BOOTCAMP
DATA
ANALYTICS

LIFE EXPECTANCY

FINAL PROJECT

.https://www.linkedin.com/in/flor-garcia-ortiz/

OBJECTIVE

Develop a model to predict the life expectancy of a person, based on given parameters of the provided dataset from WHO.

1 Data

2 Models

3 Results

1 Data

SOURCE

- Dataset from WHO available in Kaggle
- https://www.kaggle.com/code/wrecked22/life-expectancy-regression/data

EXPLORATION

- NaN
- Outliers ?
- Variables:

1. 'country' - High cardinality
2. 'year' - ¿relevant?

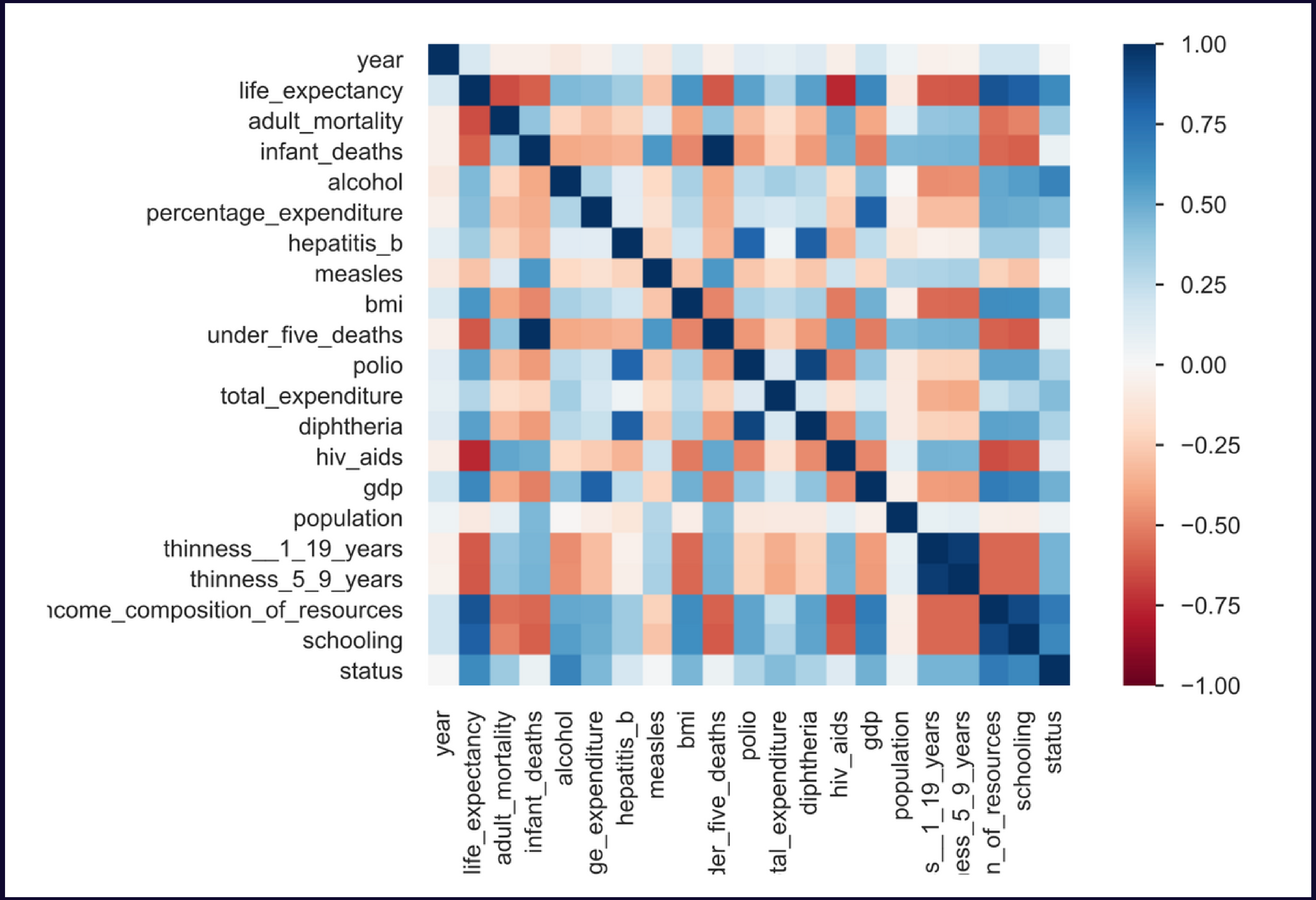
CLEANING

- Removed or transformed NaN
- Outliers affect data distribution significantly?
- Means in the case of 'country' and 'year' are equal? - ANOVA test
- What is the best way to deal with high cardinality for 'country'?

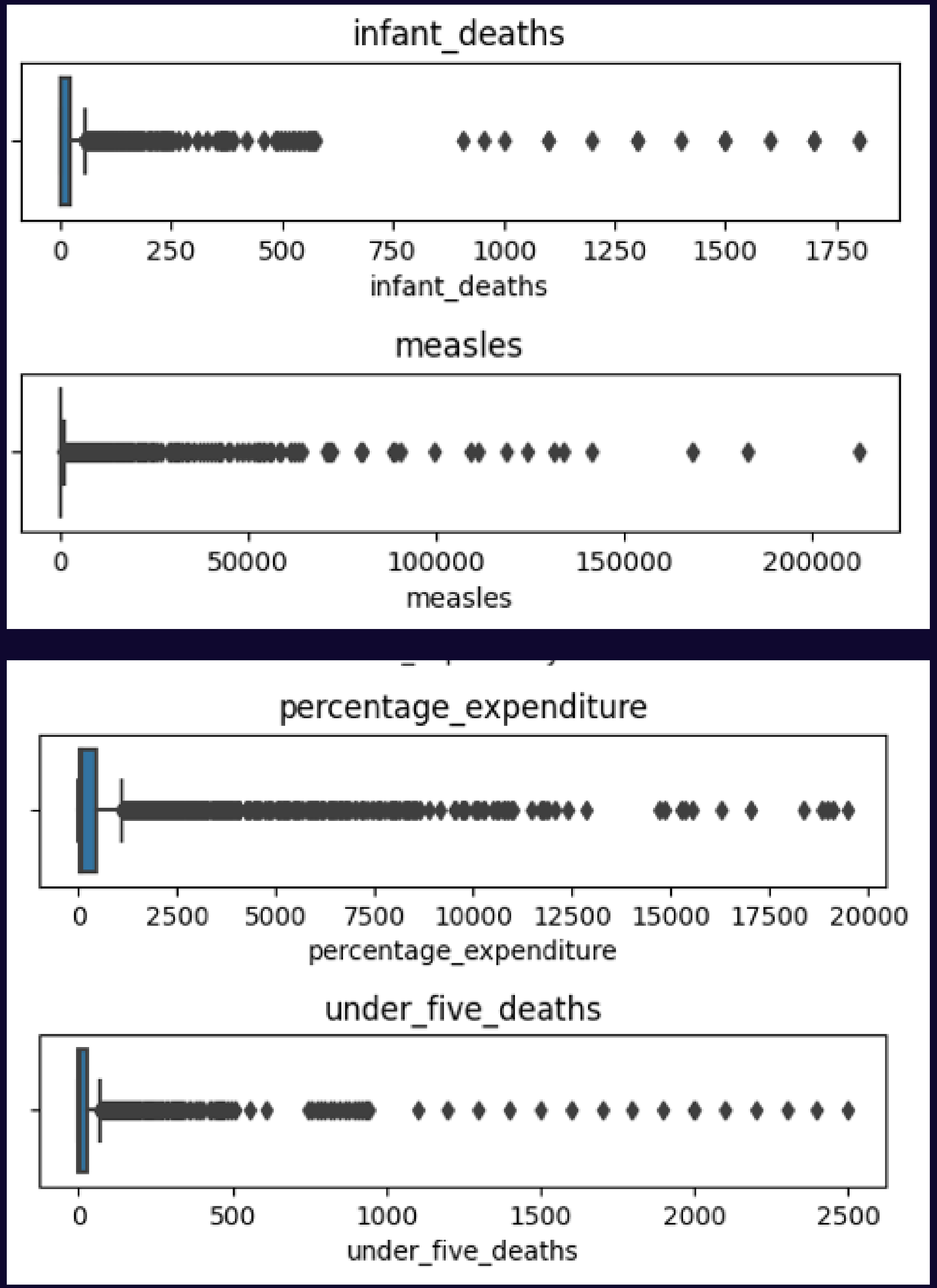
DATA PROCESSING

2938 rows
22 columns

CORRELATION MATRIX



BOX PLOT



2 Models

DATA1

1. NaNs are removed.
2. Outliers are removed.
3. 'Year variable is removed.

1444 ROWS/21 COL

DATA2

1. Countries are transformed into an ordinal categorical variable: country_rank
2. NaNs are treated with KNN.
3. Outliers are removed.

2650 ROWS/22 COL

DATA3

1. Countries are transformed into an ordinal categorical variable: country_rank
2. NaNs are removed.
3. Outliers are NOT removed.

1649 ROWS/22 COL

LINEAR REGRESSION

StandarScaler - Num
OneHotEncoder - Cat

Variable 'country'

K-NN

StandarScaler - Num
OneHotEncoder - Cat

Variable 'country_rank'

RANDOM FOREST

StandarScaler - Num
OneHotEncoder - Cat

Variable 'country_rank'

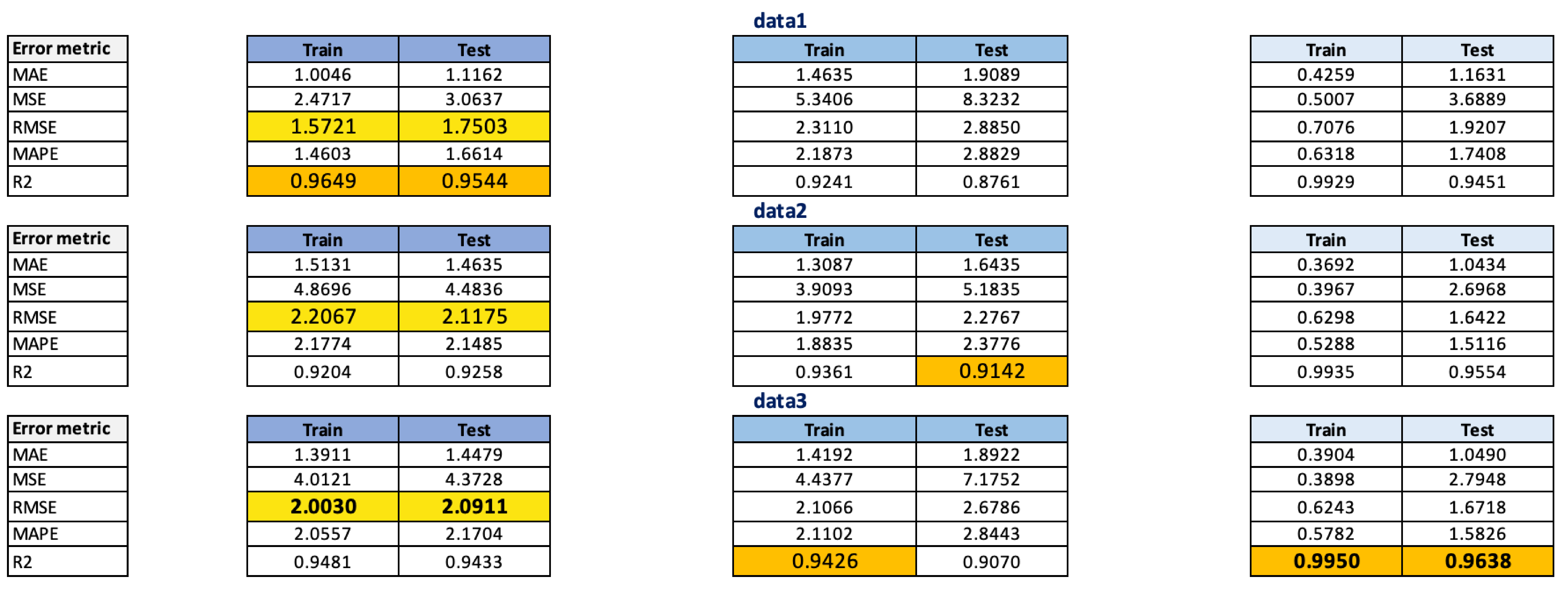
Agregar un subtítulo

3 Results

LINEAR REGRESSION

K-NN

RANDOM FOREST



data1

Error metric	Train	Test	Train	Test	Train	Test
MAE	1.0046	1.1162	1.4635	1.9089	0.4259	1.1631
MSE	2.4717	3.0637	5.3406	8.3232	0.5007	3.6889
RMSE	1.5721	1.7503	2.3110	2.8850	0.7076	1.9207
MAPE	1.4603	1.6614	2.1873	2.8829	0.6318	1.7408
R2	0.9649	0.9544	0.9241	0.8761	0.9929	0.9451

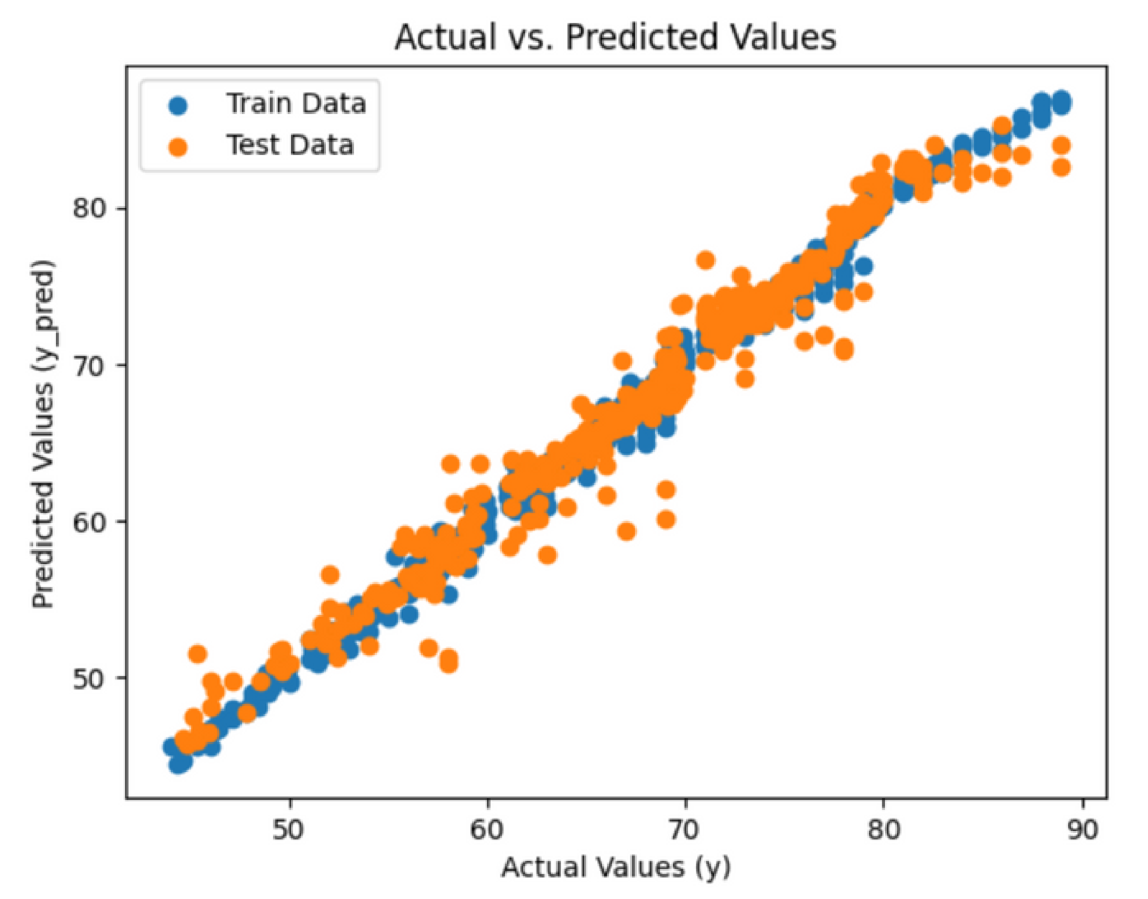
data2

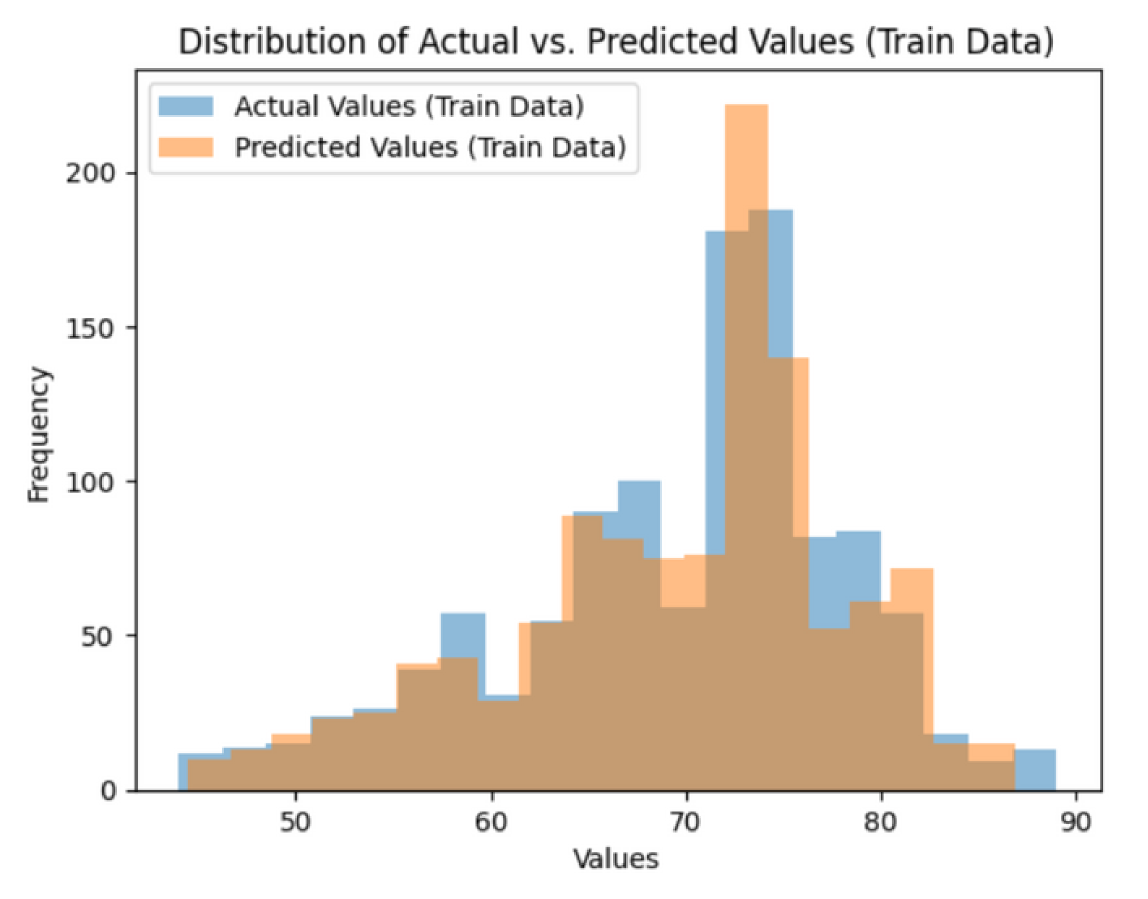
Error metric	Train	Test	Train	Test	Train	Test
MAE	1.5131	1.4635	1.3087	1.6435	0.3692	1.0434
MSE	4.8696	4.4836	3.9093	5.1835	0.3967	2.6968
RMSE	2.2067	2.1175	1.9772	2.2767	0.6298	1.6422
MAPE	2.1774	2.1485	1.8835	2.3776	0.5288	1.5116
R2	0.9204	0.9258	0.9361	0.9142	0.9935	0.9554

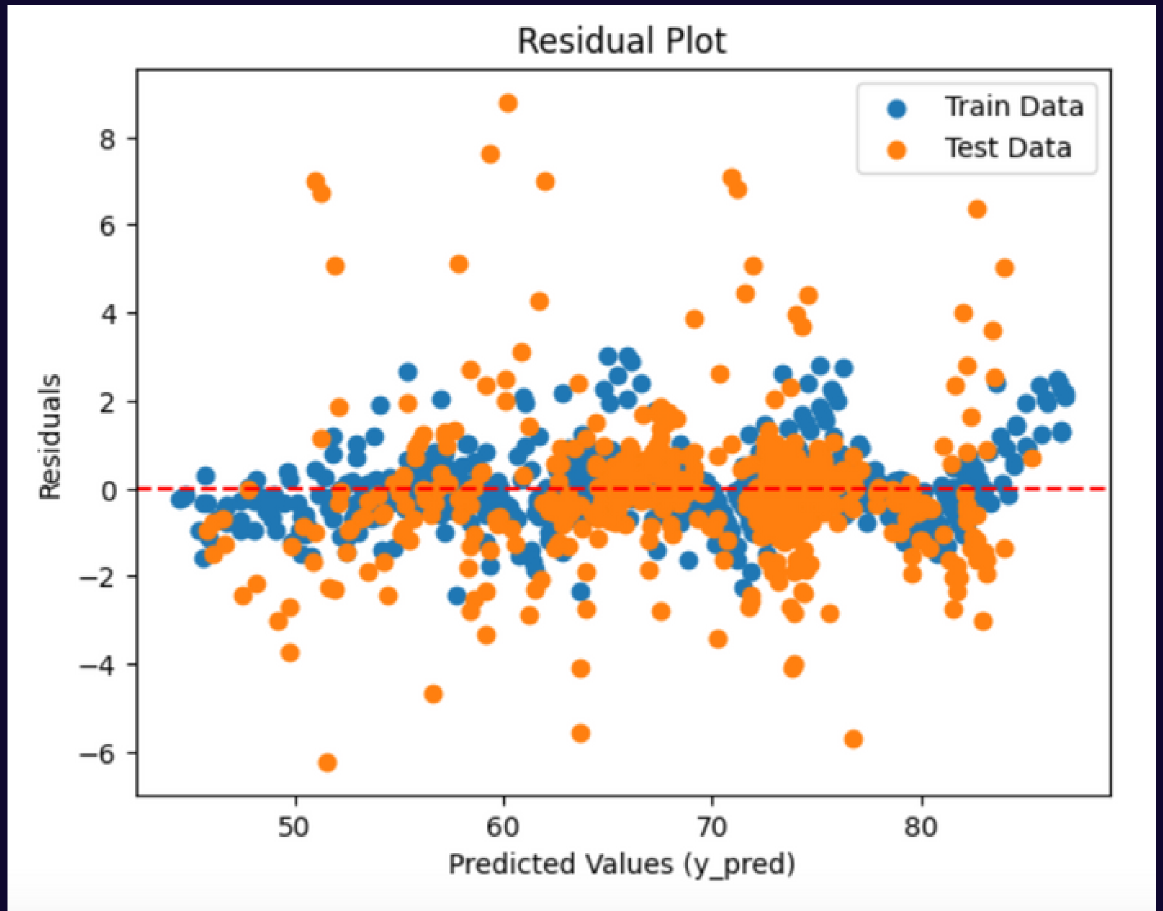
data3

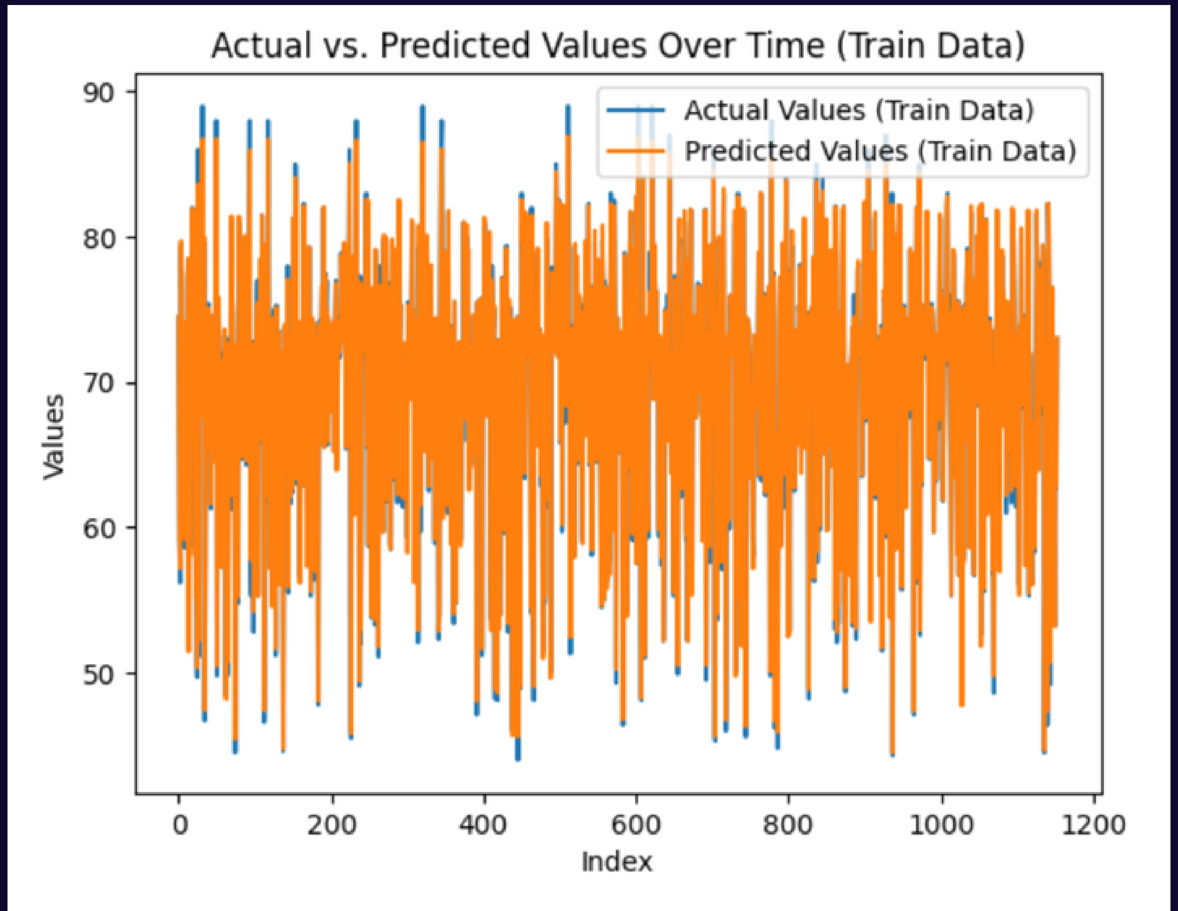
Error metric	Train	Test	Train	Test	Train	Test
MAE	1.3911	1.4479	1.4192	1.8922	0.3904	1.0490
MSE	4.0121	4.3728	4.4377	7.1752	0.3898	2.7948
RMSE	**2.0030**	**2.0911**	2.1066	2.6786	0.6243	1.6718
MAPE	2.0557	2.1704	2.1102	2.8443	0.5782	1.5826
R2	0.9481	0.9433	0.9426	0.9070	**0.9950**	**0.9638**

DATA 3 – RANDOM FOREST









Conclusions

NaN

Categorical

Overfitting

Multifactorial

Country

Gender

THANK YOU FOR YOUR ATTENTION