GLOBAL SOLUTION - DATA SCIENCE

Tema: O Futuro do Trabalho em Dados e Inteligência Artificial
Base: https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries/data
Entrega: Link do notebook no **Google Colab** (um por grupo)
Formação: Grupos de até 3 alunos

Contexto
A área de Ciência de Dados é uma das que mais crescem no mundo, impulsionada pela Inteligência Artificial e pela transformação digital.
Compreender como variam os salários, cargos e tipos de contrato é essencial para planejar a carreira no futuro do trabalho.
Neste desafio, o grupo aplicará os conceitos de amostragem e estatística descritiva estudados ao longo do semestre, utilizando dados reais sobre profissionais de Data Science ao redor do mundo.

Instruções Gerais

- Todo o trabalho deve ser executado no Google Colab.
- As questões dissertativas devem ser respondidas de forma objetiva em células Markdown (texto explicativo logo abaixo do código).
- Cada questão deve conter:
    - O código Python executado;
    - O resultado exibido (tabela ou gráfico);
    - E a resposta dissertativa, clara e concisa.
- Envie o link do notebook Colab completo, com os nomes e RMs de todos os integrantes.

Etapas do Desafio

Dataset

Colunas principais:

- work_year – Ano de referência do salário.
- experience_level – Nível de experiência (EN, MI, SE, EX).
- employment_type – Tipo de contrato (FT, PT, CT, FL).
- job_title – Cargo.
- salary_in_usd – Salário anual em dólares.
- employee_residence – País de residência do profissional.
- remote_ratio – Percentual de trabalho remoto (0, 50 ou 100).
- company_location – Localização da empresa.
- company_size – Porte da empresa (S, M, L).

Com base nas colunas acima, o grupo deverá compreender padrões salariais por região, cargo e perfil, criar modelos de regressão linear e logística para prever salários e probabilidades de alto rendimento e interpretar resultados com métricas estatísticas claras e recomendações estratégicas.

Parte 1 – Análise Descritiva e Exploratória (EDA)

Objetivo:
Identificar o comportamento geral dos salários e fatores relacionados ao perfil profissional.

Métricas obrigatórias:
Média, mediana e desvio-padrão de salary_in_usd.
Distribuição percentual de experience_level, employment_type e company_size.
Top 5 cargos (job_title) com maior média salarial.
Correlação entre remote_ratio, company_size e salary_in_usd.
Boxplot de salary_in_usd por experience_level e company_location.

Perguntas executivas:

1. Qual é a média e o desvio-padrão do salary_in_usd para cada categoria de experience_level? Qual nível apresenta maior variabilidade salarial e o que isso indica sobre o mercado?
2. Qual tipo de contrato (employment_type) apresenta maior média salarial? Essa diferença se mantém entre portes de empresa (company_size)?
3. Compare a média de salary_in_usd por company_location. Quais países se destacam por maiores ou menores salários? Que fatores econômicos podem justificar essa diferença?
4. Analise a correlação entre remote_ratio e salary_in_usd. Há indícios de que o trabalho remoto impacte positivamente ou negativamente o salário? Explique.
5. Quais cargos (job_title) aparecem entre os cinco mais bem remunerados? Eles correspondem a posições consolidadas ou emergentes?

Parte 2 – Modelagem Preditiva (Regressão Linear e Logística)

Objetivo:
Investigar os fatores que influenciam o salário e prever probabilidades de altos rendimentos com base em variáveis do dataset.

Métricas obrigatórias:
R^2, RMSE e MAE (para o modelo linear).
Acurácia, precisão, recall, F1-score, curva ROC e AUC (para o modelo logístico).
Análise dos coeficientes e interpretação dos sinais.
Odds ratio (e^β) para variáveis significativas da regressão logística.

Perguntas executivas:

6. Quais variáveis explicam melhor as diferenças salariais segundo a regressão linear? O modelo apresenta um R^2 satisfatório para explicar o comportamento de salary_in_usd?
7. Analise os erros RMSE e MAE obtidos. O modelo linear apresenta desvios grandes? Que tipo de melhoria poderia reduzir esses erros?
8. No modelo logístico, quais variáveis aumentam significativamente as chances de um profissional ter salary_in_usd acima da média? Interprete os resultados do odds ratio.
9. A curva ROC e o valor de AUC indicam que o modelo logístico tem boa capacidade preditiva? O modelo apresenta indícios de overfitting ou generalização adequada?
10. Se você fosse consultor de RH, quais conclusões práticas e recomendações estratégicas apresentaria à diretoria com base nos resultados combinados dos modelos linear e logístico?