Supervised, Unsupervised, Time Series Machine Learning Models for an Economics Dataset

Agenda

- Introduction
- Data cleaning
- Features
- Data exploration
- Feature selections
- Machine learning Models
- Conclusions
- Future work
- Reference

Introduction

In this research, an economics data set has been chosen.

There are 45 countries and 8 economic features for every country.

The data is a time series data from 2012 to mid 2020. Also, it is seasonally adjusted. The data for 2020 is incomplete so it has not been used.

Unsupervised machine learning model to cluster countries in 8 groups has been used. Also, supervised machine learning has been used to see which features affects more on GDP in general. Finally, Time series model used to predict Industrial production for two years (2019 and 2020) based on past three years (2016 to 2018). The results has been compared to the real outcome.

Data cleaning

- 8 Countries did not have GDP, to have more reliable results, these countries dropped.
- For every 12 months of years, the related GDP allocated.
- The years dropped, so the research works on every 12 month of any year.

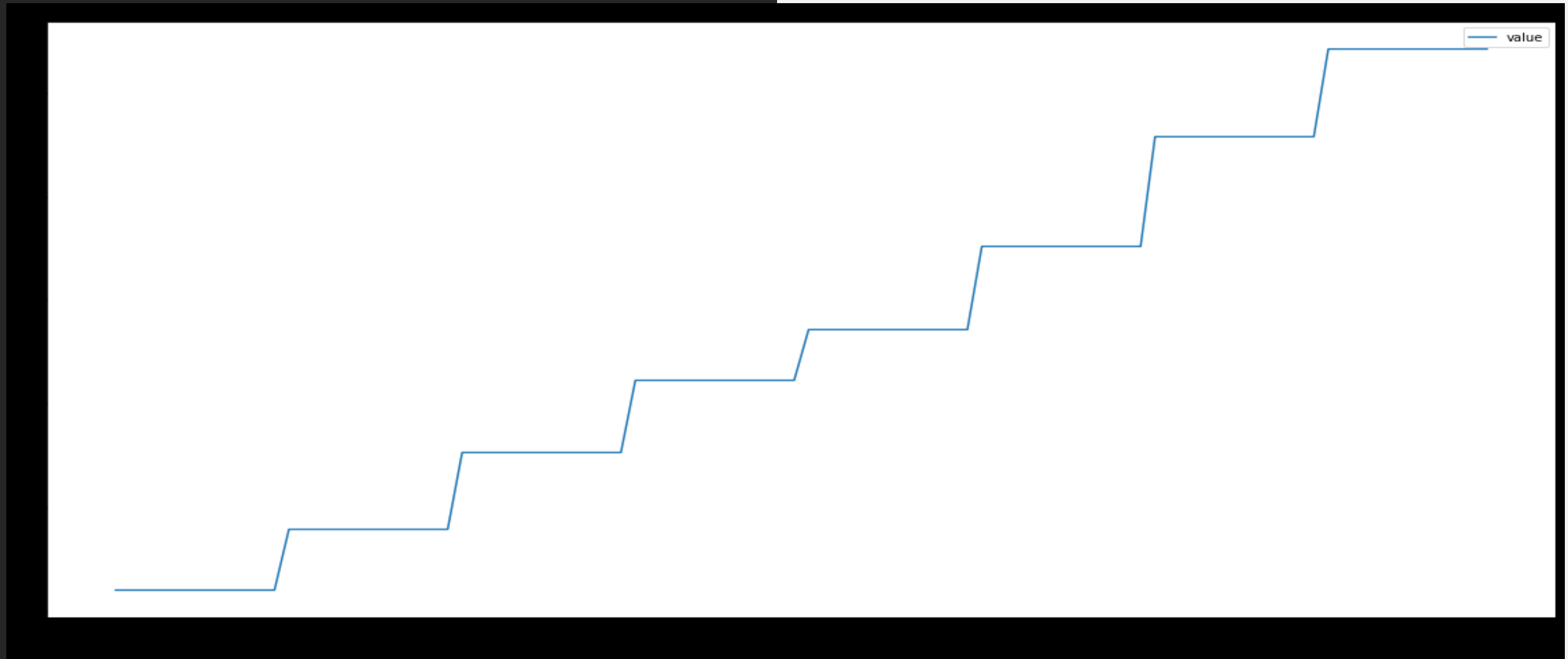
Features

In dataset there are features for every countries including:

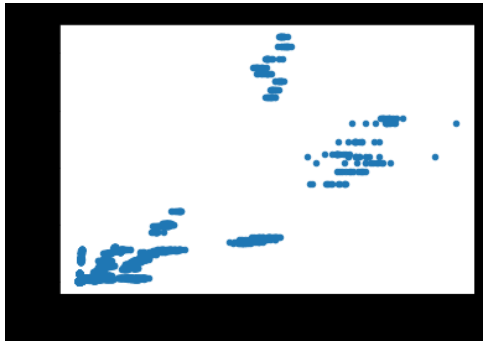
- GDP
- Export rate
- Import rate
- Stock market value
- Industrial production
- Unemployment rate
- Exchange rate

Data exploration

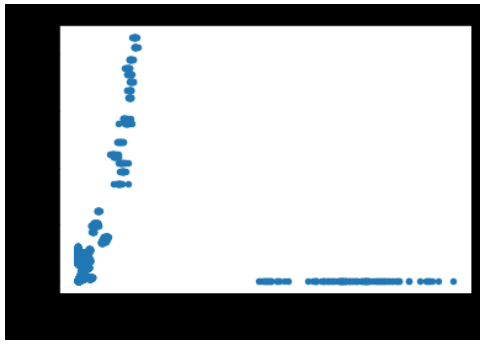
GDP in USA 2012 TO 2019



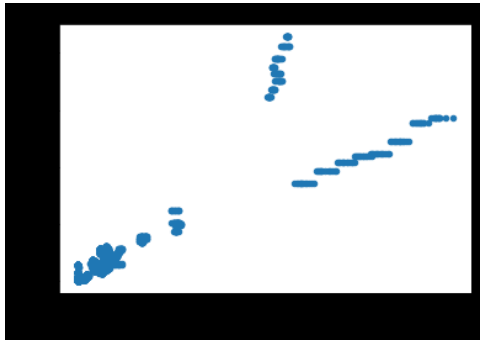
Data exploration



Export and GDP

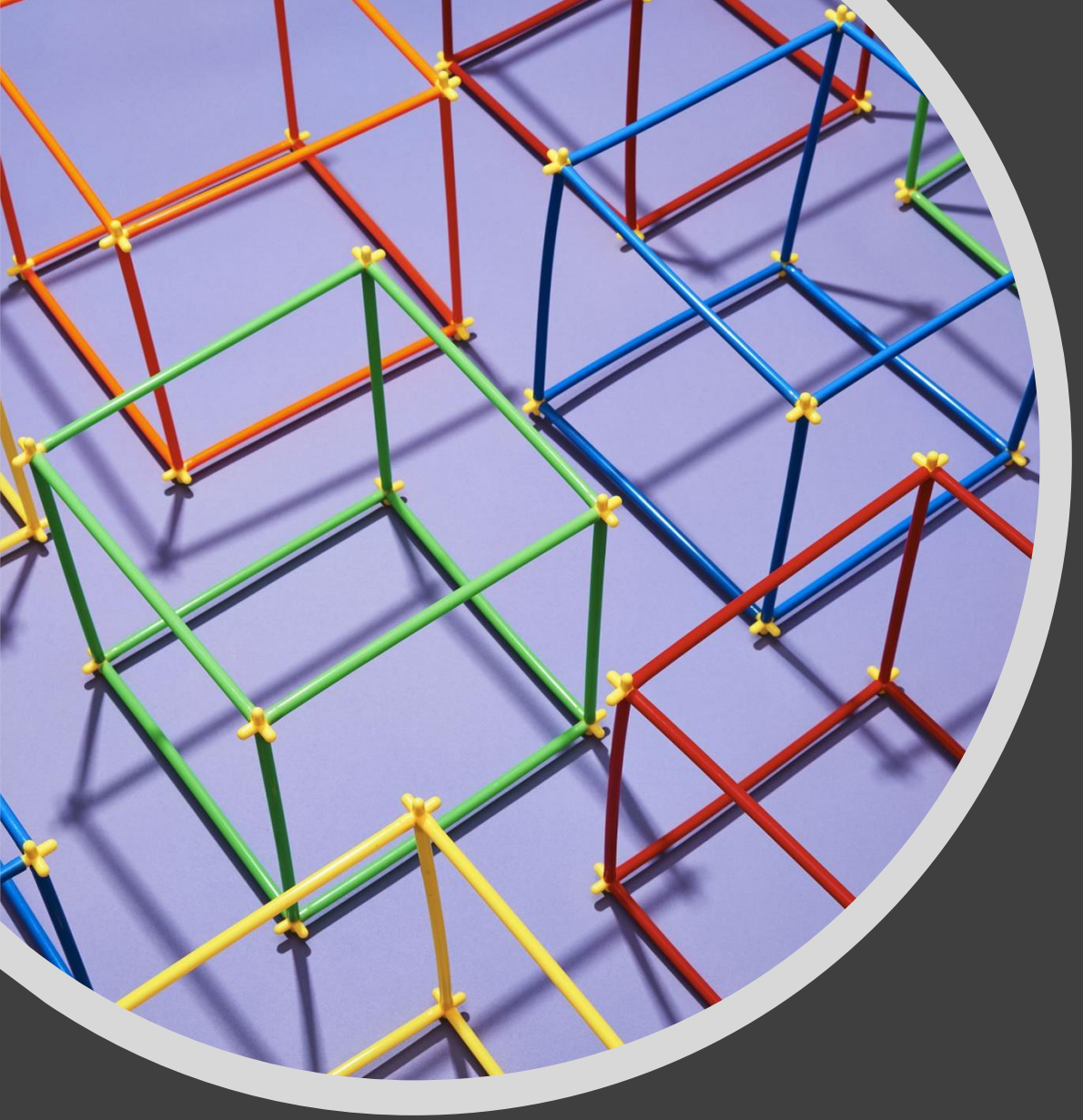


Import and GDP



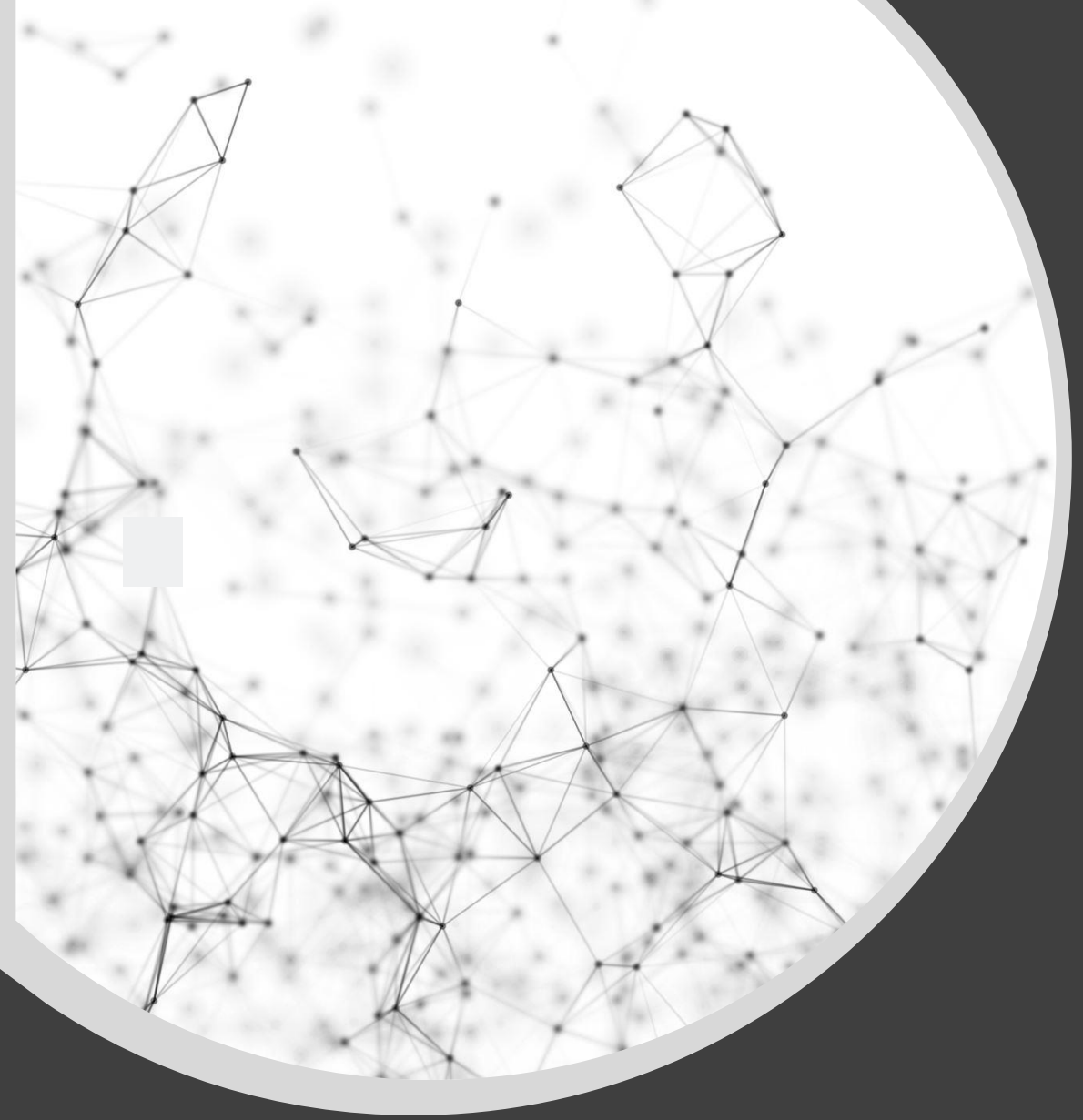
Industrial Production and GDP

Data exploration



Feature selections

In Series Column in the data set the features has been selected and changed to be the columns.



Machine learning Model

Agglomerative Clustering

Silhoette score

- N clusters=2 -> silhoette score is 0.68
- N clusters=3 -> silhoette score is 0.33
- N clusters=6 -> silhoette score is 0.39
- N clusters=8 -> silhoette score is 0.40

Agglomerative Clustering

Series	Exchange rate, new LCU per USD extended backward, period average,,	Exports Merchandise, Customs, current US$, millions, seas. adj.	GDP,current US$,millions,seas. adj.,	Imports Merchandise, Customs, current US$, millions, seas. adj.	Industrial Production, constant US$, seas. adj.,,	Stock Markets, US$,,,	Unemployment rate,Percent,,,
label							
Cluster: 0	292.655044	10953.866428	658455.424844	5503.587955	13102638651.968311	58.681050	13.508178
Cluster: 1	54.011763	19926.657803	1075598.548835	14422.951496	17513407310.377792	109.249535	5.952004
Cluster: 2	6.478809	189753.755962	11405243.471750	156546.040596	405396333333.333313	93.188648	3.967544
Cluster: 3	23.282956	9291.494888	352259.912559	10057.890618	8735637678.534626	198.382613	4.055490
Cluster: 4	0.377005	0.710520	31935.734368	912569.515112	0.000000	84.320634	3.920000
Cluster: 5	1.000000	130802.543750	18622100.000000	194896.399441	275631416666.666687	183.127274	5.455208
Cluster: 6	53.206803	89127.483574	4370674.255625	79607.241785	112644466458.182297	142.362601	3.793537
Cluster: 7	22042.605388	15172.052759	201869.304900	14964.309661	2649475762.114583	112.524468	2.181250

Agglomerative Clustering

row_0	Cluster: 0	Cluster: 1	Cluster: 2	Cluster: 3	Cluster: 4	Cluster: 5	Cluster: 6	Cluster: 7
Country								
Australia	0	0	96	0	0	0	0	0
Bahrain	0	60	0	0	0	0	0	0
Brazil	0	0	4	0	92	0	0	0
Canada	0	0	96	0	0	0	0	0
China	96	0	0	0	0	0	0	0
Colombia	0	0	0	53	43	0	0	0
Denmark	0	0	64	0	0	32	0	0
Finland	0	0	26	0	70	0	0	0
France	0	0	0	0	96	0	0	0
Germany	0	0	0	0	0	0	0	96
Greece	0	88	0	8	0	0	0	0
Hong Kong China	0	0	96	0	0	0	0	0
India	0	0	0	0	72	0	0	0
Israel	0	0	83	0	13	0	0	0
Italy	0	0	0	0	96	0	0	0
Japan	0	0	0	0	0	0	0	96
Jordan	0	0	0	96	0	0	0	0
Mexico	0	0	96	0	0	0	0	0
Morocco	0	0	0	0	96	0	0	0
New Zealand	0	0	86	0	0	10	0	0
Norway	0	0	96	0	0	0	0	0
Peru	0	0	63	0	33	0	0	0
Philippines	0	0	13	0	0	83	0	0
Portugal	0	0	0	64	32	0	0	0
Saudi Arabia	0	0	96	0	0	0	0	0
Singapore	0	0	96	0	0	0	0	0
Thailand	0	0	0	0	0	96	0	0
Turkey	0	0	6	0	90	0	0	0
United Kingdom	0	0	96	0	0	0	0	0
United States	96	0	0	0	0	0	0	0
Vietnam	0	0	0	0	0	0	96	0

RandomForestRegressor

Mean Abs Error Percentage = 0.00145

Model score (Train) = 0.9828

Model score (Test) = 0.9825

FEAT	IMPORTANCE	
3	Industrial Production, constant US$, seas. adj.,,	0.931695
0	Exchange rate, new LCU per USD extended backwa...	0.031082
1	Exports Merchandise, Customs, current US$, mil...	0.016838
2	Imports Merchandise, Customs, current US$, mil...	0.011515
5	Unemployment rate,Percent,,,	0.006300
4	Stock Markets, US$,,,	0.002570

Unemployment Rate in USA

Time Series

Date	Observed	Predicted	Difference Percentage
2019/01	4.0	3.777	0.055750
2019/02	3.8	3.770	0.007895
2019/03	3.8	3.747	0.013947
2019/04	3.6	3.752	-0.042222
2019/05	3.6	3.755	-0.043056
2019/06	3.7	3.752	-0.014054
2019/07	3.7	3.750	-0.013514
2019/08	3.7	3.750	-0.013514
2019/09	3.5	3.752	-0.072000
2019/10	3.6	3.752	-0.042222
2019/11	3.5	3.752	-0.072000
2019/12	3.5	3.752	-0.072000

Time Series Industrial Production

Date	Observed	Predicted	Difference Percentage
2019/01	2.904960e+11	2.898940e+11	0.002072
2019/02	2.890310e+11	2.889172e+11	0.000394
2019/03	2.893420e+11	2.889195e+11	0.001460
2019/04	2.875160e+11	2.881392e+11	-0.002168
2019/05	2.881430e+11	2.871840e+11	0.003328
2019/06	2.882770e+11	2.864755e+11	0.006249
2019/07	2.877700e+11	2.862352e+11	0.005333
2019/08	2.897990e+11	2.858490e+11	0.013630
2019/09	2.887920e+11	2.846193e+11	0.014449
2019/10	2.876170e+11	2.837612e+11	0.013406
2019/11	2.902860e+11	2.836573e+11	0.022835
2019/12	2.892670e+11	2.834470e+11	0.020120

Conclusion

In the research, 31 countries are clustered in 8 clusters based on their Export, Import, Exchange rate, Stock market value, Industrial production, Unemployment rate, GDP. For instance, China and USA are in one group.

The most important feature for predicting GDP is Industrial production.

The model predicted very well the Unemployment rate and Industrial production.

The research did not used 2020 data for two reason. First, the data just has information for 7 month. Additionally, due to Covid 19 the prediction should not be correct.