# 4: DATA PREPROCESSING

## NAME: P JAYAPRADHA

## ROLL NO: 230701130

import pandas as pd

import numpy as np

from sklearn.preprocessing import LabelEncoder, StandardScaler

df = pd.read\_csv('Hotel\_Dataset.csv')

print("Original Dataset:")

print(df.head())

df.replace({'Bill': { -1: np.nan, -99999: np.nan, 0: np.nan},

'NoOfPax': {-1: np.nan, 0: np.nan},

'EstimatedSalary': {-99999: np.nan, 0: np.nan},

'Rating(1-5)': { -1: np.nan}},

inplace=True)

df = df.drop\_duplicates()

df['Bill'] = df['Bill'].fillna(df['Bill'].mean())

df['NoOfPax'] = df['NoOfPax'].fillna(df['NoOfPax'].mode()[0]) # Mode for categorical-like column

df['EstimatedSalary'] = df['EstimatedSalary'].fillna(df['EstimatedSalary'].mean())

df['Rating(1-5)'] = df['Rating(1-5)'].fillna(df['Rating(1-5)'].mode()[0])

label\_encoder = LabelEncoder()

df['Hotel'] = label\_encoder.fit\_transform(df['Hotel'])

df['FoodPreference'] = label\_encoder.fit\_transform(df['FoodPreference'])

df = pd.get\_dummies(df, columns=['Age\_Group'], drop\_first=True)

scaler = StandardScaler()

df[['Bill', 'EstimatedSalary']] = scaler.fit\_transform(df[['Bill', 'EstimatedSalary']])

print("\nPreprocessed Dataset:")

print(df.head())

df.to\_csv('Preprocessed\_Hotel\_Dataset.csv', index=False)





