# Prediction rule ensembles

## balancing accuracy and interpretability in statistical prediction



Marjolein Fokkema, Leiden University

# Example dataset

Penninx et al. (2011): Predicting chronic depression

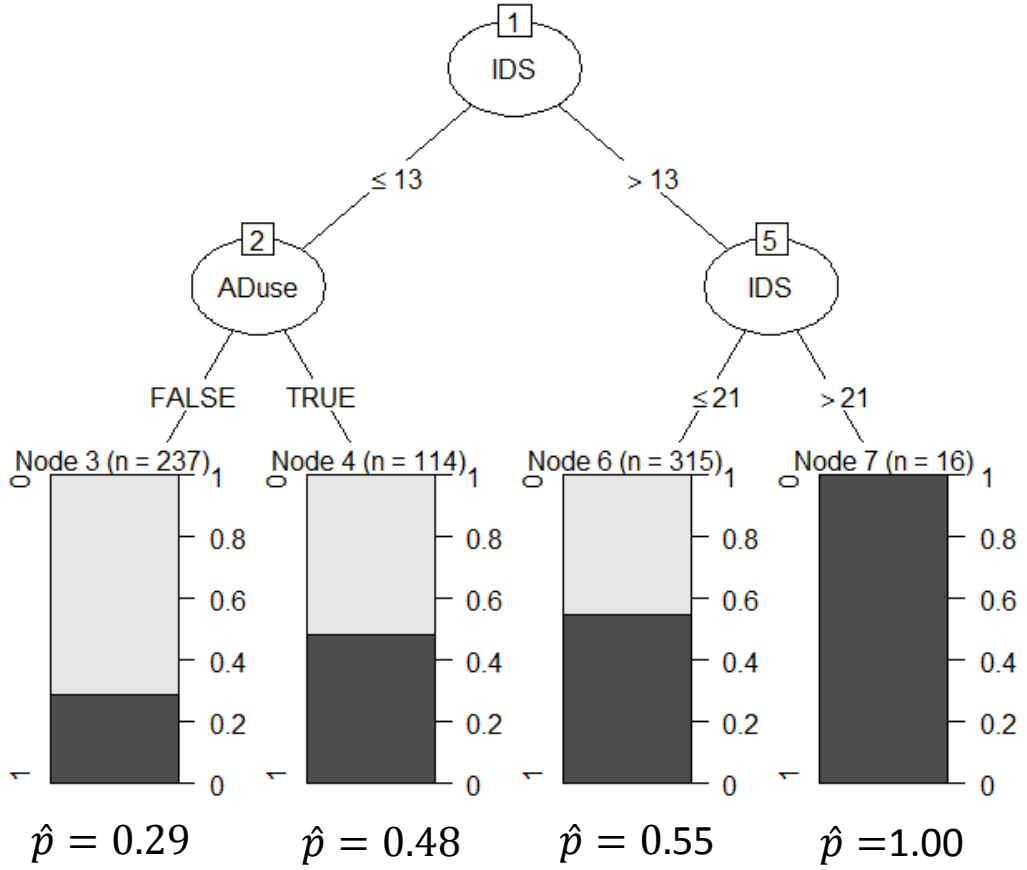
Sample: Respondents with current depressive disorder (N = 682)

Response: Depression diagnosis at two-year follow-up

20 possible predictors (baseline):

- gender, age, years of completed education
- presence of anxiety disorder(s)
- IDS (depressive symptoms)
- Receiving pharmacotherapy, psychotherapy
- BAI and FQ (anxiety symptoms)
- ….

# Conditional inference tree (Hothorn et al., 2006)

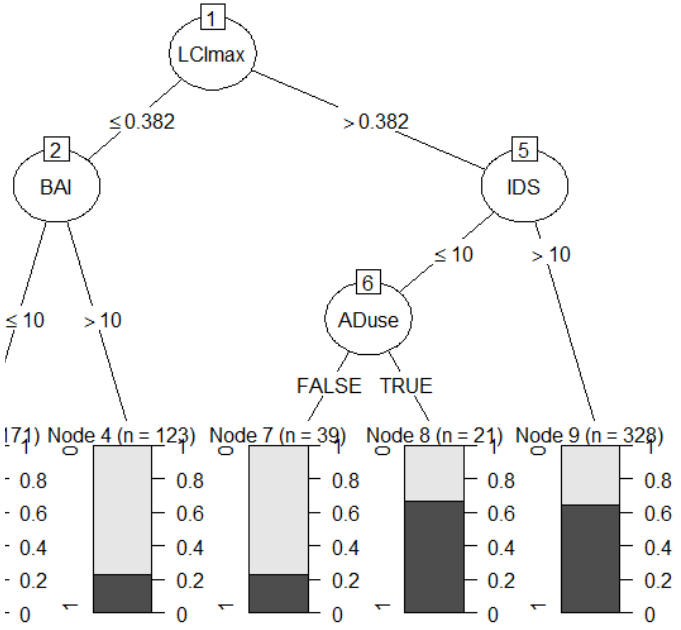


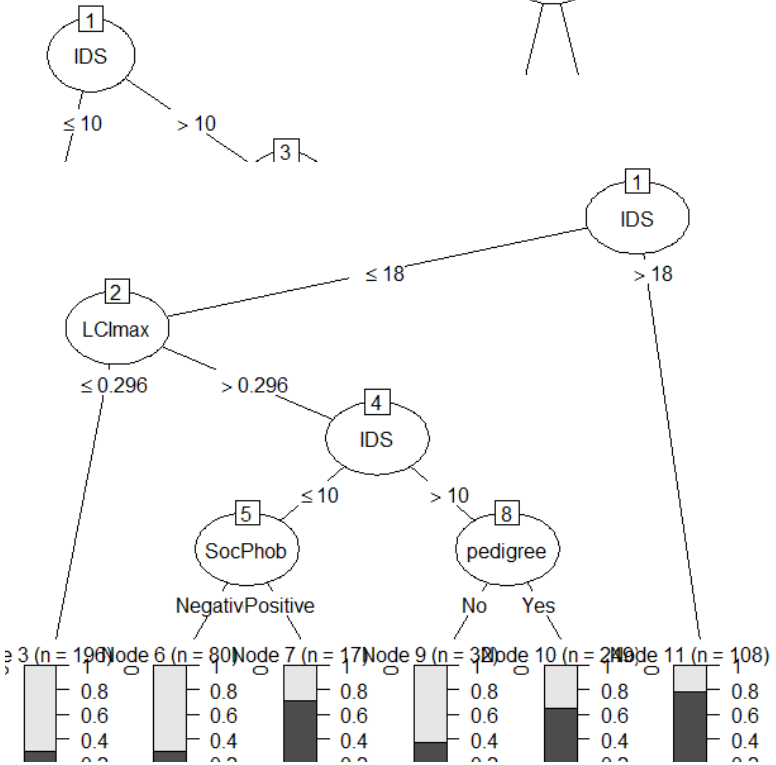
# Single trees

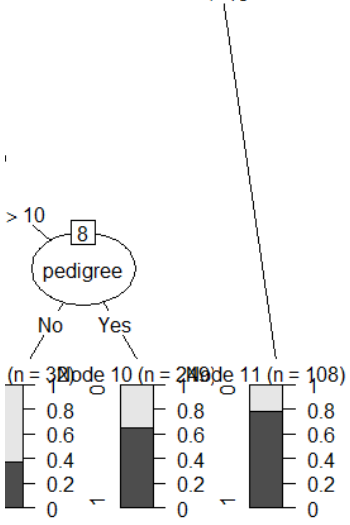
Good: Easy to interpret and apply

Bad: Not most accurate method

Ugly: Unstable



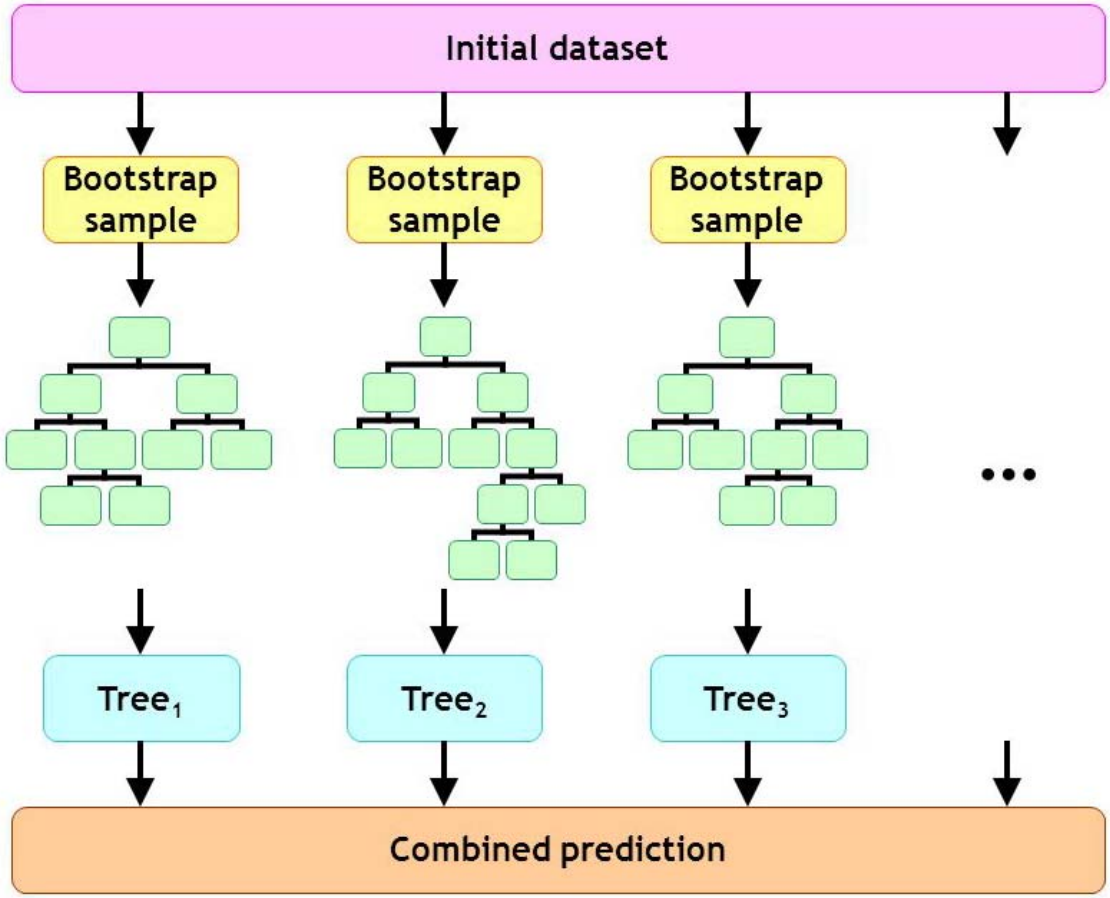
# Single trees

Good: Easy to interpret and apply

Bad: Not most accurate method

Ugly: Unstable

# Tree ensembles



# Tree ensembles

+ High predictive accuracy

- Difficult to interpret and apply
  - Many (complex) trees
  - Prediction requires lots of computation and information

Solution: Prediction rule ensembles

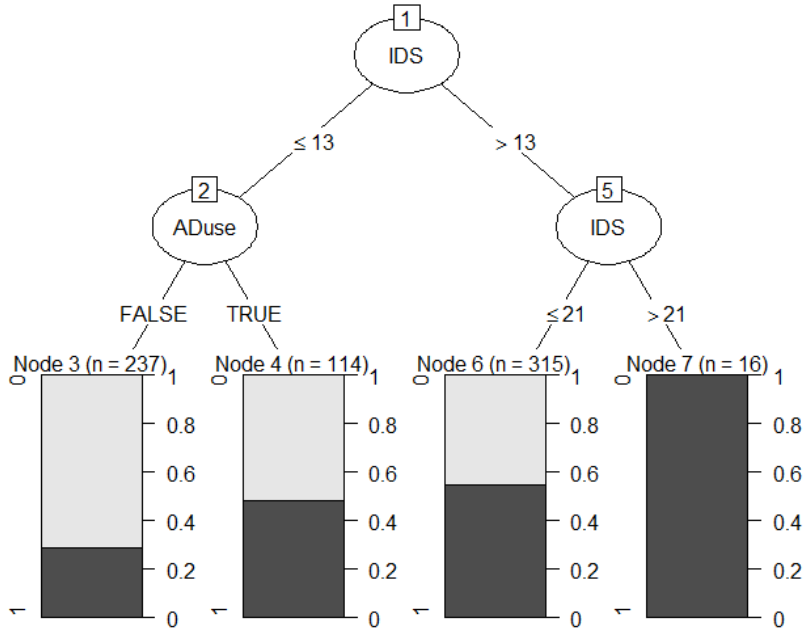
- Only keep parts that contribute most to accuracy.
- E.g., Rulefit (Friedman & Popescu, 2008); Node Harvest (Meinshausen, 2010); …

# RuleFit algorithm (Friedman & Popescu, 2008)

1) Take subsamples from training data
2) Grow tree on each sample
   - Classification and regression tree (CART) algorithm
   - Boosting (learning rate > 0)
3) Create initial ensemble
   - Include every node from every tree as a rule and
   - Include original predictor variables as linear terms
4) Select final ensemble by sparse regression on training data
   - Lasso, ridge or elastic net

# From trees to rules

$$r_2(\mathbf{x}) = I(IDS \leq 13)$$
$$r_3(\mathbf{x}) = I(IDS \leq 13) \cdot I(ADuse = FALSE)$$
$$r_4(\mathbf{x}) = I(IDS \leq 13) \cdot I(ADuse = TRUE)$$
~~$$r_5(\mathbf{x}) = I(IDS > 13)$$~~
$$r_6(\mathbf{x}) = I(IDS > 13) \cdot I(IDS \leq 21)$$
$$r_7(\mathbf{x}) = I(IDS > 13) \cdot I(IDS > 21)$$



# From trees to rules

$r_2(\mathbf{x}) = I(IDS \leq 13)$
$r_3(\mathbf{x}) = I(IDS \leq 13) \cdot I(ADuse = FALSE)$
$r_4(\mathbf{x}) = I(IDS \leq 13) \cdot I(ADuse = TRUE)$
$r_6(\mathbf{x}) = I(IDS > 13) \cdot I(IDS \leq 21)$
$r_7(\mathbf{x}) = I(IDS > 13) \cdot I(IDS > 21)$
…
$l_1(\mathbf{x}) = IDS$
$l_2(\mathbf{x}) = ADuse$
…

$$F(\mathbf{x}) = \hat{\alpha}_0 + \sum_{m=1}^{M} \alpha_m f_m(\mathbf{x})$$

| IDS | ADuse | … | $r_2$ | $r_3$ | $r_4$ | $r_6$ | $r_7$ | … |
|---|---|---|---|---|---|---|---|---|
| 5 | FALSE | … | 1 | 1 | 0 | 0 | 0 | … |
| 15 | FALSE | … | 0 | 0 | 0 | 1 | 0 | … |
| 18 | TRUE | … | 0 | 0 | 0 | 1 | 0 | … |
| 25 | TRUE | … | 0 | 0 | 0 | 0 | 1 | … |
| … | … | … | … | … | … | … | … | … |

# RuleFit algorithm (Friedman & Popescu, 2008)

1) Take subsamples from training data
2) Grow tree on each sample
   - Classification and regression tree (CART) algorithm (Breiman et al., 1984)
   - Boosting with learning rate > 0
3) Create initial ensemble
   - Include every node from every tree as a rule, and/or
   - Include original predictor variables as linear terms
4) Select final ensemble by sparse regression on training data
   - Lasso, ridge or elastic net

# R package **pre**
# (Fokkema & Christoffersen, 2017)

1) Take subsamples from training data
2) Grow tree on each sample
   - Unbiased recursive partitioning (Hothorn et al., 2006)
   - Boosting (learning rate > 0)
   - Random forest (mtry < p)
3) Create initial ensemble
   - Include every node from every tree as a rule, and/or
   - Include predictor variables as linear predictors
4) Select final ensemble by sparse regression on training data
   - Lasso, ridge or elastic net

+ support for multivariate, categorical, count and survival responses

+ ….

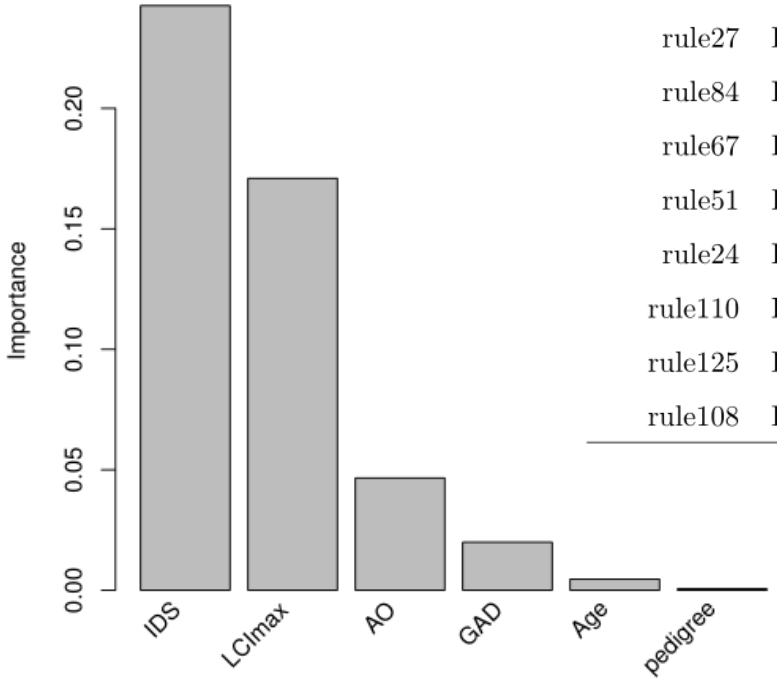
# **pre** applied to predict chronic depression

| Term | Description | Coeffient | SD | Importance |
|---|---|---|---|---|
| (Intercept) | 1 | -0.221 | 0.000 | 0.000 |
| rule3 | IDS > 10 & LCImax > 0.2632 | 0.224 | 0.494 | 0.111 |
| rule27 | IDS > 13 & LCImax > 0.3621 | 0.213 | 0.477 | 0.102 |
| rule84 | IDS <= 16 & AO > 17 | -0.175 | 0.489 | 0.086 |
| rule67 | IDS > 10 & LCImax > 0.3276 | 0.140 | 0.500 | 0.070 |
| rule51 | LCImax > 0.26 & IDS > 9 | 0.122 | 0.487 | 0.059 |
| rule24 | IDS <= 16 & GAD %in% c("Negative") | -0.080 | 0.499 | 0.040 |
| rule110 | IDS > 10 & Age > 22 | 0.020 | 0.459 | 0.009 |
| rule125 | IDS <= 17 & AO > 13 | -0.015 | 0.496 | 0.007 |
| rule108 | IDS > 14 & pedigree %in% c("Yes") | 0.002 | 0.478 | 0.001 |

Table 1: Predicting chronic depression - Default settings.

# Variable importances

Variable selection: Unselected predictor variables obtain importance of 0.




| Term | Description | Coeffient | SD | Importance |
|---|---|---|---|---|
| (Intercept) | 1 | -0.221 | 0.000 | 0.000 |
| rule3 | IDS > 10 & LCImax > 0.2632 | 0.224 | 0.494 | 0.111 |
| rule27 | IDS > 13 & LCImax > 0.3621 | 0.213 | 0.477 | 0.102 |
| rule84 | IDS <= 16 & AO > 17 | -0.175 | 0.489 | 0.086 |
| rule67 | IDS > 10 & LCImax > 0.3276 | 0.140 | 0.500 | 0.070 |
| rule51 | LCImax > 0.26 & IDS > 9 | 0.122 | 0.487 | 0.059 |
| rule24 | IDS <= 16 & GAD %in% c("Negative") | -0.080 | 0.499 | 0.040 |
| rule110 | IDS > 10 & Age > 22 | 0.020 | 0.459 | 0.009 |
| rule125 | IDS <= 17 & AO > 13 | -0.015 | 0.496 | 0.007 |
| rule108 | IDS > 14 & pedigree %in% c("Yes") | 0.002 | 0.478 | 0.001 |

Table 1: Predicting chronic depression - Default settings.

# (Non-)negativity constraints

Identify only higher-risk (lower-risk) subgroups?

Can force estimated coefficients to be positive (or negative):

| Term | Description | Coeffient | SD | Importance |
|---|---|---|---|---|
| (Intercept) | 1 | -0.352 | 0.000 | 0.000 |
| rule34 | IDS > 13 & LCImax > 0.3621 | 0.228 | 0.477 | 0.109 |
| rule4 | IDS > 10 & LCImax > 0.2632 | 0.213 | 0.494 | 0.105 |
| rule82 | IDS > 10 & LCImax > 0.3276 | 0.155 | 0.500 | 0.077 |
| rule64 | LCImax > 0.26 & IDS > 9 | 0.107 | 0.487 | 0.052 |
| rule134 | IDS > 14 & pedigree %in% c("Yes") | 0.043 | 0.478 | 0.021 |
| rule136 | IDS > 10 & Age > 22 | 0.016 | 0.459 | 0.007 |
| rule8 | IDS > 10 & LCImax > 0.2778 | 0.001 | 0.497 | 0.001 |

Table 2: Predicting chronic depression - Non-negativity constrained solution.

# Confirmatory rule(s)

Rule or linear term known a-priori to be relevant?

Can be forced in by not applying penalty to coefficient of that term:

| Term | Description | Coeffient | SD | Importance |
|---|---|---|---|---|
| (Intercept) | 1 | -0.352 | 0.000 | 0.000 |
| ADuse == TRUE | ADuse == TRUE | 0.283 | 0.498 | 0.141 |
| rule3 | IDS > 10 & LCImax > 0.2632 | 0.230 | 0.494 | 0.114 |
| rule84 | IDS <= 16 & AO > 17 | -0.193 | 0.489 | 0.094 |
| rule27 | IDS > 13 & LCImax > 0.3621 | 0.190 | 0.477 | 0.091 |
| rule67 | IDS > 10 & LCImax > 0.3276 | 0.132 | 0.500 | 0.066 |
| rule51 | LCImax > 0.26 & IDS > 9 | 0.088 | 0.487 | 0.043 |
| rule24 | IDS <= 16 & GAD %in% c("Negative") | -0.054 | 0.499 | 0.027 |

Table 3: Predicting chronic depression - Confirmatory rule included.

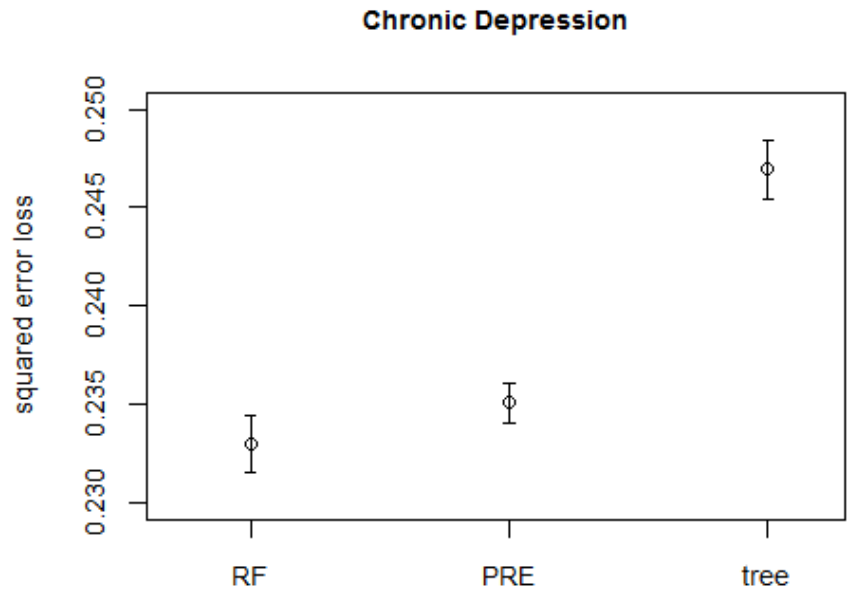
# Resolution

Does **pre** kill the bad?

Does the good survive?

# Resolution

Fokkema & Strobl (submitted):




Fokkema (in press):

Predictive accuracy:

random forest > pre > RuleFit > linear lasso > single tree

Complexity:

random forest > linear lasso > RuleFit > pre > single tree

# Discussion

But still some instability: Training data changes -> fitted PRE changes

Predictive accuracy similar, but different terms selected

Effron (2019): Prediction is easy, attribution is difficult

Future work:

- Better (i.e., more sparse, more stable) rule and variable selection:
  - Alternatives to lasso / glmnet
  - Modeling variability due to sampling

# Thank you!

m.fokkema@fsw.leidenuniv.nl

Fokkema, M. & Christoffersen, B. (2019). **pre**: Prediction Rule Ensembles. **R** package version 0.7.1. url: https://CRAN.R-project.org/package=pre

Fokkema, M. (in press). Fitting prediction rule ensembles with **R** package **pre**. *Journal of Statistical Software*. preprint: https://arxiv.org/abs/1707.07149

Fokkema, M. & Strobl, C. (under review). *Fitting prediction rule ensembles to psychological research data: An introduction and tutorial*. Working paper: https://arxiv.org/abs/1907.05302

# References


Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). Classification and regression trees. Wadsworth, New York.

Effron, B. (2019). *Prediction, estimation, and attribution*. Keynote at Conference in honor of Aad van der Vaart's 60th birthday, Leiden, The Netherlands. url: http://pub.math.leidenuniv.nl/~schmidthieberaj/publications/TalksAad/Efron.pdf

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics, 2*(3), 916-954.

Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*(3), 651-674.

Meinshausen, N. (2010). Node harvest. *The Annals of Applied Statistics*, 4(4), 2049-2072.

Penninx, B. W. J. H., Nolen, W. A., Lamers, F., Zitman, F. G., Smit, J. H., Spinhoven, P., . . ., Beekman, A. T. F. (2011). Two-year course of depressive and anxiety disorders: Results from the Netherlands Study of Depression and Anxiety (NESDA). *Journal of Affective Disorders, 133*(1), 76-85.