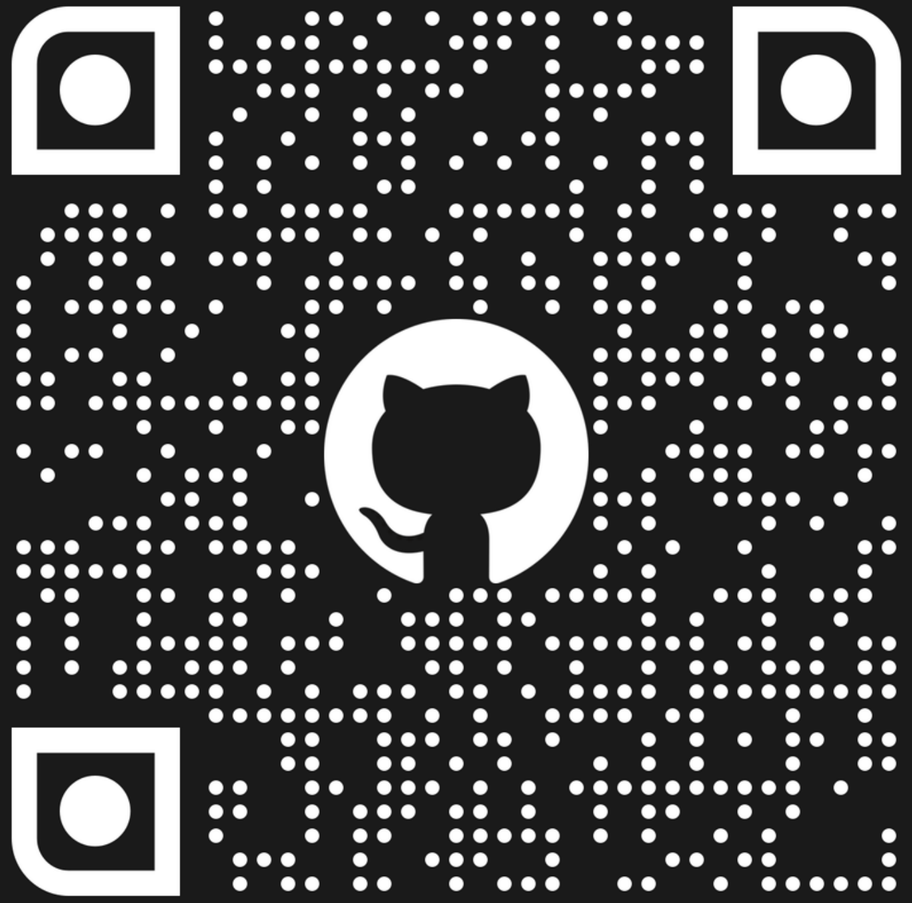
MACHINE LEARNING

Group Project

MYTHE

CONTENT

1



OR

Click Here

2

NUMERICAL DATASET

Life Expectancy

• BRIEF

THIS REPORT SHOWCASES A MACHINE LEARNING MODEL DESIGNED TO FORECAST **LIFE EXPECTANCY.** THE PREDICTIVE MODEL EMPLOYS BOTH **LINEAR REGRESSION** AND **KNN** ALGORITHMS FOR THE TASK

• DATASET DESCRIPTION

This study aims to address crucial gaps in past life expectancy research by considering previously overlooked factors such as immunization and human development index. Unlike prior studies that focused mainly on demographic and income-related variables, this research emphasizes a broader range of factors, spanning immunization, mortality, economy, and social aspects.
The primary objective is to leverage machine learning models, including mixed effects and multiple linear regression, to identify factors contributing to variations in life expectancy. By analyzing these diverse factors, the study aims to assist countries in pinpointing areas of focus to effectively enhance their population's life expectancy.

• DATASET DESCRIPTION

Before Preprocessing

SIZE OF DATA : 333.44 KB
NUMBER OF FEATURES : 22 FEATURE
NUMBER OF SAMPLES : 2937 SAMPLE

FEATURES ILLUSTRATION

FEATURES ILLUSTRATION

- **Country:** The Name of The Country
- **Year:** The Year of Study
- **Status:** Developed or Developing Country status
- **Life Expectancy:** Life Expectancy in Age
- **Adult Mortality:** Adult Mortality Rates of both Sexes
- **Infant Deaths:** No. of Infant Deaths per 1000 population
- **Alcohol:** Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
- **Percentage Expenditure:** Expenditure on health as a percentage of Gross Domestic Product per capita(%)

FEATURES ILLUSTRATION

- **Hepatitis B:** Hepatitis B immunization coverage among 1-year-olds (%)
- **Measles:** Measles - number of reported cases per 1000 population
- **BMI:** Average Body Mass Index of entire population
- **Under-five deaths:** Number of under-five deaths per 1000 population
- **Polio:** Polio (Pol3) immunization coverage among 1-year-olds (%)

FEATURES ILLUSTRATION

- **Total expenditure:** General government expenditure on health as a percentage of total government expenditure
- **Diphtheria:** Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
- **HIV/AIDS:** Deaths per 1 000 live births HIV/AIDS (0-4 years)
- **GDB:** Gross Domestic Product per capita (in USD)
- **Population :** Population of the country
- **Thinness 1-19 years:** Prevalence of thinness among children and adolescents for Age 10 to 19 (%)

FEATURES ILLUSTRATION

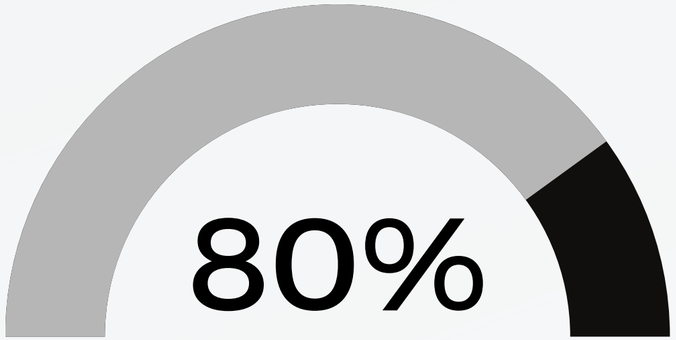
- **thinness 5-9 years:** Prevalence of thinness among children for Age 5 to 9(%)
- **Income composition of resources:** Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- **Schooling :** Number of years of Schooling(years)

DATA TYPES

float64(16), int64(4), object(2)

IMPLEMENTATION DETAILS

DATA SPLITTING



TRAIN

TEST

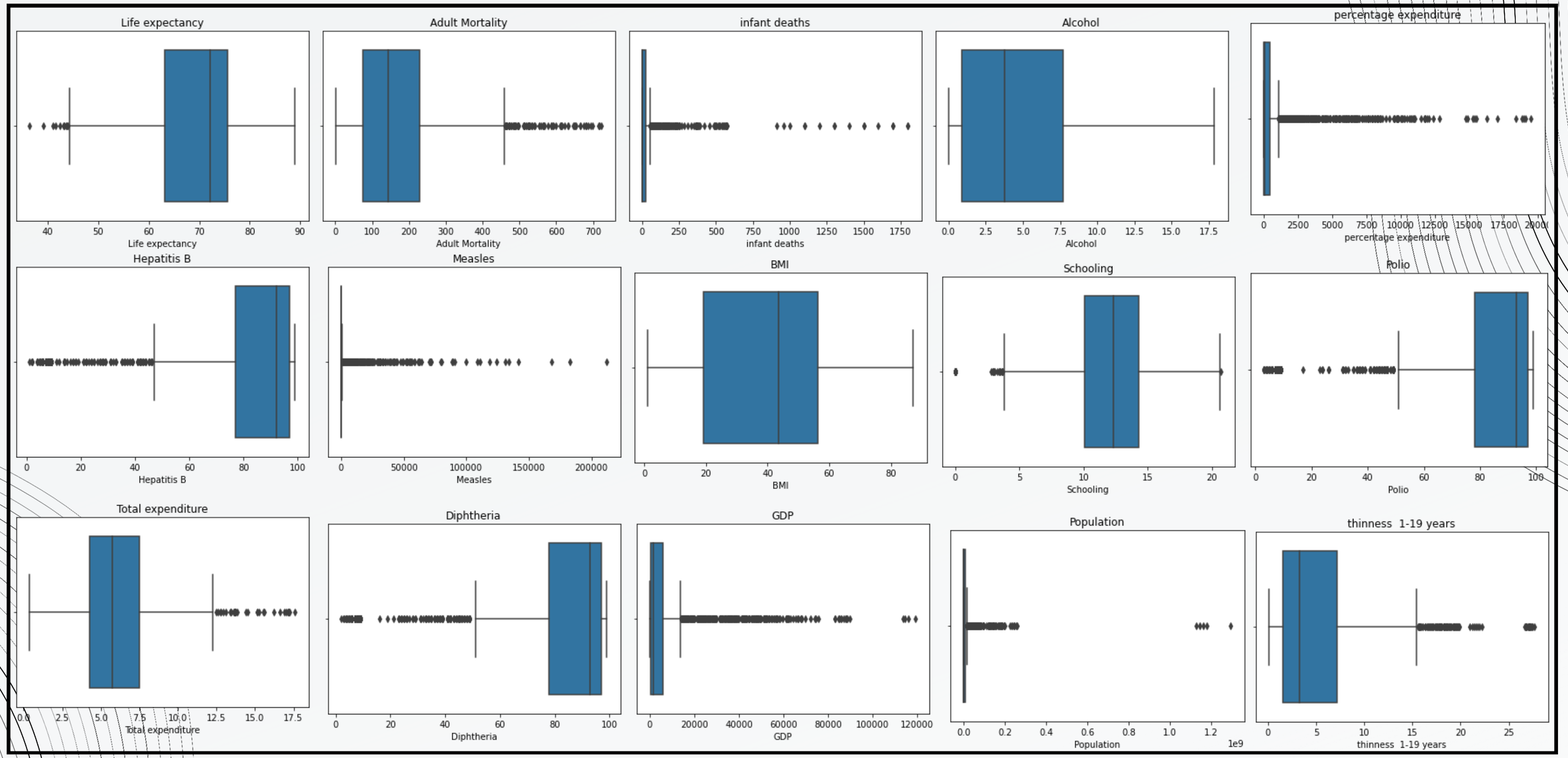
ANALYZING

- Data Import and Preprocessing:
  - Reads and explores the dataset using Pandas, including examining data samples, information, and statistical summaries.
  - Handles null values, strips column names, and visualizes distributions and outliers in numerical features.
  - Explores categorical features with bar and pie charts, depicting counts and proportions.
- Exploratory Data Analysis (EDA):
  - Investigates relationships between features and the target variable (life expectancy).
  - Visualizes temporal trends in life expectancy and infant deaths over the years.
  - Studies correlations among numerical features and their impact on life expectancy.

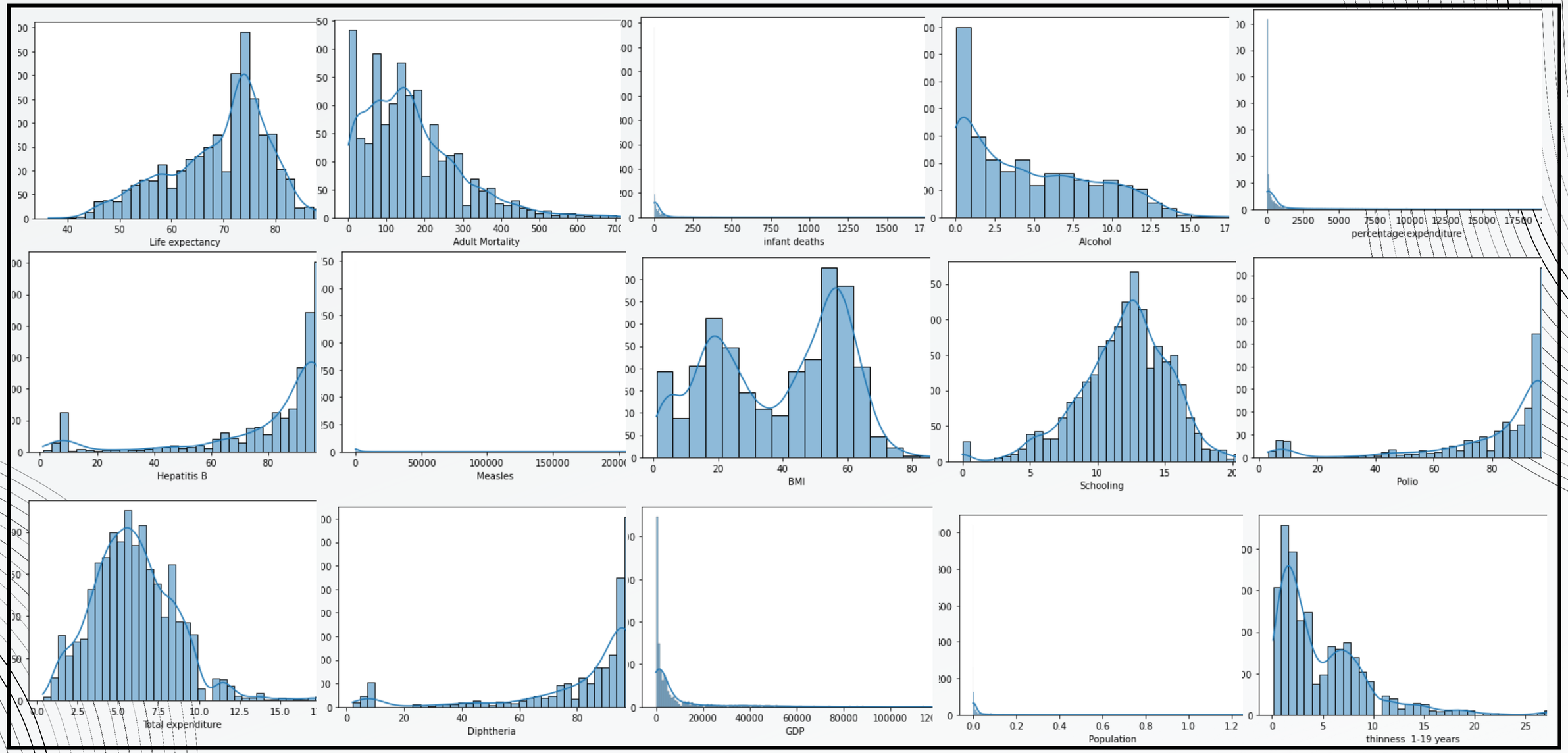
ANALYZING

- Data Preprocessing and Feature Engineering:
  - Addresses null values by imputing with median values.
  - Handles outliers using an Interquartile Range (IQR) method.
  - Drops specific columns and encodes categorical features for modeling.
- Modeling:
  - Splits data into training and testing sets.
  - Standardizes numerical features.
  - Applies Linear Regression for predicting life expectancy, evaluates model performance, and visualizes predictions against actual values.
  - Determines feature importance using coefficient analysis and visualizes the linear regression feature importance.
- Model Evaluation and Improvement:
  - Assesses model performance using Root Mean Squared Error (RMSE), R-squared score, and cross-validation.
  - Implements K-Nearest Neighbors (KNN) Regression, optimizing for the best 'k' value using GridSearchCV, and evaluates its performance.

OUTLIERS Before...

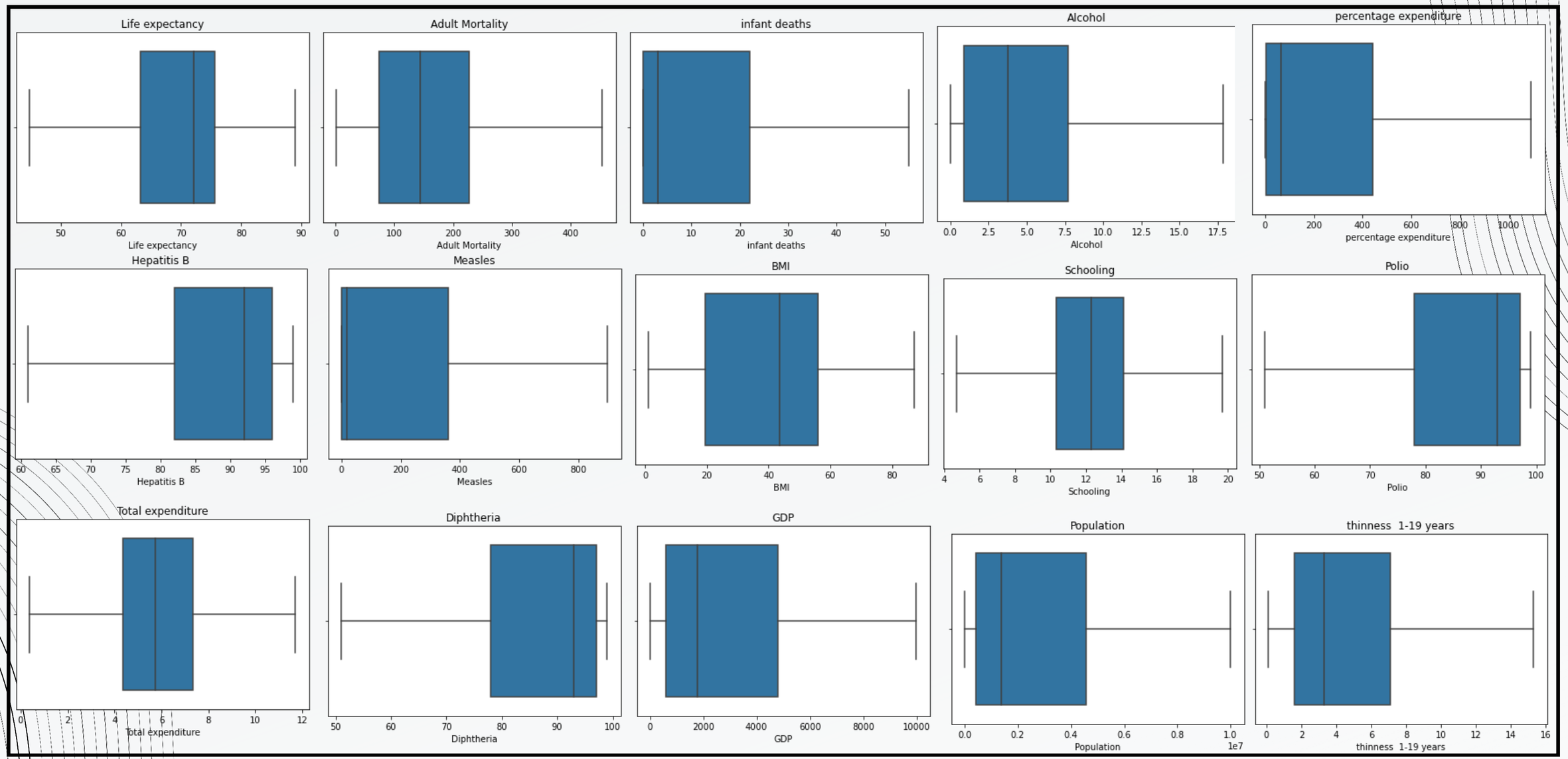


DISTRIBUTION Before...

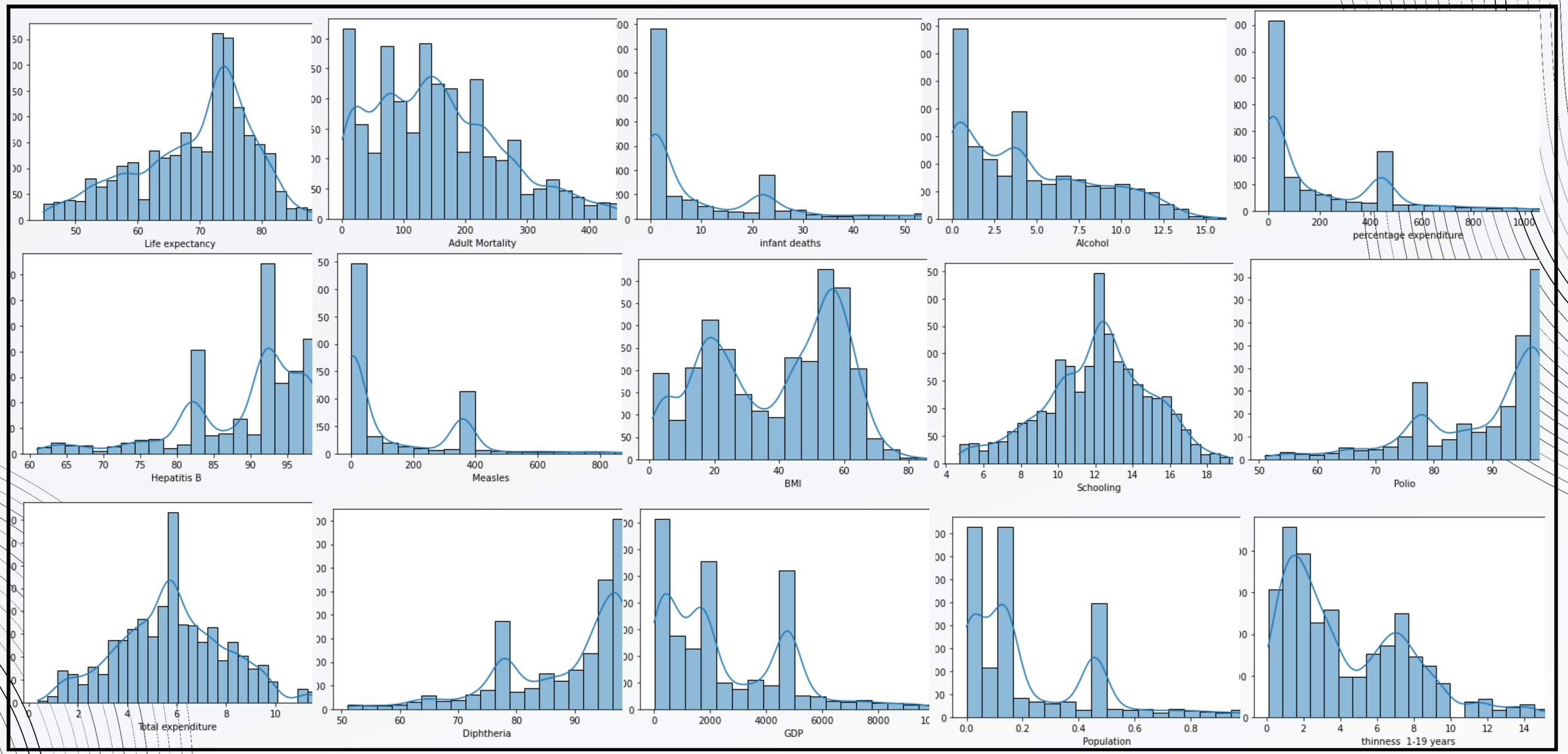


RESULT DETAILS

OUTLIERS After...



DISTRIBUTION After...



LINEAR REGRESSION

- Train accuracy 79%
- Test accuracy 80%
- RMSE for training : 4.23
- RMSE for testing :4.16
- R2_score 88%

CROSS VALIDATION

- Cross Validation score : 78.4%
- The training/validation ratio is 4:1 for each fold

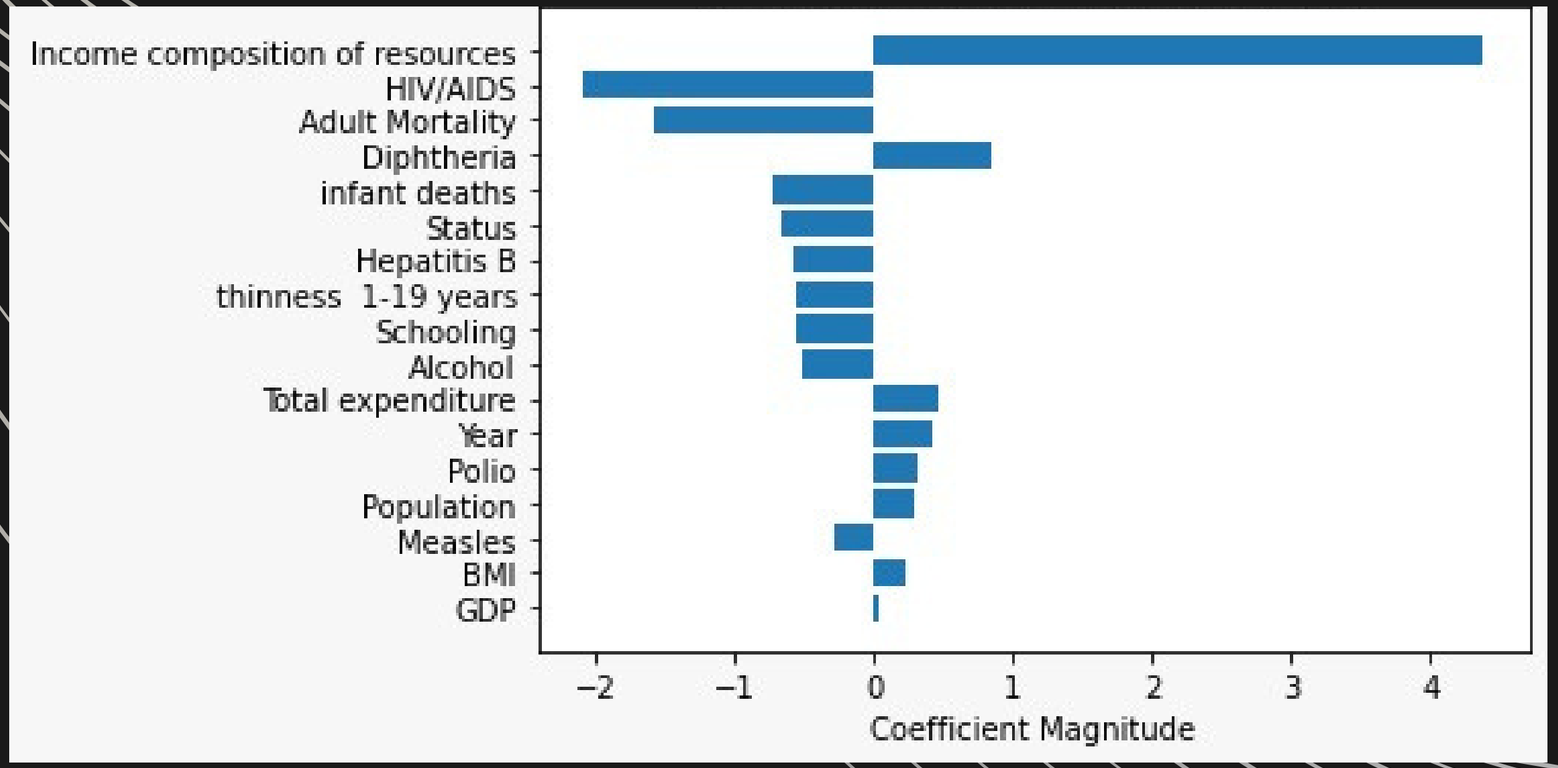
KNN

- Train accuracy 94%
- Test accuracy 88%
- RMSE for training : 2.21
- RMSE for testing : 3.23
- n_neighbors : 3

CROSS VALIDATION

- Cross Validation score : 86%
- The training/validation ratio is 4:1 for each fold

FEATURES IMPORTANCE



3

IMAGE DATASET

Stanford Dogs Dataset

• BRIEF

THIS SECTION SHOWCASES A MACHINE LEARNING MODEL DESIGNED TO FORECAST **DIVERSE IMAGES OF DOGS ACROSS 120 BREEDS.** THE PREDICTIVE MODEL EMPLOYS BOTH **LOGISTIC REGRESSION** AND **KMEANS** ALGORITHMS FOR THE TASK

• DATASET DESCRIPTION

The Stanford Dogs dataset contains images of 120 breeds of dogs from around the world. This dataset has been built using images and annotation from ImageNet for the task of fine-grained image categorization. It was originally collected for fine-grain image categorization, a challenging problem as certain dog breeds have near identical features or differ in colour and age.

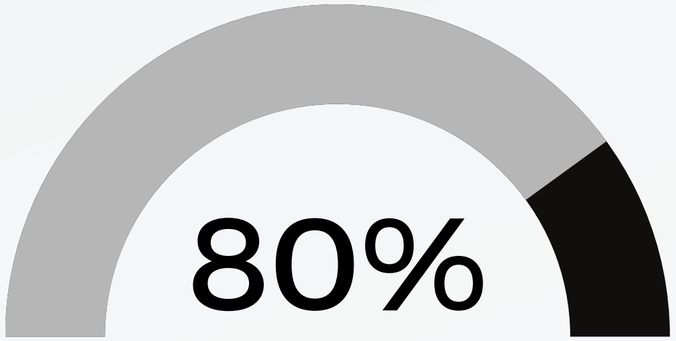
• DATASET DESCRIPTION

Before Preprocessing

SIZE OF DATA : 788.05 MB
NUMBER OF CLASSES: 33 CLASS
NUMBER OF IMAGES: 20,580

IMPLEMENTATION DETAILS

DATA SPLITTING



TRAIN

TEST

ANALYZING

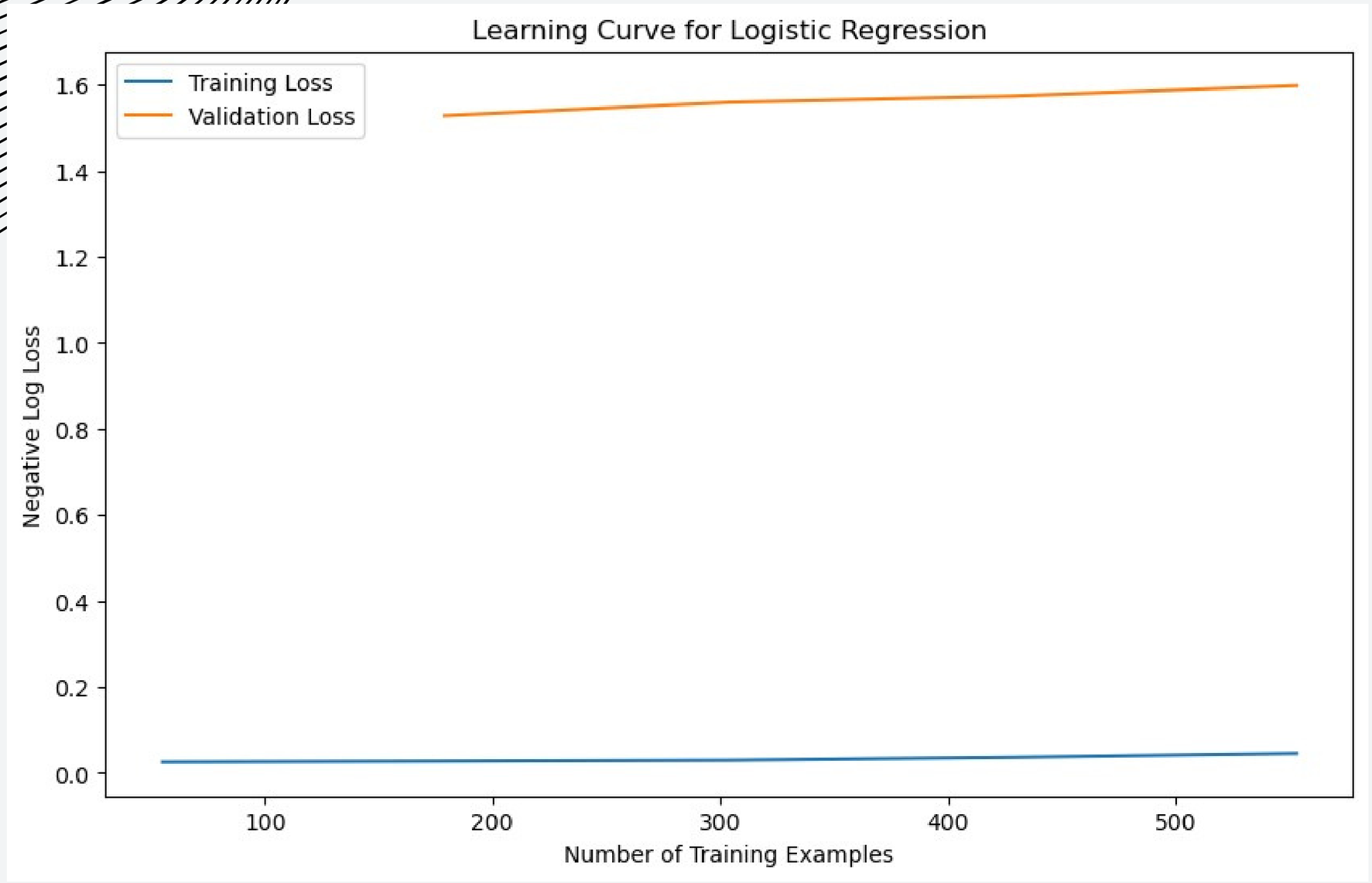
1. Data Preparation: The code reads images and annotations from specified directories, extracts bounding box coordinates, and preprocesses the images for analysis.
2. K-Means Clustering: Utilizes K-Means clustering to cluster flattened image arrays based on pixel values, generating clusters and visualizes them using Principal Component Analysis (PCA).
3. SIFT Feature Extraction: Extracts SIFT (Scale-Invariant Feature Transform) features from the images, displays original and resized images with detected SIFT features, and performs clustering based on SIFT descriptors.

ANALYZING

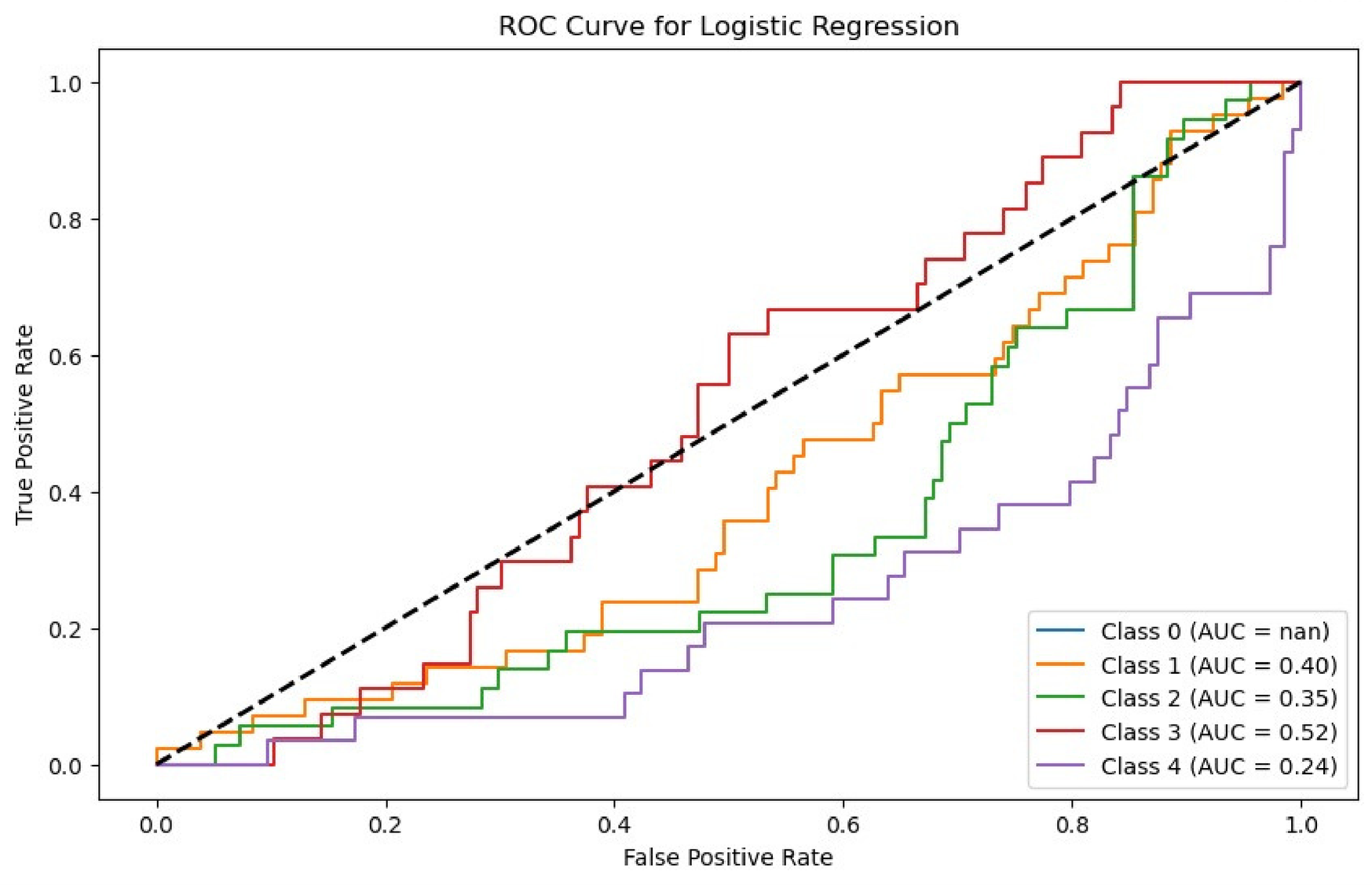
1. SIFT Features Analysis: Processes the SIFT descriptors, pads them for uniformity, performs K-Means clustering on the modified descriptors, and displays images from different clusters to visualize grouping.
2. Clustering Evaluation: Evaluates clusters' characteristics, including the number of images in each cluster, using indices and visualizations.
3. Elbow Method and PCA: Employs the Elbow Method to determine an optimal number of clusters, then applies PCA to visualize the clusters in a 2D space.

RESULT DETAILS

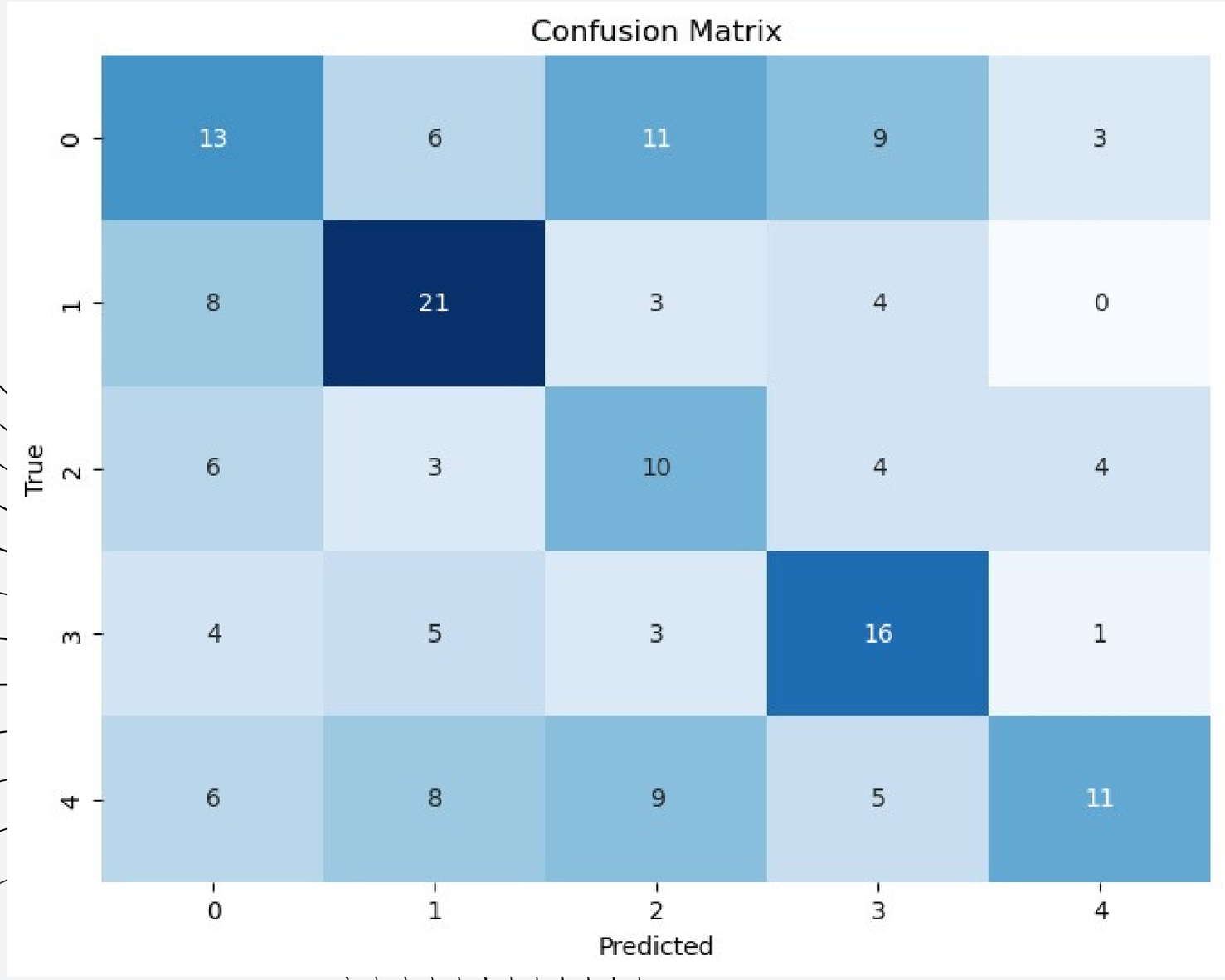
LOSS CURVE


Learning Curve for Logistic Regression
Training Loss
Validation Loss
Negative Log Loss
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
100
200
300
400
500
Number of Training Examples

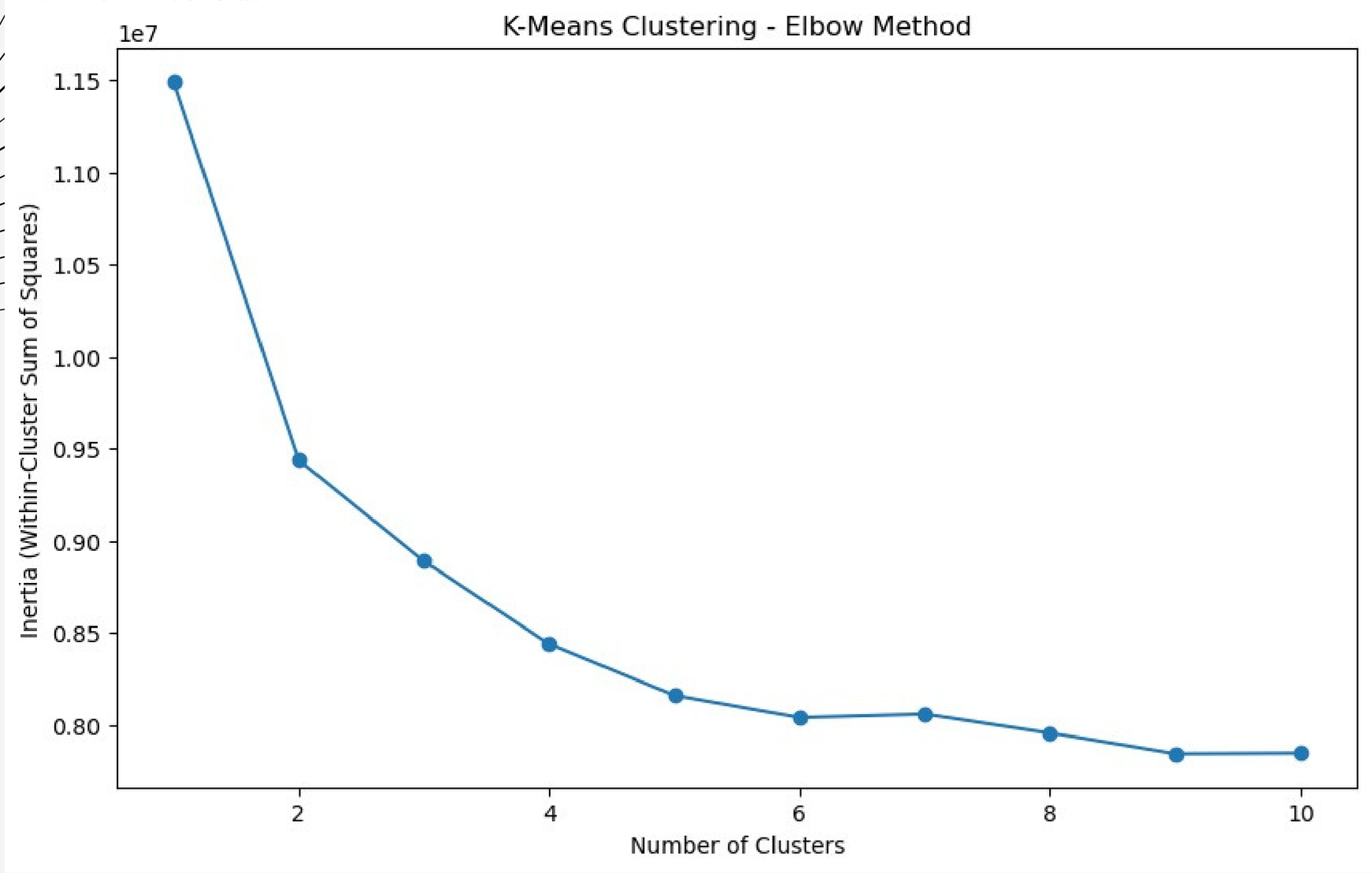
ROC CURVE


ROC Curve for Logistic Regression
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = nan)
Class 1 (AUC = 0.40)
Class 2 (AUC = 0.35)
Class 3 (AUC = 0.52)
Class 4 (AUC = 0.24)

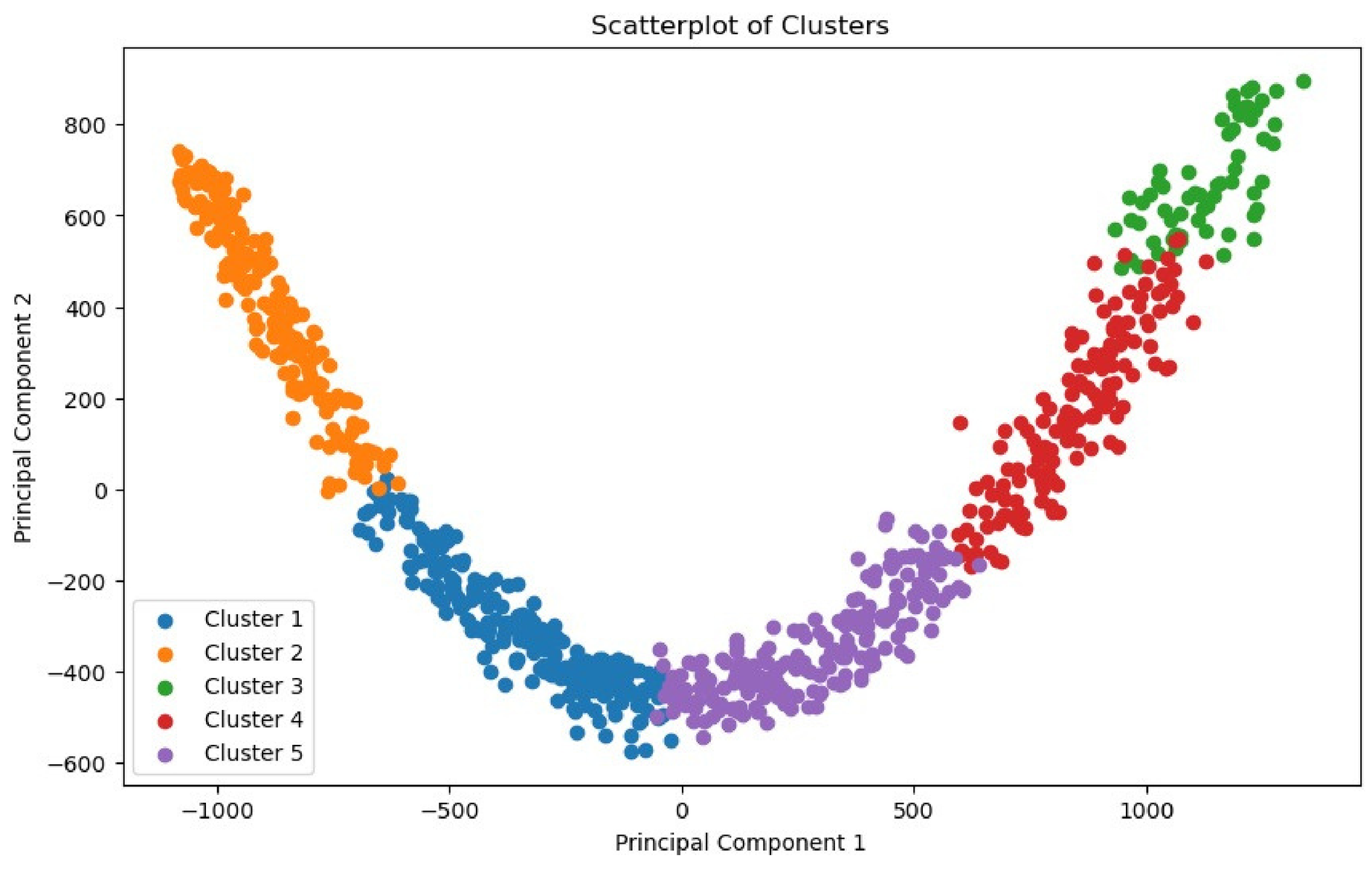
CONFUSSION MATRIX



ELBOW METHOD



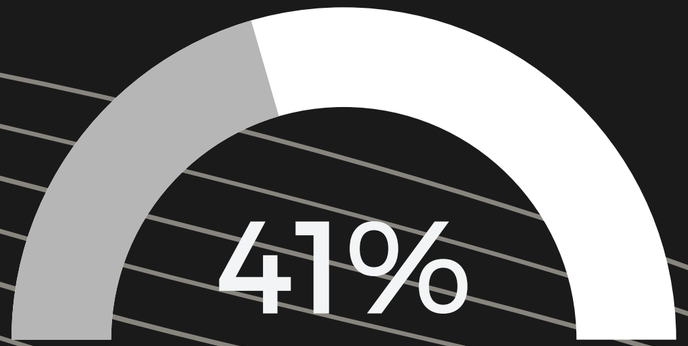
SCATTER PLOT



AFTER CLUSTRING

ACCURACY

LOGISTIC REGRISSION



OUR TEAM

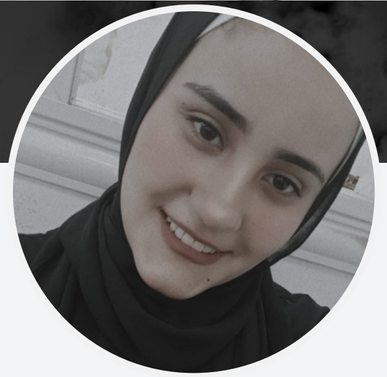
MUHAMMAD YASSER

MOHAMMAD TAREK

MOHAMED HUSSIAN

ASSER HASSAN

MADIHA SAEID



ESRAA MOHAMED

HANIA RUBY

THANKS