_{j=1}^{L} z_j^2 \frac{(\sum_i z_i y_i)^2}{(\sum_i z_i^2)^2} = \frac{1}{L}\|z\|_1^2 \frac{\|z\|_2^2(\bar{z}^T y)^2}{\|z\|_2^4}$$

with $\bar{z}$ equal to z after normalisation. By Lemma 1, the expectation of the dot product $(\bar{z}^T y)^2$ is equal to $\frac{1}{L}$. Further, the 2-norm of z can be cancelled from the numerator and denominator. Finally, using the L_1, L_2 norm inequality, $\|z\|_2 \leq \|z\|_1 \leq \sqrt{L}\|z\|_2$, we can then tightly bound the worst case introduced variance as,

$$\mathbb{E}[\epsilon^2] = \frac{1}{L^2}\|z\|_1^2\|z\|_2^{-2} \leq \frac{1}{L}$$

□

Lemma 1. *Given two vectors $y, \bar{z}$ uniformly distributed on the unit hypersphere $\mathcal{S}^{L-1}$, the expectation of their dot product is zero and variance,*

$$\mathbb{E}[(\bar{z}^T y)^2] = \frac{1}{L}.$$

Proof. As the inner product is rotation invariant when applied to both $\bar{z}$ and y, let us denote the vector $\bar{z}$ as $[1, 0, \ldots, 0]$ without loss of generality. The first element of the vector y, denoted by y_0, is thus equal to $\bar{z}^T y$. The probability density mass of the random variable y_0 is proportional to the surface area lying at a height between y_0 and $y_0 + dy_0$ on the unit hypersphere. That proportion occurs within a belt of height dy_0 and radius $\sqrt{1 - y_0^2}$, which is a conical frustum constructed out of an S^{L-2} of radius $\sqrt{1 - y_0^2}$, of height dy_0, and slope $\frac{1}{\sqrt{1-y_0^2}}$. Hence, the probability is proportional to,

$$P(y_0) \sim \frac{(\sqrt{1 - y_0^2})^{D-2}}{\sqrt{1 - y_0^2}} dy_0 = (1 - y_0^2)^{\frac{L-3}{2}}$$

Substituting $u = \frac{y_0+1}{2}$. we find that,

$$\begin{aligned} P(u)du &\sim \left(1 - (2u-1)^2\right)^{\frac{L-3}{2}} d(2u-1) \\ &= 2^{L-2}(u - u^2)^{\frac{L-3}{2}} du \\ &= 2^{L-2} u^{\frac{L-1}{2}-1}(u-1)^{\frac{L-1}{2}-1} du. \end{aligned}$$

Note that this last simplification of $P(u)$ is equal to the probability density function of the Beta distribution with both shape parameters equal $\alpha = \beta = \frac{L-1}{2}$. The variance of the Beta distribution is,

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{1}{4L}.$$

Rescaling to find the variance of y_0, we arrive at $\frac{1}{L}$. As the expectation of $\mathbb{E}[\bar{z}^T y] = 0$ due to the properties of symmetry, $\mathbb{E}[(\bar{z}^T y)^2] = \frac{1}{L}$.

This is an interesting result as it implies that, if the model is not exploiting biases in the generative distribution evenly across all of the leaves of the tree, i.e., $\|z\|_1 = \sqrt{L}\|z\|_2$, then the resulting predictions receive the greatest expected absolute perturbation when averaged over all possible y.

For the explicit kernel representation, the expected absolute perturbation bound can be analysed whereby each leaf holds an even number of observations. In such a scenario, $m_i = m$ is equal for all leaves $i \in 1, \dots, L$. Substituting this into the equations for ρ, X_2 and X_3, we can find that the bounded expected perturbation is equal to,

$$\mathbb{E}[\epsilon^2] \leq \left(\frac{\sigma_n^2}{m + \sigma_n^2}\right)^2 \frac{1}{L}$$

□

For the sake of conciseness, the full derivation of the above is left to the reader but follows the same steps as the compressed kernel representation.

7. Combinations of Fair Trees

While it is intuitive to say that ensembles of trees with GFE constraints preserve the GFE constraint, for the sake of completeness, this is now shown more formally. Random forests [36], extremely random trees (ExtraTrees) [37] and tree bagging models [38] combine tree models by averaging over their predictions. Denoting the predictions of the trees at point x as $f_i(x)$ for each $i \in 1, \dots, T$, where T is the number of trees, we can easily show that the combined difference in expectation marginalised over the space is equal to zero,

$$\begin{aligned} 0 &= \int (p_A(x) - p_B(x)) \sum_{i=1}^{T} f_i(x) \mathrm{d}x \\ &= \sum_{i=1}^{T} \int (p_A(x) - p_B(x)) f_i(x) \mathrm{d}x = \sum_{i=1}^{T} 0. \end{aligned}$$

It can also be easily shown that modelling residual errors of the trees with other fair trees, such as is the case for boosted tree models [39], also results in fair predictors. These concepts are not limited to tree methods either and the core concepts set out in this paper of constraining kernel matrices can have applications in models such as deep Gaussian process models [40].

8. Experiments

8.1. Synthetic Demonstration

The first experiment was a visual demonstration to better communicate the validity of the approach. The models examined are ExtraTrees, Gaussian processes and a single hidden layer perceptron. They endeavour to model an analytic function, $f(x) = x\cos(\alpha x^2) + \sin(\beta x)$, with observations drawn from two beta distributions, $p_A(x)$ and $p_B(x)$, respectively. The parameters of the two beta distribution are presented in Table 2. Figure 2 shows the effect of perturbing the models using the approach presented to constrain the expected means of the two populations. The figure shows the greater is the disparity between $p_A(x)$ and $p_B(x)$, the greater is the perturbation in the inferred function. Both the compressed and explicit kernel representation lead to very similar plots for the tree-based models, thus only the compressed kernel representation algorithm has been shown for conciseness. Note, in the case of the ExtraTrees model, each tree was individually perturbed before being combined. Further, in the case of the perceptron, a GMM was fit to the data in the inferred latent space rather than in the original input space.

Table 2. Parameters of Beta distributions used to create synthetic samples.

Group	$p_A(x)$	$p_B(x)$
α	2	3
β	3	2

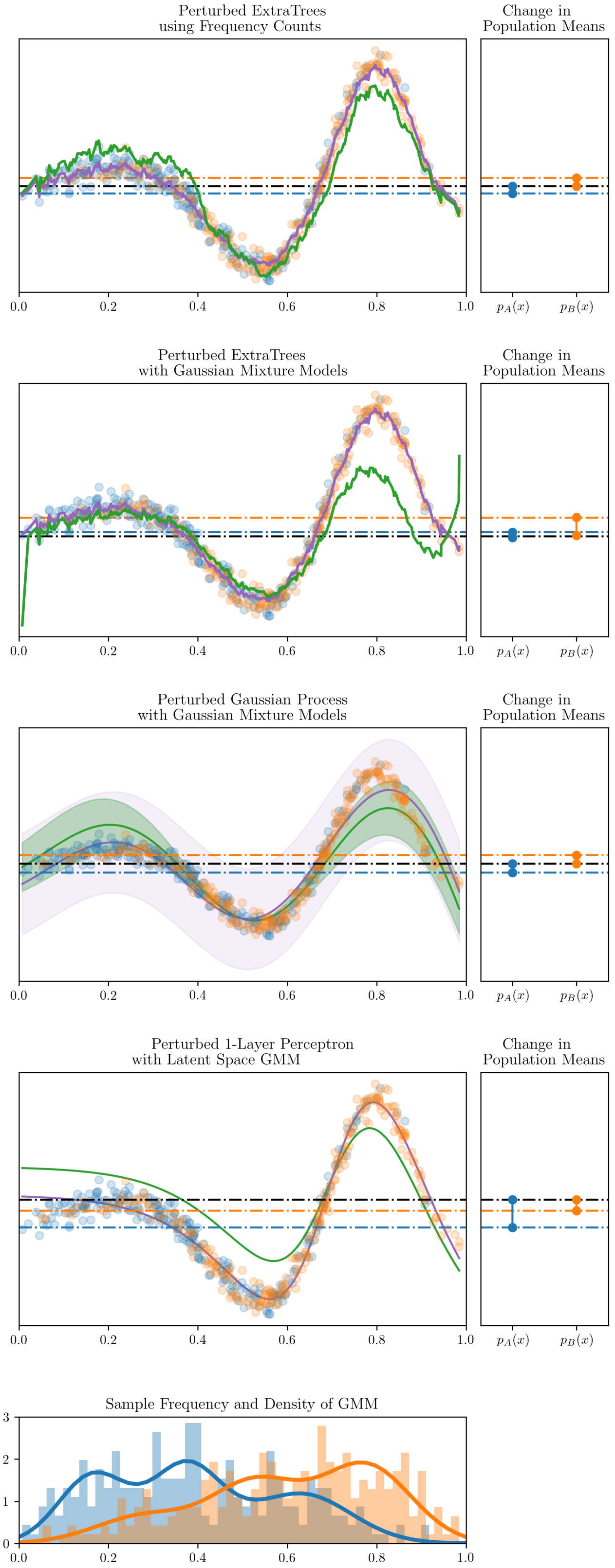

Figure 2. The figure shows synthetic data of two populations, $p_A(x)$ (blue) and $p_B(x)$ (orange). The main plots show the observations and the perturbation to the respective models. Purple functions identify the original inferred functions and green indicates the fair perturbed inferred functions. Below, the main plots show a normalised histogram of the observations for the $p_A(x)$ and $p_B(x)$ populations, respectively, along with the PDF of the Gaussian mixture model of their respective densities. To the right shows how the expected mean of the two populations have been perturbed to be equal.

A downside to group fairness algorithms more generally, as pointed out in [7], is that candidate systems which impose group fairness can lead to qualified candidates being discriminated against. This can be visually verified as the perturbation pushes down the outcome of many orange points below the total population mean in order to satisfy the constraint. By choosing to incorporate group fairness constraints, the practitioner should be aware of these tradeoffs.

8.2. ProPublica Dataset—Racial Biases

Across the USA, judges, and probation and parole officers are increasingly using algorithms to aid in their decision making. The ProPublica dataset (https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis) contains data about criminal defendants from Florida in the United States. It is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm [41], which is often used by judges to estimate the probability that a defendant will be a recidivist, a term used to describe re-offenders. However, the algorithm is said to be racially biased against African Americans [42]. To highlight the proposed algorithm, we first endeavoured to use a random forest to approximate the decile scoring of the COMPAS algorithm and then perturbed each tree to remove any racial bias from the system.

The two subpopulations we considered constraining are thus African American and non-African American. We encode the COMPAS algorithms decile score into an integer between zero and ten such that minimising L_2 perturbation is an appropriate objective function. The fact the decile scores are bounded in $[0, 10]$ was not taken into account. The random forest used 20 decision trees as base estimators and the explicit kernel representation version of the algorithm was used for the sake of demonstrative purposes.

Figure 3 presents the marginal distribution of predictions on a 20% held out test set before and after the GFE constraint was applied. It is visible that both the expected outcome for African Americans is decreased and for non-African Americans is increased. Notice that, while the means are equal, the structure of the two of distributions are quite different, indicating that GFE constraints still allow greater flexibility than more strict group fairness such as that described in Section 1. The root square difference between the predicted points before and after perturbation was 0.8. Importantly, the GFE constraint described in this work was verified numerically with the average outputs recorded as shown in Table 3.

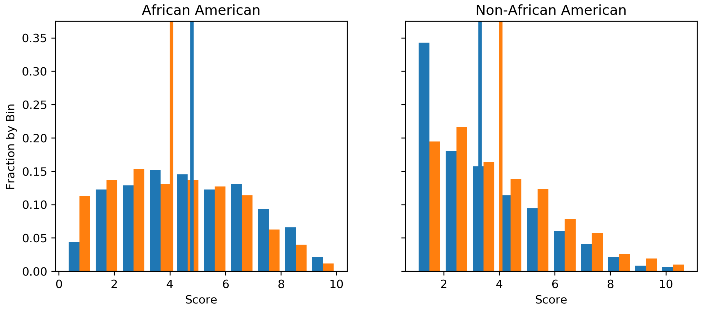

Figure 3. The figure shows the output distribution of decile scores for African Americans and non-African Americans before (blue) and after (orange) the mean equality constraint was applied. We can see that the respective means (vertical lines) become approximately equal after the inclusion of the constraint using the empirical input distribution.

Table 3. Mean score before and after GFE perturbation.

	Unconstrained	**Constrained**
African Am.	4.82	4.41
non-African Am.	3.26	4.41

8.3. Intersectionality: Illinois State Employee Salaries

The Illinois state employee salaries (https://data.illinois.gov/datastore/dump/1a0cd05c-7d17-4e3d-938d-c2bfa2a4a0b1) since 2011 can be seen to have a gender bias and bias between veterans and non-veterans. The motivation of this experiment was to show how we can deal with intersectionality issues (multiple compounding constraints) such as if one wished to predict a fair salary for future employees based on current staff. Gender labels were inferred using the employees' first names, parsed through the gender-geusser Python library. GFE constraints were applied between all intersections of gender and veteran/non-veterans, the marginals of gender and the marginals of veteran/non-veterans. Figure 4 visualises the perturbations to the marginals of each demographic intersection due to the GFE constraints. The train–test split was set as 80–20% and the incorporation of the GFE constraints increase the root mean squared error from $12,086 to $12,772, the cost of fairness. The only difference to allow for intersectionality is the z is no longer a vector, but rather a matrix with a column for each constraint. Thus,

$$f(\bar{x}) = \mathrm{y}_j + z_j(z^T z)^{-1} z^T y.$$

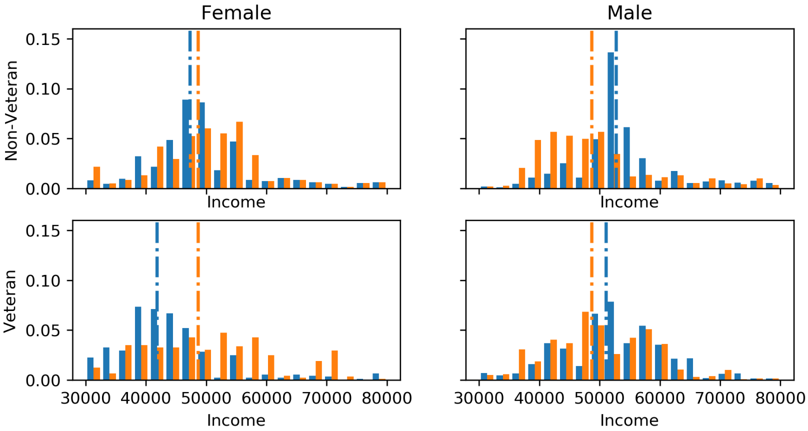

Figure 4. The above figure shows the effect of multiple GFE constraints acting on a single regression task. Blue signifies the original model and orange the perturbed model. The dashed horizontal lines signify the mean before and after perturbation.

9. Conclusions

This work offers an easily implementable approach to constrain the means of kernel regression, which has direct applicability to decision tree regression, Gaussian process regression, neural network regression, random forest regression, boosted trees and other tree-based ensemble models.

Author Contributions: Conceptualisation, J.F. and A.A.A.; Methodology, J.F.; Software, J.F.; Supervision, M.O. and S.R.; Visualisation, J.F.; Writing—original draft, J.F.; and Writing—review and editing, M.O. and S.R.

Funding: S.R. would like to thank the UK Royal Academy of Engineering and the Oxford-Man Institute of Quantitive Finance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. The Executive Office of the President of the United States; Podesta, J. *Big Data: Seizing Opportunities, Preserving Values*; White House, Executive Office of the President: Washington, DC, USA, 2014.
2. Gajane, P. On formalizing fairness in prediction with machine learning. *arXiv* **2017**, arXiv:1710.03184.
3. Weisskopf, T.E. Affirmative action in the United States and India. In *A Comparative Perspective*; Routledge: London, UK; New York, NY, USA, 2004.

4. Dumont, L. *Homo Hierarchicus: The Caste System and Its Implications*; University of Chicago Press: Chicago, IL, USA, 1980.
5. Deshpande, A. Affirmative action in India. In *Race and Inequality*; Routledge: Abingdon- on-Thames, UK, 2017; pp. 77–90.
6. Wu, X.; Kumar, V.; Quinlan, J.R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Philip, S.Y.; et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **2008**, *14*, 1–37. [CrossRef]
7. Luong, B.T.; Ruggieri, S.; Turini, F. k-NN as an implementation of situation testing for discrimination discovery and prevention. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 502–510.
8. Kamiran, F.; Calders, T. Classifying without discriminating. In Proceedings of the 2009 2nd International Conference on Computer, Control and Communication, Karachi, Pakistan, 17–18 February 2009; pp. 1–6.
9. Kamishima, T.; Akaho, S.; Sakuma, J. Fairness-aware learning through regularization approach. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, 11 December 2011; pp. 643–650.
10. Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; Roth, A. A convex framework for fair regression. *arXiv* **2017**, arXiv:1706.02409.
11. Calders, T.; Karim, A.; Kamiran, F.; Ali, W.; Zhang, X. Controlling attribute effect in linear regression. In Proceedings of the 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, 7–10 December 2013; pp. 71–80.
12. Calders, T.; Verwer, S. Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* **2010**, *21*, 277–292. [CrossRef]
13. Raff, E.; Sylvester, J.; Mills, S. Fair Forests: Regularized Tree Induction to Minimize Model Bias. *arXiv* **2017**, arXiv:1712.08197.
14. Kamiran, F.; Karim, A.; Zhang, X. Decision theory for discrimination-aware classification. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, 10–13 December 2012; pp. 924–929.
15. Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; Weinberger, K.Q. On fairness and calibration. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5680–5689.
16. Kleinberg, J.; Ludwig, J.; Mullainathan, S.; Rambachan, A. Algorithmic fairness. *Am. Econ. Assoc. Pap. Proc.* **2018**, *108*, 22–27. [CrossRef]
17. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, MA, USA, 8–10 January 2012; pp. 214–226.
18. Kilbertus, N.; Carulla, M.R.; Parascandolo, G.; Hardt, M.; Janzing, D.; Schölkopf, B. Avoiding discrimination through causal reasoning. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 656–666.
19. Kusner, M.J.; Loftus, J.; Russell, C.; Silva, R. Counterfactual fairness. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4066–4076.
20. Kilbertus, N.; Ball, P.J.; Kusner, M.J.; Weller, A.; Silva, R. The Sensitivity of Counterfactual Fairness to Unmeasured Confounding. *arXiv* **2019**, arXiv:1907.01040.
21. Zafar, M.B.; Valera, I.; Rodriguez, M.; Gummadi, K.; Weller, A. From parity to preference-based notions of fairness in classification. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 229–239.
22. Dwork, C.; Immorlica, N.; Kalai, A.T.; Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 119–133.
23. Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.S.; Pontil, M. Empirical risk minimization under fairness constraints. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 2791–2801.
24. Raff, E.; Sylvester, J.; Mills, S. Fair forests: Regularized tree induction to minimize model bias. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 13 February 2018; pp. 243–250.

25. Balcan, M.F.; Dick, T.; Noothigattu, R.; Procaccia, A.D. Envy-Free Classification. *arXiv* **2018**, arXiv:1809.08700.
26. Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; Roth, A. Rawlsian fairness for machine learning. *arXiv* **2016**, arXiv:1610.09559.
27. Bechavod, Y.; Ligett, K.; Roth, A.; Waggoner, B.; Wu, Z.S. Equal Opportunity in Online Classification with Partial Feedback. *arXiv* **2019**, arXiv:1902.02242.
28. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3315–3323.
29. Zafar, M.B.; Valera, I.; Gomez Rodriguez, M.; Gummadi, K.P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 1171–1180.
30. Kamiran, F.; Calders, T.; Pechenizkiy, M. Discrimination aware decision tree learning. In Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, NSW, Australia, 13–17 December 2010; pp. 869–874.
31. Jidling, C.; Wahlström, N.; Wills, A.; Schön, T.B. Linearly constrained Gaussian processes. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 1215–1224.
32. Rasmussen, C.E. Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning*; Springer: Berlin, Germany, 2004; pp. 63–71.
33. O'Hagan, A. Bayes-Hermite quadrature. *J. Stat. Plan. Inference* **1991**, *29*, 245–260. [CrossRef]
34. Osborne, M.; Garnett, R.; Ghahramani, Z.; Duvenaud, D.K.; Roberts, S.J.; Rasmussen, C.E. Active learning of model evidence using Bayesian quadrature. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 46–54.
35. Rokach, L.; Maimon, O.Z. *Data Mining with Decision Trees: Theory and Applications*; World Scientific: Singapore, 2008; Volume 69.
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
38. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
39. Elith, J.; Leathwick, J.R.; Hastie, T. A working guide to boosted regression trees. *J. Anim. Ecol.* **2008**, *77*, 802–813. [CrossRef] [PubMed]
40. Damianou, A.; Lawrence, N. Deep Gaussian processes. In Proceedings of the Artificial Intelligence and Statistics, Scottsdale, AZ, USA, 29 April–1 May 2013; pp. 207–215.
41. Dieterich, W.; Mendoza, C.; Brennan, T. *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*; Northpoint Inc.: Santa Maria, Philippines, 2016.
42. Dressel, J.; Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **2018**, *4*, eaao5580. [CrossRef] [PubMed]