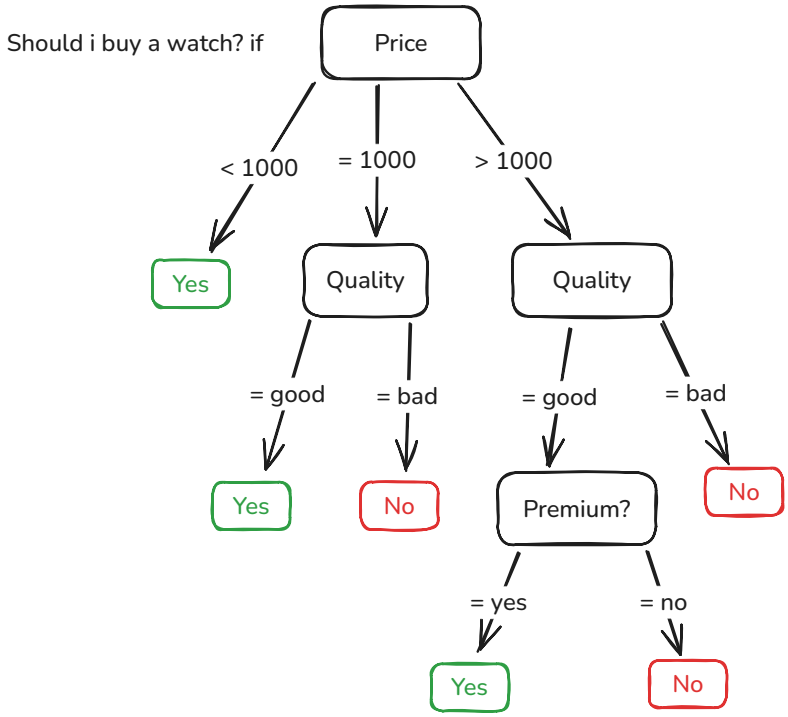
Data Science (Calculations) - Jaish Khan

Decision Tree --> Rules (Induction)

We have this decision tree



To convert this into "Rules" go from left-to-right and write each of the branches:

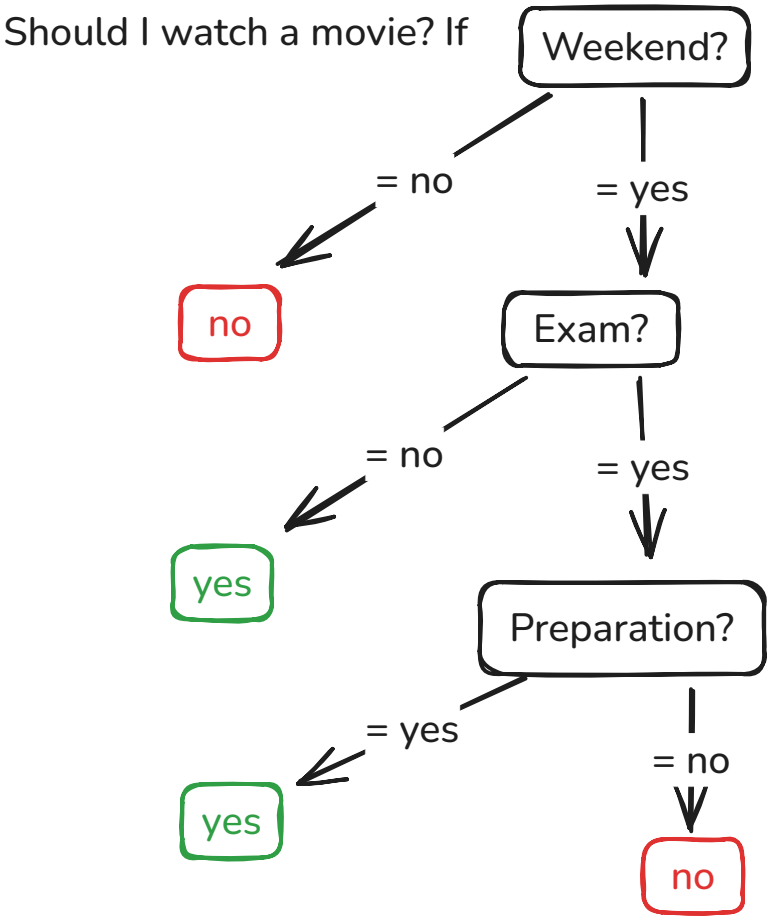
1. If `Price < 1000` then **yes**
2. If `Price = 1000 AND Quality = good` then **yes**
3. If `Price = 1000 AND Quality = bad` then **no**
4. If `Price > 1000 AND Quality = good AND Premium = yes` then **yes**
5. If `Price > 1000 AND Quality = good AND Premium = no` then **no**
6. If `Price > 1000 AND Quality = bad` then **no**

Rules (Induction) --> Decision Tree

We have these rules for this question **Should I watch a movie?**

1. If `Weekend = yes AND Exam = yes AND Preparation = yes` then **yes**
2. If `Weekend = yes AND Exam = yes AND Preparation = no` then **no**
3. If `Weekend = yes AND Exam = no` then **yes**
4. If `Weekend = no` then **no**

Just go line-by-line and draw each branch:



Distance Calculation

There are two types of distances

1. The Euclidean distance between two points $A(x_1, y_1)$ and $B(x_2, y_2)$ in a 2D space is the straight-line distance, calculated using the formula:

$$d_{\mathrm{e}} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Example: Points $A(1, 2)$ and $B(4, 6)$

$$d_e = \sqrt{(4-1)^2 + (6-2)^2} = \sqrt{32 + 42} = \sqrt{9 + 16} = \sqrt{25} = 5$$

2. The Manhattan distance between the same points $A(x_1, y_1)$ and $B(x_2, y_2)$ is calculated as:

$$d_{\mathrm{m}} = |x_2 - x_1| + |y_2 - y_1|$$

Example: Points $A(1, 2)$ and $B(4, 6)$

$$d_{\mathrm{m}} = |4 - 1| + |6 - 2| = 3 + 4 = 7$$

Choosing between Regression, Classification and Clustering

When do we choose each of these

	Regression	Classification	Clustering
When?	You have to predict a "number" value.	You have to separate data into "groups/classes".	You have to find "groups" in unlabeled data.
Input Data is labeled?	Labeled Data	Labeled Data	Unlabeled Data
Target Variable is?	Continuous	Categorical	No Target

K-Nearest Neighbors

Problem: We want to classify a new data point $P(7, 6)$ into one of two classes: **Class A** or **Class B** and We have the following dataset:

Point (x, y)	Class
(1, 1)	A
(2, 2)	A
(3, 3)	A
(8, 8)	B
(9, 9)	B
(10, 10)	B

1. **Choose the value of k**
   Let $k = 3$. This means we'll look at the **3 nearest neighbors** to classify the new point.
2. **Calculate distances** using Euclidean Distance formula
   Use the **Euclidean Distance** formula to calculate the distance between P(7, 6) and all other points:

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

P(x, y)	Class	Distance to P(7,6)P(7, 6)P(7,6)
(1, 1)	A	$\sqrt{(7-1)^2 + (6-1)^2} = \sqrt{36 + 25} = 7.81$

P(x, y)	Class	Distance to P(7,6)P(7, 6)P(7,6)
(2, 2)	A	$\sqrt{(7-2)^2+(6-2)^2} = \sqrt{25+16} = 6.40$
(3, 3)	A	$\sqrt{(7-3)^2+(6-3)^2} = \sqrt{16+9} = 5.00$
(8, 8)	B	$\sqrt{(7-8)^2+(6-8)^2} = \sqrt{1+4} = 2.24$
(9, 9)	B	$\sqrt{(7-9)^2+(6-9)^2} = \sqrt{4+9} = 3.61$
(10, 10)	B	$\sqrt{(7-10)^2+(6-10)^2} = \sqrt{9+16} = 5.00$

3. **Identify the k nearest neighbors**
   Sort the points by distance (ascending order) and pick the 3 nearest neighbors:
   In our case, the three nearest points are:

- (8,8) ⟶ 2.24 belongs to Class A
- (9,9) ⟶ 3.61 belongs to Class A
- (3,3) ⟶ 5.00 belongs to Class B

Since, there are 2 nearest for Class A vs 1 nearest for Class B; The majority class is **Class A**. Hence now our point $P(7, 6)$ belongs to Class A.