

JIAHUI BONG, CELINE TAN, RHEA KENNETH Tutorial Group: DSF3



TABLE OF CONTENTS



01.Introduction

02. Data Cleaning

03. EDA + ML (Global)

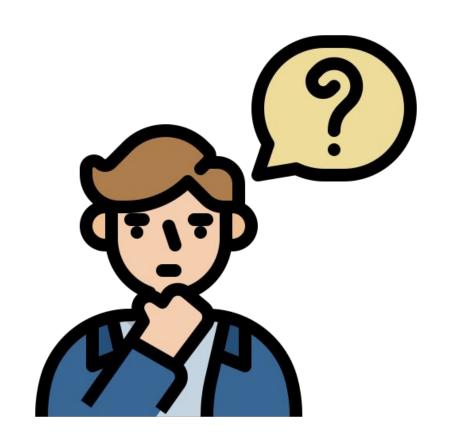
04. EDA + ML (Regional)

05. Conclusion



01

Introduction







Motivation

- → Stereotype Thinking: Developed countries have higher life expectancy while less developed countries have lower life expectancy
- → Eg. Africans have lower life expectancy due to malnutrition or other diseases
- → What factors contribute to such inequality?



Dataset Used





New Notebook

Lagrand → Download (245 kB)

:

WHO national life expectancy

Subset of factors influencing average life expectancy



- → Dataset from kaggle
- → Made up of GHO (Global Health Observatory) and UNESCO (United Nations Educational Scientific and Culture Organization) Information





Does life expectancy inequality really exists and if yes, what are the indicators of such scenarios and are there possible solutions to narrow the inequality?





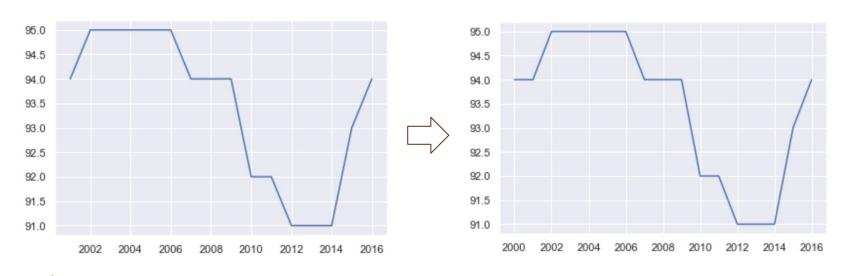
02

Data Cleaning





Linear Interpolation Method





Bar Plot of Number of Null Values vs Countries Some countries have all null values



Fill_na method

- → Datas are in time series
- Use linear interpolation method to fill in the null values
- → If there are more than 10 years of values are missing then will leave that country's data unfilled (which will be removed later after EDA)





03 Global Life Expectancy





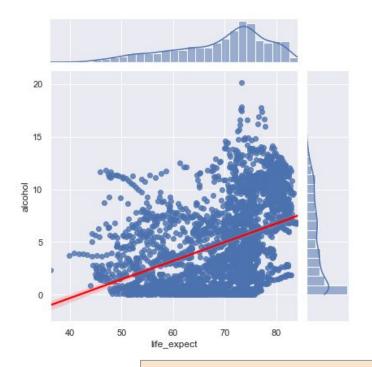
Exploratory data analysis

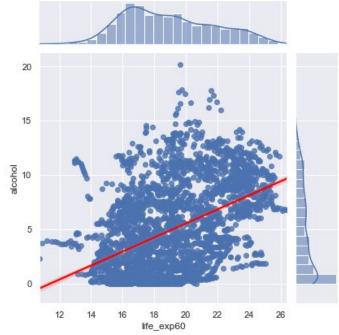
- 1) Heatmap
- 2) Insights
- 3) Estimated top predictors of life expectancy



															u	<u> </u>						
		life_expect	1.00	0.89	-0.95	-0.93	-0.86	0.40	0.60	-0.57	0.62	0.40	0.64	0.65	0.64	0.83	0.67	0.58	0.61	0.27	0.03	-0.51
		life_exp60	0.89	1.00	-0.78	-0.77	-0.66	0.47	0.46	-0.56	0.55	0.30	0.51	0.50	0.50	0.68	0.59	0.57	0.67	0.38	0.00	-0.34
а	dult	_mort_rate	-0.95	-0.78	1.00	0.81	0.76	-0.25	-0.52	0.46	-0.57	-0.33	-0.53	-0.54	-0.53	-0.73	-0.57	-0.53		-0.17	-0.06	0.68
ir	nfan	t_mort_rate	-0.93	-0.77	0.81	1.00	0.91		-0.64		-0.60		-0.72	-0.73		-0.85	-0.65	-0.53		-0.23	-0.00	0.34
а	ge1-	4mort_rate	-0.86	-0.66	0.76	0.91	1.00	-0.31	-0.62		-0.54	-0.42	-0.70	-0.70	-0.69	-0.80	-0.55	-0.41		-0.15	-0.03	0.29
		alcohol	0.40	0.47	-0.25	-0.45	-0.31	1.00	0.28		0.14	0.14	0.30	0.30	0.30	0.41	0.56	0.27	0.53	0.37	-0.04	-0.04
		brni	0.60	0.46	-0.52	-0.64	-0.62	0.28	1.00	-0.69	0.81	0.36	0.47	0.44	0.45	0.67	0.50	0.41	0.46	0.24	-0.17	-0.16
		age5-19thinness	-0.57	-0.56	0.46	0.58		-0.49	-0.69	1.00		-0.34	-0.44	-0.42	-0.42		-0.59	-0.36	-0.57	-0.39	0.26	0.19
	aç	ge5-19obesity	0.62	0.55	-0.57	-0.60	-0.54	0.14	0.81	-0.55	1.00	0.34	0.40	0.38	0.38	0.59	0.39	0.57	0.43	0.20	-0.04	-0.24
		hepatitis	0.40	0.30	-0.33		-0.42	0.14	0.36	-0.34	0.34	1.00	0.65	0.65	0.67	0.40	0.26	0.23	0.22	0.11	-0.11	-0.05
		measies	0.64	0.51	-0.53	-0.72	-0.70	0.30	0.47	-0.44	0.40	0.65	1.00	0.92	0.92	0.66	0.44	0.35	0.40	0.20	-0.01	-0.11
		polio	0.65		-0.54	-0.73	-0.70	0.30	0.44	-0.42	0.38	0.65	0.92	1.00	0.96	0.66	0.45	0.36	0.42	0.22	-0.03	-0.11
		dphtheria	0.64			-0.72	-0.69	0.30	0.45	-0.42	0.38	0.67	0.92	0.96	1.00		0.42	0.35	0.41	0.22	-0.03	-0.09
		Basic_water	0.83	0.68	-0.73	-0.85	-0.80	0.41	0.67		0.59	0.40	0.66	0.66	0.65	1.00	0.64	0.52	0.51	0.20	0.03	-0.29
		doctors	0.67		-0.57	-0.65	-0.55	0.56	0.50	-0.59	0.39	0.26	0.44	0.45	0.42	0.64	1.00	0.44	0.65	0.45	-0.08	-0.34
		gni_copita	0.58		-0.53	-0.53	-0.41	0.27	0.41	-0.36	0.57	0.23	0.35	0.36	0.35	0.52	0.44	1.00	0.41	0.11	-0.04	-0.19
		ggbe-d	0.61	0.67		-0.58		0.53	0.46	-0.57	0.43	0.22	0.40	0.42	0.41		0.65	0.41	1.00	0.68	-0.08	-0.09
		dhe_gdp	0.27	0.38	-0.17	-0.23	-0.15	0.37	0.24	-0.39	0.20	0.11	0.20	0.22	0.22	0.20	0.45	0.11	0.68	1.00	-0.07	0.05
		une_pop	0.03	0.00	-0.06	-0.00	-0.03	-0.04	-0.17	0.26	-0.04	-0.11	-0.01	-0.03	-0.03	0.03	-0.08	-0.04	-0.08	-0.07	1.00	-0.10
		une_hiv	-0.51	-0.34	0.68	0.34	0.29	-0.04	-0.16	0.19	-0.24	-0.05	-0.11	-0.11	-0.09	-0.29	-0.34	-0.19	-0.09	0.05	-0.10	1.00
			lfe_expect	18_exp60	dul_mort_rate	ant_mont_rate	1-4_mort_rate	alcohol	Ē	95-19thirness	aped-19obesity	hepatits	reasies	olog	dpitheria	basic_water	doctors	gni capita	p-syddi	de ad	dogram	une hiv

Alcohol





Median of alcohol consumption

4.02 liters per capita



Insights

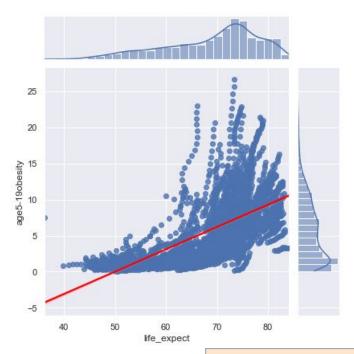
Expectation

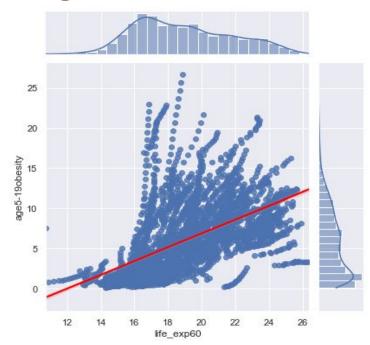
Alcohol is associated with health issues and has a **negative relation** with life expectancy

Reality

Alcohol in **moderation** can have beneficial health effects leading to longevity

Obesity rates from ages 5-19





Median of obesity rate:

5.20 %

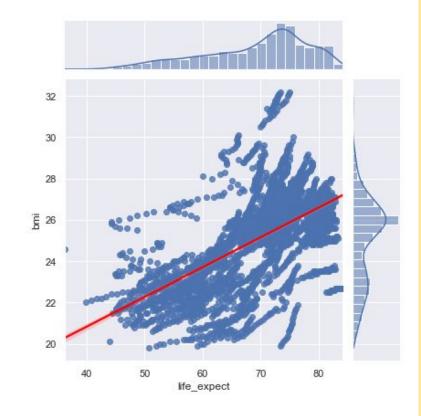
Insights

Expectation

Obesity is considered a health issue and is expected to have a **negative relation** with life expectancy

Reality

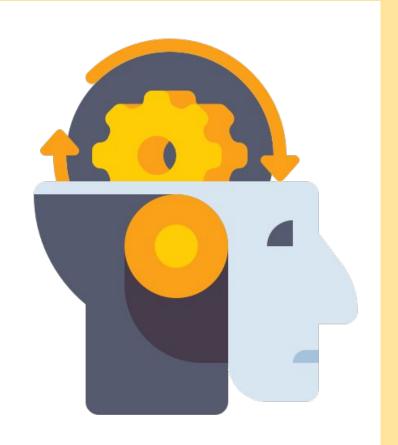
- The BMI of those >30 makes up only about 5.97% of population on average.
- 2) BMI also indicates **muscle mass** and **not only body fat**





Machine Learning

- Multivariate Linear Regression
- 2) Random Forest Regressor
- 3) Gradient Boosting Regressor





RESPONSE: Life Expectancy At Birth

Train Data

Score	Linear Regression	Random Forest	Gradient Boosting
R-squared value (R²)	0.986	0.999	0.999
Mean Squared Error (MSE)	1.140	0.056	0.028

Test Data

Score	Linear Regression	Random Forest	Gradient Boosting
R-squared value (R²)	0.986	0.995	0.996
Mean Squared Error (MSE)	1.160	0.361	0.316

RESPONSE: Life Expectancy At Birth

Train Data

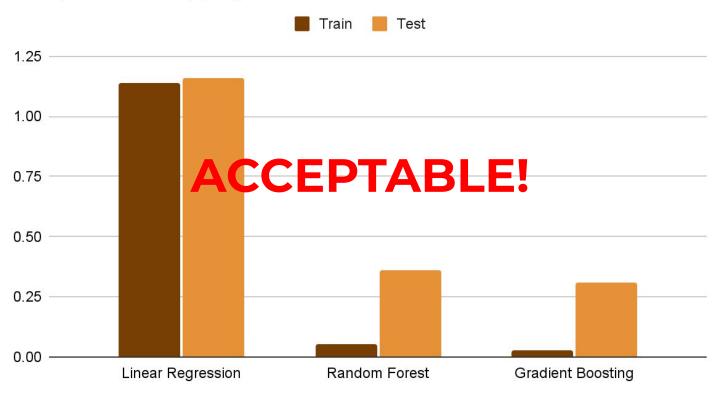
Score	Linear Regression	Random Forest	Gradient Boosting
R-squared value (R²)	0.986	0.999	0.999
Mean Squared Error (MSE)	1.140	0.056	0.028

Test Data

Score	Linear Regression	Random Forest	Gradient Boosting
R-squared value (R²)	0.986	0.995	0.996
Mean Squared Error (MSE)	1.160	0.361	0.316

Comparing MSE for train and test data

Response = life_expect



RESPONSE: Life Expectancy At Age 60

Train Data

Score	Linear Regression	Random Forest	Gradient Boosting
R-squared value (R²)	0.824	0.990	0.999
Mean Squared Error (MSE)	1.482	0.082	0.001

Test Data

Score	Linear Regression	Random Forest	Gradient Boosting
R-squared value (R²)	0.848	0.972	0.983
Mean Squared Error (MSE)	1.412	0.254	0.159

Comparing MSE for train and test data

Linear Regression



Random Forest

Gradient Boosting

0.0

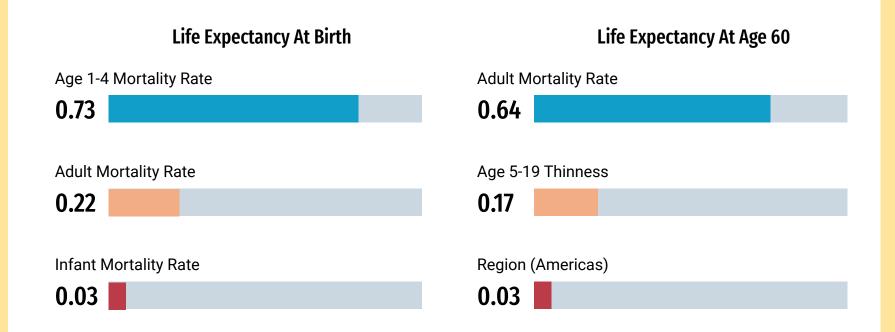
Top Predictors

Top 3 predictors for each response variable inferred from GBR





Gradient Boosting Regressor





04 Regional Life Expectancy

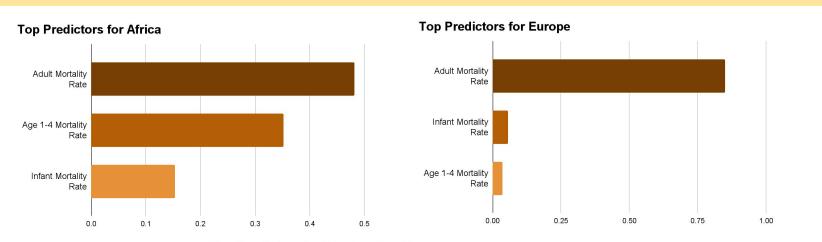




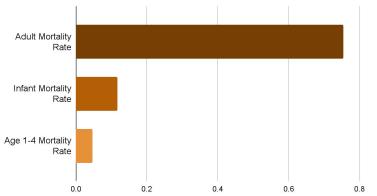
Exploratory data analysis

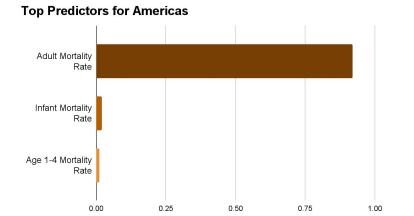
Top predictors for each region's life expectancy at birth inferred from GBR



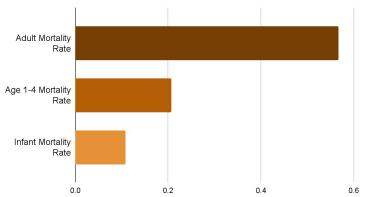






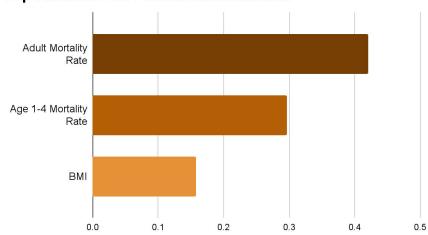






EASTERN MEDITERRANEAN

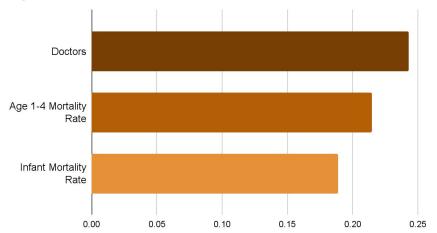
Top Predictors for Eastern Mediterranean



Score	Train Data	Test Data
R-squared value (R²)	0.999	0.991
Mean Squared Error (MSE)	0.019	0.305

SOUTH-EAST ASIA

Top Predictors for SEA



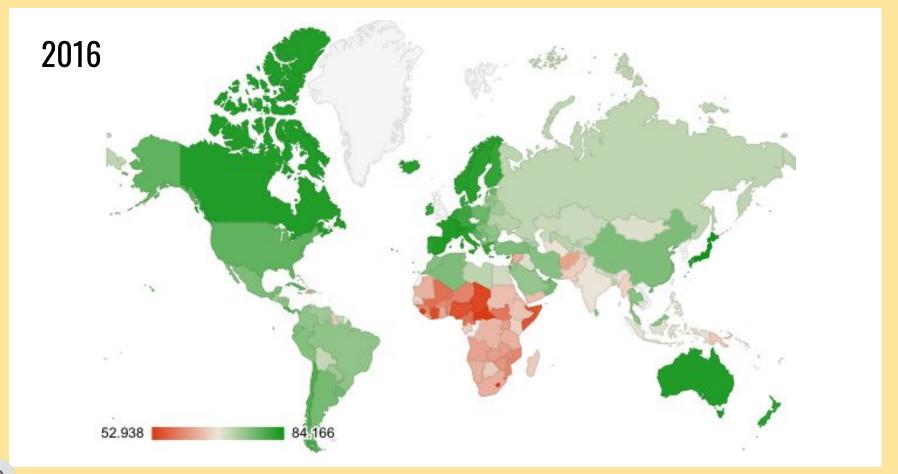
Score	Train Data	Test Data
R-squared value (R²)	0.999	0.930
Mean Squared Error (MSE)	0.004	0.643



05Conclusion









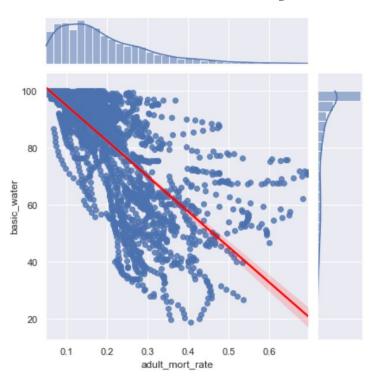
Indicators affecting life expectancy

- → Mortality rates are common top predictors across regions
- → High negative correlation between life expectancy vs mortality rate of different group age
- → Decrease mortality rate but HOW?





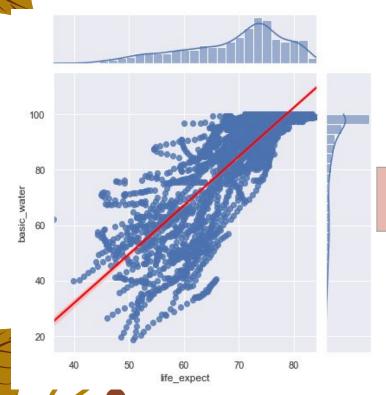




- → High negative correlation
- → Rationale: Water acts as solvent to transport nutrients to body cells



Basic water

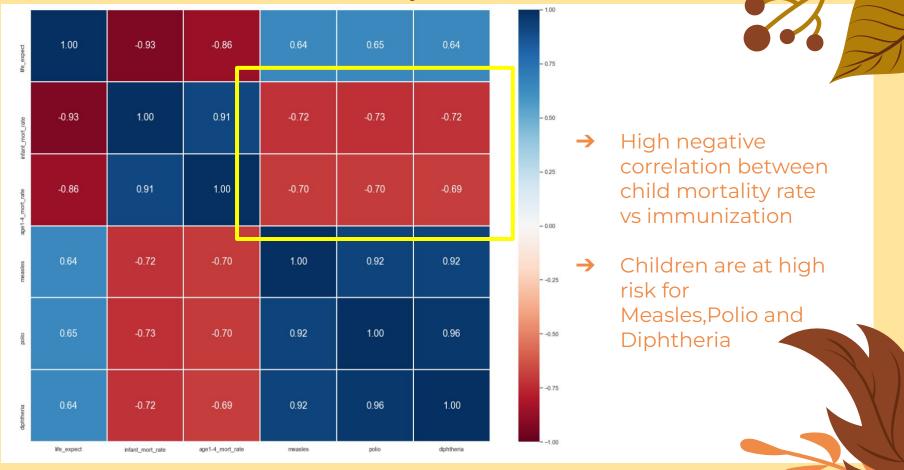


Correlation with life_exp:

0.83



Correlation between child mortality rate and immunization





Possible Implementation



Water Infrastructure

→ Government should allocate more budget to water infrastructure



Vaccination

→ Raise awareness on the importance of immunisation in children

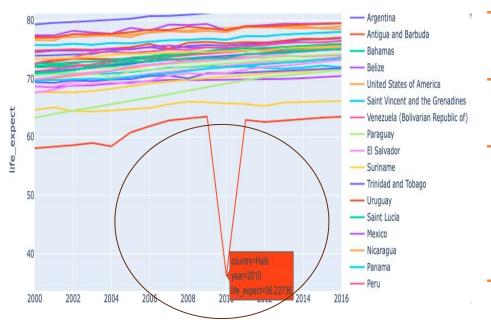
Goal

To improve life expectancy in less developed country and narrow the inequality gap





Limitation of Proposed Solution



- → Sudden drop in life expectancy in **Haiti** in **2010**
 - Caused by a 7.0 magnitudeearthquake
- → NO possible implementation when low life expectancy is caused by unpredictable circumstances
 - Need to take **geographic factors** into consideration



THANK YOU!

