Justificativa da Escolha do Banco de Dados

A escolha do banco de dados utilizado neste projeto foi fundamentada em sua relevância prática, clareza na estrutura e potencial para aplicação de algoritmos de aprendizado de máquina supervisionado. O conjunto de dados apresenta características ideais para experimentação com modelos de classificação, como Decision Tree, Random Forest e XGBoost, pois possui variáveis preditoras bem definidas e uma variável alvo binária, o que torna o problema adequado para técnicas de classificação.

Além disso, trata-se de um banco de dados com informações limpas, sem excesso de ruído ou inconsistências graves, o que favorece o foco no desenvolvimento e avaliação dos modelos de machine learning, sem a necessidade de um tratamento de dados excessivamente complexo. Isso foi essencial para que o projeto pudesse ser realizado com foco no aprendizado dos algoritmos e na interpretação dos resultados.

Outro fator decisivo foi a aplicabilidade do conjunto de dados no mundo real. Os dados refletem cenários reais de negócios, como churn de clientes, aprovações de crédito, diagnósticos médicos, entre outros. Esses cenários são amplamente estudados na área de Ciência de Dados por sua importância em processos decisórios empresariais e impacto direto nos resultados das organizações. Trabalhar com esse tipo de base permite compreender como os algoritmos de aprendizado supervisionado podem gerar valor real para empresas e instituições.

Por fim, a base de dados escolhida apresenta um equilíbrio interessante entre número de registros e número de atributos, o que permite análises estatísticas relevantes e treinar modelos com eficiência computacional, sem a necessidade de grandes recursos de hardware. Isso foi fundamental para a execução do projeto em ambiente local, como o Jupyter Notebook, com desempenho satisfatório.

Portanto, a escolha deste banco de dados não foi apenas estratégica para o domínio técnico, mas também didática, contribuindo significativamente para a compreensão dos processos de modelagem preditiva e validação de desempenho de classificadores.