Galaxies Case Study Summary Report

By Wangari Kimani

Understanding the Data

- The data contains 80 variables that characterize the demographic and socio-economic situation of 181 galaxies over a period of at most 26 years. A composite index is given that measures their well-being.
- However, the demographic and socio-economic variables that influence this index is not known. We seek to determine, what makes the galaxies better off?
- The train data had 3097 rows and 80 columns.
- I started by checking for the data relevance by checking the columns with null values and selecting the columns which contained missing values that were less than 30 %, which were only 12 columns. The rest of the columns had too many missing values which would not really be relevant when trying to fit our models because they contained very little information.

Cleaning and Analyzing the Data

Cleaning the Data

- The data contained missing values of which I handled through:
- Replacing with median.
- Using the forward fill method
- Applying the knn imputer

Analyzing the Data

- I performed a multivariate analysis by getting the correlation of the features against the well being index.
- Out of the 12 columns, 9 columns have a correlation value above 0.5 and Intergalactic Development Index (IDI), Rank has the highest correlation of 0.705557

Which variables best explain the variance of the well-being index?

I compared the variances of the features against that of the well being index and obtained the following results.

Features	Variances
galactic year	2.10050000e+04
existence expectancy index	2.70743299e-02
existence expectancy at birth	6.53498434e+01
Gross income per capita	1.32988449e+05
Income Index	3.77349585e-02

The features shown below proved to be the ones that best explain the variance of the well being index.

Features	Variances
Expected years of education (galactic years)	1.28311850e+01
Mean years of education (galactic years)	1.00929343e+01
Intergalactic Development Index (IDI)	2.95727226e-02
Education Index	3.75293975e-02
Intergalactic Development Index (IDI), Rank	2.68860707e+02

Modeling

- I decided to use 3 models, which I used to compare their respective RMSE scores and used the one with the lowest RMSE score to predict the y variable, Predicted well-Being index in the validation data.

1. Linear Regression model: RMSE - 0.03842908220489801
2. Lasso Regression model: RMSE - 0.04838327257109511
3. Ridge Regression model: RMSE - 0.03702374311033586

<u>Conclusion</u>

From the above RMSE scores, ridge model has the lowest one, meaning it is the most accurate model, which I applied in predicting the well being index values in the validation data.