Anais do Congresso Brasileiro de Iniciação Científica
(https://periodicos.unisanta.br/COB)

ISSN 2317-1316

RAG PARA PREVISÃO DE CHURN RATE: UMA APLICAÇÃO AGÊNTICA DO MODELO DE PREVISÃO ADASYN-PSO

RAG FOR CHURN RATE PREDICTION: AN AGENTIC APPLICATION OF AN ADASYN-PSO PREDICTION MODEL

André Leal Costa, Igor Silva de Carvalho, Fernando Andrade Lima Tavares, Nina Magalhães de Oliveira, Vagner dos Santos Macedo

Fatec Rubens Lara, Faculdade de Tecnologia, Curso de Ciência de Dados
E-mail para contato: igor.carvalho2@aluno.cps.sp.gov.br

RESUMO – Neste trabalho foi apresentado um ambiente RAG local com ferramentas voltadas para interações business. O objetivo foi integrar a previsão de churn a uma interface de chat, permitindo o acesso a documentos e dados de clientes. O sistema foi desenvolvido em Python, utilizando LangChain, LangGraph, Ollama, Pandas e Flask. Os testes foram realizados em um ambiente de playground, no qual foram disponibilizados arquivos para consulta. O chatbot demonstrou capacidade de pesquisar, indexar e ler documentos, contudo, modelos mais fracos necessitaram de maior direcionamento. Sendo assim, o projeto apresenta um sistema de apoio à consulta de informações e ao acesso à ferramenta de previsão de churn.

Palavras-chave: Previsão de churn; RAG; PSO; ADASYN; Chatbot.

ABSTRACT – This work presents a local RAG environment with tools for business interactions. The objective was to integrate churn prediction into a chat interface, providing access to documents and customer data. The system was developed in Python using LangChain, LangGraph, Ollama, Pandas and Flask. Tests were conducted in a playground environment containing files for consultation. The chatbot demonstrated the ability to search, index and read documents, although weaker models required more guidance. Thus, the project presents a system that supports information retrieval and access to a churn prediction tool.

Keywords: Churn prediction; RAG; PSO; ADASYN; Chatbot.

1 INTRODUÇÃO

Uma regra amplamente difundida no mundo corporativo é a de que ganhar um novo cliente é muito mais custoso do que manter aquele já existente. Por conta disso, independente do setor, as empresas sempre procuram investir extensivamente em medidas que minimizem o

ISSN 2317-1316

abandono de clientes, seja em estratégias de *marketing*, planos de fidelização ou melhoria dos produtos e serviços.

Esse abandono pode ser supervisionado pela métrica chamada de taxa de *chun*, a qual é responsável por quantificar os clientes que foram perdidos em determinado período. Essa métrica se torna ainda mais significativa quando o nicho em que a empresa está inserida é altamente competitivo, portanto, quanto mais fácil é a ação de abandono, maior é o cuidado com a análise de *churn*. É o caso das empresas de telecomunicação, por exemplo, conhecidas por fazerem parte de um cenário dinâmico e acirrado (MELIAN, 2022).

A fim de evitar que o *churn* aconteça, primeiro é necessário identificar quais são os clientes que tem maior probabilidade de abandonar a empresa, ou seja, prever quem são os possíveis *churners* e, conforme cada um, elaborar um plano de retenção. Para fazer essa previsão, é indispensável o uso de dados históricos de clientes, já que é a partir das características de *churners* passados que a classificação de novos é feita. Manter uma base de dados já é padrão para as empresas de telecomunicação, o que pode ser considerado inovador é a técnica utilizada em cima desses dados para prever o comportamento dos clientes.

Faris (2018) destaca que, porque existem diversos motivos pelos quais um cliente pode cancelar a assinatura, seja pelo favoritismo a outra empresa que oferece preços e/ou planos mais atrativos ou pelo simples fato que o cliente esqueceu de fazer a renovação, a previsão de churn se torna um problema de classificação. Melian *et al*. (2022) utilizaram dados de mais de 10 mil clientes de uma empresa de telecomunicação e investigaram diferentes técnicas estatísticas e de aprendizado de máquina para associar comportamentos ao ato de abandono. Segundo os autores, o problema de *churn* não depende de uma única variável, tornando a aplicação de algoritmos de previsão essencial já que, a olho nu, é muito mais difícil de identificar o cliente que está prestes a cancelar sua assinatura.

Apesar de ser um assunto que ganhou maior notoriedade nos últimos anos, algoritmos de aprendizado de máquina já fazem parte do contexto empresarial há muito tempo e operadoras, em vista da concorrência, possuem um papel proativo no estudo e desenvolvimento desses modelos. Os mais conhecidos, como *Support Vector Machines* (SVM), Naive Bayes, Árvores de Decisão, *K-Nearest Neighbors* (KNN), entre outros, já são amplamente estudados na área de previsão de *churn* e possuem níveis variantes de acurácia e complexidade (FARIS,

ISSN 2317-1316

2018). Apesar da alta escalabilidade e capacidade de interpretação, os algoritmos de aprendizado de máquina mais tradicionais possuem dificuldade em lidar com distribuições desbalanceadas de dados, o qual é o caso com os dados de *churners* no mercado de telecomunicação.

Por conta disso, são propostas diferentes abordagens de algoritmos de aprendizado de máquina que consigam superar o obstáculo dos dados desbalanceados. Uma delas são os algoritmos baseados em comportamento da natureza (*nature-inspired algorithms*), os quais apresentam bons resultados em casos mais complexos como é o caso da previsão de *churn*. Um exemplo desse tipo de algoritmo é chamado de *Particle Swarm Optimization* (PSO), o qual será explorado neste trabalho já que, baseado em estudos externos, apresenta resultados promissores.

Acredita-se ser importante desenvolver, não apenas o algoritmo, mas também uma maneira de apresentar os resultados computados para que a tomada de decisão seja a mais otimizada possível. Desse modo, em conjunto com o PSO, será elaborada uma interface baseada em modelos de *Large Language Models* (LLMs) que, com acesso a documentos externos, forneça a previsão de churn de forma rápida e integrada. Portanto, o produto visado é um sistema completo, o qual há o cálculo da probabilidade de *churn* e um *chatbot* especializado que transmita informações relativas a essa previsão.

Companhias de diversos setores já reconhecem os *chatbots* como o próximo grande meio de comunicação. Além de permitirem uma nova maneira de transmitir informações que seja mais rápida e disponível 24 horas por dia, *chatbots* significam um corte de custos significativos para as empresas, as quais podem redirecionar recursos - tanto financeiro quanto operacional para tarefas mais complexas (Akhtar *et al*., 2019).

Em vista disso, o desenvolvimento do *chatbot* neste trabalho procura aproveitar as vantagens dessa tecnologia e orientá-la ao funcionário da empresa de telecomunicação, e não ao cliente. Ou seja, deseja-se criar uma ferramenta de apoio ao funcionário para que este tenha acesso à análise de *churn* a qualquer momento, viabilizando uma melhora no serviço.

O *chatbot* aqui apresentado é classificado como sendo orientado a uma tarefa específica que, neste caso, é informar a probabilidade de *churn* a partir do algoritmo aplicado em cima do banco de dados de clientes. Assim, é um sistema de diálogo que busca manter uma conversa

ISSN 2317-1316

objetiva e focada em um assunto específico, sem que haja necessidade de muitas trocas entre ele e o usuário. O seu funcionamento é baseado no *framework Retrival Augmentation Generation* (RAG).

Desde a sua introdução ao mercado, os grandes modelos de linguagem (LLMs) representaram o ápice da transformação digital atual e revolucionaram diversos setores através da análise de grandes volumes de dados e da geração de texto semelhante àquele produzido por um ser-humano. O uso desses modelos nas empresas está rapidamente se tornando o novo paradigma de negócio, desde a concepção de uma ideia até sua execução.

Entretanto, de acordo com Arslan *et al.* (2024), embora os LLMs apresentem grande capacidade de processamento e geração de linguagem natural, eles podem enfrentar dificuldades em consultas específicas de determinados domínios. Isso se torna especialmente evidente para as empresas já que, porque os modelos são treinados com dados disponíveis na Internet, quando apresentada uma questão que necessita do conhecimento de documentos internos há maiores chances de o LLM alucinar e fornecer uma resposta incorreta.

Como uma maneira de solucionar o problema da alucinação e aumentar a precisão e a relevância das respostas produzidas, foi introduzido o RAG, uma abordagem que integra mecanismos de recuperação de informações externas ao processo de geração de texto (ARSLAN *et al*., 2024).

De forma simplificada, um sistema RAG envolve uma etapa de recuperação de documentos externos fornecidos e uma etapa de geração. Gao *et al*. (2026) descrevem esse processo a partir de três componentes fundamentais: indexação dos documentos, recuperação dos trechos mais semelhantes à consulta e geração da reposta pelo modelo de linguagem.

Os três componentes descrevem a base de qualquer sistema RAG, entretanto, a arquitetura tradicional pode ser ampliada para atender tarefas mais complexas e, desse modo, foram criados diversos tipos desse sistema. Um deles é chamado de Modular RAG, no qual os sistemas são organizados em módulos, submódulos e operadores, permitindo maior flexibilidade na composição dos fluxos de recuperação e geração (GAO *et al*., 2026).

Em vista disso, a abordagem RAG em sistemas internos de empresas pode ser considerada de extrema relevância, pois permite o uso dos benefícios das LLMs em seus diversos setores e

ISSN 2317-1316

domínios, incluindo aquele responsável por analisar dados de clientes e avaliar a probabilidade de *churn*.

2 MATERIAIS E MÉTODOS

Neste projeto foi desenvolvido um sistema RAG voltado à previsão de *churn*, por meio de uma pesquisa de caráter experimental, com desenvolvimento de um protótipo e avaliação empírica de seu funcionamento. Todo o projeto foi montado na linguagem Python, tanto para o tratamento dos dados quanto para o modelo de previsão e o funcionamento do agente.

Para a criação do agente, foram utilizados os módulos *LangChain* e *LangGraph*. O *LangChain* é responsável pela integração do modelo de linguagem com as ferramentas disponíveis, enquanto o *LangGraph* dá suporte ao fluxo de execução e à memória da conversa. Já a execução local do modelo é feita por meio do Ollama, sendo necessário selecionar um dos modelos previamente baixados no computador de uso. É a partir dessa integração que o chatbot pode consultar arquivos, analisar bases de dados e acionar a previsão de churn conforme a solicitação do usuário.

Para os documentos em PDF, TXT e *Markdown*, o conteúdo é dividido em trechos e indexado em uma base vetorial. O modelo *nomic-embed-text* é utilizado para gerar os *embeddings* desses trechos, permitindo que o sistema encontre conteúdos semelhantes à pergunta realizada. Assim, o RAG tem acesso às informações dos documentos disponibilizados para elaborar suas respostas. Já para o processamento dos arquivos CSV, utilizou-se da biblioteca Pandas, responsável pela leitura dos dados e pelas operações de consulta e análise das tabelas.

No que diz respeito à previsão de *churn*, foi utilizada a base “IBM Telco Customer Churn” nos testes iniciais, composta por dados históricos de clientes de telecomunicação. A variável *Churn* é utilizada para identificar os clientes que abandonaram o serviço. Antes do treinamento, os dados passam por pré-processamento, separação entre treinamento e teste e normalização das variáveis, procurando reduzir o efeito das diferenças de escala.

Juntamente com o modelo de previsão, optou-se pelo uso do *Adaptive Synthetic Sampling* (ADASYN), uma técnica de sobreamostragem que gera exemplos sintéticos da classe

ISSN 2317-1316

minoritária a partir de exemplos próximos. O motivo pelo qual essa técnica foi utilizada é explicado pelo desbalanceamento dos dados, já que há uma concentração maior de clientes que permanecem no serviço da empresa do que de abandono, a geração dos exemplos é concentrada nas regiões em que há maior dificuldade de distinguir as classes, sendo aplicada aos dados de treinamento.

Após esse processo, utiliza-se o *Particle Swarm Optimization* (PSO) para buscar combinações de pesos atribuídos às variáveis de entrada e de quantidade de neurônios da rede *Random Weight Network* (RWN). Cada combinação é representada por uma partícula, que atualiza sua posição considerando a melhor solução encontrada individualmente e pelo grupo. A RWN é responsável pela classificação, utilizando pesos de entrada definidos aleatoriamente e pesos de saída ajustados a partir dos dados de treinamento. Para a avaliação do classificador, utiliza-se validação cruzada.

A ferramenta de previsão pode ser acionada pelo agente para realizar a previsão de *churn* de um cliente específico e apresentar o resultado ao usuário por meio do chat. O modelo é treinado sob demanda, podendo ser salvo e utilizado novamente nas consultas posteriores.

Por fim, para a renderização em ambiente web, em formato de chat, foi utilizado o Flask, que é um *microframework* de desenvolvimento web para Python. Segundo Grinberg (2018), sua filosofia minimalista torna o Flask uma escolha adequada para projetos que demandam flexibilidade na arquitetura. No contexto deste projeto, o *framework* atua como camada de interface entre o usuário e o sistema RAG, sendo responsável por receber requisições do *chat*, encaminhá-las ao agente e retornar as respostas produzidas. O HTML e o CSS foram utilizados para a estrutura e a apresentação visual da interface, enquanto o JavaScript permite a interação com o chat e o envio das mensagens e dos arquivos.

Os modelos utilizados foram avaliados de maneira empírica a partir da criação de um ambiente de *playground* para a realização dos testes. O diretório de teste limitava o acesso do *bot* a um ambiente controlado. Neste ambiente, criou-se ou transferiu-se arquivos com os formatos suportados pelo RAG para verificar sua capacidade de responder aos prompts passados. Durante os testes, foi observado se o *bot* conseguia localizar os arquivos, indexar e ler os documentos e utilizar as ferramentas disponíveis, além de verificar quando era necessário

ISSN 2317-1316

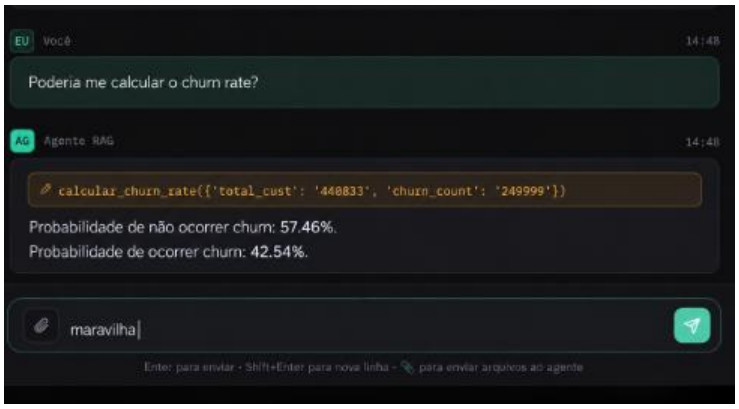
fornecer direcionamento adicional. O modo de *debugging* foi utilizado para acompanhar a execução e analisar dificuldades na ordem das operações.

3 RESULTADOS E DISCUSSÃO

O sistema desenvolvido apresenta a capacidade de realizar a previsão individual de *Churn Rate* por meio de um modelo ADASYN-PSO treinado sob demanda. A ferramenta de previsão é acionada pelo agente a partir da solicitação do usuário, e o resultado é apresentado na interface de chat.

No teste apresentado, a previsão foi realizada para todos os clientes presentes na base de dados "*Customer Churn Dataset*" extraída do Kaggle, mas a modelagem foi feita com finalidade de utilizar qualquer base de dados, visto que não foi passado um perfil de cliente, isso permitiu observar o funcionamento da ferramenta sobre o conjunto de clientes disponível, o resultado dessa execução pode ser visto na Figura 1. Além da análise de toda a base, o sistema também permite realizar a previsão para casos específicos, direcionando a consulta a um cliente selecionado.

Figura 1: Resultado da previsão de churn para os clientes da base de dados.



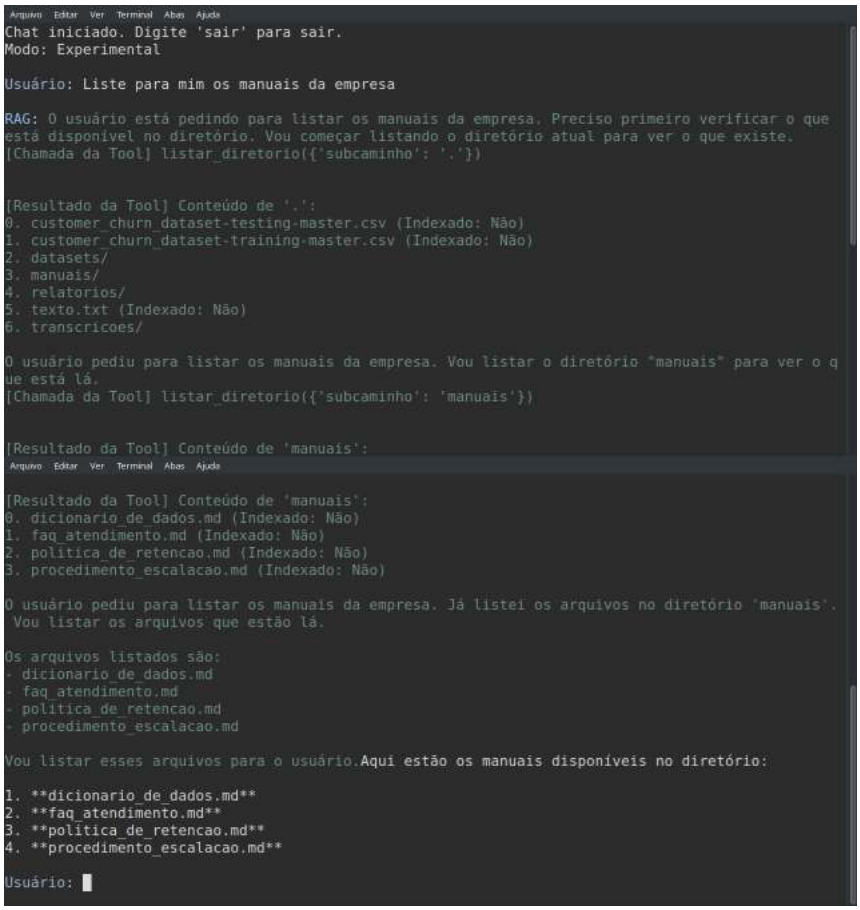
Fonte: elaboração própria (2026).

O *chatbot* demonstrou boa capacidade de pesquisar, indexar, e ler documentos. Para modelos mais potentes é desnecessária a utilização de múltiplos prompts para responder a pergunta-alvo: o bot é capaz de responder, dadas as ferramentas e a familiaridade de localização com o ambiente de playground conforme a Figura 2.

ISSN 2317-1316

A Figura 2 demonstra um exemplo em que o agente localiza e lista os manuais disponíveis no diretório, sem que o usuário precise fornecer direcionamento adicional.

Figura 2: Exemplo de resposta sem auxílio (modo Debugging)



Fonte: elaboração própria (2026).

Contudo, para modelos mais fracos, foi necessário constantemente direcionar o RAG para responder os *prompts*, em alguns casos sendo necessário fornecer diretamente o diretório de arquivos relevantes ou reforçar a ordem de operações de uso de tools, por exemplo, primeiro indexar e depois abrir um arquivo específico.

Outro problema pôde ser analisado por meio do modo de *debugging*, expondo os prompts internos do RAG (os seus “pensamentos”). Lendo-se o fluxo de raciocínio da IA, é notável a grande quantidade de recursão, exclusões e confusões acerca do que tem de ser feito e em que ordem.

ISSN 2317-1316

Melhoras futuras poderiam utilizar o fluxo de raciocínio do *bot* para criar melhorias no prompt de sistema e na descrição das ferramentas exposta ao RAG, além de criar um fluxo mais sistemático e simples de ser utilizado para leitura de arquivos.

4 CONCLUSÃO

Neste trabalho foi apresentado um ambiente de *chat* para acessar um RAG local com ferramentas voltadas para interações *business,* que *p*or meio de uma interface web feita com Flask, o usuário pode conversar com o *chatbot* para consultar documentos, analisar *datasets* CSV e solicitar a previsão de *churn*, realizada por um modelo ADASYN-PSO treinado sob demanda, durante os testes, o sistema retornou a previsão para a base de clientes e demonstrou capacidade de pesquisar, indexar e ler documentos, mas para modelos mais fracos, ainda foi necessário direcionar o *bot*, principalmente para localizar os arquivos e seguir a ordem de uso das ferramentas. Por conta dessas dificuldades, os possíveis próximos passos são otimizar o *prompt* de sistema e testar novos modelos de agentes.

5 DECLARAÇÃO DE USO DE IA

Durante a elaboração deste trabalho, utilizou-se a ferramenta de IA Claude com o modelo Sonnet 5.5 para auxiliar no desenvolvimento do projeto e na correção de escrita. O conteúdo gerado foi revisado, editado e validado pelos autores, que assumem total responsabilidade pelo texto final.

6 REFERÊNCIAS

AKHTAR, Mubashra; NEIDHARDT, Julia; WERTHNER, Hannes. The potential of chatbots: analysis of chatbot conversations. In: IEEE 21st CONFERENCE ON BUSINESS INFORMATICS (CBI), 21., 2019, Moscow, Rússia. Anais eletrônicos... [S. l.]: IEEE, 2019. DOI: https://doi.org/10.1109/cbi.2019.00052. Acesso em: 23 set. 2026.

ALI, Muhammad; REHMAN, Aziz Ur; HAFEEZ, Shamaz; ASHRAF, Muhammad Usman. Prediction of churning behavior of customers in telecom sector using supervised learning techniques. In: INTERNATIONAL CONFERENCE ON COMPUTER, CONTROL, ELECTRICAL, AND ELECTRONICS ENGINEERING (ICCCEEE), [S.I], 2018, Khartoum,

ISSN 2317-1316

Sudão. Anais eletrônicos... [S. l.]: IEEE, 2018. DOI: https://doi.org/10.1109/iccceee.2018.8515857. Acesso em: 23 set. 2026.

ARSLAN, Muhammad; GHANEM, Hussam; MUNAWAR, Saba; CRUZ, Christophe. A Survey on RAG with LLMs.In: 28th INTERNATIONAL CONFERENCE ON KNOWLEDGE-BASED AND INTELLIGENT INFORMATION & ENGINEERING SYSTEMS (KES 2024), 28., 2024, Seville, Espanha. Anais eletrônicos... [S.I]: ScienceDirect, 2024. p.3781-3790. Disponível em: www.sciencedirect.com. Acesso em: 23 set. 2026.

FARIS, Hossam. A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors. Information, v. 9, n. 11, p. 288, 2018. DOI: https://doi.org/10.3390/info9110288. Acesso em: 23 set. 2026.

GAD, Ahmed G. Particle swarm optimization algorithm and its applications: a systematic review. Archives of Computational Methods in Engineering, v. 29, p. 2531-2561, 2022. DOI: https://doi.org/10.1007/s11831-021-09694-4. Acesso em: 23 set. 2026.

GAO, Yufan; XIONG, Yun; WANG, Meng; WANG, Haofen. Modular RAG: Transforming RAG Systems into LEGO-like Reconfigurable Frameworks. arXiv, [S.I], 2024. Disponível em: https://doi.org/10.48550/arXiv.2407.21059. Acesso em: 23 set. 2026.

GRINBERG, Miguel. Flask Web Development: developing web applications with Python. 2. ed. Sebastopol: O'Reilly Media, 2018. Acesso em: 23 set. 2026.

HE, Haibo; BAI, Yang; GARCIA, Edwardo A.; LI, Shutao. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: IEEE INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN), 18., 2008, Hong Kong, China. Anais eletrônicos... [S. l.]: IEEE, 2008. p. 1322-1328. Disponível em: https://ieeexplore.ieee.org. Acesso em: 23 set. 2026.

MELIAN, Denisa; DUMITRACHE, Andreea; STANCU, Stelian; NASTU, Alexandra. Customer churn prediction in telecommunication industry: a data analysis techniques approach. Postmodern Openings, v. 13, n. 13, p. 78-104, 2022. DOI: https://doi.org/10.18662/po/13.1Sup1/415. Acesso em: 23 set. 2026.