DECISION TREE

WHAT IS MEAN BY DECISION TREE

a Supervised Machine Learning Algorithm and Visual, flowchart-like structure used for classification and regression tasks

Gini Index & Entropy

ENTROPY:

A METRIC MEASURING IMPURITY, DISORDER, OR UNCERTAINTY WITHIN A DATASET, RANGING FROM 0 (PERFECTLY PURE) TO 1 (MAXIMUM DISORDER)

$$E = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

GINI:

A METRIC USED TO DETERMINE HOW WELL A PARTICULAR FEATURE SPLITS THE DATA INTO DISTINCT CLASSES

$$Gini = 1 - \sum_{i=1}^{j} P(i)^2$$

🔹 STEP 1: DATASET LOADING

- THE DATASET IS LOADED FROM A CSV FILE USING PANDAS
- IT CONTAINS USER-RELATED FEATURES AND A TARGET COLUMN CALLED INTEREST

🔹 STEP 2: FEATURE AND TARGET SELECTION

- FEATURES (X):
- ALL COLUMNS EXCEPT INTEREST
- TARGET (Y):
- INTEREST (INDICATES WHETHER THE USER IS INTERESTED IN THE MOVIE)

🔹 STEP 3: TRAIN-TEST SPLIT

- 80% OF DATA IS USED FOR TRAINING
- 20% OF DATA IS USED FOR TESTING
- RANDOM_STATE=42 ENSURES REPRODUCIBILITY

🔹 STEP 4: DECISION TREE CLASSIFIER

- CRITERION: GINI
- → MEASURES IMPURITY IN THE DATASET
- MAX DEPTH: 5
- → PREVENTS OVERFITTING

🔹 STEP 5: MODEL TRAINING

- THE MODEL LEARNS PATTERNS FROM THE TRAINING DATA USING THE .FIT() METHOD

🔹 STEP 6: PREDICTION

- THE TRAINED MODEL PREDICTS MOVIE INTEREST ON UNSEEN TEST DATA

🔹 STEP 7: MODEL EVALUATION

- ACCURACY SCORE: MEASURES OVERALL CORRECTNESS

🔹 STEP 8: MODEL SAVING

- THE TRAINED MODEL IS SAVED USING JOBLIB
- CAN BE REUSED IN FLASK OR STREAMLIT APPLICATIONS