# MSc. Data Science & Analytics Research Scholar Test.

Mboya Jackline Achieng

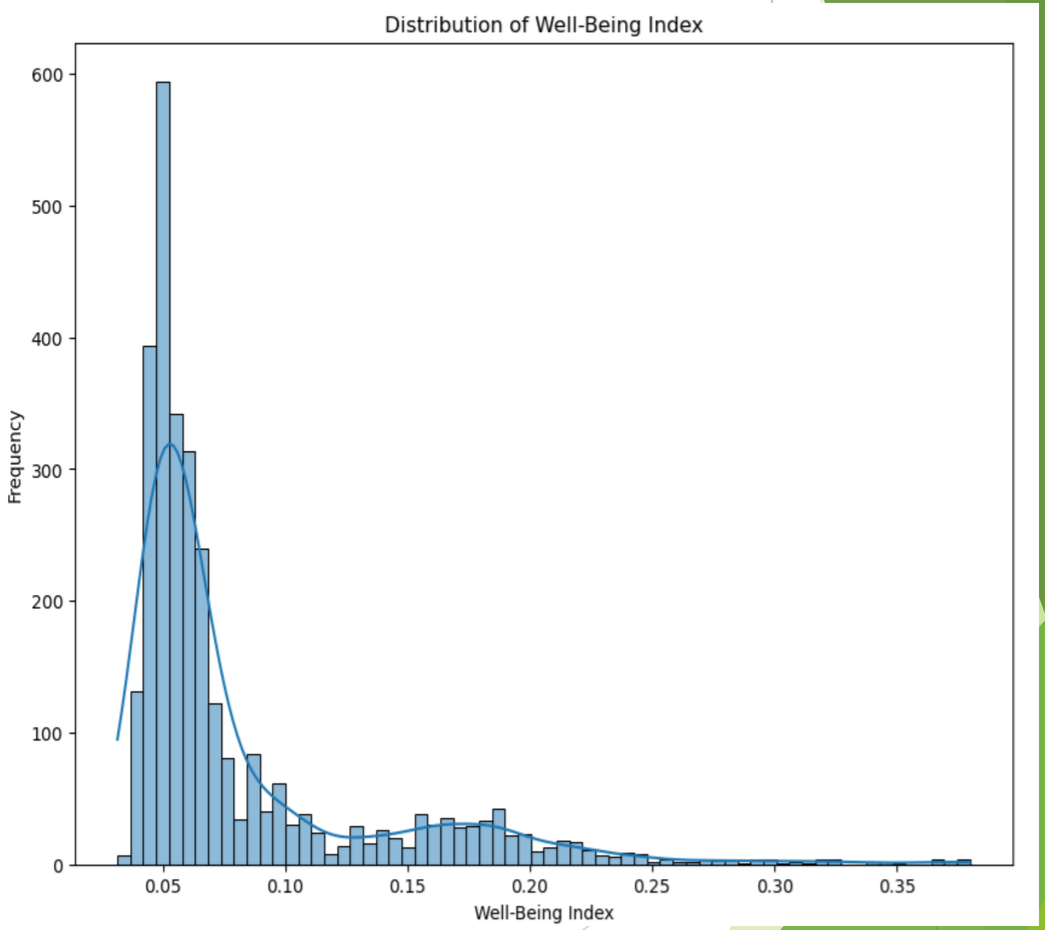
Demographic and Socio-Economic Determinants of Galaxies' Well-being

# INTRODUCTION

- **Data Description**:
  - The dataset contains information about 181 galaxies over a period of at most 26 years.
  - Each galaxy has 80 demographic and socio-economic variables.
  - The composite index is used to measure the well-being of each galaxy.
- **Objective**:
  - Identify variables that best explain the variance in the well-being index.
  - Predict future well-being values of the galaxies.

# Exploratory Data Analysis (EDA)

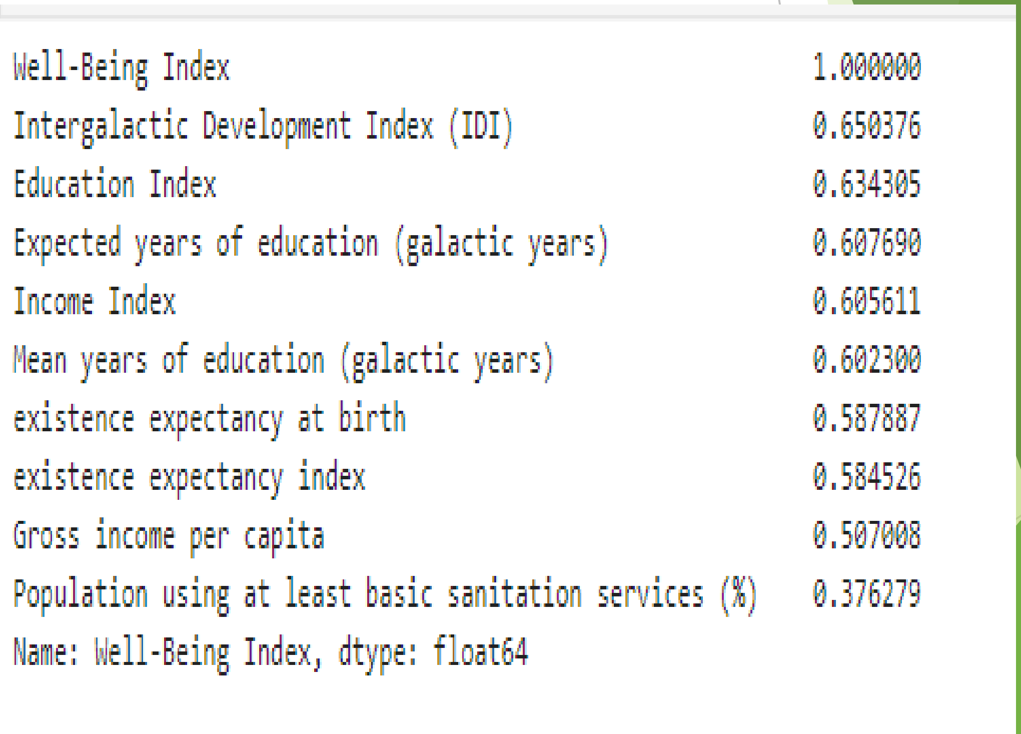
- **Dataset Overview**:
  - **Dimensions**: 3097 rows and 81 columns.
  - **Summary statistics**:
    - Mean, standard deviation, minimum, and maximum values for each variable.
    - Percentiles (25th, 50th, 75th) for understanding data distribution.
- **Key Insights**:
  - The dataset has missing values.
  - Significant right-skewness in the well-being index, indicating outliers or anomalies.



# Feature Selection

- **Identifying Key Variables**
  - **Feature Importance using Random Forest Regressor**:
    - Evaluated the contribution of each variable.
    - Important features include well-being index.
- **Dimensionality Reduction**:
  - Selected top variables depending on importance scores for model training.

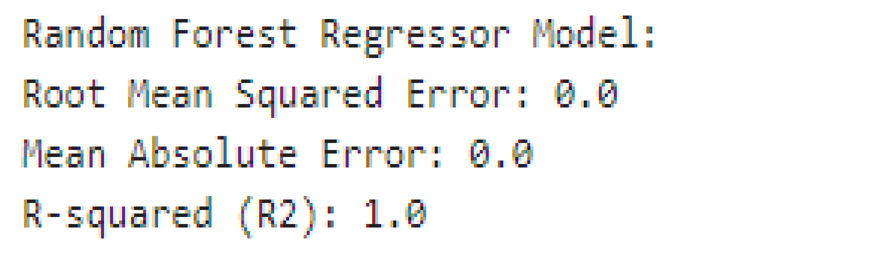
- **Top variables**



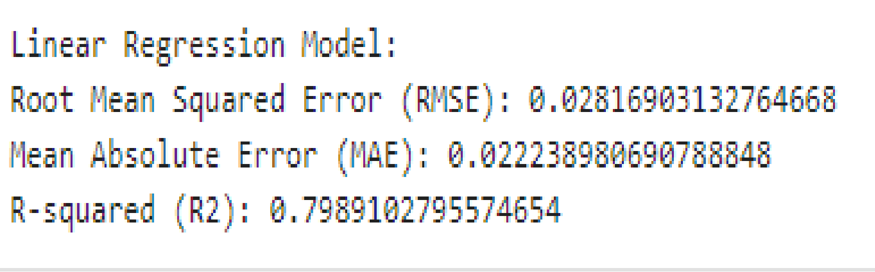
```
Well-Being Index                                      1.000000
Intergalactic Development Index (IDI)                 0.650376
Education Index                                       0.634305
Expected years of education (galactic years)          0.607690
Income Index                                          0.605611
Mean years of education (galactic years)              0.602300
existence expectancy at birth                         0.587887
existence expectancy index                            0.584526
Gross income per capita                               0.507008
Population using at least basic sanitation services (%)  0.376279
Name: Well-Being Index, dtype: float64
```

# Model Training and Evaluation

- **Random Forest Regressor**:
  - Performance Metrics: RMSE: 0.0; MAE: 0.0; R²: 1.0
  - **Insights**:
    - Perfect scores indicate potential overfitting.
    - Model might have learned noise and patterns too well.
- **Linear Regression Model**:
  - Performance Metrics: RMSE: Reasonable values; MAE: Lower values are better; R²: 0.7984
  - **Insights**:
    - Indicates good but realistic and generalizable performance.
    - 79.84% variance explained by the model.



```
Random Forest Regressor Model:
Root Mean Squared Error: 0.0
Mean Absolute Error: 0.0
R-squared (R2): 1.0
```



```
Linear Regression Model:
Root Mean Squared Error (RMSE): 0.02816903132764668
Mean Absolute Error (MAE): 0.022238980690788848
R-squared (R2): 0.7989102795574654
```

# Conclusions and Future Work

- **Key Findings**:
  - Identified critical variables impacting the well-being of galaxies.
  - Random Forest model overfitted the data;
  - Linear Regression provided more realistic and generalizable results.
  - Model predictions aligned well with actual values.
- **Future Work**:
  - Address overfitting in complex models.
  - Explore additional variables and modeling techniques.
  - Continuous validation with new data to improve prediction accuracy.