Supplementary Materials

1 The data

Figure 1 depicts the distribution of missing data for all variables included in the final model based on the dataset after controlling for long-term residence and presence of information about psychotic experiences.

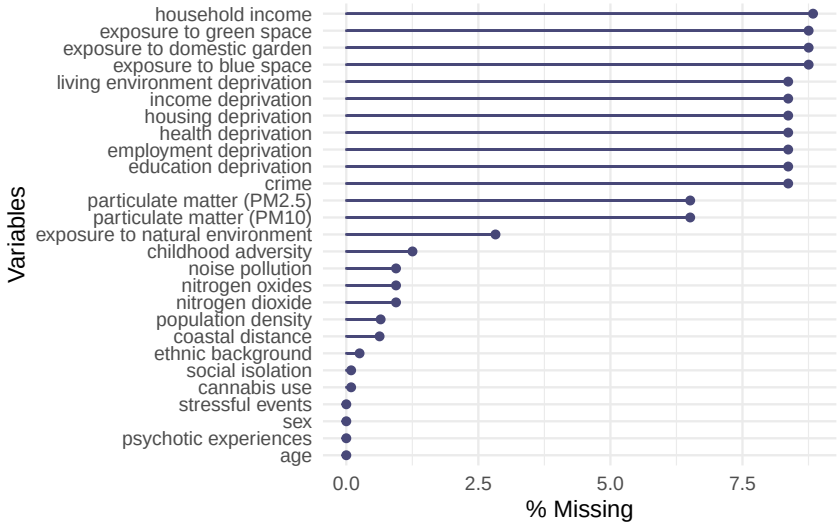
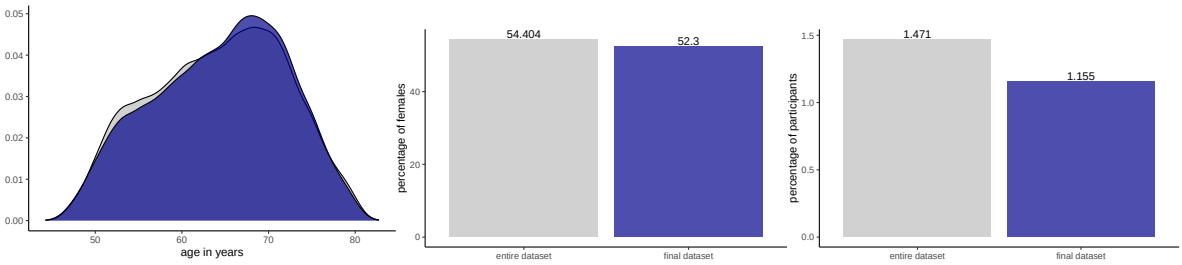


Figure 1: The distribution of missing data

To ensure generalisability of the results, a comparison was conducted between age, sex and psychotic experiences from the original dataset and the dataset that was used in the present analyses (see Figure 2).



(a) Age (b) Females (c) Psychotic experiences

Figure 2: Comparison of the original and final dataset. Data that are part of the final dataset are shown in blue.

2 Results

2.1 Predictors

We calculated Person correlations between the prediction variables and family-wise error corrected for multiple testing using the Holm-Bonferroni criterion (see Figure 3).

2.2 Feature selection

Variable importances, as selected by the models, are listed and colour-coded in Table 1.

The complete output of the Boruta algorithm is listed in Table 2.

The partial dependence plots for all variables selected by the Boruta algorithm are displayed in Figure 4.

2.3 Predictive power

Figure 8 plots the area under the curve against the number of selected variables to identify which model was the most parsimonious.

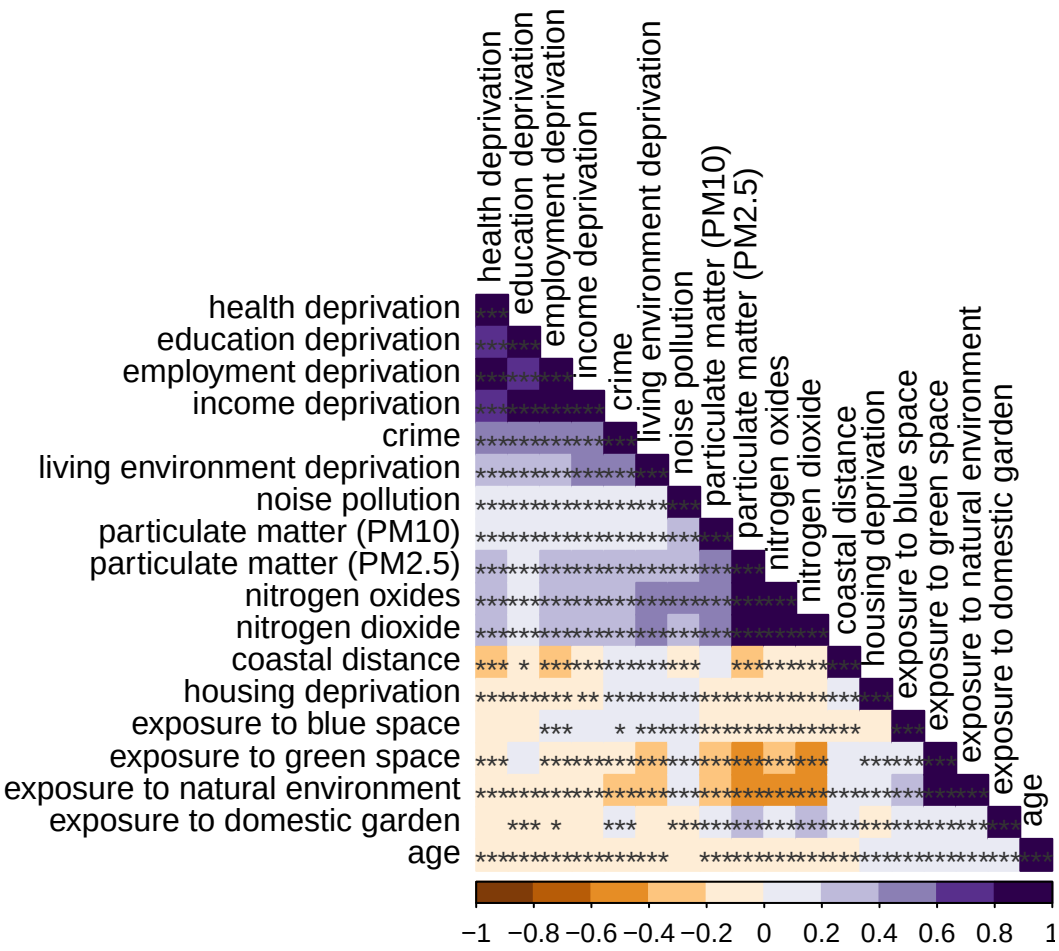


Figure 3: Correlation plot of all continous prediction variables. The p-values are Holm-Bonferroni corrected for multiple testing.

Table 1: Feature Importance

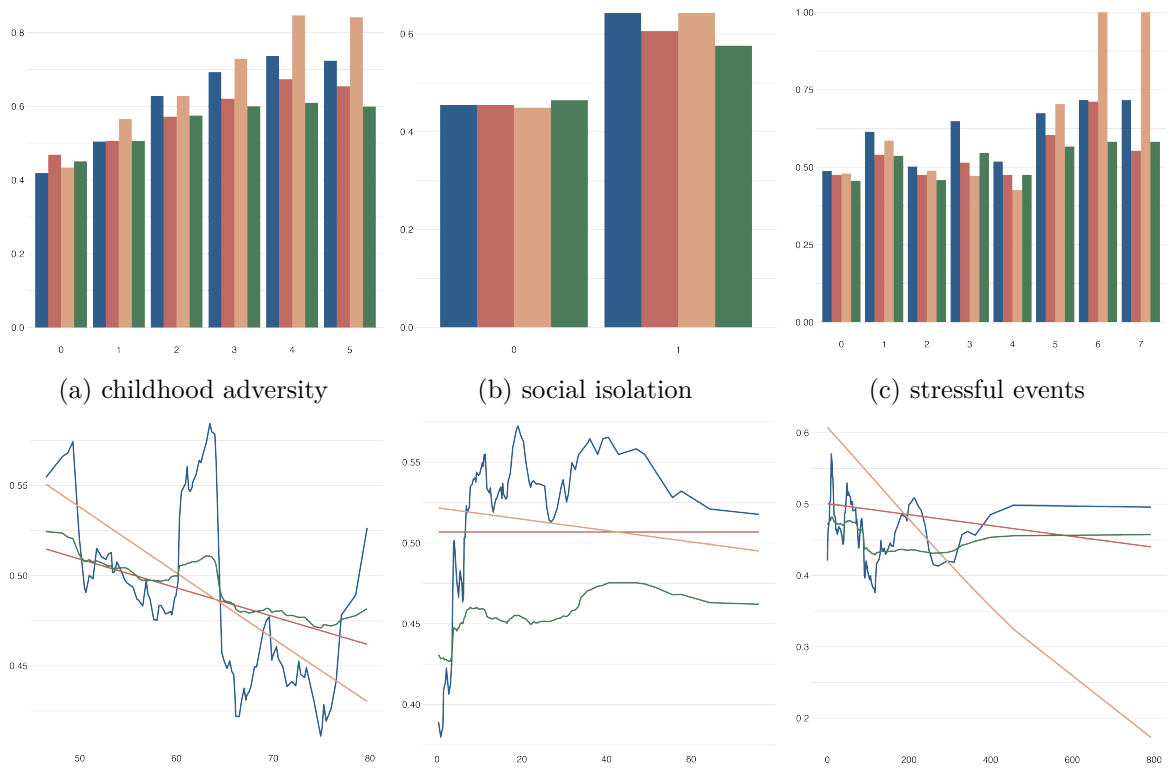
	Boruta	Elastic net regression	Random forest	Logistic regression
childhood adversity	2.34 (1.42) *	2.02 (1.3) *	0.95 (1.05)	1.57 (1.25)
stressful events	1 (1.31) *	2.63 (3.15) *	0.82 (0.84)	2.04 (3.01)
social isolation	1.68 (1.44) *	0.43 (0.12) *	0.43 (1.25)	0.34 (0.11)
household income	0.57 (0.9)	0.83 (0.48) *	0.69 (0.86)	0.64 (0.46)
age	0.98 (1.32) *	0.1 (0.07) *	1.29 (1.64)	1.13 (0.45)
exposure to natural environment	0.57 (0.68) *	0.08 (0.14) *	1.02 (0.59)	0.74 (0.43)
education deprivation	0.56 (0.66)	0.08 (0.06) *	1.11 (0.86)	0.65 (0.3)
crime	0.56 (0.95)	0.06 (0.03) *	1.06 (1.01)	0.61 (0.3)
living environment	0.6 (0.97) *	0.01 (0.01)	1.07 (1.1)	0.23 (0.27)
housing	0.29 (0.61)	0.03 (0.03) *	1.13 (1.07)	0.5 (0.28)
cannabis	0.04 (0.25)	0.24 (0.37) *	0.24 (0.38)	0.19 (0.35)
coastal distance	0.19 (0.48)	0.05 (0.05) *	1.07 (0.59)	0.69 (0.4)
exposure to blue space	0.1 (0.47)	0.05 (0.07) *	1.04 (0.47)	1.19 (0.36)

Note:
This table lists the features' mean variable importance, along with their standard deviations, across the cross-validation (CV). Only features selected by either of the selective algorithms in more than half of the CV folds are included. The variables are ranked based on their mean predictive importance in the two feature selection algorithms. To facilitate comparability between the models, the feature importance was scaled. Columns 2 to 5 are colour-coded to indicate the mean variable importance, ranging from low (light yellow) to high (purple). In the first column, the colours reflect the mean feature importance across the four models. Note that for logistic regressions, categorical variables' levels have been summarised into a single feature to ensure comparability.

* Feature was selected in more than half of the CV folds.

Table 2: Variable importance in the Boruta algorithm

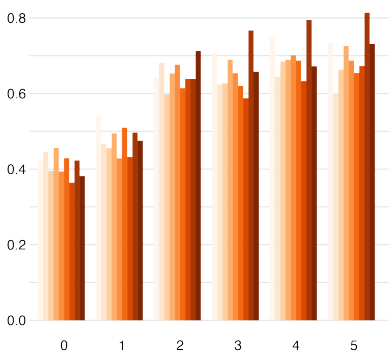
Features	meanImp	medianImp	minImp	maxImp	normHits	confirmed	rejected	sd	decision
childhood adversity	10.75	10.76	6.20	15.29	1.00	10	0	2.95	Confirmed
social isolation	7.72	7.71	4.08	11.70	0.90	9	1	3.00	Confirmed
stressful events	4.60	4.60	1.01	8.09	0.73	9	1	2.73	Confirmed
age	4.53	4.54	0.76	8.50	0.74	9	1	2.75	Confirmed
living environment	2.76	2.70	-0.67	6.21	0.51	7	3	2.01	Confirmed
household income	2.62	2.65	-0.17	5.45	0.38	5	5	1.87	Rejected
exposure to natural environment	2.60	2.60	-0.34	5.76	0.41	6	4	1.41	Confirmed
education deprivation	2.59	2.54	-0.42	6.02	0.39	5	5	1.36	Rejected
crime	2.58	2.58	-0.38	5.58	0.41	5	5	1.98	Rejected
nitrogen oxides	2.14	2.14	-0.85	5.18	0.33	4	6	1.35	Rejected
particulate matter (PM2.5)	2.11	2.15	-0.55	4.67	0.29	3	7	1.39	Rejected
nitrogen dioxide	2.02	2.01	-0.43	4.66	0.32	4	6	1.48	Rejected
income	1.98	1.95	-0.71	4.92	0.26	3	7	1.25	Rejected
employment	1.87	1.79	-0.37	4.45	0.26	3	7	1.13	Rejected
exposure to green space	1.39	1.42	-0.93	3.41	0.08	1	9	1.03	Rejected
housing	1.35	1.33	-1.25	4.21	0.17	1	9	1.28	Rejected
coastal distance	0.86	0.90	-1.59	2.95	0.10	0	10	0.99	Rejected
health deprivation	0.76	0.74	-1.23	2.96	0.08	1	9	1.27	Rejected
exposure to blue space	0.45	0.51	-1.65	2.30	0.05	1	9	0.97	Rejected
particulate matter (PM10)	0.38	0.37	-1.62	2.23	0.06	1	9	0.94	Rejected
cannabis	0.17	0.19	-1.53	1.97	0.00	0	10	0.52	Rejected
population density	-0.04	-0.10	-1.68	1.85	0.00	0	10	0.47	Rejected
ethnic background	-0.04	0.01	-1.68	1.48	0.00	0	10	0.54	Rejected
sex	-0.18	-0.12	-1.75	1.42	0.00	0	10	0.43	Rejected
noise pollution	-0.20	-0.14	-1.99	1.63	0.00	0	10	0.44	Rejected
exposure to domestic garden	-0.44	-0.50	-1.95	1.36	0.00	0	10	0.60	Rejected



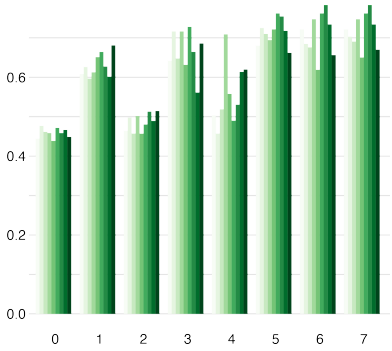
(a) childhood adversity (b) social isolation (c) stressful events

(d) age (e) living environment (f) exposure to natural environment

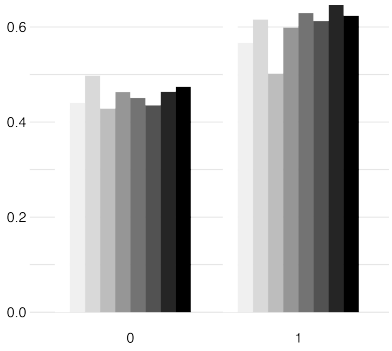
Figure 4: Partial dependence plots for the variable selected by the Boruta algorithm (blue) for an exemplary cross-validation fold. To ensure comparability, the predictors were not standardised for the logistic regression models. The y-axis gives the predicted probability of having psychotic experiences. The elastic net model is depicted in red, the random forests in green and the unpenalised logisitc regression in orange. Outliers in continous variables, i.e., deviations of 2 standard deviations from the mean, are not depicted.



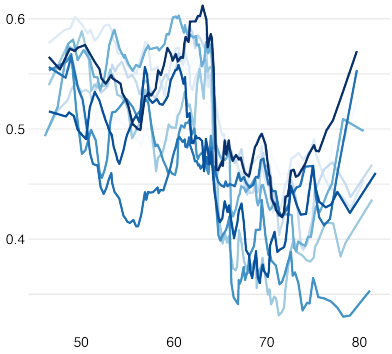
(a) childhood adversity



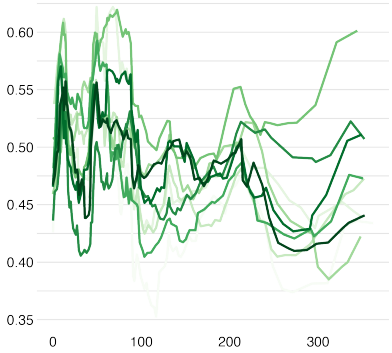
(b) stressful events



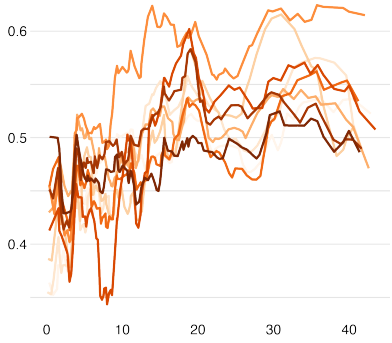
(c) social isolation



(d) age

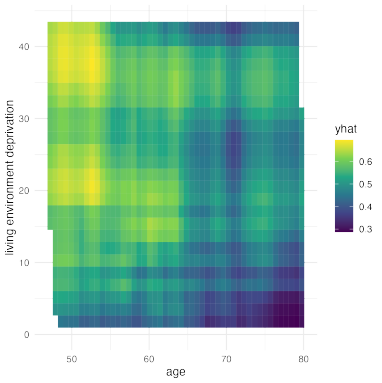


(e) exposure to natural environment

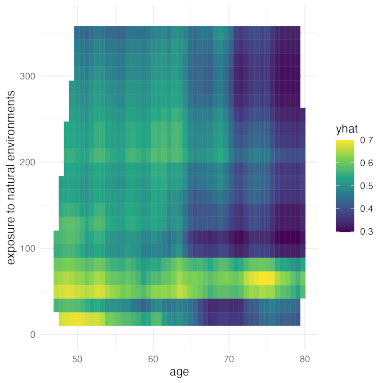


(f) living environment deprivation

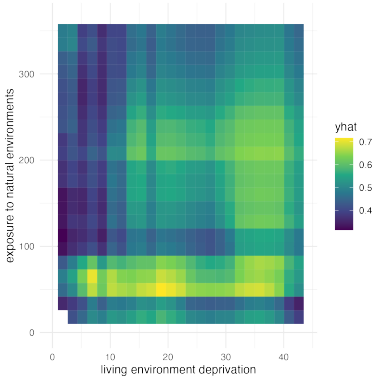
Figure 5: Partial dependence plots for all folds of the cross-validation based on the Boruta models.



(a) living environment deprivation x age

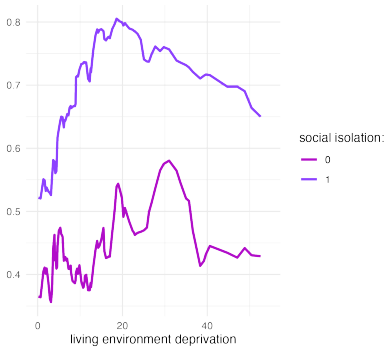


(b) exposure to natural environments x age

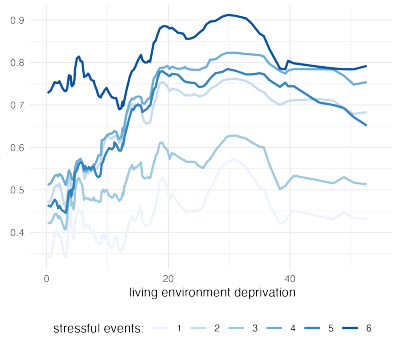


(c) exposure to natural environments x living environment deprivation

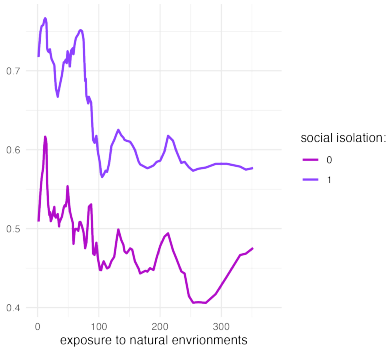
Figure 6: Two-dimensional partial dependence plots for an exemplary cross-validation fold.



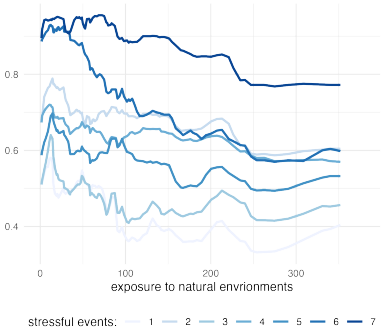
(a) exposure to natural environments x social isolation



(b) exposure to natural environments x stress



(c) living environment deprivation x social isolation



(d) living environment deprivation x stress

Figure 7: Grouped partial-dependence profiles for exposure to natural environements and living environment deprivation grouped by social isolation and the experience of stressful events for an exemplary cross-validation fold.

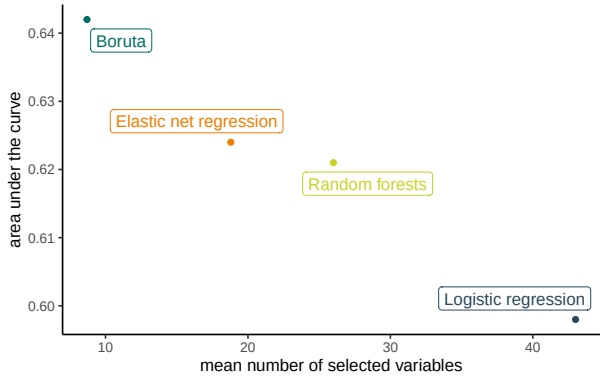


Figure 8: This figure shows the area under the curve against the number of variables to illustrate the parismony of the different trained models.