



Centro de
Informática
UFPE

Predição de Aprovação de Acesso a Recursos

Aprendizado de Máquina e Ciência de Dados

Gabriel Braz (gbcs)

Problema e Relevância

- **Problema:** Classificar solicitações de acesso como "aprovado" ou "negado"
- **Dados:**
 - 32.769 solicitações
 - 9 variáveis categóricas de alta cardinalidade
- **Relevância:**
 - Segurança: reduzir acessos não autorizados
 - Eficiência: automatizar decisões
 - Conformidade: rastreabilidade para auditorias

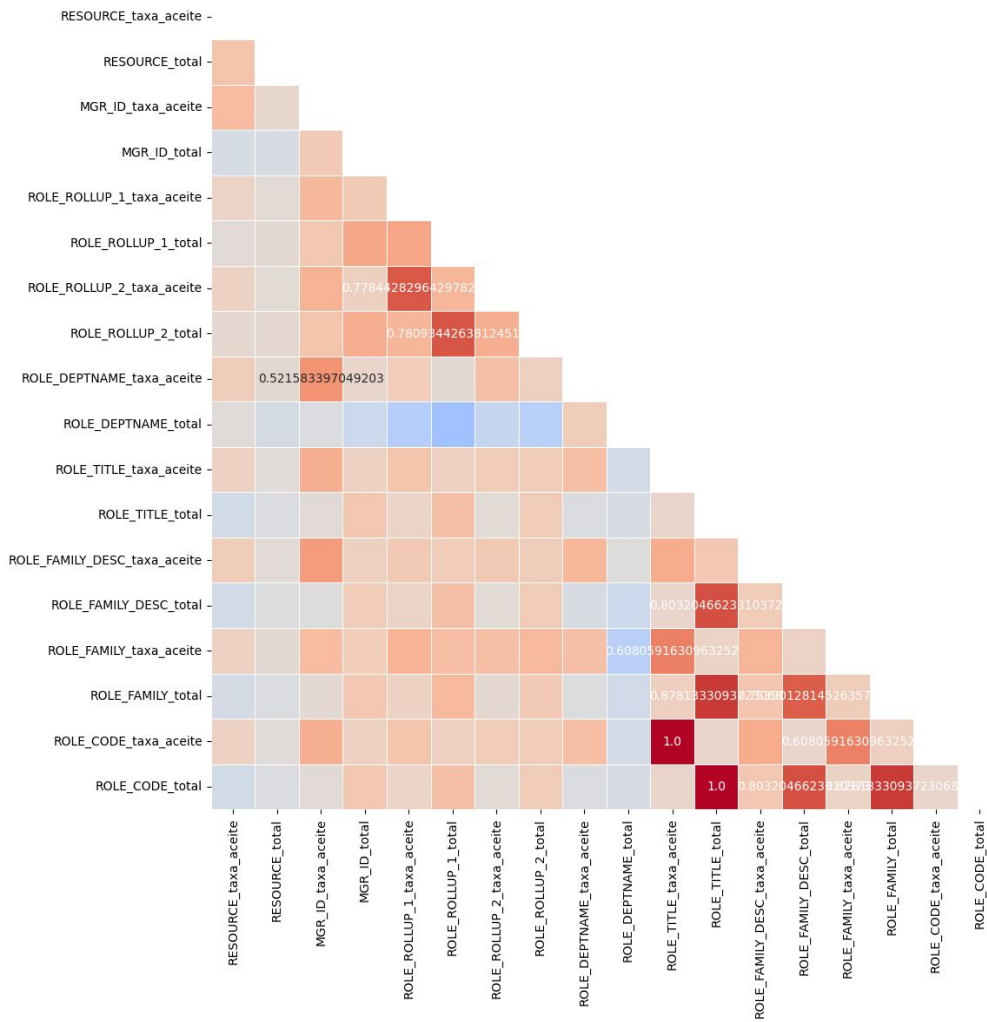
Desafios dos Dados

- **Alta cardinalidade:**
 - **RESOURCE** com 7.518 valores únicos
- **Desbalanceamento:**
 - **94,5%** aprovados
 - **5,5%** negados
- **Variáveis categóricas:**
 - **todas são IDs numéricos**
- **Sem valores faltantes ou duplicatas**

A Solução

- **Transformação proposta:**
 - Taxa histórica de aceite por valor categórico
 - Frequência total de cada valor
- **Resultado:**
 - 9 variáveis categóricas
 - → 18 features numéricas
- **Benefícios:**
 - captura de contexto
 - redução dimensional
 - sem explosão de dimensionalidade

Matriz de Correlação Linear das Features



Matriz de
Correlação
Linear

Pré-processamento Aplicado

- Normalização Min-Max (todas as features entre 0 e 1)
- Balanceamento com Random UnderSampling
 - Antes: 31.000 vs 1.897
 - Depois: 1.897 vs 1.897 (balanceado)
- Divisão estratificada:
 - 80% treino
 - 20% teste

Modelos Testados

- **KNN**
 - Simples
 - baseado em vizinhança
- **Random Forest**
 - Ensemble
 - robusto
 - com importância de features
- **MLP**
 - Rede neural
 - captura não-linearidades

Resultados Comparativos

Modelo	Acurácia	Precisão (0 - 1)	Recall (0 - 1)	F1-Score (0 - 1)
KNN	90%	90% - 89%	89% - 90%	90% - 90%
Random Forest	94%	91% - 96%	97% - 91%	94% - 93%
MLP	90%	89% - 90%	91% - 89%	90% - 90%

Análise Estatística e Significância

- Random Forest vs KNN/MLP:
 - diferença de 4% na acurácia
 - Valor-p < 0,001 (significativo)
- KNN vs MLP:
 - diferença não significativa
 - (p = 0,85)
- Conclusão:
 - Random Forest foi estatisticamente superior nos testes

Trade-offs e Impacto Operacional

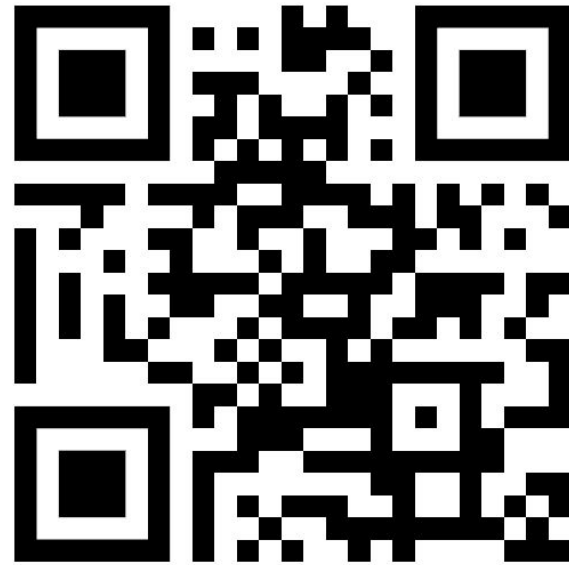
- **Segurança:**
 - Recall de 97% para "negado"
 - → bom controle
- **Conveniência:**
 - Precisão de 96% para "aprovado"
 - → menos falsos positivos
- **Equilíbrio ideal para ambiente corporativo**

Conclusões e Insights

- Feature engineering funcionou:
 - taxas históricas + frequência foram efetivas
- Random Forest foi o melhor modelo:
 - 94% de acurácia
 - robusto
 - interpretável
- Limitações:
 - perda de dados no undersampling
 - sem validação temporal
- Próximos passos:
 - validação temporal
 - testes com oversampling
 - explicação de decisões (XAI)



Muito
obrigado!



<https://portal.cin.ufpe.br/>



@cinufpe



Centro de Informática UFPE



Centro de Informática UFPE



@CInUFPE



Centro de
Informática
UFPE