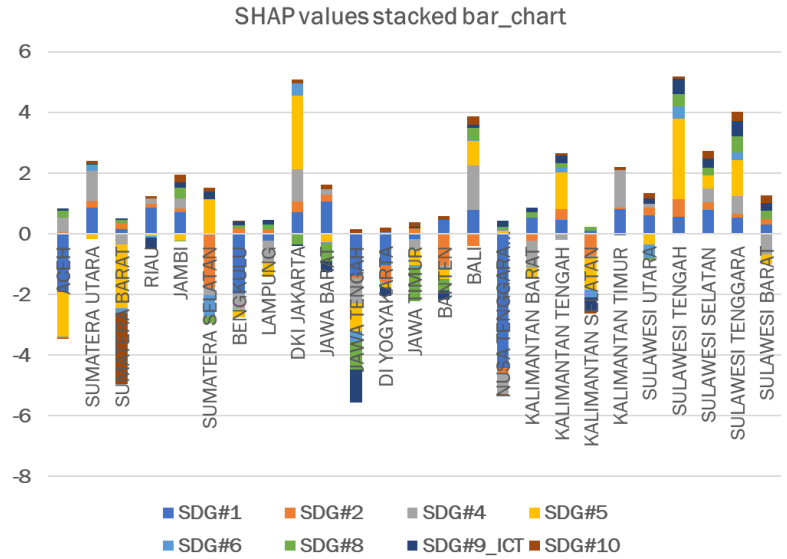
国際会議でのAI分析発表プレゼンの手法

2024年6月29日

学習院大学経済学部

白田由香利



☆1: 自分の名前と所属は覚えてもらうために大きな声ではっきりと

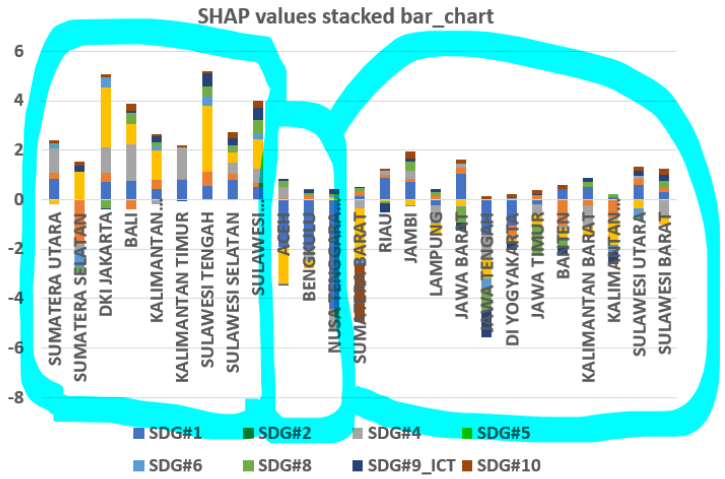
- Today I would like to talk about XXXXX. My name is Prof Shirota from Gakushuin University. I am a professor of Economics Faculty.
- **発表の目的は，自分が発見した成果を皆さんに披露したいため**
- **プレゼンは「皆さん是非私の話を聞いてください」「私は皆さんにお話しをしたいのです」と，話したいオーラを出す.**
- **気合いを入れてから壇上へ．気合いの入っていない発表は，極端に言えば，する必要が感じられない**

☆2: CONTENTSは必ず述べる

- 目次はきちんと説明する.
- ~~本日はこのような内容で話します~~，はダメ
- Firstly, Secondly, Finally I will conclude this presentation.
- 目次項目が変わるたびにこの同じ目次ページを入れること

他人の話は，わかりづらい.
構成をきちんと話して，
今何を話しているのか，
しつこく説明しないと
相手は分からない

Contents



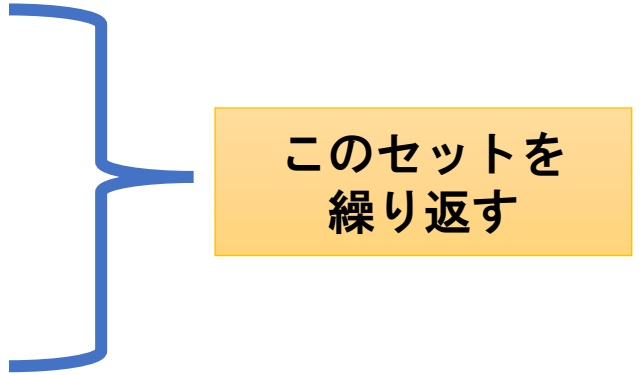
Let me move on to the next topic which is XXX,と言いながらページをめくる

☆3:　CONTENSの中に，DATAとMETHODSは必ず入れる

- 分析ではデータと手法
- 悪い例：回帰なのかクラスタリングなのか，言わずにSHAPの値を説明する．
- 悪い例：回帰分析で何がターゲットなのか説明変数なのか言わない
- データの出典は必須．同じ実験が誰でも可能なように．

- 基本の構成

1. Research Objective
2. Data
3. Methods
4. Result
5. Evaluation
6. Conclusions



☆4:　分析の目的を明確に述べる

目的例

- ●教育効果を上げるためにどこに投資すべきか
- ●Gender equalityの進んでいる地域の特性を探求したい
- ●棚卸資産回転率が株価成長に貢献していることを証明したい
- ●成長率の高い企業の特質を発見したい

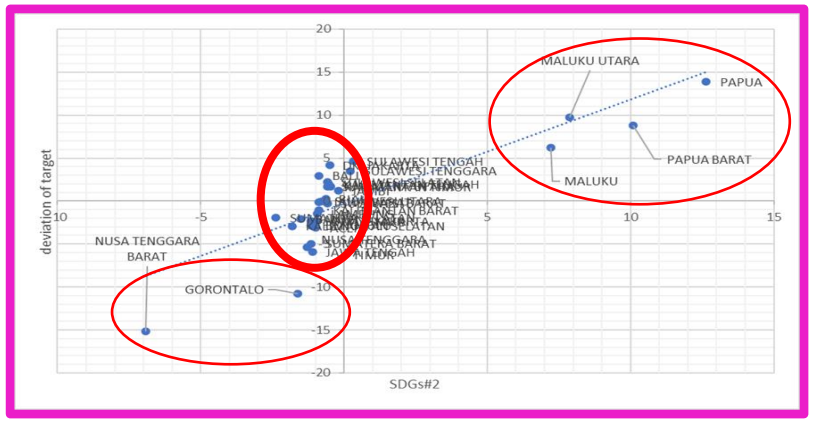
- 目的がはっきりしていないと手法とデータが適切であるか判断できない
- この目的のために，このようなアプローチで分析をします，という論理的筋道

☆5:　視覚的に見せる．グラフで勝負！

- **一番重要な結果のグラフ**は，タイトルページや，CONTENTSの余白に張り付けておく．
- 発表開始までの待ち時間に「これは何だろう」と興味をもたせる．
- グラフがよくできていれば，少々英語が下手でも伝わる

☆6:グラフは軸の説明と単位が必須項目

- そもそも，何のグラフか必ず言うこと．
- 発表でも，軸の説明を口頭でも必ず行う
- This axis shows と言って，軸をレーザーポインタでなぞる



- 「結果はこうなります」だけではダメ
- 他人はどこがすごいのか，言われなくては分かりません．
- The largest target value is this というようにポイントする．
- 注目点は，予め，円で囲むか，矢印をつけておく．ひと手間を惜しんではいけない．

☆7:　問いかけを効果的に入れて，話についてこさせる

- Among the 3 clusters, which cluster has the highest performance?
- Which pair offers higher similarity?
- What is this data?
- **考えさせて，理解させる．この間が重要．**

- **値を２つ並べられても，大きい方が優れているのか，小さいほうが優れているのか，聞いている人には判断がつかない．**
- **どっちが優れているの？？？**
- **しっかりと「こちらが良い」と言う．**

☆8:　原稿を読まない．観客に背中をむけない．

- ぼそぼそ原稿を読むだけでは観客は寝るか，ケータイを始める．
- そもそも，何か伝えたいという積極的態度がないのがいけない．
- 自分が伝えたいことを，相手に伝える．
- 英語が話せないなら，英語をｐｐｔに書き込んでおく
- 目を合わせなくてもよいので，Are you with me? と心の中で念じながら，時々，聴衆のほうを見るともなく見る．

☆9:　英語は伝わるように発音する

- 発音の訓練は時間がかかるので日頃からトレーニングする
- イギリス英語よりもアメリカ英語のほうが伝わりやすい
  ZOOMプレゼンで痛感した.

- 間違った発音でも，大きく自信をもって話そう
- 声が大きいほうが，よく伝わる

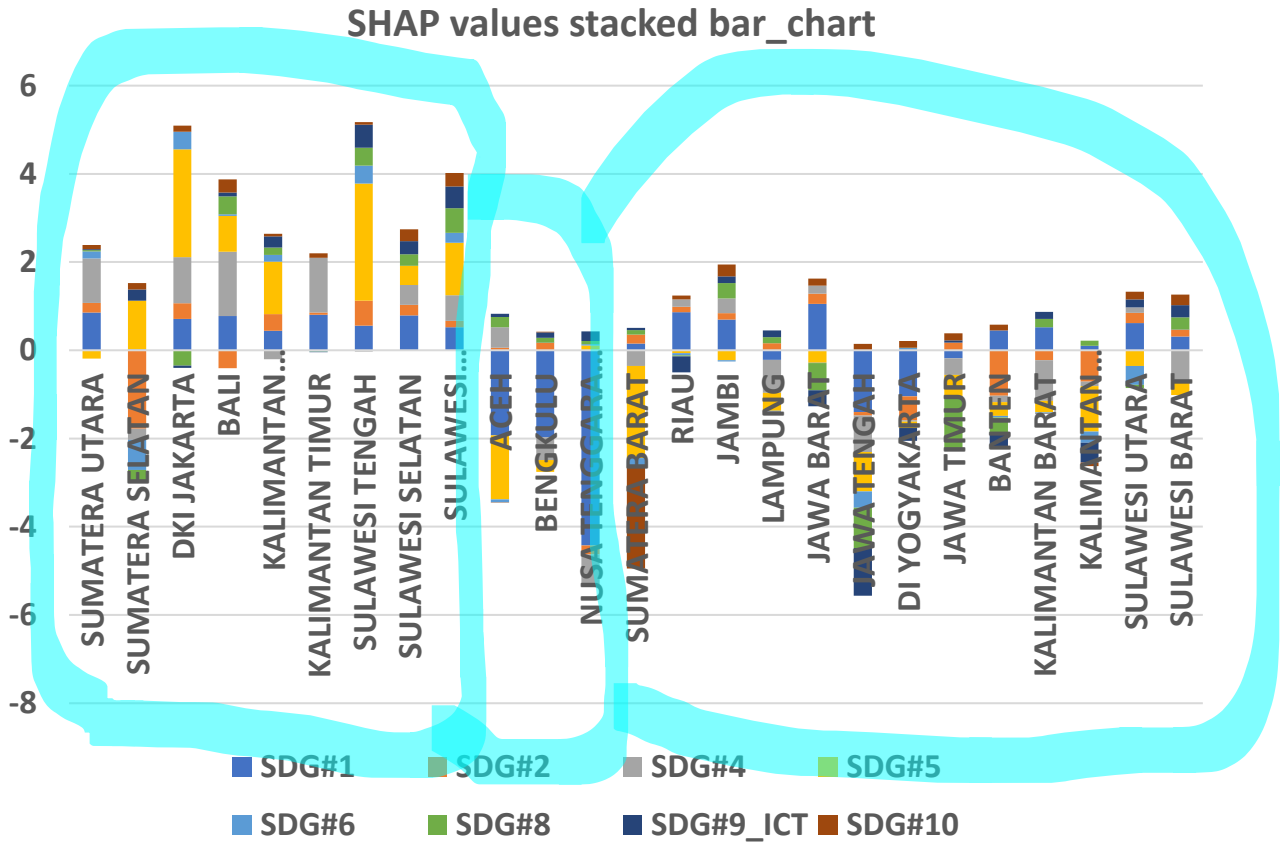
☆10:　要所要所でそこまでのCONCLUSIONS 最後のCONCLUSIONSは時間たっぷりとる

- ゆっくり話していると時間がなくなる
- CONCLUSIONSに時間をたっぷり残す
- CONCLUSIONSでは，そこだけ聞いても話の内容がわかるように要にして完，にまとめる
- 重要なグラフはべたべた張り付ける
- まとめは，今後のプロジェクト資金がとれるかとれないかの瀬戸際と思って（そうでなくとも）真剣に伝える.

Shapley-based Analysis of Dominant Factors for Well-being Achievement by Indonesia Provinces

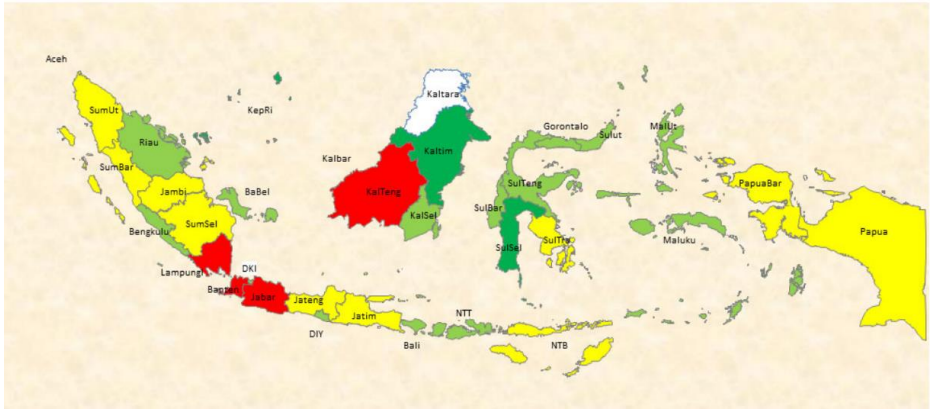
Contents

1. Research Objective
2. Data Selection
3. Regression: XGBOOST
4. SHAP
5. Clustering by SHAP Values
6. Conclusion



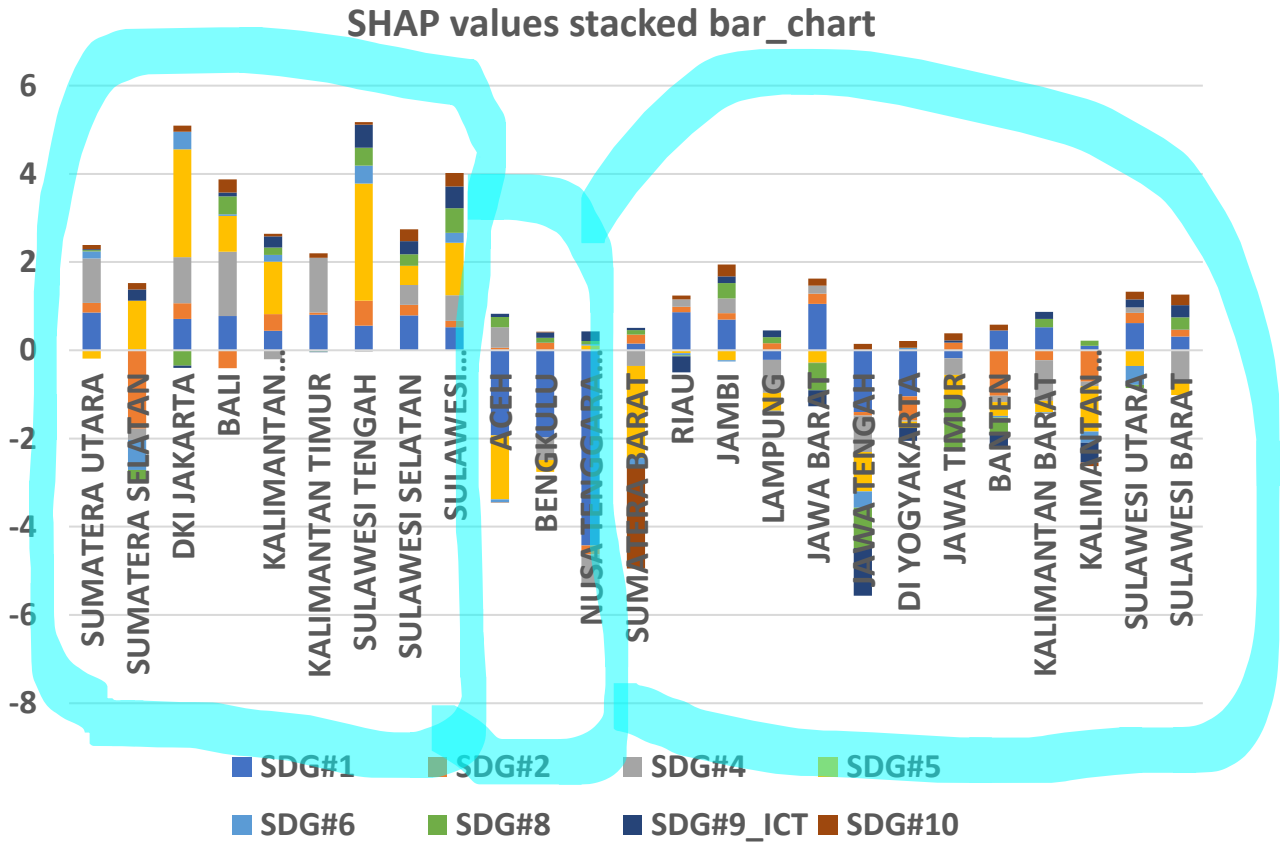
Research Objective and Approach

- Among 17 SDG scores, which factors are dominant to increase well-being scores?
- Effective investment by the government
- Data: data by provinces from ***BADAN PUSAT STATISTIK***
- Method: AI-based regression XGBoost
- XAI (eXplainable AI) tool: SHAP



Contents

1. Research Objective
2. Data Selection
3. Regression: XGBOOST
4. SHAP
5. Clustering by SHAP Values
6. Conclusion



Data Selection

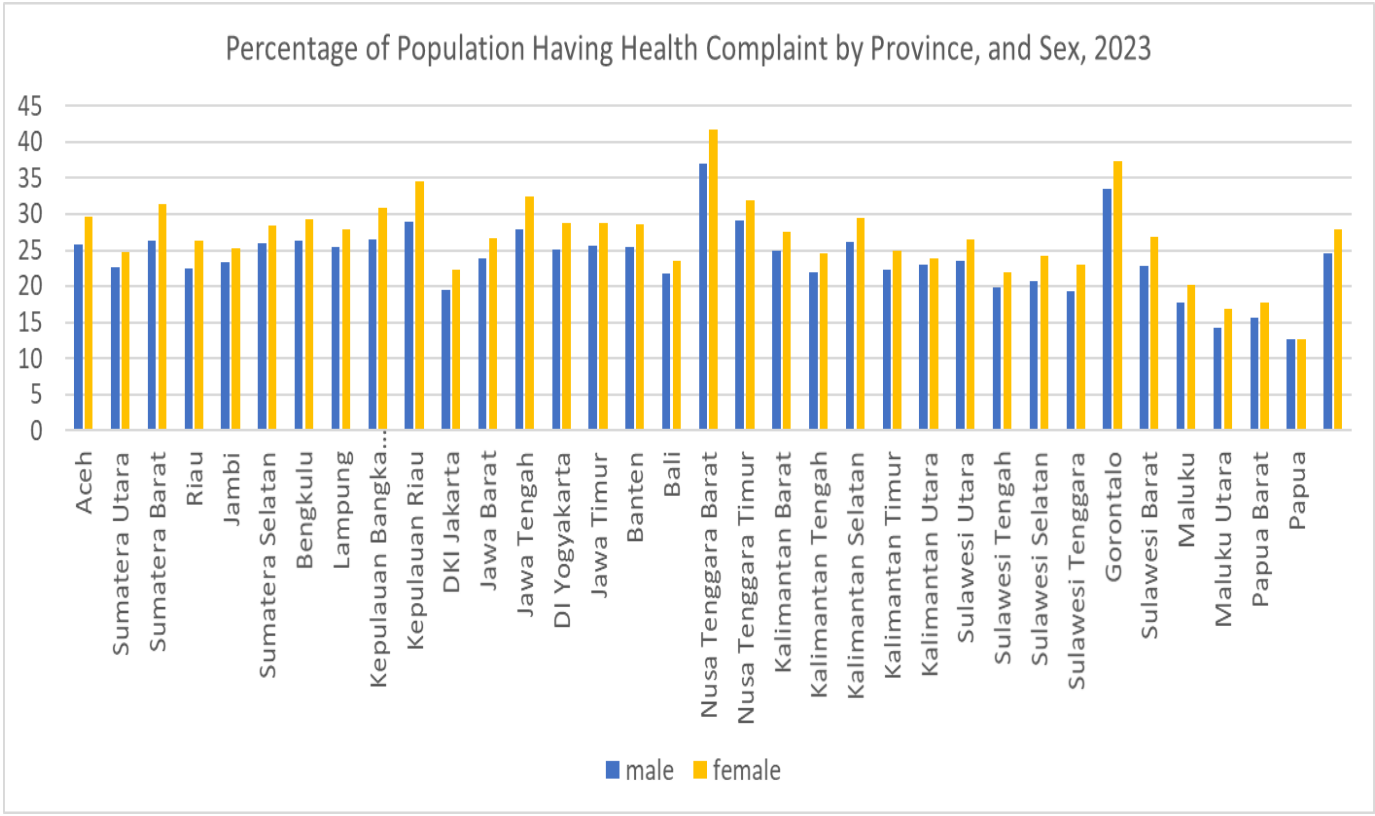
- SDGs#1: [NotPoor%]=100 – [Percentage of Poor Population (P0) by Province and Area (Percent), cited from [1]]
- SDGs#2: NotHungryLevel, [Daily Average Consumption of Calorie and Protein per Capita]
- SDGs#3: [Well-being%]=100 - [Percentage of Population Having Health Complaint by females]
- SDGs#4: [Completion Rate by Educational Level and Province: Senior High School %]
- SDGs#5: [Gender Empowerment Index]
- SDGs#6 [Households Using Safely Managed Sanitation Services (%)]
- SDGs#8: [GDP%] = [[2010 Version] Distribution of GRDP to Total GRDP of 34 Provinces at Current Market Prices by Province (Percent)]
- SDGs#9: [ICT%} = [Proportion of Adults (Aged 15-59 Years) with Information and Communication Technology Skills (Percent)]
- SDGs#10: [NoEconomicDisparityLevel] = (-1) Standardized value of Gini index

Cited: https://www.unicef.org/indonesia/reports/sustainable-development-goals-provincial-briefs

Percentage of Population Having Health Complaints by females as Well-being index

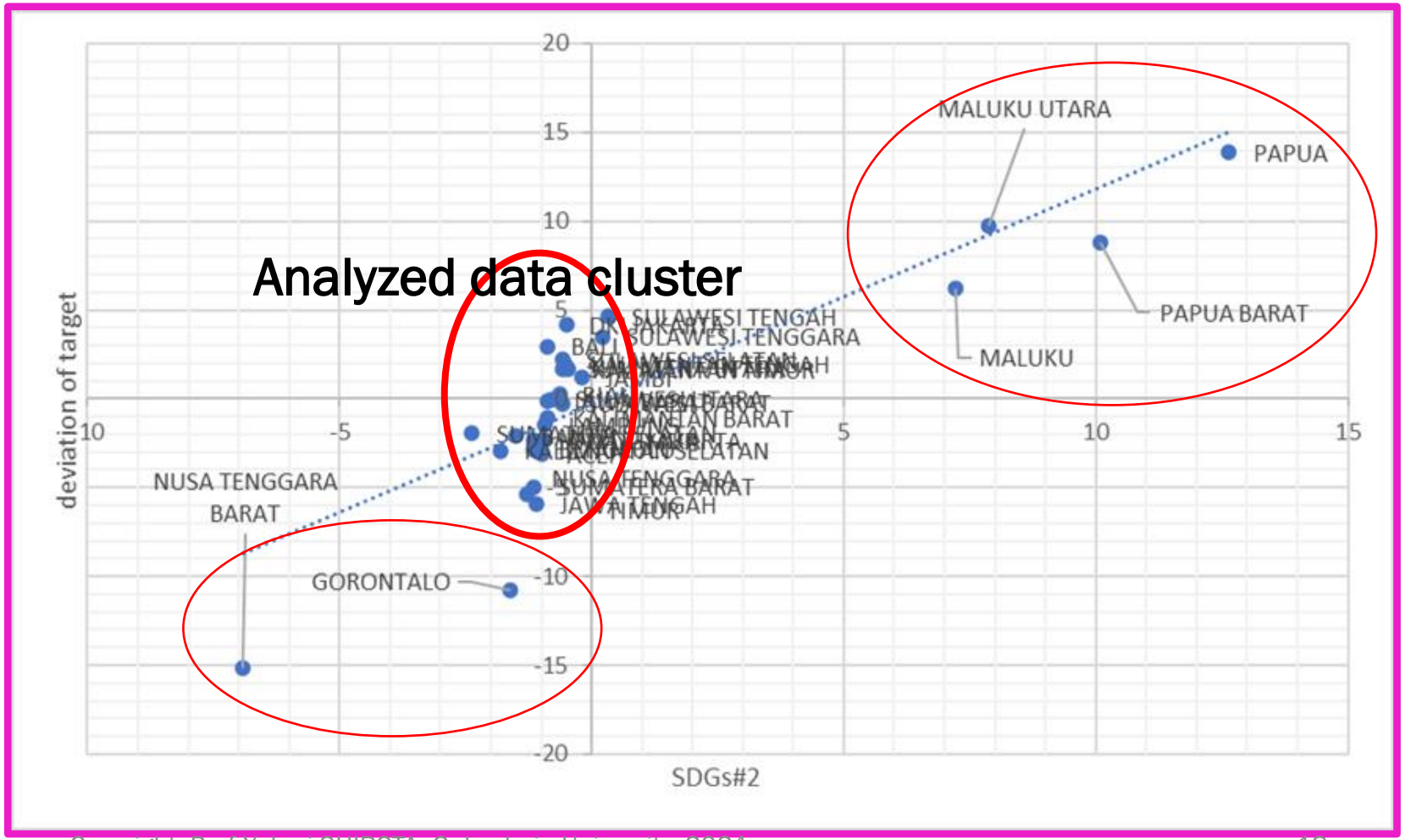
- Comparison between complaints by male/female
- In every province higher % of complaints by female than by male
- Perhaps females' positions are more vulnerable
- Target variable:

[Well-being%]=
"100 - this female %"



SDGs#2[NotHungry%] vs [Well-being%] in 2023

- Papua has the highest well-being%
- Focused on the center cluster 25 provinces



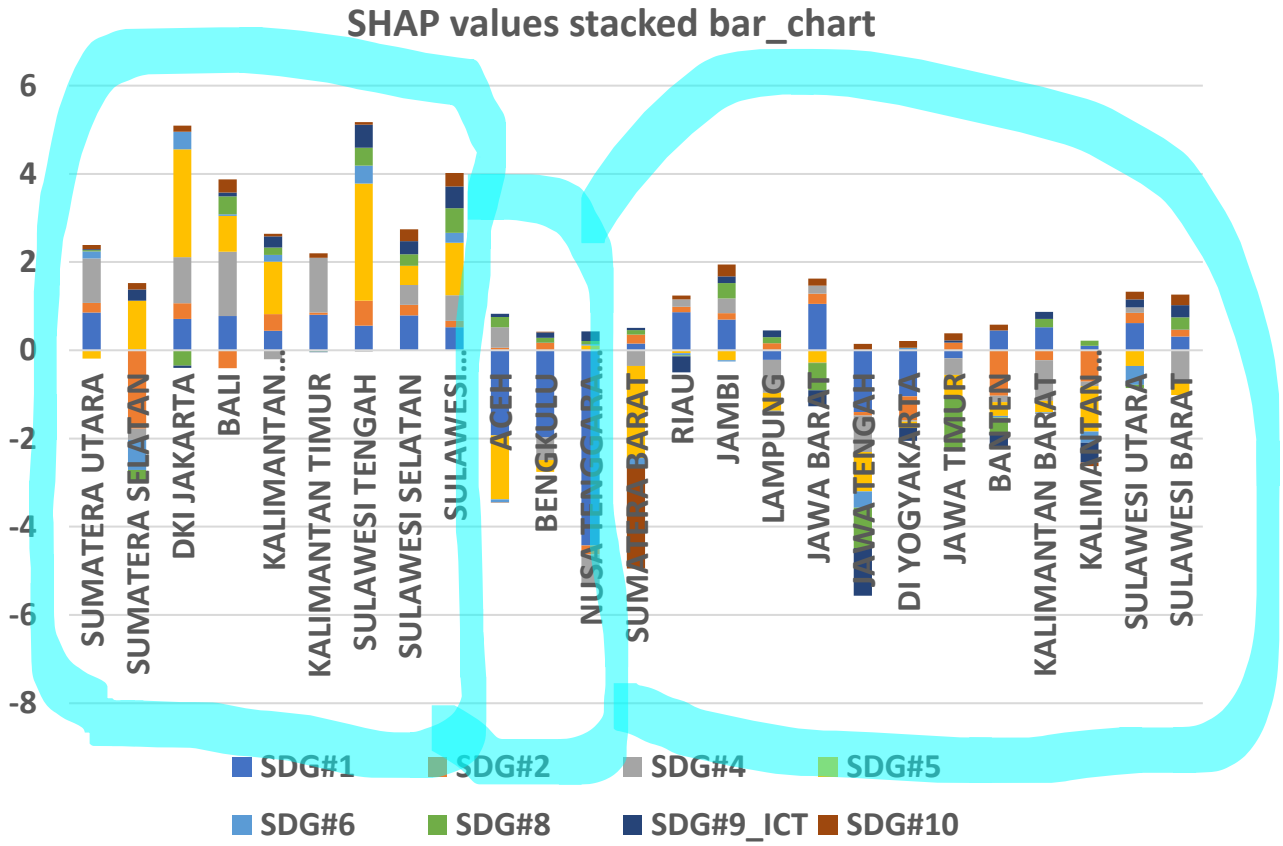
Correlation Coefficients among Variables before Regression

- 8 Explanatory Variables
- No explanatory var. highly correlated to TARGET

	SDG#1	SDG#2	TARGET :SDG#3	SDG#4	SDG#5	SDG#6	SDG#8	SDG#9_I CT	SDG#10
SDG#1	1.00								
SDG#2	0.49	1.00							
TARGET:SDG#3	0.39	0.24	1.00						
SDG#4	0.48	0.43	0.29	1.00					
SDG#5	0.10	0.13	0.23	0.03	1.00				
SDG#6	0.43	0.41	0.18	0.72	0.07	1.00			
SDG#8	0.24	0.14	0.11	0.31	0.11	0.47	1.00		
SDG#9_ICT	0.70	0.34	0.25	0.81	0.13	0.63	0.47	1.00	
SDG#10	-0.10	-0.20	-0.16	-0.46	-0.41	-0.41	-0.54	-0.51	1.00

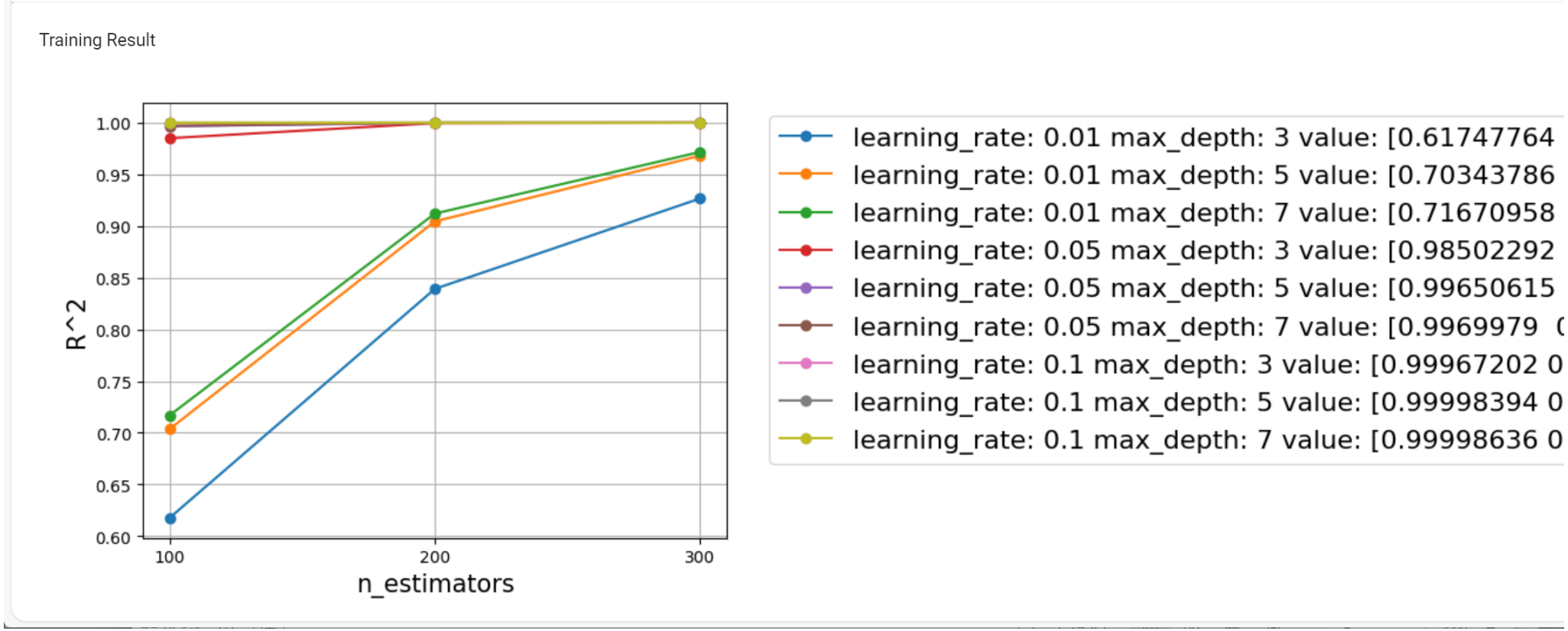
Contents

1. Research Objective
2. Data Selection
3. Regression: XGBOOST
4. SHAP
5. Clustering by SHAP Values
6. Conclusion



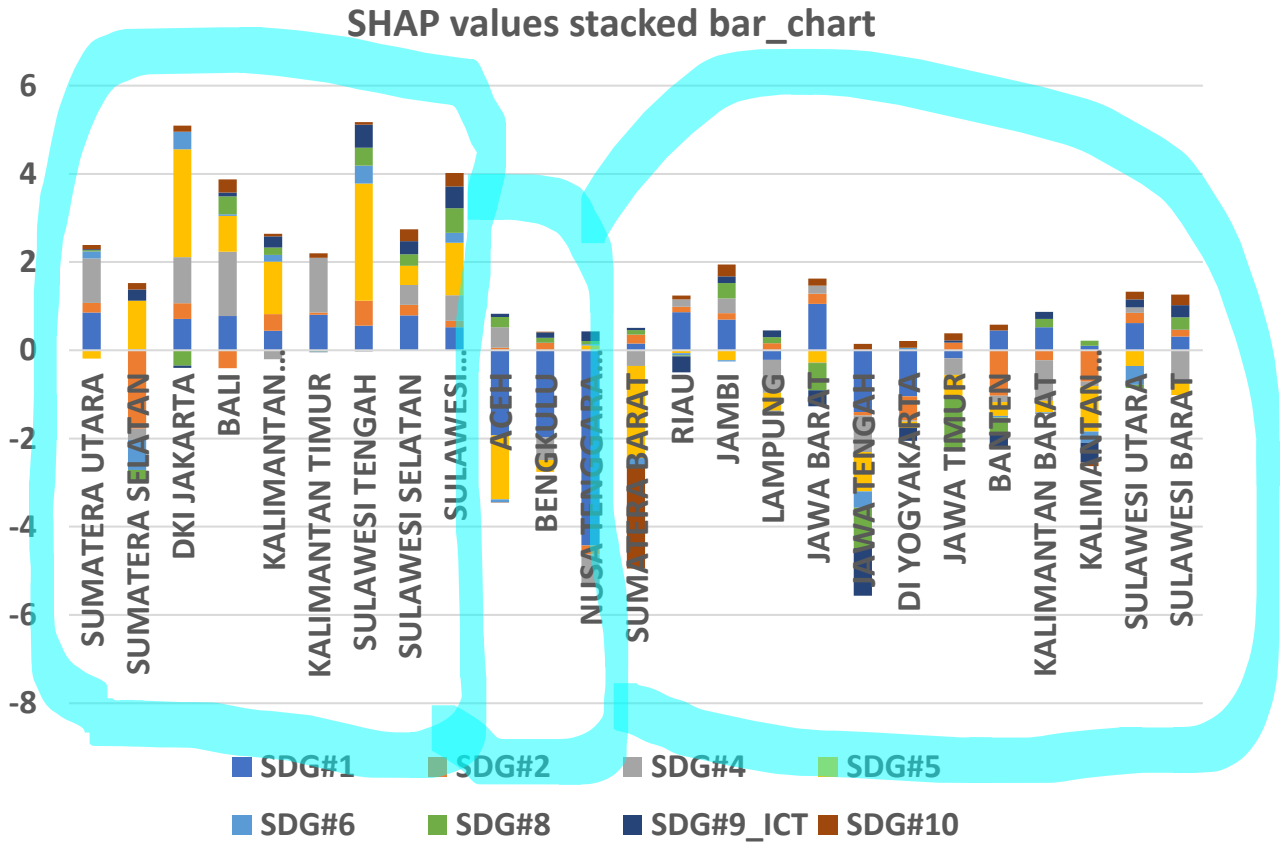
XGBoost Regression

- R^2: Almost 1.0 by the best parameter set



Contents

1. Research Objective
2. Data Selection
3. Regression: XGBOOST
4. SHAP
5. Clustering by SHAP Values
6. Conclusion



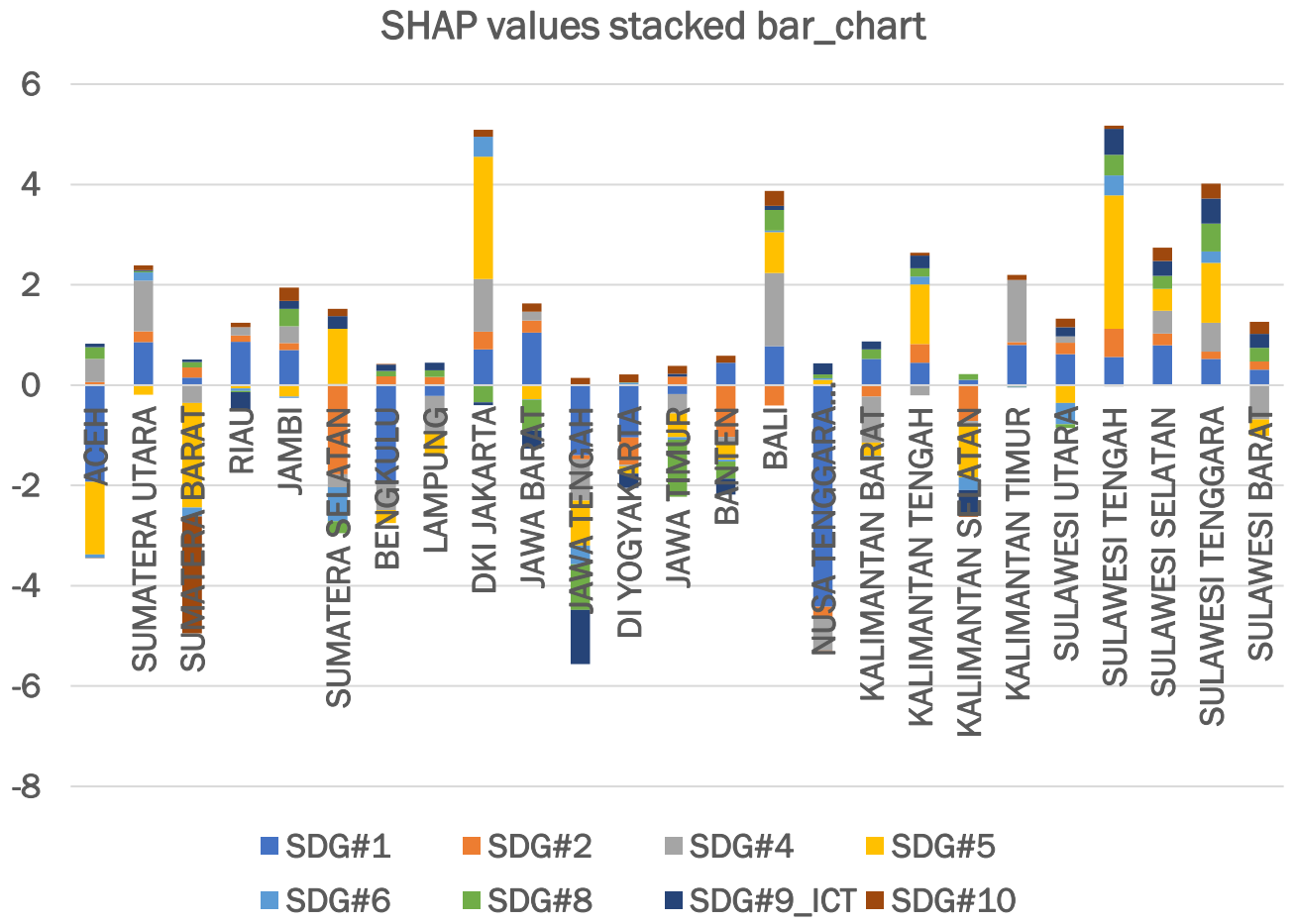
SHAP

- Using the resultant regression model f(X), SHAP values can be calculated.

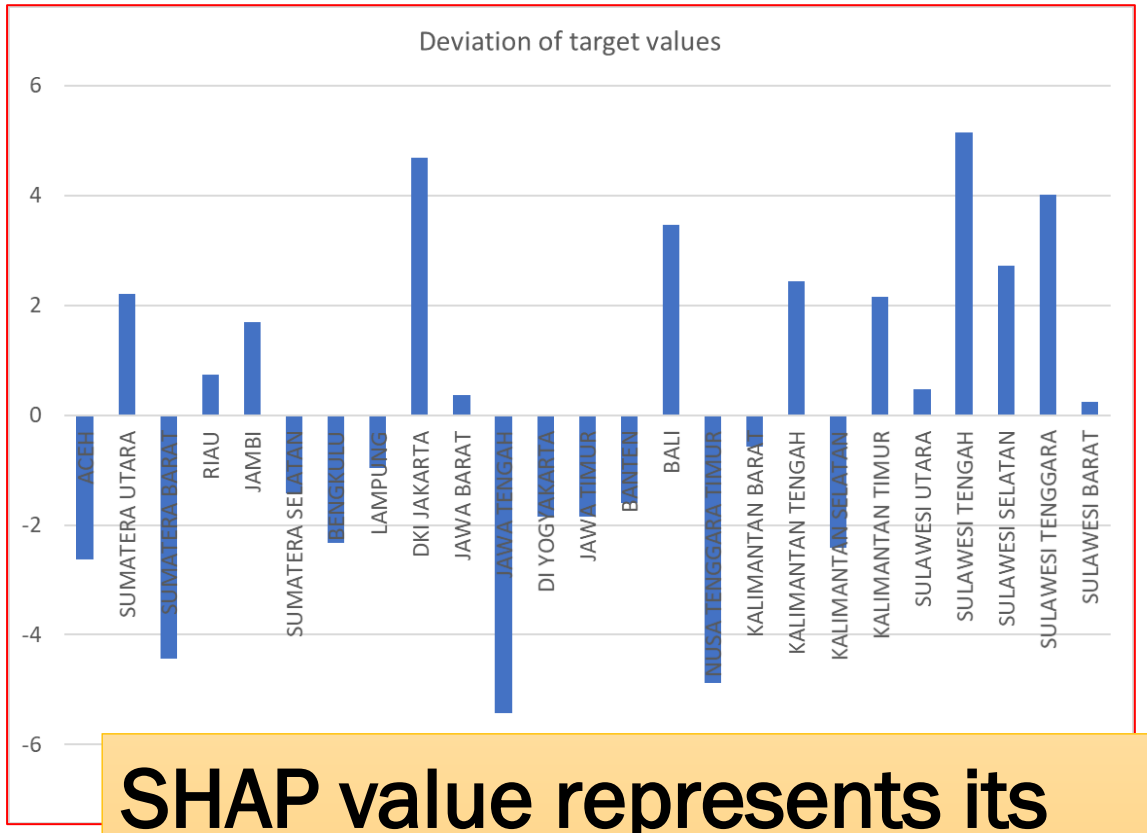
- DBKDA 2023 tutorial video by Shirota and Basabi **Theoretical Explanation and Case Studies of Shapley Values in Machine Learning Regression**
- https://www.youtube.com/watch?v=ml214YlY0oc
  https://www.iaria.org/conferences2023/TutorialsDBKDA23.html

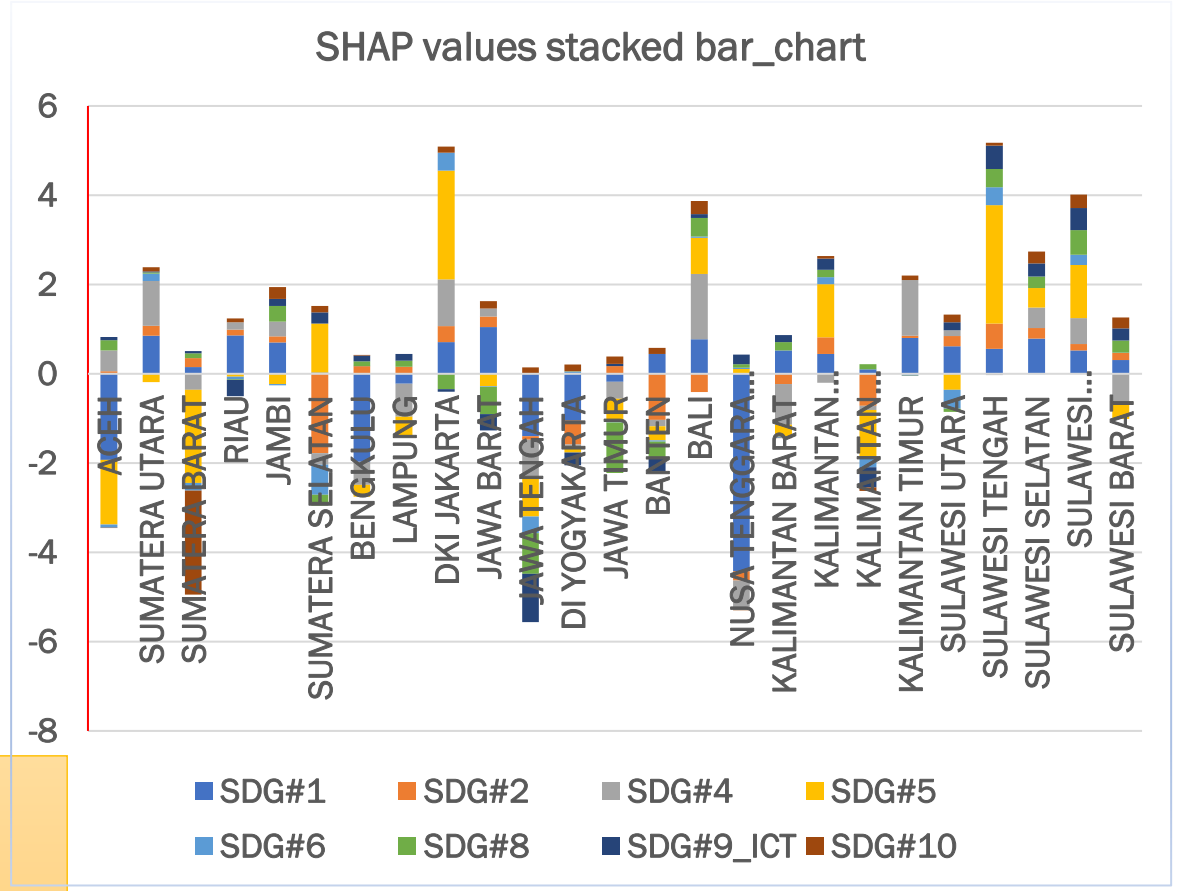
SHAP

- Each province data has 8 SHAP values.
- Y-axis: deviations of target values (target – average)
- SHAP values may be negative
- The sum of the 8 SHAP values becomes the deviation of the target value in each province.



Target Deviation and Stacked Bar_chart of SHAPs

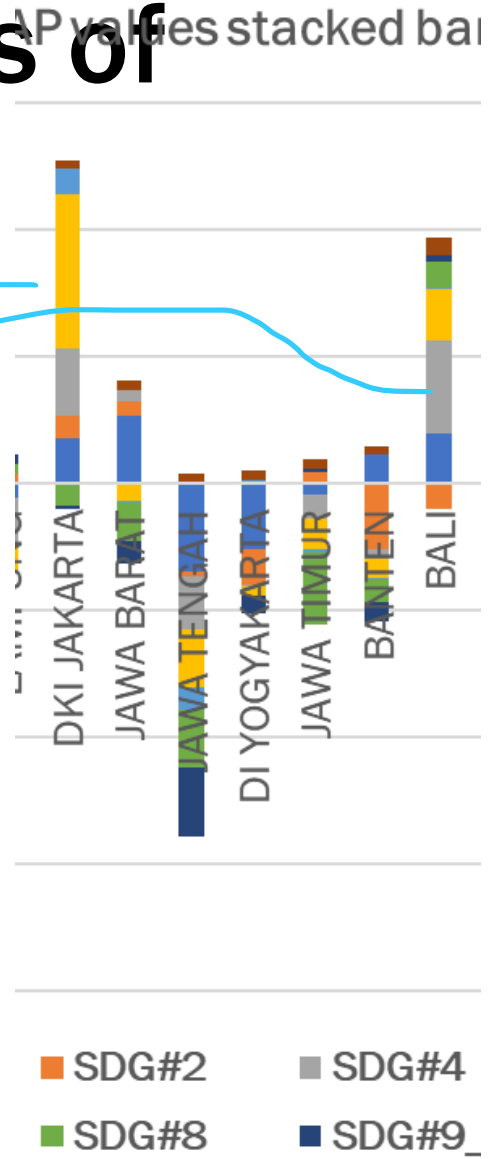




SHAP value represents its contribution to the target value

SHAP shows Characteristics of Each Province

- DKI Jakarta: dominant factor is SDG#5 (gender)
- Bali: dominant factor is SDG#4 (education)
- The 8 variables' contributions are up to the province.
- Using its characteristics, measure the effectiveness of each index variable



Which var is the dominant factor to well-being? Correlation coefficients among SHAP and target

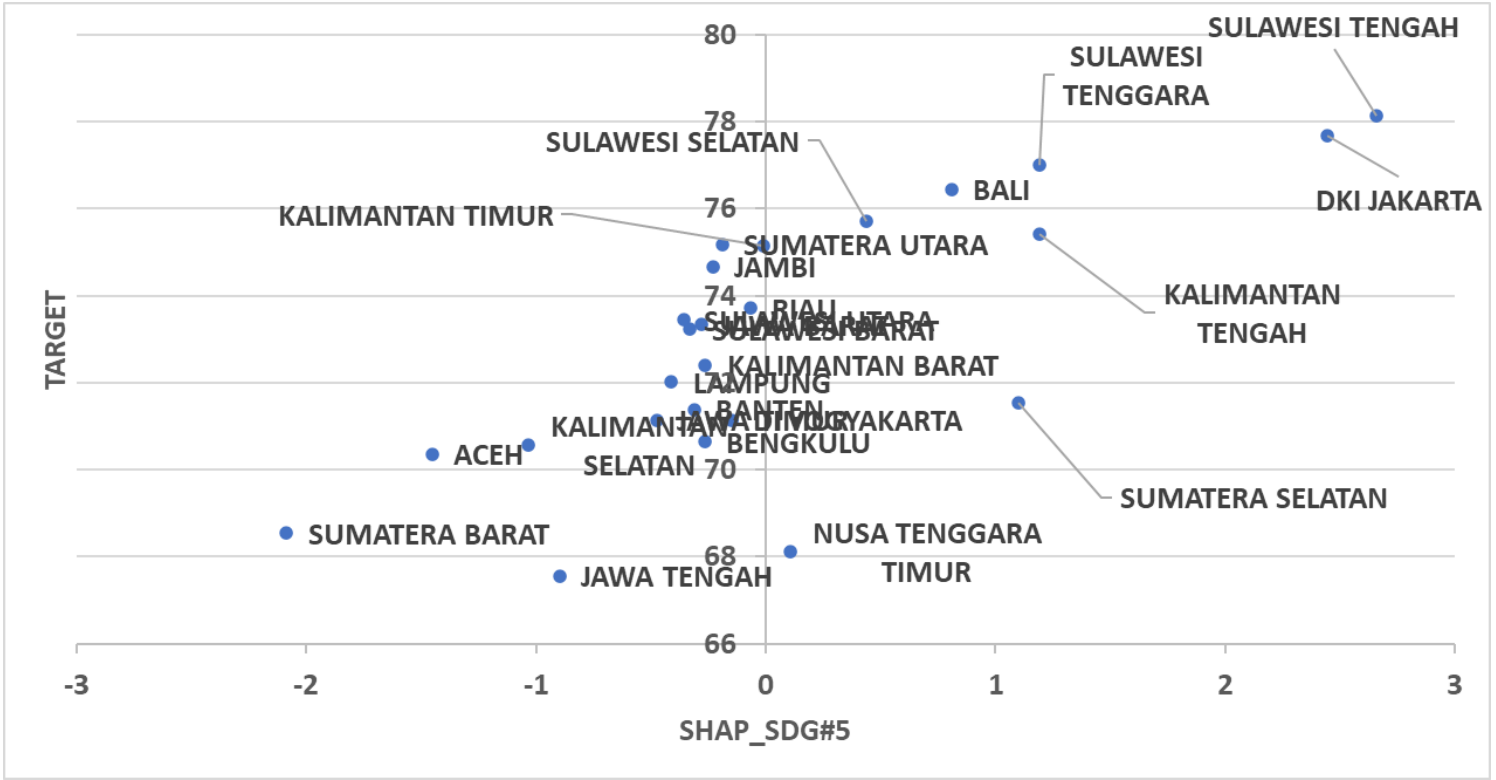
1. SDG#5 (gender) has the highest value 0.76
2. SDG#1 (NoPoor) 0.68
3. SDG#4 (Education) 0.66

SHAP/TARGET	SDG#1	SDG#2	SDG#4	SDG#5	SDG#6	SDG#8	SDG#9_ICT	SDG#10	TARGET:SDG#3
SDG#1	1.00								
SDG#2	0.13	1.00							
SDG#4	0.44	0.13	1.00						
SDG#5	0.25	0.09	0.33	1.00					
SDG#6	0.12	0.60	0.34	0.51	1.00				
SDG#8	0.08	0.13	0.23	0.19	0.32	1.00			
SDG#9_IC	0.09	0.22	0.12	0.43	0.33	0.62	1.00		
SDG#10	0.06	-0.09	0.21	0.45	0.14	-0.05	0.01	1.00	
TARGET:S	0.68	0.36	0.66	0.76	0.60	0.40	0.49	0.40	1.00

Scatter Plot between SDG#5_SHAP and target：0.76

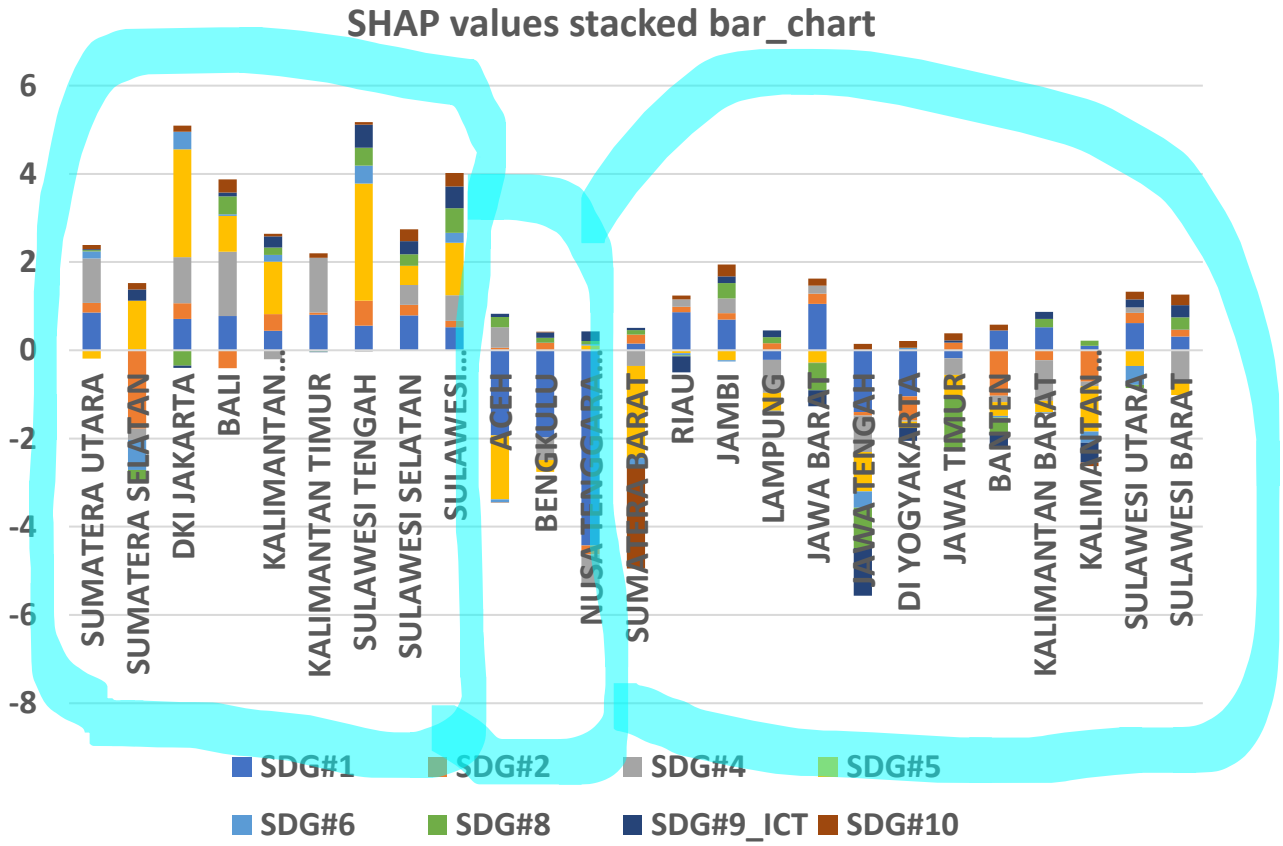
- Highest SDG#5_SHAP

1. SULAWESI TENGAH
2. DKI JAKARTA
3. SULAWESI TENGGARA
4. KALIMANTAN TENGA



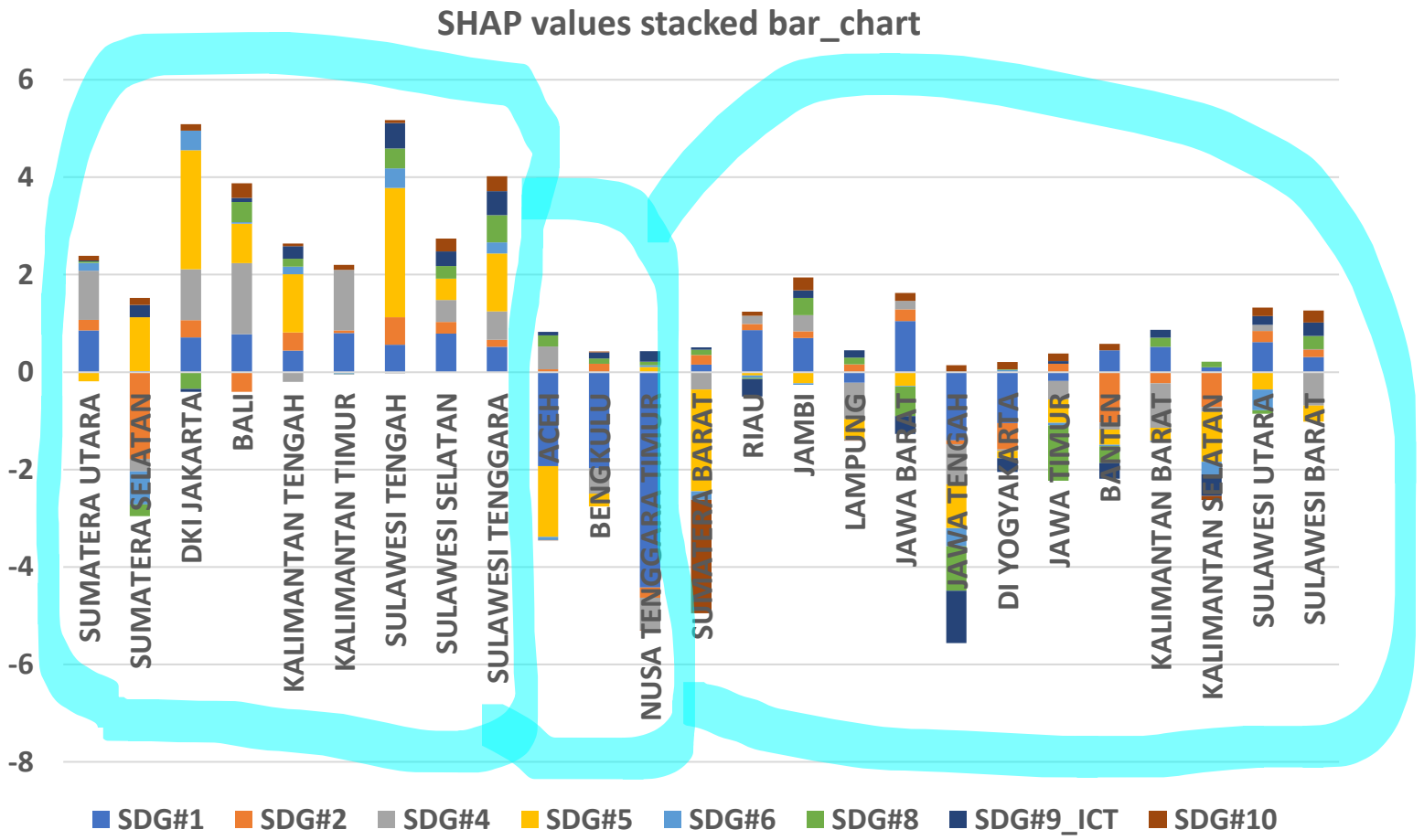
Contents

1. Research Objective
2. Data Selection
3. Regression: XGBOOST
4. SHAP
5. Clustering by SHAP Values
6. Conclusion



Clustering by SHAP distribution pattern

- 3 clusters



Clustering by 8 SHAP Values

- K-means
- The number of clusters k=3

Name	SDG#1_sh	SDG#2_sh	SDG#4_shapvalue	SDG#5_sh	SDG#6_sh	SDG#7_sh	SDG#8_sh	SDG#9 IC	SDG#10_s
ACEH	-1.32654	0.212461	0.5003764	-1.70384	0.008896	0.096279	-0.04466	-0.04872	-0.3157
SUMATER	0.323192	0.05647	1.4388171	-0.173	0.041875	0.303295	0.332676	-0.05031	-0.07491
SUMATER	0.0088	-0.00718	-0.69242096	-2.03179	-0.05423	0.001329	-0.10816	-0.06959	-1.48705
RIAU	0.097818	0.431496	-0.12935348	-0.19001	0.020559	0.410955	0.248114	-0.03034	-0.12018
JAMBI	0.228608	0.543912	-0.04496461	-0.07918	-0.02021	0.449312	0.210442	-0.02564	0.425004
SUMATER	0.064943	-0.82872	-0.20721847	0.22851	-0.63826	0.17615	0.037068	-0.05305	-0.21058
BENGKUL	-1.43052	-0.09703	-0.38009563	-0.09457	-0.15745	0.115303	-0.0582	-0.04536	-0.18407
LAMPUNG	0.200032	-0.05438	-0.4760555	-0.17452	-0.15737	0.158617	-0.04391	-0.07555	-0.33855
DKI JAKAR	0.499201	0.294244	1.0982153	1.402713	0.380063	0.298922	0.132814	0.239914	0.341256
JAWA BAR	0.225541	-0.0375	-0.10031061	-0.27548	0.002049	0.118123	-0.06808	-0.00415	0.49775
JAWA TEN	-0.53864	0.072779	-0.39650568	-1.18475	-0.00879	-3.34553	-0.00563	-0.0394	0.025324
DI YOGYA	0.508485	0.034989	0.340477	0.265698	0.020533	-3.0429	-0.17284	0.088556	0.115681
JAWA TIM	0.06426	-0.08567	-1.0101215	-0.94568	-0.0576	0.050973	-0.15073	-0.06531	0.359153
BANTEN	0.07224	-0.39842	-1.0393312	-0.1532	0.074785	0.09938	-0.21814	-0.03623	-0.00139
BALI	0.66559	-0.07583	1.5194385	0.158591	0.090098	0.347438	0.210376	0.258025	0.293991
NUSA TEN	-2.48124	-0.44546	-0.4487205	-1.06701	-0.12665	0.108841	-0.12152	0.06561	-0.35443
KALIMAN	0.265338	-0.06634	-0.30994222	-0.20193	-0.09781	0.320106	-0.04288	-0.04889	-0.3891
KALIMAN	0.226552	0.148986	-0.047996495	1.689522	0.187477	0.354901	0.071199	-0.04646	-0.14595

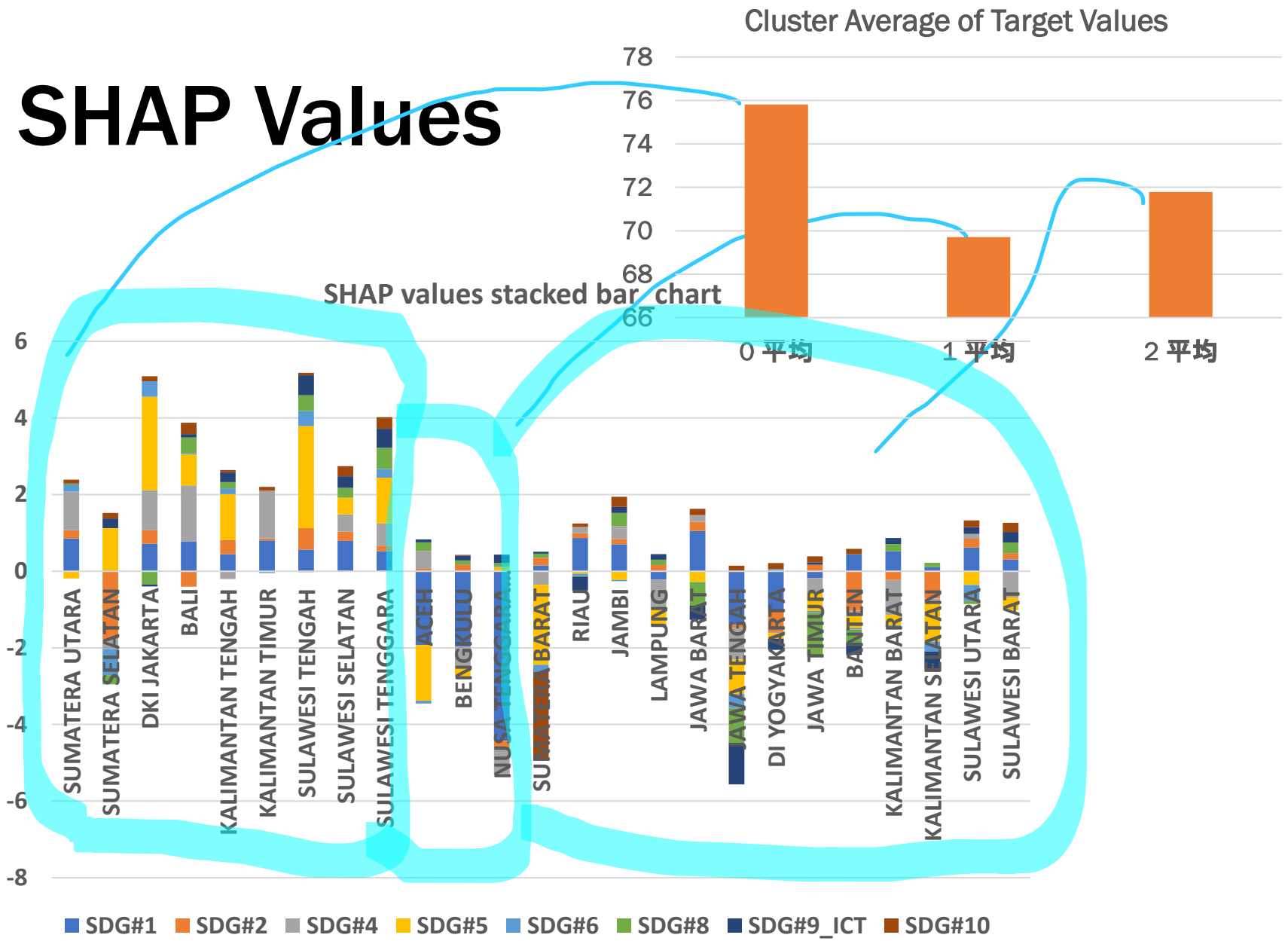
Find Average SHAP Values of Each Cluster

- Resultant LABEL ID

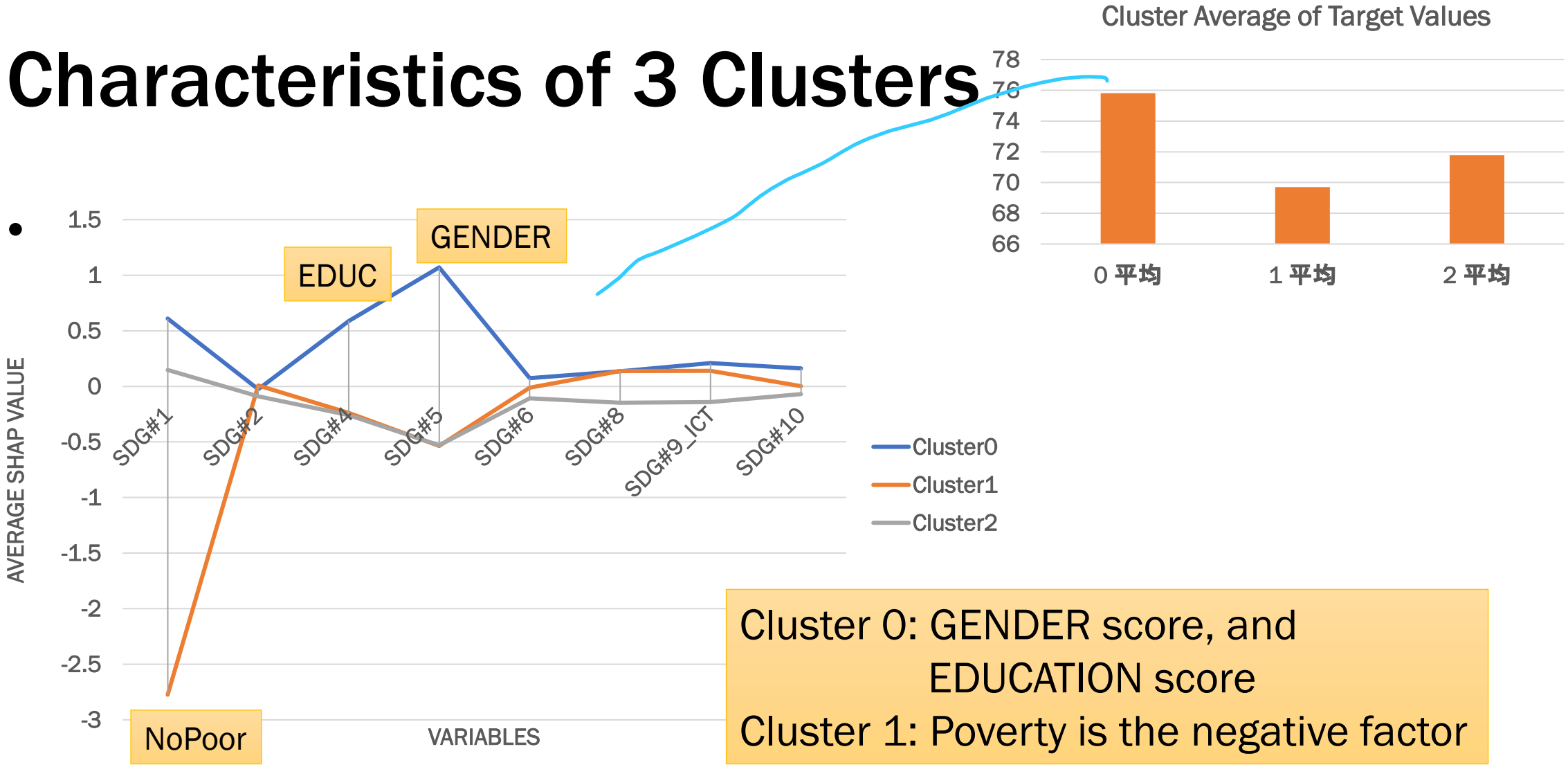
Name	LABEL_ID	SDG#1_sh	SDG#2_sh	SDG#4_shapvalue	SDG#5_sh	SDG#6_sh	SDG#7_sh	SDG#8_sh	SDG#9 IC	SDG#10_s	SDG#3
ACEH	2	-1.32654	0.212461	0.5003764	-1.70384	0.008896	0.096279	-0.04466	-0.04872	-0.3157	70.36
SUMATER	0	0.323192	0.05647	1.4388171	-0.173	0.041875	0.303295	0.332676	-0.05031	-0.07491	75.18
SUMATER	2	0.0088	-0.00718	-0.69242096	-2.03179	-0.05423	0.001329	-0.10816	-0.06959	-1.48705	68.54
RIAU	0	0.097818	0.431496	-0.12935348	-0.19001	0.020559	0.410955	0.248114	-0.03034	-0.12018	73.72
JAMBI	0	0.228608	0.543912	-0.04496461	-0.07918	-0.02021	0.449312	0.210442	-0.02564	0.425004	74.67
SUMATER	2	0.064943	-0.82872	-0.20721847	0.22851	-0.63826	0.17615	0.037068	-0.05305	-0.21058	71.55
BENGKUL	2	-1.43052	-0.09703	-0.38009563	-0.09457	-0.15745	0.115303	-0.0582	-0.04536	-0.18407	70.65
LAMPUN	2	0.200032	-0.05438	-0.4760555	-0.17452	-0.15737	0.158617	-0.04391	-0.07555	-0.33855	72.02
DKI JAKA	0	0.499201	0.294244	1.0982153	1.402713	0.380063	0.298922	0.132814	0.239914	0.341256	77.67
JAWA BA	0	0.225541	-0.0375	-0.10031061	-0.27548	0.002049	0.118123	-0.06808	-0.00415	0.49775	73.34
JAWA TEN	1	-0.53864	0.072779	-0.39650568	-1.18475	-0.00879	-3.34553	-0.00563	-0.0394	0.025324	67.56
DI YOGYA	1	0.508485	0.034989	0.340477	0.265698	0.020533	-3.0429	-0.17284	0.088556	0.115681	71.14
JAWA TIM	2	0.06426	-0.08567	-1.0101215	-0.94568	-0.0576	0.050973	-0.15073	-0.06531	0.359153	71.14
BANTEN	2	0.07224	-0.39842	-1.0393312	-0.1532	0.074785	0.09938	-0.21814	-0.03623	-0.00139	71.38
BALI	0	0.66559	-0.07583	1.5194385	0.158591	0.090098	0.347438	0.210376	0.258025	0.293991	76.45
NUSA TEN	2	-2.48124	-0.44546	-0.4487205	-1.06701	-0.12665	0.108841	-0.12152	0.06561	-0.35443	68.11
KALIMAN	2	0.265338	-0.06634	-0.30994222	-0.20193	-0.09781	0.320106	-0.04288	-0.04889	-0.3891	72.41

3 Clusters' SHAP Values

•



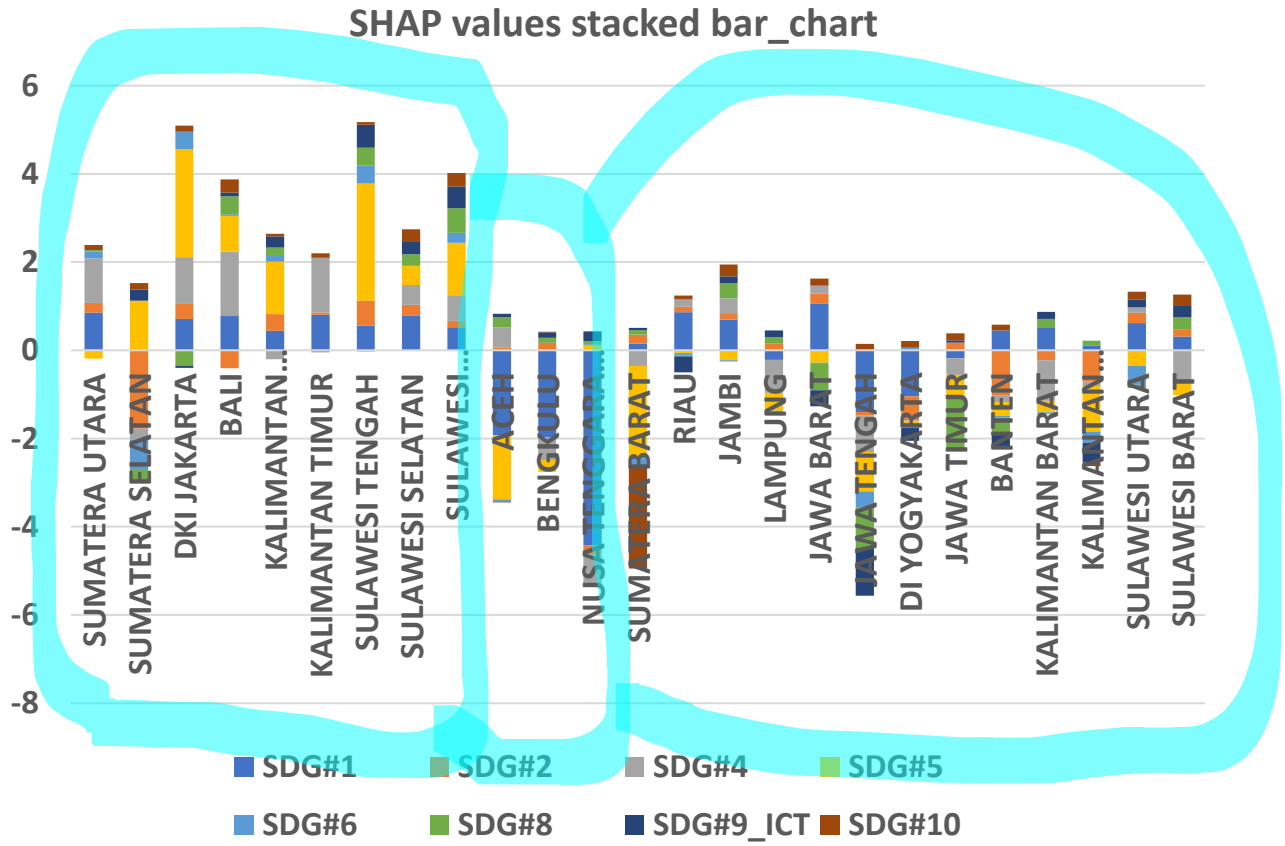
Characteristics of 3 Clusters



- Cluster 0: GENDER score, and EDUCATION score
- Cluster 1: Poverty is the negative factor

Contents

1. Research Objective
2. Data Selection
3. Regression: XGBOOST
4. SHAP
5. Clustering by SHAP Values
6. Conclusion



Conclusions

- SDG#3 well-being is target
- Dominant factors for well-being
- XGBoost and SHAP

1. SDG#5 (gender) 0.76
2. SDG#1 (NoPoor) 0.68
3. SDG#4 (Education) 0.66

- Clustering by SHAP
  - High well-being cluster: high gender_SHAP
  - Low well-being cluster's problem: poverty