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	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
H1B Visa Dataset .	Missing values in critical columns like `CASE_STATUS`, `PREVAILING_WAGE`, and `SOC_NAME`	High	Drop rows with missing values for critical columns; fill missing `SOC_NAME` with the mode.
H1B Visa Dataset .	Duplicate entries for certain applications, leading to potential bias in analysis	Moderate	Identify duplicates using `CASE_NUMBER` and remove them from the dataset.

H1B Visa Dataset	Inconsistent data formats in `FULL_TIME_POSITION` (e.g., 'Y' vs 'N' for full-time).	Low	Standardize the values by mapping them to numeric (1 for 'Y', 0 for 'N').
H1B Visa Dataset	Outliers detected in `PREVAILING_WAGE` (e.g., extremely low or high wages).	Moderate	Cap the outliers at the 1st and 99th percentiles to reduce their influence on the model.
H1B Visa Dataset	Categorical variables not encoded for modeling (e.g., `SOC_NAME`).	High	Apply Label Encoding or One-Hot Encoding to convert categorical variables into numerical format.
H1B Visa Dataset	Data type inconsistencies in numeric fields (e.g., `YEAR` stored as float).	Moderate	Convert `YEAR` and other numeric columns to appropriate data types (e.g., int).