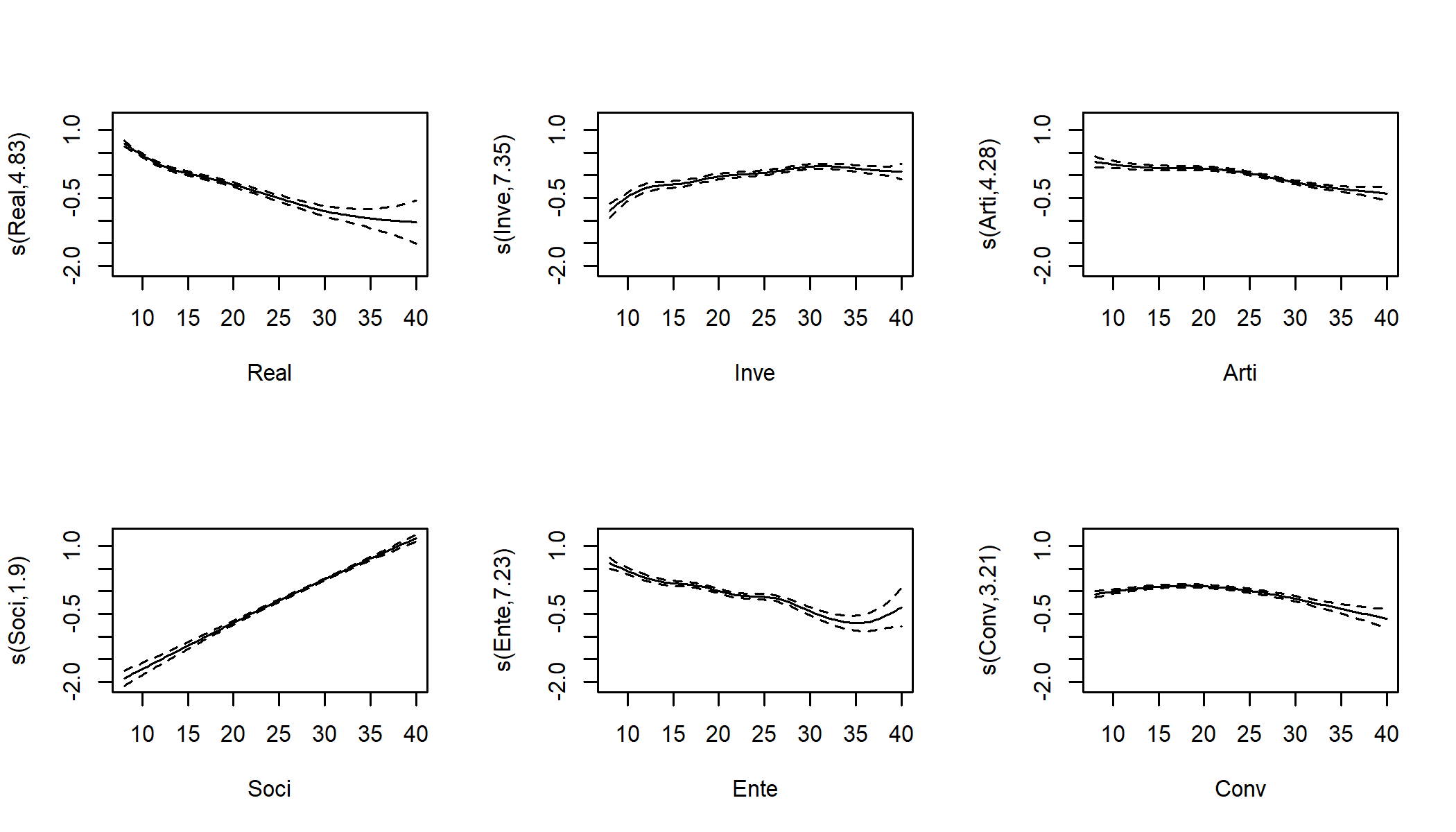
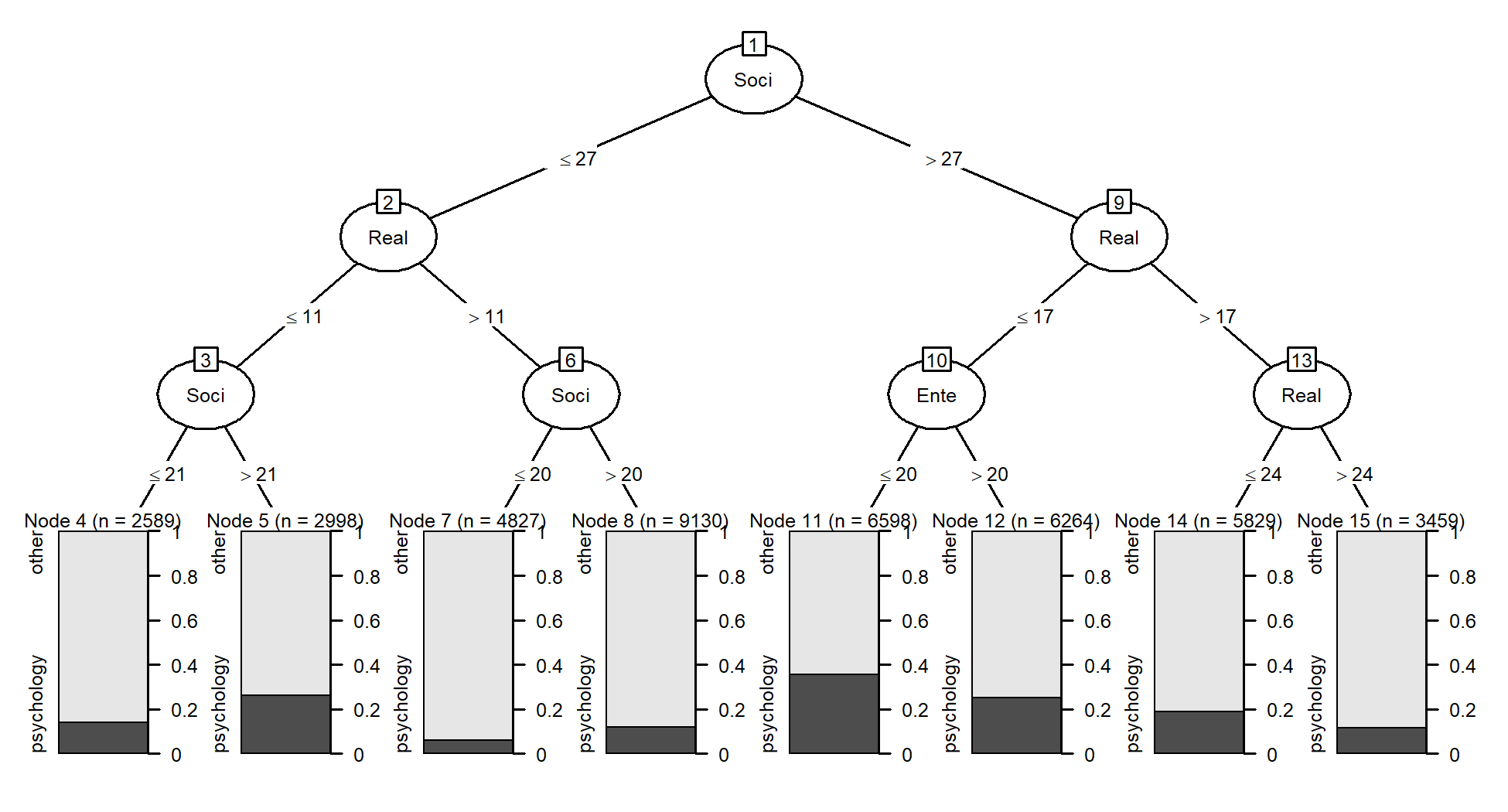
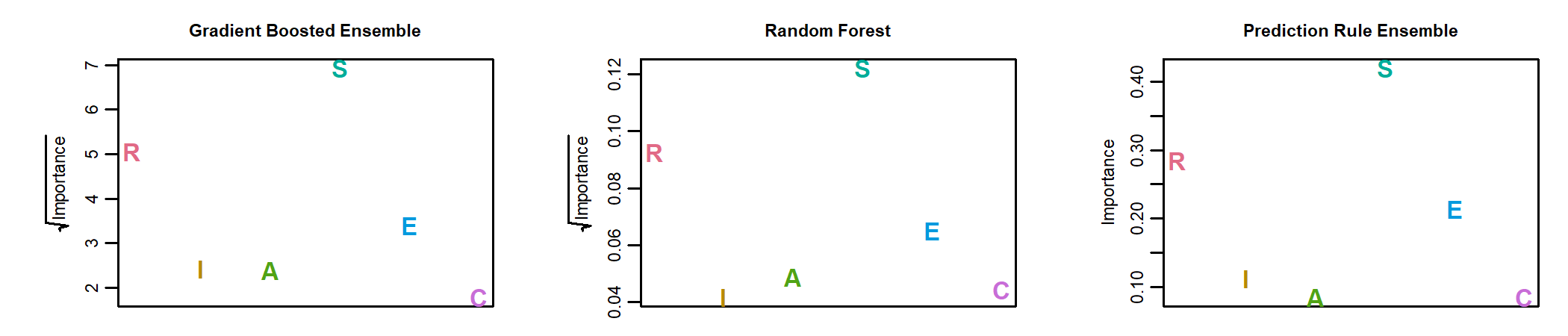
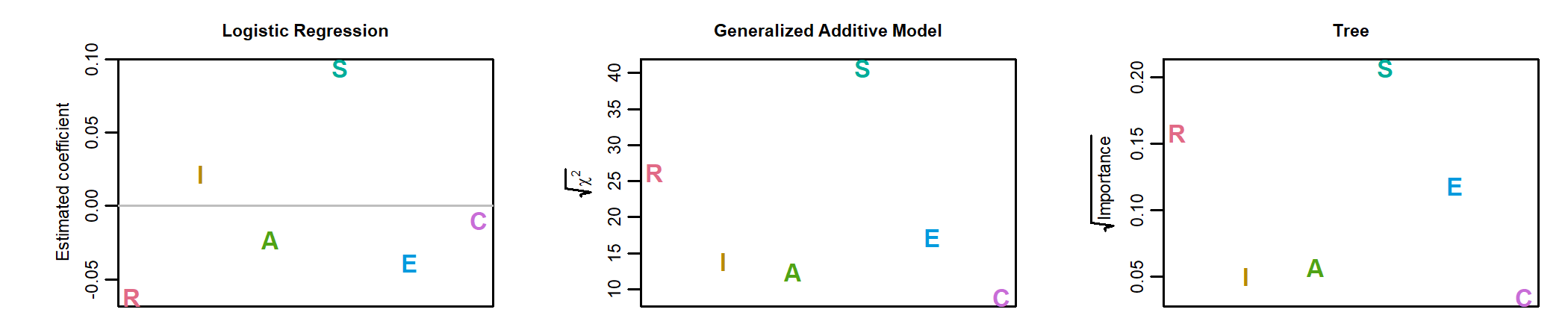
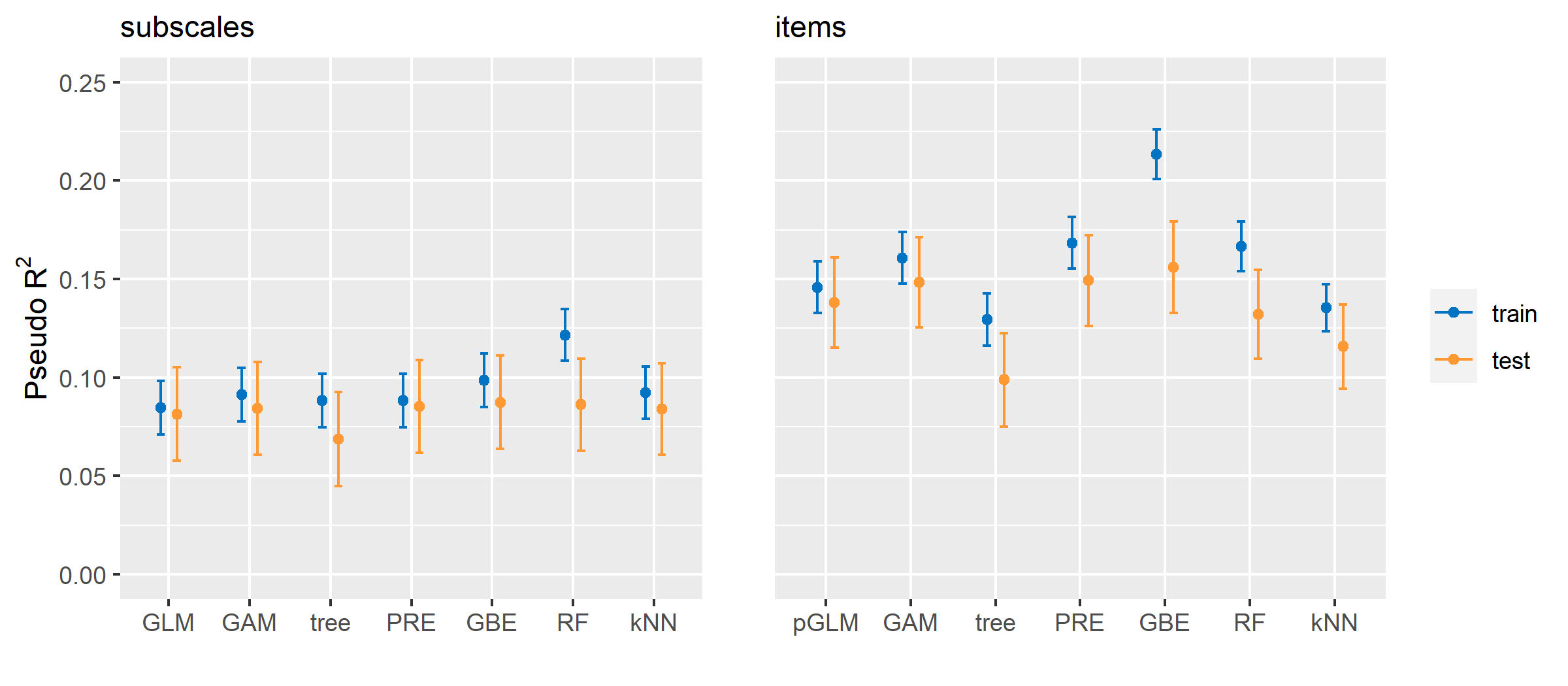
**Figure 1**  
*Fitted smoothing spline curves for each of the RIASEC subscales*  *Note.*  Values on the -axis reflect the effect on the log-odds of having completed a university major in psychology.

**Figure 2**  
*Conditional inference tree pruned to a depth of three* 

**Figure 3**  
*Variable contributions for each of the models fitted using RIASEC subscale scores as predictors*  *Note.* Coefficients in the logistic regression and importance measures of the prediction rule ensemble are on the scale of standard deviations. Importance measures for the other methods are on the scale of variances; for those methods, the square roots are plotted.

**Figure 4**  
*Predictive accuracy on train and test observations for each of the models fitted on subscale scores (left panel) and items scores (right panel)*  *Note.* (p)GLM = (penalized) logistic regression; GAM = generalized additive model with smoothing splines; PRE = prediction rule ensemble; GBE = gradient boosted tree ensemble; RF = random forest; kNN = k nearest neighbours.

**Table 1**  
*Six most important rules in the prediction rule ensemble*

|  |  |
| --- | --- |
| **Description** | **Coefficient** |
| Soci > 27 & Ente <= 31 & Conv <= 30 | 0.182 |
| Soci > 23 & Ente <= 29 & Real <= 24 | 0.181 |
| Real > 10 & Soci <= 35 | -0.175 |
| Real <= 22 & Soci > 19 & Inve > 18 | 0.138 |
| Inve > 10 & Real <= 13 | 0.120 |
| Conv <= 23 & Arti <= 29 & Soci > 21 | 0.112 |