import pandas as pd

from sklearn.preprocessing import LabelEncoder

# Load the dataset

file\_path = "/root/fake\_job\_postings.xlsx"  # Update with your file path

df = pd.read\_excel(file\_path, engine="openpyxl") # Use the file\_path variable here

# ... (rest of your code remains unchanged)

# Drop irrelevant columns

df\_cleaned = df.drop(columns=["job\_id", "department", "salary\_range"])

# Fill missing values

text\_columns = ["company\_profile", "description", "requirements", "benefits"]

df\_cleaned[text\_columns] = df\_cleaned[text\_columns].fillna("")

categorical\_columns = ["employment\_type", "required\_experience", "required\_education", "industry", "function"]

df\_cleaned[categorical\_columns] = df\_cleaned[categorical\_columns].fillna("Unknown")

# Encode categorical features

label\_encoders = {}

for col in categorical\_columns:

    le = LabelEncoder()

    df\_cleaned[col] = le.fit\_transform(df\_cleaned[col])

    label\_encoders[col] = le

# Save the preprocessed dataset

preprocessed\_file\_path = "/content/preprocessed\_fake\_job\_postings.csv"

# Use to\_csv to save as a CSV file

df\_cleaned.to\_csv(preprocessed\_file\_path, index=False)

print(f"Preprocessed dataset saved as {preprocessed\_file\_path}")