**Introduction to Data Science**

**Semester Project**



**Session: 2021**

**Submitted by:**

**Wali Muhammad 2021-SE-39**

**Submitted to:**

**Dr. Irfan Yousuf**

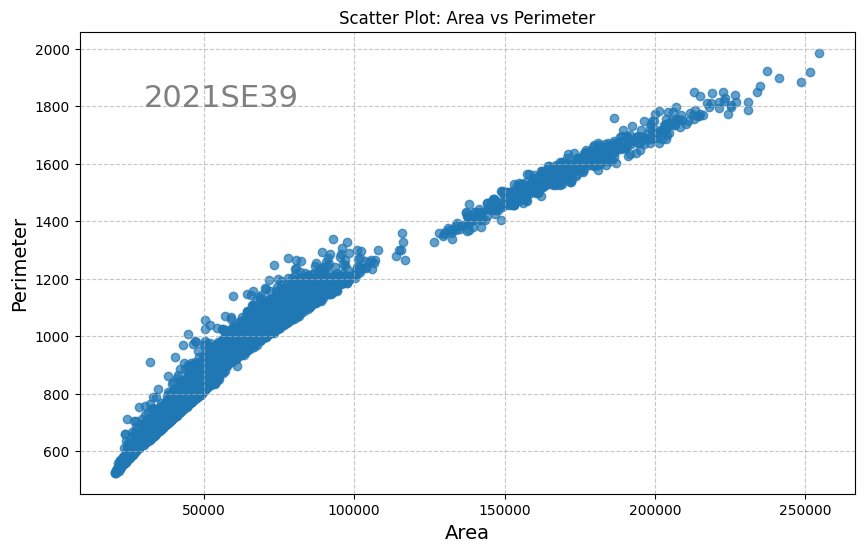
Department of Computer Science,

**University of Engineering and Technology, New Campus Lahore**

**Exploratory Data Analysis**

**Bivariate Analysis**

1. **Scatter Plot for ‘Area vs Perimeter’**

****

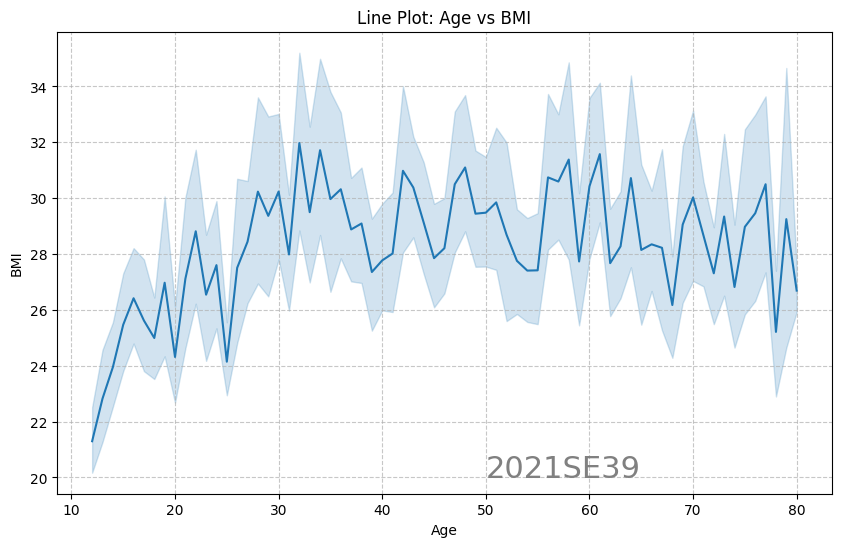
Plot 1.

This scatter plot shows a strong, positive, linear association between Area and Perimeter of Dry bean. The axes are clearly labeled, with the x-axis representing area and the y-axis representing perimeter value. There don't appear to be any outliers in the data.

By looking at scatter plot, a steep ongoing pattern emerges. The data points on the scatter plot form a dense cluster along a diagonal line that ascends from the bottom left to the top right. This alignment suggests a strong positive correlation between area and perimeter. The ‘Area’ is the area of a bean zone and the number of pixels within its boundaries and the ‘Perimeter’ is bean circumference which is defined as the length of its border. So, as the area of an object increases, so does its perimeter. It means larger dry beans tend to have longer perimeters. The most of the datapoints have a strong density in the region of 10000-100000. If we see the graph upward trend, it is braked by a few points in between 120000-130000. Within the general pattern, there are no striking outliers, no data points that deviate significantly from the linear trend. Each bean’s area and perimeter align harmoniously, contributing to the overall coherence of the scatter plot. The absence of outliers reinforces the reliability of this relationship

In conclusion, the scatter plot of area of dry bean versus perimeter of dry bean visually represents how strong and linear relation is in between the area and the perimeter of the dry bean. The scattered data points and the central reference line help grasp the graphical content of the dataset. The overall analysis of the data represented shows that, as the area of bean increases the perimeter also increase.

1. **Line Plot for ‘BMI vs Age’**

****

Plot 2.

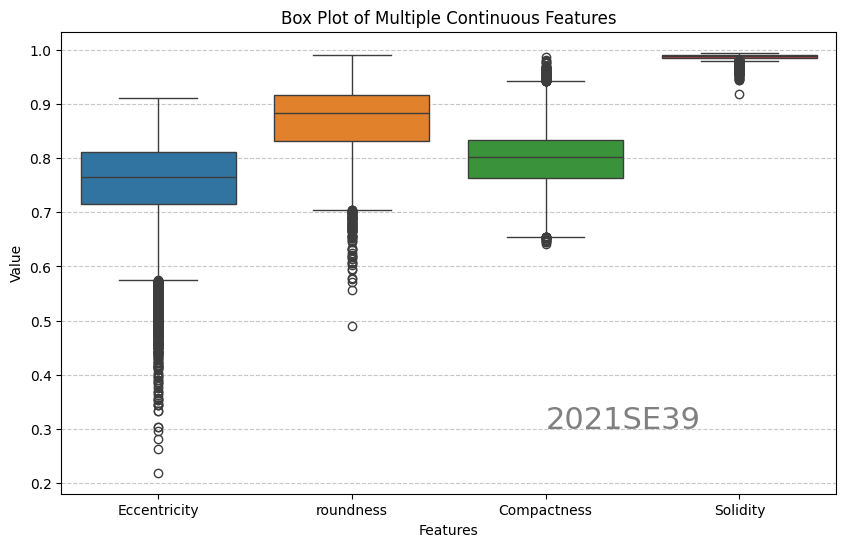
The line plot depicts the tangled relationship between Age and BMI (Body Mass Index). This relationship is nonlinear and moderate in strength. It exhibits a complex pattern, initially showing a positive association followed by a moderate negative trend. The BMI is plotted on y-axis and the Age is plotted on y-axis. The plot seems to have no outliers.

The graph reveals a wavy pattern instead of a straight line, indicating that the relationship between BMI and age is more complex than initially thought. It serves as a visual representation, highlighting how age influences BMI. Initially, there's a steady increase with some downwards in BMI from around age 11 to 35, spanning a range of approximately 20 to 32 BMI units. After the age 35 to onwards the BMI pattern has some sharp upwards and downwards. This analysis aims to uncover underlying patterns, offering insights into the complex interplay between age and BMI. Upon closer examination, fluctuations become apparent between ages 35 and 50, where BMI have consistent shifts. Subsequently, there's a slight decline in BMI from age 60 to onwards. Although this downward trend isn't particularly strong, there's considerable variation in BMI between ages 33 and 79, ranging from 26 to 32. Overall, there appears to be a moderate association between age and BMI, with sharp shifts showing that some other factors also affect the BMI with age.

In conclusion, the graph illustrates a fluctuating relationship between BMI and age, with BMI initially increasing until around age 33 at its peak. This indicates that BMI tends to increase from teenage to adult, followed by a gradual decrease from adult age to senior age.

**Multivariate Analysis**

1. **Box Plot for ‘Eccentricity, Roundness, Compactness and Solidity’**

****

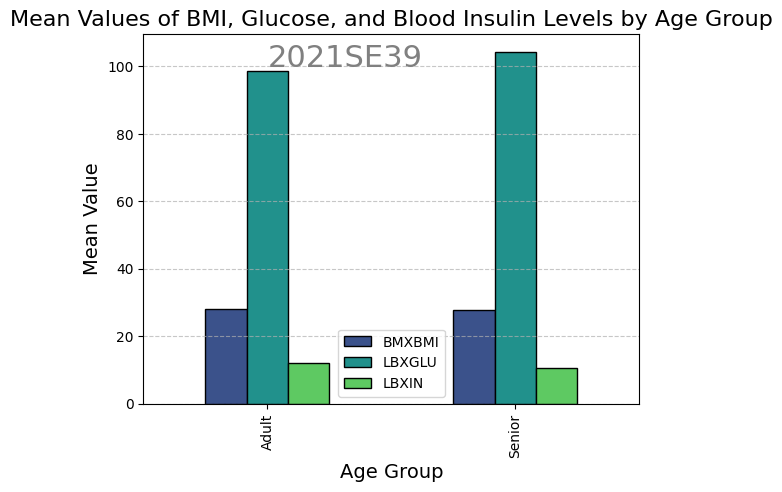
Plot 3.

In this data exploration, box plot of multiple continuous features of a dry bean dataset: eccentricity, roundness, compactness, solidity. Eccentricity has a streak of outliers below the minimum whisker and same in the case of roundness. Compactness has a few outliers below and some are above the maximum values. Solidity has nearly zero variance with a few outliers below the minimum.

This Box plots help show the middle, spread, and unusual values in the data visually. Examining the box plot for eccentricity, roundness, compactness, solidity, we find their distributions are skewed towards lower values, indicating more outliers on the lower side. Among them, ‘eccentricity’ has the widest range, followed by ‘roundness’ and ‘compactness’. By close look at the plot the solidity is bound between 0.98 with some outliers below. The median line of solidity lies outside of all the other variables interquartile range. Compactness has a covered the area from 0.65 to 0.94 leaving some outliers above and below. Its median line lies at the edge of the eccentricity interquartile range. The roundness interquartile range and the middle line is above from eccentricity and compactness but below than the solidity, leaving a few points behind the lower whisker. The eccentricity has wide spread from 0.58 to 0.91 having a stream of outliers below the lower whisker. The outliers have noticeable spread from 0.2 to the smallest value of eccentricity. The spread of eccentricity shows the unusual behavior of the ellipse having the same moments as the region of the bean. The roundness has a small negative skewness and the Solidity has all the interquartile, upper and lower bounds closely packed.

In conclusion, these box plots give us a clear picture of the data's central tendency, variability, and any unusual values. Looking at eccentricity, roundness, compactness, and solidity, we notice their distributions tend to skew towards lower values, indicating the presence of more outliers on the lower end. Among them, eccentricity stands out with the widest range, followed by roundness and compactness. Solidity has a very less variance and median line staying at the top of all other features.

1. **Bar Plot for average values of ‘BMI, Glucose and Insulin Level’**

****

Plot 4.

In this data analysis, a bar plot shows average values of key variables—BMI (Body Mass Index), Glucose and Insulin levels in the representative body—in the (NHANES) age prediction dataset. Among all the features Glucose has the highest mean value, the Insulin has the lowest value and the BMI stays in the middle of the glucose and insulin level in both age group.

Examining the variables reveals clear connections. The adults have a smaller glucose mean value then the senior aged. Meaning as age tend to increase, consequently, the amount of glucose increases. Similarly, there's a moderate positive correlation between BMI and Glucose level in the age group. The positive correlation between the glucose is pretty much clear but the BMI has nearly neglectable difference. Interestingly, insulin has a weak negative correlation, as the age group changes from adult to senior, hinting that higher glucose level, resulting in smaller insulin level. The difference in BMI and Glucose associated with each age group with respect to insulin level illustrate that the BMI and Glucose has a moderate negative correlation with insulin. The glucose mean value has an increment from 98 to 105 in the age group while the incline mean slightly decreases from 12 to 10 while moving in the age group from Adult to Senior.

Concluding the analysis, the bar plot efficiently conveys mean values, exposing correlations, trends and trend in age groups. The BMI level remains constant in age group, the glucose average has an increment from adult to senior while the insulin tends to slightly go down following the pattern from adult to senior.

**Classification Results**

**Dataset 1: Result**

|  |  |  |
| --- | --- | --- |
|  | KNN | Naïve Bayes |
| **Confusion Matrix** |  |  |
| **Accuracy** | **0.735** | **0.765** |
| **Precision** | **0.746** | **0.767** |
| **Recall** | **0.735** | **0.765** |
| **F1-Score** | **0.738** | **0.765** |

Dataset used for this classification results is **dry\_bean\_dataset.csv**. The dataset has a total of 17 column having 16 continuous type feature columns and one target column. The target column has 7 classes named as: BARBUNYA, BOMBAY, CALI, DERMASON, HOROZ, SEKER, SIRA. The dataset has a total of 13612 rows in which 20% of data is used for testing purpose.

By training kNN and Naïve Bayes on dataset, the Naïve Bayes shows an accuracy of 0.765 and the kNN shows the accuracy of 0.735. **The Naïve Bayes has the highest accuracy and precision, recall and f1-score**. The confusion matrix has 7 rows and column indicating the 7 classes in dataset. The completed report of both models is as follow:

|  |  |
| --- | --- |
| KNN | Naïve Bayes |
|  |  |

**Dataset 2: Result**

|  |  |  |
| --- | --- | --- |
|  | KNN | Naïve Bayes |
| **Confusion Matrix** |  |  |
| **Accuracy** | **0.980** | **0.958** |
| **Precision** | **0.966** | **0.903** |
| **Recall** | **0.946** | **0.919** |
| **F1-Score** | **0.956** | **0.911** |

Dataset used for this classification results is **NHANES\_age\_prediction.csv**. The dataset has a total of 10 column having 8 feature columns and one target column with one Id column. The target column has only 2 classes named as: ‘Adult’ and ‘Senior’. The dataset has a total of 2279 rows in which 20% of data is used for testing purpose.

By training kNN and Naïve Bayes on this dataset, surprisingly both the models show an accuracy above 95%. The Naïve Bayes shows an accuracy of 0.958 and the kNN shows the accuracy of 0.980. **The kNN has the highest accuracy and precision, recall and f1-score**. The confusion matrix has 2 rows and column indicating the 2 classes in dataset. The completed report of both models is as follow:

|  |  |
| --- | --- |
| KNN | Naïve Bayes |
|  |  |

**\_\_\_\_\_\_\_\_\_\_\_\_**