# Mid-term assignment

**Introduction:** This report compares the performance of two regression models, Linear Regression and K-nearest neighbors (KNN), on a dataset. The analysis focuses on the impact of various factors such as feature selection, scaling, outlier filtering, and collinearity handling on the models' performance.

**Models:**

1. Linear Regression Model:
2. K-Nearest Neighbors (KNN) Model:

**Metrics:**

1. Root Mean Squared Error (RMSE)

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \hat{x}_i\right)^2}{N}}$$

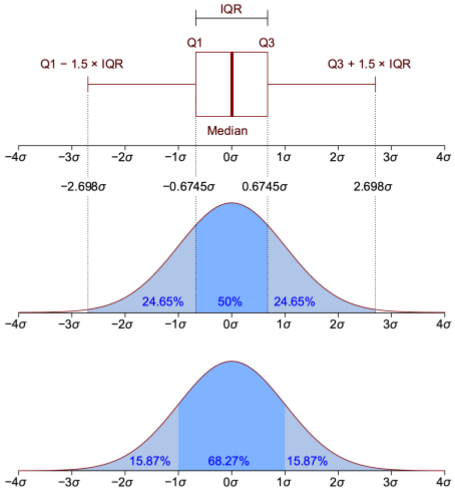
2. R-squared
   - R-squared is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

$$R^2 = 1 - \frac{RSS}{TSS}$$

$R^2$ = coefficient of determination
RSS = sum of squares of residuals
TSS = sum of squares of residuals

**Methods Used**

1. Outliner Filtering



- The Interquartile Range (IQR) method is commonly used for outlier detection and filtering. It involves calculating the IQR, which is the range between the first quartile (Q1) and the third quartile (Q3) of the data distribution.
- Implementation:
  - Calculate Quartiles: The first quartile (Q1) and third quartile (Q3) of the data distribution are calculated.
  - Upper and lower bounds are defined as Q3 + 1.5 * IQR and Q1 - 1.5 * IQR, respectively.
  - Any data point outside the defined thresholds is considered an outlier.

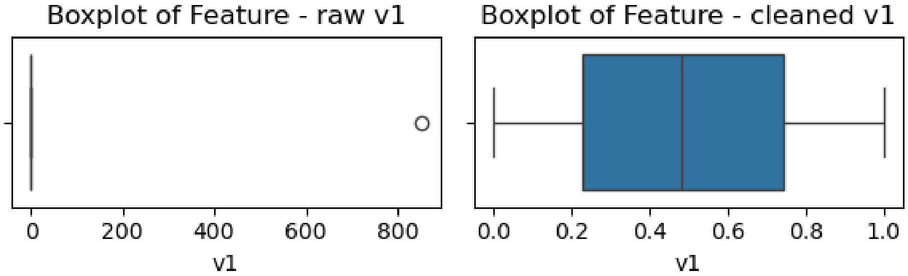


*Figure 1: Feature v1 after removing outliners.*

2. Feature Importance
   - **SelectKBest** is a feature selection technique available in scikit-learn. It operates based on univariate statistical tests to select the top k features that have the strongest relationship with the target variable.
   - Score Calculation**: SelectKBest** calculates a score for each feature based on its relationship with the target variable using the chosen scoring function.
   - Feature Ranking**:** It ranks the features based on their scores, with higher scores indicating stronger relationships with the target variable.
   - Feature Selection**:** The top k features with the highest scores are selected and retained, while the remaining features are discarded.

3. Scaling
   - Standardization (Z-score normalization): This method scales the features to have a mean of 0 and a standard deviation of 1. It is achieved by subtracting the mean from each feature and dividing by the standard deviation.
   - Normalization (Min-Max scaling): This method scales the features to a fixed range, typically between 0 and 1. It is achieved by subtracting the minimum value from each feature and dividing by the range (maximum value - minimum value).

4. Collinearity handling
   - Variance Inflation Factor (VIF) method to detect collinearity among predictor variables. VIF quantifies how much the variance of an estimated regression coefficient is inflated due to multicollinearity. A high VIF value (>10) indicates strong collinearity among predictors.

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{Xj|X-j}$ is the R2 from a regression of $X_j$ onto all the other predictors. If $R^2_{Xj|X-j}$ is close to one, then collinearity is present, and so the VIF will be large.

- VIF ratio calculated from the training data:

| **VIF** | **Value** |
|---|---|
| **v1** | **5447.914179** |
| v2 | 1.004091 |
| **v3** | **251.062339** |
| v4 | 1.003422 |
| **v5** | **84596.222118** |
| v6 | 1.007476 |
| **v7** | **110179.362532** |
| v8 | 1.006334 |
| v9 | 1.010076 |
| **Y** | **251.104069** |

**Results:**

1. Linear Regression
   - The best configuration is:
     - Outliner Filter
     - Feature Selection: 3 features
     - Scaler: StandardScaler
     - K-fold cross-validation: 5 folds

| **No Features** | **Features** | **Average RMSE** | **Average R2** |
|---|---|---|---|
| 1 | ['v3'] | 14.9405 | 0.99596 |
| 2 | ['v3', 'v9'] | 14.9402 | 0.99596 |
| **3** | **['v3', 'v7', 'v9']** | **14.9123** | **0.99598** |
| 4 | ['v3', 'v5', 'v7', 'v9'] | 14.9281 | 0.99597 |
| 5 | ['v1', 'v3', 'v5', 'v7', 'v9'] | 14.9567 | 0.99595 |
| 6 | ['v1', 'v3', 'v4', 'v5', 'v7', 'v9'] | 14.9562 | 0.99595 |
| 7 | ['v1', 'v3', 'v4', 'v5', 'v7', 'v8', 'v9'] | 14.9777 | 0.99594 |

| **No Features** | **Features** | **Average RMSE** | **Average R2** |
|---|---|---|---|
| 8 | ['v1', 'v2', 'v3', 'v4', 'v5', 'v7', 'v8', 'v9'] | 14.9897 | 0.99594 |
| 9 | ['v1', 'v2', 'v3', 'v4', 'v5', 'v6', 'v7', 'v8', 'v9'] | 14.9923 | 0.99593 |

2. KNN Regression

- The best configuration is:
    - Outliner Filter
    - Feature Selection: 1 feature
    - Scaler: None
    - K-fold cross-validation: 10 folds

| **No Features** | **Features** | **Average RMSE** | **Average R2** |
|---|---|---|---|
| **1** | **['v3']** | **1.5055** | **0.99996** |
| 2 | ['v3', 'v9'] | 5.3188 | 0.99948 |
| 3 | ['v3', 'v7', 'v9'] | 15.2102 | 0.99577 |
| 4 | ['v3', 'v5', 'v7', 'v9'] | 15.1907 | 0.99578 |
| 5 | ['v1', 'v3', 'v5', 'v7', 'v9'] | 15.1940 | 0.99578 |
| 6 | ['v1', 'v3', 'v4', 'v5', 'v7', 'v9'] | 15.2813 | 0.99573 |
| 7 | ['v1', 'v3', 'v4', 'v5', 'v7', 'v8', 'v9'] | 15.2813 | 0.99573 |
| 8 | ['v1', 'v2', 'v3', 'v4', 'v5', 'v7', 'v8', 'v9'] | 15.2392 | 0.99575 |
| 9 | ['v1', 'v2', 'v3', 'v4', 'v5', 'v6', 'v7', 'v8', 'v9'] | 15.2414 | 0.99575 |