## Candidate Information:

Name: ANKUR KUMAR

Email: ai.ankurkr@gmail.com

Phone: 9263025968

Date: 20-06-2024

### **Section 3: Analytical and Problem-Solving Skills**

To clean and prepare a dataset with missing values, duplicates, and inconsistencies, I would follow these steps:

Data Importing :

* Load the Dataset and Inspect the structure, types, stats of data using df.info().
* Identify the Missing Values, Duplicated and Inconsistencies using df.isnull(), df.duplicated().

**Handling Missing Values**:

* **Imputation**: Fill missing values using methods such as mean, median, or mode for numerical data, and the mode or a placeholder for categorical data or can use KNN Imputer to impute the missing values.
* Remove rows or columns if more than 60% Data has missing values or if imputation is not appropriate.

**Handling Duplicates**:

* Identify duplicate rows based on unique identifiers or a combination of columns using df.duplicated().
* Remove duplicate rows to avoid skewing the analysis.

**Handling Inconsistencies**:

* Standardize categorical values.
* Check for data entry errors and correct them based on domain knowledge or external validation.
* Detect Outliers of the Data using Boxplot, Z-score or IQR and Treat them By Imputation, Capping, Transformation or can Remove Outliers if treatment is appropriate.

**Data Transformation**:

* Normalize or scale numerical features if necessary if using ML algorithm.
* Encode categorical variables using methods such as one-hot encoding or label encoding.

**Feature Engineering**:

* Create new features if they add value to the analysis or remove extra features which don’t have any use in the analysis.