

# 통계학세미나 1 final project

건강데이터를 이용한 비지도/지도 학습 시각화



222STG25 이다은

2023.12.14

- 목차 -

1. 데이터

1-1. 데이터 소개

1-2. 전처리

(1) 범주화

(2) 이상치

(3) 상관관계

(4) 최종 데이터

1-3. 데이터 특징

(1) 분포

(2) 통계량 : mean

2. 비지도 학습

2-1. PCA

2-2. k-means

3. 지도학습

3-1. LDA

3-2. Random forest

3-3. 성능비교

4. 결론 (28p)

## 1.data

### 1.1 데이터 소개

분석에 사용된 데이터 Life\_Expectancy\_Data.csv는 여러 국가를 기반으로 한 면역 요인, 사망률 요인, 경제적 요인, 사회적 요인 및 기타 건강 관련 요인이 포함되어 있다. 총 22개의 변수, 1649개의 데이터로 구성되어 있으며 결측치는 존재하지 않는다.

변수명	정의	설명	유형
Country	국가	다양한 국가의 이름	범주형
Year	연도	데이터를 관찰한 연도 2000 ~ 2015	연속형
Status	국가의 상태	Developed : 선진국 Developing : 개발도상국	범주형
Life Expectancy	기대수명	기대 수명(나이)	연속형
Adult Mortality	성인 사망수	15세~60세 사이의 성인 1000명당 사망자 수	연속형
Infant Deaths	영아 사망수	유아 1000명당 사망자 수	연속형
Alcohol	알코올 소비량	1인당 알코올 소비량(리터)	연속형
Percentage Expenditure	지출 비율	GDP대비 보건 예산 지출 비율(%)	연속형
Hepatitis B	B형 간염	1세 아동의 B형 간염 예방 접종률(%)	연속형
Measles	홍역	인구 1000명 당 홍역 예방 접종률(%)	연속형
BMI	BMI	인구 평균 체질량 지수	연속형
Under-Five Deaths	5세 미만 사망자 수	5세 이하 아동의 1000명당 사망자 수	연속형
Polio	소아마비	1세 아동의 소아마비 예방접종률(%)	연속형
Total Expenditure	총 지출	정부 총 예산 대비 보건 분야 예산(%)	연속형
Diphtheria	디프테리아	1세 아동의 디프테리아 예방접종률(%)	연속형
HIV/AIDS	HIV/AIDS	1000명당 HIV/AIDS으로 인한 사망률	연속형
GDP	국내총생산	1인당 GDP	연속형
Population	인구	국가 총 인구	연속형
Thinness 1-19 Years	저체중(1-19세)	10-19세 청소년의 저체중 비율	연속형
Thinness 5-9 Years	저체중(5-9세)	5-9세 어린이의 저체중 비율	연속형
Income Composition of Resources	ICOR	소득 분배 및 자원 접근성을 반영하는 종합 지수	연속형
Schooling	교육	평균 교육 기간	연속형

[table 1. 데이터 변수]

\*데이터 출처: Kaggle (<https://www.kaggle.com/datasets/uom190346a/health-and-demographics-dataset>)

## 1.2 데이터 전처리

정확한 분석을 진행하기 위해 재범주화, 이상치 및 상관관계 확인을 통해 불필요한 정보를 제거하는 전처리를 진행하였다.

### (1) 범주화

기존 데이터에 변수 중 건강, 면역, 사망, 경제 및 사회 요소에 차이를 줄 것이라고 예상되는 변수를 새롭게 범주화 하였다.

#### ① country\_g

기존 변수 country는 133 개의 국가 이름을 나타내는 범주형 변수이다. 나라마다 여러 요소에서 차이가 존재할 것이라고 예상하였으나, 나라의 이름이 매우 많아 분석을 위해 6 개의 대륙(Africa /Asia/Europe/ Middle\_East/Oceania/ South\_America )으로 재범주화 하였다. 이때, North America 에 해당하는 국가가 Canada 하나이므로, canada 는 South America 값을 부여하였다.

level	
Africa	Algeria, Angola, Botswana, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Comoros, Djibouti, Equatorial Guinea, Eritrea, Ethiopia, Gabon, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Nigeria, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, South Africa, Swaziland, Togo, Tunisia, Uganda, Zambia, Zimbabwe
Asia	Bangladesh, Cambodia, China, India, Indonesia, Kazakhstan, Malaysia, Maldives, Mongolia, Myanmar, Nepal, Pakistan, Philippines, Sri Lanka, Tajikistan, Thailand, Timor-Leste, Turkmenistan, Uzbekistan
Europe	Albania, Armenia, Austria, Azerbaijan, Belarus, Belgium, Bhutan, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Estonia, France, Georgia, Germany, Greece, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Montenegro, Netherlands, Poland, Portugal, Romania, Russian Federation, Serbia, Spain, Sweden, Ukraine, Cabo Verde
Middle_East	Afghanistan, Iraq, Israel, Jordan, Lebanon, Syrian Arab Republic, Turkey
Oceania	Australia, Fiji, Kiribati, Papua New Guinea, Samoa, Solomon Islands, Tonga, Vanuatu

South_America	Argentina, Belize, Benin, Brazil, Chile, Colombia , Costa Rica , Dominican Republic, Ecuador, El Salvador, Guatemala, Guyana, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama, Paraguay, Peru, Suriname, Trinidad and Tobago, Uruguay, Canada
---------------	---

[table 2. Country\_g]

### ② year\_g

기준 변수 year 은 데이터가 관찰된 연도를 나타내는 연속형 변수로, 2000 ~ 2015 의 값을 갖는다. 연속형 변수로 취급하기에는 값이 16 개밖에 되지 않으며, 2000년대 초반보다는 2015년에 가까울수록 건강, 경제 등의 요소에 차이가 존재할 것이라고 예상하였다. 따라서 연도의 기간을 묶어 총 3 개(y\_0004 / y\_0509 / y\_1015)로 그룹화하였다.

level	
y_0004	2000년 ~ 2004년에 해당하는 경우
y_0509	2005년 ~ 2009년에 해당하는 경우
y_1015	2010년 ~ 2015년에 해당하는 경우

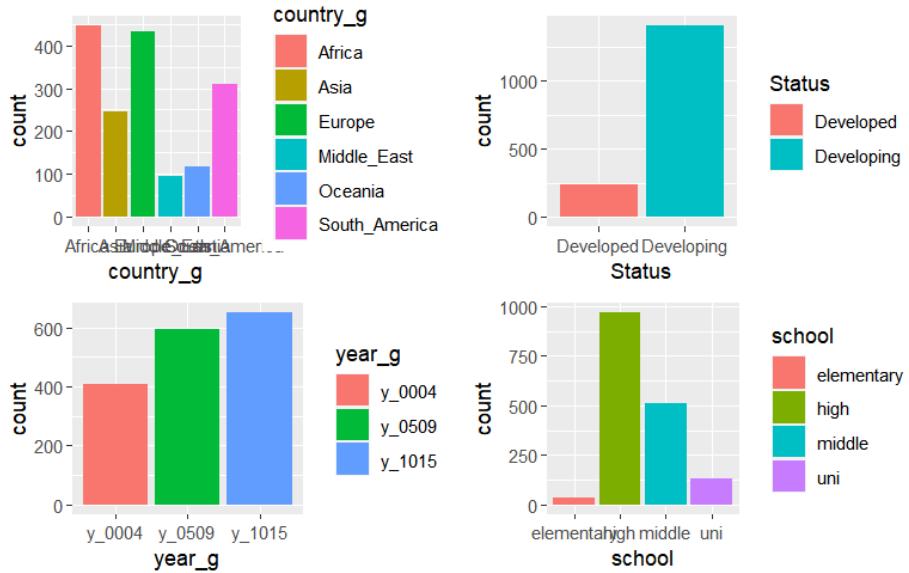
[table 3. year\_g]

### ③ school

기준 변수 schooling은 평균 교육기간을 나타내는 연속형 변수이다. 평균 교육 기간이 해당 국가의 경제 및 사회적 발전 수준을 나타낼 수 있는 변수 중 하나라고 생각하여 평균 교육 기간에 따라 건강/면역/경제 및 사회적 요인에 차이가 존재하는지 알아보기자 4개((Elementary,middle,high,uni)로 그룹화 하였다.

level	
Elementary	평균 교육기간이 6년 미만인 경우
High	평균 교육기간이 6년 이상 ~ 11년 미만인 경우
Middle	평균 교육기간이 11년 이상 ~ 16년 미만인 경우
Uni	평균 교육 기간이 16년 이상인 경우

[table 4. school]



[figure 1. 범주형 변수]

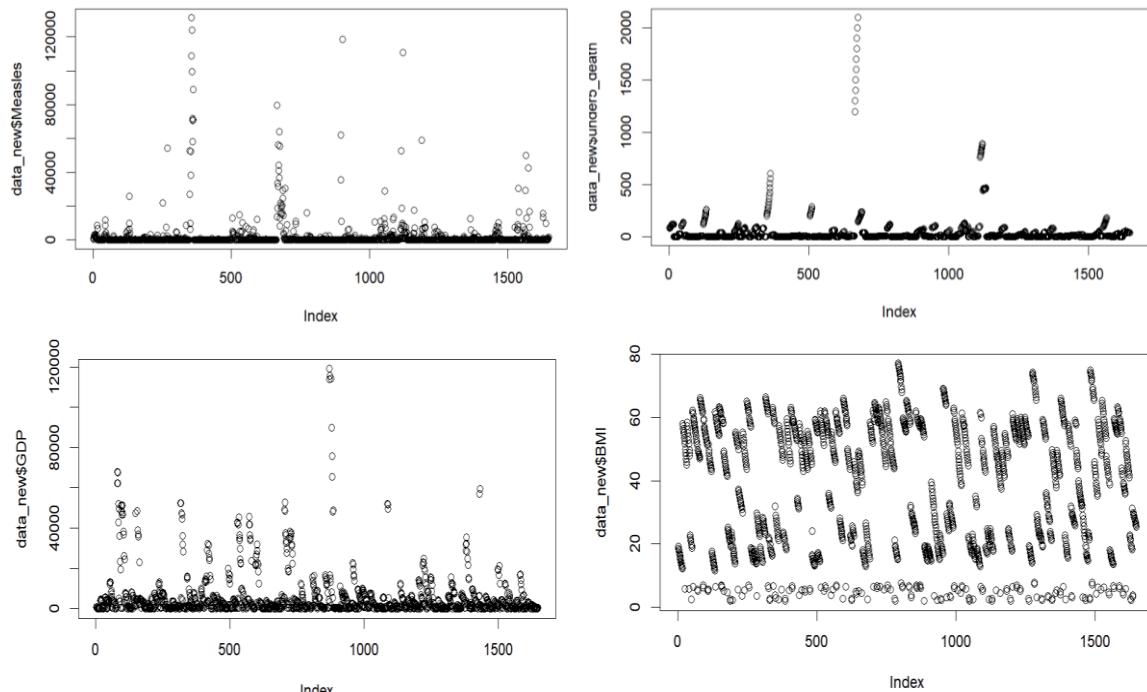
## (2) 이상치

Plot을 통해 데이터를 살펴보며 이상치를 제거하였다. 이때 사용되는 데이터가 변수 개수에 비해 전체 데이터양이 적은 편이라 Tukey Fences 방법을 적용하였을 때 전체적으로 삭제되는 데이터가 매우 많으므로, 변수 정의를 이용하여 이상치를 판단 및 제거를 진행하였다.

BMI의 경우 세계보건기구 비만의 정의 및 분류에 의하면 40 이상이 'Very severe'로 분류된다. 데이터에는 40 이상인 경우가 874개, 50 이상인 경우가 615개이며, 이러한 값은 이상치라고 판단하였다. 이상치가 전체 데이터의 높은 비율을 차지하므로 변수를 제거하였다.

변수명	변수 정의	이상치 정의	갯수	
Measles	1000 명 당 홍역 예방 접종자 수	1000 이상의 값	292	이상치 제거
under5_death	5 세 이하 아동의 1000 명당 사망자 수	1000 이상의 값	15	이상치 제거
GDP	1 인당 GDP	100 이하의 값	116	이상치 제거
BMI	인구 평균 체질량 지수	- 40 이상의 값 (874) - 50 이상의 값 (615) (세계보건기구 비만의 정의 및 분류)		변수 제거

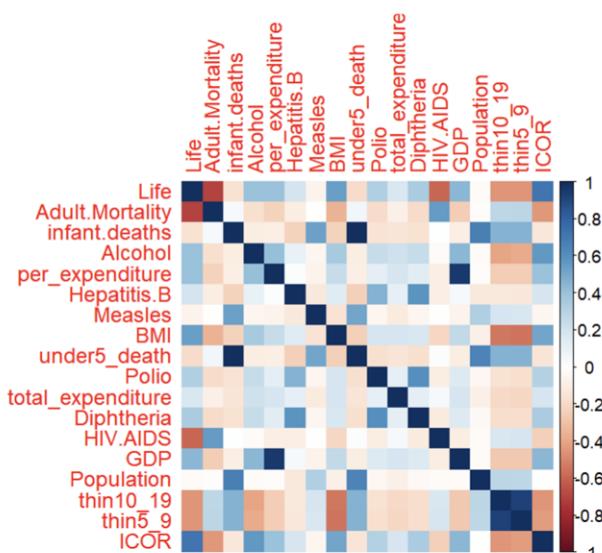
[table 5. 이상치]



[figure 2. 이상치 plot]

### (3) 상관관계

변수가 여러 개인 경우, 변수 간의 상관관계가 높으면 다중공산성의 문제가 발생할 수 있으므로 상관관계를 확인하였다. 그 중 0.95 이상의 높은 관계성을 가지는 변수쌍 2 개를 발견하였고, infant.deaths 와 per\_expenditure 를 제거하였다.



[figure 3. corrplot]

corr	var
positive	infant.deaths ↔ under5_death GDP ↔ per_expenditure thin10_19 ↔ thin5_9
negative	Life ↔ Adult.Mortality Life ↔ HIV.AIDS
0.95>	infant.deaths ↔ under5_death GDP ↔ per_expenditure

[table 6. 상관관계]

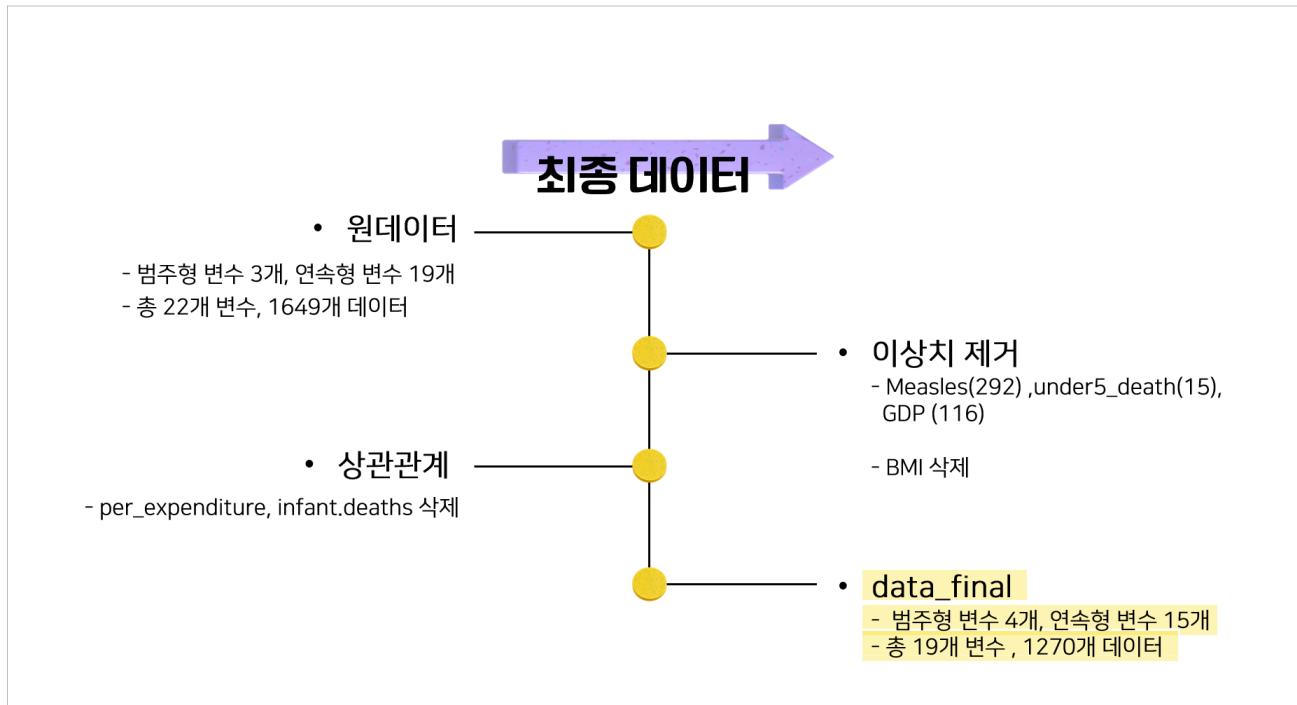
#### (4) 최종 데이터

데이터 전처리 과정을 통해 최종적으로 사용할 데이터를 구성하였다. 이상치 제거로 329 개의 데이터, 변수 3 개가 제거되어 총 19 개의 변수, 1270 개의 데이터로 구성되어 있으며, 결측치는 존재하지 않는다. 이 중 범주형 변수는 country\_g , year\_g , Status , School 으로 4 개이다. 또한 분석의 편의를 위해 변수 몇 개의 변수명을 변경하였다.

변수마다 갖는 값의 범주가 극단적으로 다른 경우, 스케일이 큰 변수의 영향력이 상대적으로 크게 작용하는 문제가 발생한다. 따라서 이후의 비지도 및 지도 학습에는 19 개의 변수에 대해 스케일링을 한 데이터를 이용하였다.

변수명	정의	설명	유형
country_g	대륙	Africa, Asia, Europe, Middle_East, Oceania, South_America	범주형
year_g	연도	데이터를 관찰한 연도 y_0004, y_0509, y_1015	범주형
Status	국가의 상태	Developed : 선진국 Developing : 개발도상국	범주형
Life Expectancy	기대수명	기대 수명(나이)	연속형
Adult Mortality	성인 사망수	15세~60세 사이의 성인 1000명당 사망자 수	연속형
Alcohol	알코올 소비량	1인당 알코올 소비량(리터)	연속형
Hepatitis B	B형 간염	1세 아동의 B형 간염 예방 접종률(%)	연속형
Measles	홍역	인구 1000명 당 홍역 예방 접종률(%)	연속형
Under-Five Deaths	5세 미만 사망자 수	5세 이하 아동의 1000명당 사망자 수	연속형
Polio	소아마비	1세 아동의 소아마비 예방접종률(%)	연속형
total_expenditure	총 지출	정부 총 예산 대비 보건 분야 예산(%)	연속형
Diphtheria	디프테리아	1세 아동의 디프테리아 예방접종률(%)	연속형
HIV/AIDS	HIV/AIDS	1000명당 HIV/AIDS으로 인한 사망률	연속형
GDP	국내총생산	1인당 GDP	연속형
Population	인구	국가 총 인구	연속형
thin10_19	저체중(1-19세)	10-19세 청소년의 저체중 비율	연속형
thin5_9	저체중(5-9세)	5-9세 어린이의 저체중 비율	연속형
ICOR	ICOR	소득 분배 및 자원 접근성을 반영하는 종합 지수	연속형
School	교육	평균 교육 기간 Elementary, high, middle, uni	범주형

[table 7. 최종 데이터]

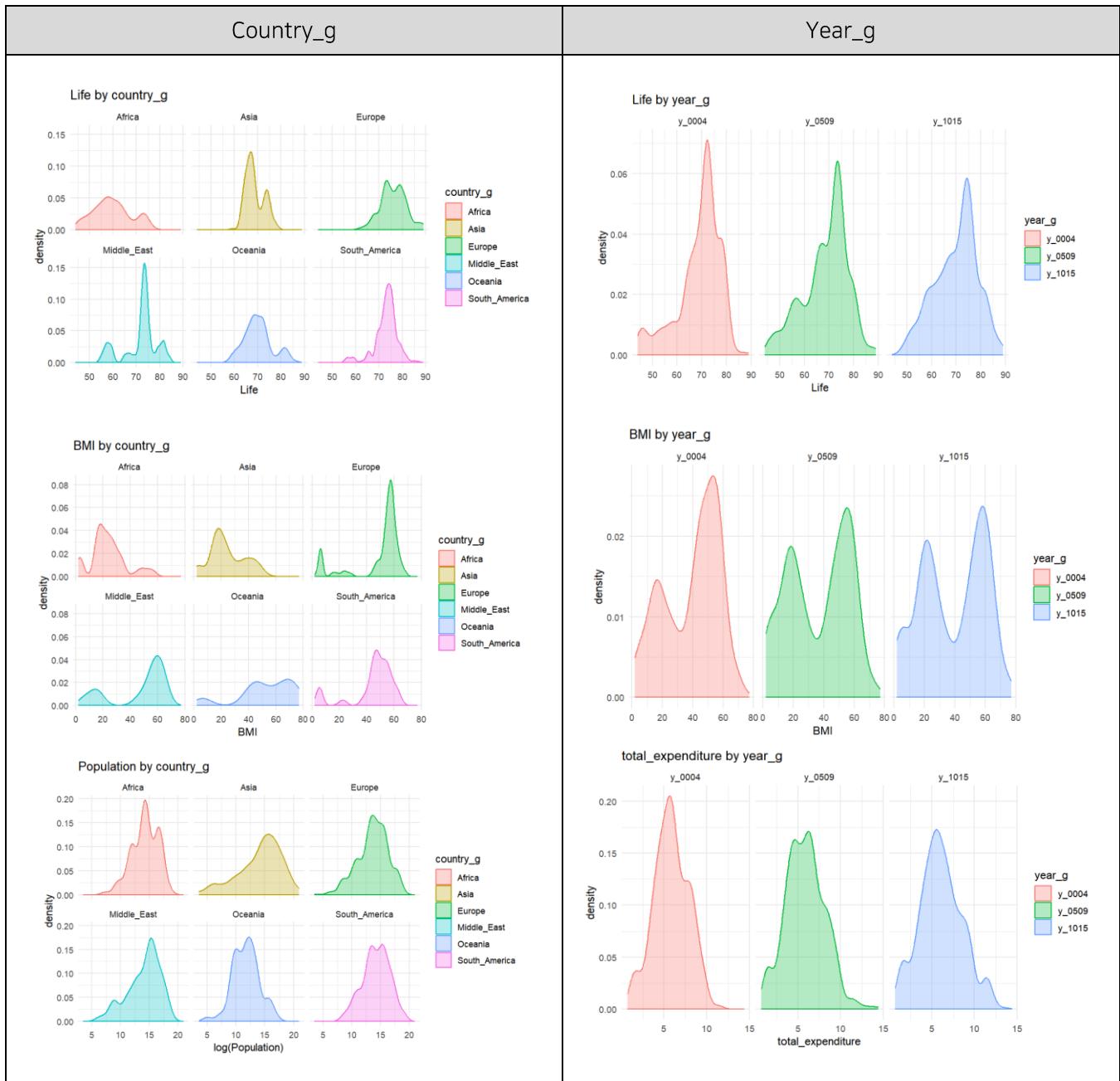


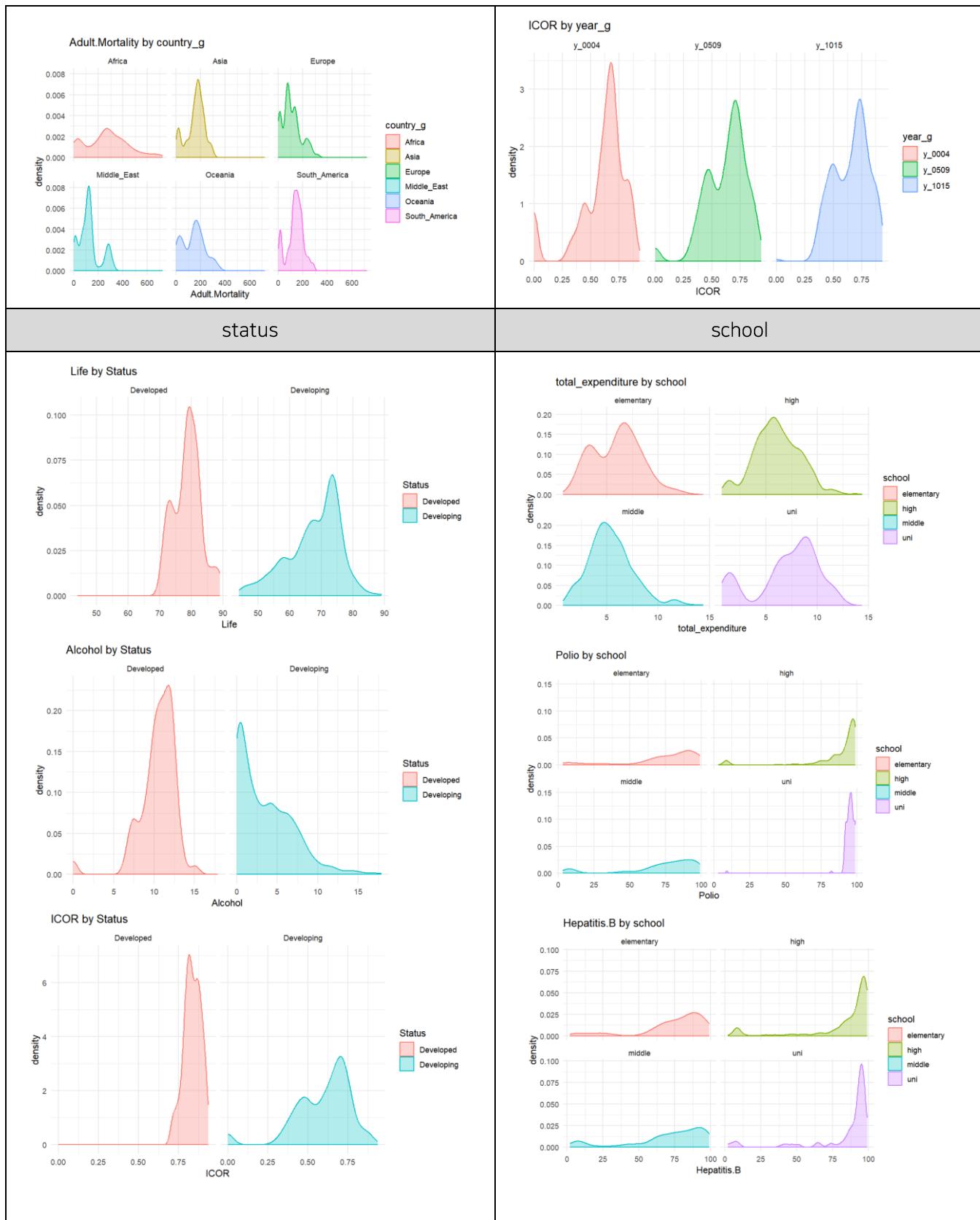
[figure 4. 데이터 전처리 과정]

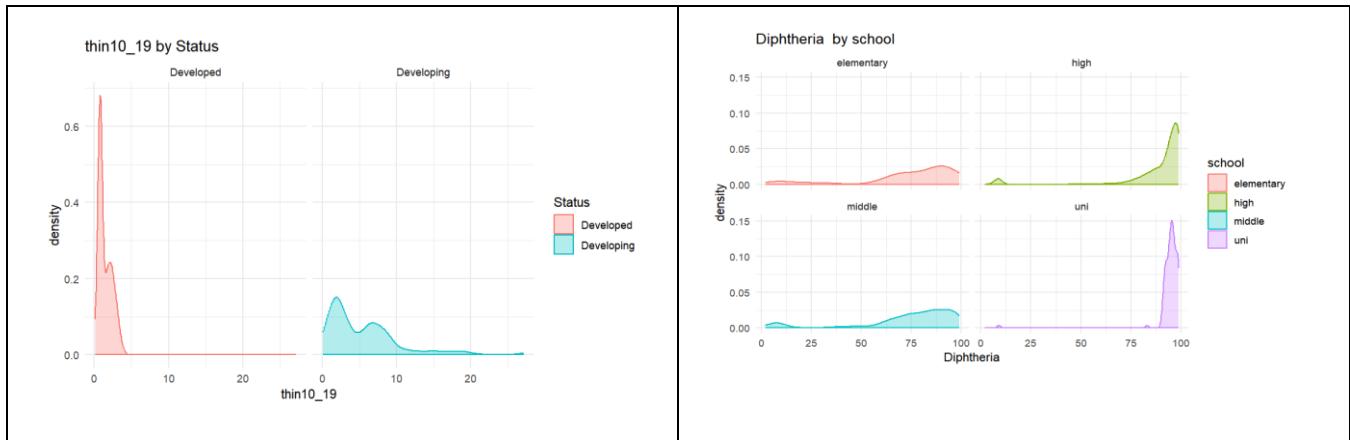
### 1.3 데이터 특징

#### (1) 분포

country\_g, year\_g, status, school의 각 레벨에 의해 다른 건강, 면역, 경제 및 사회적 의미를 가진 변수들이 차이를 보일 것이라고 예상하였고, 실제로 차이가 존재하는지 살펴보기 위해 원데이터를 이용하여 분포를 살펴보았다. 4개의 범주형 변수 중 country\_g, status는 레벨 간 뚜렷한 분포 차이가 존재했으며, year\_g, school의 레벨 간에는 유의미한 차이가 없는 것으로 판단된다. 따라서 후의 분석 진행은 country\_g 와 status 를 이용하여 건강, 면역, 경제 및 사회적 변수에 그룹 간 차이가 존재하는지 살펴보았다.







[table 8. 범주형 변수의 분포]

## (2) 통계량 means

### ① country\_g

contry	Life	Adult. Mortality	Alcohol	Hepatitis.B	Measles	under5_ death	Polio	total_ expenditure
Africa	60.56	263.38	2.64	76.68	142.48	33.96	78.70	5.74
Asia	68.69	168.87	2.05	90.41	150.54	19.68	91.03	4.43
Europe	76.34	105.69	8.77	83.61	69.61	1.46	91.91	6.69
Middle_East	74.02	104.87	1.28	82.44	211.27	15.25	80.11	6.80
Oceania	70.82	132.67	2.34	74.50	23.87	1.17	79.33	5.49
South_America	73.46	133.89	5.13	81.15	24.23	11.7	85.19	6.54
	Diphther ia	HIV.AIDS	GDP	Population	thin10_19	thin5_9	ICOR	
Africa	79.46	6.63	1997.80	6213260	7.12	7.17	0.50	
Asia	91.80	0.16	2305.88	8640778	7.57	8.12	0.60	
Europe	92.76	0.13	13046.96	7198992	2.49	2.55	0.76	
Middle_East	77.60	0.10	6212.73	9235640	4.87	4.788	0.65	
Oceania	73.98	0.17	6098.87	966653	1.17	1.11	0.58	
South_America	87.08	0.41	4996.55	11999832	2.51	2.44	0.69	

[table 9. 대륙 별 평균값]

대륙별로 각 변수의 평균 값을 구하여 특성을 살펴본 결과, Africa 는 Adult.Mortality , under5\_death , HIV.AIDS 와 같은 사망 요인과 부정적인 건강 지표가 높게 나타나며, GDP 와 기대 수명이 가장 낮았습니다. Asia 의 경우, 아프리카에 이어 비슷한 경향을 보이며, 두 번째로 부정적인 건강 지표가 높다. Middle\_East 는 Adult.Mortality 과 HIV.AIDS 은 낮지만, under5\_death 이 높은 편이다. Oceania 는 다른 지역에 비해 under5\_death 과 thin10\_19 , thin5\_9 이 다른 대륙에 비해 매우 낮게 나타난다. Europe 은 Life 가 가장 길고, 알코올 소비량이 높은 편이며, 경제적으로 안정되어 있다. 이러한 특성들을 종합해보면, 아프리카와 아시아에서는 건강과 경제적인 측면에서 부정적인 면이 보이며, 유럽은 상대적으로 건강하고 경제적으로 안정된 상태라고 판단할 수 있다.

## ② status

Status	Life	Adult. Mortality	Alcohol	Hepatitis.B	Measles	under5_ death	Polio	total_ expenditure
Developed	78.73	82.90	10.54	87.43	76.85	0.93	94.42	7.06
Developing	68.84	174.28	3.68	79.81	90.18	16.89	83.43	5.92
	Diphtheria	HIV.AIDS	GDP	Population	thin10_19	thin5_9	ICOR	
Developed	94.67	0.10	19530.34	7196869	1.42	1.44	0.83	
Developing	83.80	2.13	3831.99	7914268	4.65	4.71	0.61	

개발도상국의 여부로 각 변수의 평균 값을 구하여 특성을 살펴본 결과, Developing 는 Developed 에 비해 확실히 Adult.Mortality , under5\_death , HIV.AIDS , thin10\_19, thin5\_9 와 같은사망 및 부정적인 건강 지표가 높게 나타나는 것을 확인할 수 있다. 그에 비해 Developed 는 Life와 면역 요인을 나타내는 변수가 높게 나타나는 모습이다.

## 2. 비지도학습

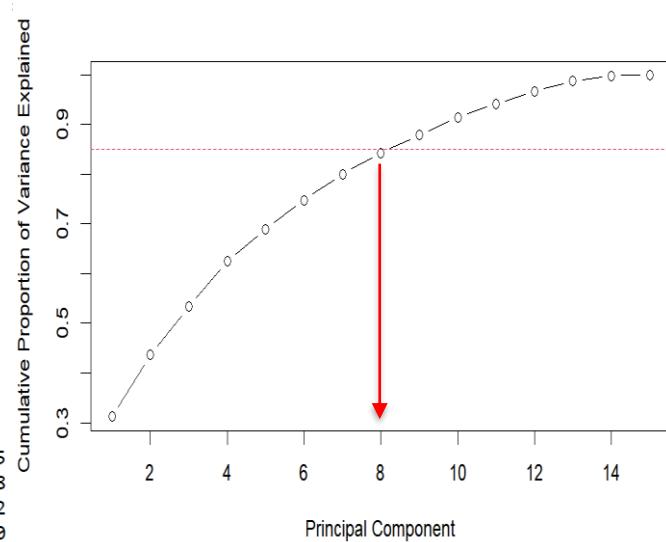
### 2-1. PCA

비지도 학습인 k-means 를 이용하여 클러스터링을 진행하기 전, 사용되는 데이터에 변수가 범주형을 제외하고는 15 가 있어 차원 축소를 위해 PCA를 사용하여 데이터의 차원을 줄였다. 누적 분산의 정도를 기준으로 삼아 PCA 결과를 살펴봤을 때, 전체 데이터 분산의 약 85%를 설명할 수 있는 8 개의 주성분을 선택하는 것이 적절하다고 판단하였다.

Importance of components:

	Comp.1	Comp.2	Comp.3
Standard deviation	2.1618776	1.3690937	1.20445463
Proportion of Variance	0.3118265	0.1250596	0.09679028
Cumulative Proportion	0.3118265	0.4368862	0.53367645
	Comp.4	Comp.5	Comp.6
Standard deviation	1.16763614	0.98683582	0.93630825
Proportion of Variance	0.09096323	0.06497416	0.05849093
Cumulative Proportion	0.62463968	0.68961384	0.74810477
	Comp.7	Comp.8	Comp.9
Standard deviation	0.88715531	0.79458951	0.73856811
Proportion of Variance	0.05251098	0.04212467	0.03639418
Cumulative Proportion	0.80061576	0.84274042	0.87913460
	Comp.10	Comp.11	Comp.12
Standard deviation	0.72685688	0.64304691	0.60541402
Proportion of Variance	0.03524915	0.02758901	0.02445433
Cumulative Proportion	0.91438375	0.94197277	0.96642710
	Comp.13	Comp.14	Comp.15
Standard deviation	0.57348661	0.38256296	0.167199568
Proportion of Variance	0.02194307	0.00976465	0.001865182
Cumulative Proportion	0.98837017	0.99813482	1.000000000

[figure 5. PCA 결과]



[figure 6. PCA Cumulative Proportion of Variance plot]

각 주성분을 살펴 본 결과, Comp.1 은 주성분을 구성함에 있어서 변수 간의 값에 차이는 존재하지만, population 을 제외하고는 유난히 큰거나 작은 값을 갖지 않는 경향을 볼 수 있다. 따라서 제 1 주성분은 모든 변수를 종합적으로 반영하고 있다고 판단된다. Comp.2 는 Hepatitis B(B 형 간염 접종률), Polio(1 세 아동의 소아마비 예방접종률), Diphtheria (1 세 아동의 디프테리아 예방접종률) 등이 특히 면역 요인과 관련된 변수가 주성분을 구성함에 있어서 높은 값을 갖는 것을 확인할 수 있다. Comp.3 은 1000 명당 HIV/AIDS 으로 인한 사망률, 1 인당 알코올 소비량(리터), 5 세 이하 아동의 사망자 수, 성인 사망수 등 특히 사망 정보와 관련되어 있는 것을 알 수 있다.

또한, Life, b 형 간염 예방접종률, 소아마비 예방접종률, 디프테리아 예방접종률, gdp 등은 양의 값을 갖는 반면 성인사망률, 5 세 미만 사망률, 에이즈 사망률은 음의 값을 가지므로 주성분을 구성하는데 서로 상반되는 영향을 미치는 것을 확인할 수 있다.

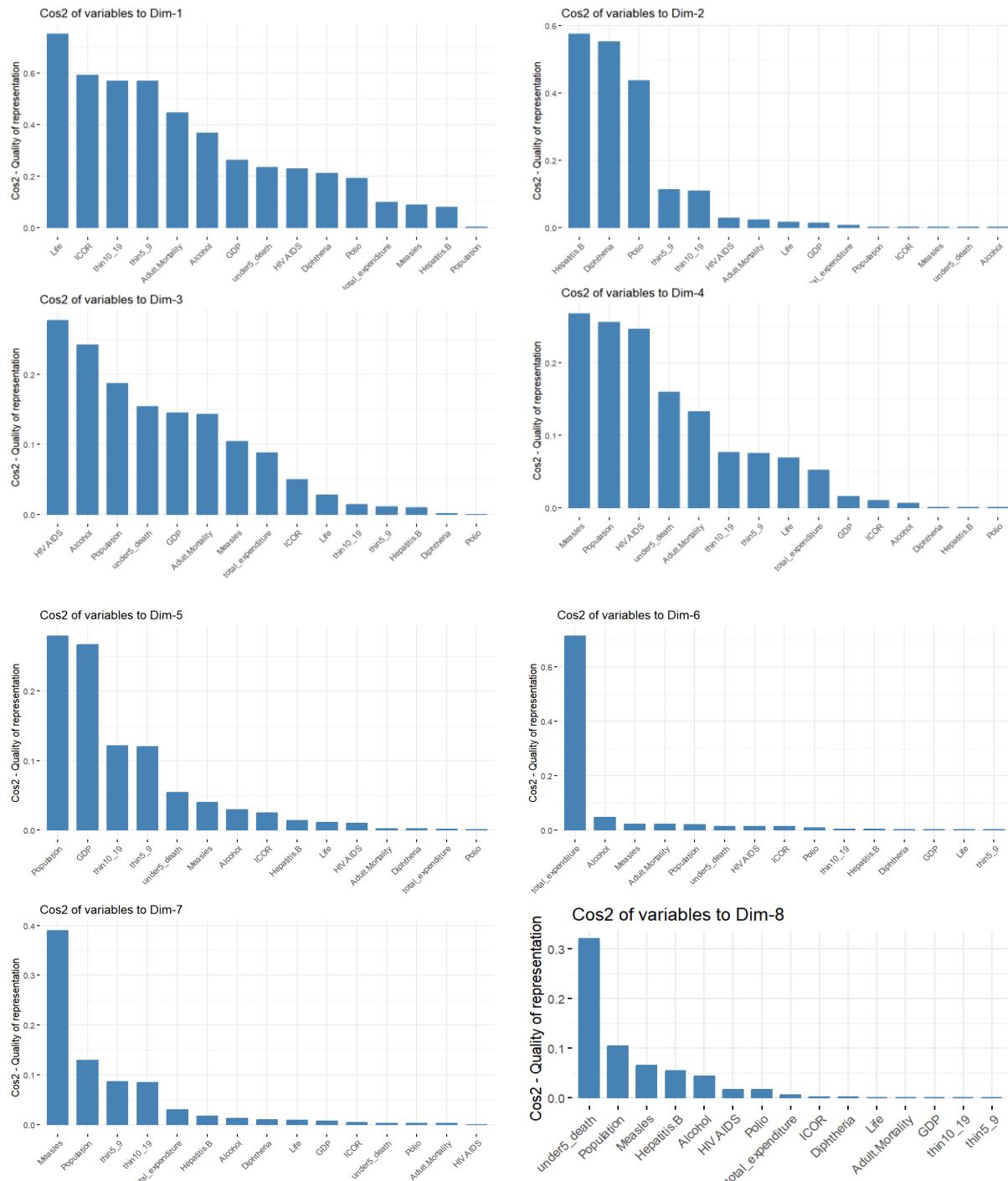
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Life	0.40066313	0.08989306	0.137509727	0.22422300	0.10446389	0.04084902	0.10419573	0.037476366
Adult.Mortality	-0.30850718	-0.10996242	0.313882581	-0.31163244	-0.04423502	-0.15269768	-0.05748745	0.036608913
Alcohol	0.28013881	0.02019302	0.408185388	-0.06777879	0.17320508	-0.23082068	0.12781177	-0.262630201
Hepatitis.B	0.12952056	0.55390286	0.080265784	0.01187229	-0.11871667	0.05089359	-0.14799788	-0.294397108
Measles	-0.13658545	0.0222158	-0.267563850	0.44287862	0.20243602	0.15405949	-0.70372006	0.322268951
under5_death	-0.22292443	0.02209065	0.325352089	0.34170748	-0.23487083	0.12330395	-0.06048649	-0.711016975
Polio	0.20238432	0.48217671	-0.004237598	0.01105742	-0.02523550	-0.09225994	-0.06033891	0.161401435
total_expenditure	0.14466996	-0.06644961	-0.246061991	-0.19477310	-0.03336527	0.90171616	0.19302473	0.094305809
Diphtheria	0.21215367	0.54255210	-0.030678301	0.01251563	-0.04266236	-0.04260810	-0.11272465	0.045786021
HIV.AIDS	-0.22031058	-0.12196182	0.436842775	0.42482412	0.09778889	-0.12054358	0.01208992	0.163878132
GDP	0.23642261	0.08474746	-0.315689369	0.10570145	0.52333885	-0.04199246	0.09158528	-0.027874195
Population	0.01496666	-0.02678657	-0.358698764	0.43283180	-0.53569748	-0.14834336	0.40560956	0.405887860
thin10_19	-0.34850327	-0.24071270	0.098221932	0.23604421	0.35273592	0.05434886	0.32670647	0.021148490
thin5_9	-0.34840548	-0.24523288	0.087404625	0.23414428	0.35084303	0.03655792	0.33119536	0.003997728
ICOR	0.35537707	0.02590021	-0.184709424	0.08497822	0.15934440	-0.11915834	0.07504585	-0.048239383

[figure 7. PCA component]

component	Function
Comp.1	$0.40066313 * \text{Life} - 0.30850718 * \text{Adult.Mortality} + 0.28013881 * \text{Alcohol} + \dots + 0.35537707 * \text{ICOR}$
Comp.2	$0.08989306 * \text{Life} - 0.10996242 * \text{Adult.Mortality} + 0.02019302 * \text{Alcohol} + \dots + 0.02590021 * \text{ICOR}$
Comp.3	$0.137509727 * \text{Life} - 0.313882581 * \text{Adult.Mortality} - 0.408185388 * \text{Alcohol} + \dots - 0.184709424 * \text{ICOR}$
Comp.4	$0.22422300 * \text{Life} - 0.31163244 * \text{Adult.Mortality} - 0.06777879 * \text{Alcohol} + \dots - 0.08497822 * \text{ICOR}$

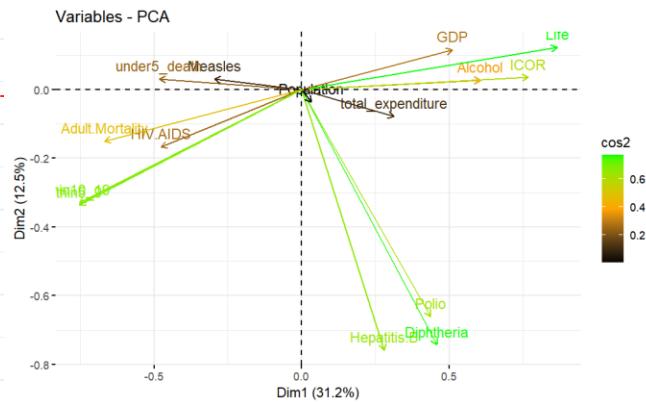
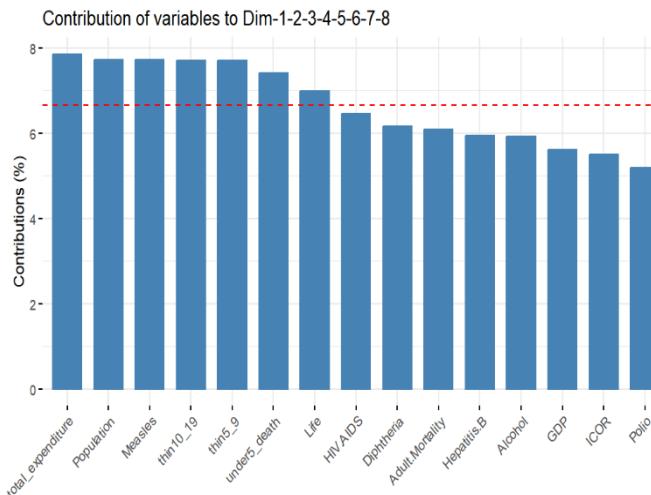
[table 10. 주성분 함수]

각 주성분에 의해서 높게 설명되는 변수들을 시각화 한 그래프이다. 결과는 앞의 PCA component 결과와 동일하다.



[figure 8. 각 PCA component]

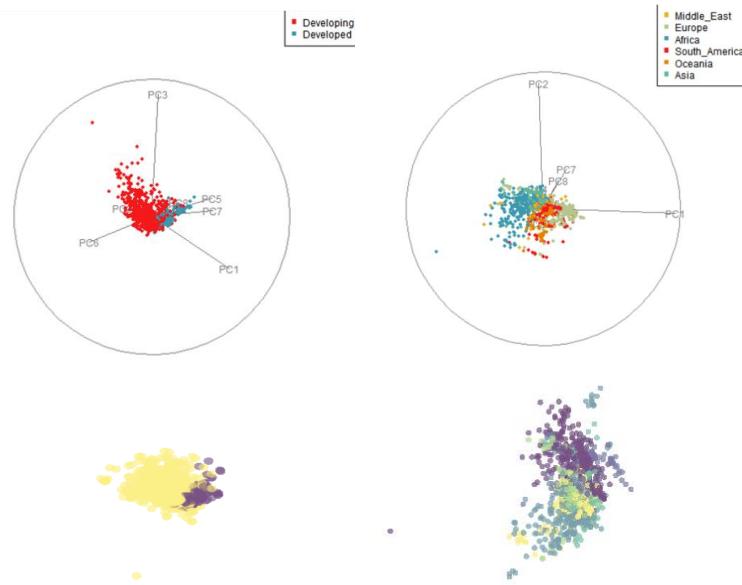
전반적으로 살펴보면, 주성분을 구성하는데 Total Expenditure, Population, Measles, Thin10\_19, Thin5\_9, 그리고 Life가 유의미하게 높은 기여도를 가졌음을 확인할 수 있다. 또한, Biplot을 통해 (GDP, Alcohol, ICOR, Life), (under5 death, Measles), (Adult Mortality, HIV/AIDS, Thin5\_9, Thin10\_19), (Total Expenditure, Polio) 변수 간의 상관관계가 높은 것을 알 수 있다.



[figure 9. 전체 PCA component]

[figure 10. PCA variables]

PCA를 통해 차원축소를 한 데이터를 사용했을 때의 결과를 원데이터와 비교하면, 원자료의 모든 변수 22개를 이용했을 때 보다 주성분 분석을 통한 8개의 주성분을 이용했을 때 클러스터링 결과가 비교적 뚜렷한 것을 볼 수 있다. 특히 변수의 레벨이 6개인 범주형 변수 country\_g를 사용한 경우에 더 큰 차이를 확인할 수 있다.



[figure 9. PCA 클러스터링 결과 비교]

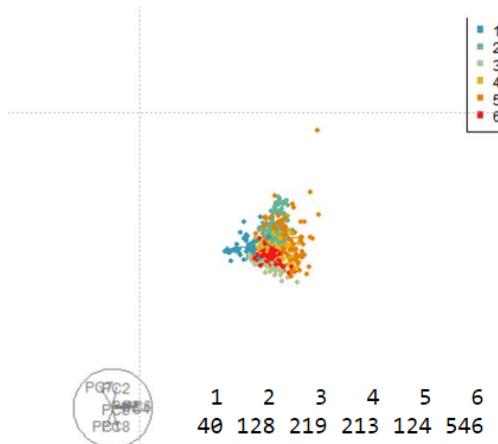
## 2-2. k-means

앞서 진행한 PCA 결과를 이용해, 주성분 8 개로 구성된 데이터로 k-mean 를 진행하였다. 원자료와 클러스터 결과를 비교하기 위해 K-means 실행 시 클러스터 개수를 각각 6 개, 2 개로 설정하였다.

### ①country\_g (level 6 개)

투어 결과를 살펴보면 각 군집끼리는 잘 뭉쳐져 있으나, 군집 간의 거리가 매우 가까워 경계부분에서는 데이터가 섞여있는 모습이며, 클러스터 1 부터 6 까지 분류된 데이터의 개수에 차이가 많이 나는 결과를 확인할 수 있다.

5 번 군집에 포함되어 있는 이상치는 원데이터에서 population 등의 몇몇의 변수 값이 다른 나라들보다 매우 큰 india 의 데이터로 예상되어, 5 번 군집이 원자료에서 india 가 속한 Asia 군집이라고 가정하였다. 이후 두 군집이 비슷한 경향을 갖는지 확인을 위해 country 별 평균을 비교한 결과, 평균값이 유사하지 않았다. 따라서 k-means 를 이용한 클러스터링은 군집이 비교적 잘 나눠지지만, 실제로 원자료와 대조하여 비교하는 것은 어렵다고 판단된다.



[figure 10. kmeans 결과 1]

country_g	Life	Adult.Mortality	Alcohol	Hepatitis.B	Measles	under5_death	Polio
Africa	60.56981	263.3896	2.646169	76.68506	142.48052	33.961039	78.70779
Asia	68.69565	168.8783	2.054870	90.41739	150.54783	19.686957	91.03478
Europe	76.34420	105.6981	8.779596	83.61725	69.61995	1.469003	91.91644
Middle_East	74.02571	104.8714	1.287286	82.44286	211.27143	15.257143	80.11429
Oceania	70.82018	132.6789	2.344679	74.50459	23.87156	1.174312	79.33945
South_America	73.46094	133.8990	5.132997	81.15825	24.23906	11.760943	85.19529

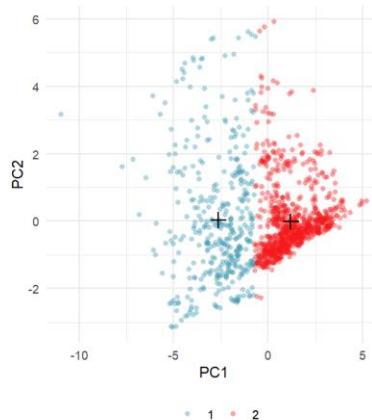
[figure 11. 원자료 country\_g 평균]

k_cl	Life	Adult.Mortality	Alcohol	Hepatitis.B	Measles	under5_death	Polio
1	49.38500	522.90000	4.731750	76.40000	115.95000	26.75000	82.80000
2	69.97969	157.07031	2.914687	41.52344	36.39062	9.140625	50.42969
3	80.06895	72.94977	10.497032	81.16438	94.73516	1.141553	94.47945
4	66.16854	194.11737	1.985164	87.42723	64.30986	12.666667	89.86854
5	59.06210	253.58065	1.790403	69.31452	426.68548	69.016129	65.36290
6	72.73407	131.02930	4.895037	90.97436	27.41026	7.714286	92.79487

[figure 12. kmeans country\_g 평균]

## ② Status (level 2 개)

앞서 클러스터를 6 개로 설정했을 때보다 비교적 명확하게 군집이 분리된 것을 확인할 수 있다. 마찬가지로 이상치가 india라고 가정한 후, 파란 1 번 군집의 평균값과 india 의 status 값인 developing 데이터들의 평균값을 비교한 결과는 유사하지 않았다.



[figure 13. kmeans 결과 2]

Status	Life	Adult.Mortality	Alcohol	Hepatitis.B	Measles	under5_death	Polio
Developed	78.73836	82.90868	10.546758	87.43379	76.85388	0.9315068	94.42466
Developing	68.84358	174.28449	3.681256	79.81637	90.18554	16.8924833	83.43673

[figure 14. 원자료 status 평균]

k_cl	Life	Adult.Mortality	Alcohol	Hepatitis.B	Measles	under5_death	Polio	total_expenditure
1	61.59066	254.6237	2.191010	70.72980	173.18434	33.093434	73.77020	5.45053
2	74.60915	114.9874	6.076773	85.84211	49.23913	5.552632	90.56979	6.43286

[figure 15. kmeans status 평균]

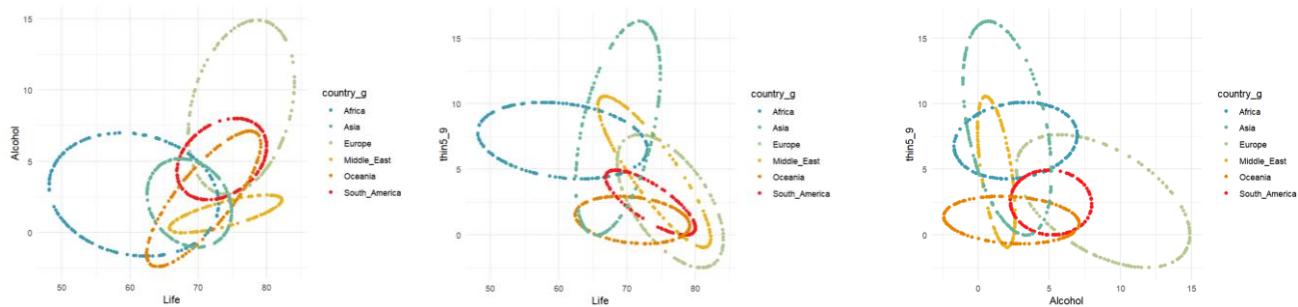
### 3. 지도학습

#### 3-1. LDA

LDA를 진행을 위해 기존 데이터를 스케일링 한 후, train data(7)와 test data(3)를 분할 생성하였다. 이후 train data를 이용하여 LDA 모델을 학습하고, test 데이터를 이용해 예측하였다.

##### ①country\_g (level 6 개)

country\_g에 따른 각 그룹의 분포를 살펴보면 대부분 클러스터의 모양이 일치하지 않는 모습이다. 따라서 LDA의 가정을 만족하지 못하는 상태라고 판단된다.



범주형 변수 country\_g를 사용한 경우 판별 함수는 총 5개가 생성되었으며, 선형 판별 함수의 계수를 살펴보면 Life, Alcohol, Measles, Under-Five Deaths, Total\_expenditure, HIV/AIDS, GDP, Thin5\_9 등이 각 판별 함수에 큰 영향을 주는 변수임을 확인할 수 있다. 따라서 종합적으로 고려하였을 때, 데이터를 6개의 대류으로 분류할 때, 18개의 변수 중 Life, Alcohol, Measles, Under-Five Deaths 즉, 면역, 사망, 경제 요인 등 다양한 요인에 의해 영향을 받는 것을 알 수 있다.

Coefficients of linear discriminants:

	LD1	LD2	LD3	LD4	LD5
Life	0.66363745	0.11092100	-1.08709528	-0.03333450	0.4885071353
Adult.Mortality	-0.06938632	0.11243123	-0.06207634	0.03605237	0.3343136600
Alcohol	0.78137459	1.07830980	0.55492180	-0.14198530	-0.2444321160
Hepatitis.B	-0.14899431	-0.11389347	-0.24302136	0.12291429	-0.1059340768
Measles	-0.10657479	0.04401924	-0.41531916	-0.11490825	0.6485131410
under5_death	-0.16830298	0.07934980	0.06139814	-0.45050834	0.7373377833
Polio	-0.04398315	0.06646687	-0.06414036	0.14117889	-0.2669777310
total_expenditure	0.08782725	0.00113375	-0.24138303	0.67448155	0.0009371451
Diphtheria	-0.04189410	0.40503667	0.33794720	0.21998205	0.4490421973
HIV.AIDS	-0.02276926	-0.15922303	0.01397435	0.51723286	-0.0577191967
GDP	-0.15525406	-0.13310303	0.21224396	0.13662121	0.5900587982
Population	0.01426769	-0.01198278	-0.07808703	-0.06182466	0.3353263969
thin10_19	0.05479890	0.02745610	0.31057770	-1.84829029	0.0646652390
thin5_9	0.77532848	0.95014026	-0.66713689	1.76862146	-0.1743824324
ICOR	-0.27778863	0.15589999	-0.44757193	-0.23920713	0.2246127176

Proportion of trace:

LD1	LD2	LD3	LD4	LD5
0.4527	0.2496	0.1882	0.0826	0.0268

[figure 16. LDA country\_g 결과]

LD	Function
LD1	0.66363745*Life + … + 0.78137459*Alcohol + … + 0.77532848*thin5_9
LD2	1.07830980*Alcohol + … + 0.95014026*thin5_9
LD3	0.55492108*Alcohol + … -0.66713689*thin5_9
LD4	-0.67448155*total_expenditure + … - 0.51723286*HIV.AIDS + … + 1176862146*thin5_9
LD5	-0.6485131410*Measles + 0.7373377833*under5_death + … -0.5900587982

[table 11. Country\_g 판별함수]

판별분석은 예측 정확도 67%으로 높지 않은 결과를 나타낸다. 원데이터의 결과를 이용하여 matrix를 생성한 결과, Asia, South America에서 특히 오분류된 데이터가 많았으며 전반적으로 모든 경우에서 오분류가 발견된다.

투어를 이용해 시각화한 결과에서도 같은 군집끼리 뭉쳐져있는 것이 보이기는 하지만 명확하지 않으며, 특히 Aisa(하늘색)와 south America(빨간색)의 경우 데이터가 넓게 퍼져 있는 모습을 볼 수 있다. Boundary를 시각화한 결과 역시 뚜렷하지 않은 군집 결과를 나타낸다. 이는 앞서 확인한 것처럼, LDA의 기본 가정을 만족하지 못했기 때문이라고 생각되며, 현재의 데이터를 분류하고자 할 때 LDA가 적절하지 않은 방법이라고 판단된다.

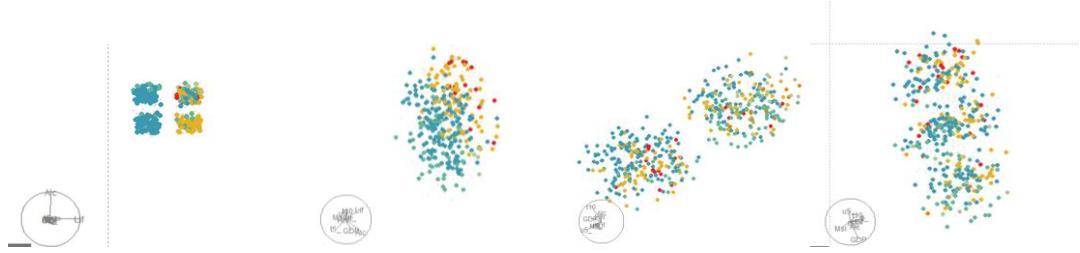
Confusion matrix						
	Test.pc					
	Africa	Asia	Europe	Middle_East	Oceania	South_America
Africa	68	19	8	8	4	2
Asia	0	18	0	3	7	7
Europe	0	7	74	4	2	13
Middle_East	2	0	0	18	2	3
Oceania	0	3	7	0	22	0
South_America	5	2	9	2	8	56

[table 12. LDA Confusion Matrix- country\_g]

Overall statistics	
Accuracy	0.06667
95% CI	(0.6169,0.7139)
No Information Rate	0.2572
P-value [Acc > NIR]	0.000000000000000022
Kappa	0.586
McNemar's Test P-Value	0.00000001025

[table 13. LDA statistics - country\_g]

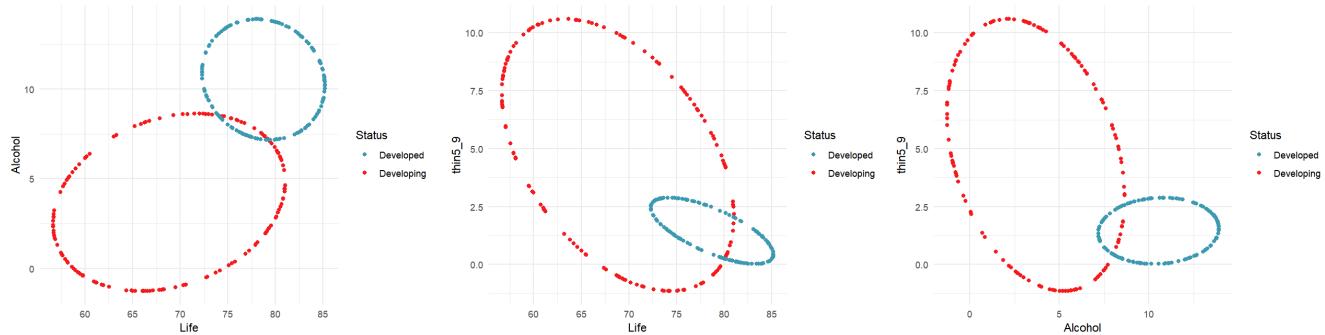
[figure 17. LDA tour – country\_g]



[figure 18. LDA boundary – country\_g]

## ② Status (level 2 개)

Status에 따른 각 그룹의 분포를 살펴본 결과, 비교적 두 클러스터의 모양이 일치하는 모습이다. 따라서 이 경우에는 LDA의 가정을 만족한다고 판단할 수 있다.



범주형 변수 Status를 사용한 경우, 판별 함수는 1개가 생성되었으며, 판별 함수에 큰 영향을 주는 변수로는 Alcohol, Life, GDP, Measles, Population가 있다. 따라서 데이터를 2개의 상태로 분류할 때는 18개의 변수 중 Alcohol, Life, GDP, Measles, Population 즉, 건강, 사회, 면역 요인에 의해 영향을 받는 것을 알 수 있다.

Coefficients of linear discriminants:

	LD1
Life	-0.371446228
Adult.Mortality	0.098060431
Alcohol	-1.000452367
Hepatitis.B	-0.114732328
Measles	-0.187044504
under5_death	0.024762911
Polio	-0.005123114
total_expenditure	-0.020853887
Diphtheria	0.056541771
HIV.AIDS	-0.081653804
GDP	-0.360047079
Population	0.153360339
thin10_19	-0.063246462
thin5_9	0.064000051
ICOR	0.063832779

LD	Function
LD1	$\begin{aligned} & -0.371446228 * \text{Life} + \dots -1.000452367 * \text{Alcohol} + \dots \\ & -0.187044504 * \text{Measles} \\ & + \dots -0.360047079 * \text{GDP} + 0.153360339 * \text{population} + \dots \end{aligned}$

[table 12. 판별함수- status]

[figure 16. LDA 결과 - status]

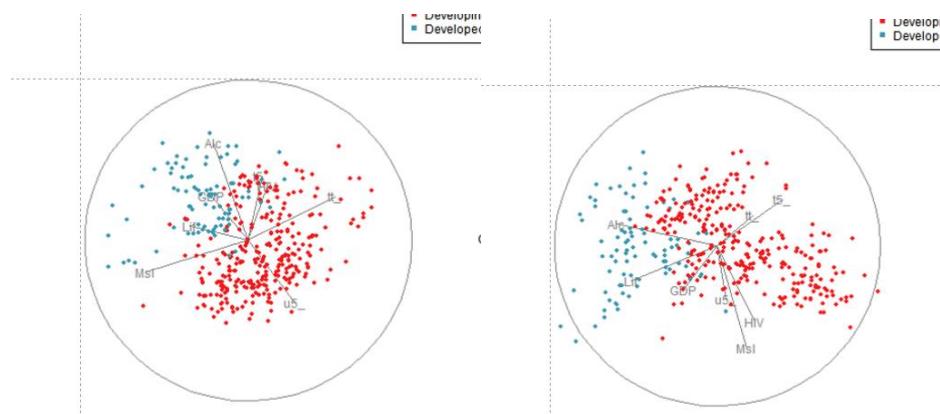
Status를 이용한 판별분석은 예측 정확도 92%으로 높은 결과를 나타낸다. 원데이터의 결과를 이용하여 matrix를 생성한 결과, developing 이 developed로 잘못 예측된 경우가 많았다. 투어를 이용해 시각화한 결과에서도 country\_g를 이용했을 때 보다 분류가 잘 된 모습을 볼 수 있는데, 이것은 country\_g를 이용했을 때 보다 데이터가 LDA의 가정을 만족하며, 클러스터의 개수 역시 2개로 적은 편이기 때문이라고 생각된다.

Confusion matrix			
		Test.pc2	
		Developed	Developing
Developed		74	3
Developing		25	279

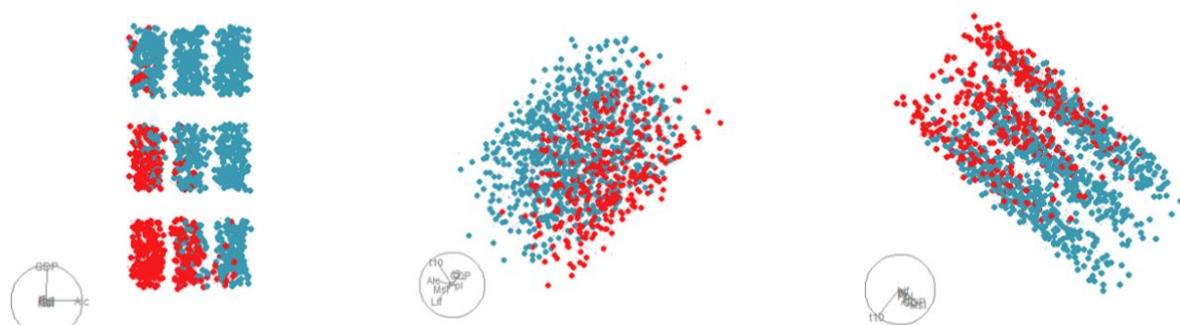
[table 13. Confusion Matix - status]

Overall statistics	
Accuracy	0.9265
95% CI	(0.8955, 0.9506)
No Information Rate	0.7402
P-value [Acc > NIR]	<0.00000000000000022
Kappa	0.7941

[table 14. LDA statistics – status]



[figure 17. LDA tour - Status]



[figure 19. LDA boundary - Status]

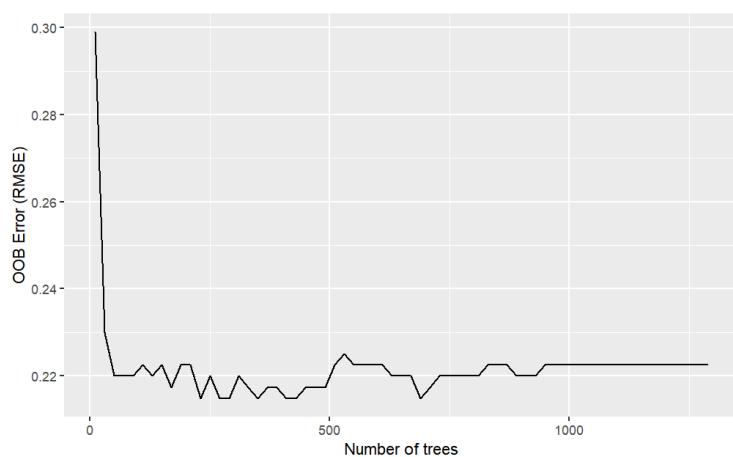
### 3-2. Random Forest

Random forest를 진행을 위해 기존 데이터를 스케일링 한 후, train data(7)와 test data(3)를 분할 생성하였다. 이후 train data를 이용하여 Random forest 모델을 학습하고, test 데이터를 이용해 예측하였다.

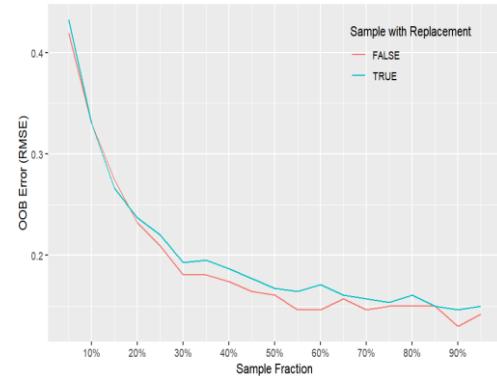
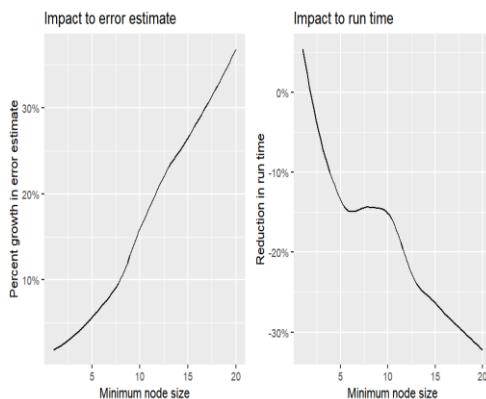
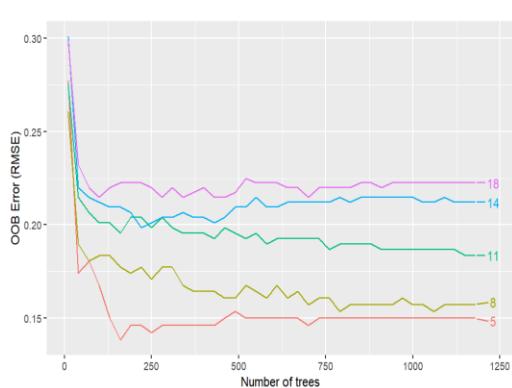
#### ①country\_g (level 6 개)

Random Forest (RF) 모델의 성능을 향상시키기 위해, RMSE 값을 가장 크게 만드는 방향으로 파라미터 탐색을 진행하였다. Tree 의 개수를 변화시키면서 RMSE 값의 변화를 그래프로 확인한 결과, ntree 값이 230에서 690 사이일 때 RMSE 가 0.2147539로 가장 작았습니다. ntree 값이 증가함에 따라 안정적인 결과를 얻을 수 있으므로, ntre를 690으로 선택하였다. 그 다음, ntree 값을 고려하여 다른 파라미터들의 최적값을 찾아 mtry=5, node size=5, sample fraction=0.8 (replace = FALSE)로 선택하였다. 최종적으로, ntree=690, 찾아 mtry=5, node size=5, sample fraction=0.8 (replace = FALSE)을 기준으로 조금씩 값을 변경해가며 RF 의 정확도를 가장 높일 수 있는 최적 파라미터를 확인하였고, ntree=690, mtry=5, node size=3, sample fraction = 0.9 (replace = FALSE)으로 설정하였다.

trees	rmse
230	0.2147539
270	0.2147539
290	0.2147539
350	0.2147539
410	0.2147539
430	0.2147539
690	0.2147539
170	0.2173571
330	0.2173571
370	0.2173571



[figure 20. Rf ntree – country\_g]



[figure 21. Rf mtry/ node size/ sample fraction – country\_g]

파라미터 조정 후 rf 모델의 정확도를 살펴보면, 전과 비교하여 accuracy 가 0.9843에서 0.9921로 향상되었으며, 오분류 개수도 6 개에서 3 개로 줄어든 것을 확인할 수 있다. 조정 후 대부분의 대륙에서는 잘 분류되었지만, 실제로는 south America 인데 Africa 로 오분류된 경우 1 개, Europe 이지만 south America 로 오분류 된 경우가 2 개 존재한다.

Confusion matrix						
	Africa	Asia	Europe	Middle_East	Oceania	South_America
Africa	90	0	0	0	0	2
Asia	0	33	1	0	0	0
Europe	0	1	111	0	0	0
Middle_East	0	0	0	21	0	0
Oceania	0	0	0	0	39	0
South_America	0	0	2	0	0	81

[table 15. Rf confusion matrix before - country\_g]

Overall statistics	
Accuracy	0.9843
95% CI	(0.966,0.9942)
No Information Rate	0.2992
P-value [Acc > NIR]	<0.00000000000000022
Kappa	0.98

[table 15. Rf statistics before - country\_g]

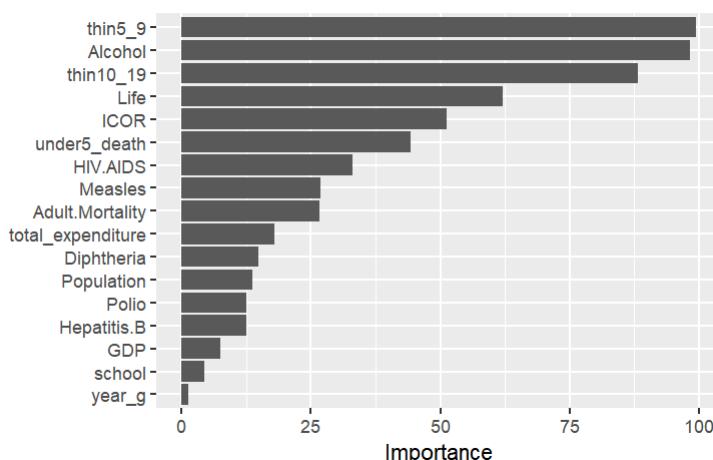
Confusion matrix						
	Africa	Asia	Europe	Middle_East	Oceania	South_America
Africa	90	0	0	0	0	0
Asia	0	34	0	0	0	0
Europe	0	0	112	0	0	2
Middle_East	0	0	0	21	0	0
Oceania	0	0	0	0	39	0
South_America	1	0	0	0	0	82

[table 16. Rf confusion matrix after - country\_g]

Overall statistics	
Accuracy	0.9921
95% CI	(0.9772,0.9984)
No Information Rate	0.294
P-value [Acc > NIR]	<0.00000000000000022
Kappa	0.99

[table 17. Rf statistics after - country\_g]

이후 rf 모델에서 impurity 를 감소시키는 것을 기준으로 변수의 중요도를 살펴봤을 때, Thine 5-9, Alcohol, Thine10\_19, Life 변수가 큰 중요도를 갖는 것을 확인할 수 있다.



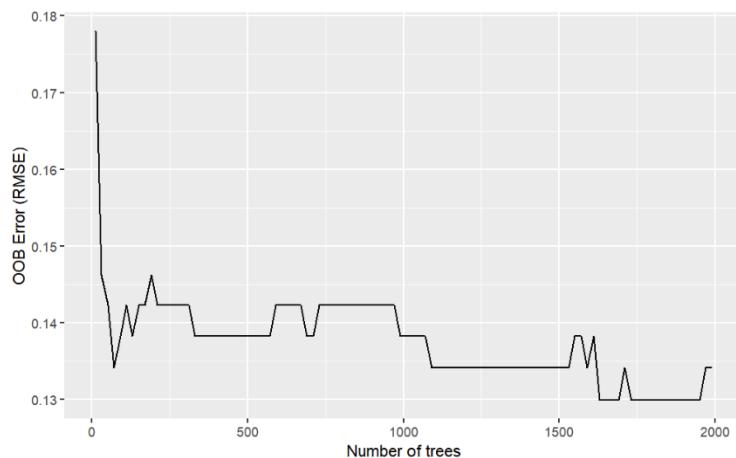
[figure 22. Rf importance - country\_g]

Variable	Importance
thin5_9	99.137072
Alcohol	93.981948
thin10_19	84.275387
Life	59.311441

[table 18. Rf importance - country\_g]

## ② Status (level 2 개)

위와 마찬가지로 rf 모델의 성능을 향상시키기 위해 rmse 를 가장 작게 갖도록 하는 방향으로 파라미터 탐색 과정을 진행하였다. 파라미터의 값을 확인한 결과, ntree=1610, mtry=8, nodesize=5, sample fraction=0.8 (replacement False)로 선택하였다. 이 값을 조금씩 변경해가며 rf 의 정확도가 가장 높게 나오도록 설정한 ntrees = 900, mtry = 5, min.node.size = 5, sample.fraction = 0.90(replace = FALSE)이다.



trees	rmse
710	0.1382845
990	0.1382845
1010	0.1382845
1030	0.1382845
1050	0.1382845
1070	0.1382845
1550	0.1382845
1570	0.1382845
1610	0.1382845
50	0.1422936

[figure 23. Rf ntree – Status]

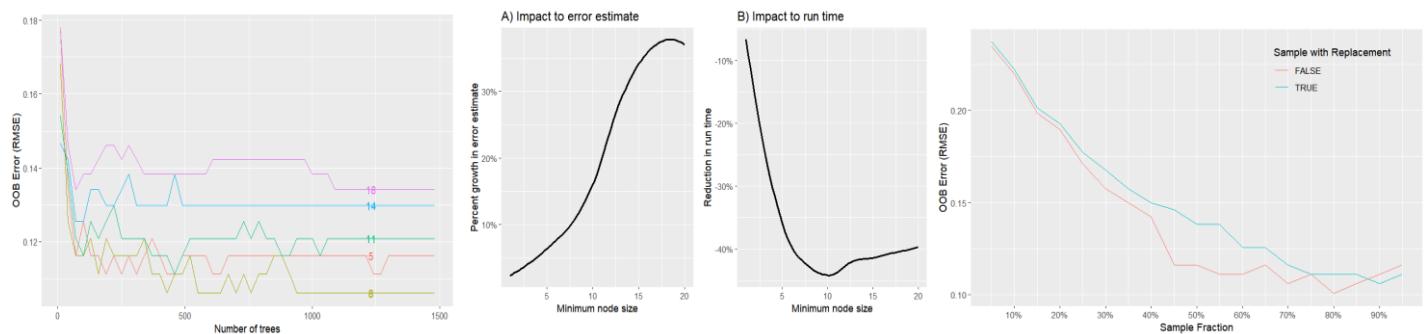


Figure 1[figure 24. Rf mtry/ node size/ sample fraction – Status]

파라미터를 조정한 후 rf 모델의 정확도를 살펴보면, 전과 비교하여 accuracy 가 0.9869 에서 0.9948 로 향상되었으며, 오분류 개수도 5 개에서 2 개로 줄어든 것을 확인할 수 있다. 오분류 된 데이터는 실제로는 Developed 인데 Developing 로 오분류된 경우 1 개, Developing 이지만 Developed 로 오분류 된 경우가 1 개 존재한다.

Confusion matrix		
	Developed	Developing
Developed	66	3
Developing	2	310

[table 19. Rf confusion matrix before - Status]

Overall statistics	
Accuracy	0.9921
95% CI	(0.9772,0.9984)
No Information Rate	0.294
P-value [Acc > NIR]	<0.000000000000000022
Kappa	0.99

[table 20. Rf statistics before- Status]

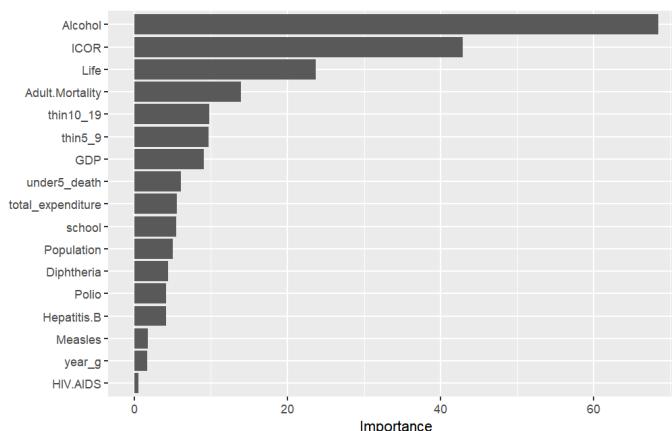
Confusion matrix		
	Developed	Developing
Developed	68	1
Developing	1	311

[table 21. Rf confusion matrix after - Status]

Overall statistics	
Accuracy	0.9921
95% CI	(0.9772,0.9984)
No Information Rate	0.294
P-value [Acc > NIR]	<0.000000000000000022
Kappa	0.99

[table 21. Rf statistics after- Status]

이후 rf 모델에서 impurity 를 감소시키는 것을 기준으로 변수의 중요도를 살펴봤을 때, Alcohol, ICOR, Life, Adult.Mortality 변수가 큰 중요도를 갖는 것을 확인할 수 있다.



[figure 25. Rf importance - Status]

Variable	Importance
Alcohol	68.3772738
ICOR	42.8567924
Life	23.6674802
Adult.Mortality	13.9111507

[table 22. Rf importance - Status]

### 3-3. LDA vs RF

지도 학습인 LDA 와 randomforest 의 정확도를 비교해 본 결과, LDA 를 이용해 6 개의 대륙으로 분류하고자 했을 때를 제외하고는 0.9 이상의 높은 정확도를 보였다. LDA 의 가정인 정규성을 만족하지 않는 데이터이므로 LDA 에 의해서는 분류가 잘 되지 않지만, 변수 Status 는 레벨이 2 개로 적어 정확도가 비교적 높게 나온 것으로 생각할 수 있다. Randomforest 의 경우는 모든 경우에서 0.99 의 매우 높은 정확도로 데이터를 분류하였다.

또한 모든 모델에서 중요도의 순서는 다르지만, 공통적으로 Life, Alcohol 을 중요한 변수로 사용한 것이 확인된다. 이를 통해 데이터를 6 개의 대륙, 혹은 개발도상국의 여부로 나누고자 할 때 Life, Alcohol 가장 영향력 있으며 유의미하게 사용되었다고 판단된다.

LDA		Random forest	
Country_g(6)	Status(2)	Country_g(6)	Status(2)
0.6667	0.9265	0.9921	0.9948
Life, Alcohol, Measles, Under-Five Deaths, Total_expenditure	Alcohol, Life, GDP , Measles ,Population	Thine 5-9, Alcohol, Thine10_19, Life	Alcohol, ICOR, Life , Adult.Mortality

[table 23, LDA vs RF]

## 4. 결론

다양한 국가를 기반으로 한 면역, 사망률, 경제적, 사회적, 그리고 기타 건강 관련 요인으로 구성된 데이터를 이용하여 건강 지표들이 국가의 상태에 따라 유의미한 차이가 있는지 탐색하기 위해 클러스터링 및 분류 분석을 수행하였다. 국가를 대륙으로 범주화한 country\_g, 국가의 경제적 상태를 나타내는 status, 사회적 상태를 나타내는 school 을 범주형 변수로 생성하였고, 시간의 흐름에 따라 건강 지표에 변화가 있을 것이라고 예상하여 연도를 기간으로 나눈 year\_g 도 재범주화 하였다. 하지만 변수에 따른 분포를 살펴본 결과, school 과 year\_g 에 따른 분포 차이가 존재하지 않아 country\_g 와 Status 만을 사용하여 k-means, LDA, Random Forest 분석을 각각 진행하였고, 결과를 시각화하여 확인하였다.

### k-means

K-means 를 하기 전, PCA 를 이용하여 차원축소를 하였다. 누적분산을 기준으로 8 개의 주성분을 선택했는데, 제 1 주성분은 모든 변수를 종합적으로 반영하고, 제 2 주성분은 특히 면역 요인, 제 3 주성분은 사망 정보와 관련되어 있다. 이를 바탕으로한 k-means 결과를 살펴보면, 각 군집끼리 잘 뭉쳐져 있으나, 군집 간의 거리가 매우 가까워 경계부분에서는 데이터가 섞여 있는 모습이다. 이상치의 통계량값을 이용하여 실제 country\_g/Status 값과 kmeans 의 군집 결과를 비교해보았지만, 같은 군집일 것이라고 예상된 두 군집의 경향이 매우 다른 것으로 확인되어 원자료와의 대조비교는 어렵다고 판단된다.

### LDA

분석을 진행하기 전에, Country\_g 에 따른 각 그룹의 데이터 분포를 살펴본 결과 대부분 클러스터 형태가 매우 상이하여 LDA 의 기본 가정을 만족하지 못하는 것으로 판단되었다. 판별 함수는 총 5 개가 생성되었으며, 6 개의 대륙으로 데이터를 분류할 때, 18 개의 변수 중 면역, 사망, 경제 요인 등 다양한 요인(Life, Alcohol, Measles, Under-Five Deaths )에 의해 영향을 받는 것을 알 수 있다. 이 판별분석은 예측 정확도 67%으로 높지 않았고, 특히 Asia, South America에서 오분류된 데이터가 많았다. 투어를 이용해 시각화한 결과, 같은 군집끼리 뭉쳐있지만 경계가 명확하지 않았으며, 특히 Aisa(하늘색)와 south America(빨간색)의 경우 데이터가 넓게 퍼져 있는 것으로 나타났다. 이는 LDA 의 기본 가정을 만족하지 못했기 때문이라고 판단되어, 현재의 데이터를 분류하고자 할 때 LDA 가 적절하지 않은 방법이라고 생각된다. 그러나 Status 의 경우에는 각 그룹의 분포가 비교적 LDA 의 기본 가정을 만족하는 모습을 보인다. 판별 함수는 1 개가 생성되었으며, 데이터를 개발도상국의 여부에 따라 분류할 때 18 개의 변수 중 건강, 사회, 면역 요인(Alcohol, Life, GDP, Measles, Population)에 의해 영향을 받는 것을 알 수 있다. 예측 정확도는 92%으로 높은 결과를 나타냈는데, 데이터가 LDA 기본 가정을 잘 만족하고, 클러스터의 개수 역시 2 개로 적은 편이기 때문이라고 생각된다.

### Random Forest

Random Forest (RF) 모델의 성능을 향상시키기 위해, RMSE 값을 가장 크게 만드는 방향으로 파라미터 탐색을 진행하였다. 튜닝 후 정확도가 미세하지만 향상된 것을 확인하였고, south America 인데 Africa 로, Europe 이지만 south America 로 오분류된 데이터가 발견되었다. Status의 경우에는 Developed 와 Developing 가 서로 오분류되는 경우가 하나씩 발생하였다. rf 모델에서 impurity 를 감소시키는 것을 기준으로 변수의 중요도를 살펴봤을 때, 각각 Thine 5-9, Alcohol, Thine10\_19, Life 와 Alcohol, ICOR, Life, Adult.Mortality 변수가 큰 중요도를 갖는 것을 확인할 수 있다.

종합적으로 살펴보면, 비지도 학습인 LDA에 비해 지도 학습인 RF의 성능이 확실히 좋았으며, 공통적으로 데이터를 6개의 대류, 혹은 개발도상국의 여부로 나누고자 할 때 Life, Alcohol 가장 영향력 있으며 유의미하게 사용되었다는 것을 발견할 수 있었다. 또한 최종적으로 사용한 19개의 변수 중 15개가 연속형 변수로, 클러스터링 및 분류를 진행할 때 비교적 많은 정보를 이용하였다고 생각된다. 따라서 분석 결과에 데이터의 특성이 잘 반영되었지만, 분류가 명확하게 되지 않은 한계가 존재한다. LDA와 RF 외의 더 다양한 분석 방법을 시도해보고, 다른 성능 측정 지표들을 이용하여 살펴보면 데이터로부터 더 의미 있는 결과를 얻을 수 있을 것이라고 예상된다.