Encoding Nominals

Encoding nominals with integers is wrong, because the ordering and size of the integers are meaningless
e.g. blue 1, green 2, orange 3 — x

One hot encoding is better

explode nominal attributes into many binary attributes, one for each discrete value

Use sklearn.preprocessing.OneHotEncoder
Use pandas get_dummies

Use the one hot encoded features in place of

Encoding with Many Classes

What about having many classes in a feature
- Feature with 50 states
- Feature with 190 countries

Define a hierarchical structure

Example: Zip code — use regions → states → city as the hierarchy and choose a specific level to encode the zip column

Try to group levels by similarity to reduce the overall number of groups