Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	LTVIP2024TMID25012
Project Title	Predictive Modeling for H1B Visa Approval Using Machine Learning
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description

Data Overview	The dataset consists of 3,002,458 entries and 11 columns, including key attributes such as CASE_STATUS, EMPLOYER_NAME, SOC_NAME, JOB_TITLE, FULL_TIME_POSITION, and PREVAILING_WAGE. **Basic Statistics**: □ Total Rows: 3,002,458 □ Total Columns: 11 □ Unique Values in CASE_STATUS: • CERTIFIED: 2,615,623 • CERTIFIED-WITHDRAWN: 202,659 • DENIED: 94,346 • WITHDRAWN: 89,799 • Other statuses (PENDING, REJECTED, INVALIDATED): Minimal occurrences **Data Types**: o Object: 6 columns (categorical data) o Float: 4 columns (numerical data) o Int64: 1 column (likely an index)
Univariate Analysis	**Description**: Univariate analysis was conducted to explore individual variable characteristics: • CASE_STATUS: Most common value is CERTIFIED. • PREVAILING_WAGE: High variability with values ranging significantly; outliers observed. • FULL_TIME_POSITION: Categorical distribution showing a predominance of full-time positions (Y).

Bivariate Analysis	**Description**: ○ **Relationships**: The correlation between PREVAILING_WAGE and CASE_STATUS shows that higher wages correlate positively with approval. ○ **Scatter Plots**: Generated to visualize the relationship between wage and approval status, revealing trends in the data.
Multivariate Analysis	**Description**: Patterns involving multiple variables were analyzed: • A combination of FULL_TIME_POSITION, PREVAILING_WAGE, and SOC_NAME appears to provide a more robust prediction model for visa approval outcomes. • **Visualization**: PCA (Principal Component Analysis) was applied to reduce dimensionality and identify key features.
Outliers and Anomalies	**Description**: Outliers were identified in the PREVAILING_WAGE column. For instance: ○ Extreme low values (e.g., below $20,000) and high values (e.g., above $250,000) were capped at the 1st and 99th percentiles to maintain model integrity.
Data Preprocessing Code Screenshots	
Loading Data	import pandas as pd df = pd.read_csv("path/to/h1b_dataset.csv") print(df.shape) print(df.head())

Handling Missing Data	# Check for missing values
	missing_values = df.isnull().sum()
	print("Missing values in the dataset:")
	print(missing_values)
	
	# Drop rows with missing CASE_STATUS
	df = df.dropna(subset=['CASE_STATUS']).

Data Transformation	# Transform CASE_STATUS to numeric values df['CASE_STATUS'] = df['CASE_STATUS'].map({ 'CERTIFIED': 1, 'DENIED': 2, 'CERTIFIED-WITHDRAWN': 3, 'WITHDRAWN': 4 })

Feature Engineering	# Create new feature based on SOC_NAME
	def classify_soc_name(row):
	if 'software' in row['SOC_NAME'].lower():
	return 'IT'
	return 'Other'
	
	df['SOC_CATEGORY'] = df.apply(classify_soc_name, axis=1)

Save Processed Data	df.to_csv("processed_h1b_data.csv", index=False) print("Processed data saved successfully.")