TITLE	**IMPROVING THE GREEDY NATURE OF TREE ALGORITHMS**
GROUP(S)	Methods & Statistics, Institute of Psychology, Leiden University
PROJECT TYPE	Combined internship & thesis
SUPERVISOR(S)	Dr. Elise Dusseldorp, Dr. Marjolein Fokkema
CONTACT INFO	m.fokkema@fsw.leidenuniv.nl elise.dusseldorp@fsw.leidenuniv.nl
BEGIN DATE	March 2022
END DATE	Latest Sept 2022
DESCRIPTION	Decision trees are highly interpretable prediction models. However, they are relatively unstable, which is mostly due to their greedy, local splitting approach. That is, the optimal split in the current node is found by minimizing the current value of the loss function. This yields low computational burden, but may not yield a globally optimal tree structure. One way to improve the performance and stability of tree methods is creating an ensemble of trees (e.g. random forest). Although ensembling successfully improves predictive accuracy of trees, it also yields black box models, requiring additional tools to explain the predictions of the ensemble. In these projects, we evaluate and improve the performance of single tree algorithms. Two or more students will work together in a group, and each will focus on one of the following topics: • **Exact trees (Meulman, Dusseldorp & Van Os, 2011)** do not employ a greedy search, but aim to find a globally optimal tree using a dynamic programming algorithm. We recently developed an R implementation of the algorithm. We will assess and compare the performance of exact trees with that of Bayesian Additive Regression Trees (BART; Chipman et al., 2010) and Evolutionary Trees (evtrees; Grubinger et al., 2014). BART can be seen as a Bayesian version of gradient boosting, and has shown comparable top-ranking predictive accuracy in many comparisons; evtrees is an algorithm for deriving single trees, that also aims to find a globally optimal tree using an heuristic approach. • **Generalized linear mixed-effects model trees (GLMM trees; Fokkema et al., 2018)** is a tree algorithm that takes into account nested (i.e., longitudinal, multilevel, hierarchical) data structures (Fokkema et al., 2021). Taking into account nested structures improves predictive performance, stability and interpretability of the resulting trees. Yet, GLMM trees rely on greedy splitting, offering further avenues for improvement. For example, Breiman and Shang (1996) proposed the born-again tree method, where an individual tree is trained on predictions generated from a highly accurate black box (e.g., gradient boosted tree ensemble; but similar approaches have been developed using neural networks) , instead of the original response variable. This improved stability and predictive accuracy of CART trees. In this project, we aim to develop a born-again GLMM tree approach, where we use (multilevel) BART

	as the black box, and fit GLMM trees to BART's predictions.
RELEVANT LITERATURE	Breiman, L., & Shang, N. (1996). Born again trees. *University of California, Berkeley, Berkeley, CA, Technical Report*, *1*(2), 4. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.367.9399&rep=rep1&type=pdf Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266-298. https://doi.org/10.1214/09-AOAS285 Craven, M. & Shavlik, J. (1995). Extracting tree-structured representations of trained networks. Advances in Neural Information Processing Systems 8, 24–30. https://proceedings.neurips.cc/paper/1995/file/45f31d16b1058d586fc3be7207b58053-Paper.pdf 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. https://doi.org/10.1080/10503307.2020.1785037 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. https://doi.org/10.3758/s13428-017-0971-x Grubinger, T., Zeileis, A., & Pfeiffer, K. P. (2014). evtree: Evolutionary learning of globally optimal classification and regression trees in R. *Journal of Statistical Software*, *61*, 1-29. https://doi.org/10.18637/jss.v061.i01 Meulman, J.J, Dusseldorp, E, & van Os, B.J. (2011). *An exact dynamic programming algorithm for regression trees*. (M van der Heijden, B Koren, R.D van der Mei, & J.A.J van Vonderen, Eds.) *Jan Karel Lenstra, the traveling science man: liber amicorum* (pp. 198–208). CWI. https://pure.tue.nl/ws/portalfiles/portal/3473039/723521.pdf
USED SKILLS	Programming, improving algorithms, running simulation experiments, running benchmark experiments using existing datasets
REQUIREMENTS	R programming experience, Statistical Learning course completed.