DATA COLLECTION AND PREPROCESSING PHASE

Date	5 July 2024
Team ID	740102
Project Title	Medical Cost Prediction
Maximum Marks	2 Marks

Data Quality Report

A Data quality report evaluates the accuracy, completeness, consistency, timeliness, validity, and uniqueness of a dataset, identifying issues and recommending actions to improve data reliability and support informed decision-making.

Data source	Data Quality Issue	Severity	Resolution Plan
Medical cost prediction (insurance) kaggle dataset	Inconsistent data format for categorical features like region.	Low	Standardize the data format for categorical features to ensure consistency across the dataset.
	Outliers in features like charges, bmi potentially skewing the prediction model.	Moderate	Apply outlier detection techniques like IQR or z-score method to identify and handle outliers, either by removing them or by applying a transformation like log transformation.

	Inconsistent data formats for categorical features like region.	Low	Standardize the data format for categorical features to ensure consistency across the dataset.
	Biased representation of age groups	Medium	Apply oversampling technique to balance age groups