Decision Trees - 1

21 March 2025 16:57

Dataset: Airtel

- Label/Target Variable: Attrition
- Attrition 'Yes' means left the company
- Attrition 'No' means working in the company

If company finds out **at early stage** that an employee is about to leave, it has two choices:

1. Let him/her leave.
    a. Company will have to initiate hiring process.
    b. Advantage: Company doesn't need to bargain with the employee.
    c. Drawbacks: Hiring itself is very expensive, Usually the new employee demands more salary, The new recruit might not as good fit in company's culture/with clients as previous employee.
2. Convince her/him not to leave.
    a. Drawback: Airtel might need to bargain with the employee & may need give more salary
    b. Advantages: Many times salary hike is not what the employee was looking for, ditch the expensive and risky hiring process.

But option-2 is only possible when we can predict attrition of a given employee before he/she leaves or even think to leave the company.

Our job as a ML Engineer will be the followings:

1. To find the probability of Attrition for an employee whose data/'features' is given to our model - Logistic Regression
2. Identifying the most important factors (features) for attrition - Model Interpretation

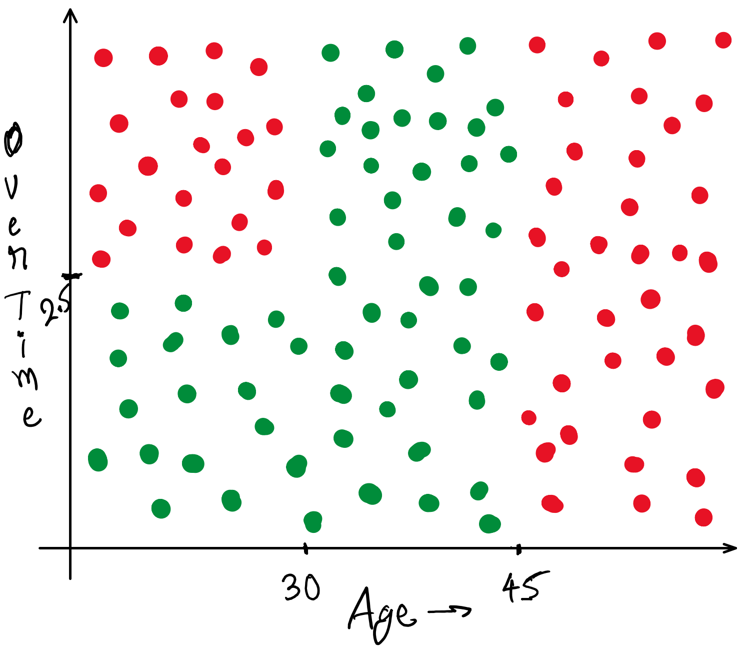
$$\vec{w} \cdot \vec{x} + w_0 = 0 \Rightarrow w_1 x_1 + w_2 x_2 + \dots + w_n x_n + w_0 = 0$$

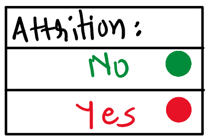
If w2 > w1 then feature X2 is more important factor that feature X1

Standard Procedure to develop an ML model: (Home work)

1. Acquire the data - Data Ingestion Pipeline
2. Preprocessing of the data -
   a. Data cleaning -
      i. Handling missing values
      ii. Removing duplicates
      iii. Dropping the unnecessary columns: Over18, EmployeeCount
   b. Encoding -
      i. One Hot Encoding to columns: BusinessTravel, EducationField, MaritalStatus
      ii. Label Encoding to columns: Attrition, Gender, Overtime
      iii. Target Encoding to columns: Department, JobRole
   c. Treatment for outliers
   d. Feature Engineering -
      i. Reducing dimensionality using statistics (e.g. VIF)
      ii. Create new, more meaningful/relevant features
   e. Data Rebalancing - using SMOTE
   f. Feature Scaling -
      i. Normalization OR
      ii. Standardization
   g. EDA
3. Creating an ML Model

Solving this problem with a different approach:

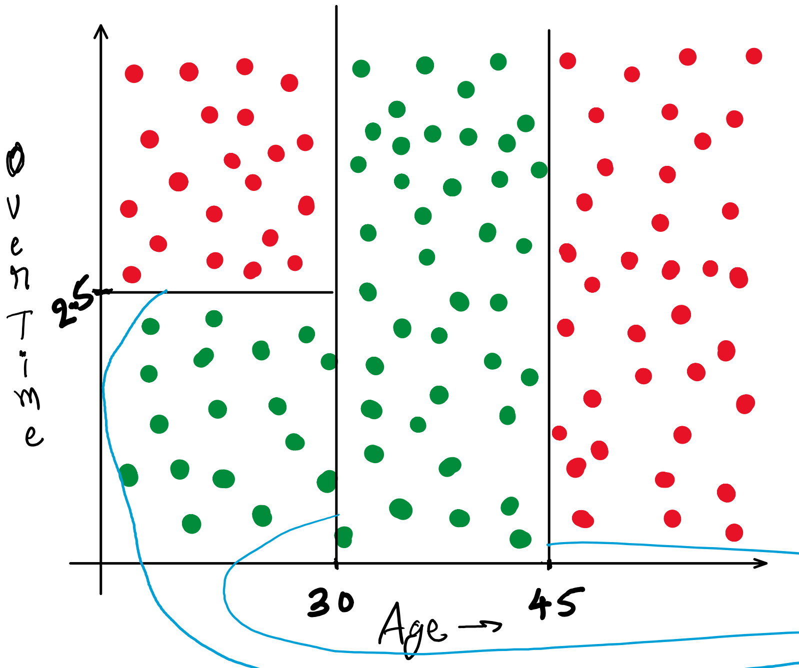




Options that we have:

- As we see, the data is not linearly separable so we cannot use linear regression.
- We may use a polynomial regression model but it is less preferred due to its complex nature & risk of overfitting.
- kNN - No because most multinational companies have lacs of employees and kNN will become very slow.

Another approach: Let's ask a few questions to our new datapoint (query point)

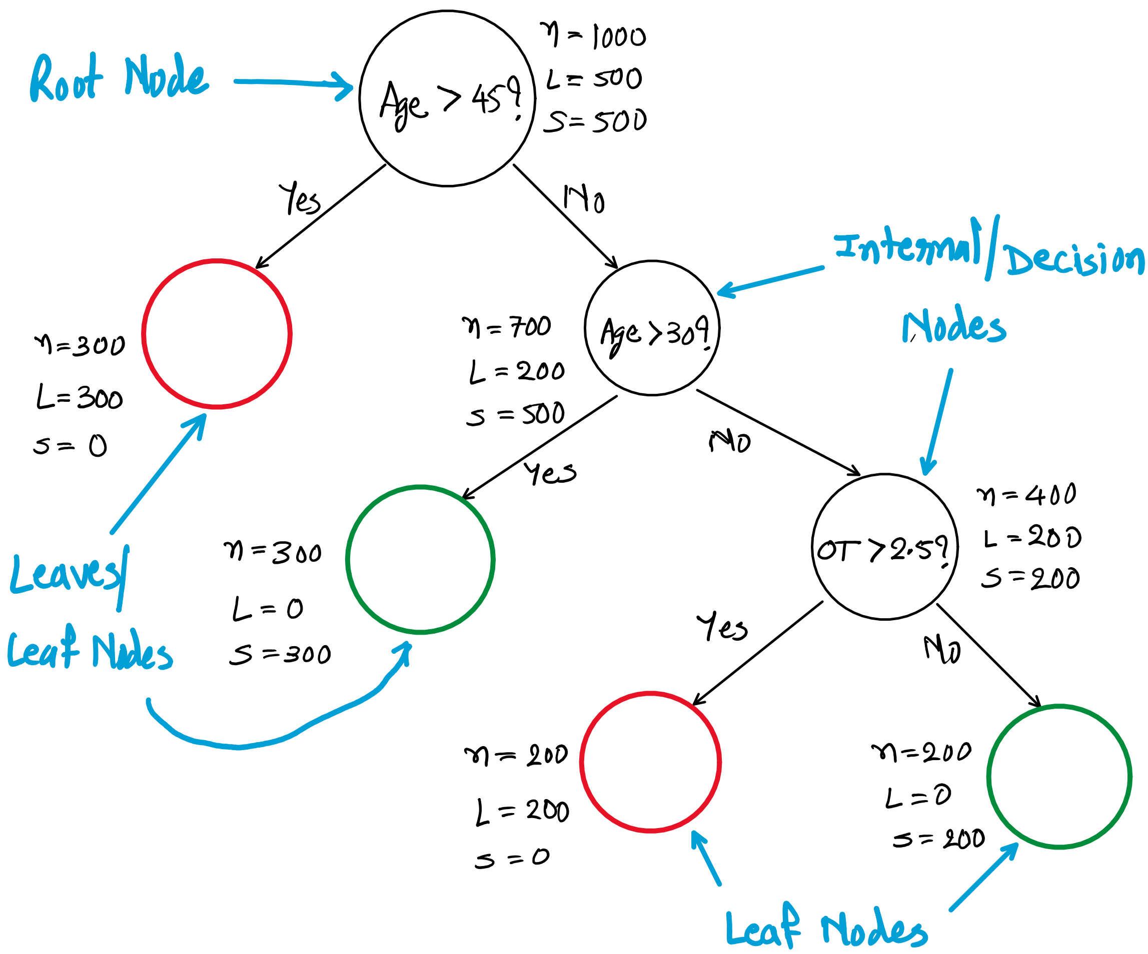


Age > 45?
- Yes - Leave
- No - Age > 30?
  - Yes - Stay
  - No - Overtime > 2.5?
    - Yes - Leave
    - No - Stay

Axis-Parallel Decision Boundaries

But, how can we implement any of these options logically? Ans: using if-else

But first let's try to draw it in the form of a "Tree" as below:



Can I also start with a different question? For example,

OT > 2.5?
- Yes => Age > 30?
  - Yes => Age > 45?
    - Yes => Leave
    - No => Stay
  - No => Leave
- No => Age > 45?
  - Yes => Leave
  - No => Stay

If I call this way of questioning "Option B" and the previous one as "Option A", which one do you think we should start with/is better?

This is going to be a very important point for decision trees because, if we end up starting with a "wrong question" then it might become extremely expensive in terms of computation.

Let's understand this concept using an example.

f1	f2	f3	f4	f5	f6	Y

$y^+ = 100$

$n = 200$

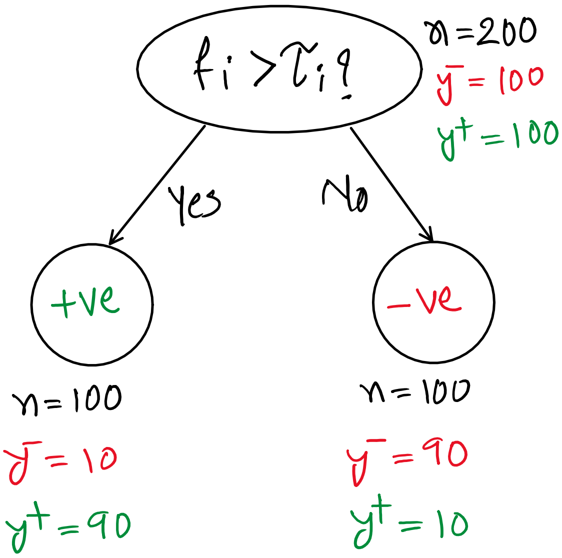
$\bar{y} = 100$

we always ask any question about on of the features only (eg, age, overtime etc). Also we compare that feature with some "threshold" (eg, age > 45 or overtime > 2.5)

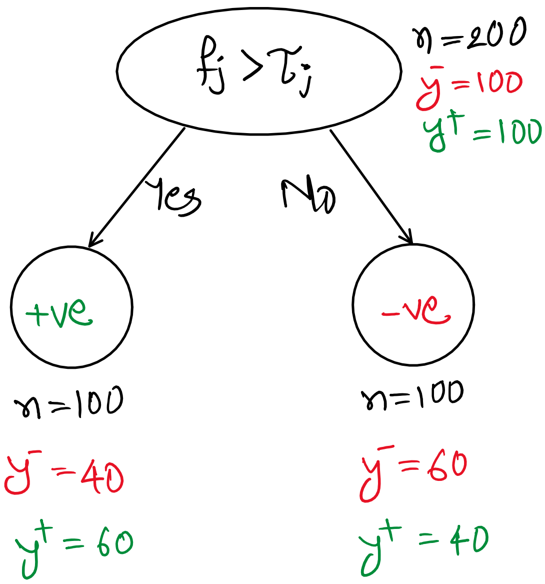
Therefore, in the above table, let's say we chose a question about feature f_i and compared it with the threshold τ_i

Vs.

We chose another question about feature f_j and compared it with the threshold τ_j

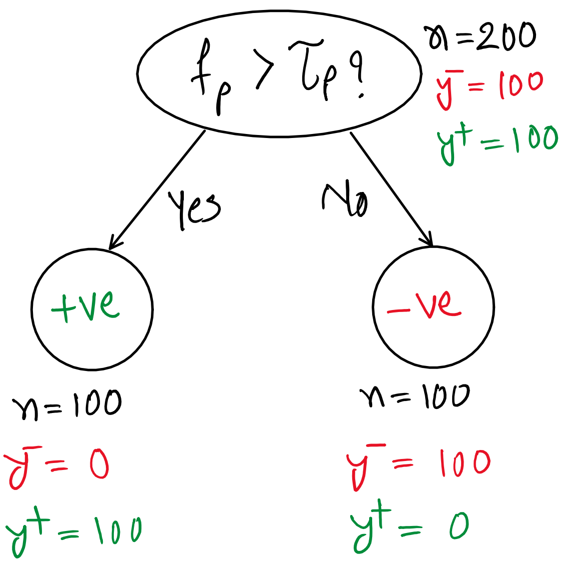


No. of misclassified points = 20

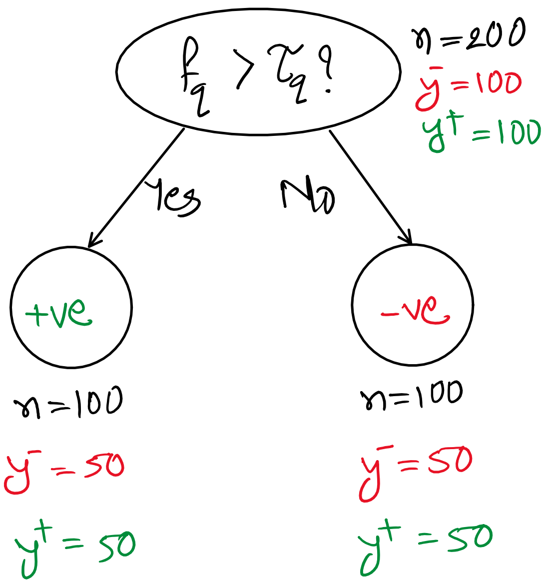


No. of misclassified points = 80

Let's consider another pair of questions-



No. of misclassified points = 0



No. of misclassified points = 100

Pure Homogenous Nodes or Regions

Pure Heterogenouse Nodes or Regions

While the regions/nodes we get by asking question about f_i and f_j, are slightly homogenous or slightly heterogenous. Therefore, there is a need to quantify this "Homogeneity" / "Heterogeneity".

Solution to this is: **Entropy**

Entropy is measure of impurity. Higher the value of entropy, more is the impurity (less pure). For our purpose, entropy of a node x is denoted by H(x) and given by:

$$\boxed{H(x) = -\sum p(x_i) \cdot \log_2 p(x_i)}$$

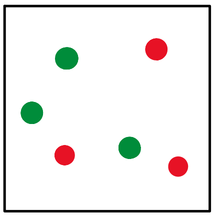
$$\therefore H(x) = -\left[p(x_i \in y^+) \cdot \log_2 p(x_i \in y^+) + p(x_i \in y^-) \cdot \log_2 p(x_i \in y^-) \right]$$

But, $p(x_i \in y^-) = 1 - p(x_i \in y^+)$

$$\therefore H(x) = -\left[p(x_i \in y^+) \cdot \log_2 p(x_i \in y^+) + (1 - p(x_i \in y^+)) \cdot \log_2 (1 - p(x_i \in y^+)) \right]$$

Recall log loss? But this has no connection to log loss it just looks like it.

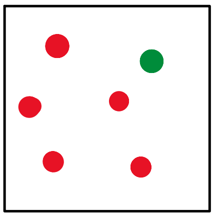
Example:

A 

$p(x_i \in y^+) = 0.5$

$p(x_i \in y^-) = 0.5$

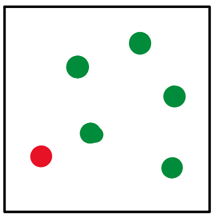
$$H(A) = -(0.5 * \log_2 0.5 + 0.5 * \log_2 0.5)$$
$$= -(-0.5 - 0.5)$$
$$\boxed{H(A) = 1}$$

B 

$p(x_i \in y^+) = 1/6$

$p(x_i \in y^-) = 5/6$

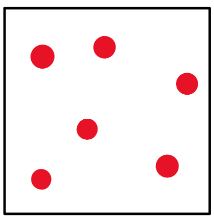
$$H(B) = -\left(\frac{1}{6} * \log_2 \frac{1}{6} + \frac{5}{6} * \log_2 \frac{5}{6}\right)$$
$$\boxed{H(B) = 0.65}$$

C 

$p(x_i \in y^+) = 5/6$

$p(x_i \in y^-) = 1/6$

$$H(B) = -\left(\frac{5}{6} * \log_2 \frac{5}{6} + \frac{1}{6} * \log_2 \frac{1}{6}\right)$$
$$\boxed{H(B) = 0.65}$$

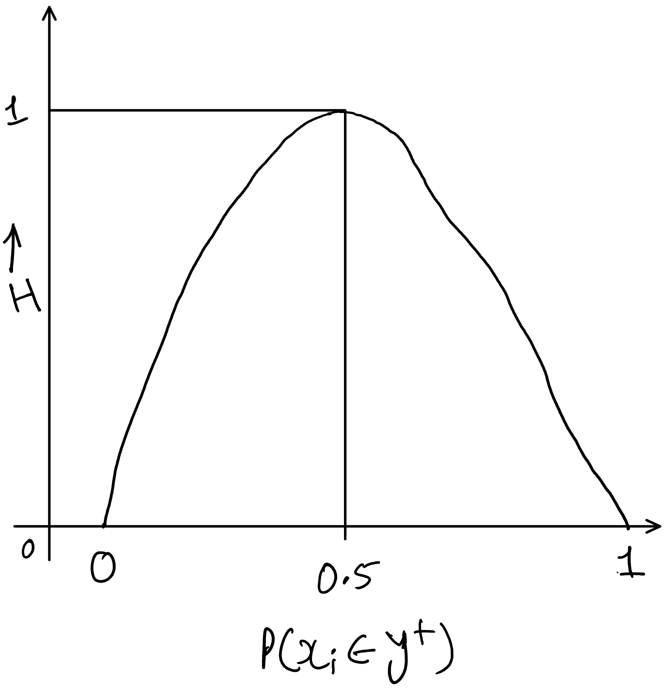
D 

$p(x_i \in y^+) = 0$

$p(x_i \in y^-) = 1$

$$H(B) = -(0 * \log_2 0 + 1 \cdot \log_2 1)$$
$$= -(0 + 0)$$
$$\boxed{H(B) = 0}$$

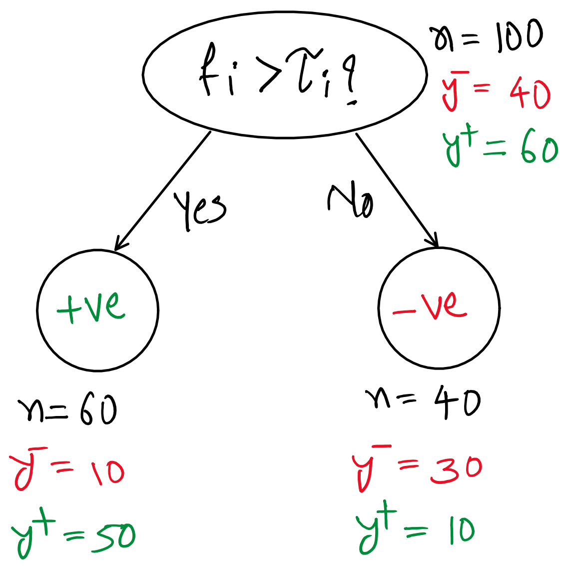
If we plot graph of Entropy vs. P(xi $\in$ y+), it will look like this:

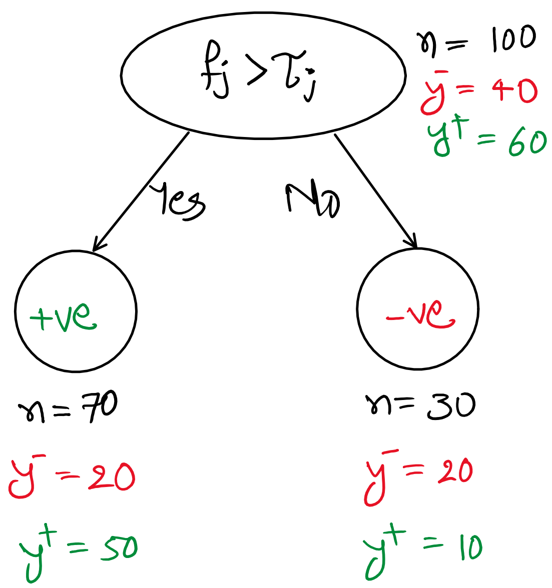


After understanding the Entropy, let's come back to our original problem:

Which feature should I consider questioning first?
Or
Which question is better than the other?
Or
Which question should I ask first?

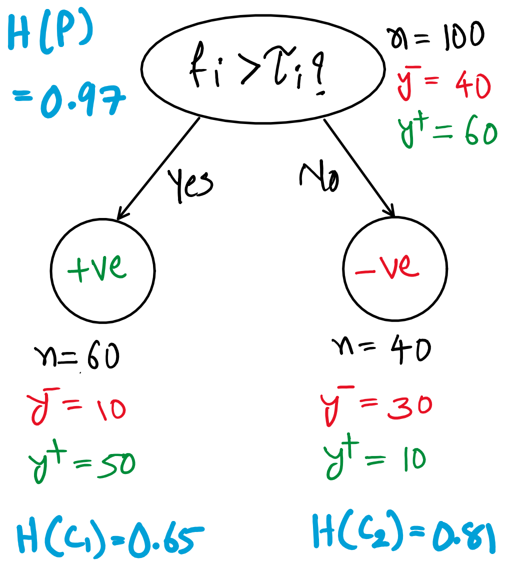
Let's try to solve this problem using Entropy.

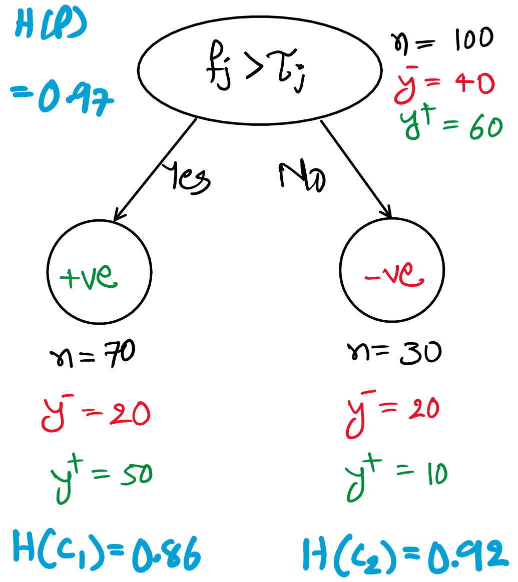




But we will get 3 Entropy values for each question. One at the parent level & 2 at the child level then Which value of entropy should we consider as the entropy of that question?

Ans: We will find the "drop in the entropy" by subtracting entropy at child-level from the entropy at parent-level and for that, we first need to do average of child-level entropies by taking their weighted mean.





Weighted mean entropy at child-level $= \frac{n_1}{n} \cdot H(C_1) + \frac{n_2}{n} H(C_2)$

$\therefore$ <u>For Q-1</u>:

$$H(C) = \frac{60}{100} \times 0.65 + \frac{40}{100} \times 0.81$$

$$= 0.39 + 0.324$$

$$\boxed{H(C) = 0.714}$$

<u>For Q-2</u>:

$$H(C) = \frac{70}{100} \times 0.86 + \frac{30}{100} \times 0.92$$

$$= 0.602 + 0.276$$

$$\boxed{H(C) = 0.878}$$

Drop in the entropy with feature i

= 0.97 - 0.714
= 0.256

Drop in the entropy with feature j

= 0.97 - 0.878
= 0.092

This drop in the Entropy is also known as **Information Gain**.
Hence, the question with more Information Gain is better to start with.