**Introduction**

In the field of predictive modeling, the relationship between accuracy and interpretability is often seen as a trade-off. An ensemble of trees, like for example, a random forest, often provides highly accurate results but the final predictive model is too complex for human understanding (James et al., 2013). A high-accuracy, low-interpretability model like this will from now on be referred to as a *black box* model.

As the usage of black box models has become more commonplace it is increasingly important to understand how these models make their predictions (Bodó, & Janssen, 2021; Varshney & Alemzadeh, 2017). As pointed out by Rudin (2019): “the lack of transparency and accountability of predictive models can have (and has already had) severe consequences”. In the USA, for example, people have been incorrectly denied parole, or incorrectly released on bail (Wexler, 2017). Recently, the Dutch government came under scrutiny in an event called “Dutch childcare benefits scandal” (Dutch: “Toeslagenaffaire”), where, between 2013 and 2019, an estimated 26,000 parents were wrongly accused of making fraudulent benefit claims, (van Bruxvoort & van Keulen, 2021; Huisman, 2020). The European Digital Rights (EDRi) organization, a network of more than 40 European civil organizations, advocate for more regulation and transparency on artificial intelligence (AI) decision-making systems in the European Union, especially in the area of law enforcement and criminal justice. With the recent developments in language models like chatGPT by OpenAI (2022), it seems that the growth of the importance of algorithms in society is not slowing down (Future of Life Institute, n.d.). While highly accurate black box models are very useful, the call for models with higher interpretability is valid and cannot be ignored. The road to interpretability has two directions: Post-hoc explanation tools, and inherently interpretable models.

**Self driving cars, radiology, protein structures**

In practice, post-hoc explanation techniques that attempt to interpret black box models are used often. Local Interpretable Model-agnostic Explanations (LIME; Ribeiro, et al., 2016) and Shapley additive explanations (SHAP; Lundberg & Lee, 2017) are popular examples. Kaur et al. (2020) however, point out that while these tools are useful in uncovering issues with datasets or models, they are often over-trusted and misused for interpretation. Rudin (2019) argues that instead of trying to explain black models, the way forward is to create *glass box* models that are not only accurate, but also inherently interpretable. She disregards the trade-off between accuracy and interpretability as a myth. She shows that in numerous domains, highly interpretable glass box models exist that have accuracy close to, equal, or higher than black box models. Furthermore, she notes that:

“Generally, in the practice of data science, the small difference in performance between machine learning algorithms can be overwhelmed by the ability to interpret results and process the data better at the next iteration. In those cases, the accuracy/interpretability tradeoff is reversed – more interpretability leads to better overall accuracy, not worse.”

That said, high accuracy is of course still needed. Perfectly interpreting a model that is not accurate would not be a fruitful enterprise. In the current thesis, I will attempt to improve the accuracy of the Generalized Linear Mixed-Model (GLMM) tree model. As explained below, this glass box model is easily interpretable, as it results in a single decision tree. It can also produce more accurate results as compared to decision trees created using the Classification And Regression Trees (CART) mechanism (Fokkema et al., 2018; Hajjem et al., 2017; Sela & Simonoff, 2012). GLMM trees however, fall short on the accuracy of black box models. Using the *born-again tree* algorithm, combined with a Bayesian tree ensemble method, I will attempt to improve the accuracy of the GLMM tree to be comparable to that of a black box model, while maintaining the high interpretability.

**Born-Again Trees**

In 1996, Breiman and Shang introduced a manner to use a highly accurate black box model to improve the accuracy of an interpretable glass box model: the Born-Again (BA) tree algorithm. The BA approach allows the user to create a single tree model, which has an accuracy that is close to a black box model, but is much easier to interpret and can be used by humans to make predictions using only an image of the tree and the predictor variables occurring in the tree.

Breiman and Shangs (1996) BA tree algorithm follows the following four steps:

1. A black box model is fitted on the original predictor variables (**X**) to predict original outcome variable (**Y**).
2. Based on **X**,a large number of observations are artificially generated by resampling rows from **X** and columnwise permutation to create **X­gen**.
3. The black box model is applied to **X­gen** to obtain predictions **Ygen** for the observations generated in step 2.
4. A single tree is fitted on **X­gen** to predict **Ygen** resulting in a BA tree.

Whether predictive accuracy of the BA tree fitted in step 4) is improved, while its interpretability is retained, largely depends on three factors that are potentially controlled by the user:

* The accuracy of the black box model (steps 1 and 3)
* Whether the data generation method yields a realistic set of observations artificial observations **Xgen** (step 2)
* The interpretability and accuracy of the glass box model fitted on the artificial dataset (step 4)

In their study, Breiman and Shang (1996) used bagging and arcing (both CART-based tree ensembles) as black box models, smearing (a resampling method explained below) as data-generation method, and a CART tree as the glass box model. They then compared the accuracy (measured in MSE) and the interpretability (measured in number of nodes) of their BA CART tree with the regular CART tree and the black box models. They found that BA CART trees did increase the accuracy over regular CART trees, but that it came at the cost of interpretability, which worsened significantly. Below, I will describe how the three aspects are varied in the current study in an attempt to retain the high interpretability.

***The glass box model: GLMM tree***

Breiman and Shang (1996) used the CART algorithm to create the BA Tree. While a single CART is useful for decision making, as it functions as a flow-chart, the accuracy of the tree is usually low (Fernández-Delgado et al., 2014; Gacto et al., 2019; Zhang et al., 2017). Another limitation of the CART algorithm is that is does not account for multilevel data.

Multilevel models are useful for analysing longitudinal or multilevel datasets, where observations are nested within higher-level units. They account for the correlated nature of observations by estimating random effects, yielding more accurate standard errors and reducing type-I and -II errors. Examples of using multilevel models include analysing patient outcomes in different treatment centres, examining changes in mental health over time for a group of individuals, and analysing educational achievement of students nested within different schools.

The Generalized Linear Mixed-Model (GLMM) tree model has recently been developed which accounts for correlated structures in decision-tree analyses, and has been shown to yield more accurate, as well as less complex trees than gained from the CART algorithm (e.g., Fokkema et al., 2018; Hajjem et al., 2017; Sela & Simonoff, 2012). I will be using this GLMM tree model in the current study as the BA tree. This however requires the use of a black box model that also allows for multilevel analysis, which is explained below.

***The Black box model: M-BART***

Previous studies focusing on black box models in the context of BA trees, found accurate results with boosted tree ensembles (Breiman & Shang, 1996; …), support vector machines (BRON BRON), and neural networks (Craven & Shavlik, 1995; …). In the current study, I will focus only on tree ensembles. When multiple trees are combined to form tree ensembles, as in the BART, Random-Forest, Arcing, Bagging, or Boosting algorithms, the predictive accuracy is greatly improved and the risk of over-fitting is reduced, at the cost of dramatically decreasing the interpretability. Random forests, for example, use bootstrap aggregating (bagging) to build many decision trees and then average their predictions (Breiman, 2001) and have been shown to be effective in a wide range of applications, including image classification and bioinformatics (Caruana & Niculescu-Mizil, 2006).

In the current study, I will use Bayesian Additive Regression Trees (BART) ensembles. The BART model uses a Bayesian framework to learn the ensemble, using prior distributions on the tree structure and the model parameters. Additionally, BART models result in a Posterior Probability Distribution (PPD), rather than a single point estimate (Chipman, et al., 2010). BART based ensembles have been shown to outperform Boosting, Neural Networks, Random Forests, and a host of other machine learning algorithms in terms of predictive accuracy (Chipman et al., 2010; Dorie et al., 2019).

Though the BART model assumes independence among observations, it can be altered to allow for multilevel analysis (Sparapani et al., 2021; Tan, 2018; Wundervald et al., 2022) to create the Multilevel BART (M-BART) model. In the current thesis I will be utilizing this model, allowing the BA tree to be a multilevel tree as well. The PPD resulting from the BART models will be used in the data generation method.

***The Data Generation Method: Smearing and PPD sampling***

*Smearing*

In order for the BA tree to have a high accuracy, it needs a large sample size to build the tree (Breiman & Shang 1996). Breiman and Shang therefore suggest manufacturing data to increase the sample size. In their study, they apply *smearing* on the original set of predictor variables X, to create a new dataset **X­gen** with size ngen.The smearing algorithm requires the user to a-priori specify the probability that a value in the data matrix will be permuted: *palt*. The smearing algorithm repeats the following steps:

1. Sampling: A row *i* (*i* = 1, … , *N)* is randomly selected from **X**.
2. Permutation: for every predictor *j* (*j* = 1, … , *P)*, a draw is taken from a binomial distribution with p = *palt* and n = 1, which determines whether the current value *xij* is retained, or replaced by a random draw from x***j***.
3. Step 1. and step 2. are repeated for ngen amount of times to create **Xgen**.
4. Using the black box model defined earlier, **Xgen** is used to create **Ygen**.

Where *N* is the number of participants, *P* is the number of predictors, and *palt* is a threshold number that is set beforehand between 0 and 1. Note that for *palt* = 1, we are performing permutation and when *palt* = 0, we are sampling random participants as is. *palt* thus represents the degree of permutation. *palt* can be optimized for more accurate results. Breiman and Shang (1996) use *palt* = 0.25 and *palt* = 0.50 in their study, but they do not go into detail on the effect of *palt* on the accuracy or the interpretability. In the current thesis, I will test different values of *palt* (*palt* = 0, *palt* = 0.25, *palt* = 0.50) to get a better understanding of its effect. Furthermore, I will introduce a novel data generation method: PPD sampling.

*PPD sampling*

There has not been significant improvement to smearing as the data generation method, but the use of permutation may lead to unrealistic data patterns and has been criticized (Hooker et al., 2021). For this reason, PPD sampling can be utilized to create **Xgen­** and **Ygen­** that more closely resemble the original data. The PPD sampling algorithm repeats the following steps:

1. Sampling: A row is randomly selected from **X*i*** (*i* = 1, … , *N)*.
2. A random prediction is sampled from the PPD of the corresponding row from step 1.
3. Step 1. and step 2. are repeated for Ngen amount of times to create **Xgen** and **Ygen**.

Data generated in this manner should more closely resemble the original data and thus lead to a more accurate BA tree. As stated before, this is possible when using a M-BART or BART model as a black box.

***Research Questions and Hypotheses***

The main goal of the current paper is to study how BART-based BA GLMM trees perform to “regular” GLMM trees. As stated before, I prioritize interpretability over accuracy when comparing these models, but accuracy should still be compared. The first two research questions are thus:



**RQ1:** Does a BART-based BA approach improve the predictive accuracy of GLMM trees?

**RQ2:** Does a BART-based BA approach improve the interpretability of GLMM trees?

I will also compare BART-based BA GLMM trees to smearing-based BA GLMM trees, as proposed by Breiman and Shang (1996). This leads us to the third research question:

**RQ3:** In terms of predictive accuracy and interpretability, does a BART-based BA approach outperform a smearing-based approach?

Finally, I will study two variables whose effect on interpretability and accuracy, although briefly discussed by Breiman and Shang (1996), is not yet clear. These variables are N­gen, and *palt*. Breiman and Shang (1996) mention N­gen should be large, but not how large or what effects its’ size has. The same can be said for *palt*. It is only mentioned that the authors used *palt* = 0.25 and *palt* = 0.50. The final research question is thus:

**RQ4** In terms of predictive accuracy and interpretability, what effect does varying Ngen and *palt* in BA trees have?

**Methods**

***Datasets***

The hypotheses were tested on four different benchmark datasets described below. The datasets all had a multilevel structure. Because of computational limitations, random samples of N = 1000 (when possible) were taken for analyses. For time-series data, random patients were sampled instead of random observations, so that every level had enough data.

*AIDS Clinical Trials Group Study 175 (ACT)*

This dataset was created by Hammer et al. (1996) measures human immunodeficiency virus type 1 (HIV-1) infected patients and measures the amount of CD4 T cells present in the blood as outcome measure. A number of drugs, demographics and confounding variables are present to predict CD4 count. Multiple measurements were taken over time within patients, creating a multilevel structure. There are 6417 observations from 2139 patients 24 variables. A sample of N = 1000 was taken, containing 377 patients.

*Safety*

This dataset made publicly available by Hox et al. (2017). A sample of 100 streets are selected, and on each street a random sample of 10 persons are asked how often they feel unsafe while walking that street. The question about feeling unsafe is asked using three answer categories: 1 = never, 2 = sometimes, 3 = often. Predictor variables are age and gender; street characteristics are an economic index (standardized Z-score) and a rating of the crowdedness of the street (7-point scale). The multilevel structure comes from the person that are nested in the streets. No sample was taken as the original data has N = 1000 already.

*Marriage*

This is a simplified version of a dataset created by Lax and Phillips (2009) where support for gay marriage in the USA is predicted based on age, education, gender, Christianity and political vote. 6525 Participants are nested within the 49 states they live in. A sample of N = 1000 containing all 49 states was taken.

*Early Childhood Longitudinal Study Kindergarten Class of 2010-11 (ECLS-K*)

Mulligan et al. (2016) collected data about kindergarteners over the course of five years. The data includes information on selected child and family characteristics, such as poverty status, parental education, family type, and primary home language. This is used to predict the children's knowledge and skills in math, reading and science, in separate datasets. As multiple measurements are taken within the children, we get a multilevel structure. For the math and reading datasets, samples of N = 1000 containing 200 children were taken. For the science dataset a sample of N = 999 containing 333 children were taken, as the science test was only measured at three ages.

We also transformed the dataset to include school wide statistics. We calculated the percentage of males, the mean Social Economic Score, the mean race, and mean certification per school. This new data is called ECLSK-School was used to predict the reading score of the individual children, nested in schools, and is not longitudinal. From this dataset a sample of N = 999 containing 146 schools was taken.

***BART model***

BART models are created by first creating K trees (usually 200) with a single root node. The mean of this node is sampled from a prior, which is a normal distribution where 95% of the means lie between y­min and ymax (when using shrinkage parameter *k* = 2). The model then calculates the residuals and goes into the second iteration, where the trees (number of splits, or means) are randomly permuted based on the priors. A tree can grow extra nodes of prune off nodes based on a beta distribution prior. Usually, = 0.95 and = 2 is set so that most trees end up having size 2 or 3 (55% and 28% respectively). The number of iterations is usually set to 1000. Every iteration, trees that improve the fit are favored. The end result is not a single per observation but a distribution of predicted values for every iteration. This distribution is called the Posterior Probability Distribution (PPD). The first 200 iterations are usually very inaccurate and are thus removed. These are called the burn-in samples. For a more detailed explanation of the BART model, read Chipman et al. (2010).

All parameters can be optimized by cross-validation, but this is computationally very demanding. As the “standard” parameters are usually effective, it is recommended to use them instead (Chipman et al., 2010; Sparapani et al., 2021). Using the standard parameters is not only computationally advantageous, but is a more data-centric approach that is less likely to lead to overfitting (Carnegie, 2020).

*Ngen*

Breiman and Shang (1996) state that a larger size of Ngencould lead to more accurate trees. In their own preliminary experiments, they generated data that was a factor 10 to 20 times the original training set (i.e., Ngen *­*= 10 to Ngen *­*= 20). They found trees that were more accurate, but so large that they were not interpretable. For their main experiment, they used Ngen *­*= 1. They further stated that their results were not sensitive to the value of Ngen*.* In the current thesis,I will attempt to replicate these results by varying Ngen between 1, 5, and 10, and studying the result on interpretability and accuracy. For computational purposes, I weighted the artificial data (**Xgen, Ygen**)when fitting the BA Trees equally with wi = 1 / Ngen.

***Procedures***

First, samples of N = 1000 (or 999, see *Datasets* section above) were taken from each dataset because of computational limitations. The data was then randomly split in ten equal folds to use in 10-fold Cross Validation (10-CV). When a fold was assigned as test set, it was checked whether all factor levels in the test set were also present in the training set. If not, the data was moved to the training set. Next, three algorithms were applied to the data:

1. GLMM tree.
2. M-BART.
3. BA GLMM tree.

To create the M-BART based BA GLMM tree, artificial datasets were created according to the BA-tree algorithm described in the Introduction. That is, a set of artificial observations **Xgen** was generated using resampling and permutation. The response variable **Ygen** was sampled from the PPD of the M-BART ensemble (in the case of PPD sampling), or generated using the M-BART model (in the case of smearing). The GLMM tree algorithm was the applied to this artificial dataset (**Xgen**, **Ygen**). The parameters used for resampling and permutation (Ngen and *palt*) were varied according to the design presented below:

* Ngen: The sample size of the artificial dataset. Ngen is varied between [1, 5, 10] times the size of the original dataset.
* *palt*: A parameter of the smearing data-generation method. *palt* is varied between [0.00, 0.25, 0.50].

This gives us three BART-based BA GLMM tree models: [**M**BAbart.1, **M**BAbart.2, **M**BAbart.3], and different smearing-based GLMM tree models: [**M**BAsmear.1, **M**BAsmear.2, …, **M**BAsmear.9]. The generated datasets were weighted to compensate for increased sample size leading to increased power and overfitting. Namely, the weights were [1, , ] corresponding to the size of the artificial dataset.

The running of the analyses was repeated for all 10 train/test splits. The 10-CV was repeated 10 times, resulting in 10 × 10 = 100 different values of MSE and Tree Size for every model on every sample.

***Comparisons***

The models were compared on interpretability and accuracy. These are measured in Tree Size, and MSE respectively.

*Tree Size*

The size of trees is taken as a measure for interpretability. It is defined as the number of splits in the tree. Tree size will not be computed for the BART ensembles as they contain a very large number of trees.

*R2*

To compare the predictive accuracy of the models, the mean squared error (MSE) will be computed on test observations. This is divided by the variance of the dependent variable (y) to obtain a measure for R2 (Consonni et. al, 2010), which can be compared across different datasets and models. R2 measures how much of the variance of y is explained by the prediction of the model. Strictly, it ranges from 0 (y cannot be explained by prediction of model), to 1 (y is perfectly explained by prediction of model). This version of R2 however, can reach below 0 indefinitely when the MSE is higher than the variance of y. It is precisely 0 when the MSE is equal to the variance of y.

***Software***

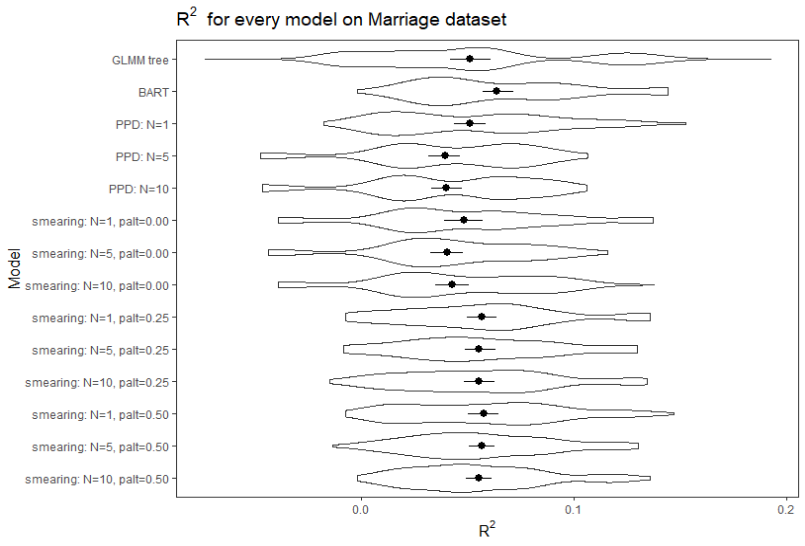
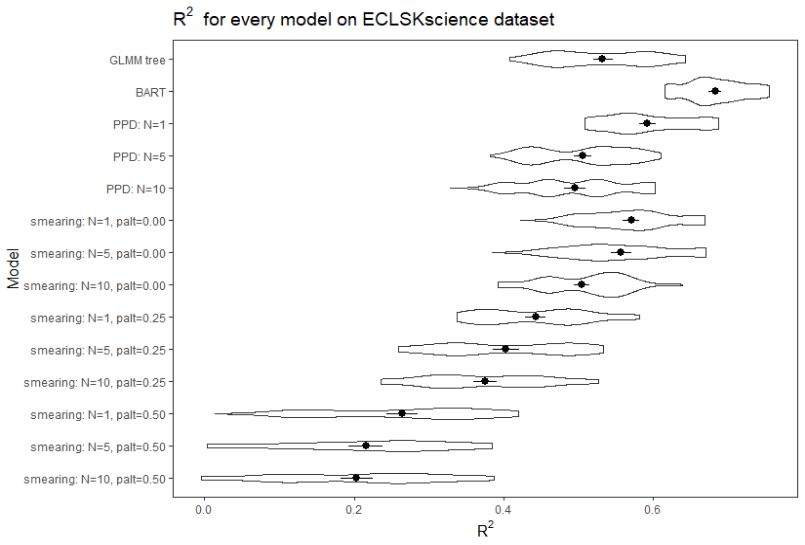
All analyses were performed using R Statistical Software (v4.1.1; R Core Team, 2021). The BART models were calculated using the rbart\_vi function in the dbarts package (v0.9-20; Dorie, 2021). The GLMM trees were made using the lmertree function in the glmertree package (v0.2-0; Fokkema et al., 2018).

**Results**

The conclusions drawn form the results depend on whether or not an accurate model was found to fit the data. When the predictive accuracy of the black box model (e.g., M-BART) is very low (R2 is close to, or equal to 0), as is the case for the Marriage and ATC datasets, the glass box models (e.g., GLMM tree, BA GLMM tree) all have similarly low predictive accuracy. Furthermore, the BA GLMM trees do not improve the predictive accuracy of the GLMM trees. The predictive accuracy in R2 of every model for the Marriage dataset is displayed in figure 1. When the predictive accuracy of the black box model is high(er), the glass box models also have higher predictive accuracy. The highest performing BA GLMM trees outperform GLMM trees in this case. In Figure 2, the predictive accuracy of every model trained on the ECLSKscience dataset is displayed.

Note that the results can be divided into two groups: 1) Datasets where an accurate model was found, and 2) Datasets where an accurate model was not found. Results in group 1) behave similarly to the ACT dataset, and results in group 2) behave similarly to the ECLSKscience dataset described above. All of these results can be found in Appendix 1 and Appendix 2. For brevity’s sake, I will only use the ECLSKscience and ACT datasets to illustrate the current results section.

Figure 1: R2 for every model on Marriage dataset

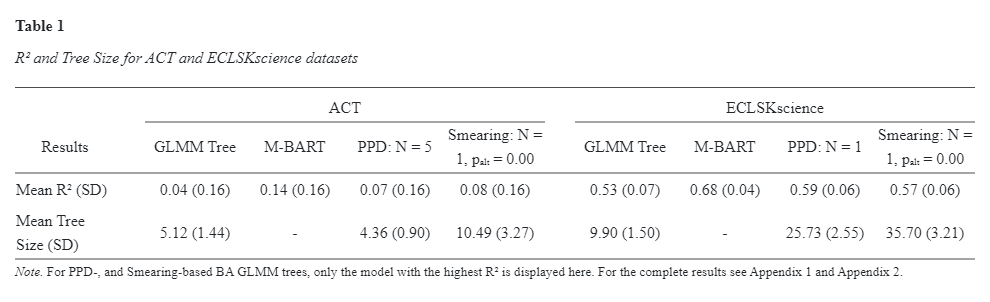
 ******

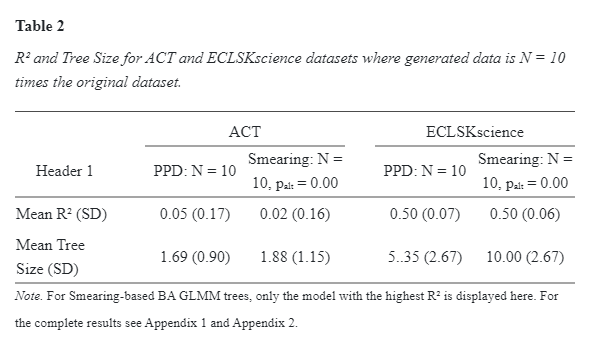
***Predictive Accuracy***

To answer the first research question, and statistically test whether it is possible for BA GLMM trees to improve the predictive accuracy of GLMM trees, the mean R2 of the GLMM tree was compared to the mean R2 of the most accurate PPD-based BA GLMM tree and the mean R2 of the most accurate smearing-based BA GLMM tree for every dataset. Tukey’s Honest Significant Difference (HSD) tests revealed that the PPD-based BA GLMM trees had significantly higher mean R2 than GLMM trees on the ECLSKmath (*p* < .001), ECLSKreading (*p* = .018), ECLSKscience (*p* < .001), and Safety (*p* < .001) datasets. There was no significant difference on the ACT, ECLSKschool, and Marriage datasets.

Using the same Tukey’s’ HSD test, we can answer the third research question and statistically test whether PPD-based GLMM trees improve the predictive accuracy of smearing-based GLMM trees. Only on the ECLSKmath (*p* < .001), and the ECLSKscience (*p* = .031) did PPD-based GLMM trees have a statistically higher R2 than smearing-based GLMM trees. It is worth noting however, that smearing-based GLMM trees only have significantly higher R2 than GLMM trees on the ECLSKscience (*p* < .001) dataset. The R2 and tree size of the ACT and ECLSKscience datasets are shown in table 1.

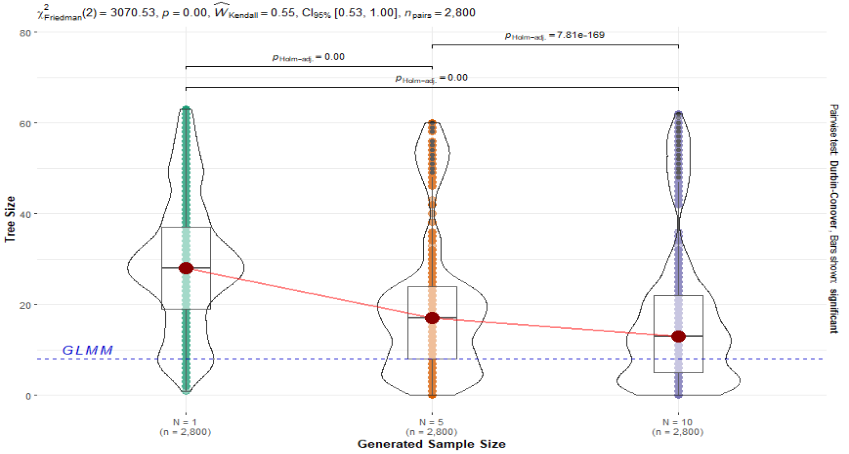
***Tree Size and N***

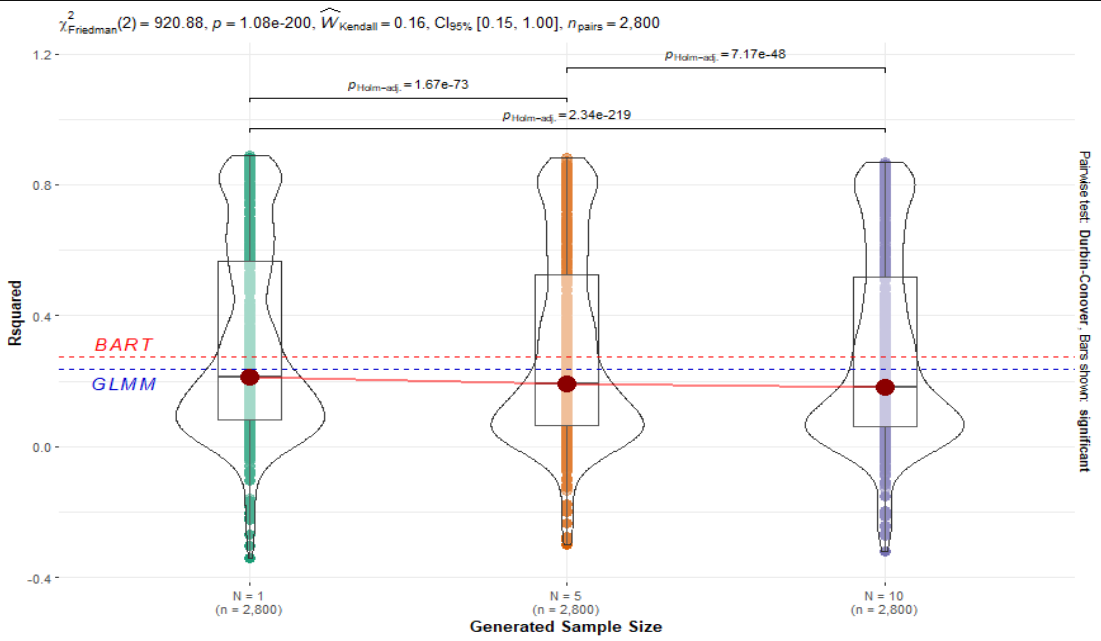
As seen in table 1, the tree size of the BA GLMM trees increases drastically compared to the GLMM tree (except for PPD: N = 5). There is a direct reverse relationship with Tree Size and N. That is, the higher N, the higher the Tree Size. As seen in Appendix 2, when choosing N = 1 for BA GLMM trees, the tree size is always higher than for GLMM trees. When we choose N = 10, as seen in table 2, the tree size is significantly lower than for BA GLMM trees than for GLMM trees. The R2 on the other hand, only increases a little amount. In figure 3 we can see this relation more clearly. 

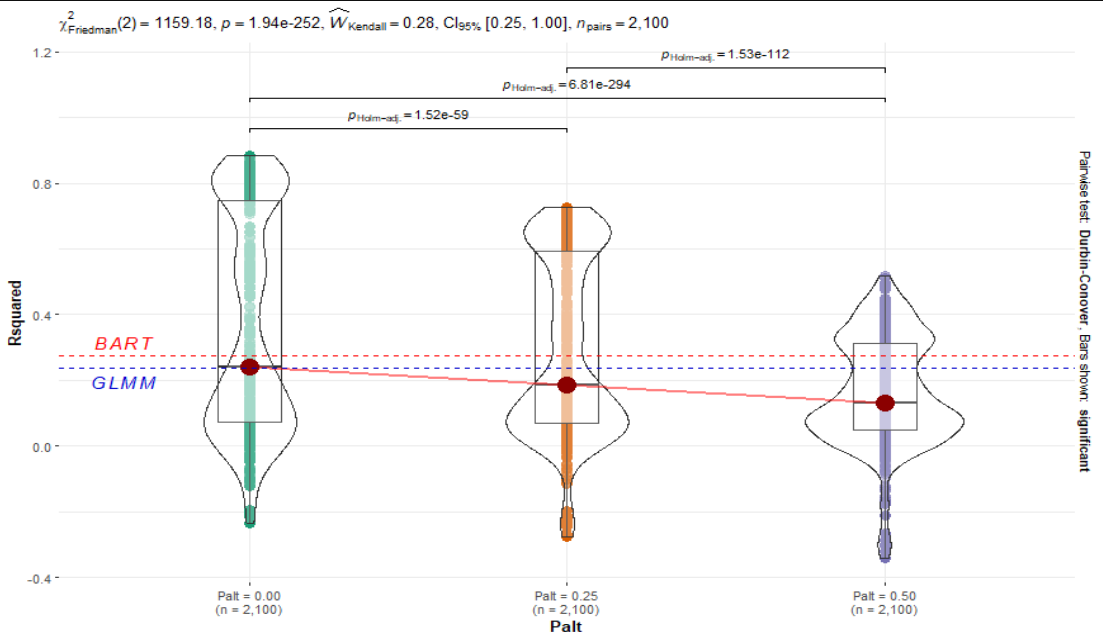


Friedman’s non-parametric repeated measures ANOVA revealed that the tree size consistently got significantly lower when N was increased (*X2*(2)= 3070.53, *p* < .001). The effect size (W­kendall­ = .55) indicates a moderate effect (Landis & Koch, 1977). Figure 4 shows that *R2* also consistently decreases when N increases (*X2*(2)= 920.88, *p* < .001). The effect size however (W­kendall­ = .16) indicates a slight effect (Landis & Koch, 1977). This suggests that increasing N could be useful when interpretability is important, while mostly maintaining the accuracy.

***palt***

For smearing-based BA GLMM trees, varying palt has a clear effect. Figure 5 shows that *R2* consistently decreases when N increases (*X2*(2)= 1159.18, *p* < .001). The effect size (W­kendall­ = .28) indicates a fair (slight/medium) effect (Landis & Koch, 1977). Figure 6 reveals that palt consistently results in the lowest tree size (*X2*(2)= 857.65, *p* < .001) with (W­kendall­ = .20) indicating a fair effect (Landis & Koch, 1977). While this seems to suggest that setting palt= 0 always leads to the best results in terms of accuracy and predictability, it is worth noting that for N = 1 and palt= 0 leads to both the highest *R2* and highest tree size (see Appendix 3). 

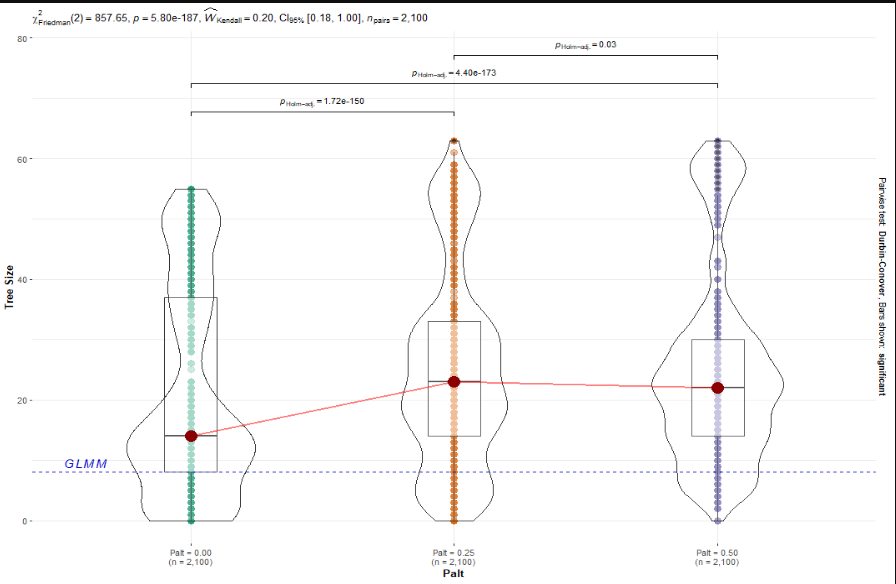


****** **Discussion**

The results have provided evidence to reject the null-hypothesis that BA GLMM trees and GLMM trees have equivalent accuracy (measured in *R2*). Tukey’s HSD test revealed that PPD-based GLMM trees obtained higher *R2* than GLMM trees on four out of seven datasets, which were datasets where the M-BART model was able to make an accurate prediction. On the three datasets where the M-BART model did not find an accurate prediction, there was no significant difference in *R2*. This is likely because the BA GLMM trees are modelled based on the predictions from the M-BART model. When the black box model does not produce accurate predictions it makes no sense to calculate further models based on those predictions. I thus recommended to only create BA tree models when a black box model can find an accurate prediction.

The results showed that the tree size for BA GLMM trees increased to uninterpretable levels (often reaching > 40) when using the most accurate model with N = 1. When increasing N, however the tree size decreases and can get lower than the GLMM tree. This increase in interpretability comes at the cost of lowering the accuracy, but this decrease is a small one. Researchers using BA tree models should thus take into account their goals and thus vary N based on the importance of interpretability in their study. As stated in the introduction I believe that there is an inherent value to interpretability that can outweigh a small decrease in accuracy. Using a high value of N could lead to a very informative tree, even if the accuracy is slightly lower.

There is slight evidence to reject the null-hypothesis that PPD-based BA GLMM trees would have equivalent accuracy to smearing-based BA GLMM trees on the same datasets. Tukey’s HSD test revealed that PPD-based GLMM trees obtained higher *R2* than smearing-based BA GLMM trees on two out of seven datasets. It is noted that smearing-based BA GLMM trees only outperformed GLMM trees on one out of seven datasets. When setting palt = 0, the smearing-based BA GLMM trees consistently had the highest interpretability and highest accuracy. This contrasts the findings by Breiman and Shang (1996) that found palt = 0.50 to lead to the most accurate results. The current study thus leads to the conclusion that the smearing should not be considered as a data-generated method for BA trees. This is because setting palt = 0 means that we are not actually smearing but resampling from the original data with replacement. Setting palt > 0 most likely introduces too much random variance in the data, leading it to more inaccurate results.



**Limitations**

*Relative MSE*

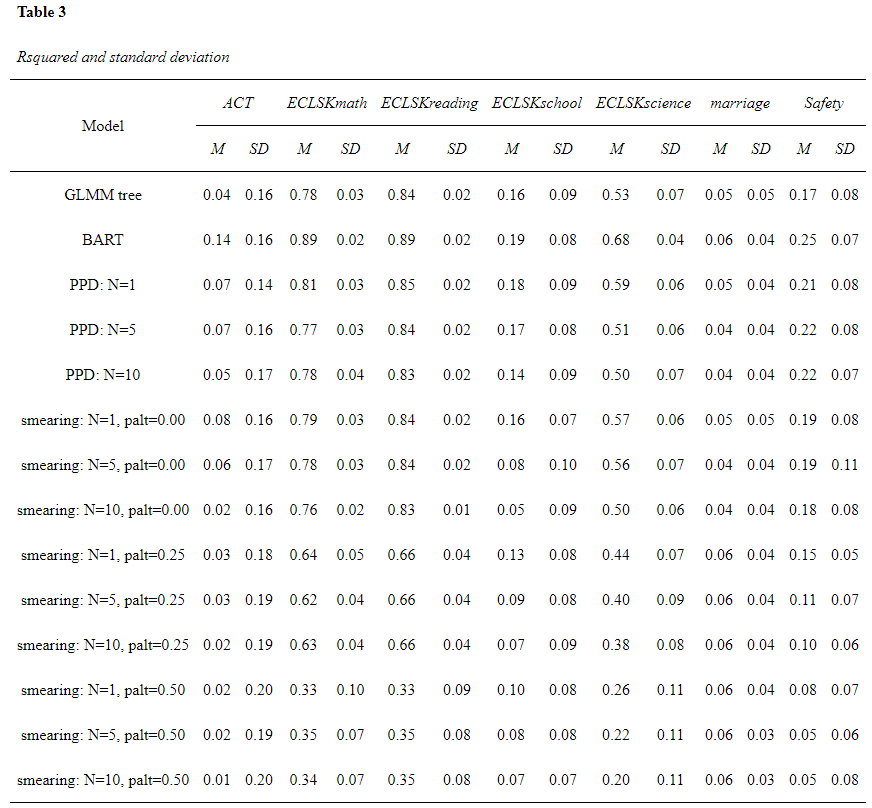
As stated before, I used relative MSE to compare the MSE obtained in different datasets. Relative MSE, as defined by Chipman et. al (2010) to be MSE/min(MSE), makes the assumption that for every split, every possible value of MSE is known.

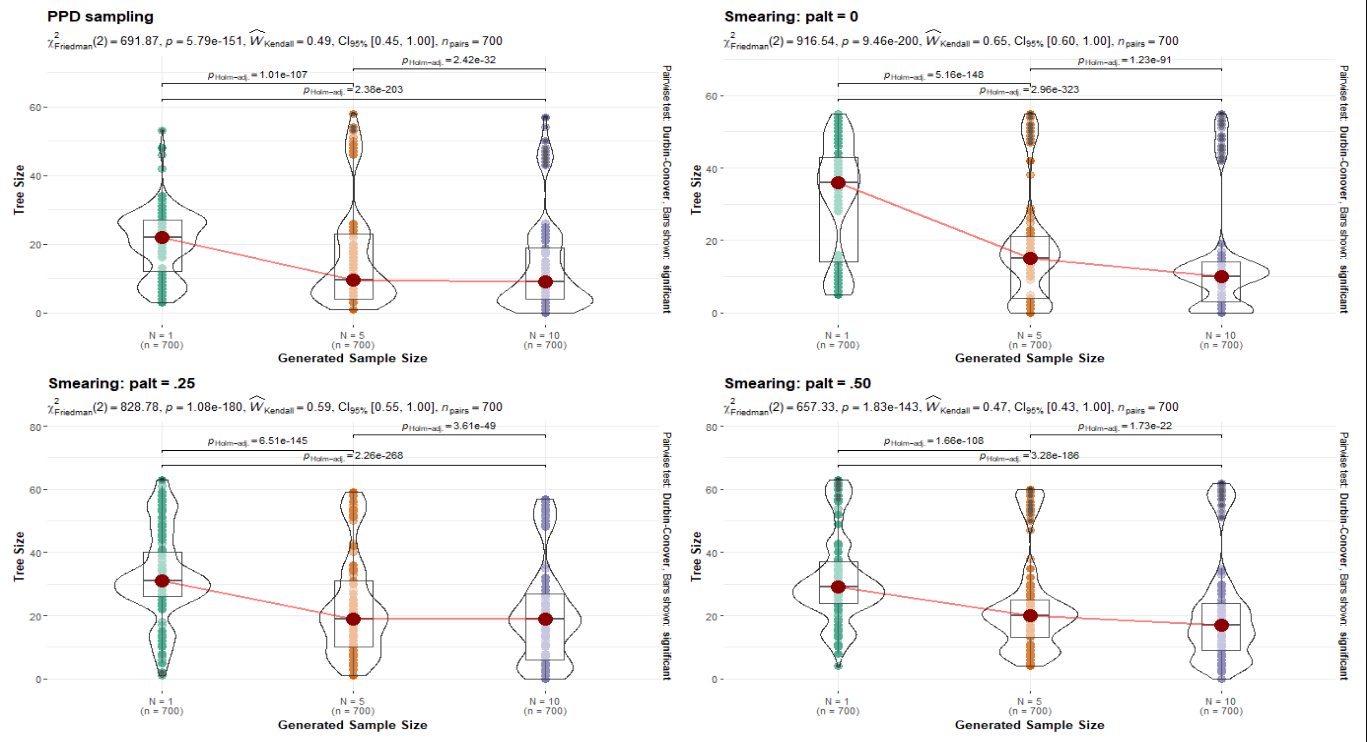
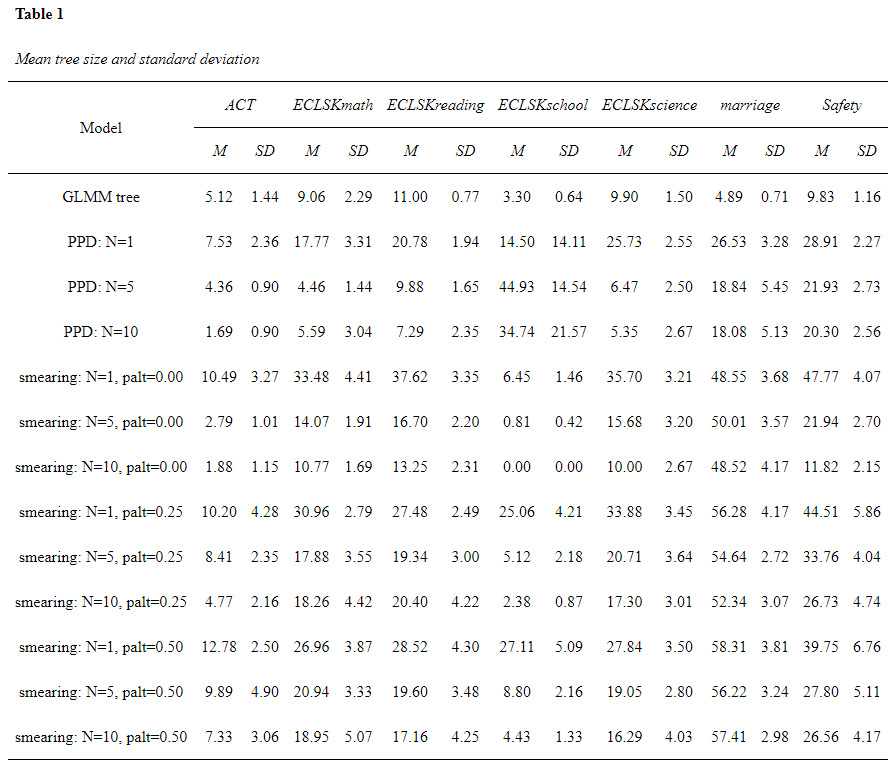
*Variance*

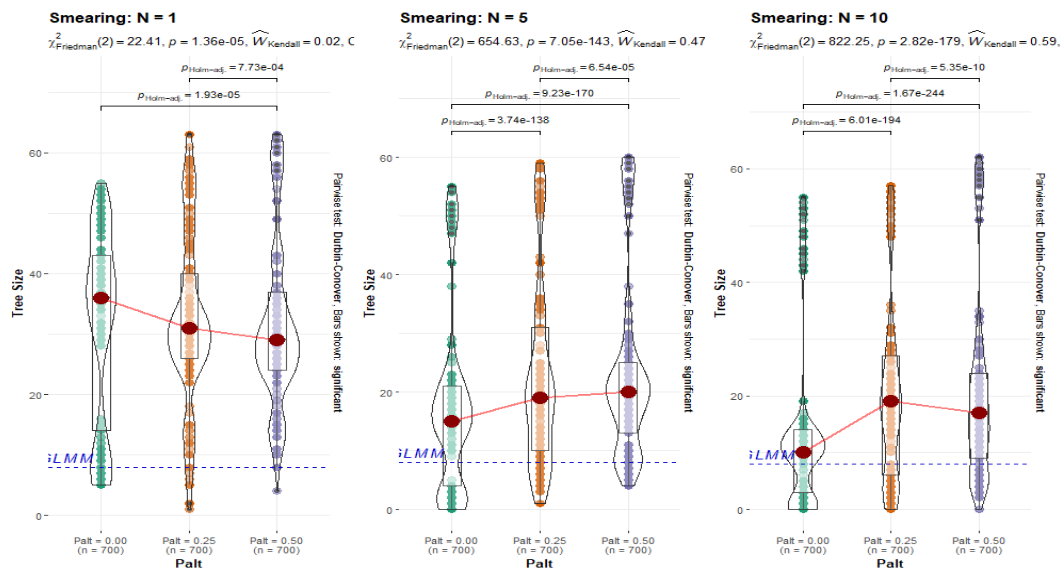
As can be seen in Figure 1, there is high variance in the obtained tree sizes. This is partly due to the fact that on the datasets ECLSKschool and ECSLKmath BA GLMM trees are created that have zero nodes. This happens because the GLMM tree makes its prediction using only the multilevel variance. This could skew the results in Figure 2, making it look like some BA GLMM trees have lower tree size on average than can actually be expected.

*Weights*

Bespreek ook dat N\_gen en sampling weights niet onderling gevarieerd zijn. Dus dat andere sampling weights mogelijk andere resultaten laten zien voor wat betere en minder goede N\_gen waardes zijn.





****

**References**

Bodó, B., & Janssen, H. (2021). Here Be Dragons–Maintaining Trust in the Technologized Public Sector. *Amsterdam Law School Research Paper*, *2021-23*

Breiman, L., & Shang, N. (1996). Born again trees. *University of California, Berkeley, Berkeley, CA, Technical Report*, *1*(2), 4.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

van Bruxvoort, X., & van Keulen, M. (2021). Framework for Assessing Ethical Aspects of Algorithms and Their Encompassing Socio-Technical System. *Applied Sciences*, *11*(23), 11187.

Carnegie, N. B., (2020, May 13). *Introduction to Bayesian Additive Regression Trees for Causal Inference* [Targeted Learning Webinar Series]. Putnam Data Sciences. <https://www.youtube.com/watch?v=9d5-3_7u5a4&t=2093s>

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168).

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266-298.

Consonni, V., Ballabio, D., & Todeschini, R. (2010). Evaluation of model predictive ability by external validation techniques. *Journal of chemometrics*, *24*(3‐4), 194-201.

Craven, M., & Shavlik, J. (1995). Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, *8, 24–30*.

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, *34*(1), 43-68.

Fair Trials and European Digital Rights (EDRi), (2022, March 1). *Civil society calls on the EU to ban predictive AI systems in policing and criminal justice in the AI Act*. Retrieved from <https://edri.org/our-work/civil-society-calls-on-the-eu-to-ban-predictive-ai-systems-in-policing-and-criminal-justice-in-the-ai-act/>

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The journal of machine learning research*, *15*(1), 3133-3181.

**Fokkema, M.** (2020). Fitting prediction rule ensembles with R package pre. Journal of Statistical Software, 92(12), 1-30. [http://doi.org/10.18637/jss.v092.i12](http://www.google.com/url?q=http%3A%2F%2Fdoi.org%2F10.18637%2Fjss.v092.i12&sa=D&sntz=1&usg=AOvVaw3CYd7ANIfiRaqJxtcrfvbz)

Fokkema, M., Edbrooke-Childs, J., & Wolpert, M. (2021). Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy Research*, *31*(3), 329-341.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior research methods*, *50*(5), 2016-2034.

Gacto, M. J., Soto-Hidalgo, J. M., Alcalá-Fdez, J., & Alcalá, R. (2019). Experimental study on 164 algorithms available in software tools for solving standard non-linear regression problems. *IEEE Access*, *7*, 108916-108939.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., ... & Merigan, T. C. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, *335*(15), 1081-1090.

Hooker, G., Mentch, L., & Zhou, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, *31*(6), 1-16.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Huisman, P. (2020). Hoe de toeslagenaffaire kon gebeuren. *Management Kinderopvang*, *26*(2), 36-37.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020, April). Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Lax, J. R., & Phillips, J. H. (2009). How should we estimate public opinion in the States?. *American Journal of Political Science*, *53*(1), 107-121.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020, April). Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).

Mulligan, G. M., McCarroll, J. C., Flanagan, K. D., & Potter, D. (2016). Findings from the Third-Grade Round of the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K: 2011): First Look. NCES 2016-094. *National Center for Education Statistics*.

OpenAI. (2022, November 30., *ChatGPT: Optimizing Language Models for Dialogue.* Retrieved January 25, 2023, <https://openai.com/blog/chatgpt/>

Future of Life Institute. (n.d.). *Artificial Intelligence.* Retrieved January 25, 2023, from <https://futureoflife.org/cause-area/artificial-intelligence/>

R Core Team (2021). R: A language and environment for statistical computing*. R Foundation for Statistical Computing,* Vienna, Austria. URL <https://www.R-project.org/>

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Sparapani, R., Spanbauer, C., & McCulloch, R. (2021). Nonparametric machine learning and efficient computation with bayesian additive regression trees: the BART R package. *Journal of Statistical Software*, *97*, 1-66.

Tan, Y. V. (2018). *Novel Applications and Extensions for Bayesian Additive Regression Trees (BART) in Prediction, Imputation, and Causal Inference* (Doctoral dissertation).

Varshney, K. R., & Alemzadeh, H. (2017). On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, *5*(3), 246-255.

Vincent Dorie (2021). dbarts: Discrete Bayesian Additive Regression Trees Sampler. R package version 0.9-20. URL <https://CRAN.R-project.org/package=dbarts>

Wexler, R. (2017). When a computer program keeps you in jail. *New York Times*.

Wundervald, B., Parnell, A., & Domijan, K. (2022). Hierarchical Embedded Bayesian Additive Regression Trees. *arXiv preprint arXiv:2204.07207*.