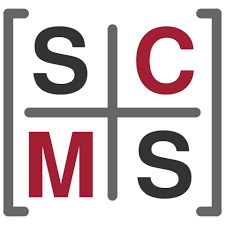
**** ****

**Savitribai Phule Pune University**

**Department of scientific computing modeling and simulation**

**MINI – PROJECT**

**MACHINE LEARNING**

**Performing EDA and Regression**

**Sakshi Ladkat**

**Msc FY**

**MS2408**

**About Dataset**

The **Fish Market dataset** available on Kaggle is a collection of data pertaining to seven common fish species sold in markets. It includes various measurements such as weight, length, height, and width for each fish, making it suitable for tasks like predictive modeling and regression analysis.

**Resource**

**https://www.kaggle.com/datasets/vipullrathod/fish-market**

The dataset contains 159 rows and 7 columns, with no missing values. Of these 7 columns, 6 are numeric, representing variables such as **length1**, **length2**, **length3**, **height**, and **width**. The 7th column is categorical, likely representing different categories or types of fish. The **response variable** is the **weight of the fish**, which is being predicted using the predictor variables, which are the different measurements of the fish (lengths, height, and width). This setup suggests a regression problem, where the goal is to predict the weight of the fish based on these physical dimensions.

**Exploratory Data Analysis**

**Histogram :**

* From your observation, it appears that the **length1** column follows a normal distribution, as suggested by its histogram. However, the histograms of the other columns (length2, length3, height, and width) do not appear to follow a normal distribution. This could imply that the data for these variables may be skewed, have outliers, or follow another distribution type.
* Since the density curve fit suggests that the data is skewed and has outliers, this further confirms that the distribution of predictor variables (**length1**,**length2**, **length3**, **height**, and **width**) is not normal. Skewness indicates that the data might have a long tail on one side, while the presence of outliers suggests that some extreme values are distorting the overall distribution.

**Boxplot :**

**-** The analysis of the dataset reveals several key insights. First, the **length** variable shows significant variability, as indicated by the boxplot, while the **height** and **width** variables exhibit less variability. This suggests that the fish in the dataset have a wide range of lengths, but their height and width are more consistent across observations. Upon examining the relationship between the predictor variables, it was found that **length** is strongly correlated with both **height** and **width**.

-Specifically, smaller fish tend to have smaller height and width, while larger fish exhibit larger dimensions, indicating a proportional relationship between these physical characteristics. This correlation could be further quantified by calculating the correlation coefficients among these variables.

- Additionally, the analysis revealed the presence of outliers, particularly in the species **Roach** and **Smelt**, where extreme values in length, height, and width are observed. These outliers could represent rare cases or data quality issues that warrant further investigation.

**Quantile – Quantile Plot :**

**-** The **Quantile-Quantile (Q-Q) plot** analysis further confirms the presence of skewness in the data, as well as a **curved light tail**, indicating deviations from normality. This pattern suggests that the data is not symmetrically distributed, and there are **outliers** present. A curved Q-Q plot typically indicates **non-linearity**, meaning the data does not follow a normal distribution. The presence of a **light tail** suggests fewer extreme values than expected under normality, while the skewness suggests that some variables have an uneven distribution with a longer tail on one side.

-The **Q-Q normal plot (qqnorm)** also displayed a skewed pattern, further confirming that the data does not follow a normal distribution. This deviation from normality is evident as the data points do not align along the reference diagonal line in the Q-Q plot.

**Pairplot:**

The **pairplot** analysis reveals strong correlations among multiple predictor variables. Specifically, **length1 and length2** are highly correlated, indicating that these two measurements share a linear relationship. Similarly, **length1 and length2 with length3** exhibit the same correlation pattern, suggesting that all three length measurements are proportionally related. Furthermore, **length1, length2, and length3** show a consistent correlation with **height and width**, reinforcing the idea that larger fish tend to have greater height and width. Additionally, **height and width are also correlated**, which aligns with the observations from the **boxplot**, where fish with **maximum height tend to have maximum width**, while those with **lesser height have lesser width** across different species. However, the presence of **outliers**, particularly in certain species, further emphasizes the variability within the dataset. The strong multicollinearity among length, height, and width variables suggests that **dimensionality reduction techniques** like **Principal Component Analysis (PCA)** or **feature selection** may be necessary to avoid redundancy and improve model performance.

**Variance :**

After fitting individual **linear models**, the **residual analysis** reveals a clear pattern, indicating **non-constant variance (heteroscedasticity)** in the data. This violates the assumption of homoscedasticity, which is crucial for reliable linear regression models. The **fitted vs. residual plot** further supports this observation, as the residuals exhibit a structured pattern rather than being randomly scattered, suggesting that the model fails to capture some underlying relationships in the data. Additionally, the **scale-location plot** confirms these insights by showing an increasing or decreasing trend in residual variance, reinforcing the presence of heteroscedasticity. These findings suggest that applying **data transformations** (such as log or Box-Cox transformations) or using **heteroscedasticity-robust regression models** (such as weighted least squares or generalized least squares) may improve model accuracy and make residuals more randomly distributed.

**Summary :**

In summary, after performing **model selection** and fitting **linear, logarithmic, and polynomial regression models**, the results indicate that the **best model** is the **linear regression model**, which achieves an **R² score of 0.8843**. This suggests that the model explains approximately **88.43% of the variance** in fish weight based on the selected predictor variables. The polynomial regression model did not provide any significant improvement over the linear model, indicating that a higher-degree polynomial does not contribute additional predictive power. The final selected model uses **length2, length3, and height** as the most important predictors for estimating fish weight, confirming that these features hold the strongest relationship with the target variable. Given the high R² value, this model can be effectively used for fish weight prediction, though further improvements could be explored by addressing **heteroscedasticity, potential multicollinearity, and outliers**.