**<World Indicators Project Overview>**

**201533661 이승수, 201533645 배성재**

**<Contents>**

1. Objective Setting: Country code’KOR’의 insight 도출
2. Data Curation: 주로 Country.csv, Indicators.csv, 추가로 (CountryNotes+FootNotes,Series+SeriesNotes),
3. Data Inspection
4. Data Rreprocessing
5. Decision Tree: determine split value for continuous variables, Pruning
6. Logistic Regressioin
7. SVM: feature reduction
8. Data Analysis
   1. Decision Tree, Logistic Regression, SVM
9. Evaluation

Evaluation methods: k-fold cross-validation로 feature engineering이전/이후 시간,정확도 비교.

1. Deployment
2. **Objective Setting**

Clustering nation which has CountryCode=’KOR’ by decades, and evaluate Korea’s development trends.

1. **Data Curation**

**Kaggle:**

**(1960~2010)** <https://www.kaggle.com/worldbank/world-development-indicators>

**(1960~2019)** <https://www.kaggle.com/laiyipeng/2019-world-development-indicator-wdi>

1. **Data Inspection**

<테이블 간 구조>

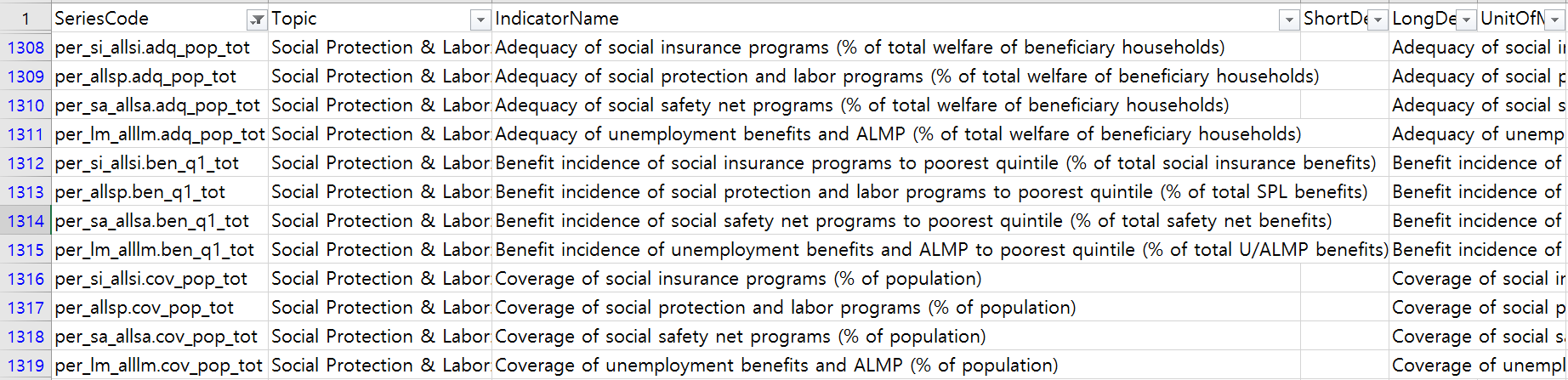
1. Data Preprocessing

<row/columns 삭제>

**[Series.csv]**

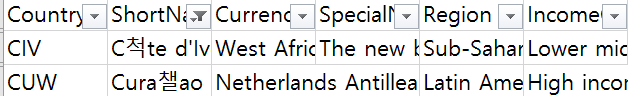
column: null인 것 삭제(OtherNotes, RelatedSourceLink, OtherWeblink, RelatedIndicators)

row: SeriesCode중 형식에 어긋나고 사용되지 않는 12개 삭제



한글/오타 등 삭제

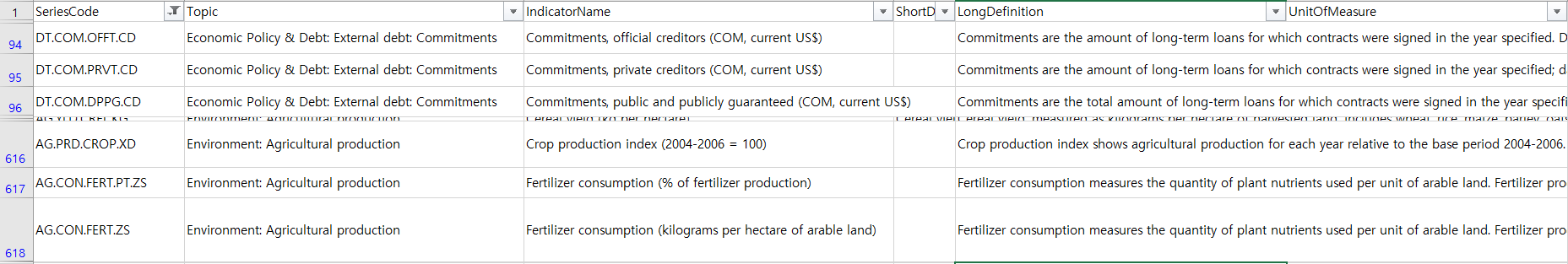




[Series.SeriesCode]

-대체적으로 Industry(맨앞)+code+measureUnit(맨뒤) 구조로 구성, 꼭 이 형식을 따르는 건 아님.

-NGram: code의 길이와 표현 방식은 조금씩 다 다르고 HIV/AIDS 등 같은 것도 다르게 표기하는 경우가 있어서 n-gram으로 code를 어떻게 해설할지 정해줘야함. 어려우면 industry+unit으로 code를 clustering하는 것도 가능할 듯?



위 그림에서 DT코드는 개념상 Economic-Debt 순으로 배열되는게 맞는데 여기선 그냥 DT.으로 사용.

마지막 unit은 대체적으로 각각 맞는 경향. 단위별 길이/넓이 등으로 단위만 통일시켜주면 될 것.

1.Industry: length=49

|  |
| --- |
| -Economy: 'DT'(debt),  임시Economy: 'BN', 'BM', 'BG', 'BX'(Economic Policy & Debt: Balance of payments: 공통), 'DC',  -Financial Sector: 'FP', 'FS', 'FM'(Monetary holdings (liabilities)), 'FR'(Interest rates), 'FI', 'FD', 'FB'(Bank,Assets),  -Agriculture: 'AG'(agriculture),  -Environment: 'EA'(Environment:Agriculture production), 'EG'(Environment: Energy production & use), 'EN'(Environment: Emissions), 'ER',  -: 'EP'(Infrastructure: Transportation),  -Social Protection: 'SE’(Education), 'SH'(Health), 'SN'(Health), ’SP'(Health:Population), 'SG'(Health:Risk factors), 'SI'(Poverty), 'SL'(Social Protection & Labor), 'SM'(Migration), 'ST'(Travel,Trade,Tourism),  -: 'NY', 'NE', 'NV', (Economic Policy & Debt: National accounts:까지 공통)  -: 'PX'(Financial Sector: Exchange rates & prices), 'PA'(Economic Policy & Debt: Purchasing power parity), 전체적으로 수 적음.  -나머지: 'DC', 'MS', 'VC', 'CM', 'LP', 'GB', 'GC', |

분류:

Agriculture:

Trade:

2.measure unit: length=138,

|  |
| --- |
| -무게:  -개수:  -넓이:  -길이:  -백분율(%):  -돈: |

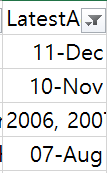
'CE', 'DOCS', 'DDAY', 'MMRT', 'METG', 'MO', 'KG', 'MT', 'PRDC', 'KH', 'ANEM', 'TOTL', 'GD', 'IMRT', 'GINI', 'LNDP', 'DPST', 'PPP', 'XQ', 'PP', 'CD', 'RISK', 'NAHC', 'EX', 'INCD', 'DPRT', 'DFCT', 'TENR', 'NRES', 'K6', 'DTHS', 'P5', 'CG', 'TM', 'K3', 'UR', 'SC', 'GAPS', 'NAGP', 'PSGR', 'DURS', 'P6', 'P1', 'FE', '10', 'UN', 'ARVL', 'TCHR', 'TIME', 'XD', 'WFRT', '0014', 'PROC', 'HA', 'IN', 'K2', 'REFG', 'PCAP', 'NMRT', 'RESD', 'TFRT', 'ZG', 'MM', '20', 'NREG', 'P3', 'ENRL', 'WT', 'PRVT', 'OR', 'MA', 'TRNR', 'GAP2', 'NO', 'LEND', 'P2', '2DAY', 'DPND', 'KM', 'AGES', 'PC', 'ORTH', 'MTHD', 'SECR', 'URHC', 'SETS', 'DT', 'IDPT', 'PAYM', 'DPPG', 'PC40', 'MAIN', 'MCTY', 'TU', 'RUHC', 'OFFT', 'FCRF', 'ATLS', 'RU', 'LE', 'ZS', 'HE', 'VO', 'NE', 'OUTG', 'UNER', 'GROW', 'GC', 'YG', 'RUGP', 'ACSN', 'SRCE', 'KN', 'RINR', 'RF', 'CN', 'DETH', 'OVRL', 'NENR', 'PUBL', '05', 'BBND', 'MD', 'ENRR', 'OE', 'K1', 'WD', 'M3', 'KD', 'MEAS', 'DNST', 'MORT', 'URGP', 'REER', 'LCTY', 'NETM', 'KT', 'OL'

**[Country.csv]** -> country\_cleansed.csv -> country\_cleansed\_ver2.csv(date타입 변경)

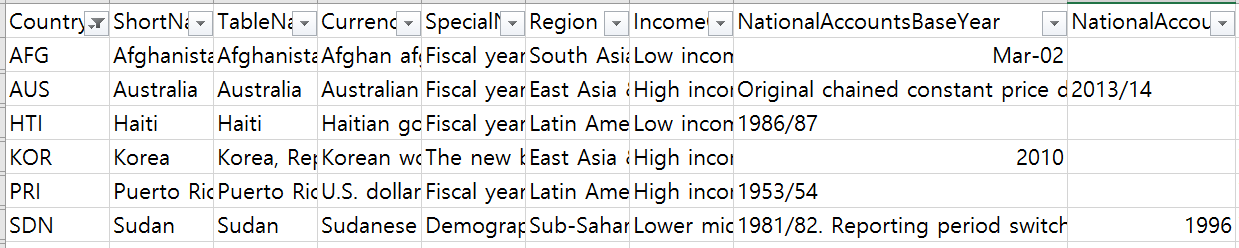
Feature 형식변환 ex)1986/87->1986, 2013/14->2013, Original…->삭제,

Incom…,National..Year 두개 하나로 함침(최신거로)

year자리에 ‘Rolling’ -> NaN

연도 형식은 파이썬에서 읽히지 않아 손으로 변경, 여기 날짜데이터 손상됐으니깐 수정하기\*\*

여러year은 2002, 2003 형식으로: 나중에 각각의 평균으로 넣어주고 date형식도 int로 바꿔줘야함



\*고려사항(SDN\_ NationalAccountsBaseYear)Reporting period switch from fiscal year to calendar year from 1996. Pre-1996 data converted to calendar year. ->

**[Indicators.csv]** ->Indicators\_bindWithYear.csv

**Series**.csv에서 분류한 분야별 코드별로

|  |
| --- |
| **[경제/금전]**  -Economy: 'DT'(debt),  임시Economy: 'BX'(Economic Policy & Debt: Balance of payments: 공통), 'DC',  -Financial Sector: 'FM'(Monetary holdings (liabilities)), 'FR'(Interest rates), 'FI', 'FD', 'FB'(Bank,Assets),  -: 'NV', (Economic Policy & Debt: National accounts:까지 공통)  (1960~1979)  (중복)  DC.DAC.TOTL.CD Net bilateral aid flows from DAC donors, Total (current US$)  FM.AST.DOMS.CN Net domestic credit (current LCU)  FM.LBL.BMNY.CN Broad money (current LCU) //해외자본  NV.IND.TOTL.ZS Industry (including construction), value added (% of GDP)  NV.IND.MANF.ZS Manufacturing, value added (% of GDP)  (1980~1999)  BX.KLT.DINV.CD.WD Foreign direct investment, net inflows (BoP, current US$)  BX.KLT.DINV.WD.GD.ZS Foreign direct investment, net inflows (% of GDP)  +중복  (2000~2009)  +중복  **[인프라/무역]**  - Infrastructure계통: 'IT'(Infrastructure: Communications), 'IS'(Infrastructure: Transportation), 'IP'(Infrastructure: Technology),  - Private Sector계통: 'IC'(Private Sector & Trade: Business environment), 'IE'(Private infrastructure investment),'IQ'(Public Sector: Policy & institutions),  -Trade: 'TM', 'TX', 'TT', 'TG',  (1960~1979)  없음  <중복>  TG.VAL.TOTL.GD.ZS Merchandise trade (% of GDP)  TM.VAL.FUEL.ZS.UN Fuel imports (% of merchandise imports)  TX.VAL.FUEL.ZS.UN Fuel exports (% of merchandise exports)  ->기름 수출/수입 지표  TM.VAL.MRCH.CD.WT Merchandise imports (current US$)  TX.VAL.MANF.ZS.UN Manufactures exports (% of merchandise exports)  ->생산품 수출/수입 지표  (1980~1999)  중복+  IP.TMK.TOTL Trademark applications, total  IS.AIR.PSGR Air transport, passengers carried  IT.CEL.SETS.P2 Mobile cellular subscriptions (per 100 people)  ->100명당 스마트폰 보급율(결측치 0으로)  (2000~2019)  중복+  IT.NET.USER.ZS Individuals using the Internet (% of population)  IS.SHP.GOOD.TU Container port traffic (TEU: 20 foot equivalent units)  TX.VAL.ICTG.ZS.UN ICT goods exports (% of total goods exports)  **[농업/환경/건강]**  -Agriculture: 'AG'(agriculture),  -Environment:'EG'(Environment: Energy production & use), 'EN'(Environment: Emissions), 'ER', 'EP'(Infrastructure: Transportation),  (1960~1979)  <중복>  AG.LND.AGRI.ZS Agricultural land (% of land area)  AG.PRD.LVSK.XD Livestock production index (2004-2006 = 100)  EN.ATM.CO2E.KT CO2 emissions (kt)  EN.POP.DNST Population density (people per sq. km of land area)  (1980~1999)  +중복  EN.ATM.CO2E.GF.KT CO2 emissions from gaseous fuel consumption (kt)  (2000~2019)  +중복  **[건강/복지/노동/여가]**  -Social Protection: 'SE’(Education), 'SH'(Health), 'SN'(Health), 'SI'(Poverty), 'SL'(Social Protection & Labor), 'ST'(Travel,Trade,Tourism),  (1960~1979)  SE.SEC.CUAT.LO.ZS Educational attainment, at least completed lower secondary, population 25+, total (%) (cumulative)  (중복)  SH.DTH.IMRT Number of infant deaths  SL.UEM.TOTL.NE.ZS Unemployment, total (% of total labor force) (national estimate)  SP.DYN.TFRT.IN Fertility rate, total (births per woman)  SP.DYN.TO65.MA.ZS Survival to age 65, male (% of cohort)  (1980~1999)  +중복  SH.IMM.IDPT Immunization, DPT (% of children ages 12-23 months)  SL.EMP.1524.SP.NE.ZS Employment to population ratio, ages 15-24, total (%) (national estimate)  SL.AGR.EMPL.ZS Employment in agriculture (% of total employment) (modeled ILO estimate)  SL.GDP.PCAP.EM.KD GDP per person employed (constant 2011 PPP $)  SL.TLF.TOTL.IN Labor force, total  (2000~2019)  SH.XPD.CHEX.PC.CD Current health expenditure per capita (current US$)  ST.INT.ARVL International tourism, number of arrivals  **[인구수]**  (1960~1979)  SP.POP.TOTL Population, total  SP.POP.TOTL.FE.IN Population, female  SP.POP.TOTL.MA.IN Population, male  ->여성/남성 비율  SP.POP.0014.TO Population ages 0-14, total  SP.POP.1564.TO Population ages 15-64, total  SP.POP.65UP.TO Population ages 65 and above, total  ->생산가능인구:1564/나머지(~14,65+)  SP.RUR.TOTL.ZS Rural population (% of total population)  SP.URB.TOTL.IN.ZS Urban population (% of total population)  ->Urban/Rural 비율  (1980~1999)  (2000~2019) |

1. Data Analysis
2. Evaluation
3. Deployment

[경제/금전]

-Economy: 'DT'(debt),

임시Economy: 'BN', 'BM', 'BG', 'BX'(Economic Policy & Debt: Balance of payments: 공통), 'DC',

-Financial Sector: 'FP', 'FS', 'FM'(Monetary holdings (liabilities)), 'FR'(Interest rates), 'FI', 'FD', 'FB'(Bank,Assets),

-: 'NY', 'NE', 'NV', (Economic Policy & Debt: National accounts:까지 공통)

-: 'PX'(Financial Sector: Exchange rates & prices), 'PA'(Economic Policy & Debt: Purchasing power parity), 전체적으로 수 적음.

BX.KLT.DINV.CD.WD Foreign direct investment, net inflows (BoP, current US$)

(1960~1979)

DC.DAC.TOTL.CD 중복:Net bilateral aid flows from DAC donors, Total (current US$)

FM.AST.DOMS.CN 중복:Net domestic credit (current LCU)

FM.LBL.BMNY.CN 중복:Broad money (current LCU) //해외자본

NV.IND.TOTL.ZS 중복:Industry (including construction), value added (% of GDP)

NV.IND.MANF.ZS 중복:Manufacturing, value added (% of GDP)

(1980~1999)

BX.KLT.DINV.WD.GD.ZS Foreign direct investment, net inflows (% of GDP)

+중복

(2000~2009)

BX.KLT.DINV.WD.GD.ZS Foreign direct investment, net inflows (% of GDP)

+중복

[인프라/무역]

- Infrastructure계통: 'IT'(Infrastructure: Communications), 'IS'(Infrastructure: Transportation), 'IP'(Infrastructure: Technology),

- Private Sector계통: 'IC'(Private Sector & Trade: Business environment), 'IE'(Private infrastructure investment),'IQ'(Public Sector: Policy & institutions),

-Trade: 'TM', 'TX', 'TT', 'TG',

(1960~1979)

없음

<중복>

TG.VAL.TOTL.GD.ZS Merchandise trade (% of GDP)

TM.VAL.FUEL.ZS.UN Fuel imports (% of merchandise imports)

TX.VAL.FUEL.ZS.UN Fuel exports (% of merchandise exports)

TM.VAL.MRCH.CD.WT Merchandise imports (current US$)

TX.VAL.MANF.ZS.UN Manufactures exports (% of merchandise exports)

(1980~1999)

중복+

IP.TMK.TOTL Trademark applications, total

IS.AIR.PSGR Air transport, passengers carried

IT.CEL.SETS.P2 Mobile cellular subscriptions (per 100 people)

->100명당 스마트폰 보급율(결측치 0으로)

(2000~2019)

중복+

IT.NET.USER.ZS Individuals using the Internet (% of population)

IS.SHP.GOOD.TU Container port traffic (TEU: 20 foot equivalent units)

TX.VAL.ICTG.ZS.UN ICT goods exports (% of total goods exports)

[농업/환경/건강]

-Agriculture: 'AG'(agriculture),

-Environment:'EG'(Environment: Energy production & use), 'EN'(Environment: Emissions), 'ER', 'EP'(Infrastructure: Transportation),

(1960~1979)

<중복>

AG.LND.AGRI.ZS Agricultural land (% of land area)

AG.PRD.LVSK.XD Livestock production index (2004-2006 = 100)

EN.ATM.CO2E.KT CO2 emissions (kt)

EN.POP.DNST Population density (people per sq. km of land area)

(1980~1999)

+중복

EN.ATM.CO2E.GF.KT CO2 emissions from gaseous fuel consumption (kt)

(2000~2019)

+중복

EN.ATM.CO2E.GF.KT CO2 emissions from gaseous fuel consumption (kt)

[건강/복지/노동/여가]

-Social Protection: 'SE’(Education), 'SH'(Health), 'SN'(Health), 'SI'(Poverty), 'SL'(Social Protection & Labor), 'ST'(Travel,Trade,Tourism),

(1960~1979)

SE.SEC.CUAT.LO.ZS Educational attainment, at least completed lower secondary, population 25+, total (%) (cumulative)

SE.SEC.CUAT.PO.ZS Educational attainment, at least completed post-secondary, population 25+, total (%) (cumulative)->이걸 축적하면 위에거(이건 5년단위 측정)

(중복)

SH.DTH.IMRT Number of infant deaths

SL.UEM.TOTL.NE.ZS Unemployment, total (% of total labor force) (national estimate)

SP.DYN.TFRT.IN Fertility rate, total (births per woman)

SP.DYN.TO65.MA.ZS Survival to age 65, male (% of cohort)

(1980~1999)

+중복

SH.IMM.IDPT Immunization, DPT (% of children ages 12-23 months)

SL.EMP.1524.SP.NE.ZS Employment to population ratio, ages 15-24, total (%) (national estimate)

SL.AGR.EMPL.ZS Employment in agriculture (% of total employment) (modeled ILO estimate)

SL.GDP.PCAP.EM.KD GDP per person employed (constant 2011 PPP $)

SL.TLF.TOTL.IN Labor force, total

(2000~2019)

SH.XPD.CHEX.PC.CD Current health expenditure per capita (current US$)

ST.INT.ARVL International tourism, number of arrivals

[인구수]

(1960~1979)

SP.POP.0014.TO Population ages 0-14, total

SP.POP.1564.TO Population ages 15-64, total

SP.POP.65UP.TO Population ages 65 and above, total

SP.POP.TOTL Population, total

SP.POP.TOTL.FE.IN Population, female

SP.POP.TOTL.MA.IN Population, male

SP.RUR.TOTL.ZS Rural population (% of total population)

SP.URB.TOTL.IN.ZS Urban population (% of total population)

(1980~1999)

(2000~2019)