CENTRO UNIVERSITÁRIO IBMR

Pedro Augusto Costa Cordeiro Cunha Ribeiro

CONSTRUÇÃO DE UM *PIPELINE* DE DADOS PARA ANÁLISE DE *CHURN*

Rio de Janeiro
2024

Pedro Augusto Costa Cordeiro Cunha Ribeiro

CONSTRUÇÃO DE UM *PIPELINE* DE DADOS PARA ANÁLISE DE *CHURN*

Trabalho de Conclusão de Curso apresentado ao Departamento de Engenharia de Produção do Centro Universitário IBMR como requisito parcial para a obtenção do título de Engenheiro de Produção.
Orientador: Michael Leoni
Coorientador:

Rio de Janeiro
2024

Pedro Augusto Costa Cordeiro Cunha Ribeiro

CONSTRUÇÃO DE UM *PIPELINE* DE DADOS PARA ANÁLISE DE *CHURN*

Este Trabalho de Conclusão de Curso foi julgado adequado e aprovado, em sua forma final, pelo Curso de Graduação em Engenharia de Produção, do Centro Universitário IBMR.

Rio de Janeiro, 06 de outubro de 2024

Orientador

Banca Examinadora

Rio de Janeiro
2024

Construção de um *pipeline* de dados para análise de *churn*

Pedro Augusto Costa Cordeiro Cunha Ribeiro[1]

Resumo

O presente trabalho tem como objetivo desenvolver e apresentar a construção de um pipeline de dados para a análise de churn em uma instituição de ensino superior, utilizando uma abordagem de análise por coortes. A proposta envolve a criação de uma base de dados simulada, abrangendo o período de 2010 a 2023, com informações relevantes como matrículas, cancelamentos, cursos, e motivos de evasão. O pipeline é responsável por realizar o tratamento, a limpeza e a integração dos dados, permitindo uma análise descritiva e preditiva da retenção de alunos. Os resultados obtidos demonstram a importância de um pipeline de dados bem estruturado para facilitar a tomada de decisões estratégicas voltadas à retenção de estudantes e melhoria contínua de processos institucionais.

Palavras-chave: *Churn*; Pipeline; Análise de *Cohort*; Análise de dados; Engenharia de dados

Absctract

This work aims to develop and present the construction of a data pipeline for churn analysis in a higher education institution, using a cohort analysis approach. The proposal involves creating a simulated database, covering the period from 2010 to 2023, with relevant information such as enrollments, cancellations, courses, and dropout reasons. The pipeline is responsible for data processing, cleaning, and integration, allowing descriptive and predictive analysis of student retention. The results demonstrate the importance of a well-structured data pipeline to facilitate strategic decision-making aimed at student retention and continuous improvement of institutional processes. Keywords: churn, data pipeline, cohort analysis, student retention, higher education.

Keywords: Churn; Pipeline; Cohort Analysis; Data analytics; Data engineering

[1] Engenharia de Produção pelo Centro Universitário IBMR, Rio de Janeiro, Brasil. E-mail: paugustoribeiro@gmail.com

SUMÁRIO

1 INTRODUÇÃO

Na contemporaneidade empresarial, onde a informação se torna vital para o êxito organizacional, a Engenharia de Dados surge como um pilar essencial para a sustentabilidade operacional. A integração estratégica de ferramentas de gestão de dados tem sido uma abordagem crucial para otimizar processos, além de proporcionar uma visão mais aprimorada e abrangente das operações.

Nos últimos anos, a análise de dados evoluiu de uma simples prática operacional para se consolidar como uma ferramenta estratégica indispensável para as organizações. Com o crescimento exponencial de dados gerados diariamente, a capacidade de coletar, processar e interpretar informações de forma eficiente tem proporcionado uma vantagem competitiva significativa. A análise de dados permite insights profundos sobre o comportamento dos consumidores, otimiza processos internos, melhora a eficiência operacional e apoia a inovação. Além disso, ferramentas avançadas de análise preditiva possibilitam a antecipação de tendências e a identificação de oportunidades e riscos com maior precisão. Em um ambiente de negócios cada vez mais dinâmico e competitivo, a habilidade de transformar dados em decisões assertivas tornou-se um diferencial crucial para o sucesso e a sustentabilidade das organizações a longo prazo.

No contexto educacional, o gerenciamento eficaz de alunos e a fidelização têm ganhado destaque, considerando o aumento da concorrência entre instituições e a necessidade de otimização dos recursos acadêmicos. Com a crescente digitalização do setor, as instituições de ensino têm acesso a um volume significativo de dados que, se bem tratados e analisados, podem fornecer insights valiosos sobre o comportamento e as necessidades dos alunos.

A taxa de *churn* impacta diretamente a rentabilidade das instituições de ensino, sendo um indicador crucial da satisfação dos alunos e da eficiência das estratégias de marketing e retenção. No entanto, a ausência de um pipeline de dados eficiente pode dificultar a identificação de padrões e a compreensão das razões que levam os alunos a abandonarem o curso.

Diante desse cenário, o desenvolvimento de um pipeline de dados robusto torna-se essencial para monitorar e analisar métricas de evasão, permitindo que as instituições identifiquem as turmas mais propensas ao cancelamento e adotem estratégias proativas de retenção. A implementação de processos eficazes de coleta, tratamento e análise de dados possibilita a segmentação dos alunos e a personalização de ofertas, aumentando as chances de retenção.

Este trabalho tem como objetivo explorar a construção de um pipeline de dados para análise de *churn* utilizando a metodologia de *cohort*, com foco em transformar dados brutos em informações estratégicas. Serão abordados os benefícios da análise descritiva e a importância de um processo estruturado na tomada de decisões, com o intuito de otimizar as estratégias de retenção e melhorar a experiência do aluno no setor educacional.

1.1 OBJETIVO

O objetivo deste trabalho é desenvolver um pipeline de dados eficiente para análise de *churn* em uma instituição de ensino superior, utilizando uma base de dados simulada para identificar padrões de cancelamento e auxiliar na tomada de decisões estratégicas de retenção de alunos.

Os objetivos específicos são:

a) Criar e estruturar uma base de dados simulada de matrículas e cancelamentos de alunos ao longo de 10 anos;
b) Construir um pipeline de dados;
c) Aplicar técnicas de análise de *cohort* para avaliar a retenção e cancelamento de alunos;
d) Gerar visualizações que facilitem a interpretação dos resultados e a implementação de estratégias de retenção;
e) Validar a eficácia do pipeline e dos métodos utilizados para análise de *churn*.

1.2 METODOLOGIA

O presente trabalho foi conduzido com uma abordagem prática, focando na implementação de um pipeline de dados para a análise de *churn* no contexto

educacional. Para isso, inicialmente, foram geradas bases de dados fictícias em Python, simulando informações de matrícula, incluindo: curso, grupo, status, cancelamento, gênero, idade, motivo de cancelamento, modalidade, turno, mensalidade, desconto, valor de pagamento, data de matrícula e data de cancelamento.

Após a criação das bases, estruturou-se um *pipeline* abrangendo as etapas de coleta e tratamento das informações. Este pipeline foi desenvolvido utilizando ferramentas de programação, com ênfase em Python e na biblioteca Pandas, assegurando eficiência no manuseio e processamento dos dados.

Em seguida, os dados foram compactados e unificados em um único arquivo CSV, facilitando a integração com o Power BI para a confecção de um dashboard que facilita identificar o perfil do aluno com maior propensão ao cancelamento.

Posteriormente, a análise de *cohort* foi aplicada para segmentar as informações coletadas em grupos, permitindo a identificação de padrões relevantes. Foi elaborado um relatório em Excel para apresentar a matriz de *cohort*, visando identificar safras com baixas taxas de retenção ou altas taxas de cancelamento, oferecendo uma visão clara sobre o comportamento dos alunos.

Finalmente, com base nos insights gerados, foram propostas estratégias de retenção personalizadas, com o intuito de melhorar a experiência acadêmica e reduzir a taxa de *churn* na instituição.

2 REVISÃO BIBLIOGRÁFICA

A análise de *churn*, um conceito amplamente estudado no setor empresarial, refere-se à taxa de abandono de clientes ou usuários de um serviço. Kumar e Petersen (2012) destacam que entender o *churn* é essencial para desenvolver estratégias de retenção eficazes, uma vez que a perda de clientes impacta diretamente a receita e a sustentabilidade das organizações. No contexto educacional, o *churn* se manifesta como evasão escolar, onde estudantes deixam de se matricular ou abandonam seus cursos, prejudicando tanto sua trajetória acadêmica quanto o desempenho institucional.

Vários estudos têm explorado a aplicação de técnicas analíticas para prever e mitigar o *churn*. No setor de saúde, Martins e Silva (2020) utilizaram um modelo preditivo para identificar clientes de planos de saúde com maior risco de cancelamento. Esse estudo demonstrou como a análise de *churn* pode ser uma ferramenta poderosa para a retenção, ao permitir intervenções direcionadas com base em dados históricos e comportamentais.

No campo do marketing, a análise de *cohort* tem sido amplamente utilizada para segmentar e entender o comportamento de grupos específicos de clientes ao longo do tempo. Fader (2012) explica que essa técnica permite que as organizações rastreiem como diferentes coortes reagem a variáveis específicas, como mudanças em produtos ou campanhas promocionais, oferecendo insights valiosos para a tomada de decisões estratégicas.

A aplicação dessas metodologias no setor educacional tem sido explorada por instituições como o Instituto Federal do Rio Grande do Norte (IFRN) e o Instituto Federal da Paraíba (IFPB). No IFRN, Barros et al. (2020) implementaram um pipeline de dados para prever a evasão escolar, utilizando técnicas de modelagem preditiva para identificar estudantes em risco. A análise de *cohort* foi empregada para segmentar os estudantes com base em diferentes variáveis, como ano de ingresso e curso, permitindo uma análise detalhada das taxas de retenção e evasão ao longo do tempo.

Silva et al. (2020) no IFPB adotaram uma abordagem semelhante, utilizando um pipeline de dados para analisar os fatores associados à evasão escolar. A análise de *cohort* revelou padrões de comportamento em diferentes coortes de estudantes, destacando a importância de variáveis socioeconômicas e acadêmicas na decisão de permanecer ou abandonar o curso. Os insights obtidos permitiram a implementação de estratégias de retenção mais eficazes, como programas de apoio financeiro e acadêmico.

Esses estudos demonstram que, assim como no setor empresarial, a análise de *churn* e *cohort* no contexto educacional oferece uma abordagem robusta para entender e melhorar a retenção de alunos. A integração de pipelines de dados e um bom gerenciamento de banco de dados, junto com

técnicas analíticas não apenas facilita a identificação de padrões de evasão, mas também permite o desenvolvimento de intervenções proativas e personalizadas, contribuindo para a melhoria contínua das práticas institucionais.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 PIPELINES DE DADOS

Um pipeline de dados consiste em uma série de processos automatizados que viabilizam a coleta, o tratamento e a análise de dados de maneira contínua e eficiente. A estruturação adequada de um pipeline é essencial para assegurar a qualidade e a relevância dos dados processados, permitindo análises mais precisas e informativas.

A construção de um pipeline robusto geralmente envolve três etapas principais: extração, transformação e carregamento (ETL). Durante a extração, os dados são coletados de diversas fontes, como bancos de dados, APIs e arquivos. A fase de transformação é dedicada à limpeza, normalização e preparação dos dados para análise. Finalmente, os dados transformados são carregados em sistemas de armazenamento, como *Data Warehouses*, onde podem ser acessados para análises futuras.

Zikopoulos et al. (2012) destacam que pipelines de dados eficazes são cruciais para iniciativas de *business intelligence*, pois permitem a captura e tratamento de dados, facilitando análises contínuas e a geração de insights acionáveis. No contexto deste trabalho, o pipeline foi utilizado para processar e compactar dados simulados de matrículas e cancelamentos de alunos, assegurando a integridade dos dados e sua integração eficiente com ferramentas de visualização.

3.2 *CHURN*

O conceito de *churn*, ou taxa de evasão, refere-se ao abandono ou cancelamento de um serviço por parte dos clientes. Esse indicador é amplamente utilizado para medir a lealdade e a retenção dentro de organizações. Kumar e Petersen (2012) destacam a importância de modelos preditivos na identificação precoce de padrões de *churn*, permitindo que as organizações implementem estratégias preventivas de retenção.

No contexto educacional, o *churn* é evidenciado pela desistência de alunos, seja por não renovarem suas matrículas ou por abandonarem o curso. Segundo Kumar e Shah (2004), a compreensão das causas do *churn* educacional é crucial, dado que a perda de alunos não apenas afeta financeiramente as instituições, mas também impacta sua reputação e capacidade de atrair novos estudantes. Fatores como insatisfação com o ensino, falta de engajamento e mudanças nas circunstâncias pessoais são frequentemente citados como razões para a evasão, ressaltando a necessidade de intervenções direcionadas para mitigar esses fatores.

3.3 CÁLCULO DE *CHURN*

A taxa de *churn* é uma métrica crucial para a avaliação da retenção, seja em empresas ou instituições educacionais. O *churn* pode ser calculado usando a seguinte equação:

Figura 1 – Fórmula de *churn*

$$Churn = \left(\frac{\mathrm{C}c}{\mathrm{C}t}\right) \times 100$$

$\mathrm{C}c$ = Número de clientes que cancelaram durante o período.

$\mathrm{C}t$ = Número total de clientes no início do período

Este cálculo fornece uma visão clara sobre a saúde da base de clientes ou alunos de uma instituição, permitindo a implementação de estratégias direcionadas para minimizar o churn. No ambiente educacional, a aplicação dessa métrica é fundamental para monitorar a eficácia de iniciativas de retenção e para ajustar estratégias conforme necessário.

3.4 ANÁLISE DE *COHORT*

A análise de *cohort* é uma técnica estatística utilizada para examinar o comportamento de grupos específicos de indivíduos, denominados *cohorts*, que compartilham características ou experiências comuns ao longo do tempo. Fader (2012) destaca que essa abordagem permite identificar tendências e padrões que podem não ser evidentes em análises globais.

A aplicação dessa técnica envolve várias etapas. Primeiro, a definição do *cohort*, que pode ser baseada em critérios como a data de ingresso ou características demográficas. Em seguida, os dados relevantes são coletados para cada cohort. A análise temporal é então conduzida para avaliar como o comportamento de cada grupo evolui ao longo do tempo, revelando padrões de retenção e churn (Chen et al., 2019).

Ramaswamy e Ozcan (2018) afirmam que a análise de cohort auxilia na segmentação eficaz da base de clientes ou usuários, proporcionando insights que informam decisões estratégicas e a personalização de ofertas. No contexto educacional, essa técnica é particularmente útil para entender a dinâmica de retenção de estudantes e para identificar oportunidades de melhoria nos processos institucionais. Neste trabalho, a análise de cohort foi utilizada para segmentar os alunos simulados e avaliar suas taxas de retenção e cancelamento ao longo do tempo.

4 APRESENTAÇÃO DE RESULTADOS

A análise de dados realizada neste trabalho revelou informações cruciais sobre o comportamento dos alunos e suas taxas de retenção. A seguir, são apresentados os principais resultados obtidos através do pipeline de dados implementado, da análise de *cohort* e das visualizações geradas no Power BI.

4.1 PIPELINE DE DADOS EM PYTHON

4.2 TAXAS DE *CHURN* EM EXCEL

A taxa de *churn* calculada para o período analisado foi de XX%. Essa métrica indica a proporção de alunos que cancelaram suas matrículas em relação ao total de matrículas no início do período. A identificação de fatores que contribuem para esse abandono foi um passo fundamental para entender as dinâmicas de retenção.

4.3 MATRIZ DE *COHORT* EM EXCEL

A matriz de cohort, elaborada no Excel, foi fundamental para identificar padrões de retenção e cancelamento ao longo dos anos. A matriz segmentou os

alunos em diferentes coortes com base no ano de ingresso, permitindo uma análise detalhada das taxas de retenção mês a mês.

Essa análise revelou tendências importantes, como:

- Taxas de Retenção por Cohort: Alunos que ingressaram em anos mais recentes mostraram taxas de retenção superiores, sugerindo melhorias nos programas de apoio e adaptação das estratégias institucionais.
- Padrões de Cancelamento: A matriz destacou períodos críticos durante os quais a taxa de cancelamento foi mais alta, fornecendo insights sobre momentos que requerem atenção especial, como o início e o meio do semestre letivo.

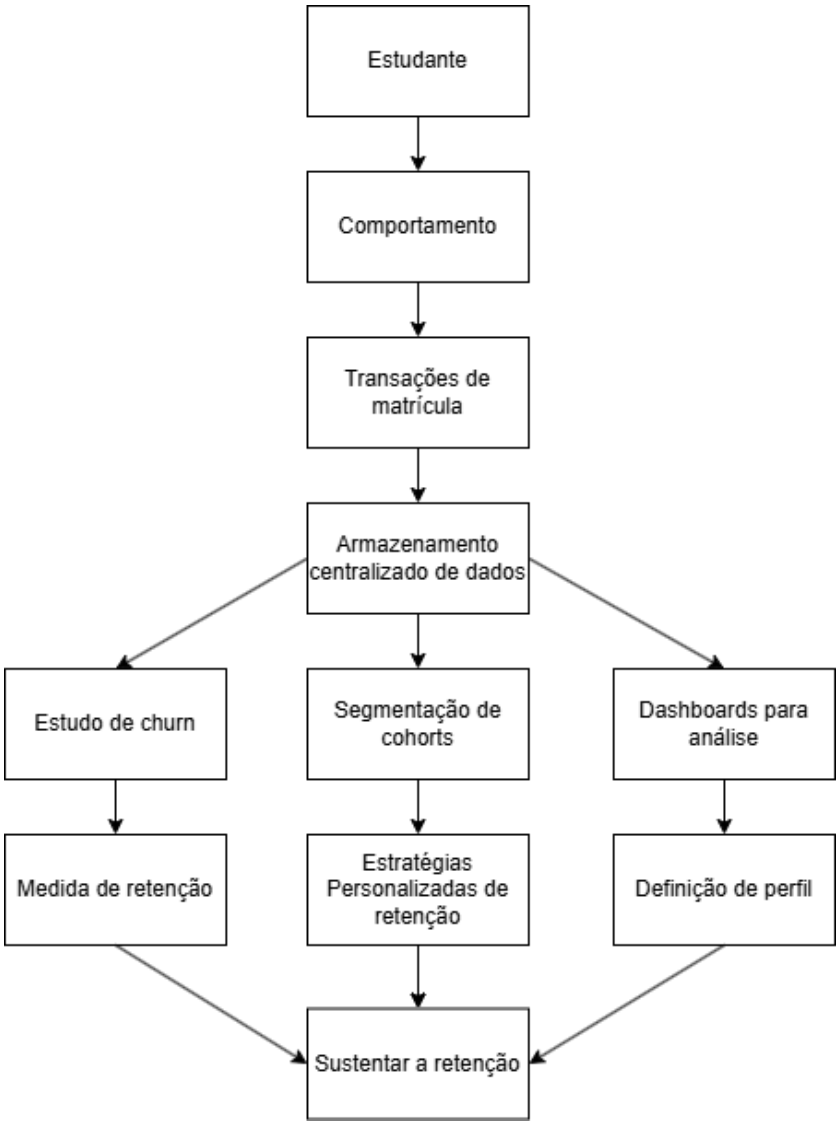
A análise de *cohort* permitiu segmentar os alunos com base em características como ano de ingresso e modalidade de ensino. Os resultados indicaram que os alunos que ingressaram no segundo semestre apresentaram uma taxa de retenção de XX%, enquanto aqueles que começaram no primeiro semestre apresentaram uma taxa de XX%. Essas informações são essenciais para adaptar as estratégias de marketing e retenção em períodos específicos.

4.4 *DASHBOARD* EM POWER BI

O *dashboard* criado no Power BI fornece uma visualização clara das taxas de retenção, categorizadas por curso e modalidade. Através de gráficos interativos, foi possível identificar quais cursos apresentaram maiores índices de cancelamento e quais fatores estavam associados a essas taxas. As principais visualizações incluem:

- **Gráfico de Cancelamento por Curso:** Demonstra a porcentagem de alunos que permaneceram matriculados em cada curso ao longo dos semestres.
- **Gráfico de Cancelamento por Motivo:** Apresenta os motivos mais comuns para o cancelamento, permitindo à instituição focar em áreas que necessitam de melhorias.

Figura 2 – Fluxo de processos



5 CONSIDERAÇÕES FINAIS

Este trabalho apresentou a importância do desenvolvimento de um pipeline de dados para a análise de *churn* por safra em instituições de ensino. Através da implementação de técnicas de análise de *cohort*, foi possível identificar padrões de comportamento que impactam diretamente a retenção de alunos.

Os resultados obtidos destacam a necessidade de um suporte acadêmico mais robusto e a implementação de estratégias específicas para diferentes turmas. A utilização de ferramentas de visualização de dados também se mostrou eficaz para a comunicação das métricas de retenção à equipe acadêmica.

Para trabalhos futuros, recomenda-se a continuidade da análise de dados e a inclusão de novas variáveis, como feedback dos alunos e taxas de participação em atividades extracurriculares. Essas informações podem contribuir ainda mais para a formulação de estratégias de retenção mais eficazes, melhorando a experiência acadêmica e a sustentabilidade financeira das instituições. Além disso, o uso de técnicas de aprendizado de máquina pode ser explorado para prever o *churn* com maior precisão, permitindo intervenções mais rápidas e personalizadas.

Em suma, a análise de *churn* é fundamental para que as instituições de ensino compreendam melhor o comportamento dos alunos, aumentem a taxa de retenção e, consequentemente, fortaleçam sua posição no mercado educacional. O sucesso da retenção de alunos não depende apenas de intervenções pontuais, mas de um compromisso contínuo com a qualidade da educação e com o apoio ao aluno ao longo de sua trajetória acadêmica.

6 REFERÊNCIAS BIBLIOGRÁFICAS

BARROS, J.; SILVA, F.; OLIVEIRA, R. Processo orientado a dados para a geração de modelos preditivos de evasão escolar. Instituto Federal do Rio Grande do Norte, 2020. Disponível em: https://repositorio.ifrn.edu.br. Acesso em: 3 nov. 2024.

CHEN, X.; WANG, Q.; LIU, Y. Cohort Analysis in Health Research: A Review and Perspective. Health Research and Policy Systems, v. 17, n. 1, p. 1-12, 2019.

ECKERSON, W. Performance Dashboards: Measuring, Monitoring, and Managing Your Business. 2. ed. New York: Wiley, 2010.

FADER, P. S. Customer Centricity: Focus on the Right Customers for Strategic Advantage. Upper Saddle River: Wharton Digital Press, 2012.

FARRIS, P. W.; BENDLE, N. T.; PFEIFER, P. E.; REIBSTEIN, D. J. Marketing Metrics: The Definitive Guide to Measuring Marketing Performance. Upper Saddle River: Pearson Education, 2010.

INMON, W. H. Building the Data Warehouse. 4. ed. Indianapolis: Wiley, 2005.

KIMBALL, R.; ROSS, M. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. 3. ed. Indianapolis: Wiley, 2013.

KUMAR, V.; PETERSEN, J. A. Statistical Methods in Customer Relationship Management. Hoboken, NJ: John Wiley & Sons, 2012.

KUMAR, V.; SHAH, D. Building and Sustaining Profitable Customer Loyalty for the 21st Century. Journal of Retailing, v. 80, n. 4, p. 317-330, 2004.

MARTINS, A.; SILVA, M. Aplicação de Modelos Preditivos na Análise de Churn no Setor de Saúde. Revista de Saúde e Tecnologia, v. 5, n. 3, p. 45-60, 2020.

RAMASWAMY, V.; OZCAN, K. Branding and Customer Engagement: A Cohort Analysis. Journal of Brand Management, v. 25, n. 2, p. 135-152, 2018.

SILVA, F.; BARROS, J.; OLIVEIRA, R. Análise de variáveis associadas à evasão escolar no IFPB. Instituto Federal da Paraíba, 2020. Disponível em: https://repositorio.ifpb.edu.br. Acesso em: 3 nov. 2024.

ZIKOPOULOS, P.; GATES, R.; MAGEE, D.; PARMAK, J. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. 1. ed. New York: McGraw-Hill, 2012.