DOCUMENTAÇÃO TÉCNICA DO SISTEMA DE PREDIÇÃO E CONTROLE DE EVASÃO ACADÊMICA

Autor: Davi Ferreira Freitas

---

1. Resumo Executivo

Este documento apresenta a arquitetura, funcionamento e fundamentos técnicos do Sistema Integrado de Predição e Inserção de Alunos Evadidos, desenvolvido para uma instituição de ensino superior.

O projeto foi criado com o objetivo de identificar, registrar e prever a evasão de alunos dos turnos EAD, híbrido e presencial, utilizando técnicas avançadas de engenharia de dados e modelos de Machine Learning.

A solução é composta por três módulos principais:

1. **Inserção de alunos evadidos**
2. **Inserção de todos os alunos ativos**
3. **Modelo de predição de evasão usando LightGBM, RandomForest e CatBoost**

Todos os módulos foram desenvolvidos e aprimoradas por Davi Ferreira Freitas.

---

2. Objetivos do Projeto

2.1. Objetivo Geral

Desenvolver um sistema automatizado para:

- **Prever a probabilidade de evasão dos alunos.**
- **Registrar automaticamente alunos evadidos no banco institucional.**
- **Criar uma base atualizada de todos os alunos ativos para consultas e análises.**

2.2. Objetivos Específicos

- **Criar pipelines estruturadas de leitura e validação de dados.**
- **Implementar três algoritmos preditivos (LightGBM, RandomForest, CatBoost).**
- **Definir thresholds personalizados conforme prioridades institucionais.**
- **Garantir integridade dos dados nos ambientes EAD, Híbrido e Presencial.**

- **Disponibilizar um conjunto de modelos e arquivos para utilização operacional.**

---

3. Arquitetura Geral do Sistema

O sistema é composto por quatro notebooks principais, cada um responsável por uma etapa crítica do processo de predição e gestão da evasão acadêmica.

3.1. MachineLearning_LightGBM.ipynb — Modelo principal para Ensino Presencial

Este notebook contém o desenvolvimento do modelo LightGBM, que apresentou o melhor desempenho para alunos do ensino Presencial, considerando:

- **Volume de dados significativamente maior**
- **Alta variabilidade comportamental**
- **Necessidade de velocidade e estabilidade no treinamento**
- **Melhor desempenho geral em métricas de Recall e AUC**

Responsável por:

- **Preparação detalhada dos dados históricos**
- **Engenharia de atributos acadêmicos, socioeconômicos e comportamentais**
- **Treinamento e validação do LightGBM**
- **Avaliação de métricas e comparação contra outros modelos**
- **Ajuste do threshold ideal para identificação de risco**
- **Exportação do modelo otimizado para produção**

3.2. MachineLearning_CatBoost.ipynb — Melhor modelo para Ensino Híbrido

Este notebook inclui o desenvolvimento do modelo CatBoost, que demonstrou o melhor desempenho para alunos do ensino Híbrido, especialmente devido a:

- **Melhor tratamento de variáveis categóricas**
- **Alta tolerância a dados heterogêneos e mistos**
- **Estabilidade superior em bases menores e mais irregulares**

Responsável por:

- **Preparação dos dados específicos do público híbrido**
- **Treinamento do modelo CatBoost**
- **Testes de precisão, recall e F1-score**
- **Seleção do threshold mais apropriado para o perfil híbrido**
- **Exportação do modelo final ajustado**

3.3. MachineLearning_RandomForest.ipynb — Melhor modelo para Ensino EAD

Notebook dedicado ao treinamento do RandomForest, que apresentou o melhor resultado para alunos do EAD, considerando:

- **Dados mais uniformes e com menor variabilidade de presença física**
- **Melhor equilíbrio entre precisão e recall nesse segmento**
- **Maior estabilidade em cenários com distribuição menos complexa**

Responsável por:

- **Limpeza e preparação da base EAD**
- **Treinamento do modelo RandomForest**
- **Avaliação comparativa com CatBoost e LightGBM**
- **Definição de threshold adequado para maximizar deteções**
- **Exportação do modelo final**

3.4. Insere_Resultado_Machine.ipynb — Inserção dos resultados no Banco de Dados

Notebook responsável por inserir no banco de dados institucional os resultados gerados pelos modelos de Machine Learning.

Responsável por:

- **Carregar o arquivo .pkl do modelo correspondente ao turno (Presencial, Híbrido ou EAD)**
- **Gerar previsões atualizadas para todos os alunos**
- **Criar estrutura de tabela padronizada de risco de evasão**
- **Inserir ou atualizar registros via SQLAlchemy**

- **Garantir integridade e consistência das informações persistidas**

---

4. Pipeline de Inserção de Dados

4.1. Entrada

- **Arquivos Excel da instituição**
- **Dados acadêmicos, cadastrais e de frequência**
- **Identificação do turno (EAD, Híbrido e Presencial)**

4.2. Etapas

1. **Carregamento com Pandas**
2. **Padronização de colunas e tipos**
3. **Remoção de duplicidades**
4. **Validação de matrícula e status**
5. **Comparação entre planilha e banco**
6. **Inserção final via SQLAlchemy**

4.3. Saída

- **Registros novos inseridos**
- **Alunos atualizados**
- **Logs de inconsistência**
- **Atualização das tabelas de alunos e evasão**

---

5. Modelos de Machine Learning Desenvolvidos

A modelagem envolveu três algoritmos principais — RandomForest, CatBoost e LightGBM — sendo cada um deles avaliado, ajustado e selecionado conforme o desempenho por modalidade de ensino (EAD, Híbrido e Presencial).

Embora todos os modelos tenham sido treinados sobre a mesma estrutura de dados, cada modalidade apresentou comportamento distinto, resultando na escolha de um modelo ideal para cada formato acadêmico.

5.1. RandomForest – Modelo selecionado para Ensino EAD

O RandomForest apresentou desempenho superior nos dados do ensino EAD, caracterizado por:

- **Estrutura de dados mais uniforme**
- **Menor variabilidade comportamental entre alunos**
- **Excelente equilíbrio entre precisão e recall**
- **Robustez em cenários com menor quantidade de features presenciais**

O modelo se destacou pela estabilidade e consistência nas previsões de evasão para alunos totalmente remotos, tornando-se a melhor escolha para esta modalidade.

5.2. CatBoost – Modelo selecionado para Ensino Híbrido

O CatBoost demonstrou o melhor desempenho para alunos do ensino Híbrido, devido a:

- **Tratamento nativo de variáveis categóricas**
- **Alta performance em bases heterogêneas**
- **Melhor adaptação ao público que alterna entre aulas online e presenciais**
- **Excelente recall e estabilidade**

Por conta da natureza mais mista (comportamento online + presencial), o CatBoost mostrou-se o modelo mais adequado ao perfil híbrido.

5.3. LightGBM – Modelo selecionado para Ensino Presencial

O LightGBM foi o modelo que apresentou melhor performance global para alunos do ensino Presencial, principalmente pela quantidade maior de dados, variabilidade comportamental e volume mais alto de registros.

Ele se destacou por:

- **Maior velocidade de treinamento**
- **Ótimo equilíbrio entre performance, recall e AUC**
- **Suporte eficiente a grandes volumes de dados**
- **Capacidade de lidar com elevado número de atributos sem perda de estabilidade**

Por esse motivo, o LightGBM foi definido como o modelo final para a modalidade Presencial.

---

6. Preparação dos Dados para os Modelos

6.1. Engenharia de Variáveis

Foram criadas e tratadas diversas variáveis relevantes para a análise de evasão, entre elas:

- **Frequência total e por disciplina**
- **Notas, desempenho histórico e reprovações**
- **Número de disciplinas ativas**
- **Modalidade de ensino (EAD, Híbrido, Presencial)**
- **Dados cadastrais e socioeconômicos (quando fornecidos)**
- **Tempo de vínculo acadêmico**
- **Indicadores financeiros (pagamentos, atrasos, inadimplência)**

Essas features foram padronizadas e preparadas de acordo com as necessidades técnicas de cada modelo.

6.2. Divisão dos Dados

A divisão adotada para o treinamento dos três modelos seguiu o padrão:

- **80% – Treinamento**
- **20% – Teste**

Garantindo avaliação justa para cada modalidade (EAD, Híbrido, Presencial).

6.3. Métricas Avaliadas

Os modelos foram comparados utilizando as principais métricas para classificação binária:

- **Accuracy**
- **Precision**
- **Recall (métrica prioritária para evasão)**

- **F1-score**
- **AUC**
- **Curvas Precision x Recall**

Cada modalidade teve thresholds ajustados de acordo com seu comportamento estatístico.

6.4. Seleção de Threshold

A partir das curvas de Precision/Recall e análises segmentadas por modalidade, definiu-se que thresholds inferiores ao padrão (0.50) proporcionaram melhor taxa de detecção de evasão.

Exemplo utilizado nos testes:

threshold = 0.34

Cada modelo final recebeu seu threshold exclusivo, conforme tuning e comportamento dos dados.

---

7. Exportação e Persistência dos Modelos

Cada modelo foi exportado em .pkl, incluindo:

- **Pipeline completo de pré-processamento**
- **Modelo treinado**
- **Threshold configurado**
- **Estrutura de previsão para uso direto em produção**

Exemplo de arquivo gerado:

modelo_lgb_com_threshold_034_real.pkl

Os arquivos são carregados pelos notebooks de operação e integração com o banco de dados institucional.

---

8. Aplicação dos Modelos por Modalidade: EAD, Híbrido e Presencial

Cada modalidade recebeu o modelo de Machine Learning mais adequado ao seu comportamento:

8.1. Ensino EAD – RandomForest

- **Maior risco estrutural de evasão**
- **Dados mais homogêneos**
- **RandomForest obteve melhor recall e estabilidade**

8.2. Ensino Híbrido – CatBoost

- **Variabilidade entre presença física e atividades remotas**
- **CatBoost lidou melhor com dados mistos e categóricos**
- **Melhor desempenho geral para o perfil híbrido**

8.3. Ensino Presencial – LightGBM

- **Maior volume de dados e variabilidade comportamental**
- **LightGBM entregou melhor AUC, velocidade e recall**
- **Modelo final utilizado para esta modalidade**

O sistema reconhece e trata cada modalidade de forma independente, otimizando a assertividade das previsões de evasão.

---

9. Benefícios Institucionais

A solução proporciona:

- **Detecção antecipada de alunos com risco de evasão**
- **Possibilidade de intervenção proativa pelas áreas responsáveis**
- **Registro automático e padronizado das informações**
- **Redução do trabalho manual e inconsistências**
- **Base consolidada e confiável para análises estratégicas**
- **Apoio à tomada de decisão acadêmica e administrativa**

---

10. Conclusão

O Sistema de Predição e Inserção de Evasão Acadêmica, desenvolvido por Davi Ferreira Freitas, integra automação, engenharia de dados e múltiplas técnicas de Machine Learning — RandomForest, CatBoost e LightGBM — otimizadas para os turnos EAD, Híbrido e Presencial.

A solução é madura, estruturada e pronta para implantação institucional, garantindo:

- **Precisão**
- **Confiabilidade**
- **Escalabilidade**
- **Manutenção facilitada**