Expectativa de vida - Mineração de Dados - Regressões

Alexis Mariz, Leandro Diniz, Samuel Lipovetsky
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil

I. Contexto escolhido e sua relevância

A expectativa de vida é um indicador crucial para avaliar a qualidade de vida e o bem-estar da população. Ela reflete não apenas as condições de saúde, mas também fatores como acesso a serviços básicos, infraestrutura, segurança e condições socioeconômicas. Compreender os padrões e fatores que influenciam a expectativa de vida é essencial para desenvolver políticas públicas eficazes, reduzir desigualdades e promover um envelhecimento saudável e sustentável para todos.

A. Objetivo e critérios de sucesso associados ao(s) problema(s) do contexto escolhido

O objetivo principal deste trabalho é aplicar técnicas de regressão linear para prever a expectativa de vida das pessoas, além de identificar e analisar os fatores que exercem maior influência sobre sua longevidade.

B. Recursos disponíveis, requisitos, suposições, restrições, riscos e contingências

- **Recursos disponíveis:** Dados públicos sobre expectativa de vida e fatores de saúde, coletados de 193 países, incluindo informações relacionadas ao status de saúde, indicadores econômicos e outros fatores determinantes. Esses dados foram obtidos a partir do repositório da Organização Mundial da Saúde (OMS) e do site das Nações Unidas, com foco nos fatores críticos que mais impactam a longevidade.
- **Requisitos:** Tratamento e limpeza dos dados, garantindo que as informações estejam consistentes para a análise desses dados.
- **Suposições:** A qualidade e completude dos dados são adequadas para a análise pretendida, mesmo com lacunas em alguns atributos.
- **Restrições:** Muitas colunas possuem dados ausentes, especialmente nos campos relacionados a doenças, como ”Hepatitis B”, o que limita o uso desses atributos em algumas análises.
- **Riscos e contingências:** A falta de dados completos em certos atributos pode dificultar análises mais detalhadas, exigindo métodos para tratar dados ausentes ou limitações na interpretação de resultados.

C. Objetivos da mineração de dados, detalhamento da tarefa e critérios de sucesso

O objetivo principal deste trabalho é aplicar técnicas de regressão linear para prever a expectativa de vida das pessoas, identificando os fatores que mais impactam sua longevidade. A análise busca determinar quais variáveis, como condições socioeconômicas, acesso à saúde e educação, entre outras, têm maior influência na expectativa de vida. O critério de sucesso será a capacidade do modelo de previsão em estimar com precisão a expectativa de vida, além de identificar os fatores mais determinantes para a longevidade.

D. Descrição do projeto

Os dados utilizados foram disponibilizados no *Kaggle* do desafio *Life Expectancy (WHO)* de 2018. Para criar esse *dataset* foram utilizados dados públicos da *Global Health Observatory* na categoria de *World Health Organization*. Cada linha desse banco de dados contem informações relacionadas a saúde da população de um pais em um determinado ano. A tabela abaixo contem todas as vinte duas colunas que iremos utilizar nesse trabalho:

- **Life expectancy**: expectativa de vida média da população, variável-alvo para as análises de regressão.
- **Country, Year, Status**: dados relacionados ao país, como seu nome, ano que os dados foram recuperados e o nível socioeconômico do país.
- **Adult Mortality, infant deaths, under-five deaths**: índices de mortalidade da população divido por certas faixas etárias.
- **Alcohol**: consumo médio de álcool per capita.
- **percentage expenditure**: porcentagem do gasto público em saúde em relação ao PIB.
- **BMI**: índice de massa corporal médio da população.
- **Hepatitis B, Polio, Measles, Diphtheria, HIV/AIDS**: dados relacionados a doenças e infecções que atingem a população, como cobertura da vacinação, pessoas afetadas e quantidades de fatalidades que elas causaram.
- **Total expenditure**: gasto total em saúde como percentual do PIB.
- **GDP**: Produto Interno Bruto per capita.
- **Population**: tamanho da população total, usado para calcular indicadores per capita.
- **thinness 5-9 years, thinness 1-19 years**: prevalência de magreza em crianças e adolescentes de 1 a 19 w 5-9 anos.

- **Income composition of resources**: Composição de renda e acesso a recursos.
- **Schooling**: média de anos de escolaridade da população.

E. Descrição e exploração dos dados

O *dataset* utilizado possui 2938 linhas e 22 colunas, e a maior parte dos dados é numérica, com exceção de algumas colunas, como "*Status*" e "*Country*". No entanto, há uma quantidade significativa de valores ausentes em campos importantes, como "*Hepatitis B*" então colunas como essas não serão utilizadas. Os atributos serem numéricos vai facilitar o uso das técnicas de regressão a única coluna que precisou ser modificada foi a coluna status, que teve seus valores únicos "*Developing*" e "*Developed*" alterados para 0 e 1 respectivamente.

F. Visualização exploratória

Para visualizar a evolução da expectativa de vida média ao longo dos 15 anos, de 2000 a 2015, foi elaborado um *line plot*. Esse gráfico mostra um aumento progressivo na expectativa de vida durante esse período, o que pode ser atribuído a diversos fatores, como avanços na área da saúde, melhorias nas condições socioeconômicas e maior acesso a serviços básicos em várias regiões.

Fig. 1: Expectativa de vida média por ano

Foi criada uma correlação de *Pearson* entre todas as colunas do *dataset* para analisar de modo superficial como esses atributos estão relacionados entre si. É possível observar que a expectativa de vida esta fortemente relacionada ao nível de escolaridade **(r = 0.83)** e a composição de renda da população **(r = 0.91)**.

Por outro lado, a variável relacionada aos gastos com saude *("total expenditure")* apresenta uma correlação negativa perfeita **(r = -1)** com as demais variáveis, o que é atípico e pode indicar algum problema nos dados ou na escala utilizada. Já o GDP (Produto Interno Bruto) tem uma correlação positiva moderada **(r = 0.45)** com a expectativa de vida, indicando que, embora a riqueza de um país seja relevante, ela não é o único fator determinante para a longevidade

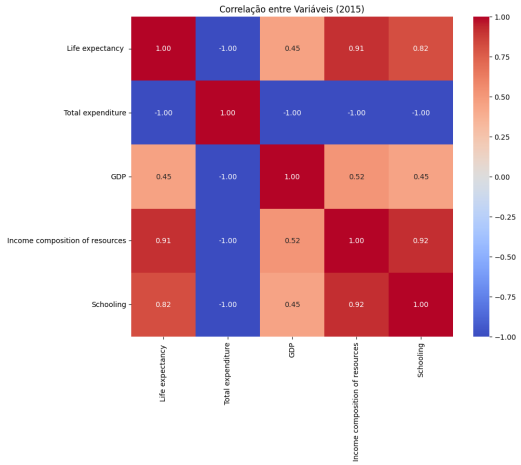

Fig. 2: Correlação de Pearson em 2015

TABLE I: Países com maior expectativa de vida e seu GDP

País	**Expectativa de vida**	**GDP**
Japan	82.53	24892.54
Sweden	82.51	29334.99
Iceland	82.44	30159.50
Switzerland	82.33	57362.87
France	82.21	26465.55

Ao analisar o top 5 dos países com maior expectativa de vida, observa-se que todos são países desenvolvidos, com um Produto Interno Bruto (PIB) significativamente superior à média de 7.483,15, que corresponde ao valor médio registrado ao longo desses anos.