Intro to Machine Learning

Preliminary Steps

1. **import libraries:** pd, np, sklearn.preprocessing (StandardScaler, OneHotEncoder), sklearn.impute (SimpleImputer), sklearn.compose (make_column_transformer, make_column_selector), sklearn.pipeline (make_pipeline), sklearn.model_selection (train_test_split), sklearn (set_config), set_config(display = 'diagram')

2. **load data**

3. **explore data:** df.info(), df.head() AND encode ordinal data with replacement dictionary

Assign & Train Test Split

4. assign target (y) and features (X)

5. X_train, X_test, Y_train, Y_test = train_test_split(X, y, random_state = 42)

Instantiations

6. **column selectors:** col_selector = make_column_selector(dtype_include = 'dtype')

7. **imputers:** strategy_imputer = SimpleImputer(strategy = 'strategy')*

8. **transformers:** scaler = StandardScaler(),
ohe = OneHotEncoder (handle_unknown = 'ignore', sparse = False)

9. **pipelines:** dtype_pipe = make_pipeline(strategy_imputer, scaler/ohe)**

10. **tuples:** dtype_tuple = (dtype_pipe, col_selector)

11. **column transformer:** preprocessor = make_column_transformer(dtype_tuple, other_dtype_tuple)**

Train & Transform

12. preprocessor.fit(X_train)

13. X_train/test_processed = preprocessor.transform(X_train/test)

Inspect

14. print(np.isnan(X_train/test_processed).sum().sum(), 'missing values in training/testing data')

15. print('All data in X_train/test_processed are', X_train/test_processed.dtype)

16. print('shape of data is', X_train_processed.shape)

17. X_train_processed

* mean, median, mode, or most_frequent

** to see pipelines or preprocessor: display()