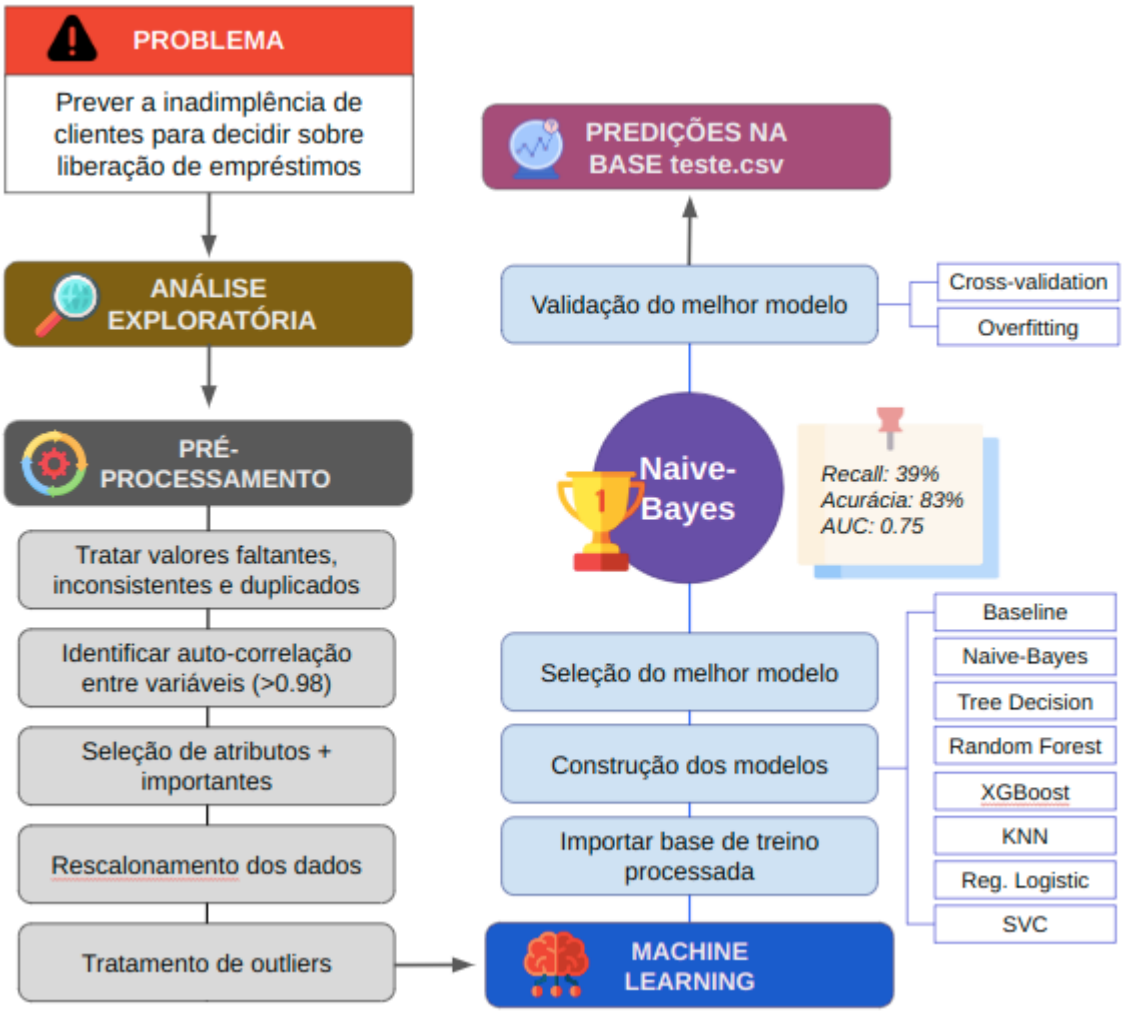
DOCUMENTAÇÃO - CASE DATARISK

Objetivo: Realizar uma análise preditiva de inadimplência dos clientes para tomar decisões sobre liberação de empréstimos.

Link do repositório na web: https://github.com/nmonalisa/case_datarisk

Estrutura de pastas:
/dados: contém os conjuntos de dados utilizados no treinamento dos modelos
/modelos.
/notebooks: contém o código com: (1) análise exploratória de dados, (2) processamento dos dados, e (3) construção dos modelos de machine learning e previsão para a base teste.csv fornecida pela equipe.
/predict: contém o arquivo teste.csv contendo a coluna "inadimplente", referente à previsão do modelo.

Etapas de realização do trabalho:



Conclusão da análise:

Após tentar diferentes abordagens e modelagens com os dados, o modelo campeão foi o Naive-Bayes, com um desempenho médio de 38% para a métrica de interesse (revocação). Isso quer dizer que de todos os clientes que são inadimplentes, o modelo será capaz de prever corretamente 38% deles, subsidiando a decisão de negação de empŕestimo.

Se considerarmos outras métricas que incluem acertos do modelo para clientes adimplentes (e não só inadimplentes), o desempenho sobe para 83% total de acertos (acurácia).

O modelo campeão também foi testado em relação à possibilidade de sobre-ajuste aos dados de treino e parece ser capaz de lidar com novos dados desconhecidos em um hipotético ambiente de produção. Ou seja, não há evidências de *overfitting*.

Considero que eu tenha obtido um bom modelo, capaz de ajudar o negócio a evitar prejuízos com concessão de empréstimos a clientes inadimplentes. Talvez algumas melhorias poderiam ser obtidas com algoritmos mais complexos, como redes neurais. Porém, apesar de mais eficientes, eu considero esses modelos excessivamente complexos e de difícil explicabilidade. Portanto, preferi modelar o meu fenômeno de forma mais simples, explorando todas as etapas de um problema típico de machine learning, e obtendo resultados que são mais transparentes e facilmente interpretáveis.

O modelo campeão foi usado para fazer as predições no arquivo de teste fornecido pela equipe Datarisk. Elas se encontram em "dados/predict/teste.csv".