ATITUS
EDUCAÇÃO

Projeto Nanodegree

Nomes: Bruno Pasquetti, Gabriel Brocco, Pedro H. De Bortoli e Rafael Klein

ML é a abordagem certa? Por quê? E quanto à IA generativa?

Sim, Machine Learning é a abordagem certa porque o objetivo do projeto é prever a evasão de estudantes com base em dados em um sistema de curso síncrono.

A IA generativa pode ser útil como apoio, ajudando na explicação dos resultados, na elaboração de textos e na apresentação, mas não substitui o papel central do ML na construção do modelo preditivo.

E se fosse o contrário, quais seriam os motivos?

A IA generativa só seria a abordagem principal se o foco do projeto fosse a geração de conteúdos personalizados, como mensagens motivacionais ou relatórios em linguagem natural, ou se não houvesse rótulos nos dados e fosse necessário identificar padrões de forma não supervisionada. Como o objetivo do projeto é prever a evasão com base em dados históricos rotulados, a IA generativa não é a abordagem adequada para o problema central.

Qual é o tipo de problema? Poderia ser outro? Quais heurísticas?

O problema é de classificação (evasão / não evasão), mas poderia ser tratado como regressão (probabilidade de evasão). Se não houvesse o rótulo de evasão, seria um problema de agrupamento (clustering), para identificar perfis de alunos com comportamentos semelhantes. Optamos pela classificação devido à sua aplicabilidade imediata na tomada de decisões educacionais.

As heurísticas mais comuns incluem:

- **Correlação entre variáveis** para seleção de atributos relevantes.
- **Balanceamento de classes** caso haja desbalanceamento entre evasores e não evasores.
- **Validação cruzada** para avaliar o desempenho dos modelos.

Há correlação entre o atributo-alvo e os outros atributos?

Sim, há correlação entre o atributo-alvo (evasão) e outros atributos, especialmente aqueles relacionados à participação do aluno na plataforma, como número de acessos, envio de tarefas e participação em atividades.

Quais os dados disponíveis e quais dados poderiam agregar?

Dados disponíveis: informações descritivas, temporais, contagens de interações e escores de desempenho dos alunos na plataforma.

Dados que poderiam agregar: idade, histórico escolar, nível socioeconômico, perfil comportamental e feedbacks qualitativos, pois ajudariam a compreender fatores externos que influenciam na evasão.

Avalie as características dos dados (abundante, consistente, confiável, disponível, correto, representativo).

- **Abundância:** Baixa, o conjunto possui apenas 500 registros (400 para treino e 100 para teste), o que é considerado pequeno para problemas de classificação. Isso pode limitar o desempenho e a generalização de modelos mais complexos, como redes neurais profundas.

- **Consistência:** Moderada, embora os dados sigam um padrão de nomenclatura, há registros com valores ausentes e formatos inconsistentes em algumas variáveis, o que exige tratamento antes da modelagem.

- **Confiabilidade:** Moderada, os dados são reais, mas foram descaracterizados (reamostragem e substituição de nomes), o que reduz um pouco a fidelidade ao contexto original.

- **Disponibilidade:** Moderada, os dados estão completos, mas não estão tão bem organizados, mas acompanham um dicionário de variáveis.

- **Correção:**Moderada, os dados apresentam coerência interna entre variáveis. No entanto, é necessário tratamento de valores nulos e análise de outliers para garantir maior integridade na modelagem.

- **Representatividade:** Alta, os dados representam bem o comportamento dos alunos em relação ao engajamento e à evasão no curso online.

Qual a saída esperada? Quais as métricas de sucesso?

Saída esperada:
Um modelo preditivo capaz de identificar, com antecedência, quais alunos têm maior risco de evasão em um curso online, com base em seus dados de interação na plataforma.

Métricas de sucesso:

- **Redução da taxa de evasão** nas turmas acompanhadas com base nas previsões do modelo.

- **Aumento na taxa de intervenções pedagógicas** realizadas em tempo hábil após alertas de risco.

- **Melhora no engajamento dos alunos sinalizados**, como aumento no envio de tarefas ou frequência de acesso.

- **Satisfação dos tutores e gestores** com as previsões, medida por questionários ou feedback direto.

- **Ganho em eficiência na gestão educacional**, com priorização de casos críticos baseada nas previsões.

Que tipo de UX é útil para complementar o modelo?
Uma interface simples e intuitiva, voltada para tutores ou gestores educacionais, que permita:

- Visualizar a lista de alunos com risco de evasão.

- Destacar os principais fatores que influenciaram a previsão

- Filtrar e buscar alunos por turma, desempenho ou tempo de atividade.

- Gerar relatórios automáticos com recomendações de intervenção.

Quais os modelos iniciais?

Os modelos iniciais considerados para o problema de evasão são:

- **Regressão Logística**: modelo simples e interpretável, usado como baseline.

- **Random Forest**: modelo de árvore robusto, eficiente para lidar com dados tabulares e variáveis categóricas.

- **Rede Neural (Keras)**: adequada para capturar relações complexas e não lineares nos dados

Esses modelos serão comparados quanto ao desempenho preditivo e à interpretabilidade, com o objetivo de selecionar a melhor abordagem para prever a evasão de alunos.