Lectorial 7
Decision Trees

Machine Learning @ RMIT

RMIT UNIVERSITY

- Decision Trees are Computationally intensive methods.
- Used in situations where we have a lot of explanatory variables (descriptive features)
- The number of descriptive features are so large that we simply cannot test them all in our model.
- Tree models (decision trees, random forests) are extremely good to use in computations that in the past would have required multivariate analysis techniques.

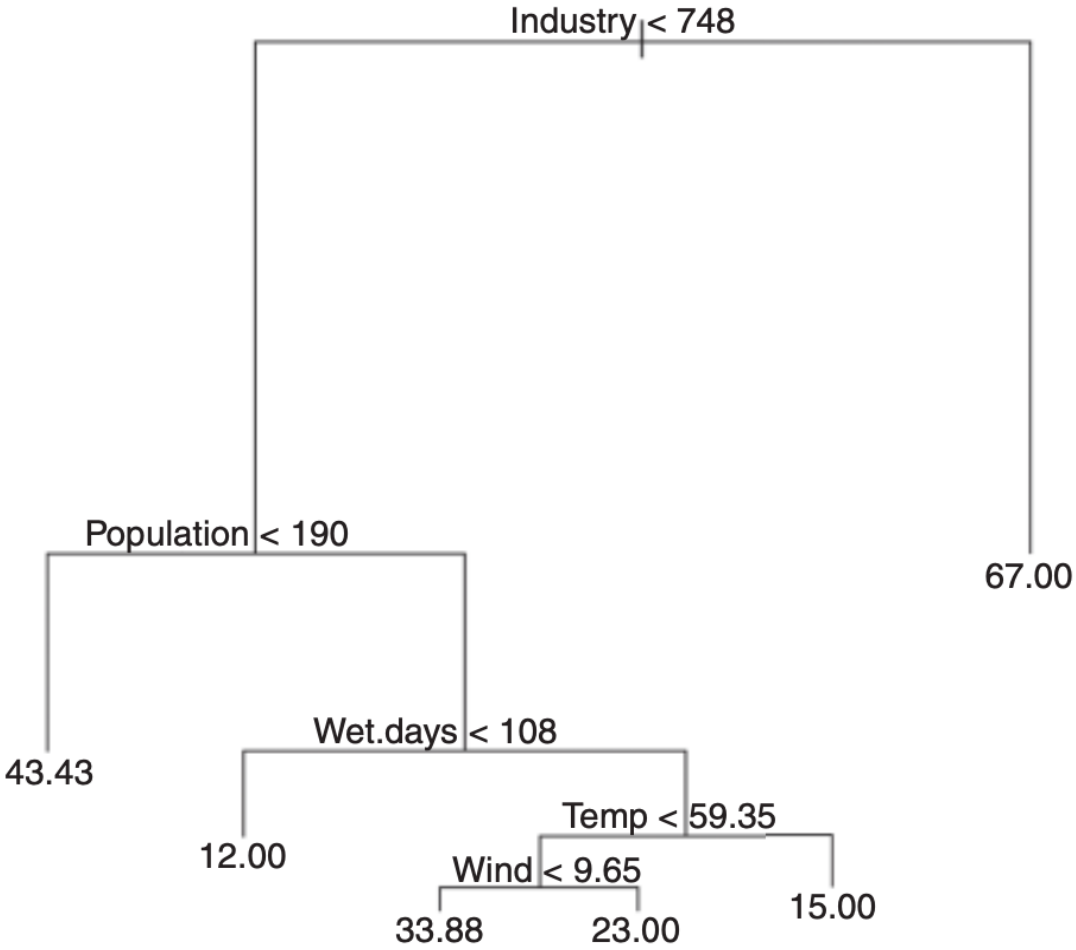
Advantages of Tree Models:

- Simple models (if you want the mathematics behind it , read Introduction to Statistical Learning by Gareth James, et al.)
- Excellent for initial data inspection
- Provides a very clear picture of the underlying structure of the data
- Provides extremely intuitive insight into the kinds of interactions between the feature variables

Example:

	Pollution	Temp	Industry	Population	Wind	Rain	Wet.days
1	24	61.5	368	497	9.1	48.34	115
2	30	55.6	291	593	8.3	43.11	123
3	56	55.9	775	622	9.5	35.89	105
4	28	51	137	176	8.7	15.17	89
5	14	68.4	136	529	8.8	54.47	116
6	46	47.6	44	116	8.8	33.36	135
7	9	66.2	641	844	10.9	35.94	78
8	35	49.9	1064	1513	10.1	30.96	129
9	26	57.8	197	299	7.6	42.59	115
10	61	50.4	347	520	9.4	36.22	147
11	29	57.3	434	757	9.3	38.98	111
12	28	52.3	361	746	9.7	38.74	121
13	14	51.5	181	347	10.9	30.18	98
14	18	59.4	275	448	7.9	46	119

15	17	51.9	454	515	9	12.95	86
16	23	54	462	453	7.1	39.04	132
17	47	55	625	905	9.6	41.31	111
18	13	61	91	132	8.2	48.52	100
19	31	55.2	35	71	6.6	40.75	148
20	12	56.7	453	716	8.7	20.66	67
21	10	70.3	213	582	6	7.05	36
22	110	50.6	3344	3369	10.4	34.44	122
23	56	49.1	412	158	9	43.37	127
24	10	68.9	721	1233	10.8	48.19	103
25	69	54.6	1692	1950	9.6	39.93	115
26	8	56.6	125	277	12.7	30.58	82
27	36	54	80	80	9	40.25	114
28	16	45.7	569	717	11.8	29.07	123
29	29	51.1	379	531	9.4	38.79	164
30	29	43.5	669	744	10.6	25.94	137
31	65	49.7	1007	751	10.9	34.99	155
32	9	68.3	204	361	8.4	56.77	113
33	10	75.5	207	335	9	59.8	128
34	26	51.5	266	540	8.6	37.01	134
35	31	59.3	96	308	10.6	44.68	116
36	10	61.6	337	624	9.2	49.1	105
37	11	47.1	391	463	12.4	36.11	166
38	14	54.5	381	507	10	37	99
39	17	49	104	201	11.2	30.85	103
40	11	56.8	46	244	8.9	7.77	58
41	94	50	343	179	10.6	42.75	125



- Follow the path from the top of the tree (called the **root**) to one of the terminal nodes known as a **leaf** by going along a succession of rules (**splits**).

- The **number** at each tip of the leaves is the **mean value** of the subset of data in belonging to that leaf end. For example the 67.00 at the righthand side terminal node indicates the **mean** SO_2 concentration of all the data that falls into that leaf terminal.

- At any node, the split that **maximally distinguishes** the target variable into the left and right branches is selected.

- This splitting of the data continues until nodes are **pure** or the **data are too sparse** (fewer than six cases : see Breiman et.al 1984)

- Each descriptive feature is assessed in turn (at each split), and the feature that explains the **greatest amount of the variation in the target variable** is selected for that split.
  The variance in the target variable is computed on the basis of a threshold value in the descriptive feature. This threshold value produces two mean values for the target variable, one mean above the threshold value and the other below the threshold.)

- If a descriptive feature is categorical (like ‘yes’, ‘no’, or levels of education), then this tree is called a classification tree.

  Else, if the terminal nodes of the tree are predicted values of a continuous variable, then it is a regression tree model

Key Questions:

- Which variables to use for the division
- How best to achieve the splits for each selected variable