SC1015 Mini-Project

A128 Team 5

Economic Data VS Fertility Rate

Ng Hai Shen Barry (U2222267H)
Tan Ke Yuan
Zhong Xing Jie (U2221276A)

Table of Contents

01 Introduction

Welcome to our project!

World Population

#1	World, 2023	> 8, 000, 000, 000 people
#2	Most Populated	China, with > 1, 450, 000, 000 people (2022)
#3	Least Populated	Niue, with only 1610 people (2021)

Overly-Populated Countries

151.86 people per square km

480.5 people per square km

Effects of Overpopulation

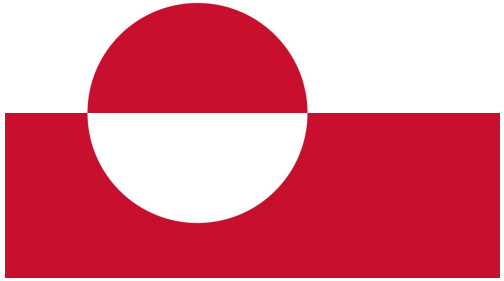
Ecological Degradation

More deforestation, spikes in pollution and harmful emissions such as toxic greenhouse gases

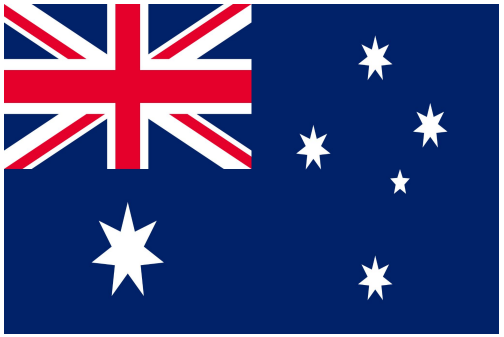
Disasters & Pandemics

Densely populated countries have higher risk of spreading infections

Under-Populated Countries



0.14 people per square km



3.44 people per square km

Effects of Underpopulation

Lower Standard of Living

Inadequate labour force to boost output and production of goods and services

Country Defence

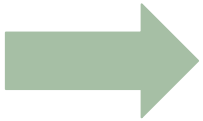
At times of war and emergency, countries may find it hard to mobilise enough people to defend

Problem Statement

“ What strategies can countries adopt to control population growth? ”

How does different economic indicators affect fertility rate?

- CapitalGDP
- GiniIndex
- LiteracyRate
- FemalePercentage
- UnemploymentRate
- PovertyRate



- FertilityRate

02 Data Preparation & Cleaning

NULL Data Processing
Removing Outliers

Data Cleaning

Remove unnecessary columns such as "Country Code', "Indicator Code" etc

- Not relevant to our analysis as it does not provide useful information

Converting "Year" column to integer data type

- Able to perform arithmetic operations on the values such as grouping the data by year

Methods to fill missing data

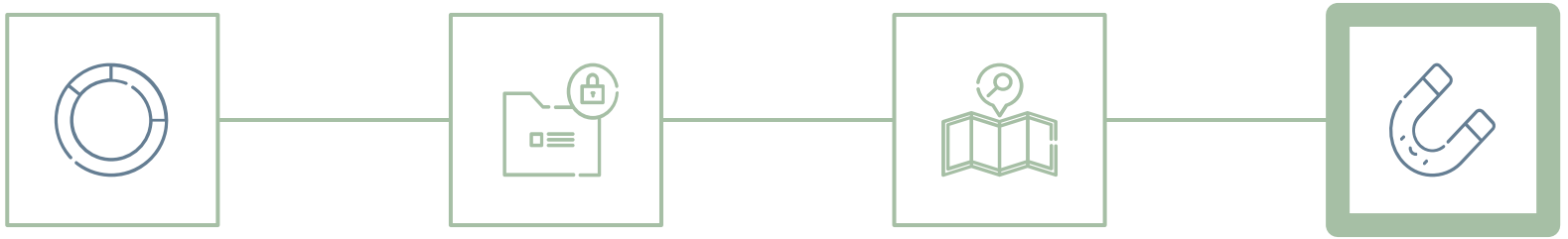
DropNA

Fill with mean

Fill with median

KNN Method

Methods to fill missing data



DropNA

Fill with mean

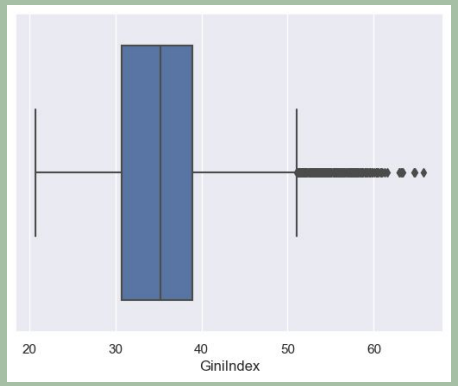
Fill with median

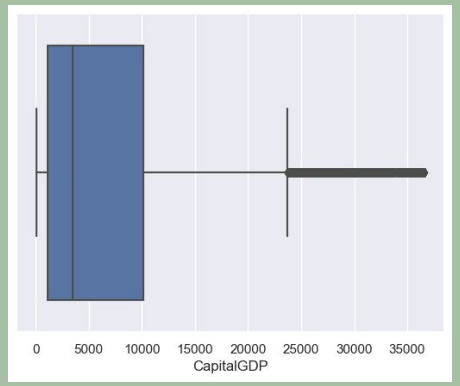
KNN Method

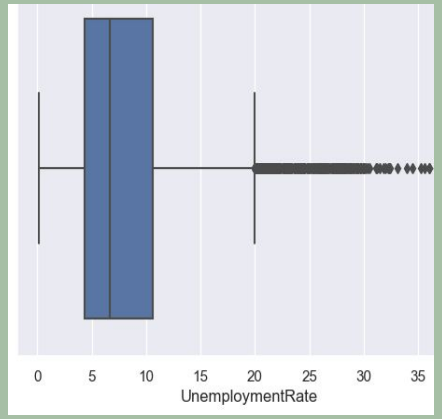
03 Exploratory Data Analysis

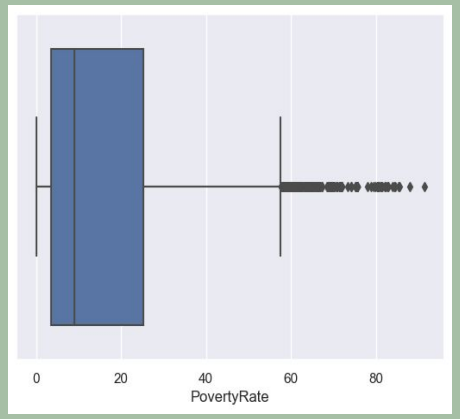
Enter subtitle here

Handling outliers

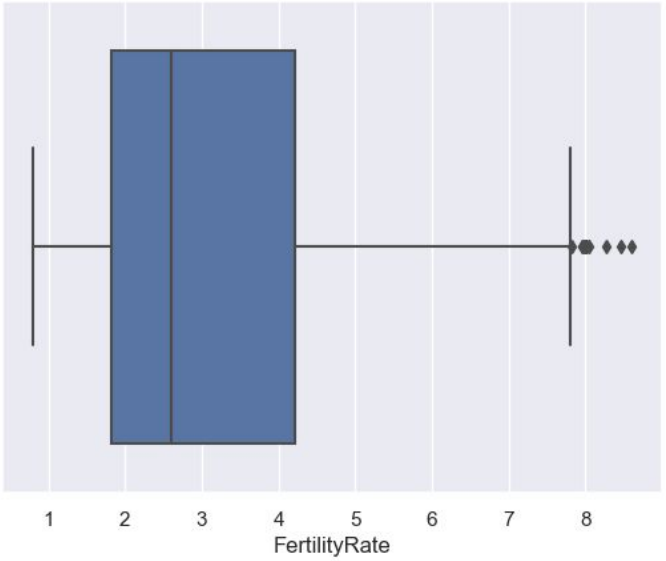




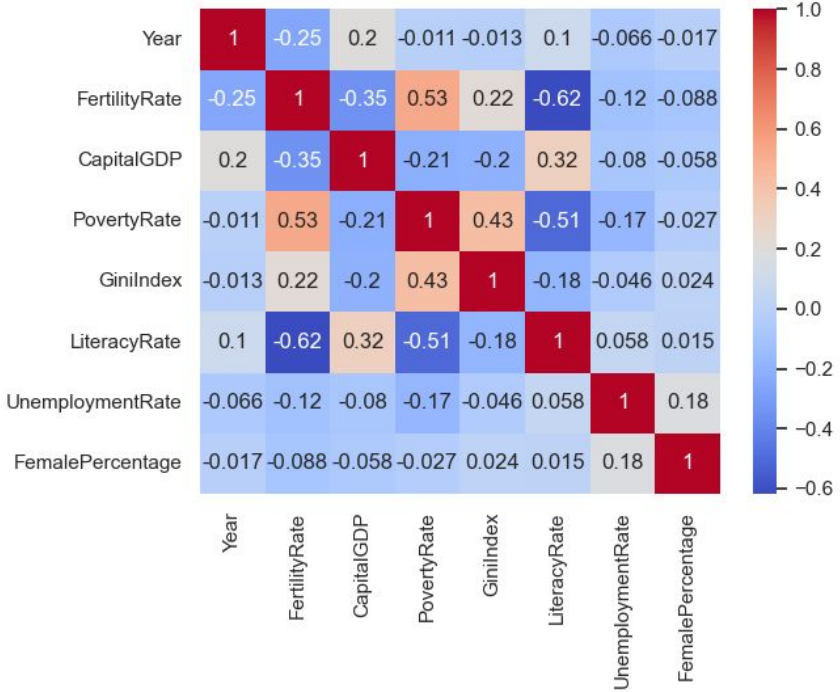




Create boxplots to observe the data first

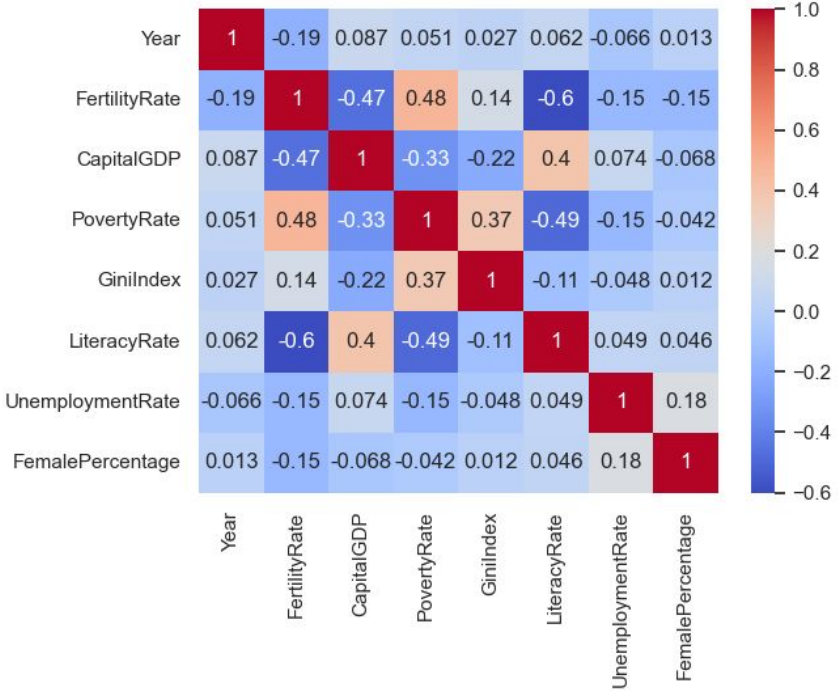


Heatmap BEFORE removing outliers



- Outliers can skew data and produce misleading results
- Majority shows moderate correlation

Heatmap AFTER removing outliers



- Even with lower correlation, removing outliers reduce influence of extreme values and skewing of data
- Provide more accurate and representative pictures

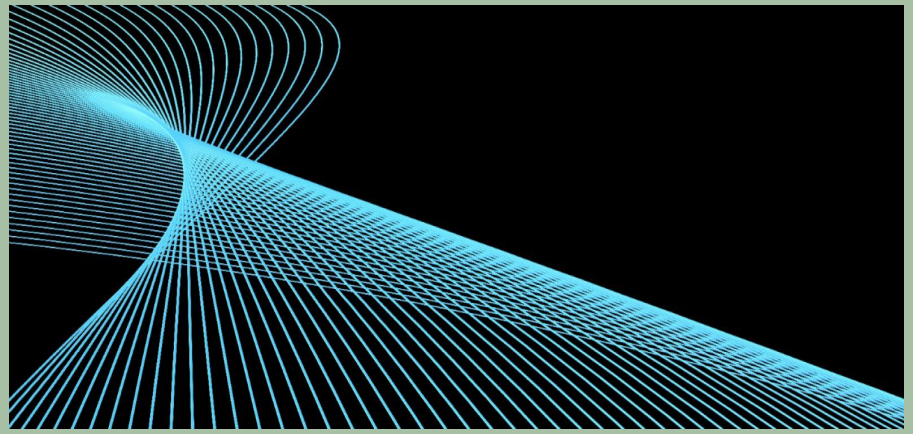
04
Machine Learning

Linear Regression
Neural Networks

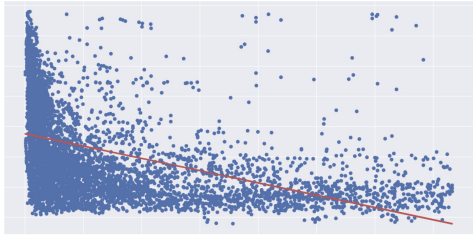
Why Linear Regression?

- Based on past, actual data
- Helps us predict which indicators increase and decrease fertility rate
- Gives us a chance to learn from our predictions and make the necessary adjustments

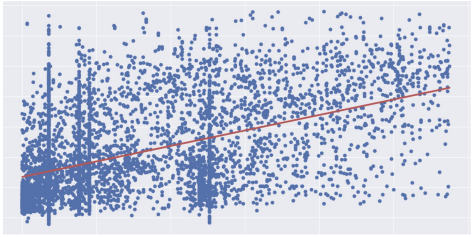
Linear Regression (Uni-variate)



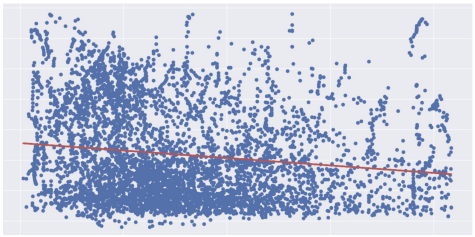
Regression Lines Against FertilityRate (y-axis)



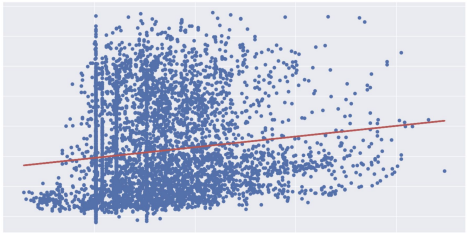
CapitalGDP



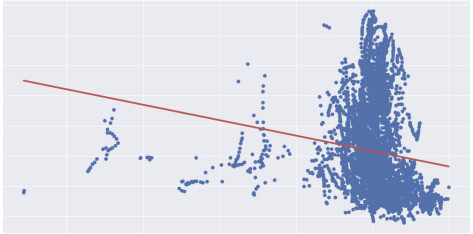
PovertyRate



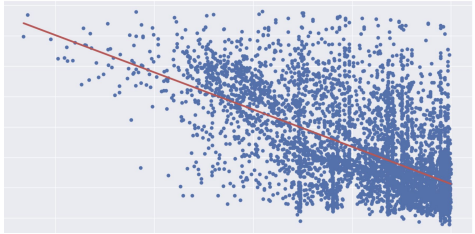
UnemploymentRate



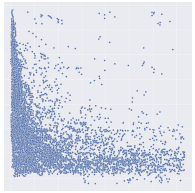
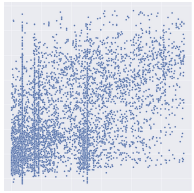
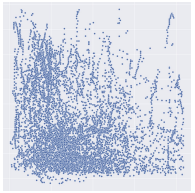
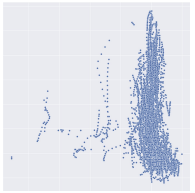
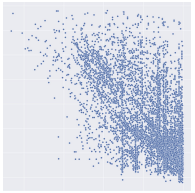
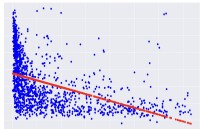
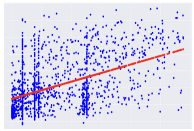
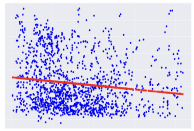
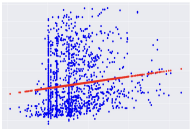
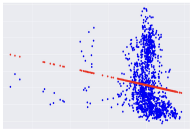
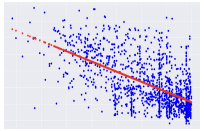
GiniIndex



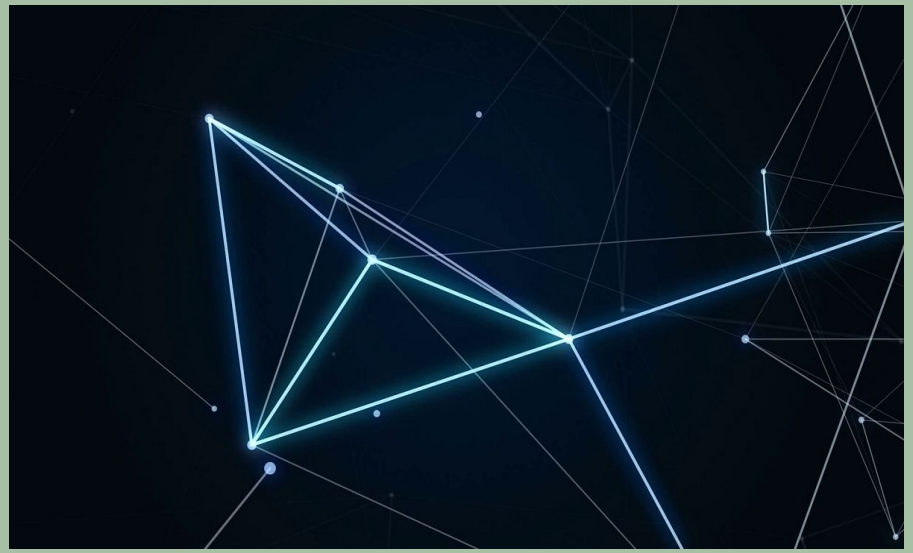
FemalePercentage



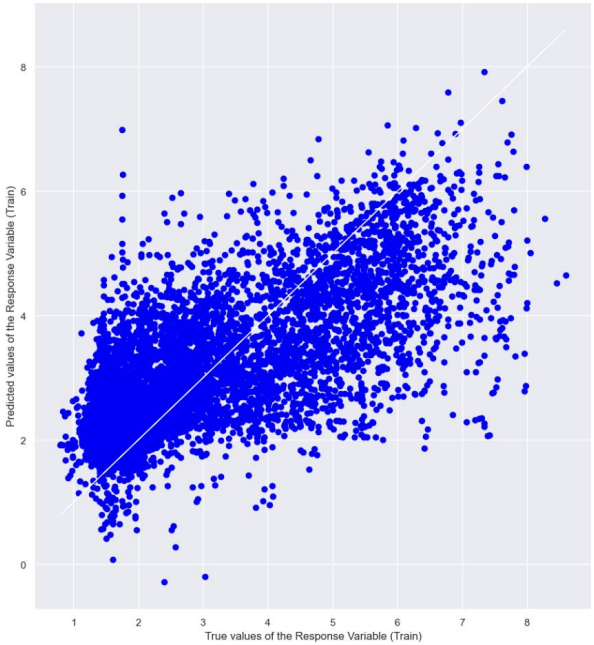
LiteracyRate

	CapitalGDP	PovertyRate	Unemployment	GiniIndex	Female %	LiteracyRate
Jointplot						
Train:Test	5743:1436	5743:1436	5743:1436	5743:1436	5743:1436	5743:1436
R^2 (Regression Line)	0.21536	0.23025	0.02149	0.01983	0.02103	0.36105
Scatterplot						
MSE	1.86103	1.81021	2.37299	2.57991	2.46809	1.46062

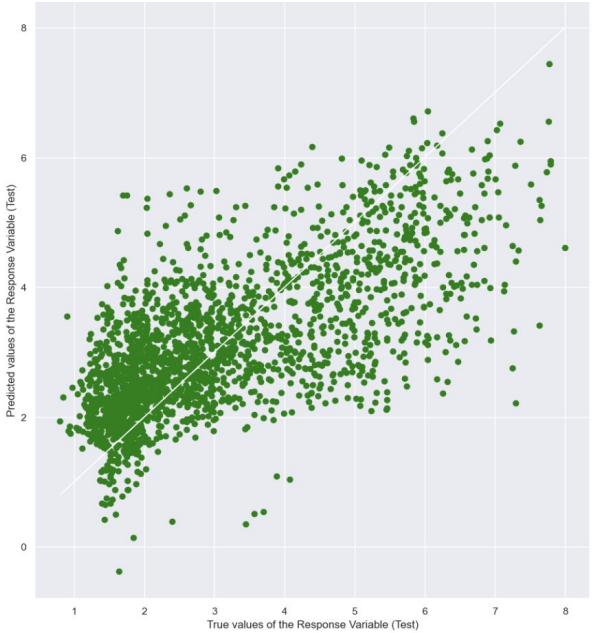
Linear Regression (Multi-variate)



True Values of Train & Test Response

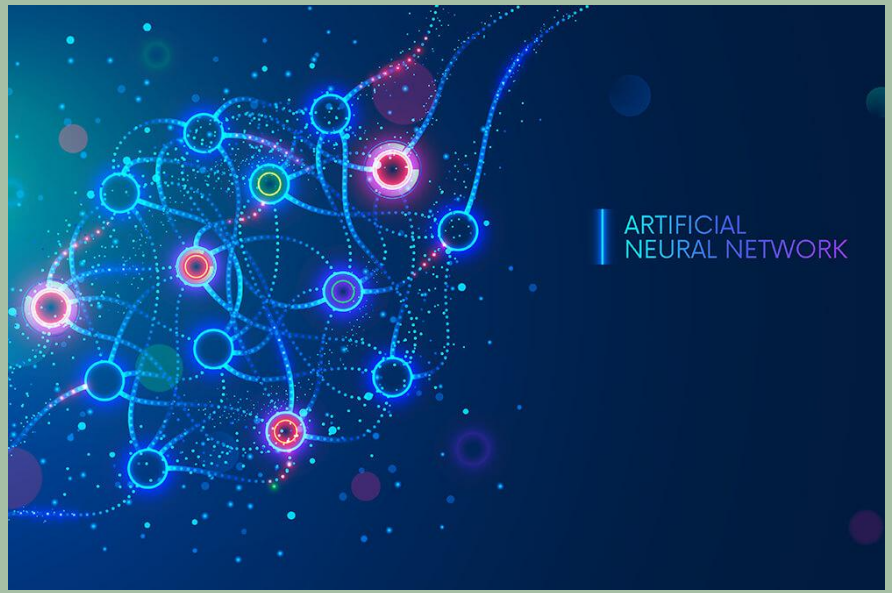


Goodness of Fit of Model Train Dataset
Explained Variance (R^2) : 0.47113274455981624
Mean Squared Error (MSE) : 1.376423938460147

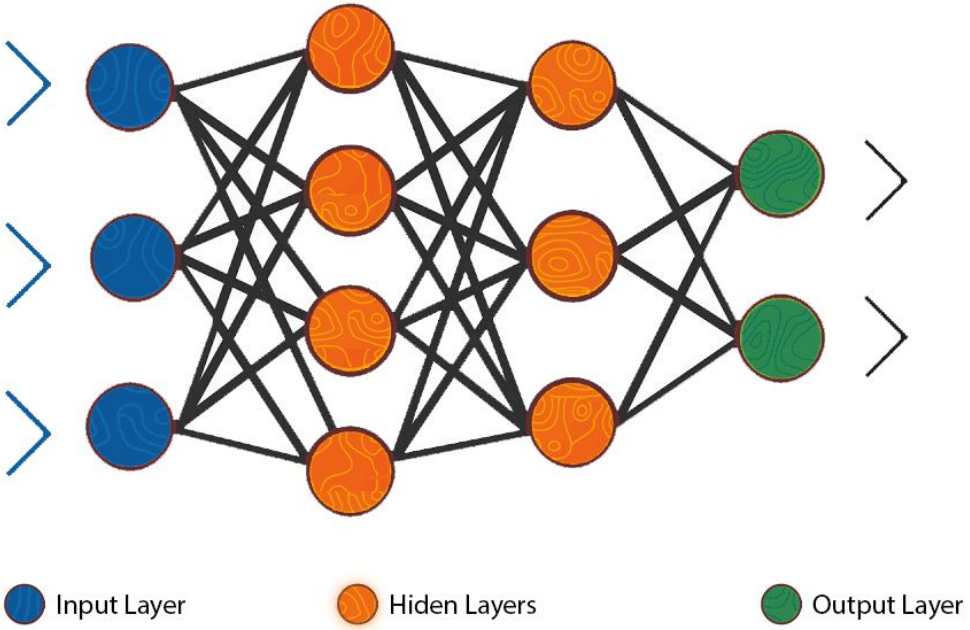


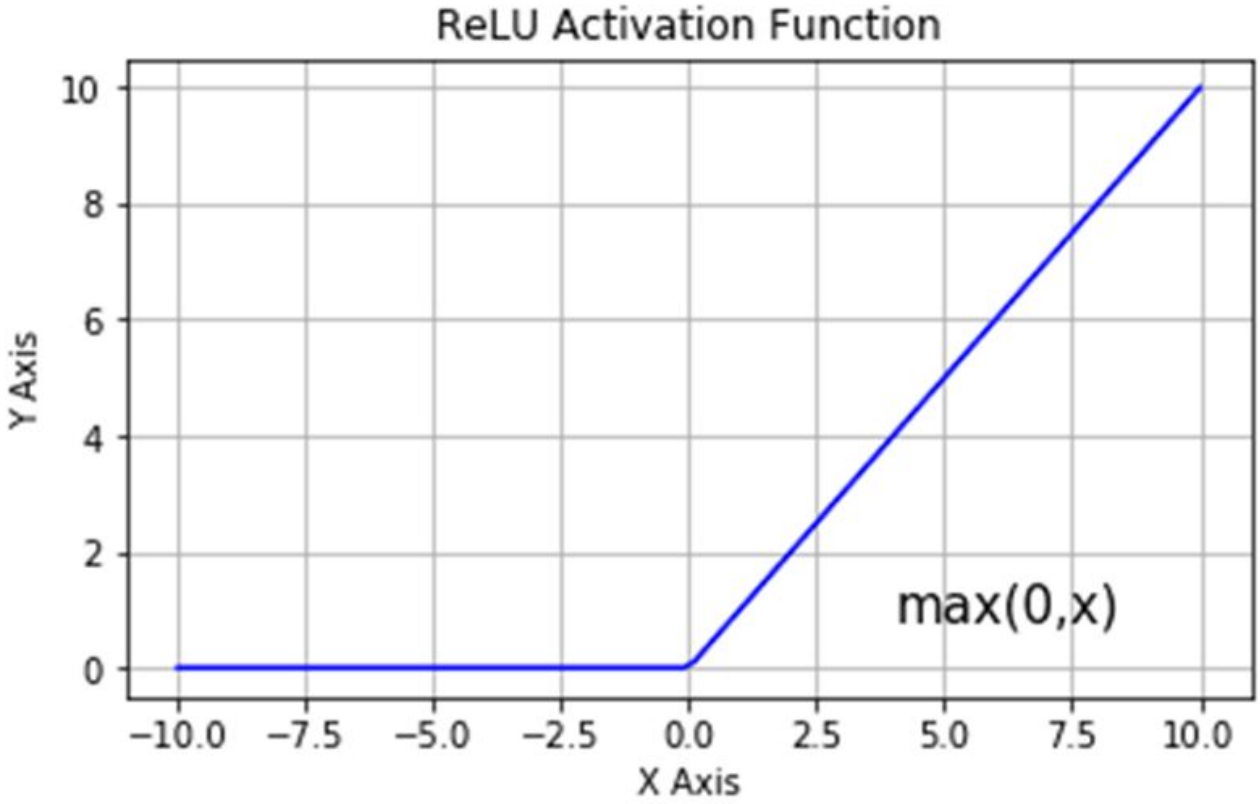
Goodness of Fit of Model Test Dataset
Explained Variance (R^2) : 0.48375506791788125
Mean Squared Error (MSE) : 1.258343668576167

Multi-Layer Perceptron (MLP)



How MLP works




ReLU Activation Function
Y Axis
X Axis
max(0,x)
10
8
6
4
2
0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Manually set hyperparameters

Layers + Neurons

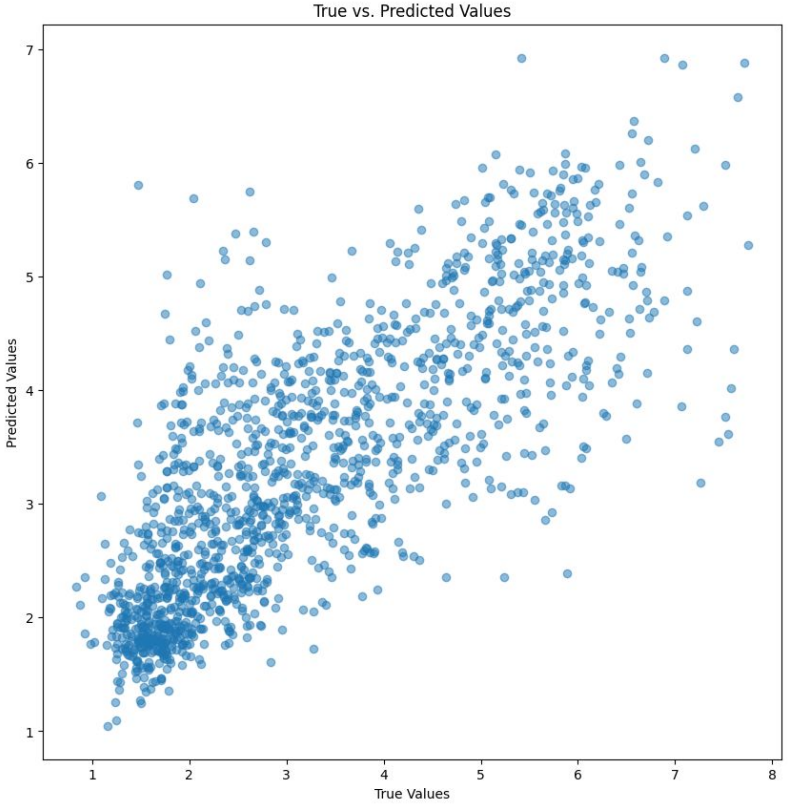
```
input_size = X.shape[1]
hidden_size1 = 16
hidden_size2 = 16
output_size = 1
```

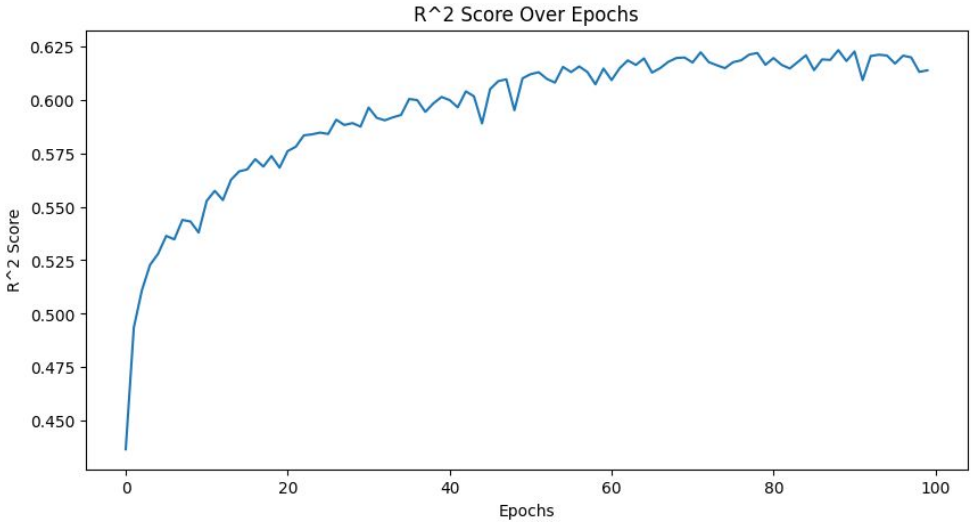
Learning rate

```
optimizer = optim.Adam(model.parameters(), lr=0.001)
```

Epoch number

```
num_epochs = 100
```


True vs. Predicted Values
Predicted Values
True Values
1
2
3
4
5
6
7
8


R^2 Score Over Epochs
R^2 Score
Epochs
0.450
0.475
0.500
0.525
0.550
0.575
0.600
0.625
0
20
40
60
80
100

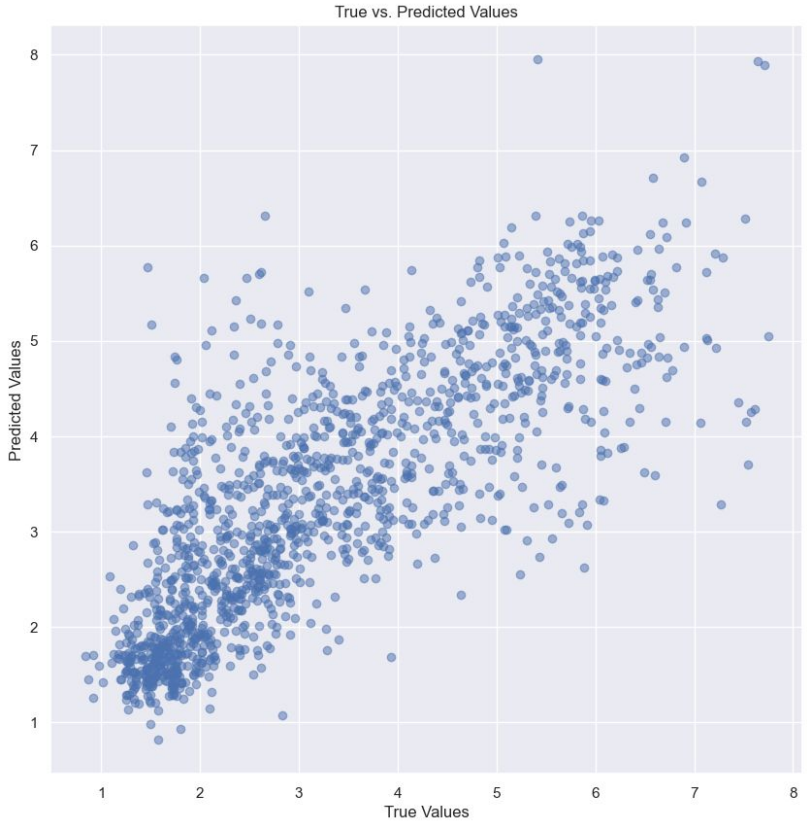
Implement GridSearch for hyperparameter tuning

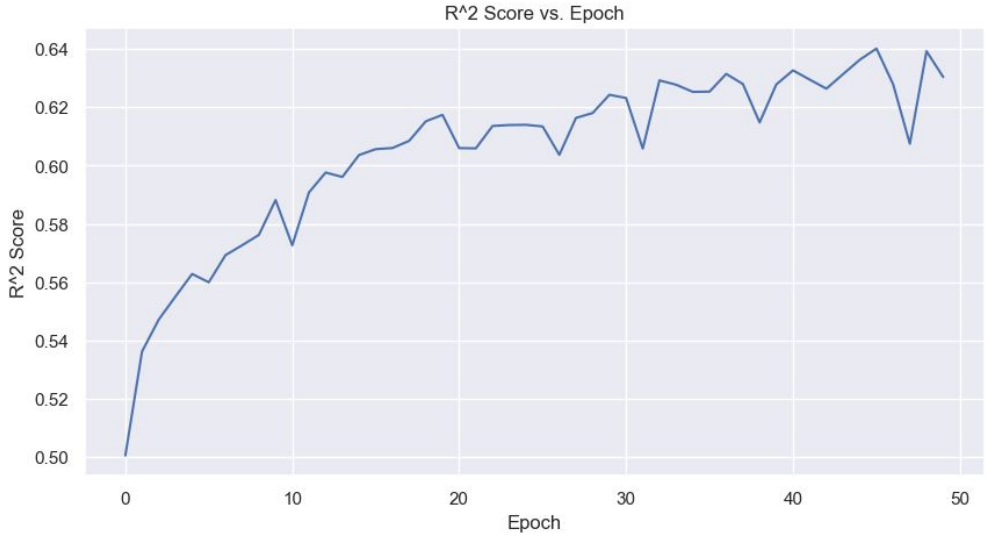
Possible hyperparameter values

```
param_grid = {
    'hidden_size1': [32, 64, 128],
    'hidden_size2': [16, 32, 64],
    'num_epochs': [50, 100, 150],
    'batch_size': [16, 32, 64],
    'learning_rate': [0.01, 0.001, 0.0001],
}
```

Best hyperparameter combination

```
Best parameters found:
  {'batch_size': 16,
   'hidden_size1': 64,
   'hidden_size2': 16,
'learning_rate': 0.001,
   'num_epochs': 50}
```


True vs. Predicted Values
Predicted Values
True Values
1
2
3
4
5
6
7
8


R^2 Score vs. Epoch
R^2 Score
Epoch
0.50
0.52
0.54
0.56
0.58
0.60
0.62
0.64
0
10
20
30
40
50

05

New Insights

What we learnt and doing something beyond this course

Data processing:

K nearest neighbours algorithm

Machine learning model

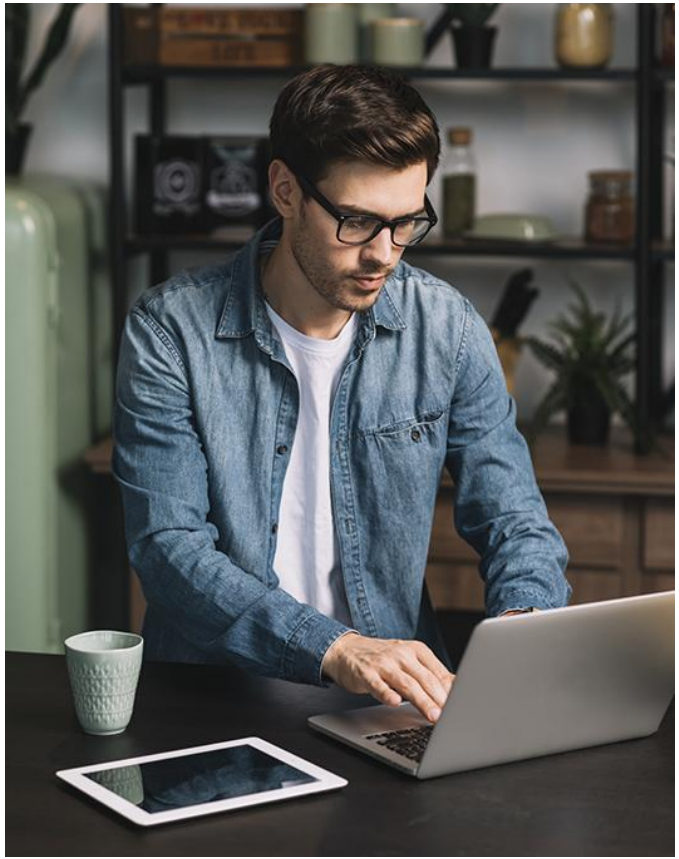
Multi-layer perceptron

Hyperparameter optimization

Grid search

06
Conclusion

Final Statement



Our outcomes

Identify significant factors influencing fertility rates

Understand the complex relationships between economic indicators and fertility rate

Thanks!