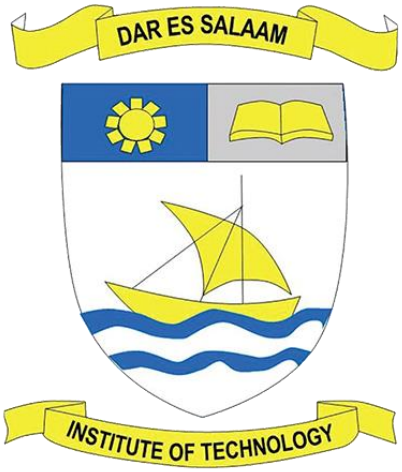
DAR ES SALAAM INSTITUTE OF TECHNOLOGY



DATA MINING AND ANALYSIS

STUDENT'S NAME: NELSON S. NYAMWERU

ADMISSION NO: 220242486009

CLASS: BENG22COE 1

INDIVIDUAL ASSIGNMENT: ASSIGNMENT 02

ANSWERS

1. Diabetes Data

1.1 Correlation Matrix and Heatmap:
The correlation heatmap shows strong correlations among several blood serum measurements (S1–S6). For example, S1 and S2 are highly positively correlated, indicating related biological processes. BMI is moderately correlated with BP and some serum measures. AGE and SEX show weak correlations with most variables. Strong correlations suggest multicollinearity.

1.2 Collinearity:
Collinearity occurs when predictor variables are highly correlated. It inflates standard errors, makes coefficient estimates unstable, and reduces interpretability, though prediction accuracy may remain acceptable.

1.3 Multivariate Linear Model:

A linear regression using all predictors estimates disease progression. MSE measures prediction error, and adjusted R^2 reflects explained variance adjusted for complexity. Some variables are not statistically significant due to multicollinearity.

1.4 Forward vs Backward Selection:

Forward selection starts with no predictors and adds variables incrementally. Backward selection starts with all predictors and removes the least significant variables.

1.5 Stepwise Selection:
Stepwise selection combines forward and backward methods. Variables are added or removed based on significance. Selected variables include AGE, SEX, BMI, BP, S1, and S2. This yields lower MSE and similar R^2 compared to the full model.

2. Titanic Dataset

2.1 Logistic vs Linear Regression:

Linear regression predicts continuous outcomes. Logistic regression predicts probabilities for binary outcomes using a logistic function.

2.2 Survival Probability:
The probability of survival equals the proportion of passengers who survived.

2.3 Survival by Group:
Females and first-class passengers have higher survival probabilities. Younger passengers also show improved survival outcomes.

2.4 Logistic Regression Model:
Passenger class and sex are highly significant predictors of survival, while age has a moderate effect. Lower class number, being female, and younger age increase survival probability.

2.5 Model Performance:
Classification accuracy from the confusion matrix measures the proportion of correctly classified passengers and indicates reasonable model performance.

3. PCA

3.1 PCA Description:

PCA reduces dimensionality by transforming correlated variables into uncorrelated components. It aids visualization, noise reduction, and handling multicollinearity.

3.2 PCA Mathematics:
PCA decomposes centered data X into eigenvectors and eigenvalues of its covariance or correlation matrix. Projections onto eigenvectors give principal components.

3.3 Dow Jones PCA:
The first principal component resembles the market factor with near-equal weights. The second captures sectoral differences.

3.4 Variance Explained:
A scree plot shows diminishing variance contributions. A limited number of components explain 95% of total variance.

3.5 Unusual Stocks:
The most distant stocks differ due to sector exposure, volatility, or firm-specific risk.