#Import required libraries

import pandas as pd

import numpy as np

from sklearn.feature_selection import VarianceThreshold

Step 1: Load the dataset

df = pd.read_csv("your_file.csv")

 Step 2: View the basic info

print("Original Data Info:") print(df.info()) print("\nMissing values per column:\n", df.isnull().sum())

Step 3: Handle missing values

Option A: Drop rows with any missing values

df_cleaned = df.dropna()

Option B (Alternative): Fill missing values with mean (use only one option)

df_cleaned = df.fillna(df.mean(numeric_only=True))

print("\nAfter handling missing values:") print(df_cleaned.info())

Step 4: Remove zero variance columns

These columns have the same value across all rows (no useful information)

selector = VarianceThreshold(threshold=0.0)
selector.fit(df_cleaned.select_dtypes(include=[np.number])) # Only numeric cols

Get column names with non-zero variance

non_zero_variance_cols =
df_cleaned.select_dtypes(include=[np.number]).columns[selector.get_support()]

df_final = df_cleaned[non_zero_variance_cols]

print("\nColumns retained after removing zero variance:")
print(non_zero_variance_cols.tolist())

Step 5: Save cleaned data to CSV (optional)

```
df_final.to_csv("cleaned_data.csv", index=False) print("\n Cleaned data saved as 'cleaned_data.csv'")
```