Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	738214
Project Title	Predicting Mental Health Illness Of Working Professionals 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
Dataset	Negative values or values less than 18 exist in the "Age" column. This data is factually incorrect and unsuitable for analysis.	High	Remove rows where the "Age" column contains values outside the valid range (18-60). This approach ensures data accuracy but might lead to data loss.
Dataset	The "Country" column has an uneven distribution, potentially causing bias in the model.	Moderate	Remove both the "Country" and "State" columns. This eliminates bias but discards potentially valuable location data. Consider alternative approaches like grouping countries into regions or excluding specific countries with

			very low representation.
Dataset	Missing data points exist in both "self_employed" and "work_interfere" columns.	Moderate	Imputation using the mode (most frequent value). This approach assumes that the most common value represents a typical scenario. Consider alternative imputation techniques depending on the data distribution (e.g., median for numerical features).
Dataset	Inconsistent formatting and variations exist in the "Gender" column entries.	Low	Standardize gender categories into "Male", "Female", and "Non-Binary". This improves data consistency and simplifies analysis. Consider including an "Other" category if there are significant variations beyond these three options.