Cálculo Numérico – Prova 2

Projeto D – Avaliação de índices de Felicidade

Contexto: O índice de felicidade é uma métrica que vai além da simples satisfação pessoal, proporcionando insights valiosos para políticas públicas, estratégias empresariais e até mesmo para a alocação de recursos em áreas como saúde e infraestrutura. Pesquisas indicam que níveis mais altos de felicidade e bem-estar nas populações estão correlacionados com menor incidência de doenças crônicas e custos mais baixos com saúde pública. Isso ocorre porque fatores como o tempo de exposição ao sol, a prática de atividades ao ar livre, e o nível de poluição impactam diretamente na saúde física e mental. Em um cenário urbano, o acompanhamento do índice de felicidade oferece uma perspectiva prática e proativa para os governos e empresas, permitindo a antecipação de demandas por recursos médicos, otimização de programas de saúde e implementação de iniciativas que favoreçam o bem-estar da população.

A partir da análise de dados de diversas cidades globais, um estudo buscou identificar a relação entre diferentes variáveis ambientais, econômicas e de estilo de vida e os níveis de felicidade. Dados como horas de luz solar, índice de poluição e expectativa de vida, por exemplo, são considerados importantes para entender como elementos externos podem afetar a felicidade e bem-estar dos cidadãos. Além disso, variáveis como o custo de vida, o nível de obesidade e o acesso a atividades ao ar livre também são avaliadas para mapear como esses fatores influenciam o bem-estar geral.

Neste estudo, você será responsável pela análise dos dados urbanos relacionados ao índice de felicidade, com o objetivo de compreender como essas diferentes variáveis afetam a saúde e o bem-estar em diferentes cidades. Ao entender esses aspectos, espera-se não apenas identificar as características das cidades mais felizes, mas também fornece recomendações práticas para melhorar a qualidade de vida nas cidades menos favorecidas.

Dados: Um conjunto de dados foi obtido de um estudo global que avaliou o índice de felicidade em várias cidades do planeta, considerando diferentes aspectos do ambiente urbano e do estilo de vida dos habitantes.

Cada coluna na tabela representa uma medida específica que pode influenciar o bem-estar e a qualidade de vida. Os dados originais estão no arquivo "healthy_lifestyle_city_2021_cleaned.csv," e incluem um total de 11 variáveis relacionadas ao ambiente, ao custo de vida e ao estilo de vida dos habitantes urbanos. Quando os dados são carregados na linha 11 do arquivo p2-d.m, as colunas seguem a ordem abaixo:

1 – City – Cidade

2 – Sunshine hours – Horas de luz solar

3 – Cost bottle of water – Custo, em libras, de uma garrafa de água

4 – Obesity Levels – Níveis de obesidade em % da população

5 – Life expectancy(years) – Expectativa de vida em anos

6 – Pollution(Index score) – Índice de poluição, que é uma notas adimensional. Maior poluição apresenta maior índice

7 – Annual avg. hours worked – Média de horas trabalhadas por ano

8 - Happiness levels(Country) – Níveis de felicidade. Os dados originais são de 0 a 10, mas fiz uma métrica onde índices acima de 7 são considerados países mais felizes (marcado como "1"), e índice menores que 7 são menos felizes (marcados como "0").

9 - Outdoor activities – Quantidade de atividades ao ar livre disponíveis

10 - No. of take out places - Quantidade de Locais de Comida para Viagem

11 - **Monthly gym membership** – Custo médio da mensalidade de academia

Dúvidas sobre os dados podem ser consultadas neste link.

Questionamentos: seu trabalho deve conter, obrigatoriamente, as análises que seguem, mas outras avaliações que o grupo julgar pertinentes podem ser consideradas. Utilize o Octave para responder:

Análise 1: Seleção das variáveis

1.1. Com base nos dados fornecidos, selecione duas variáveis que possam estar diretamente relacionadas ao nível de felicidade urbana para cidades mais felizes (marcadas como "1"). Por exemplo, expectativa de vida, índice de poluição ou horas de luz solar.

Justifique suas escolhas com uma análise visual utilizando gráficos que permitam observar a relação com a variável "Nível de Felicidade".

Sugestão: os pares de plots ao final do documento podem estar mascarando as correlações. Talvez normalizar os dados e fazer um novo plot pode ajudar.

1.2. Repita a análise em (1.1) para selecionar dois dados que representam melhor as cidades menos felizes.

1.3. Aplique análises numéricas para sustentar sua decisão sobre qual variável tem maior influência na previsão do nível de felicidade em (1.1) e (1.2). Justifique seu raciocínio com base nos resultados das análises numéricas apresentadas em aula.

Análise 2: Análise de regressão exponencial

2.1. Considere agora a relação entre cada variável e o índice de felicidade nas cidades com os menores níveis de felicidade, visando identificar padrões que indiquem uma relação não linear.

Usando os gráficos de dispersão, verifique visualmente se alguma variável apresenta uma tendência a se ajustar melhor a uma exponencial do tipo $y = ae^{bx}$. Explique a escolha da variável com base na análise visual.

2.2. Para a variável identificada em (2.1), aplique uma regressão exponencial, ajustando os dados com as duas abordagens. Apresente os gráficos da regressão obtida somente sobre os pontos de dados utilizados.

2.3. Mostre, numericamente, qual polinômio é melhor para aquele conjunto de dados. Justifique.

Análise 3: Predição

3.1. Construa dois modelos de regressão linear simples usando as duas variáveis identificadas na Análise 1. Por exemplo, um modelo pode ser da forma $y_1 = a_{0,1} + a_{1,1}x_1$ e outro $y_2 = a_{0,2} + a_{1,2}x_2$ onde x_1 e x_2 foram obtidos na Análise 1 para a previsão de cidades mais felizes. Calcule as métricas de qualidade do ajuste, como S_r, r^2 e $s_{y/x}$ para cada modelo. Qual modelo apresentou melhor ajuste? Justifique.

3.2. Implemente um terceiro modelo de regressão linear múltipla do tipo $y_3 = a_{0,3} + a_{1,3}x_1 + a_{2,3}x_2$ que combina as duas métricas utilizadas em (3.1).

Isso significa que você deve encontrar os parâmetros $a_{0,3}$, $a_{1,3}$ e $a_{2,3}$ de tal maneira que a regressão $y = a_1x_1 + a_2x_2 + a_0$ seja o melhor possível. Calcule qual a soma dos resíduos S_r, r^2 e $s_{y/x}$ para o modelo y_3. Compare-o com y_1 e y_2. Qual é o melhor? Justifique.

3.3. Com base nas métricas obtidas nos modelos de regressão linear simples e múltipla (Análise 3.1 e 3.2), avalie se o modelo ajustado permite uma previsão confiável para o índice de felicidade. Justifique sua resposta com base nos valores das métricas de ajuste.

Atenção: essa não é uma prova de aprendizado de máquina, mas de cálculo numérico. Logo, não espero uma análise utilizando métricas rebuscadas, discussões sobre modelos mais robustos ou que seu modelo performe bem. Quero que seu grupo foque na utilização do modelo construído, usando a ferramenta Octave e a interpretação dos resultados obtidos segundo a teoria dada em aula.

Códigos: o arquivo que deve resolver seu projeto é dado “p2_d.m”. Esse código já carrega os dados e faz o plot de dispersão (scatter) das métricas duas a duas. Você pode criar outros scripts .m caso precise, mas eu só devo precisar rodar “p2_d.m” para verificar os entregáveis do seu projeto.

Funções proibidas:

- polytfit, linsolve, regress, interp1, interp2, interp3, interpn, spline, fitlm, compact, fitrlinear, mvregress, mvregresslike, plsregress.
- Também é proibido resolver sistemas lineares com o método da inversa (função inv()) ou pelo operador barra invertida (\). Caso você precise, utilize a função lu() para obter a decomposição e implemente uma pequena função que resolve o sistema dado que você possui L e U.

ANEXO

Formato dos dados:

	City	Rank	Sunshine hours	Cost bottle of water	...	Happiness levels(Country)	Outdoor activities	No. of take out places	Monthly gym membership
0	Amsterdam	1	1858.0	1.92	...	1	422	1048	34.90
1	Sydney	2	2636.0	1.48	...	1	406	1103	41.66
2	Vienna	3	1884.0	1.94	...	1	132	1008	25.74
3	Stockholm	4	1821.0	1.72	...	1	129	598	37.31
4	Copenhagen	5	1630.0	2.19	...	1	154	523	32.53

Figura 1. Cinco primeiras linhas dos dados em healthy_lifestyle_city_2021_cleaned.csv

Pairplots dos dados: