Open University Learning Analytics

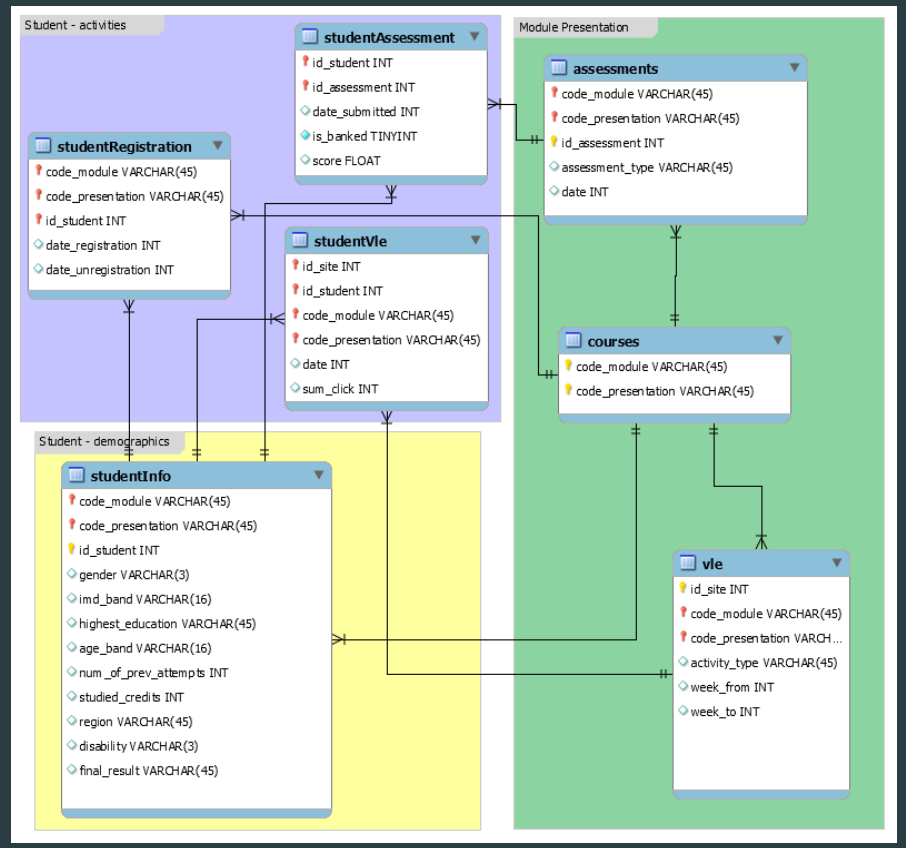
André Campolina

Agenda

- Características da base de dados
- Problema de negócio
- Descrição do modelo e método de avaliação
- Discussão dos resultados
- Próximos passos

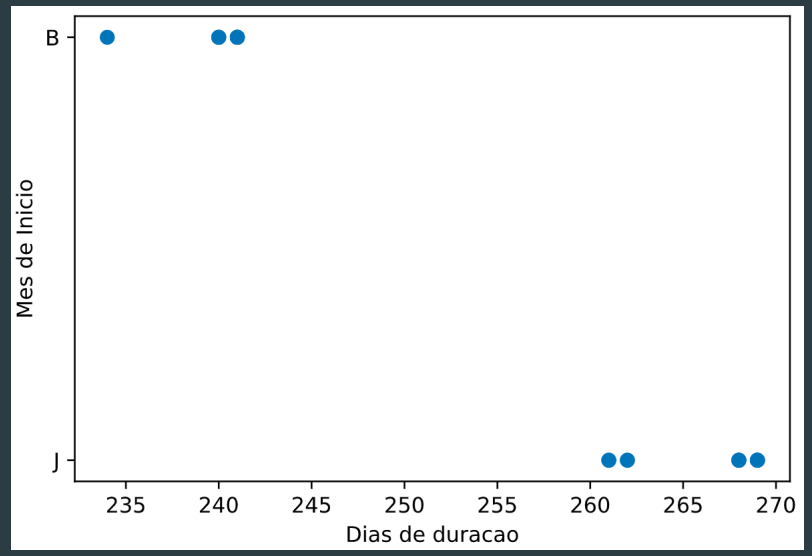
Características da base de dados

- Dados de alunos, cursos, atividades e materiais de uma universidade



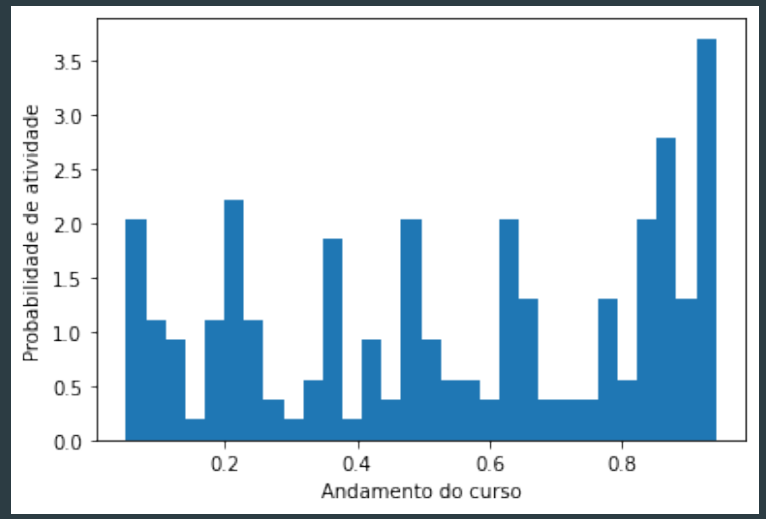
Características da base de dados

- Cursos
  - 7 cursos ofertados num total de 22 vezes
  - Ofertas acontecem em Fevereiro ou Outubro, os dados são de 2013 e 2014
  - Cursos duram em média 255 dias, mas a duração tem a ver com o mês de início



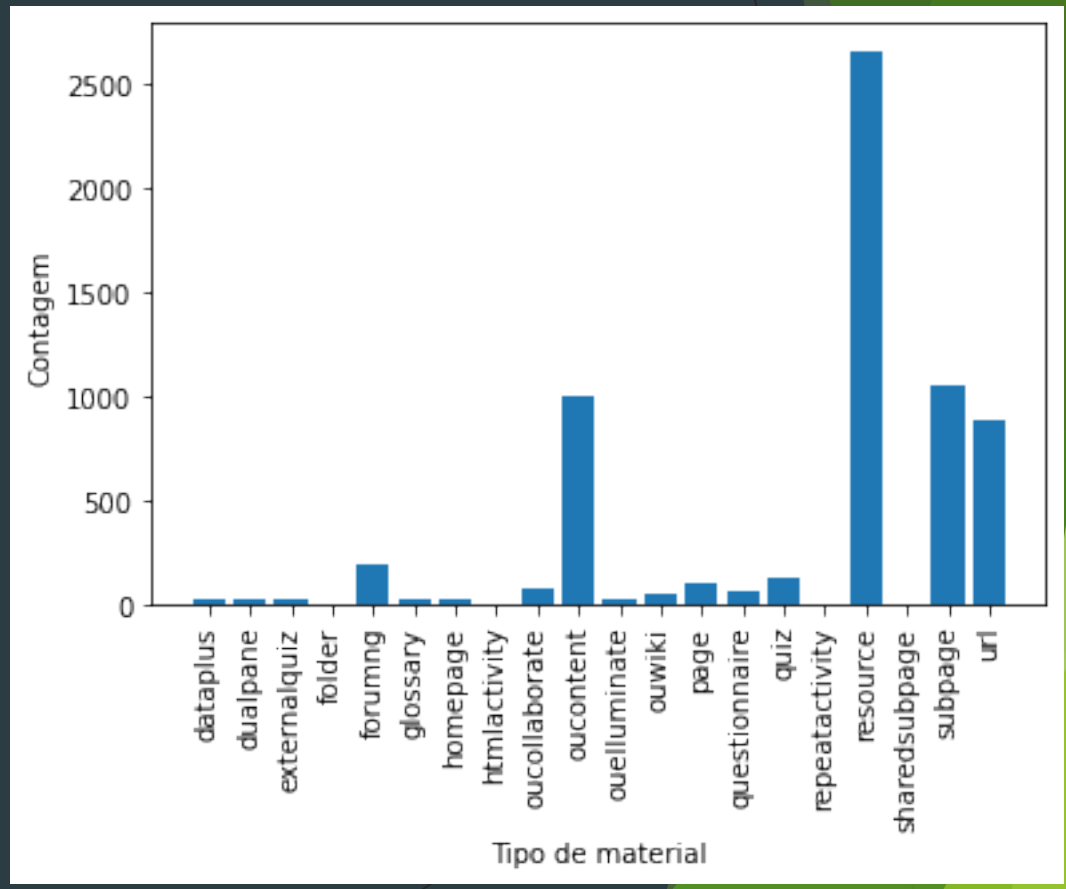
Características da base de dados

- Atividades
  - 106 atividades corrigidas por monitores (TMA), 76 atividades corrigidas por computador (CMA) e 24 exames finais
  - Atividades têm pesos atribuídos que devem somar 100%, excluindo as provas
    - Nenhum peso ultrapassa 35% e muitas atividades têm peso 0, em sua maioria CMAs
  - Cada curso tem, em média 8,27 atividades e existem picos de atividades ao longo dos cursos



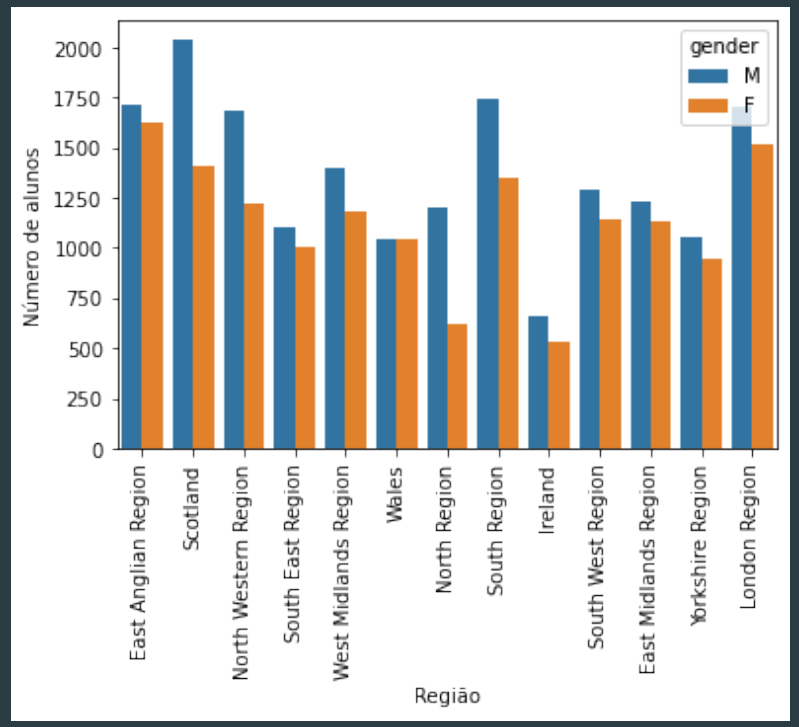
Características da base de dados

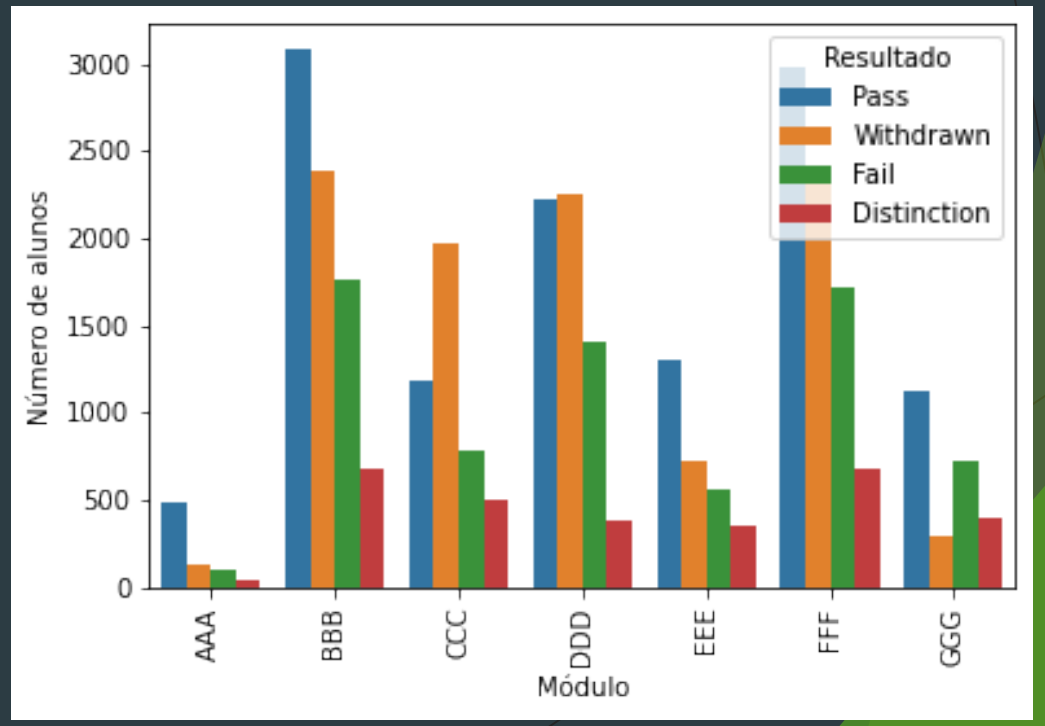
- Materiais (VLE)
  - Cada curso tem um conjunto único de materiais
  - Os mais populares são “recursos”, mas páginas internas, externas e subpáginas sõ anormalmente populares também



Características da base de dados

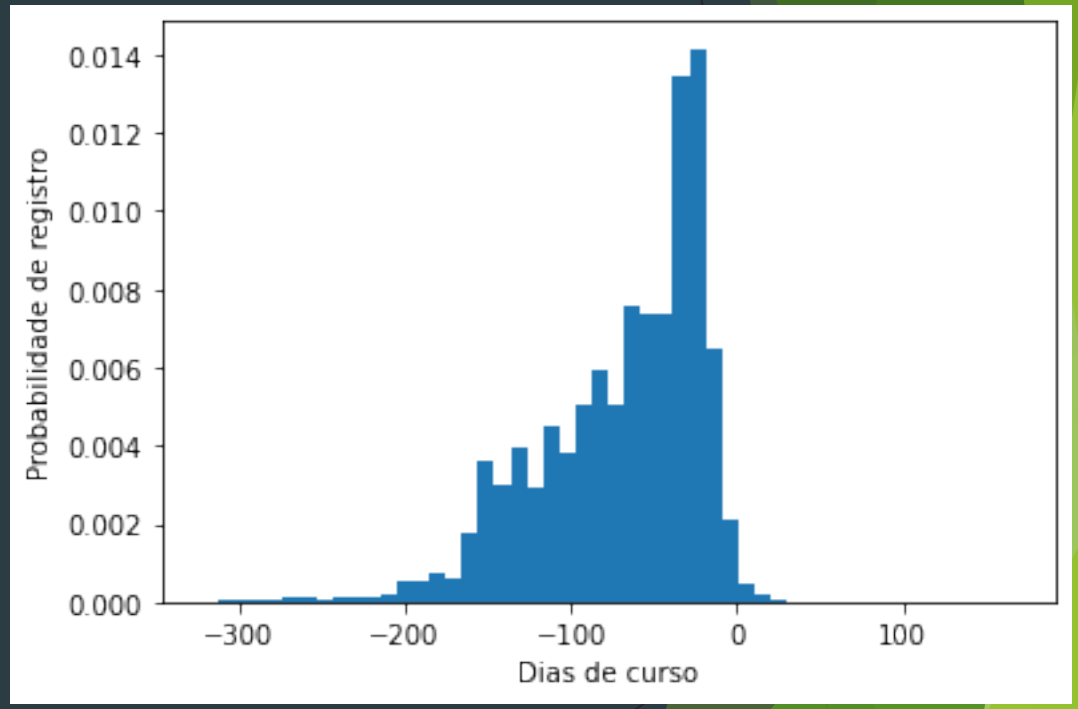
- Alunos
  - O conjunto de dados tem dados demográficos e desempenho sobre alunos





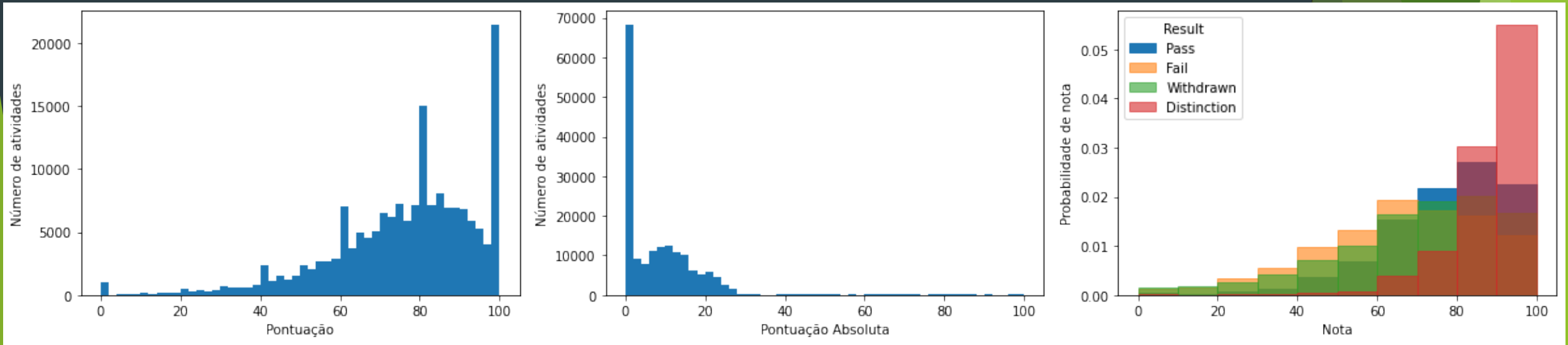
Características da base de dados

- Registro de alunos
  - Cada matrícula de aluno em um curso está registrada
  - A data de registro e de abandono (quando é o caso) é medida em número de dias a partir do início das aulas
  - Em média, alunos se registram 69 dias antes de começarem as aulas



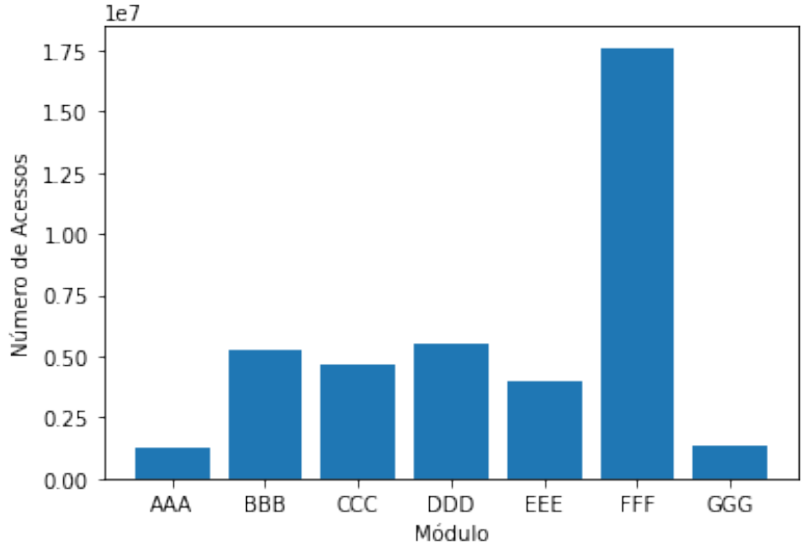
Características da base de dados

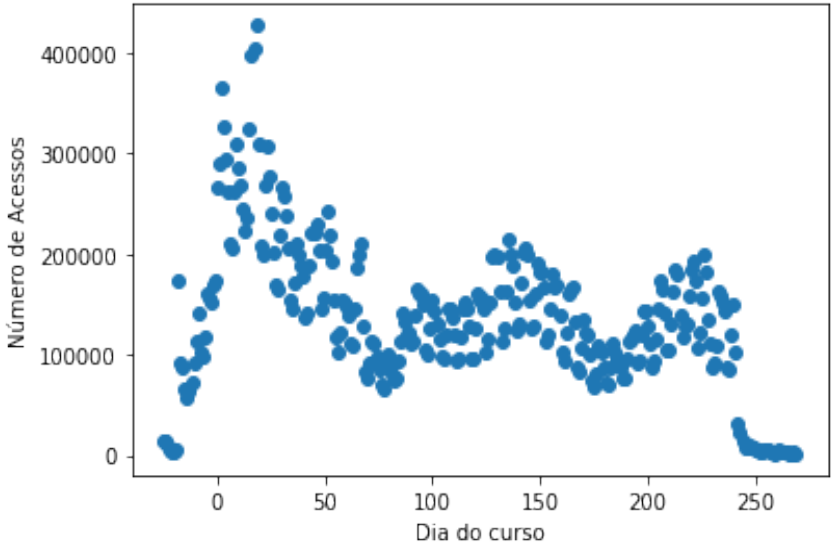
- Pontuação de alunos
  - Cada aluno recebe uma nota de 0 a 100 em toda atividade que submete
  - Atividades podem ser submetidas após a sua data limite

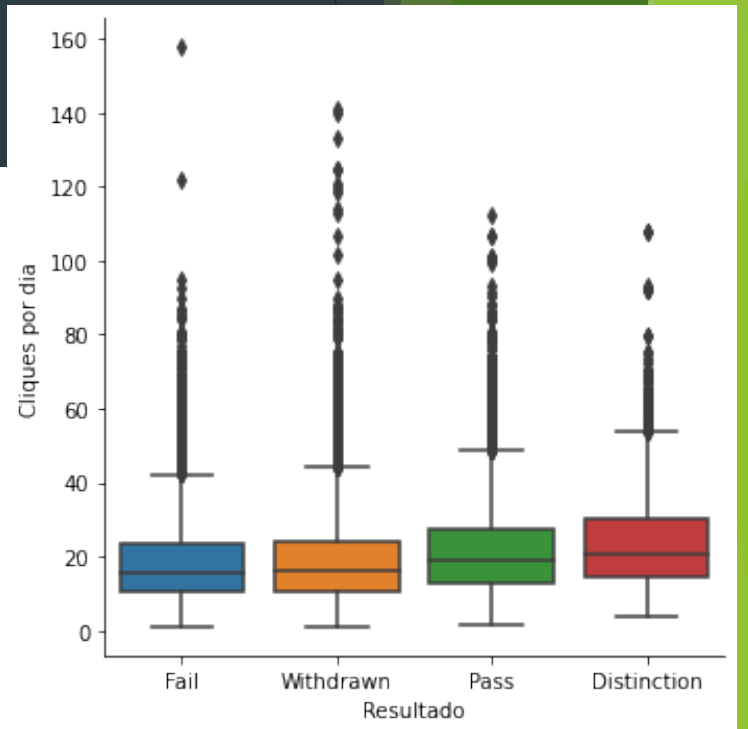


Características da base de dados

- Acesso a materiais
  - Os acessos de alunos aos materiais disponíveis ficou registrado
  - Sabemos quantos acessos cada aluno fez a um material em um dia
  - Número de acessos por dia são periódicos
  - Alunos aprovados tendem a acessar mais materiais







Problema de negócio

- Como predizer o desempenho dos estudantes nas avaliações finais?
- Ou então:

  Dado o desempenho de um aluno até o momento da prova final, ele será aprovado ou reprovado?

- Para isso, é preciso definir os dados que representarão o desempenho de um aluno até a prova final

Problema de negócio

- Dados fornecidos ao modelo:
  - Todos os dados sobre o curso atual
  - Dados demográficos do aluno, sem sua identificação única
  - Data de registro no curso
  - Data média de submissão nas atividades
    - Em relação à data limite da atividade e duração do curso
  - Nota média nas atividades
    - Notas relativas e absolutas
  - Média de cliques por dia em materiais de curso

Descrição do modelo e avaliação

- Prever a aprovação ou não de um aluno é um problema de classificação
  - Dadas as características de um aluno em um curso, ele faz parte do grupo de aprovados ou reprovados?
- Vou avaliar 3 tipos de modelo
  - Árvores de decisão
    - Usam limites lineares nos dados para diferenciar classes
  - Random forests
    - Um conjunto de árvores de decisão
  - MLP
    - Uma rede neural totalmente conectada
    - A combinação de neurônios simples pode encontrar relações não lineares

Descrição do modelo e avaliação

- Modelos de machine learning sofrem com dados de muitas dimensões
  - Por isso, é necessário reduzir o número de atributos nos dados
- Vou testar duas maneiras de reduzir dimensões
  - Selecionar N atributos mais correlacionados ao desempenho de um aluno
  - PCA: transforma os dados em N dimensões abstratas que promovem maior separação entre as dimensões originais
- Selecionei N=7

Descrição do modelo e avaliação

- Para avaliar cada combinação de modelo e dado de entrada, vou usar validação cruzada
  - 5 divisões em 80/20
- Em cada iteração, vou medir 3 métricas
  - Precisão
    - Quantas foram as classificações corretas?
  - F1-score
    - Como foi o tradeoff entre classificar corretamente e modelar corretamente as classes?
  - ROC-AUC
    - Como é a relação entre especificidade e sensibilidade?

Discussão dos resultados

	Correlação			PCA		
	Precisão	F1	ROC	Precisão	F1	ROC
Árvore de Decisão	0.776	0.795	0.780	0.734	0.757	0.740
Random Forests	0.821	0.836	0.929	0.807	0.829	0.893
Rede Neural	0.822	0.836	0.911	0.831	0.855	0.912

- Apesar de não ter o melhor ROC-AUC, a rede neural teve maior precisão. Por isso, foi o modelo selecionado

Próximos passos

- Os modelos foram todos treinados com configurações padrão
  - É possível melhorar o desempenho ajustando parâmetros individualmente
- A rede neural pode usar arquiteturas melhores que previnem overfitting e melhoram a sua precisão

Obrigado!

André Campolina