HOPE AI

Handling Multicollinearity Document

1) **Multicollinearity:**
   - When two independent variables / input columns are highly correlated.
   - If an independent variable is computed from other variables in the data set.
   - If two independent variables provide similar and repetitive results.

Types of Multicollinearity:

A. **Data-based Multicollinearity:** Arises out of the selected dataset naturally or manufacturing / origin defect of the dataset.

B. **Structural Multicollinearity:** This issue arises when researchers have a poorly designed framework for the regression analysis.

To handle multicollinearity:

➔ Remove one of the highly correlated independent variables/input columns.

➔ Combine the highly correlated columns to a single column.

➔ Using dimensionality reduction technique → PCA (principal component analysis) to reduce the number of variables/columns still retaining most of the information.

Methods to detect multicollinearity: (3 techniques)

- Correlation coefficients
- variance inflation factor
- eigenvalue method

Variance Inflation Factor(VIF):

- Measure of the amount of multicollinearity in regression analysis.
- Multicollinearity exists when there is a correlation between multiple independent variables/input columns in a multiple regression model.
- Minimal multicollinearity is required for a good model
  - If VIF value is =1, no correlation,
  - VIF<3 to 5 less correlation,
  - VIF>5 more correlated values, corrective measures to be taken.