Projeto de Conclusão do Curso Intensivo de Data Science da escola Awari.

Tendo em vista o aprendizado e a prática dos conteúdos apresentados, e, até mesmo, expandindo para além dos assuntos vistos em aula.

Realizado por Haisa Vargas Echeli.

Professora: Sara Malvar.

Mentor: João Henrique Dib Netto

Machine Learning aplicada em dados do IDH mundial

Como saber se um país é desenvolvido? Uma das melhores métricas disponíveis a nível mundial é o IDH, que, significa "Índice de desenvolvimento humano" cujo valor é calculado levando em consideração diversos fatores e estatísticas de um determinado país, e, tem como valor máximo um; de forma que quanto mais próximo de um, mais desenvolvido o País é. Levando em conta esse índice, cria-se um Ranking Mundial de países.

Geralmente os ditos "países ricos" estão no topo dessa lista. Ao observar esse ranking vemos que o Brasil se encontra na posição XX do ranking. No entanto, quais são os países que mais se assemelham matematicamente com o Brasil? É suficiente dividir os 189 países que fazem parte desse ranking em apenas quatro categorias? A posição no ranking separa os países em: altíssimo desenvolvimento humano, alto de desenvolvimento humano, médio desenvolvimento humano e baixo desenvolvimento humano. O que aconteceria se pegássemos diferentes índices, valores, porcentagens e demais métricas usadas para determinar o IDH e separássemos em diferentes temas e calculássemos por meio de um algoritmo de *machine learning* não supervisionado para determinar quantos clusters podem ser divididos os países?

Diversas perguntas podem ser feitas e respondidas ao aplicar os ditos "algoritmos não supervisionados". O presente trabalho não tem o intuito de responder todas essas questões, muito menos, a pretensão de refutar uma métrica tão consolidada quanto o IDH, apenas propõe um novo olhar sobre o assunto. Todos os cálculos feitos para chegar no IDH é o mais verossímil possível; a questão primordial desse projeto é o aprendizado e a utilização de *machine learning* em uma área na qual ainda não se utiliza tanto essas ferramentas. Sendo que a maior parte da análise feita com base nos resultados dos algoritmos é focada especificamente no Brasil.

Esse projeto não é um artigo científico propriamente dito, mas os cálculos são replicáveis, e todos os códigos estão abertos e disponíveis por meio de um repositório do GitHub.

A primeira questão a ser resolvida em um projeto de Data Science é a aquisição de uma base de dados, a fonte que serviu de base para a delimitação da base de dados usada ao longo de todo o projeto foi o *"Humano Development Report 2020. The next frontier: Human development and the Anthropocene"* produzido pela UNDP, ou seja, realizado e divulgado pelo "Programa de Desenvolvimento das Nações Unidas".

Utilizando os dados estatísticos nos quais o Relatório divulgado pela ONU foi baseado, criou-se uma base de dados reduzida montada utilizando dados e criando sete diferentes

categorias/temas. Os temas delimitados foram: Economia, Desigualdade, Gênero, População, *Expenditure* , Saúde e Meio Ambiente. Os dados selecionados para compor cada uma das categorias encontram-se dispostos na tabela abaixo:

TABELA###

'RHDI':'Ranking Human Development Index, 2019',

'HDI':'Human Development Index Value, 2019',

'LE':'Life expectancy at birth, years, 2019',

'EYS':'Expected years of schooling, years, 2019',

'MYS':'Mean years of schooling, years, 2019',

'GNI':'Gross national income per capita 2019 (2017 PPP $)',

'GDIV':'Gender Development Index Value, 2019',

'GIIV':'Gender Inequality Index Value, 2019',

'IHDI':'Gender Inequality Index Value, 2019',

'ILE':'Inequality in life expectancy, %, 2015-2020',

'IE':'Inequality in education, %, 2019',

'II':'Inequality in income, %, 2019',

'SSP':'Share of seats in parliament, % held by women, 2019',

'TP':'Total population, millions, 2019',

'TUP':'Total urban population, %, 2019',

'MA':'Median age, years, 2020',

'CHE':'Current health expenditure, % of GDP',

'GEE':'Government expenditure on education, % of GDP, 2013-2018',

'PPP':'Physicians per 10000 people, 2010-2018',

'HB':'Hospital beds per 10000 people, 2010-2019',

'VE':'Vulnerable employment, % of total employment, 2019',

'RPAE':'Rural population with access to electricity, %, 2018',

'WAFI':'Women with account at financial institution or with mobile money-service provider, % of female population ages 15 and older, 2017',

'CDEP':'Carbon dioxide emissions production per capita, tonnes, 2018',

'CDE':'Carbon dioxide emissions per unit of GDP, kg per 2010 US$ of GDP',

'FA':'Forest area, % of total land area, 2016',

'FAC':'Forest area change, %, 1990/2016',

'DMC':'Domestic material consumption per capita, tonnes, 2017',

'RLI':'Red List Index Value, 2019',

'SLF':'Skilled labour force, % of labour force, 2010-2019',

'RDE':'Research and development expenditure, % of GDP, 2014-2018'

SEPARAR EM Economia, Desigualdade, Gênero, População, *Expenditure* , Saúde e Meio Ambiente.

TABELA

A linguagem usada foi o *Phyton 3*, e, a IDE (interface) que facilitou todo o processo foi o "Jupyter Notebook", juntamente com o pacote Anaconda, possibilitando realizar todos os passos e aplicar todos os pacotes e algoritmos em máquina local.

Com a base de dados devidamente preparada, o próximo passo de todo projeto de Data Science é a manipulação e limpeza dos dados. O primeiro problema encontrado foi a ausência de valores em diversos países. Geralmente uma opção viável é retirar as "amostras" (nesse caso, países) com dados faltantes, no entanto, essa opção foi rapidamente descarta para evitar excluir qualquer país dos 189 presentes no Ranking. Para superar esse desafio, sem retirar nenhum país, foi feito o preenchimento dos dados faltantes. A maneira escolhida para preencher tais valores foi a utilização da média calculada para cada continente; dado o qual foi adicionado para cada um dos países. Criou-se assim uma tabela com cada país e seu respectivo continente, para unir a Base de Dados, e, por meio de agrupamento (*groupby)* e de uma função criada (utilizando o *for*), calcular todas as médias por continente e preencher corretamente de forma prática os valores faltantes.

Apesar de parecer "pouca coisa", na prática essa foi uma etapa complicada que demandou muita tentativa e erro. A princípio, a primeira tentativa, que pode ser visualizado no arquivo "First Draft", foi muito extensa e demandou demasiado tempo. Já o segundo código ficou muito mais simples e eficiente. Tendo em vista que o objetivo principal desse projeto é aprender na prática a realizar um projeto inteiro de Data Science, do início ao fim, essa foi uma etapa extremamente proveitosa.

Com a base de dados completa, sem nenhum valor faltante, a etapa seguinte foi garantir que cada índice, porcentagem ou valor, dentro de cada tema tivesse o mesmo peso nos algoritmos de *machine Learning*, foi indispensável passar toda a base de dados por um "Scaler" para garantir que todo dado variasse de zero à um, o *Scaler* escolhido foi o *MinMaxScaler*; inclusive nas colunas em que o valor era de no máximo um, pois, dessa forma o menor valor é calculado é zero e o maior é um, certificando a correta padronização dos dados.

A etapa seguinte foi a aplicação de dois algoritmos de *machine learning* não supervisionados. O primeiro foi a Clusterização realizada para cada tema, utilizou-se o

algoritmo chamado *K-means* onde é necessário escolher o número de Clusters. Nesse momento, ficou claro que sería preciso empregar técnicas para determinar o número clusters ideais para cada tema, utilizou-se de duas técnicas importantes: *Elbow Method* ("técnica do cotovelo") e *Silhouette Method* ("técnica da silhueta"). Levando em consideração ambas as técnicas, e, os objetivos do projeto, determinou-se o número de Clusters para cada tema.

Com esse número definido fitou-se o algoritmo *K-means* em cada uma das sete categorias. A visualizações de cada clusterização só foi possível com a redução das dimensões dos dados, porque em todas as categorias encontravam-se mais de dois parâmetros, pelo método do PCA (onde os diversos índices ou valores são transformados em apenas duas dimensões); permitindo assim a plotagem em um gráfico de duas dimensões, no qual, as cores correspondem a cada um dos Clusters.

Esses resultados serviram de base para a análise de quais países encontram-se sempre no mesmo Cluster do Brasil, independente do tema.

Em seguida, percebeu-se que seria interessante descobrir os cinco vizinhos mais próximos do Brasil em casa categoria (na verdade foram os quatro mais próximos, pois o mais próximo matematicamente é o próprio Brasil). Dessa forma foi selecionado mais um algoritmo não supervisionado: o *KNEIGHBORS*, utilizando cinco como parâmetro, como exposto à cima.