Edu2Job Predicting Job Roles from Educational Background

Dataset Name: Edu2Job\_dataset

Performing EDA for the dataset:

Code:

# -------------------------------

# EDA Report for Edu2Job Dataset

# -------------------------------

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from tkinter import Tk

from tkinter.filedialog import askopenfilename

# -------------------------------

# Step 1: Select the CSV file manually

# -------------------------------

Tk().withdraw() # Hide the root Tkinter window

file\_path = askopenfilename(

title="Select Edu2Job\_dataset.csv",

filetypes=[("CSV Files", "\*.csv")]

)

if not file\_path:

print("❌ No file selected. Please run again and choose the dataset.")

exit()

# Load dataset

df = pd.read\_csv(file\_path)

print("✅ File loaded successfully from:", file\_path)

# -------------------------------

# Step 2: Basic Info

# -------------------------------

print("\n📘 Dataset Shape:", df.shape)

print("\n📄 Column Names:", df.columns.tolist())

print("\n🧩 Data Types:\n", df.dtypes)

print("\n🔍 Missing Values:\n", df.isnull().sum())

# -------------------------------

# Step 3: Summary Statistics

# -------------------------------

print("\n📊 Descriptive Statistics:\n", df.describe(include='all'))

# -------------------------------

# Step 4: Unique Values Count

# -------------------------------

print("\n🔢 Unique Values per Column:\n", df.nunique())

# -------------------------------

# Step 5: Correlation Analysis

# -------------------------------

print("\n📈 Correlation Matrix (Numeric Columns):\n", df.corr(numeric\_only=True))

plt.figure(figsize=(6,4))

sns.heatmap(df.corr(numeric\_only=True), annot=True, cmap='Blues')

plt.title("Correlation Heatmap")

plt.show()

# -------------------------------

# Step 6: Visual Explorations

# -------------------------------

# GPA distribution

plt.figure(figsize=(6,4))

sns.histplot(df['GPA'], bins=20, kde=True)

plt.title("Distribution of GPA")

plt.xlabel("GPA")

plt.ylabel("Frequency")

plt.show()

# Job Role count

plt.figure(figsize=(8,4))

sns.countplot(y='Job Role', data=df, order=df['Job Role'].value\_counts().index)

plt.title("Job Role Distribution")

plt.show()

# Degree count

plt.figure(figsize=(6,4))

sns.countplot(x='Degree', data=df, order=df['Degree'].value\_counts().index)

plt.title("Degree Distribution")

plt.show()

# Experience Level count

plt.figure(figsize=(6,4))

sns.countplot(x='Experience Level', data=df, order=df['Experience Level'].value\_counts().index)

plt.title("Experience Level Distribution")

plt.show()

# -------------------------------

# Step 7: Insights Summary

# -------------------------------

print("\n✅ Insights:")

print("- No missing values found in the dataset.")

print("- Most common degree:", df['Degree'].mode()[0])

print("- Most common major:", df['Major'].mode()[0])

print("- Most common job role:", df['Job Role'].mode()[0])

print("- Most common industry:", df['Industry'].mode()[0])

print("- Most common experience level:", df['Experience Level'].mode()[0])

print("- Average GPA:", round(df['GPA'].mean(), 2))

Output:

✅ File loaded successfully from: C:/Users/prami/INFOSYS/Edu2Job\_dataset.csv

📘 Dataset Shape: (500, 8)

📄 Column Names: ['Degree', 'Major', 'GPA', 'Certifications', 'Skills', 'Job Role', 'Industry', 'Experience Level']

🧩 Data Types:

Degree object

Major object

GPA float64

Certifications object

Skills object

Job Role object

Industry object

Experience Level object

dtype: object

🔍 Missing Values:

Degree 0

Major 0

GPA 0

Certifications 0

Skills 0

Job Role 0

Industry 0

Experience Level 0

There are no missing values present in the dataset.

📊 Descriptive Statistics:

Degree Major GPA ... Job Role Industry Experience Level

count 500 500 500.000000 ... 500 500 500

unique 8 11 NaN ... 14 10 3

top B.A Civil Engineering NaN ... HR Executive IT/Manufacturing Mid-level

freq 75 54 NaN ... 45 58 177

mean NaN NaN 8.436660 ... NaN NaN NaN

std NaN NaN 0.867617 ... NaN NaN NaN

min NaN NaN 7.010000 ... NaN NaN NaN

25% NaN NaN 7.660000 ... NaN NaN NaN

50% NaN NaN 8.380000 ... NaN NaN NaN

75% NaN NaN 9.210000 ... NaN NaN NaN

max NaN NaN 9.990000 ... NaN NaN NaN

[11 rows x 8 columns]

🔢 Unique Values per Column:

Degree 8

Major 11

GPA 238

Certifications 283

Skills 471

Job Role 14

Industry 10

Experience Level 3

dtype: int64

📈 Correlation Matrix (Numeric Columns):

GPA

GPA 1.0









