IDS Semester Project

Supervisor

Sir Irfan Yousaf

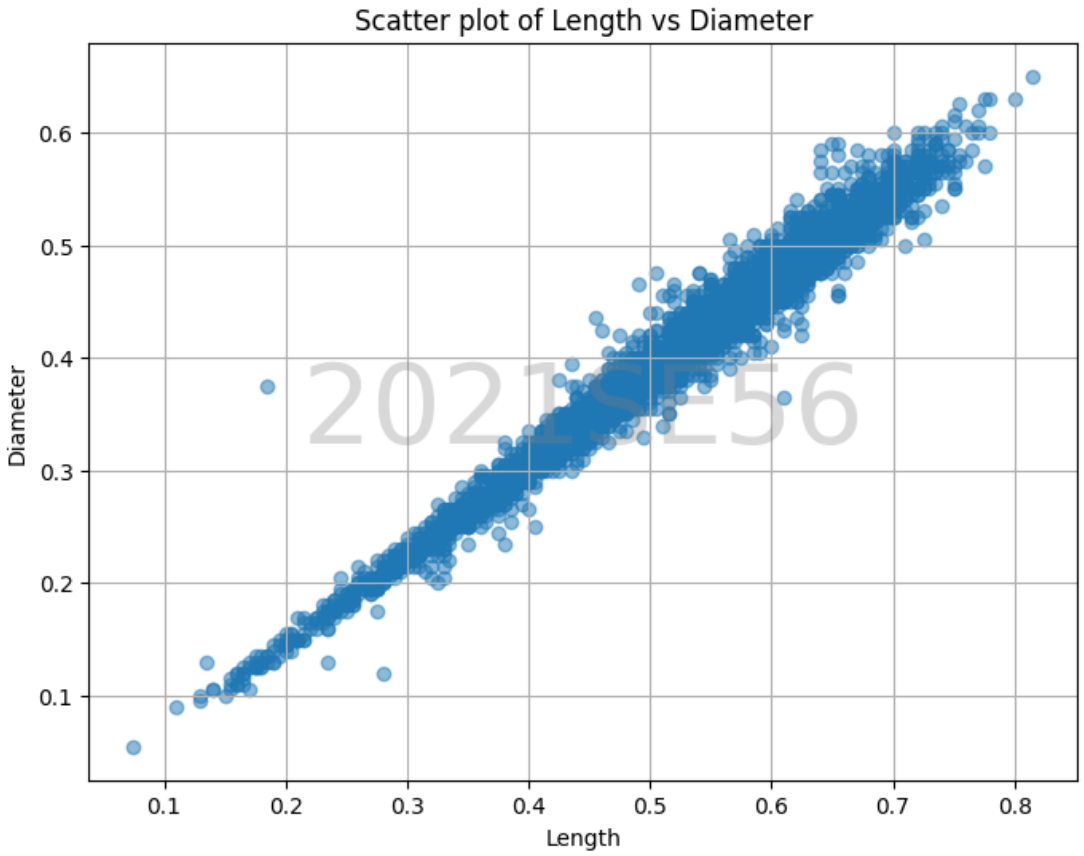
Submitted by

Farjad Waseem
{2021_SE_56}

Department of Computer Science,
University of Engineering and Technology, Lahore, New-Campus.
[28-April-2024]

Abalone Dataset EDA

Figure # 1 (**Scatter Plot - Bivariate**)



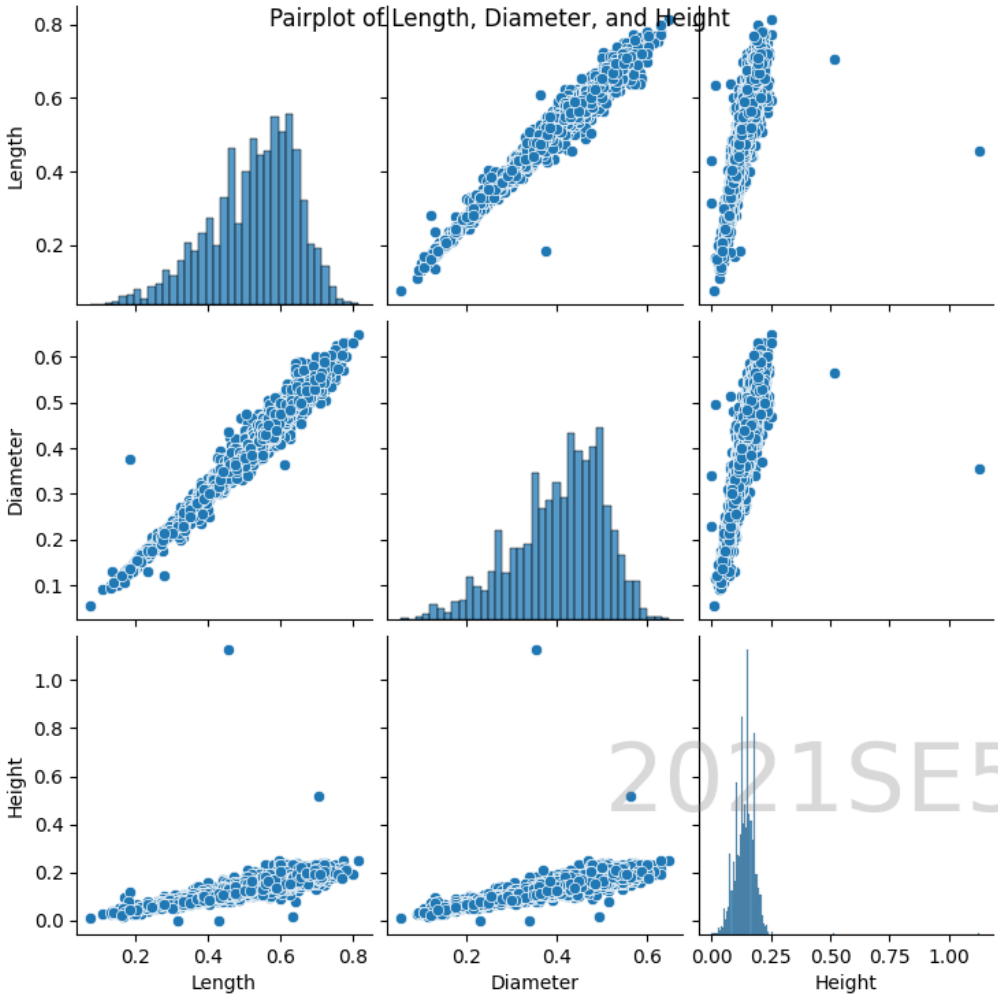
Explanation:

In this scatter plot, each point represents an abalone. The x-axis represents the length while the y-axis represents the diameter. We can observe that there's a strong positive correlation between length and diameter, which is expected as abalones tend to have a proportional shape.

A cluster of points aligned along a diagonal line indicates a strong positive correlation, suggesting that as the length of the abalone increases, its diameter also tends to increase proportionally.

The spread of points around the line indicates the variability in the relationship, which may be influenced by factors like age, species, or environmental conditions.

Figure # 2 (**Pair Plot - Multivariate**)



Explanation:

This pairplot shows scatter plots for each pair of variables (Length vs Diameter, Length vs Height, Diameter vs Height) along with histograms for each variable on the diagonal. It allows us to visualize relationships between multiple variables simultaneously.

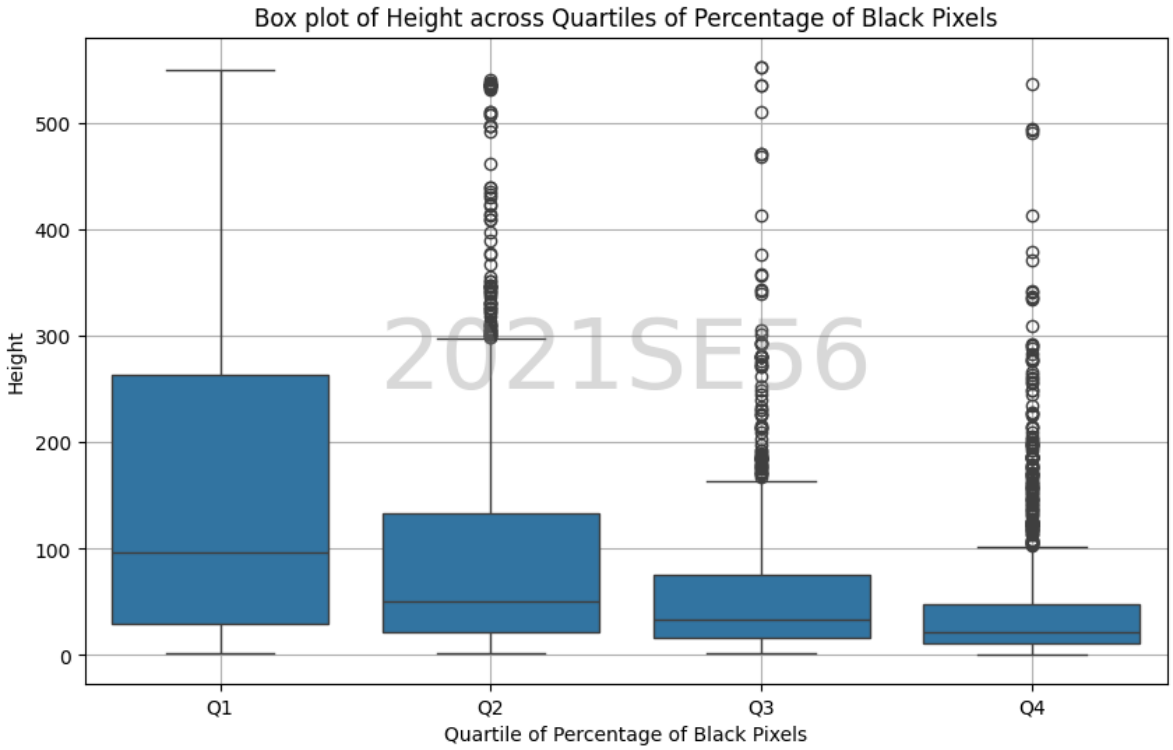
The scatter plots provide insights into the relationships between pairs of variables. For example, the scatter plot of Length vs Diameter shows a strong positive correlation, indicating that as the length of an abalone increases, its diameter also tends to increase.

The histograms along the diagonal show the distributions of individual variables, providing insights into their central tendency and spread.

Diagonal elements represent the distribution of each feature, while the off-diagonal elements represent the relationship between pairs of features.

Page-Blocks Dataset EDA

Figure # 3 (**Box Plot - Bivariate**)



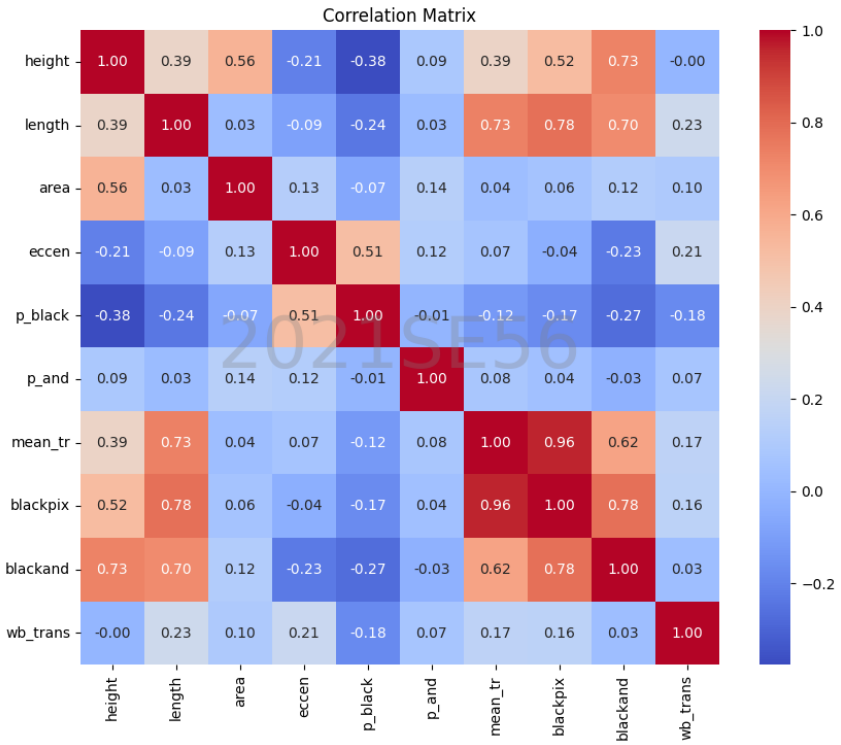
Explanation:

This box plot compares the distribution of block heights across quartiles of the 'Percentage of Black Pixels' attribute. Each box represents the height distribution within a specific quartile of the percentage of black pixels. From the plot, we can observe the following:

1. As the quartile of the percentage of black pixels increases (from Q1 to Q4), the median height of the blocks tends to decrease. This suggests a potential inverse relationship between the darkness of the content (percentage of black pixels) and the height of the blocks.
2. The interquartile range (IQR) for each quartile provides insights into the variability of block heights within different levels of black pixel percentages.
3. Outliers beyond the whiskers of each box plot indicate blocks with extreme heights compared to the majority of blocks within the quartile.

This plot helps in understanding the relationship between the percentage of black pixels and block heights, providing insights into potential patterns in the document layout.

Figure # 4 (**Heat Map - Multivariate**)



Explanation:

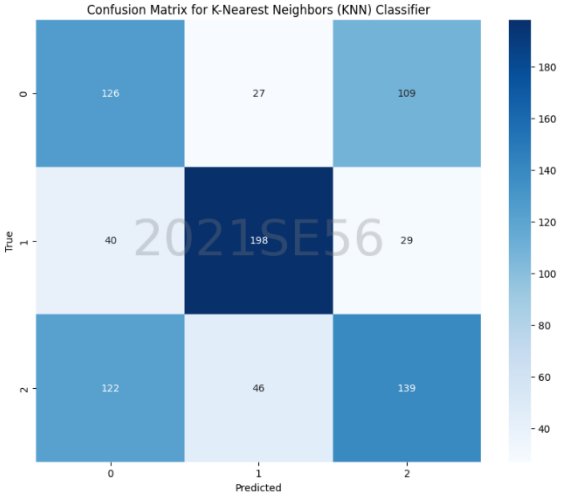
The heatmap shows the correlation between different features of a dataset used for block classification. Here's a breakdown of the correlations:

- **Height, length, and area:** These three features have a positive correlation with each other. This means that blocks with a larger height tend to also have a larger length and area, and vice versa. This makes sense intuitively as the area is calculated by multiplying the height and length.
- **Eccentricity:** Eccentricity has a weak positive correlation with height and a weak negative correlation with length. Eccentricity is the ratio of the length to the height, so a higher eccentricity value corresponds to a block that is longer than it is tall.
- **Black pixels (both original and after applying RLSA) and area:** These features have a weak positive correlation. This means that blocks with a larger area tend to have more black pixels.
- **Mean number of white-black transitions:** This feature has a weak positive correlation with both the number of black pixels (original and after applying RLSA) and the area. This means that blocks with a larger area and more black pixels tend to have more transitions between black and white pixels.

It's important to note that the correlations shown in the heatmap are relatively weak. This means that there is not a perfect linear relationship between any of the features.

KNN & Naïve Bayes

KNN (**Abalone Data**)

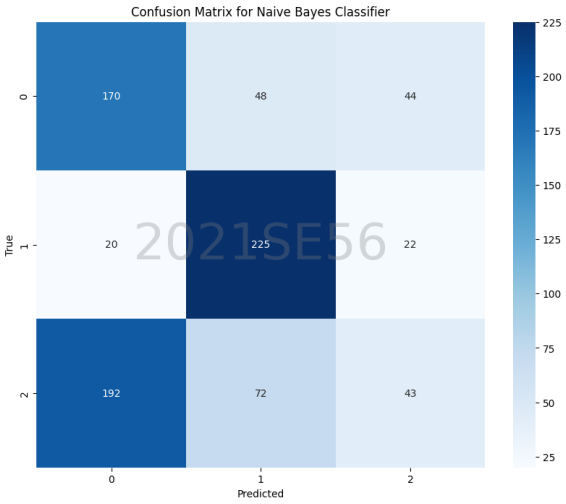


K-Nearest Neighbors (KNN) Classifier:
Accuracy: 0.5538277511961722
Precision: 0.5547328258366705
Recall: 0.5538277511961722
F1-score: 0.5534831678988429

In the specific confusion matrix, the rows represent the actual sexes (M, F, I), and the columns represent the predicted sexes by the KNN classifier. Each cell in the table shows the number of abalone that belong to a particular sex. For example, the cell in the top left corner (labeled "M 126") shows that 126 abalone were actually male (M), and the KNN classifier correctly predicted them as male.

Since the accuracy is low (around 55%), the F1-score is low, it suggests that the KNN classifier is not performing well on classifying the sex of abalone in this dataset.

Naïve Bayes (**Abalone Data**)



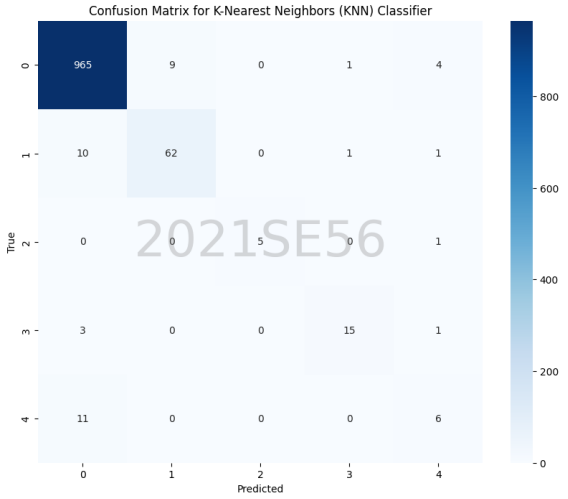
Naive Bayes Classifier:
Accuracy: 0.5239234449760766
Precision: 0.49262845115367393
Recall: 0.5239234449760766
F1-score: 0.4762115513541513

The confusion matrix shows that the Naive Bayes classifier is also not performing well on classifying the sex (M, F, I) of abalone in this dataset.

In a confusion matrix, ideally, the high values should be concentrated on the diagonal. This would indicate that the classifier is correctly classifying most of the abalone sex.

However, the low accuracy (**Accuracy: 0.5239**) suggests that the classifier is frequently misclassifying the sexes. Given these metrics, it's safe to say that the Naive Bayes classifier is not a suitable choice for classifying abalone sex in this dataset.

KNN (**Page-Blocks Data**)



K-Nearest Neighbors (KNN) Classifier:
Confusion Matrix:
Accuracy: 0.9616438356164384
Precision: 0.9593362989966747
Recall: 0.9616438356164384
F1-score: 0.96024400193974

In the confusion matrix, the rows represent the actual classes, and the columns represent the predicted classes. Each cell in the table shows the number of data points that belong to a particular class. For example, the cell in the top left corner (labeled "0 965") shows that 965 data points were actually class 0, and the KNN classifier correctly predicted them as class 0.

The diagonal of the confusion matrix shows the number of correctly classified data points. In this case, the sum of the diagonal elements is 1078, which means that the KNN classifier correctly classified 1078 out of 1088 data points (1078 / 1088 = 0.99). This is also indicated by the value **Accuracy: 0.9616** at the top of the matrix. Other metrics are shown above.

Naïve Bayes (**Abalon Data**)

Naive Bayes Classifier:
Confusion Matrix:
Accuracy: 0.908675799086758
Precision: 0.934978651589429
Recall: 0.908675799086758
F1-score: 0.9172365576617763

In the specific confusion matrix, the rows represent the actual classes, and the columns represent the predicted classes. Each cell in the table shows the number of text snippets that belong to a particular class. For example, the cell in the top left corner (labeled "0 918") shows that 918 text snippets were actually class 0, and the Naive Bayes classifier correctly predicted them as class 0.

While the Naive Bayes classifier has a high accuracy, the precision (0.93) is slightly higher than the recall (0.91), which means the model might be better at not misclassifying irrelevant snippets than finding all the relevant ones.