DAT102X: PREDICTING CHRONIC HUNGER

Prepared by: Dennis Lam

Date: 16 October 2018

EXECUTIVE SUMMARY

This document aims to present the data analysis, machine learning modeling used, results and conclusion for predicting chronic hunger in the world based on dataset provided by Food and Agricultural Organization of the United Nations (FAO). The challenge is to consider which economic, social, and political factors are indicative of trends in chronic hunger in countries around the world.

BUSINESS UNDERSTANDING

The goal of this capstone project is to predict the **annual prevalence of undernourishment** at the country level from other socioeconomic indicators. The prevalence of undernourishment expresses "the probability that a randomly selected individual from the population consumes an amount of calories that is insufficient to cover her/his energy requirement for an active and healthy life" (FAOSTAT).

Since prevalence of undernourishment is a numeric variable, regression method will be used to determine the predicted values.

WRANGLING, EXPLORATION AND CLEANING DATA

The train dataset consists of 47 variables. Total rows are 1401.

	row_id	country_code	year	agricultural_land_area	percentage_of_arable_land_equipped_for_irrigation	cereal_yield	droughts_floods_extreme_temps
0	0	889f053	2002	2.350777e+05	38.558520	935.754365	NaN
1	1	9e614ab	2012	2.300064e+04	21.282631	4031.452161	NaN
2	2	100c476	2000	9.095487e+01	4.317080	1581.935278	NaN
3	3	4609682	2013	1.008437e+05	16.636618	1127.626364	NaN
4	4	be2a7f5	2008	2.242894e+02	NaN	1418.987212	NaN
5	5	7e222a7	2014	5.196619e+04	NaN	1582.768005	NaN

Figure 1: The first 5 rows of the dataset

There are 1 categorical variable and 46 numerical variables in this dataset.

Summary statistics for each individual variables are presented for min, max , distinct count, standard deviation, mean and median in the following table.

	count	mean	std	min	25%	50%	75%	max
row_id	1401.0	7.000000e+02	4.045782e+02	0.000000	3.500000e+02	7.000000e+02	1.050000e+03	1.400000e+03
year	1401.0	2.007393e+03	4.595501e+00	2000.000000	2.003000e+03	2.007000e+03	2.011000e+03	2.015000e+03
agricultural_land_area	1385.0	3.539588e+05	1.172377e+06	2.944179	1.174577e+04	4.701980e+04	2.247874e+05	1.045780e+07
percentage_of_arable_land_equipped_for_irrigation	1153.0	2.789145e+01	2.857762e+01	0.000000	3.490956e+00	1.884623e+01	4.195478e+01	1.019063e+02
cereal_yield	1337.0	2.753178e+03	2.777815e+03	179.258873	1.424504e+03	2.221921e+03	3.296467e+03	2.797827e+04
droughts_floods_extreme_temps	75.0	1.236800e+00	1.877823e+00	0.000000	9.741960e-02	6.613793e-01	1.318327e+00	9.177338e+00
forest_area	1385.0	2.329455e+05	9.266334e+05	9.806688	4.159005e+03	2.224170e+04	1.255963e+05	8.243222e+06
total_land_area	1401.0	8.181146e+05	2.792117e+06	20.183062	2.507460e+04	1.309442e+05	6.261072e+05	2.403061e+07
fertility_rate	1387.0	3.251874e+00	1.471044e+00	0.836053	2.175432e+00	2.751553e+00	4.227445e+00	7.544631e+00
life_expectancy	1386.0	6.711405e+01	8.786850e+00	38.204140	6.167800e+01	6.985772e+01	7.370648e+01	8.477140e+01
rural_population	1401.0	2.658025e+07	1.052394e+08	0.000000	8.349747e+05	3.373348e+06	1.191299e+07	8.947322e+08
total_population	1401.0	4.499105e+07	1.546745e+08	61724.552030	1.516541e+06	7.378974e+06	2.614718e+07	1.313304e+09
urban_population	1401.0	1.840486e+07	5.150763e+07	24138.439680	8.101741e+05	3.511671e+06	1.112414e+07	4.302162e+08
population_growth	1400.0	1.636426e+00	1.299897e+00	-2.872249	9.049466e-01	1.546457e+00	2.399062e+00	1.422116e+01
avg_value_of_food_production	1234.0	2.294743e+02	1.490591e+02	3.945363	1.340173e+02	2.052890e+02	2.736998e+02	1.042484e+03
cereal_import_dependency_ratio	1084.0	3.437284e+01	5.193730e+01	-228.300258	1.126337e+01	3.504418e+01	7.172308e+01	1.019841e+02
food_imports_as_share_of_merch_exports	1148.0	3.714854e+01	6.656490e+01	0.990945	7.096791e+00	1.618082e+01	3.664659e+01	7.633824e+02
gross_domestic_product_per_capita_ppp	1362.0	1.084343e+04	1.527531e+04	573.167687	2.660430e+03	6.962375e+03	1.226570e+04	1.379537e+05
imports_of_goods_and_services	1324.0	4.547967e+01	2.284027e+01	0.065060	2.930445e+01	4.238844e+01	5.803774e+01	2.373016e+02
inequality_index	429.0	4.276917e+01	9.278521e+00	16.240718	3.487615e+01	4.308527e+01	5.072699e+01	6.448980e+01
net_oda_received_percent_gni	1237.0	6.105307e+00	1.202022e+01	-0.665359	4.284228e-01	2.168309e+00	7.538874e+00	1.891348e+02
net_oda_received_per_capita	1239.0	6.305770e+01	8.916066e+01	-49.355612	1.193211e+01	3.453597e+01	7.539141e+01	8.071926e+02
tax_revenue_share_gdp	856.0	1.640988e+01	7.863698e+00	0.057901	1.200722e+01	1.518702e+01	2.045518e+01	5.875965e+01
trade_in_services	1236.0	2.304100e+01	2.165651e+01	2.308559	1.078563e+01	1.731053e+01	2.832822e+01	2.699816e+02
per_capita_food_production_variability	1314.0	1.057109e+01	1.230408e+01	0.300291	4.173722e+00	7.024724e+00	1.198666e+01	1.060211e+02
per_capita_food_supply_variability	1229.0	3.795659e+01	2.365707e+01	2.018557	2.065735e+01	3.107293e+01	4.892649e+01	1.412757e+02
adult_literacy_rate	285.0	7.963224e+01	1.822817e+01	24.140420	6.677408e+01	8.702611e+01	9.379762e+01	1.004628e+02
school_enrollment_rate_female	795.0	8.867150e+01	1.286126e+01	35.620178	8.511887e+01	9.356546e+01	9.736060e+01	1.016180e+02
school_enrollment_rate_total	897.0	9.025370e+01	1.116576e+01	35.335727	8.701430e+01	9.464141e+01	9.764412e+01	1.017758e+02
avg_supply_of_protein_of_animal_origin	1149.0	2.796357e+01	1.598439e+01	2.957107	1.385299e+01	2.514631e+01	3.923134e+01	8.321260e+01
caloric_energy_from_cereals_roots_tubers	1149.0	5.088803e+01	1.392569e+01	22.589928	3.958169e+01	5.030516e+01	6.170234e+01	8.438812e+01
access_to_improved_sanitation	1327.0	6.505176e+01	2.842234e+01	10.337271	3.988527e+01	7.346788e+01	9.064685e+01	1.017464e+02
access_to_improved_water_sources	1339.0	8.329940e+01	1.528494e+01	30.784598	7.422783e+01	8.844126e+01	9.539357e+01	1.019713e+02
anemia_prevalence	1321.0	3.278167e+01	1.199932e+01	12.570471	2.329137e+01	3.011147e+01	4.144861e+01	6.961755e+01
obesity_prevalence	1244.0	1.276597e+01	8.360314e+00	0.699575	4.767319e+00	1.283302e+01	1.891667e+01	4.444713e+01
open_defecation	1381.0	1.170486e+01	1.513444e+01	0.000000	5.982102e-01	4.773481e+00	1.858029e+01	6.668956e+01
hiv_incidence	1030.0	2.186237e-01	5.239597e-01	0.009800	1.016419e-02	4.006811e-02	1.667658e-01	4.269284e+00
rail_lines_density	457.0	1.183129e+00	1.175000e+00	0.000000	2.977128e-01	6.081983e-01	1.869188e+00	4.867161e+00
access_to_electricity	1397.0	7.379539e+01	3.128031e+01	0.010012	5.106234e+01	8.915622e+01	9.870897e+01	1.019967e+02
co2_emissions	1317.0	8.304671e+04	2.248360e+05	100.828806	1.265778e+03	7.637910e+03	4.689573e+04	2.265183e+06
unemployment_rate	1337.0	8.580335e+00	6.645133e+00	0.491115	3.748595e+00	6.633461e+00	1.145402e+01	3.797718e+01
total_labor_force	1337.0	1.871233e+07	6.112347e+07	34906.590240	9.076810e+05	3.411048e+06	1.117916e+07	4.985771e+08
military_expenditure_share_gdp	1128.0	1.919332e+00	1.480842e+00	0.000000	1.033886e+00	1.538130e+00	2.325482e+00	1.332611e+01
proportion_of_seats_held_by_women_in_gov	1258.0	1.561846e+01	1.032428e+01	0.000000	8.575444e+00	1.309303e+01	2.161453e+01	6.477381e+01
political_stability	1266.0	-3.760201e-01	8.588882e-01	-2.781258	-9.481668e-01	-2.876587e-01	2.004494e-01	1.376322e+00
prevalence_of_undernourishment	1401.0	1.551070e+01	1.161044e+01	2.493428	5.710856e+00	1.211866e+01	2.244749e+01	5.908978e+01

Figure 2: Summary statistics

Now looking at correlation between all variables, I have shown partially of the table generated due to long columns.

	row_id	year	agricultural_land_area	percentage_of_arable_land_equipped_for_irrigation	cereal_yield
row_id	1.000000	-0.033858	-0.012665	0.038056	0.013669
year	-0.033858	1.000000	-0.019564	0.065832	0.089587
agricultural_land_area	-0.012665	-0.019564	1.000000	-0.119532	-0.066400
percentage_of_arable_land_equipped_for_irrigation	0.038056	0.065832	-0.119532	1.000000	0.295433
cereal_yield	0.013669	0.089587	-0.066400	0.295433	1.000000
droughts_floods_extreme_temps	0.069712	NaN	-0.016910	0.110269	-0.048267
forest_area	0.000519	-0.063254	0.807158	-0.131208	-0.049100
total_land_area	-0.004847	-0.037219	0.954624	-0.118240	-0.060975
fertility_rate	-0.034164	-0.103411	0.122362	-0.322452	-0.304741
life_expectancy	0.013241	0.173825	-0.138010	0.411749	0.319883
rural_population	0.002592	0.016043	0.615256	0.031571	-0.026914
total_population	0.002807	0.020510	0.652992	0.026574	-0.021191
urban_population	0.001713	0.028795	0.706455	0.014980	-0.008771
population_growth	-0.005172	0.003866	0.088559	0.045605	0.089426
avg_value_of_food_production	-0.012288	0.043498	0.047433	-0.099903	0.092945
cereal_import_dependency_ratio	0.020055	-0.005756	-0.156501	0.180983	0.010940
food_imports_as_share_of_merch_exports	-0.011792	0.058364	-0.121582	-0.096129	-0.048596
gross_domestic_product_per_capita_ppp	0.002010	0.076106	-0.013977	0.205892	0.383657
imports_of_goods_and_services	0.033825	0.039288	-0.177518	-0.001207	0.019647
inequality_index	0.057315	-0.208403	0.093274	-0.307496	0.006705
net_oda_received_percent_gni	-0.026014	-0.028510	-0.072227	-0.151468	-0.151519
net_oda_received_per_capita	-0.032414	0.144377	-0.113222	-0.132317	-0.012828
tax_revenue_share_gdp	0.023611	0.098959	-0.057528	-0.105434	-0.074736
trade_in_services	0.010279	0.013466	-0.147484	-0.109407	-0.018674
per_capita_food_production_variability	-0.002320	-0.017208	-0.021106	-0.111677	0.145605

Figure 3: Correlation Table

The next figure is the heatmap of the correlation table and rotated to portrait mode for easier reading.

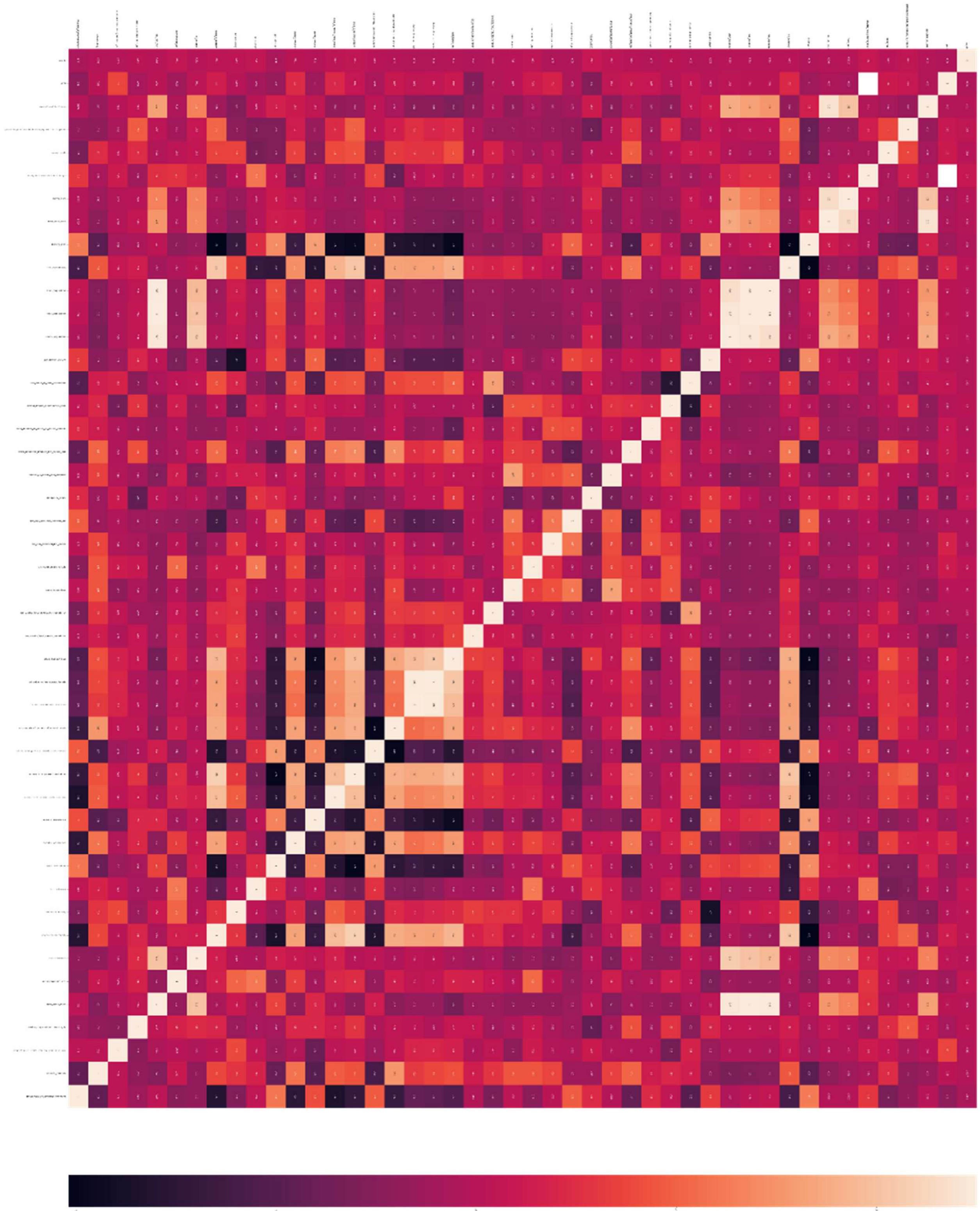


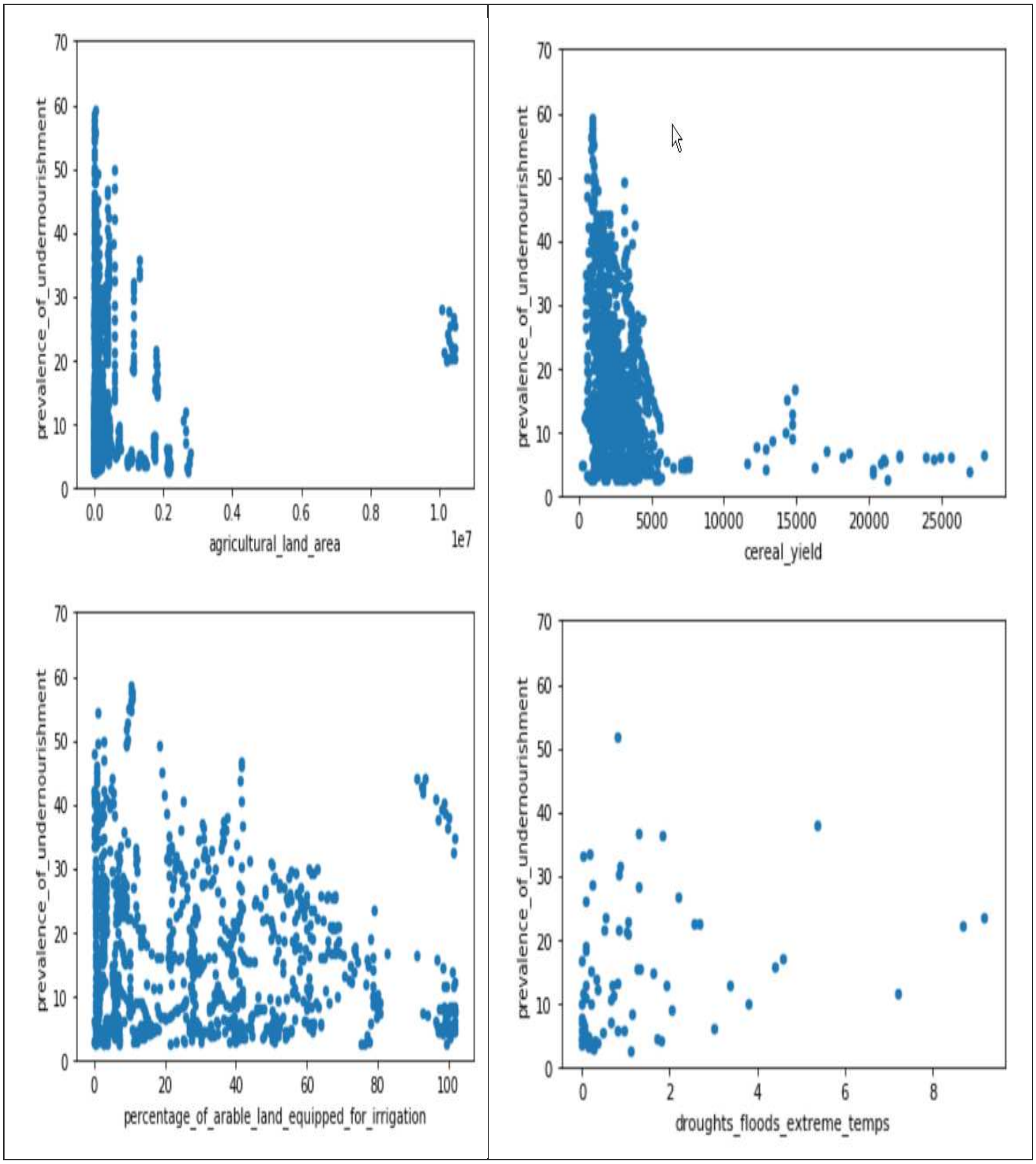
Figure 4: Heatmap of the correlation table

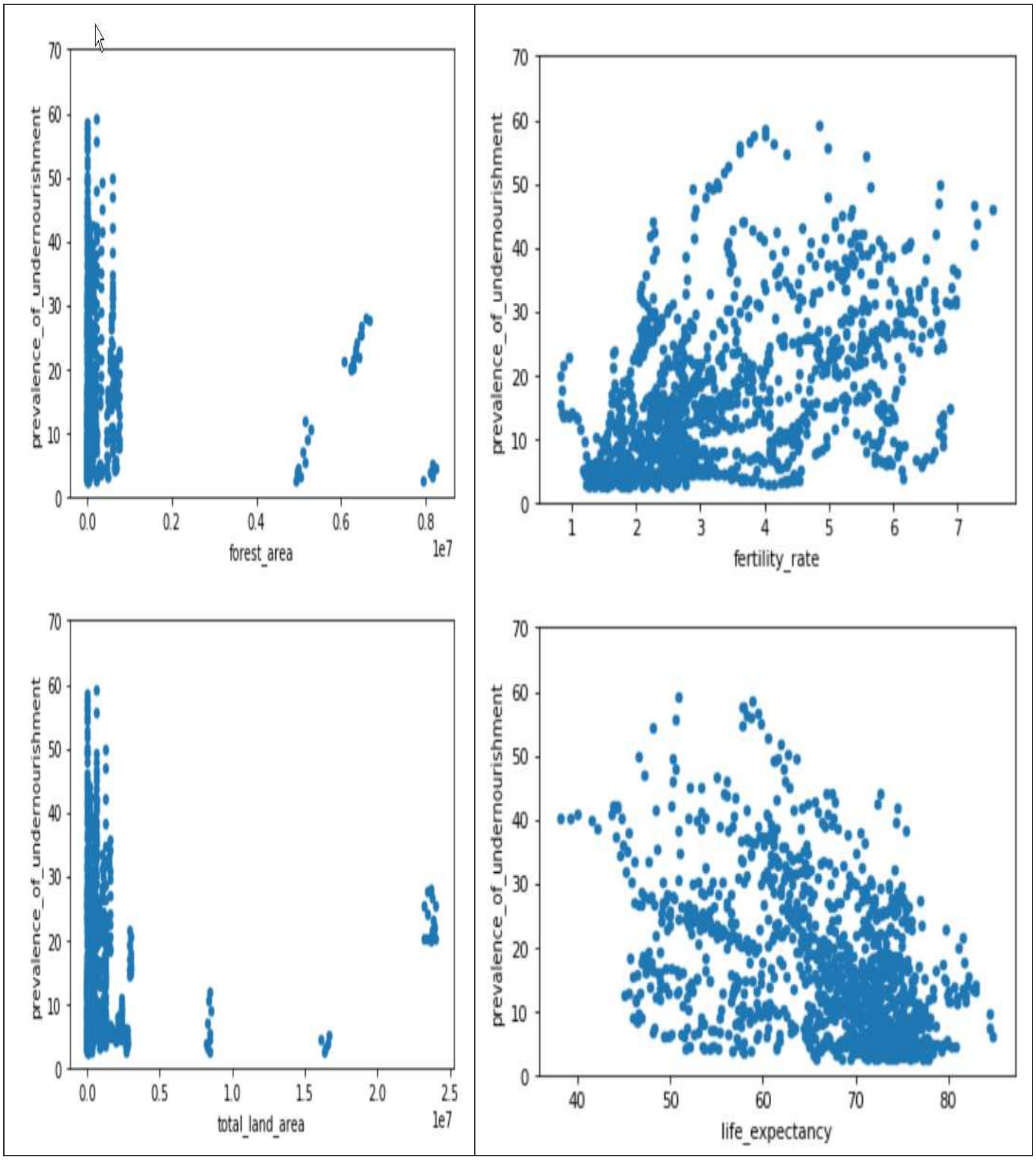
Since we are predicting the variable prevalence_of_nourishment, I summarized features which are highly correlated with it here:

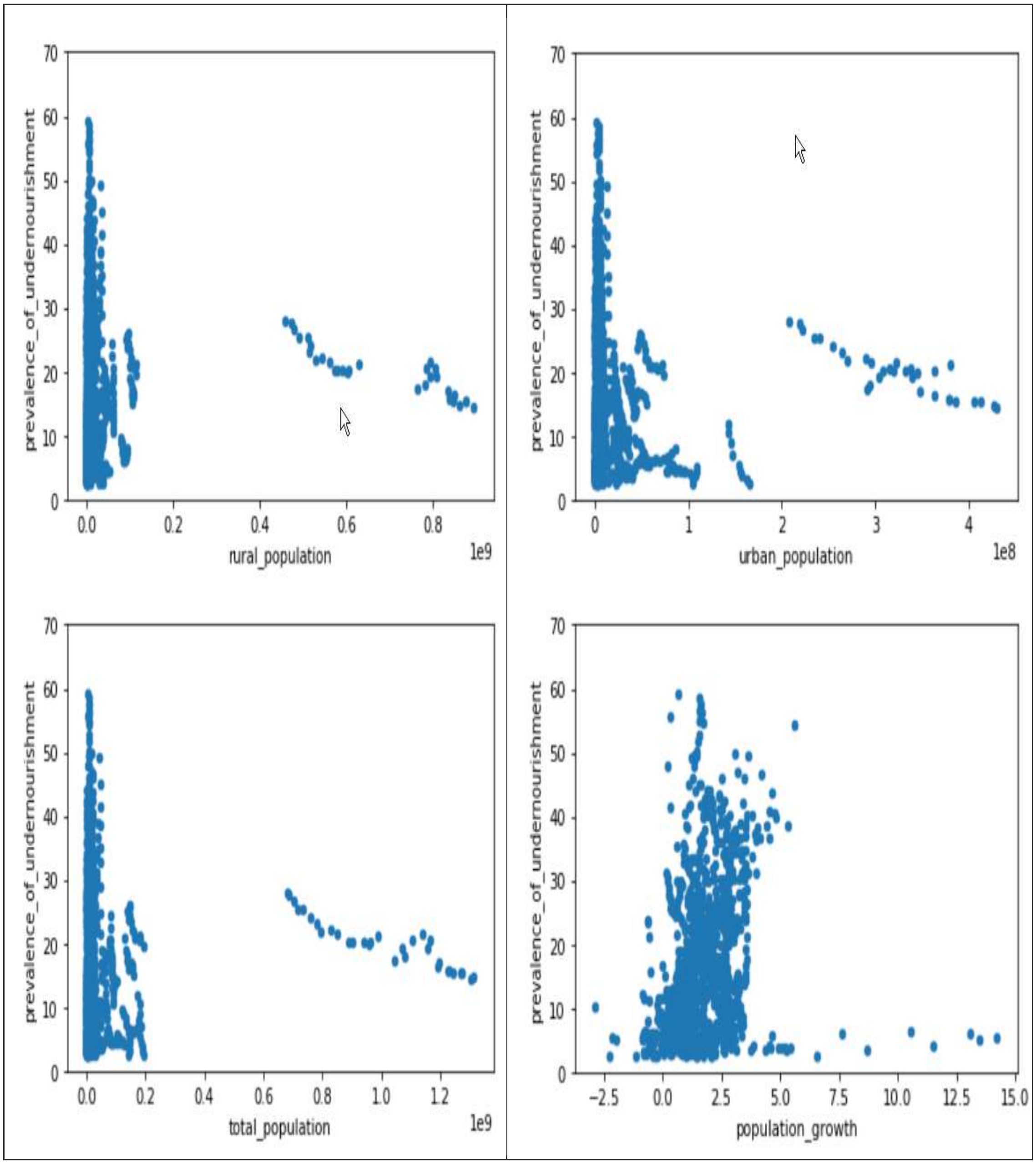
Feature Name	Correlation figure
year	-0.155572
percentage_of_arable_land_equipped_for_irrigation	-0.138048
cereal_yield	-0.249470
droughts_floods_extreme_temps	0.236992
forest_area	0.497108
population_growth	0.255205
avg_value_of_food_production	-0.389720
food_imports_as_share_of_merch_exports	0.181756
gross_domestic_product_per_capita_ppp	-0.335513
inequality_index	0.184799
net_oda_received_percent_gni	0.377888
per_capita_food_production_variability	-0.246283
adult_literacy_rate	-0.430649
school_enrollment_rate_female	-0.361153
school_enrollment_rate_total	-0.344756
avg_supply_of_protein_of_animal_origin	-0.542252
caloric_energy_from_cereals_roots_tubers	0.373514
access_to_improved_sanitation	-0.562945
access_to_improved_water_sources	-0.675150
anemia_prevalence	0.321443
obesity_prevalence	-0.600513
open_defecation	0.479173
hiv_incidence	-0.223817
rail_lines_density	-0.634076
access_to_electricity	-0.147333
co2_emissions	-0.189848
political_stability	-0.346913

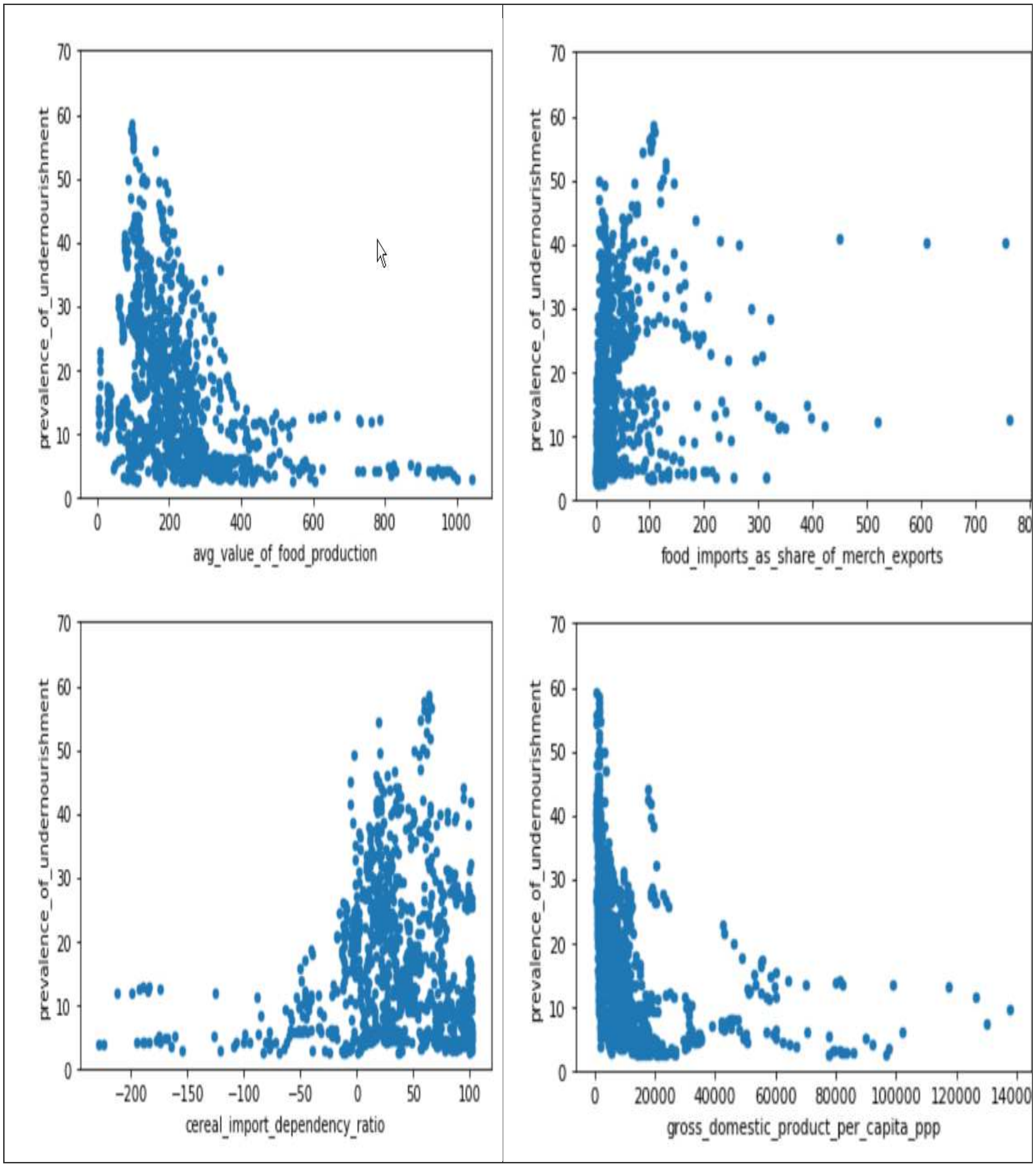
Table 1: Significant features for correlation with prevalence_of_undernourishment

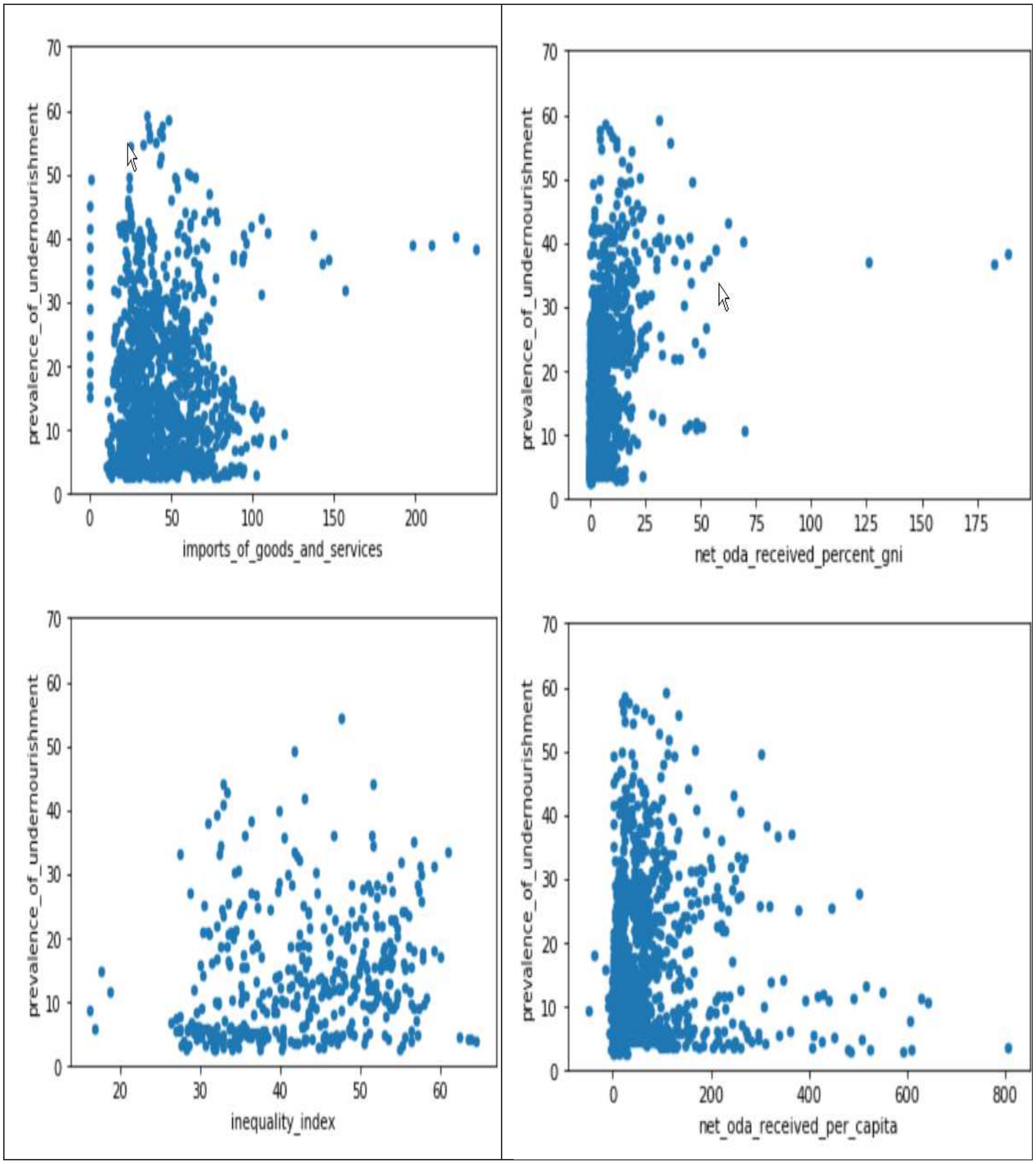
The following are scatterplot graphs to compare between each feature against undernourishment.

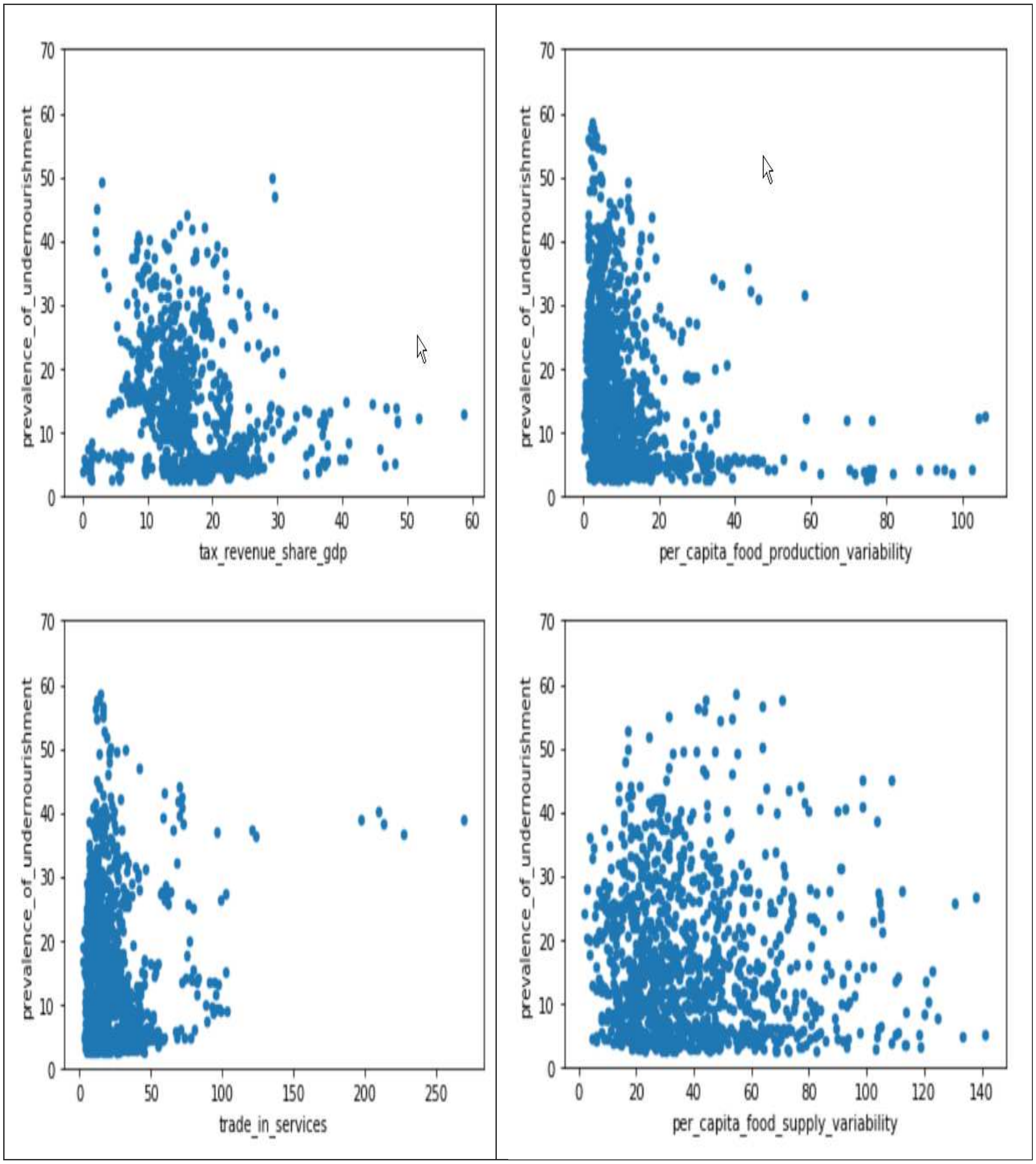

prevalence_of_undernourishment
agricultural_land_area
1e7
prevalence_of_undernourishment
cereal_yield
prevalence_of_undernourishment
percentage_of_arable_land_equipped_for_irrigation
prevalence_of_undernourishment
droughts_floods_extreme_temps

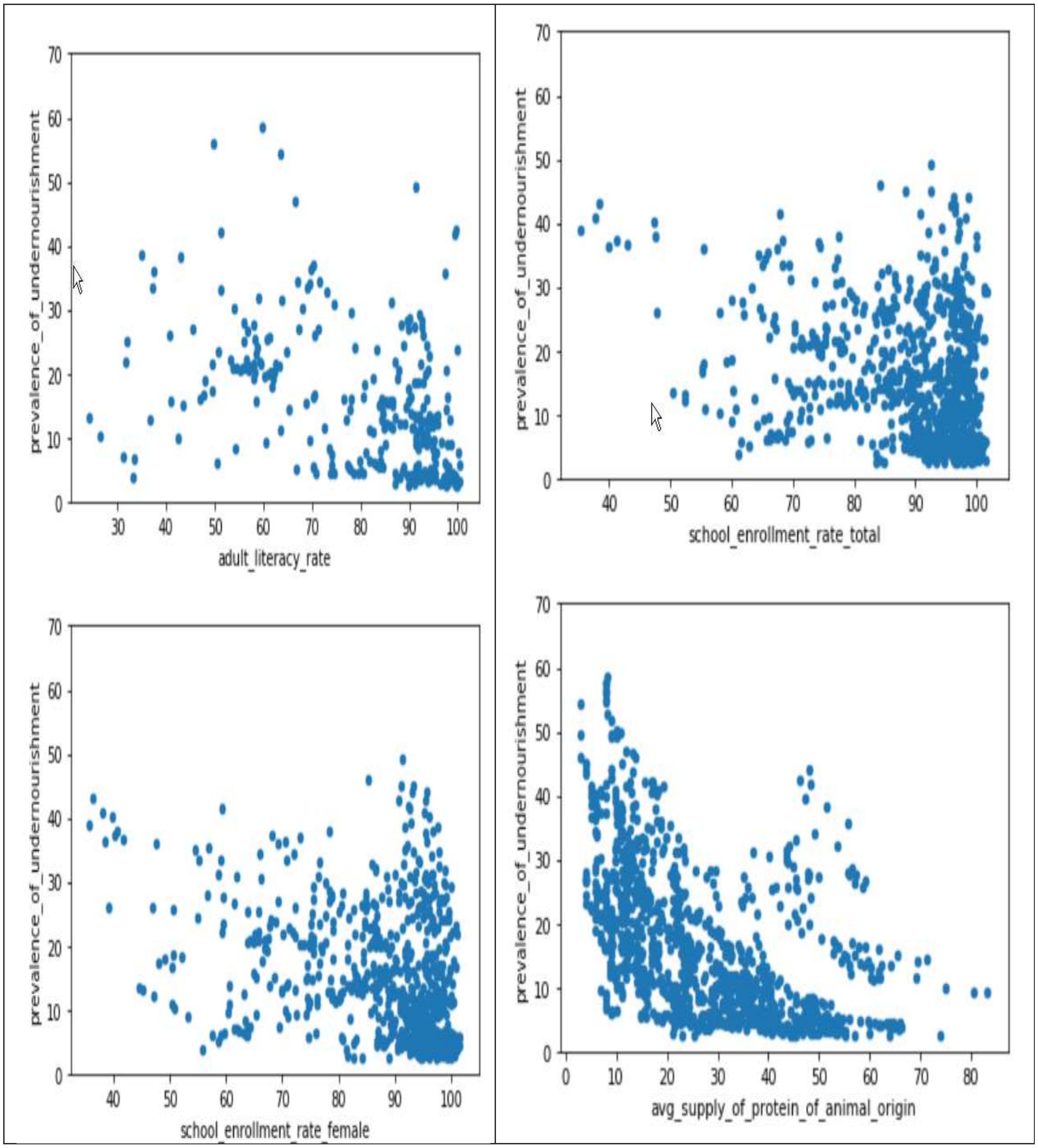

prevalence_of_undernourishment
forest_area
1e7
prevalence_of_undernourishment
fertility_rate
prevalence_of_undernourishment
total_land_area
1e7
prevalence_of_undernourishment
life_expectancy

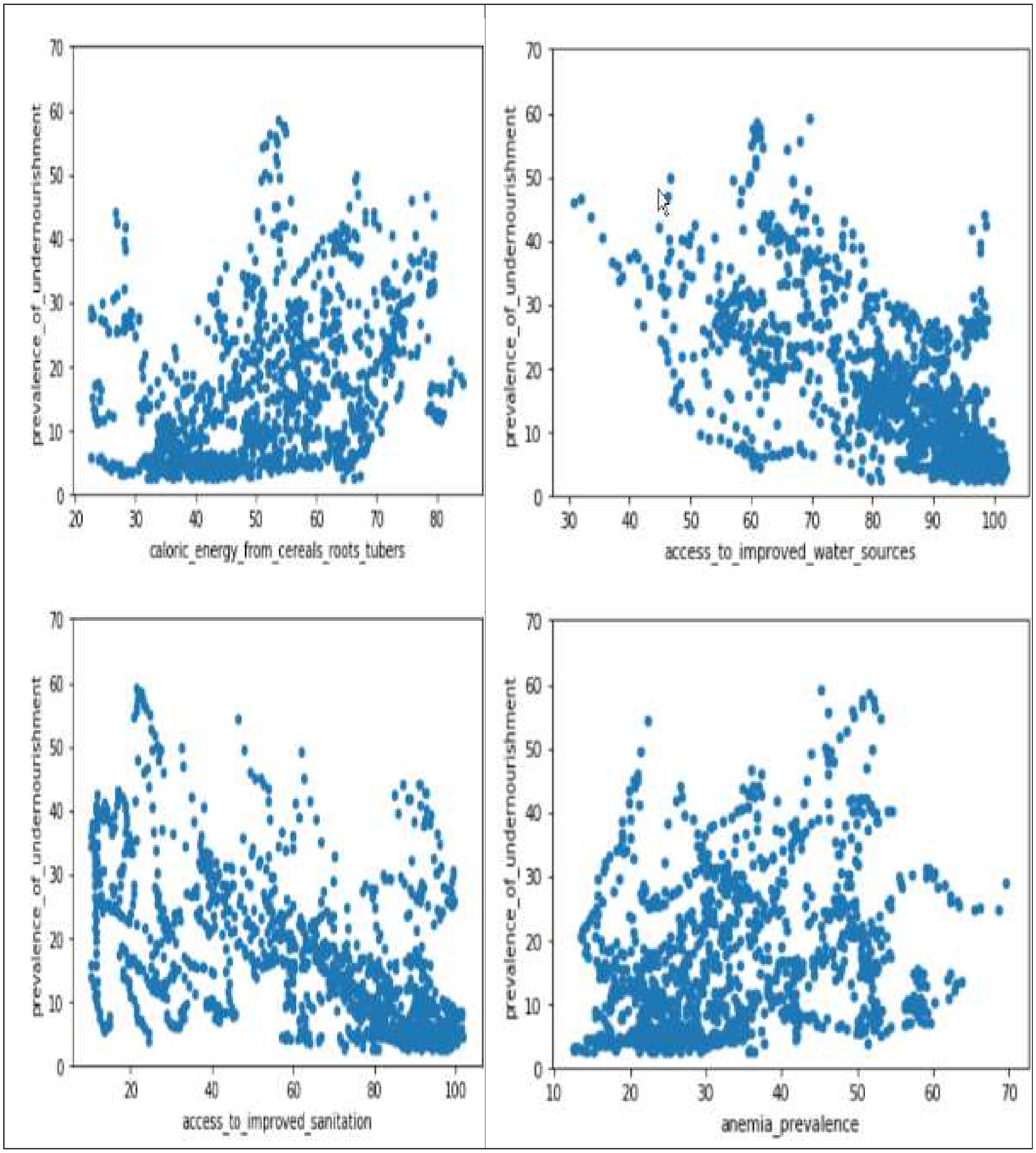

prevalence_of_undernourishment
rural_population
1e9
urban_population
1e8
total_population
1e9
population_growth

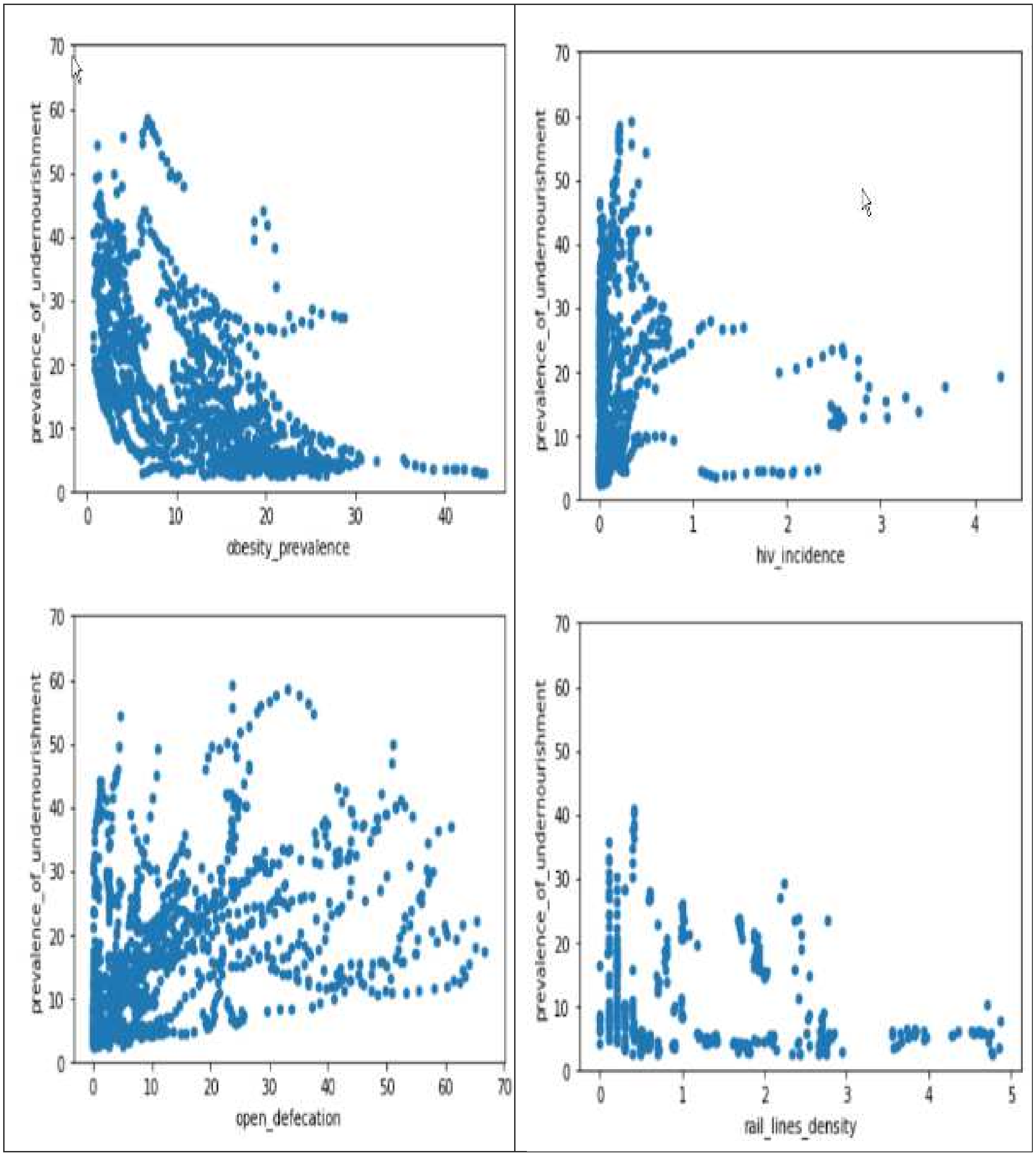

prevalence_of_undernourishment
avg_value_of_food_production
food_imports_as_share_of_merch_exports
cereal_import_dependency_ratio
gross_domestic_product_per_capita_ppp

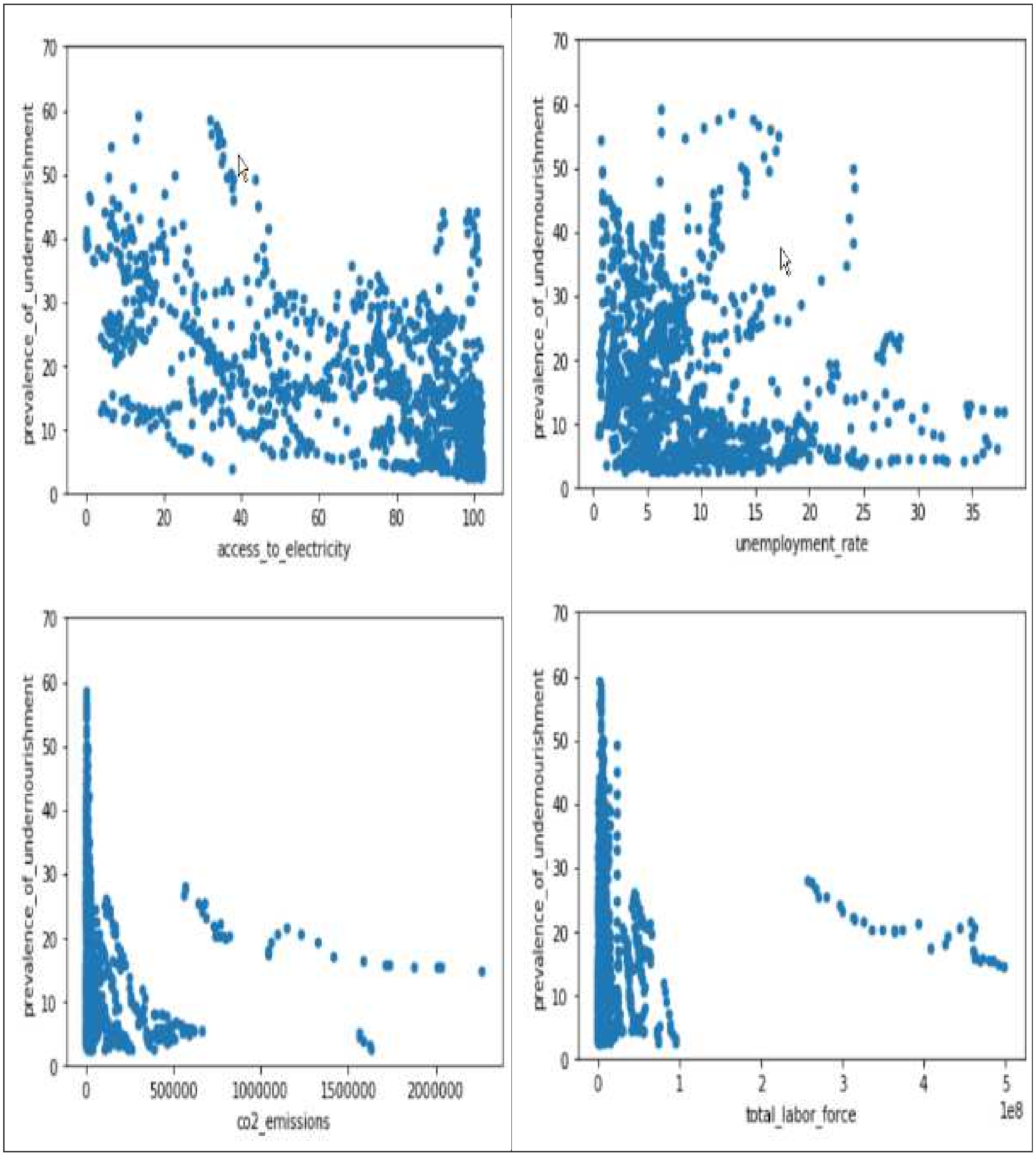

prevalence_of_undernourishment
imports_of_goods_and_services
net_oda_received_percent_gni
inequality_index
net_oda_received_per_capita


prevalence_of_undernourishment
tax_revenue_share_gdp
prevalence_of_undernourishment
per_capita_food_production_variability
prevalence_of_undernourishment
trade_in_services
prevalence_of_undernourishment
per_capita_food_supply_variability


prevalence_of_undernourishment
adult_literacy_rate
prevalence_of_undernourishment
school_enrollment_rate_total
prevalence_of_undernourishment
school_enrollment_rate_female
prevalence_of_undernourishment
avg_supply_of_protein_of_animal_origin


prevalence_of_undernourishment
caloric_energy_from_cereals_roots_tubers
prevalence_of_undernourishment
access_to_improved_water_sources
prevalence_of_undernourishment
access_to_improved_sanitation
prevalence_of_undernourishment
anemia_prevalence


prevalence_of_undernourishment
obesity_prevalence
hiv_incidence
open_defecation
rail_lines_density


prevalence_of_undernourishment
access_to_electricity
prevalence_of_undernourishment
unemployment_rate
prevalence_of_undernourishment
co2_emissions
prevalence_of_undernourishment
total_labor_force
1e8

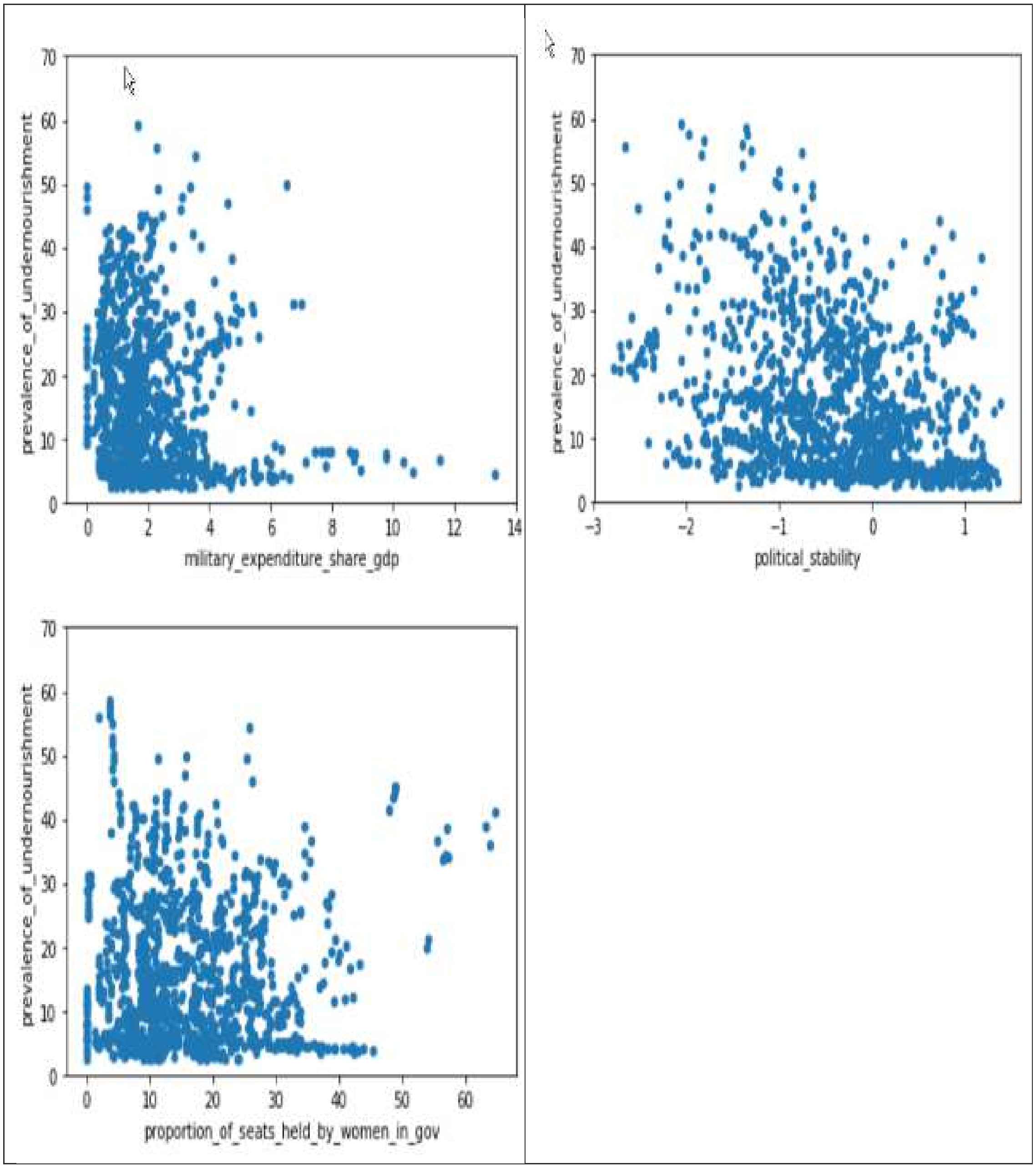


Figure 5: Comparison of independent features versus target

From all these graphs, we do not see any linearity evident, hence will need to do feature engineering and feature selection to determine which features has strong relationships with undernourishment.

The dataset has a significant percentage of NaN or no values. For example, there are 12 countries that has incomplete 16 years of data which will affect the training data stage. To complete the NaNs, I used data imputations of either zero or mean or median for each feature. This will change the distribution of each feature.

The graph below is the histogram for undernourishment. It is right skewed and majority are less than 20.

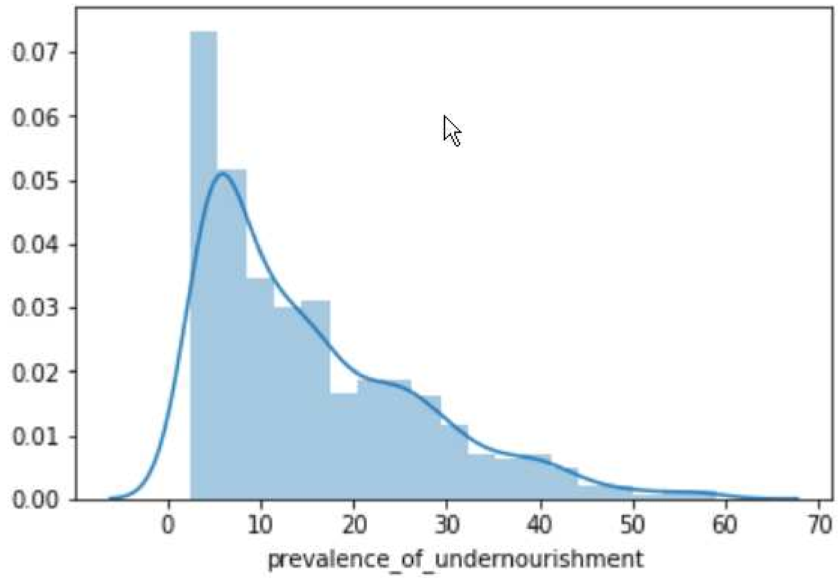


Figure 6: Distribution of prevalence_of_undernourishment

FEATURE ENGINEERING

After running some tests, below are the features will be used to predict undernourishment:

1. access_to_improved_water_sources

2. access_to_electricity

3. gross_domestic_product_per_capita_ppp

4. obesity_prevalence

5. avg_supply_of_protein_of_animal_origin

6. open_defecation

7. access_to_improved_sanitation

8. avg_value_of_food_production

The eight features are also checked for outliers and normalize on the same scale.

MODEL TRAINING

A regression model is used to predict undernourishment. Several algorithms were tested and decision forest is selected since it gives more accurate prediction.

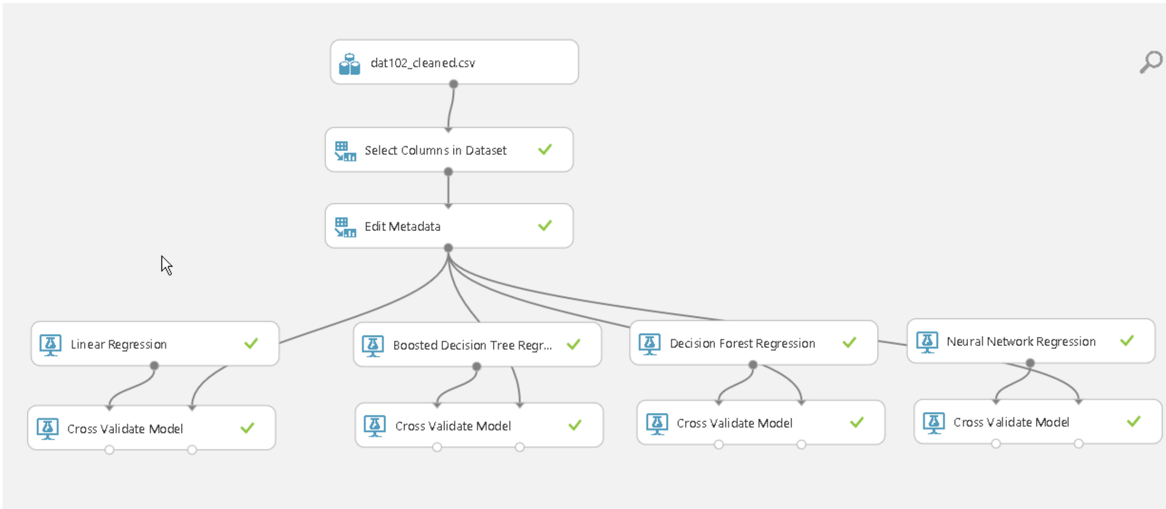


Figure 7: Cross-validation of each models to choose best algorithm

The model was trained with 80% of the data, and tested with the remaining 20%.

The Root Mean Square Error (RMSE) for the result is 2.91890.

CONCLUSION

This analysis has shown that the prevalence_of_undernourishment can be predicted using the eight features: access_to_improved_water_sources, access_to_electricity, gross_domestic_product_per_capita_ppp, obesity_prevalence, avg_supply_of_protein_of_animal_origin, open_defecation, access_to_improved_sanitation and avg_value_of_food_production are significantly effecting the label undernourishment.

To improve the model further, some suggestions can be considered:

1. Experimenting random imputation of values for missing data
2. More training data to be provided to the model
3. Eliminate or reduce NaNs inside dataset and accurate figures recorded
4. Testing with powerful algorithms like XGBoost
5. Feed the data into neural networks and do long training