

# Preprocessing with Sklearn pipeline

---

UTKARSH GAIKWAD

CLASS STARTING SHARP AT 3:07 PM



# Basic steps in machine learning

---

Step 1 : Reading dataset (Data Ingestion)



Step 2 : Perform basic data quality checks  
Check missing values and duplicates



Step 3: Separate X and Y features



Step 4: Perform preprocessing on X



Step 5: Perform train test split



Step 6: Model Building



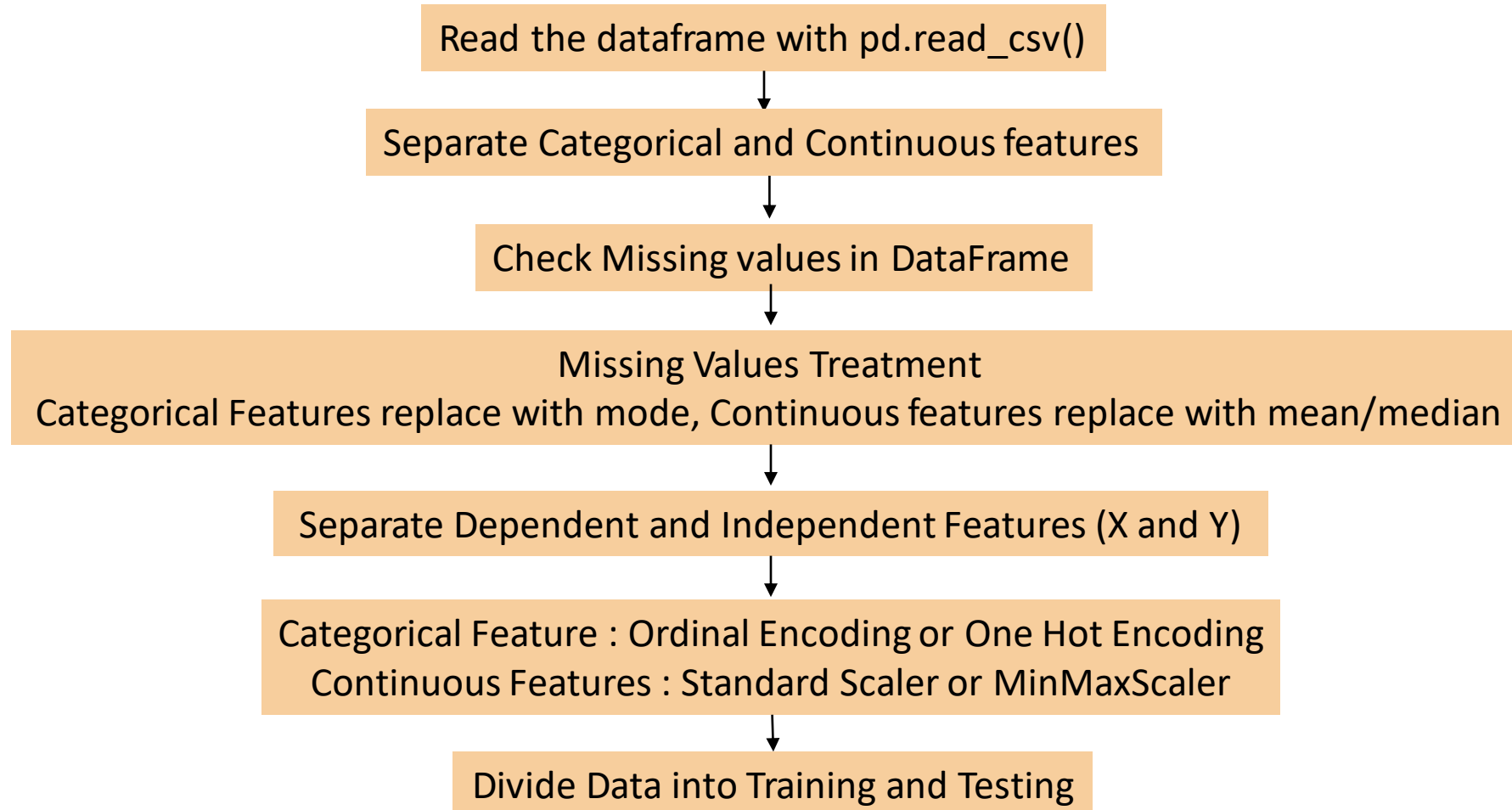
Step 7: Model Evaluation  
MSE, RMSE, MAE, R2



Step 8 : out of sample predictions

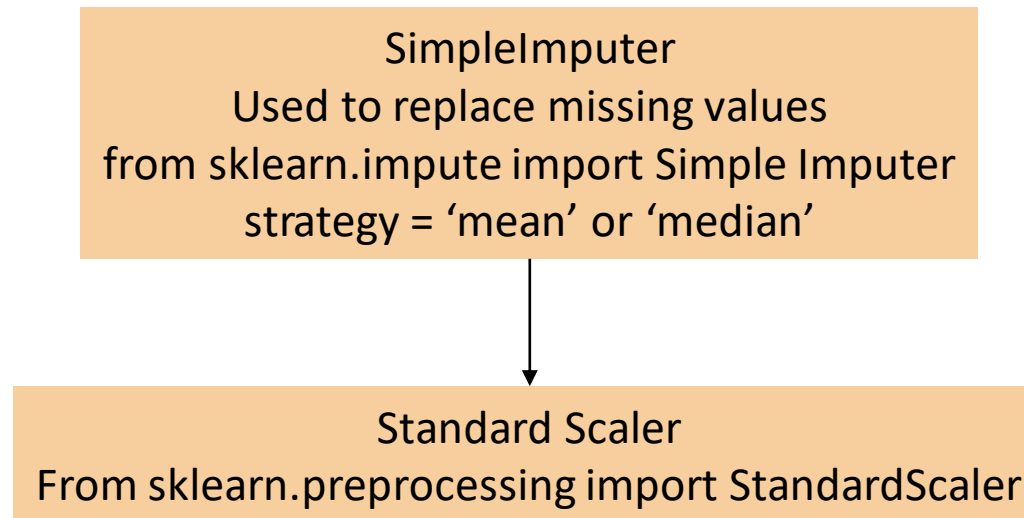
# Basic Steps in creating a Data Preprocessing

---



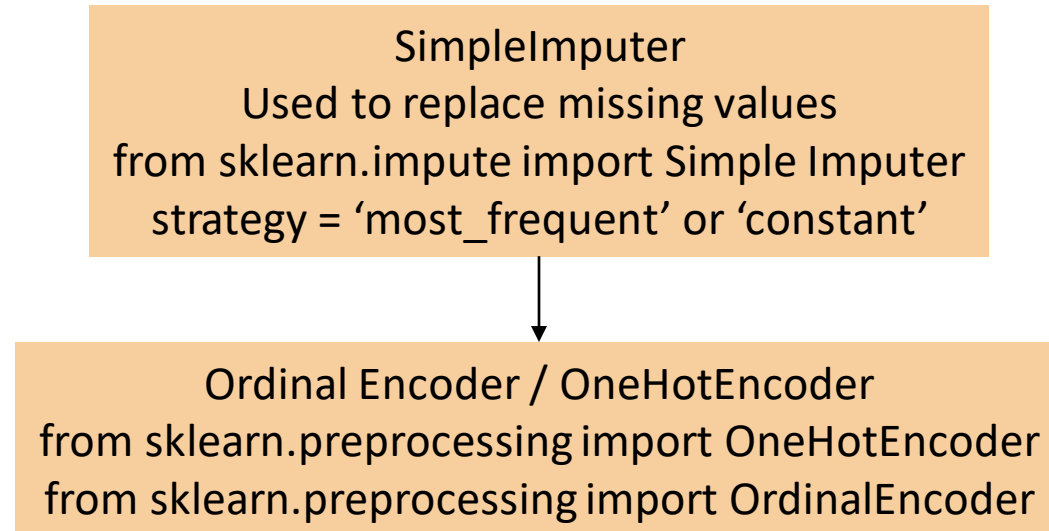
# Sklearn pipeline numeric features

---



# Sklearn pipeline for categorical features

---



# Column Transformer

---

```
from sklearn.compose import ColumnTransformer
```



```
Numeric Pipeline, Numeric features (con)
```



```
Categorical Pipeline , Categorical features (cat)
```

# Example Code with all pipeline

```
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OrdinalEncoder, StandardScaler
from sklearn.compose import ColumnTransformer
```

Import the Pipeline functions

```
# Numeric pipeline
num_pipeline = Pipeline(steps=[('imputer', SimpleImputer(strategy='mean')),
                                ('scaler', StandardScaler())])
```

Numeric Pipeline

```
# Categorical Pipeline
cat_pipeline = Pipeline(steps=[('imputer', SimpleImputer(strategy='most_frequent')),
                                ('ordinal_encoder', OrdinalEncoder())])
```

Categorical Pipeline

Combine  
Cat and  
num pipes

```
# Column Transformer
preprocessor = ColumnTransformer([('num_pipeline', num_pipeline, 'num'),
                                   ('cat_pipeline', cat_pipeline, 'cat')])
```

# Thank you

---

PING ME ON SKYPE FOR ANY QUERIES

COMPLETE THE PRACTICAL THAT WAS ALL FOR TODAY

