

دانشگاه صنعتي امیرکبیر  
(پلی تکنیک تهران)

دانشكده علوم کامپیوتر و ریاضی

پروژه سوم داده کاوی

نگارش

سید حسین محمدی

مدرس

دکتر مهدی قطعی

شهریور 1401

چكيده

در این گزارش، تلاش میکنیم میانگین طول عمر مردم یک کشور را بر اساس اطلاعت بسیار کلی مربوط به کشور مقیم را تخمین بزنیم. برای این کار از پلات، تعیین ویژگی و تحلیل های دیگر برای پیدا کردن مهمترین ویژگی ها استفاده میکنیم و یک مدل خطی برای آن میسازیم.

واژه‌های کلیدی:

Python, Regression, Feature Selection, Plots

|  |  |
| --- | --- |
| فهرست مطالب | صفحه |

[فصل اول: تبدیل داده 5](#_Toc117457259)

[1-1- حذف ستون و ردیف اضافی 5](#_Toc117457260)

[1-2- عناصر تهی 6](#_Toc117457261)

[1-3- حذف نویز 7](#_Toc117457262)

[1-4- انتخاب ستون 7](#_Toc117457263)

[فصل دوم: ایجاد مدل 9](#_Toc117457264)

[2-1- مشکل تفاوت بین کشور فقیر و پیشرفته 9](#_Toc117457265)

[2-2- ایجاد مدل 9](#_Toc117457266)

[منابع و مراجع 11](#_Toc117457267)

داده استفاده شده در این گزارش برگرفته از Kaggle میباشد و آمار های مربوط به وضعیت کشور ها در طی 16 سال را نشان میدهد. این ستون ها برای 190+ کشور وجود دارد.

معنی این ستون ها به شکل زیر است:

Country

Year

Status - Developed or Developing status

Life expectancy - Life Expectancy in age

Adult Mortality - Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)

infant deaths - Number of Infant Deaths per 1000 population

Alcohol - Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)

percentage expenditure - Expenditure on health as a percentage of Gross Domestic Product per capita(%)

Hepatitis B - Hepatitis B (HepB) immunization coverage among 1-year-olds (%)

Measles - Measles - number of reported cases per 1000 population

BMI - Average Body Mass Index of entire population

under-five deaths - Number of under-five deaths per 1000 population

Polio - Polio (Pol3) immunization coverage among 1-year-olds (%)

Total expenditure - General government expenditure on health as a percentage of total government expenditure (%)

Diphtheria - Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)

HIV/AIDS - Deaths per 1 000 live births HIV/AIDS (0-4 years)

GDP - Gross Domestic Product per capita (in USD)

Population - Population of the country

thinness 1-19 years - Prevalence of thinness among children and adolescents for Age 10 to 19 (%)

thinness 5-9 years - Prevalence of thinness among children for Age 5 to 9(%)

Income composition of resources - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

Schooling - Number of years of Schooling(years)

# فصل اول: تبدیل داده

## حذف ستون و ردیف اضافی

باید دقت کرد که هدف نهایی ایجاد یک مدل برای تخمین امید به زندگی مردم از طریق اطلاعات آن کشور هست. به همین دلیل ستون هایی مانند نام کشور، پیشرفته بودن یا نبودن، جمعیت کل و سال را به مدل نهایی نمیدهیم.

جمعیت کل نباید تاثیری رو میانگین عمر بگذارد. لذا باید توجه کرد که کشور ها با تعداد کم جمعیت ممکن هست باعث مشکل شود و به همین دلیل ردیف ها با جمعیت کمتر از 5000 را حذف میکنیم که شامل 77 ردیف (2.6%) میشود.

نام کشور قوی ترین فاکتور برای تخمین طول عمر هست اما کمکی برای ایجاد یک مدل مبتنی بر آمار های آن کشور نمی کند و برای برخی مدل باعث حواس پرتی میشود.

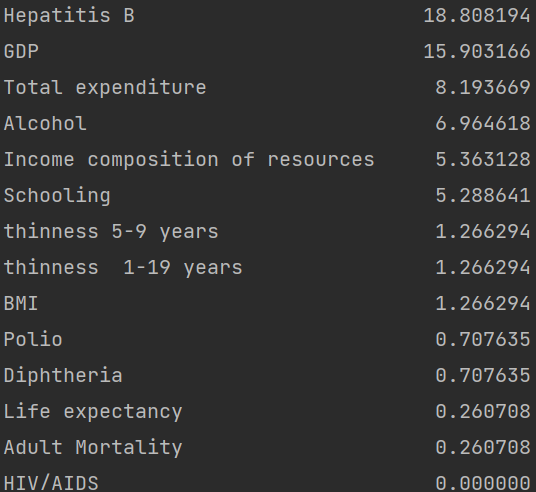
مدل نهایی ما آمار فعلی کشور را میگیرد و قرار نیست دسترسی به اطلاعات سال های قبل داشته باشد. به این دلیل و دلایل مشابه به نام کشور، این ستون را حذف میکنیم.

نهایتا ستون مربوط به پیشرفته بودن یک کشور را نیز حذف میکنیم زیرا که تعداد کمی مثبت هستند اما تاثیر زیادی دارد و ممکن هست مدل را گمراه کند.

حال مقادیر outlier شدید را با کمک zscore حذف میکنیم. اما به دلیل تفاوت شدید بین داده کشور های پیشرفته و غیر پیشرفته، آن دو را جداگانه در نظر میگیریم. مقادیر با zscore بالاتر از 4.5 برای کشور های غیر پیشرفته و بالاتر از 4 برای کشور های پیشرفته را حذف میکنیم (که شامل 176 ردیف – 6.15% داده میشود).

## عناصر تهی

درصد مقادیر تهی برای هر ستون به شکل زیر هست:

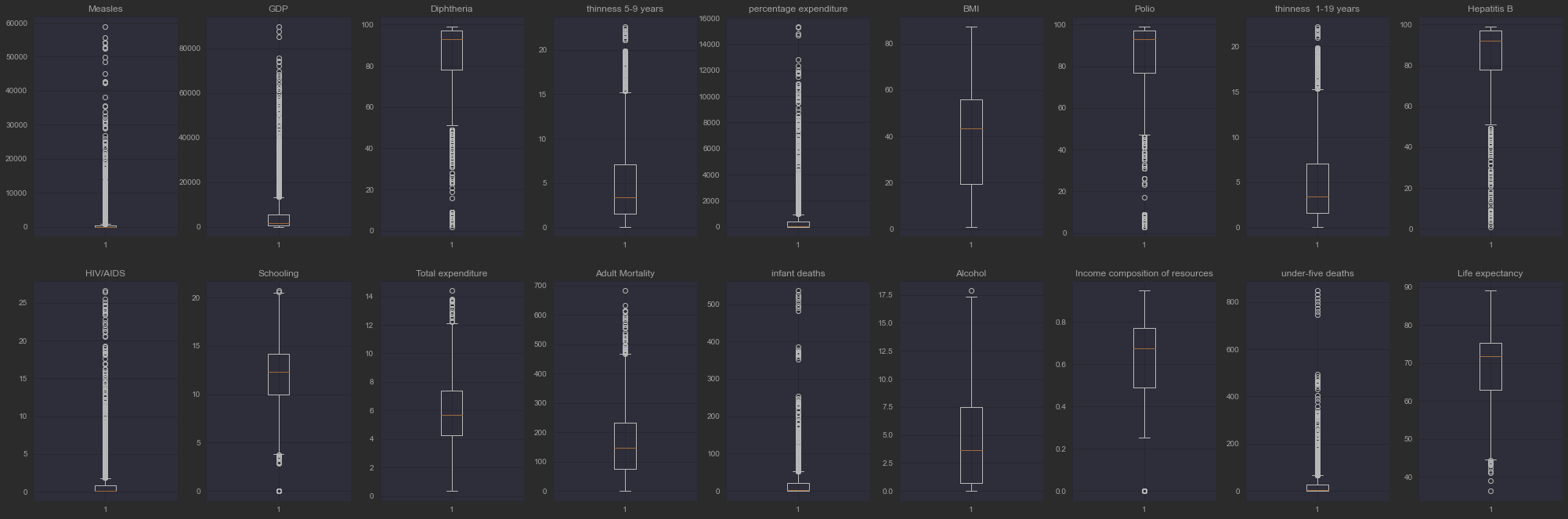


تعداد بسیاری از ستون ها دارای عناصر تهی هستند. برای حل این مساله از 2 روش استفاده میکنیم.

چون هر کشور دارای 16 ردیف میباشد، میتوان فواصل خالی بین مقادیر را با درونیابی پر کرد. ابتدا جداول جدا بر اساس کشور آن میسازیم، اگر در یک ستون تعداد مقادیر موجود بیشتر از 20% باشد تلاش برای درونیابی درجه 2 میکنیم.

عمل بالا برای پر کردن تعداد کمی از مقادریر کارآمد هست، به همین دلیل از Imputation استفاده میکنیم که به معنای ایجاد مدل برای تخمین آن ستون هست. به این مدل ها نیز ستون هایکه در بخش قبلی کنار گذاشته شد را نمیدهیم.

چون مدلی باید برای اینکار آموزش داده شود، ابتدا تبدیل هایی باید انجام شود تا به مدل توان تخمین بهتری داده شود. در پی این هدف، با پلات کردن داده مشخص میشود که تعداد زیادی از ستون، دارای outlier های بسیار دور هستند که در شکل زیر برای میتوان دید.

  
به دلیل وجود تعداد قابل توجه ای از این outlier ها در هر ستون، این ردیف را نمیتوان حذف کرد. بهتر هست یک نوع سقف نرم برای مقادیر گذاشت. یک روش آن این هست که مقادیر را با percentile 3% نرمال و سپس مقادیر بیشتر از 1 (یا کمتر از 0) را به توان 0.6 رساند. نهایتا هر ردیفی که مقداری با zscore بالاتر از 3.5 دارد را حذف میکنیم که شامل 8.6% از داده باقی هست (کشور های پیشرفته و غیر پیشرفته جداگانه در نظر گرفته میشوند).

حال سرانجام با Imputation مقادری تهی را پر میکنیم. اما این عمل را برای ستون هدف انجام نمیدهیم که از data leakage جلوگیری شود. سپس مقادیر هدف تهی را حذف میکنیم که 0.28% از داده باقی هست.

## حذف نویز

حال تلاش میکنیم مقادیر نویز را کاهش دهیم. برای این کار، PCA را حساب میکنیم و ستون ها با مجموع مقدار منفرد 2% را حذف میکنیم که شامل 2 ستون میباشد. حال با معکوس PCA، مقادیر با نویز کمتر را بدست میآوریم.

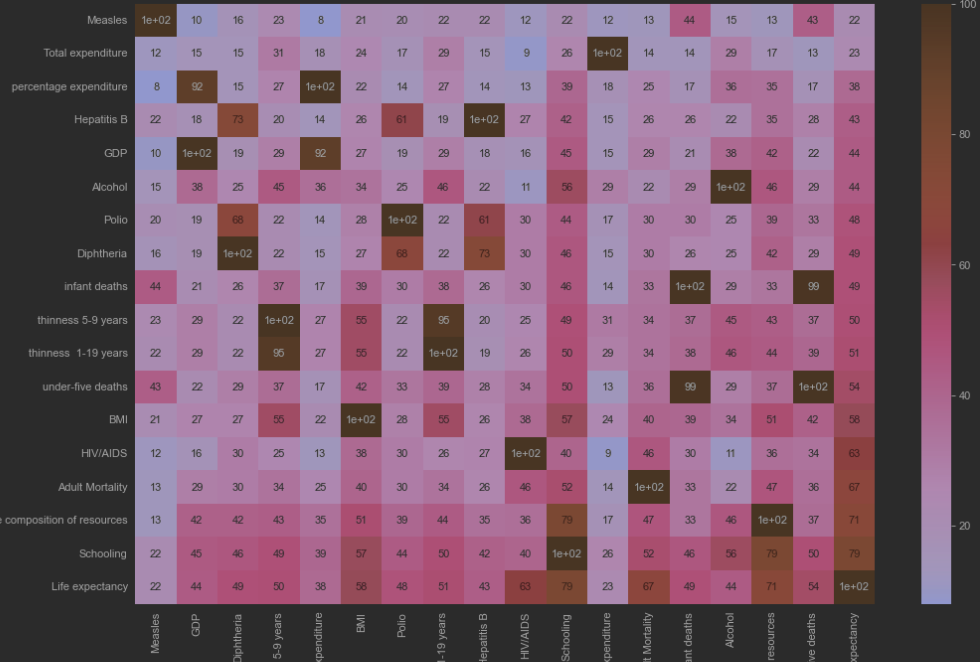
## انتخاب ستون

هدف نهایی ما استفاده از ساده ترین مدل ممکن همچون Ridge میباشد. این مدل ها به ویژگی های اضافی و تکراری حساس هستند. لذا برای بهبود عمکرد این مدل تلاش میکنیم بهترین ستون ها را برای تخمین مقدار هدف بیابیم.

در حال حاضر 17 ویژگی موجود هست که نسبتا کم هست. به همین دلیل با کمک از روش مانند Wrapper ستون ها مفید را نگه میداریم. در Wrapper، مدل مورد نظر بر بخشی از ویژگی ها اجرا میشود و با حذف و اضافه ویژگی ها، بهترین را میابد. برای انتخاب ستون هایی که به صورت کلی برای تخمین بهتر هستند، Wrapper را برای چند مدل متفاوت استفاده میکنیم و عناصری که بیشترین تعداد تکرار در خروجی Wrapper دارند را نگه میداریم.

در کنار آن از روش های filter که اطلاعات آماری را برای یافتن رابطه با ستون خروجی به صورت مستقل در نظر میگیرد و سرانجام 8 ویژگی از 17 ویژگی زیر را انتخاب میکنیم:

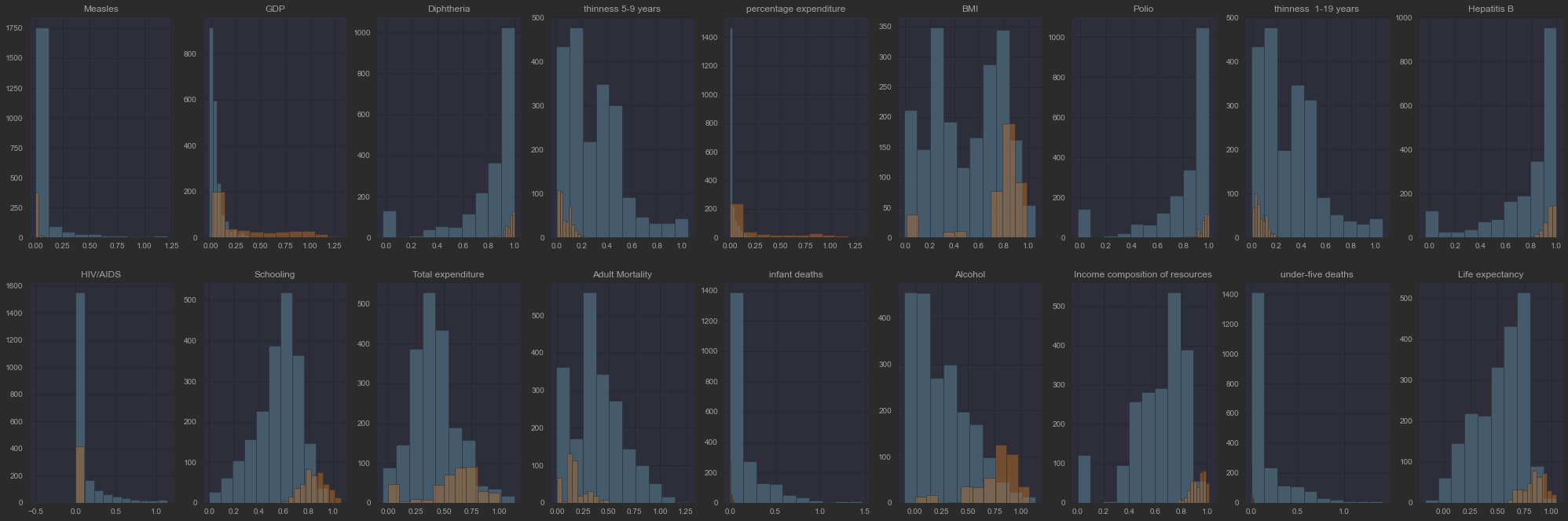
'GDP'و 'thinness 5-9 years'و 'Schooling'و 'HIV/AIDS'و 'Adult Mortality'و 'Polio'و 'Income composition of resources'و 'Alcohol'



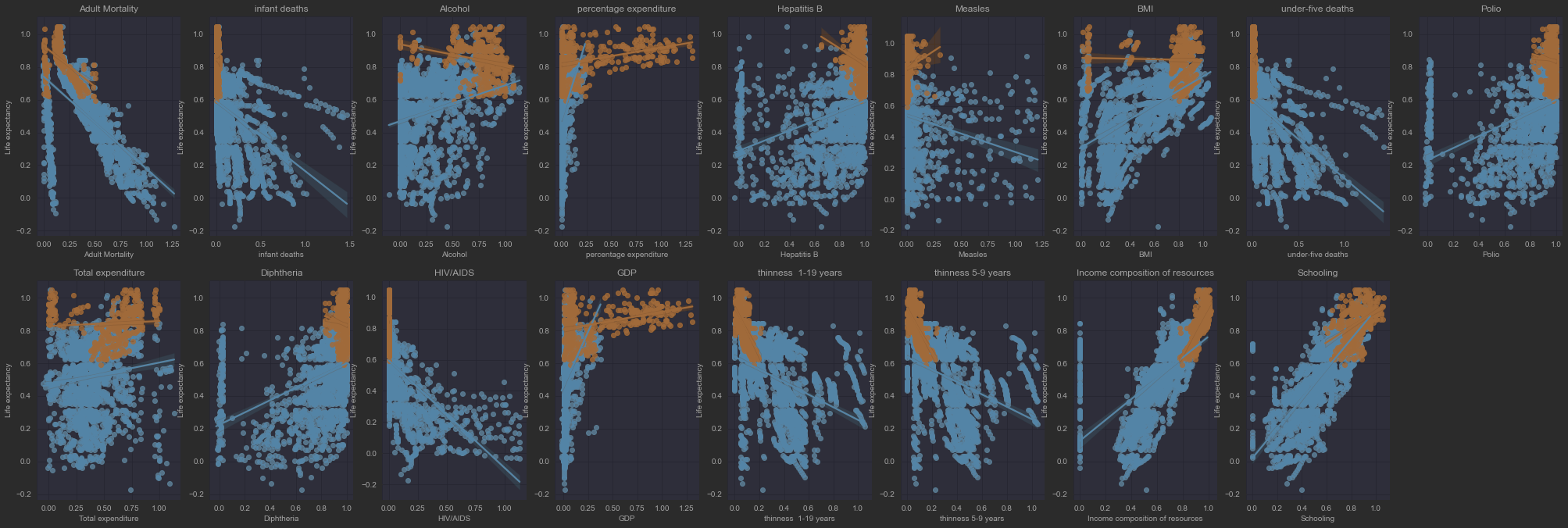
برخی از ویژگی ها با همبستگی بالا مانند bmi انتخاب نشده اند که نشانه از وجود یک تکرار در آن ویژگی هست

# فصل دوم: ایجاد مدل

## مشکل تفاوت بین کشور فقیر و پیشرفته



مشکل outlier کمتر هست اما هنوز به قدرت دیده میشود. مهمتر وجود تفاوت زیاد در بازه مقادیر بین کشور های فقیر و پیشرفته هست.



این مشکل در اشکال بالا به شدت دیده میشه. برخی از بازه ها فقط متشکل از کشور های پیشرفته هست. برای این مساله راه حل ساده ای نیافتیم.

## ایجاد مدل

یک چیز جالب این مساله بود که با انجام PCA و حتی نگه داشتن همه ویژگی ها، مدل عملکرد بدتر نشان میدهد.

این مدل را بر روی Ridge تمرین دادیم. بدون حذف نویز 0.0096 خطای MSE بود. با حذف نویز 0.0094 هست. اگر از همه ویژگی ها استفاده شود، این خطا به 0.0087 تقریبا همان هست.

برای اینکه یک تخمین برای بهترین عملکرد برای مدل بدست بیاوریم، ویژگی ها منتخب و سپس همه ویژگی ها را به خورد مدل DecisionTree میدهیم که به صورت خودکار ویژگی های اضافی را حذف میکند. تفاوت از 0.0037 به 0.0041 تغییر میکند.

اگر اهمیت ویژگی ها را از مدل DecisionTreeبگیریم، متوجه اهمیت زیادی برای نسبت تولید ثروت یک کشور از منابع اولیه دیده میشود که در لیست منتخب ها هم هست. این مساله با نظریه توصیف شده در کتاب Dictators notebook و کتاب های مشابه موافق است. دیگر ویژگی های منتخب نیز با اهمیت بالاتر از بقیه دیده میشود.

تفاوت بین خطا Ridge با DecisionTree 2.5 برابر هست.

**منابع و مراجع**

|  |  |
| --- | --- |
| [1] | Kaggle، عموما یک کلیات برای Feature Selection |



Amirkabir University of Technology  
(Tehran Polytechnic)

Department of Computer Science and Math

Data Mining Assignment #3

By

Seyed Hossein Mohammadi

Taught by

Dr. Ghaati

Sep of 2022