Encoding categorical variables

Why Encode Categorical Variables?

we convert categories into numerical representations using encoding techniques.

 ✓Label Encoding

 ✓One-Hot Encoding

✅ Label Encoding

 What it does:

Converts categories to integer labels (e.g., 0, 1, 2)

No new columns are created

	Color	Color_encoded
0	Red	2
1	Green	1
2	Blue	0
3	Red	2

✅One-Hot Encoding

 What it does:

Converts each category into a new binary column (0 or 1)

No ordering problem

	Color_Blue	Color_Green	Color_Red
0	0	0	1
1	0	1	0
2	1	0	0
3	0	0	1

When to Use What?

Situation	**Encoding Method**
Nominal data (no order)	One-hot encoding
Ordinal data (has order)	Label encoding
Tree-based model (e.g., Random Forest)	Label encoding often works well
Linear models (e.g., Logistic)	One-hot is better

When to Use What?

Ordered

These have a natural ranking or order, but the differences between categories may not be equal.

Size

Small < Medium < Large

Unordered

These are just labels or names, with no meaningful order or hierarchy.

Gender

Male, Female, Other

What is Overfitting?

Overfitting is when a machine learning model performs very well on the training data but fails to generalize to new, unseen data (like test or real-world data).

Why Overfitting Happens

1. Model is Too Complex

Example: A deep neural network used for a small dataset

Complex models can "memorize" the training data

2. Not Enough Training Data

Small datasets don’t represent real-world variability

Model thinks rare patterns are important (but they’re not)

3. Too Many Features

Irrelevant or redundant features confuse the model

Increases the chance of fitting noise

5. Lack of Regularization

Regularization (like L1, L2) discourages overfitting by penalizing complexity

Without it, model parameters grow too flexible