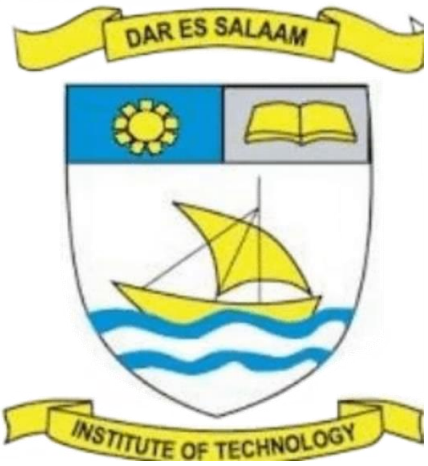
DAR ES SALAAM INSTITUTE OF TECHNOLOGY



DEPARTMENT OF COMPUTER STUDIES

BACHELOR OF ENGINEERING IN COMPUTER ENGINEERING

DATA MINING – COU 08104

ASSIGNMENT 02

NAME: ALOYCE VENANCE TUNGARAZA

REG NO: 220242425759

1. Data Transformation and Normalization

Min–Max Normalization:
Min–max normalization rescales data values into a fixed range, usually between 0 and 1. It preserves relationships among original values but is sensitive to outliers.

Z-Score Normalization:
Z-score normalization standardizes data using the mean and standard deviation. It is robust to outliers and useful when min and max values are unknown.

Decimal Scaling Normalization:
Decimal scaling normalizes data by moving the decimal point based on the maximum absolute value.

Comparison:
Min–max is simple but sensitive to outliers, z-score is statistically robust, and decimal scaling is computationally simple but less precise.

2. Dataset Understanding

Titanic Dataset:
The Titanic dataset contains passenger survival information based on class, age, sex, and family relationships, and is commonly used for classification tasks.

Diabetes Dataset:
The Diabetes dataset contains medical attributes used to predict diabetes, widely applied in healthcare analytics.

3. Stock Market Analysis and PCA

Why Use Returns Instead of Prices:
Returns are scale-independent and stationary, making them suitable for statistical analysis.

Why Use Correlation Matrix:
The correlation matrix removes scale effects, allowing fair PCA computation.

PCA Interpretation:
PCA reduces dimensionality by transforming correlated variables into uncorrelated components.

Explained Variance:
Explained variance shows how much information each principal component retains.

Conclusion

Normalization ensures fair feature comparison, while PCA simplifies complex datasets without significant information loss. These techniques are fundamental in data mining.