


# Data Analytics for Social Science
# Trees and forests

Johan A. Elkink

School of Politics & International Relations
University College Dublin

5 March 2020




# Outline

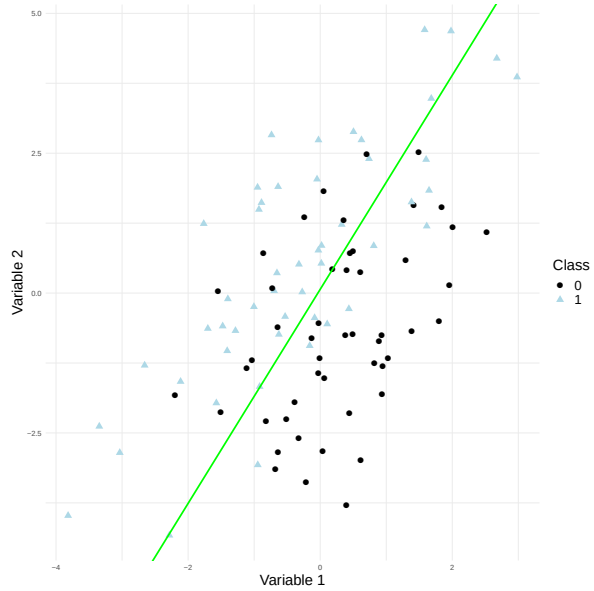



# Logistic regression

Logistic or probit regression are suitable for predicting a binary outcome variable and evaluating the relationship between an explanatory variable and a binary dependent variable. Linear discriminant analysis performs a similar function, although it is only common in a data science, i.e. predictive, context.

Common extensions allow for predicting nominal or ordinal variables.

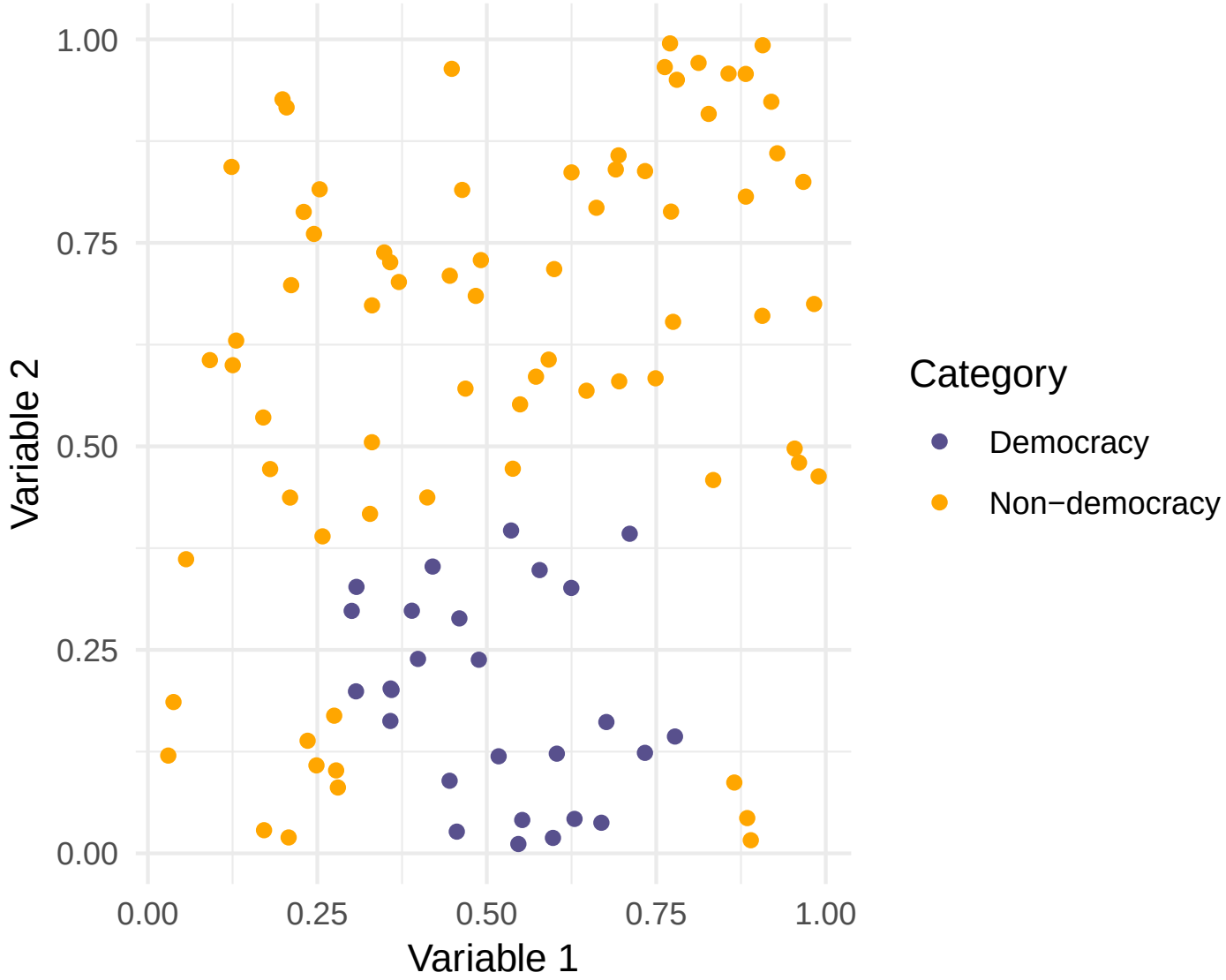
These models have in common that the starting point is a linear combination of explanatory variables, $\mathbf{X}\boldsymbol{\beta}$ or $y_i = f(\beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + ...)$, although transformations like $x_i^* = x_i^2$ can be used.






# Non-linear boundaries

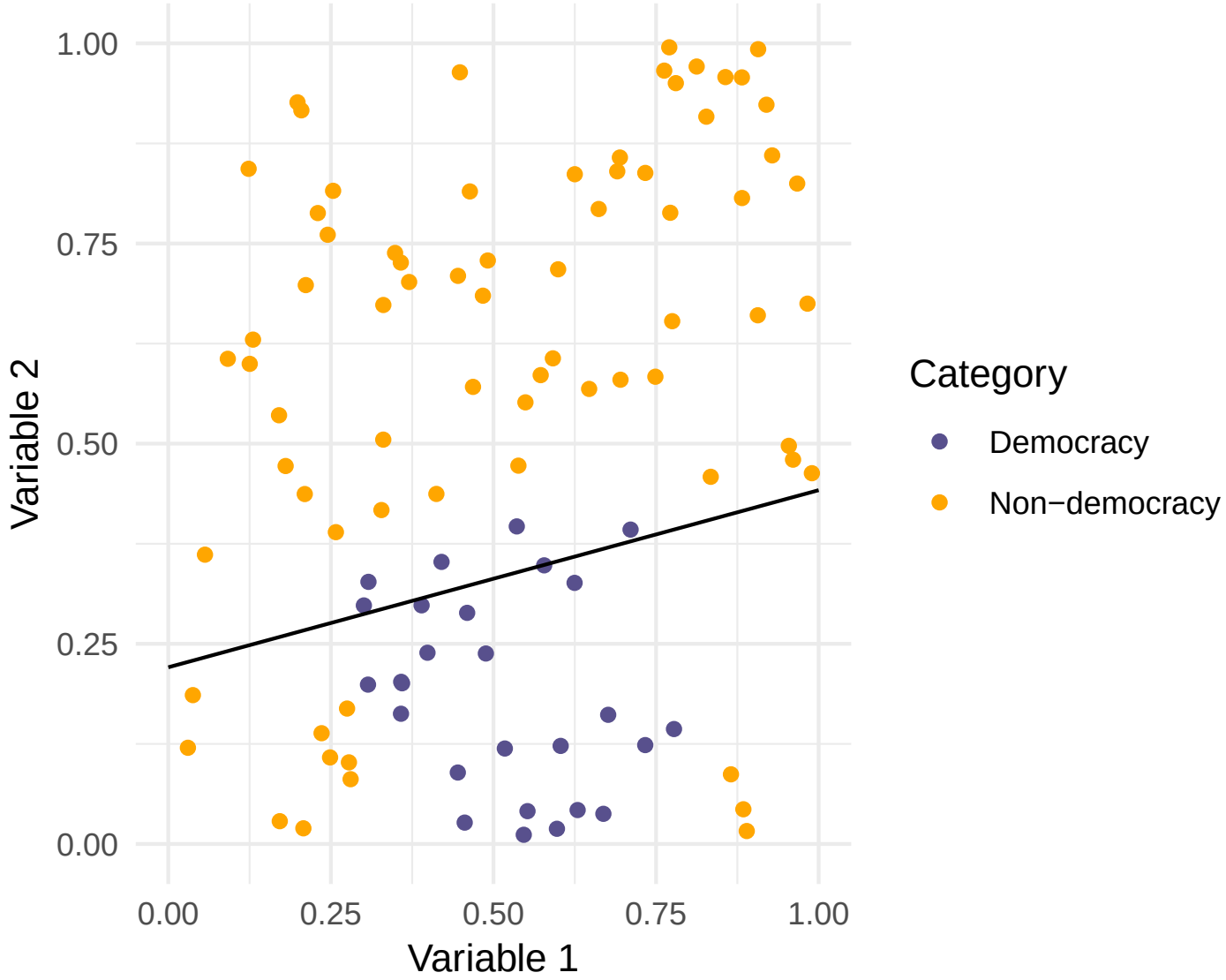
It might well be the case that the categories are not easily separable in a linear manner, however.






# Non-linear boundaries

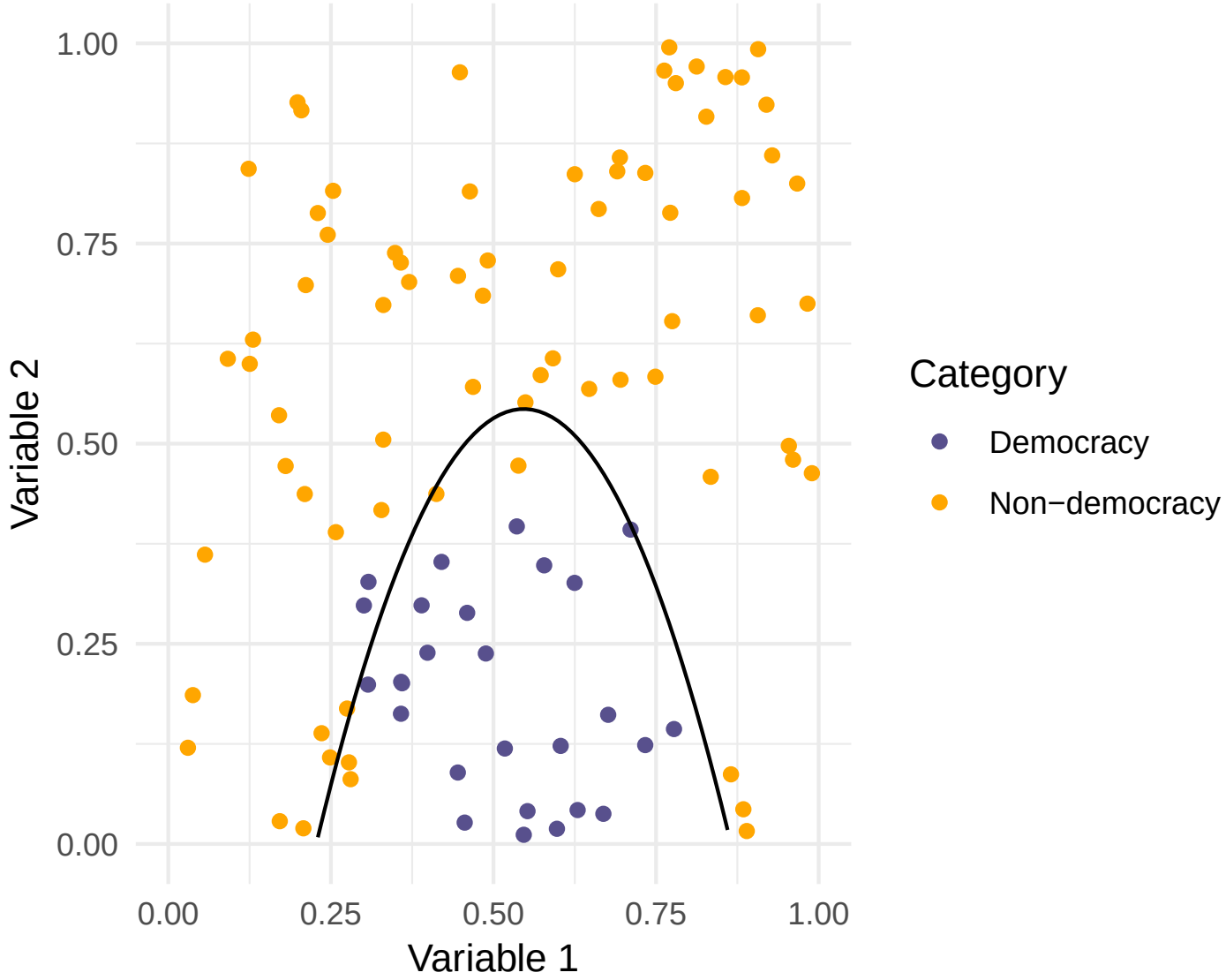
It might well be the case that the categories are not easily separable in a linear manner, however.





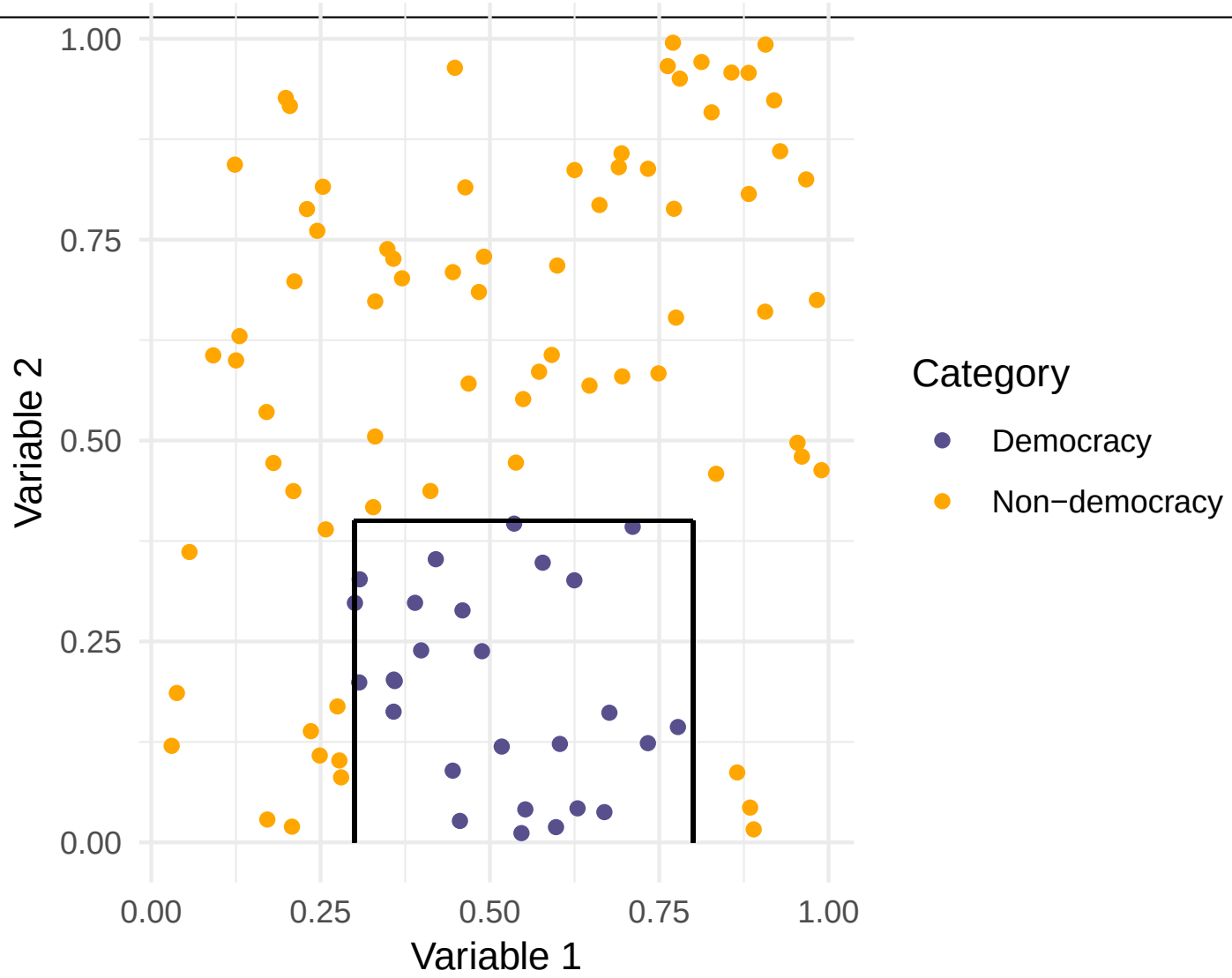

# Non-linear boundaries

It might well be the case that the categories are not easily separable in a linear manner, however.





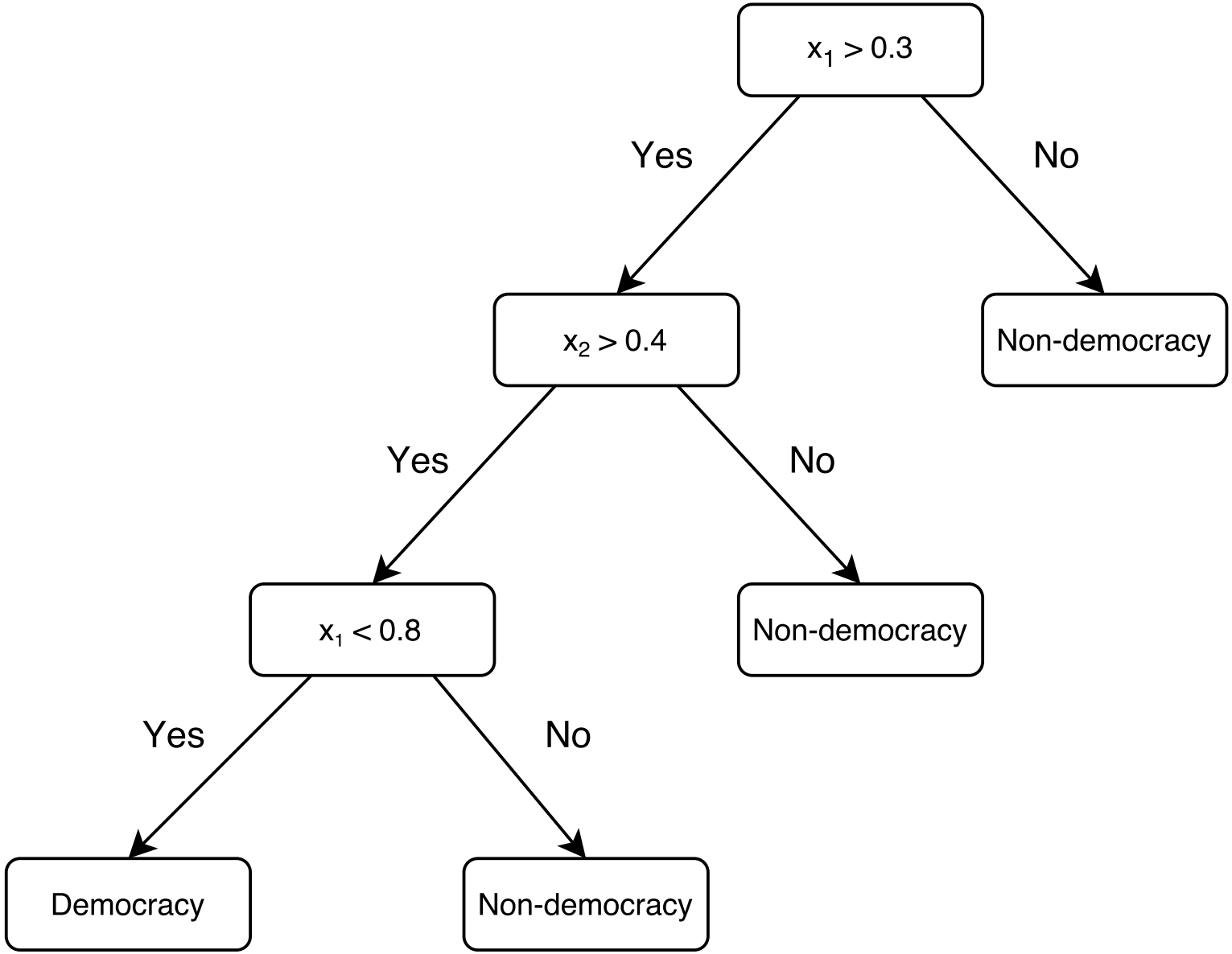

# Non-linear boundaries




In this example, the classification can be very precise, however: $y = 1$ if $x_1 > 0.3$ and $x_1 < 0.8$ and $x_2 < 0.4$, else $y = 0$.

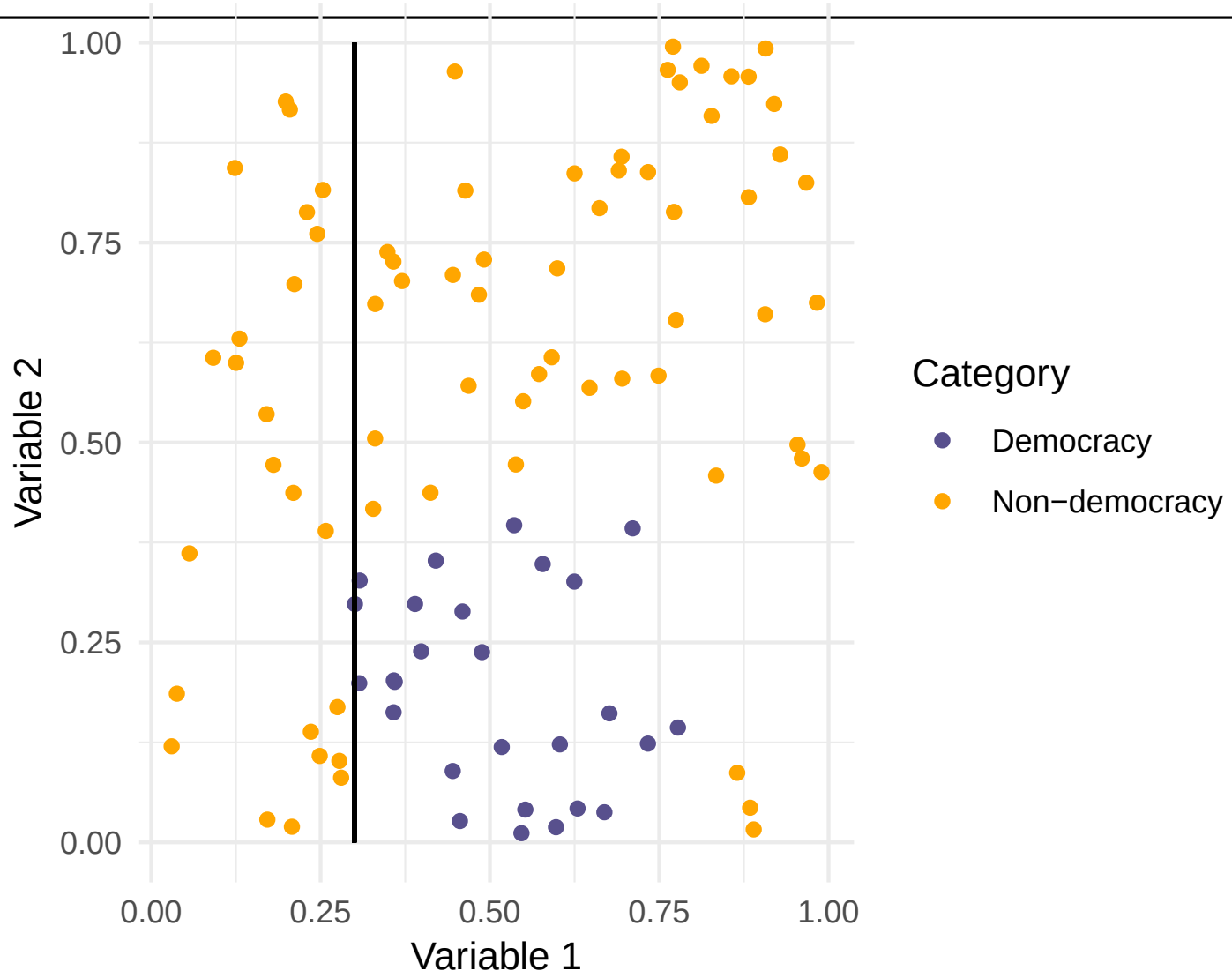



# Decision-tree





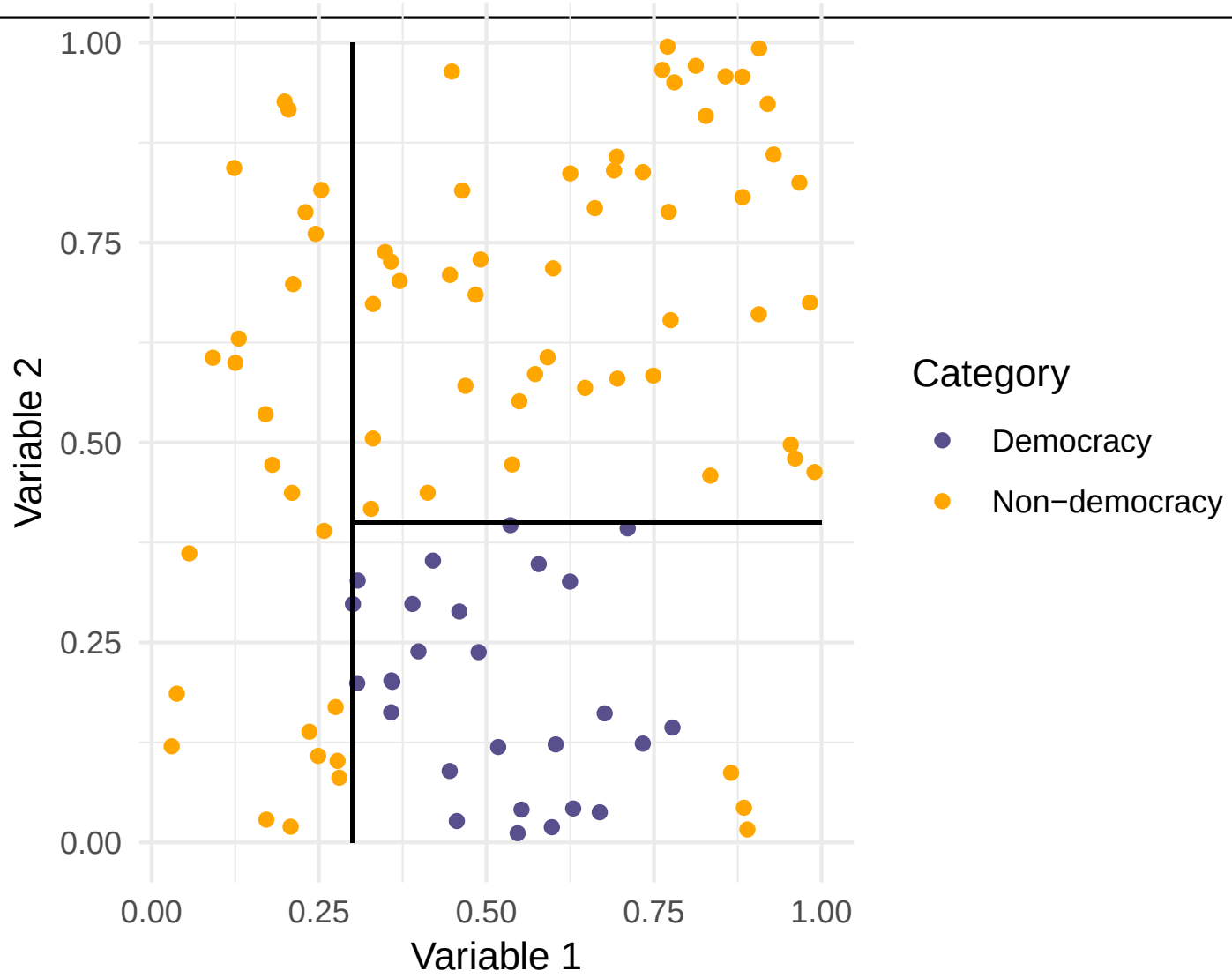

# Decision-tree




A tree classification model does exactly this—it keeps splitting segments in two.



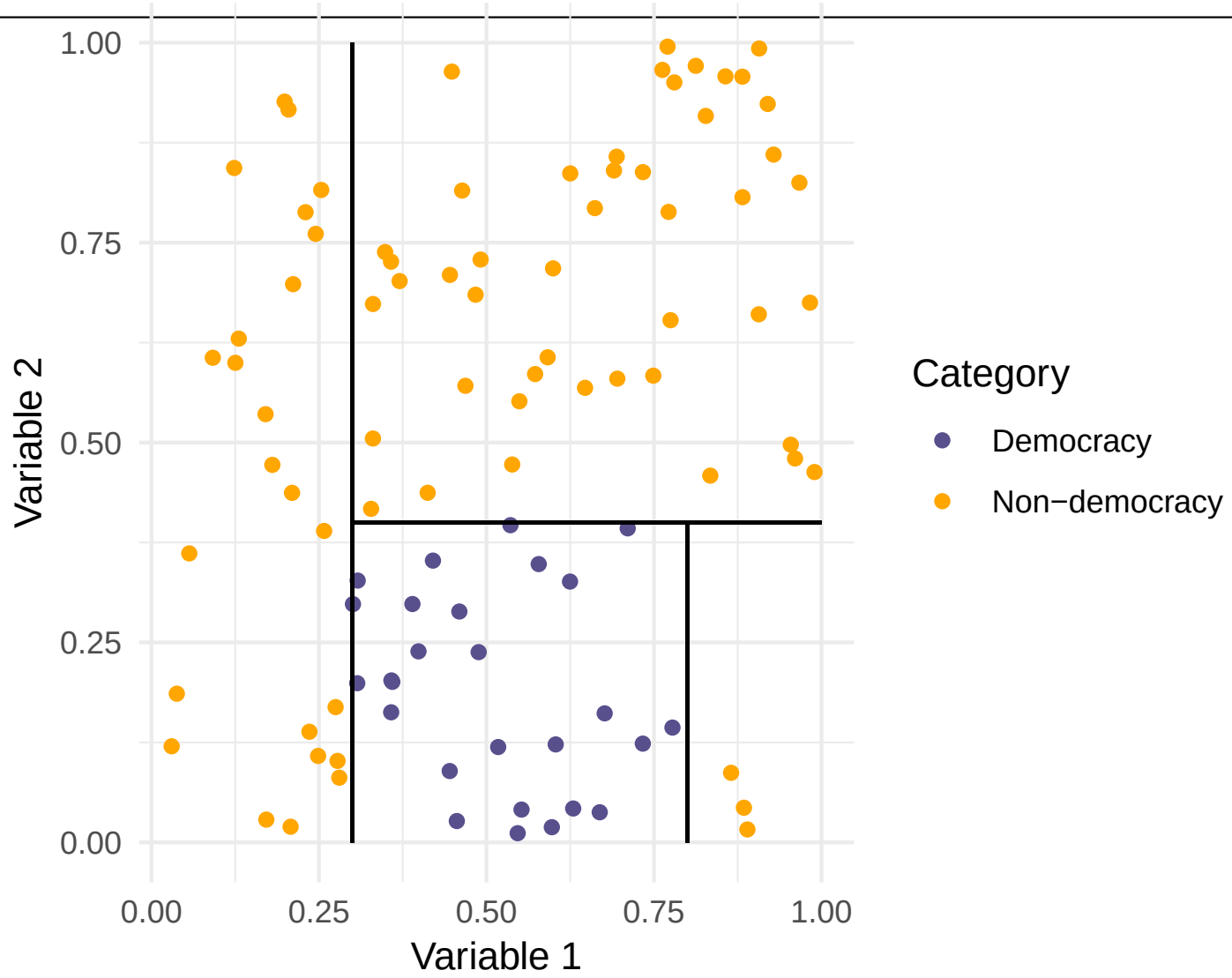

# Decision-tree




A tree classification model does exactly this—it keeps splitting segments in two.




# Decision-tree



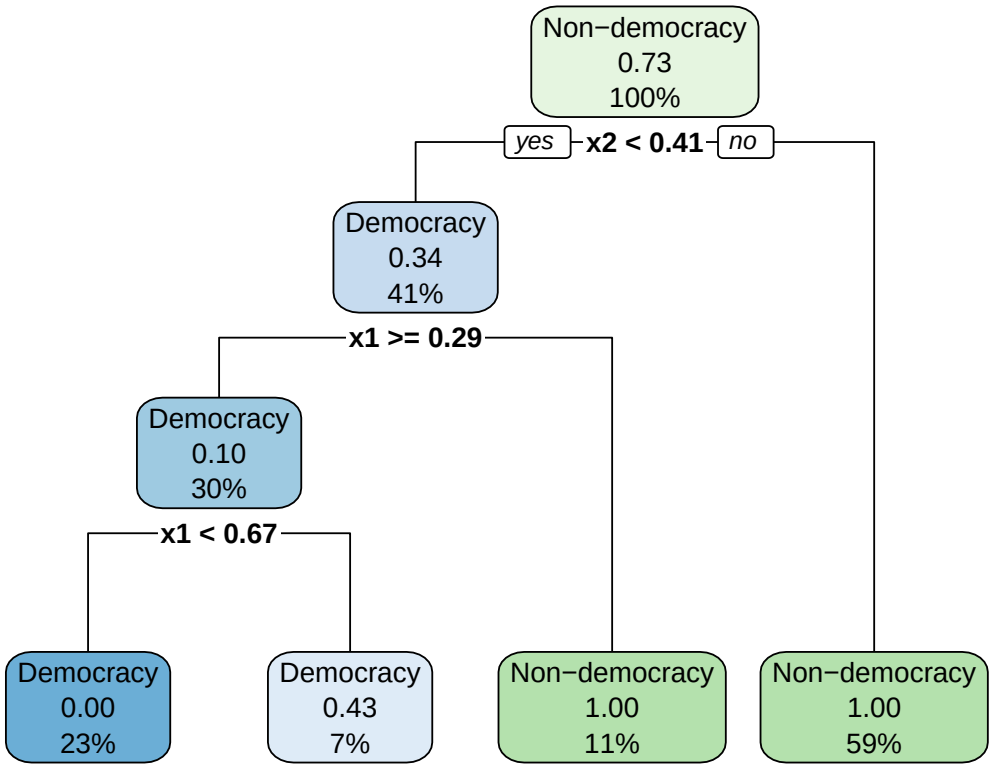

A tree classification model does exactly this—it keeps splitting segments in two.




# Classification trees

```
library(rpart)
library(plot.rpart)
t <- rpart(y ~ x1 + x2)
rpart.plot(t)
table(predict(t)[, 1], y)
```

| Pred | Dem | Non-dem |
|---|---|---|
| 0 | 0 | 70 |
| 0.57 | 4 | 3 |
| 1 | 23 | 0 |

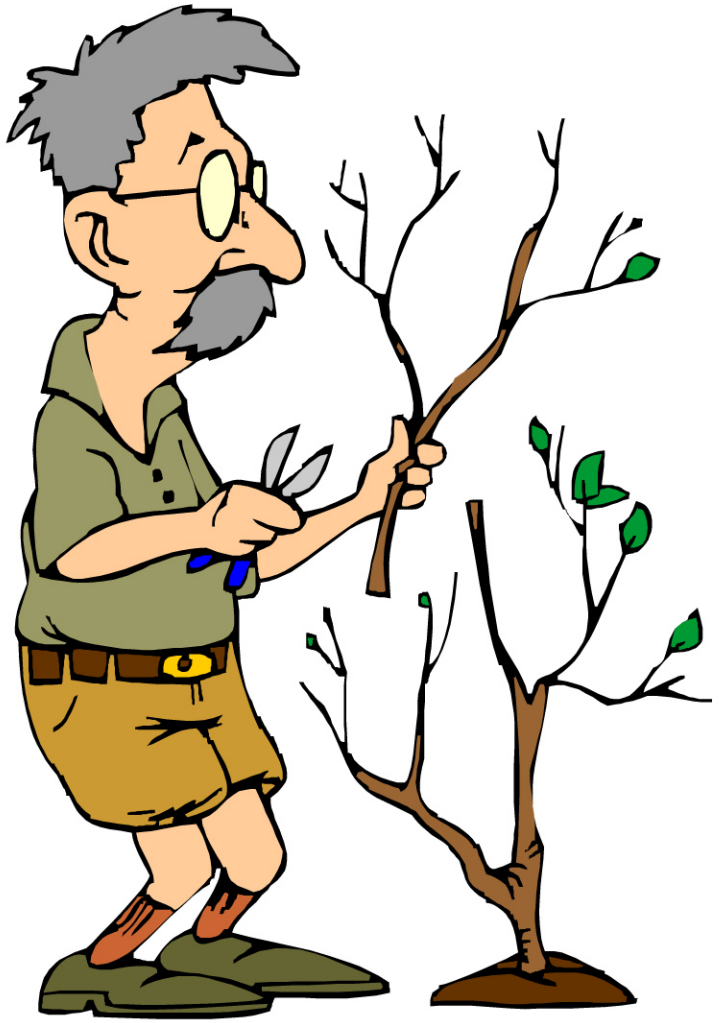





# Pruning

A tree with too many branches will be too much fine-tuned to a specific data set—this would be **overfitting**.

We therefore prune the tree by removing irrelevant subdivisions.



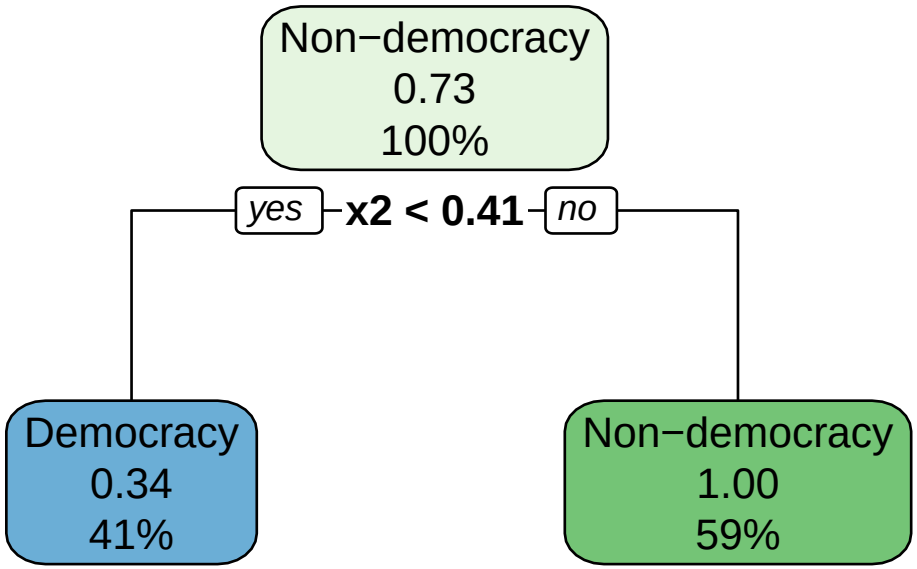
(James et al., 2013, 307–311)




# Classification trees

```
t <- prune(t, cp = .45)
rpart.plot(t)
table(predict(t)[, 1], y)
```

| Pred | Dem | Non-dem |
|---|---|---|
| 0 | 0 | 59 |
| 0.66 | 27 | 14 |






# Pros and cons of trees

- Easy to interpret and explain
- Aligns with how we think about decisions
- Easy to represent graphically
- Can directly handle categorical variables
- Can be used to predict either categorical variables (classification) or scale variables (regression)

But:

- Not generally as accurate in predictions as some other methods
- Not robust to small changes in the data

(James et al., 2013, 315–316)




# Outline




# Ensembles

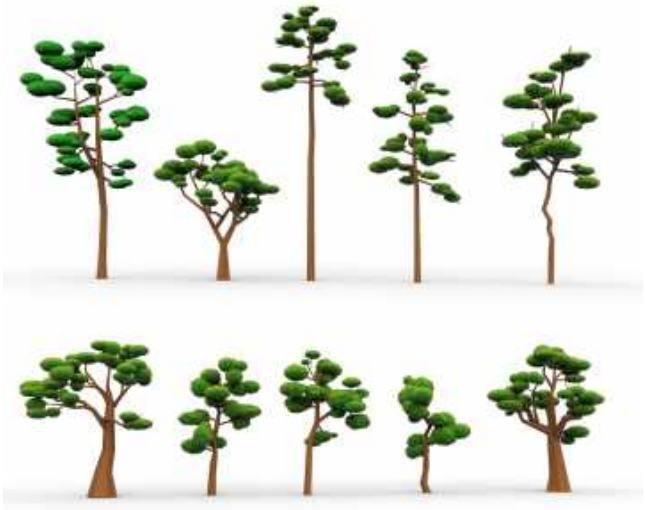
Trees—and many other methods—can be sensitive to specific outlying observations or pecularities of a particular data set.

If the model is too closely tuned on a specific data set (**overfitting**), prediction using new data worsens.

One solution is to generate many estimates based on small variations of the data set, and then take the "average" result.

This is known as **ensemble learning** and has close connections to the idea of the **wisdom of the crowd**.

(Hastie, Tibshirani and Friedman, 2009, 286)






# Bootstrapping

A standard technique for creating variations of the data set at hand is the **bootstrap**.

1. Take a random sample of the original data, of the same size, sampled *with replacement*.
2. Estimate the model on the new sample.
3. Repeat $m$ times.
4. Take the average of the estimates across $m$ estimates.

(This technique is also used to acquire standard errors in inferential statistics.)

| Data | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample 1 | 5 | 7 | 7 | 12 | 6 | 12 | 10 | 7 | 7 | 11 | 4 | 3 |
| Sample 2 | 5 | 7 | 3 | 5 | 7 | 3 | 10 | 8 | 9 | 6 | 7 | 4 |
| Sample 3 | 6 | 3 | 8 | 12 | 6 | 10 | 1 | 3 | 8 | 2 | 5 | 3 |
| Sample 4 | 11 | 3 | 12 | 1 | 2 | 3 | 9 | 1 | 7 | 9 | 4 | 5 |
| ... etc ... | | | | | | | | | | | | |




# Ensembles of trees

Three common techniques for combining trees:

**Bagging**: Take bootstrapped samples, estimate deep (not pruned) trees, and average across—for example, classify each observation by majority vote across estimates.

**Random forests**: Take bootstrapped samples, for each sample, randomly select a subset of variables, estimate deep (not pruned) trees, and average across.

**Boosting**: Estimate a small tree (few branches), take the residuals from the fitted tree, estimate a model on the residuals, add to the original model, and repeat. This slowly grows the tree.




# Forests

Ensemble estimators generate more accurate predictions, but are not so simple to interpret—no tree can be drawn.

```
library(randomForest)
t <- randomForest(y ~ x1 + x2)
table(predict(t), y)
importance(t)
```

| Pred | Dem | Non-dem |
|---|---|---|
| 0 | 0 | 73 |
| 1 | 27 | 0 |

| variable | mean decrease in Gini |
|---|---|
| $x_1$ | 17.39 |
| $x_2$ | 21.73 |

While you cannot draw the tree, you can get estimates for the **importance** of each of the independent variables in predicting the outcome.




# Bagging and boosting

**Bagging** is performed by setting the option `mtry` (the number of variables to sample each iteration) to the total number of variables:

```
t <- randomForest(y ~ x1 + x2, mtry = 2)
```

**Boosting** is a bit more involved and can be done using the `gbm` library (see James et al., 2013, 330–331).

```
library(gbm)
t <- gbm(I(as.integer(y) - 1) ~ x1 + x2,
                    distribution = "bernoulli")
table(predict(t, n.trees = 50,
                    type = "response"), y)
summary(t)
```




# Outline




# Project 2: assignment

Using the data used in Labs 4–6:

1. Make sure you include the entire first section of the lab—best to take Lab 6—including opening and merging the data, and computing the additional variables;
2. Take *democracy* as the dependent variable, which is a binary (dichotomous) variable based on the Polity IV data set (Marshall and Jaggers, 2002);
3. Focus on the effect of development aid on democracy;
4. Select some key control variables from the data, whereby it is acceptable to select the same set as in the lab regressions, but you do have to explain why these are good choices as controls;
5. Include at least one regression model and one tree-based model and perhaps some graphical descriptions of the data or key relationships.




# Project 2: essay structure

Write a 2,000-2,500 word essay (excluding captions, tables and bibliography from the count), including the graphs and tables.

Each essay should consist of

1. a short introduction,
2. a description and motivation of the data and methods used (approximately 25% of the essay),
3. the analysis including necessary graphs and tables (approximately 35%), and
4. an interpretation and conclusion (approximately 40%).

Everything needs to be properly referenced.




# Project 2: practicalities

## Literature

A good starting point for references is probably to search for "`foreign aid and democracy promotion`" in Google Scholar (`https://scholar.google.com/`).

Note that you significantly strengthen your essay by referencing two or three (at least) academic references that discuss the relationship between foreign development aid and democracy, typically in the section where you describe your model (before the analysis), but potentially also where you interpret the output.

## Markdown

Best is to make a separate Markdown file for the analysis required for the assignments, which can be copy/pasted and then edited from the lab sheets. You can potentially open the HTML file in your browser and save as, or print to, PDF.




# Project 2: submission

All assignments should be submitted electronically to `jos.elkink@ucd.ie`, consisting of either:

- a PDF file containing the essay and an R- or Rmd-file with all the commands necessary for the analysis, or
- a PDF file generated using RMarkdown with the essay and all necessary commands and results.

Recommendation: the deadline is **6 April, 1 pm**, but you should have everything now to do the assignment, so get this out of the way before other modules give assignments.




Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edition ed. New York: Springer.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With applications in R*. Springer.

Marshall, M.G. and K. Jaggers. 2002. "Polity IV project: political regime characteristics and transitions, 1800-2002.".
**URL:** *http://www.bsos.umd.edu/cidcm/polity/*