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 Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	This project aims to develop a predictive model for H1B visa approval outcomes using historical visa application data. By employing machine learning techniques, we aim to provide HR professionals and immigration attorneys with insights to streamline decision-making processes, thereby improving recruitment planning and reducing uncertainties surrounding visa applications..
Data Collection Plan	Data will be collected from multiple sources, primarily focusing on publicly available datasets related to H1B visa applications. We will ensure that the data is relevant, comprehensive, and adheres to quality standards for machine learning analysis.
Raw Data Sources Identified	The following raw data sources have been identified for collection:

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
H1B Visa Dataset	Historical data of H1B visa applications, including case status, employer, wage, etc.	Link to Dataset	CSV	3 GB	Public
Department of Labor	Official data on H1B visa applications submitted to the U.S. government, including statistics and reports.	Link to DOL	Excel	1.5 GB	Public
USCIS Immigration Data .	Comprehensive data on immigration applications processed by USCIS, including H1B	Link to USCIS Data	JSON	2 GB	Public
Employer Database	Database containing information on employers who sponsor H1B visas, including their industry.	Link to Employer Data	CSV	500 MB	Private (access required)