PREDIÇÃO DO TURNOVER EM DESENVOLVIMENTO DE SOFTWARE UTILIZANDO DADOS DO LINKEDIN E MACHINE LEARNING

PREDICTING TURNOVER IN SOFTWARE DEVELOPMENT USING LINKEDIN AND MACHINE LEARNING DATA

José Diôgo Lima da Silva
Universidade Federal de Minas Gerais - UFMG
E-mail: josediogolima59@gmail.com

Matheus Philippe da Silva
Universidade Federal de Minas Gerais - UFMG
E-mail: doumarys2@gmail.com

Recebido em 6 de fevereiro de 2026
Publicado em 19 de maio de 2026

Resumo

O crescimento do trabalho remoto e o uso do LinkedIn ampliaram as oportunidades para profissionais de tecnologia, aumentando a rotatividade (turnover) e os custos para as empresas. Este estudo buscou prever o tempo de permanência de desenvolvedores de software por meio de modelos de machine learning. A pesquisa, quantitativa e aplicada, utilizou dados de 2.367 perfis coletados no LinkedIn por web scraping, dos quais 659 foram padronizados para treinar e testar modelos de regressão Support Vector Machine (SVM). Foram comparadas uma implementação manual com *gradient descent* e outra com a biblioteca Python sklearn, avaliadas por validação cruzada usando MSE**,** MAE e R^2. A média de permanência no emprego subsequente foi de 2 anos e 2 meses, com 75% dos profissionais permanecendo menos de 3 anos. Ambos os modelos tiveram baixo poder preditivo devido ao desbalanceamento e à limitação das variáveis. Conclui-se que o uso de dados do LinkedIn é promissor, mas requer bases ampliadas, com variáveis adicionais, e integração com outras fontes, como o Glassdoor.

Palavras-chave: LinkedIn; Rotatividade; Turnover.

Abstract

The growth of remote work and the use of LinkedIn have expanded opportunities for technology professionals, increasing employee turnover and costs for companies. This study aimed to predict the length of tenure of software developers using machine learning models. The research, quantitative and applied in nature, used data from 2,367 profiles collected from LinkedIn through web scraping, of which 659 were standardized to train and test Support Vector Machine (SVM) regression models. A manual implementation using gradient descent was compared with an implementation using the Python Sklenar library, both evaluated through cross-validation using MSE, MAE, and R^2. The average tenure in the subsequent job was 2 years and 2 months, with 75% of professionals remaining for less than 3 years. Both models showed low predictive power due to data imbalance and the limited set of variables. It is concluded that the use of LinkedIn data is promising but requires larger datasets, additional variables, and integration with other sources, such as Glassdoor.

Keywords: LinkedIn; Employee turnover; Turnover.

1 INTRODUÇÃO

A expansão da internet de alta qualidade ao redor do mundo, particularmente em países

emergentes e subdesenvolvidos, impulsionou a adoção global do trabalho remoto. Como resultado, profissionais de tecnologia passaram a ter acesso a uma ampla variedade de oportunidades de trabalho remoto, o que contribuiu para uma maior rotatividade desses profissionais nas empresas. A rotatividade de proficionais, ou turnover, é definida como o movimento de entrada e saída de trabalhadores em uma empresa. Esse movimento é calculado dividindo a quantidade de demissões pelo total de funcionários da empresa (Alves, 2023).

O turnover pode ser classificado como voluntário ou involuntário. O voluntário é caracterizado pela decisão por parte do trabalhador em deixar o emprego, já a involuntária representa situações onde os trabalhadores são obrigados de deixar a empresa, essa situações podem ser demissão, doenças, morte ou aposentadoia (Sallaberry, 2021). O turnover existe em empresas de todas as áreas, uma pesquisa feita pelo Ministério do Trabalho e Emprego entre novembro de 2023 e abril de 2024 foram identificados motivos por trás do pedido de demissão por parte do trabalhador. Os três motivos mais frequentes foram outro emprego em vista, baixo valor do salário e a falta de reconhecimento no trabalho (MTE, 2024)

Ao agrupar os pedidos de demissão pela atividade exercida e analisar qual a porcentagem destes trabalhadores indicaram como motivo do pedido de demissão o fato de terem um novo emprego em vista é possível identificar quais atividades estão mais aquecidas no mercado. De todos os pedidos de demissão onde os trabalhadores atuavam na área de tecnologia da informação 59% tiveram como motivo um novo emprego em vista. (MTE, 2024)

De todas as áreas analisadas a área de tecnologia da informação foi a que apresentou um maior percentual de trabalhadores pedindo demissão por já possuir uma nova oferta de trabalho, o que mostra o quão aquecido está o mercado. (MTE, 2024). Dentre os trabalhadores da área de tecnologia da informação que pediram demissão com um novo trabalho em vista a insatisfação com o salário foi um fator bastante mencionado.

Tendo em vista os dados da pesquisa do MTE (2024) é que empresas do setor de tecnologia da informação enfrentam índices elevados de turnover com o mercado aquecido resultando em custos significativos. Isso porque, a reposição da força de trabalho envolve custos, os quais vão além das etapas de recrutamento e incluem períodos de treinamento e adaptação dos novos colaboradores ao ambiente organizacional. O cenário mais desfavorável ocorre quando um profissional, ainda em fase de adaptação, opta por deixar a empresa, resultando em desperdício do investimento realizado para integrar aquele colaborador, sem que ele tenha permanecido o suficiente para gerar retorno para a organização.

Durante o processo seletivo, profissionais de Recursos Humanos implementam dinâmicas e entrevistas para mitigar a possibilidade de que o desligamento precoce ocorra. Diante da relevância exposta, do rápido avanço da tecnologia para recrutamento e seleção através de mídias digitais, é que esse trabalho tem como objetivo analisar qual é o período de permanência no emprego para profissionais de desenvolvimento de softwares baseado em dados obtidos do Linkedin.

O Linkedin é uma plataforma social de profissionais, dos quais dispõem seus currículos.

Nessa é possível realizar networks, conhecer empresas e profissionais dos quais buscam reposicionamento no mercado. A importância da plataforma pode ser vista ao analisar os resultados de uma pesquisa feita em 2015 com mais de dois mil recrutadores americanos, cerca de um terço dos empregadores disseram que não chegariam a realizar uma entrevista com um candidato que não possuísse informações online sobre histórico profissional, dito isso, profissionais de diversas áreas percebem a importância de manter seus dados profissionais atualizados em plataformas como o Linkedin (Cho et al., 2021).

Nem todos os profissionais que criam perfis e se mantêm ativos no Linkedin estão à procura de um novo emprego, em uma pesquisa com um grupo de profissionais que abriram conta na plataforma, 90% disseram que se juntaram à plataforma para conhecer quais vagas de trabalho estão sendo criadas, levando em consideração o the unfolding model, onde uma pessoa que está motivada a obter uma evolução na sua carreira profissional e está imersa em um mercado com alternativas ela provavelmente apresentará intenção de deixar a empresa para conseguir essa evolução (Cho et al., 2021). Tendo em vista o the unfolding model, é perceptível o impacto da existência de plataformas como o Linkedin nos indicadores de turnover das empresas.

O presente trabalho visa contribuir com o desenvolvimento de soluções que auxiliem a equipe de RH em contratações mais assertivas têm impacto direto na redução da taxa de turnover nas empresas. Assim os resultados deste trabalho irão beneficiar empresas que buscam soluções para contratações mais assertivas prevendo o potencial impacto da contratação de um novo funcionário nos indicadores rotatividade ao contar com um modelo de predição do tempo de permanência de um candidato após sua contratação que auxiliará equipes de RH a contratar candidatos com maiores chances de permanecer por um longo período na empresa.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Rotatividade de funcionários ou *turnover*

A rotatividade de funcionários em uma empresa gera custos com a reposição do posto de trabalho e riscos para a operação empresarial. Após um funcionário pedir demissão, o processo de reposição não termina no momento da contratação de um novo trabalhador, pois este demanda um certo tempo de treinamento para atingir as expectativas do cargo. Assim as empresas que se antecipam adotam medidas que impactam positivamente nos indicadores de turnover se beneficiam com menor custos associados ao processo de recrutamento e reposição da força de trabalho aumentando sua competitividade a longo prazo em decorrência de uma força de trabalho mais estável (AKASHEH , 2024).

O desenvolvimento e a implementação de estratégias de RH nas empresas trazem impactos positivos nos indicadores de rotatividade de funcionários. Isso é o que aponta a pesquisa desenvolvida por Matos (2022), que avaliou o impacto da implementação de um RH estratégico no meio organizacional.

Segundo Matos (2022), a Quarta Revolução Industrial tirou o RH, assim como outras

áreas, da sua zona de conforto com a chegada de ferramentas baseadas em tecnologia. Essas ferramentas podem ser utilizadas para otimizar e melhorar o trabalho do RH em contratações assertivas, além de ajudar a enfrentar desafios como a rotatividade de funcionários.

A pandemia de COVID-19 evidenciou a importância de ferramentas e estratégias de RH citadas por Matos (2022) em seu trabalho, como por exemplo o Onboarding virtual de novos funcionários. Akdur et al. (2024) desenvolveram um trabalho para entender os impactos do processo de Onboarding virtual de profissionais de desenvolvimento de software durante a pandemia de COVID-19 na intenção de mudança de empresas por parte dos desenvolvedores.

Em seu trabalho Akdur et al. (2024), foi utilizado o Framework de Integração e Retenção Virtual (FIRV), que é um modelo desenvolvido para entender e melhorar o processo de integração virtual de novos funcionários e a retenção de talentos, especialmente em contextos como o da pandemia de COVID-19. Foram aplicados dados coletados por meio de um questionário com 150 pessoas, sendo 121 dessas pessoas identificadas como homens, com idades entre 18 e 35 anos.

Os resultados obtidos no trabalho indicam que o sucesso do onboarding virtual não possui uma relação direta com a intenção do desenvolvedor de trocar de empresa. No entanto, o sucesso do onboarding virtual impacta diretamente a satisfação no trabalho do desenvolvedor e a qualidade das relações no local de trabalho. Esses dois fatores, por sua vez, podem reduzir as chances de o desenvolvedor desejar trocar de empresa.

Embora o trabalho Akdur et al. (2024) possa perder aplicabilidade à medida que nos distanciamos do período pós-COVID-19, ele oferece oportunidades para expandir a investigação e a compreensão dos impactos de diferentes fatores nos índices de rotatividade de desenvolvedores de software. Os profissionais de software são alguns dos que têm maiores níveis de intenção de mudar de empresa em comparação com outros profissionais; no entanto, há poucos modelos teóricos que ajudem as organizações a lidar com a alta rotatividade desses profissionais (SHARMA; STOL, 2020).

Sharma e Stol (2020) desenvolveu um modelo teórico focado em profissionais de desenvolvimento de software para aprimorar os processos de integração. Foi identificado que o suporte contínuo, aliado a um ambiente propício onde os novos funcionários podem fazer perguntas sem sentir vergonha, é fundamental para o sucesso da integração, reduzindo assim a intenção de troca de empresas pelos novos colaboradores. Isso se deve ao fato de que, nas organizações de software, as tecnologias estão em constante evolução, e métodos e práticas tornam-se obsoletos continuamente.

2.2 Rotatividade de funcionários e aprendizado de máquina

A aplicação de modelos de aprendizado de máquina para a resolução de problemas tem se difundido nos últimos anos. Akasheh (2024) desenvolveu uma pesquisa de revisão da literatura sobre estudos referentes à utilização de técnicas de machine learning para a previsão da rotatividade de funcionários nas empresas. Segundo os pesquisadores Akasheh (2024)

Akasheh (2024) chegam à conclusão de que é de extrema importância que as empresas tentem identificar as causas da rotatividade de funcionários, pois isso pode impactar a satisfação dos clientes e até mesmo levar à perda de conhecimento, habilidades e memória institucional, afetando o desempenho geral da organização.

Uma das preocupações iniciais em relação ao treinamento de um modelo de aprendizado de máquina é a qualidade dos dados utilizados como dados de treino do modelo. O presente trabalho dedica grande parte dos esforços na construção de uma base de dados; no entanto, outra solução é a utilização de bases existentes, como foi feito nos trabalhos de Ajit (2016), Chakraborty (2021) e Atef, S Elzanfaly e Ouf (2022). Mesmo tendo acesso a uma base de dados padronizada, Ajit (2016) e Chakraborty et al. (2021) destacam os ruídos nos dados como um grande desafio na construção de modelos preditivos. Alguns campos precisaram ser inferidos, pois estavam em branco. Já Atef, S Elzanfaly e Ouf (2022) não teve grandes problemas com a base de dados, pois escolheu utilizar uma base disponibilizada pela IBM para o desenvolvimento de aplicações de machine learning, contendo 1470 registros de profissionais de diferentes áreas padronizados.

Em sua pesquisa, Ajit (2016) comparou o desempenho do XGBoost em relação a seis classificadores supervisionados tradicionais. O estudo utilizou uma base de dados proveniente do Sistema de Informação de Recursos Humanos de um varejista global com dados de uma força de trabalho distribuída pelos Estados Unidos, composta por pessoas em diferentes estágios de suas carreiras, com diferentes níveis de desempenho e remuneração, e de diferentes origens.

Neste estudo, foram utilizados diversos algoritmos de aprendizado de máquina, incluindo Regressão Logística, Máquinas de Vetores de Suporte (SVM), Florestas Aleatórias e Redes Neurais Artificiais, para construir modelos preditivos com base em um conjunto de dados realista de uma empresa. O conjunto de dados incluiu variáveis como satisfação no trabalho, salário, tempo de serviço, histórico de promoções, entre outros fatores. Após a construção dos modelos, as técnicas de validação cruzada e métricas como precisão, recall e F1-score foram usadas para avaliar o desempenho de cada modelo.

Os resultados indicaram que as Florestas Aleatórias e as Redes Neurais Artificiais apresentaram o melhor desempenho na previsão de rotatividade de funcionários, com níveis elevados de precisão e *recall*. A análise dos fatores de importância das variáveis revelou que a satisfação no trabalho e o histórico de promoções foram os principais preditores da rotatividade. Com base nesses resultados, recomenda-se que as empresas invistam em políticas de retenção que abordem diretamente esses fatores, a fim de reduzir a rotatividade e melhorar a retenção de talentos.

Dos três trabalhos analisados no referencial teórico, o trabalho desenvolvido por Atef, Elzanfaly e Ouf (2022) foi o que contou com uma base de dados completa, sem grandes problemas de padronização e informações, algo que o presente trabalho não consegue extrair com tanta precisão utilizando apenas os dados do LinkedIn, como o tempo gasto no trânsito para o deslocamento ao trabalho. O estudo de Atef, S Elzanfaly e Ouf (2022) aponta que o tempo gasto no deslocamento para o trabalho, a baixa idade, o salário e o estado civil solteiro são

fatores que tornam os trabalhadores mais propensos a trocar de emprego.

Em seu trabalho, Atef, S Elzanfaly e Ouf (2022) comparou dois modelos de aprendizado de máquina: Random Forest (RF) e K-Nearest Neighbors (KNN). A pesquisa revelou que o modelo KNN apresentou melhor desempenho preditivo para estimar a probabilidade de rotatividade de funcionários antes da contratação. Esse modelo foi considerado mais eficaz para auxiliar os gestores de RH durante o processo de recrutamento, proporcionando uma ferramenta útil para prever a probabilidade de um funcionário deixar a empresa.

Todos os trabalhos analisados usaram bases de dados prontas poupando tempo de desenvolvimento, porém a construção de uma base de dados permitiria a identificação de novas características para serem utilizadas no processo de treinamento e facilitaria o desenvolvimento de novas pesquisas aprimorando não os modelos apresentados e a metodologia de coleta de dados para a construção da base.

3 METODOLOGIA

Este estudo adota uma abordagem quantitativa, caracterizada como uma pesquisa aplicada, com o objetivo de prever o tempo de permanência de candidatos após a contratação. De acordo com Yin (2015), pesquisas aplicadas são especialmente adequadas para estudos que buscam resolver problemas práticos com aplicações diretas em contextos específicos. Nesse caso, o foco é prever a permanência de colaboradores no setor de tecnologia.

Quanto à estratégia de investigação, este estudo utiliza um desenho experimental, conforme recomendado por Creswell (2014) para pesquisas que buscam estabelecer relações causais ou explorar fenômenos em condições controladas. A escolha de um modelo experimental é relevante neste trabalho, pois permite investigar os fatores que influenciam a retenção de profissionais de tecnologia, facilitando o teste de alternativas de coleta e tratamento de dados, além de contribuir para o treinamento de modelos de machine learning.

Profissionais da área de desenvolvimento de software foram selecionados como foco desta pesquisa devido ao seu uso ativo da rede social LinkedIn, o que permite o acesso a dados relevantes sobre histórico profissional e tendências de permanência no setor. A coleta de dados foi realizada a partir de perfis públicos de LinkedIn, conforme os termos de uso da plataforma, visando a análise de variáveis como tempo de permanência em empresas anteriores, frequência de mudanças de emprego e interações na plataforma.

As seções subsequentes descrevem detalhadamente os procedimentos adotados para coleta, tratamento e análise dos dados, assim como os métodos específicos para treinamento do modelo de machine learning, objetivando a predição de tempo de permanência com base nos dados coletados. Para que este objetivo seja alcançado foram identificados os seguintes objetivos específicos: Montar uma base de dados com dados coletados de perfis públicos do Linkedin, Definir e Implementar um modelo de machine learning e comparar modelo implementado com implementações disponíveis em bibliotecas públicas.

3.1 Coleta de dados

Para a coleta de dados, foram desenvolvidos códigos em Python utilizando técnicas de web scraping para extrair dados públicos de usuários do LinkedIn, já que a plataforma não fornece nenhum tipo de base de dados ou API pública para coleta em massa.

Python é uma linguagem de programação criada em 1991 de alto nível, reconhecida por ser uma linguagem fácil de aprender. Ela é amplamente utilizada por desenvolvedores e cientistas da área de análise e aprendizado de máquina principalmente pelo fato de que a maior parte das bibliotecas de machine learning disponíveis no mercado serem desenvolvidas com base na linguagem (RASCHKA, 2020).

Uma API pública é uma interface de programação disponibilizada por uma organização para que desenvolvedores externos possam interagir com dados e serviços da organização. Como o LinkedIn não possui uma API pública, o acesso ao site e aplicativo é a única forma de interagir com os dados e serviços da plataforma. Assim foram utilizadas técnicas de web scraping que consiste em extrair dados de páginas web de forma automatizada, todas as páginas web consistem em um arquivo HTML, que em conjunto com outros componentes, orientam o browser como exibir os dados e quais ações serão tomadas. Todos os dados exibidos em um site estão contidos em um marcador. Para coletar dados de um uma página web com técnicas de web scraping é preciso identificar um padrão nos marcadores da página relacionando a um dado.

Os dados foram coletados durante o mês de junho de 2024 e houve a intervenção manual apenas para a seleção de filtros aplicados para a pesquisa dos perfis que foram aplicados para reduzir a influência das conexões do perfil utilizados para coleta de dados visto que a ferramenta de busca de perfis do LinkedIn é fortemente influenciada pela rede de conexões de uma conta, o que tornou necessário definir algumas métricas para selecionar quais perfis seriam incluídos na base de dados.

3.1.1 Definição dos filtros de pesquisa no LinkedIn

Nesta seção, definimos os filtros utilizados para selecionar perfis com o intuito de minimizar a influência das conexões do perfil usado para coleta de dados. A seguir, apresentamos os filtros aplicados:

- Localização: Brasil
- Filtro de Conexão de Terceiro Grau: Considera apenas perfis que são conexões de terceiro grau com o perfil usado para a coleta de dados. Esse filtro assegura que a influência direta das conexões do perfil de coleta seja minimizada, selecionando perfis mais distantes e, portanto, menos suscetíveis a viés imediato.
- Campo de busca: Considera perfis que aparecem no resultado da pesquisa onde o campo de busca contém o texto "software developer". O termo "software developer" foi definido com o intuito de obter uma quantidade maior de dados com informações relevantes, uma vez que este é um perfil de profissional que frequentemente utiliza o LinkedIn para busca de empregos.

- Empresa onde perfil já tenha trabalhado: O perfil precisa ter trabalhado em pelo menos uma das empresas presentes em uma lista de empresas que será definida na subseção 3.1.2. Este filtro foi aplicado para diminuir a influência das empresas mais frequentes nos perfis com grau de conexão mais próximo do perfil usado para coleta de dados.

3.1.2 Seleção de empresas anteriores dos trabalhadores

Foi definido um grupo de empresas com base no ranking do Great Places to Work, um estudo anual que ranqueia as melhores empresas para se trabalhar. Entre os anos de 2021 e 2023, o Great Places to Work divulgou um ranking das 10 melhores grandes empresas com mais de dez mil funcionários para se trabalhar. A Tabela 1 mostra a frequência com que cada empresa apareceu no ranking.

Tabela 1 – Frequência de listagem das 10 melhores grandes empresas com mais de 10 mil funcionários no Ranking do Great Place to Work (2021-2023)

Empresas	Frequência
Magazine Luiza	3
Itaú Unibanco	3
Vivo	3
Accenture do Brasil	3
Arcos Dorados (McDonald's)	3
Banco Santander (Brasil) S.A.	2
Localiza	2
Porto Seguro	2
Ambev	1
Whirlpool	1
DHL Supply Chain	1
Grupo Boticário	1
Mercado Livre	1
Sicredi	1
Gazin	1
IBM Brasil	1

Fonte: (GREAT PLACE TO WORK, 2021, 2022 e 2023)

Foram selecionadas as empresas que apareceram com maior frequência no rank entre os anos de 2021 e 2023 pois elas apresentam sinais de consistência em medidas que promovem um bom ambiente de trabalho para seus funcionários.

A última etapa de seleção das empresas foi identificar a representatividade dos funcionários de tecnologia no quadro geral de funcionários. Os dados da representatividade de funcionários na área de tecnologia são apresentados na Tabela 2 e foram obtidos no linkedin das empresas no início de junho de 2024. Os valores se referem a perfis cuja a localidade igual a Brasil e a função na empresa relacionado a "Information Technology".

Tabela 2 – Perfis de Empresas Relacionados a Tecnologia da Informação

Empresas	Total de perfis	Perfis TI	% de Perfis TI sobre o Total
Magazine Luiza	8.932	2.112	23,65%
Itaú Unibanco	40.149	6.838	17,03%
Vivo	94.112	8.307	8,83%
Accenture do Brasil	33.447	671	2,01%
Arcos Dorados (McDonald's)	7.109	63	0,89%

Fonte: Elaborado pelos autores.

Pela baixa representatividade de profissionais de TI a empresa Arcos Dourados foi removida do grupo deste trabalho sendo as seguintes empresas as selecionadas. As empresas selecionadas foram:

- Accenture do Brasil
- Itaú Unibanco
- Magazine Luiza
- Vivo

3.1.3 Busca dos perfis

Cada pesquisa pela página web do LinkedIn retorna uma página com 10 perfis e um total de 100 páginas, totalizando um número máximo de mil perfis por busca. Levando essa característica em consideração, as pesquisas na página web do LinkedIn foram realizadas quatro vezes com o intuito de coletar um número maior de perfis. Em cada uma dessas pesquisas, uma empresa da lista da subseção 3.1.2 era selecionada no campo 'Empresa anterior'. Assim, o número máximo de perfis coletados foi de 4 mil perfis.

Foi desenvolvido um algoritmo de web scraping em Python para a coleta de dados das páginas web com os resultados das pesquisas.

Para cada um dos links de pesquisa (ACCENTURE, s.d.), (ITAÚ, s.d.), (MAGAZINE, s.d.) e (VIVO, s.d.) com os filtros configurados de acordo com os critérios da subseção 3.1.1, o algoritmo coletava todas as URLs dos perfis públicos por meio de web scraping. A Tabela 3 apresenta a quantidade de perfis públicos obtidos neste secção relacionados a cada uma das empresas selecionadas na secção anterior.

Tabela 3 – Quantidade de perfis públicos coletados por empresa

Empresas	Quantidade
Magazine Luiza	231
Itaú Unibanco	690
Vivo	667
Accenture do Brasil	779

Fonte: Elaborado pelos autores.

3.1.4 Extração de dados das URLs

Com o intuito de obter o histórico das empresas em que os trabalhadores estiveram, bem

como o tempo de permanência em cada empresa, foi desenvolvido um algoritmo de web scraping em Python para acessar cada uma das 2.367 URLs obtidas na subseção 3.1.3.

Ao acessar os perfis das URLs obtidas na subseção 3.1.3 Foram identificados vários perfis com alguns dados incompletos ou preenchimento de dados fora do padrão mapeado pelo algoritmo de web scraping da subseção 3.1.4 levando a uma perda considerável de dados a serem utilizados para treinamento do modelo. A Tabela 4 apresenta a quantidade de perfis que seguiram o padrão mapeado pelo algoritmo de web scraping desenvolvido pelo autor. Estes perfis foram usados pelas seções seguintes para treinamento do modelo de aprendizagem de máquina.

Tabela 4 – Quantidade de perfis seguindo padrão mapeado pelo algoritmo de web scraping

Empresas	Quantidade
Magazine Luiza	66
Itaú Unibanco	271
Vivo	257
Accenture do Brasil	251

Fonte: Elaborado pelos autores.

3.1.5 Ajuste de dados

Cada um dos perfis obtidos na subseção 2.1.4 foram transformados em uma lista contendo o tempo de permanência em cada uma das empresas nas quais o perfil trabalhou. Como o modelo irá prever o tempo de permanência de um determinado perfil com base no tempo em que esse perfil trabalhou em outras empresas, o trabalho atual dos perfis foi desconsiderado, pois o tempo de trabalho na empresa atual não possui um término, uma vez que ainda está vigente.

Esta etapa resultou em uma perda de aproximadamente 22% de perfis, resultando em um total de 659 listas, onde cada lista possui valores decimais que representam o tempo trabalhado em cada uma das empresas pelo perfil.

3.2 Modelo de aprendizagem

Para cada uma das listas obtidas na subseção 3.1.5, foram extraídos os dados de entrada para o treinamento dos modelos SVM. As seguintes características e alvos utilizados para treinar o modelo de machine learning são descritos a seguir na Tabela 5.

Tabela 5 – Representação dos dados de um perfil usado na base de dados

Empresas	Início	Fim		Tempo
0	04/2022	Atualmente	Desconsiderado	
1	11/2019	03/2022	Alvo	2.5
2	07/2019	11/2019	Característica 1	0.33
3	08/2018	06/2019	Característica 2	0.91
4	10/2017	05/2018	Característica 3	0.66
5	01/2017	09/2017	Característica 4	0.75

Fonte: Elaborado pelos autores.

A Tabela 5 representa como os dados de um perfil são extraídos, o emprego atual é desconsiderado pois não houve um processo de demissão. Para o perfil apresentado na tabela o tempo no próximo emprego é de 2.5 equivalente a 2 anos e seis meses. A quantidade de empregos é 4 que são a quantidade de empregos anteriores ao alvo. E a quantidade de anos trabalhados é de 2.65 anos, que é a soma dos tempos trabalhados nos 4 empregos. Tendo em vista esta configuração, o modelo é treinado para prever o alvo com base nas características.

- Alvo: tempo trabalhado no emprego anterior ao atual.
- Característica 1: quantidade de empresas trabalhadas excluindo a empresa atual e a empresa usada como alvo.
- Característica 2: tempo total trabalhado excluindo a empresa atual e a empresa usada como alvo.

A redes neurais é um modelo de aprendizagem de máquina que em seu treinamento busca minimizar o erro de predição dos seus dados assim o modelo é propenso ao overfitting não apresentando resultados satisfatórios em dados que fogem muito do padrão da base de treinamento. Com o intuito de melhorar a generalização dados ausentes na base de treinamento foi introduzido em 1995 o *Support vector machines* (SVM), que ao contrário das redes neurais, o SVM procura minimizar o erro de generalização dos dados ao invés do erro dos dados de treinamento Fung 2020.

Os modelos SVM possuem dois tipos, o primeiro deles é o *Support vector classification* que é usado em problemas de classificação onde o modelo encontra hiperplanos capazes de dividir um determinado grupo de dados em classes. O segundo tipo do modelo SVM é o *Support vector regression* que aplicado em problemas de regressão onde o modelo tenta encontrar uma dependência funcional entre os dados da base de dados Fung 2020.

Dados os objetivos deste trabalho o modelo machine learning escolhido para a predição do tempo de permanência do candidato a vaga foi o modelo de regressão linear SVM que tentará encontrar uma função que represente dependência entre os dados da base de dados.

Foram feitas duas implementações do modelo de regressão SVM, a primeira o SVM foi implementado manualmente, já a segunda implementação utilizando uma biblioteca em Python sklearn, que é uma biblioteca de código fonte aberto utilizada amplamente em aplicações de machine learning em Python, a sklearn foi projetada para facilitar a utilização de modelos de aprendizagem de máquina fornecendo implementações prontas e eficientes.

A implementação manual foi aplicado uma função de ajuste iterativo chamado *Gradient Descent* que ajusta os pesos dos modelos de forma gradativa com o intuito de minimizar a função de custo. A função de ajuste *Gradient Descent* torna a implementação fortemente dependente da escolha da taxa de aprendizagem e quantidade de interações.

Uma taxa de aprendizagem muito alta pode fazer com que o modelo não encontre uma solução ótima, já o inverso, uma taxa de aprendizagem muito baixa, pode fazer com que o modelo demore muito mais tempo e interações para encontrar uma solução ótima.A escolha

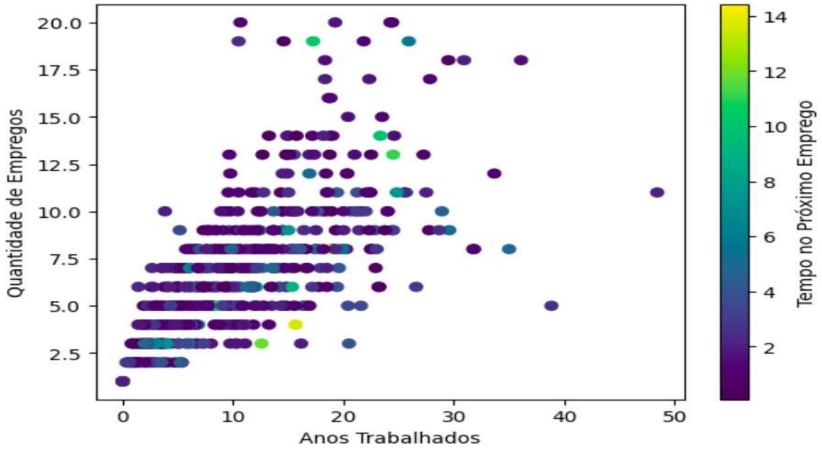
dos valores iniciais para a taxa de aprendizagem e quantidade de interações foi feita de forma empírica sendo a taxa de aprendizagem definida com valor de 0.001 e a quantidade de iterações igual a 1000. A segunda implementação do SVM utilizando a biblioteca sklearn foi definida apenas como um parâmetro, o kernel linear, que tende a otimizar o tempo de execução do modelo eliminando a necessidade de ajustes manuais. A validação cruzada com 5 grupos foi utilizada nas duas implementações com o intuito de treinar o modelo com diferentes agrupamentos de dados de treino e teste.

4 RESULTADOS

Os resultados referente a implementação manual do SVM serão identificados como implementação 1 e a implementação utilizando a biblioteca sklearn será identificada como implementação 2.

A base de dados conta com um total de 659 perfis obtidos após o processo de extração e tratamento dos dados. A média de anos trabalhados é dos 659 perfis é de aproximadamente 9 anos e 10 meses sendo 2 anos e 2 meses o tempo médio aproximado do tempo de permanência do trabalhador no próximo emprego. A quantidade média de empregos trabalhados ao longo dos anos é de aproximadamente 6.4 empregos diferentes ao longo dos anos.

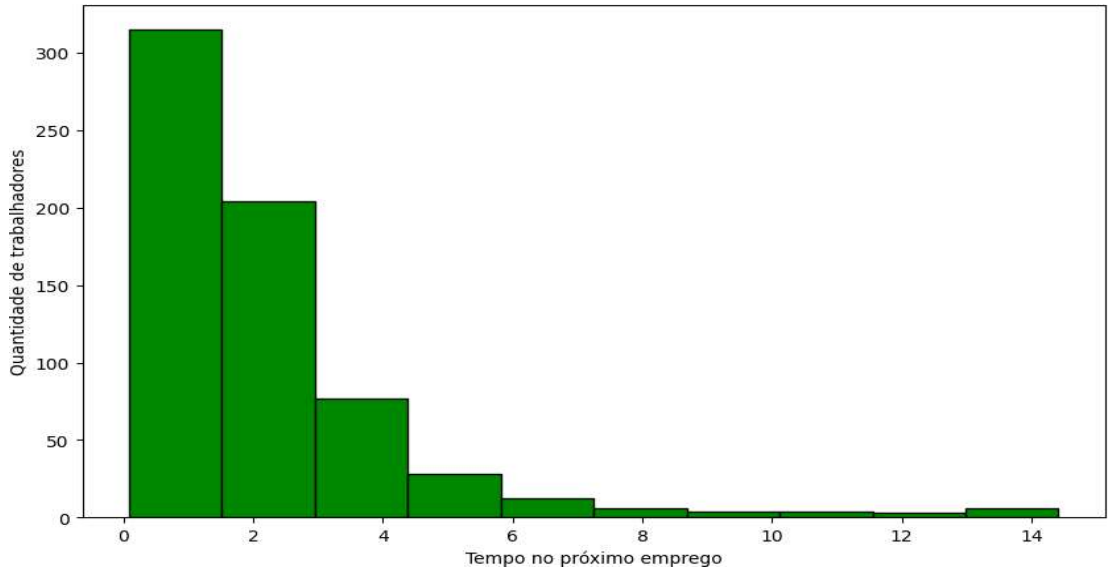
Figura 1 – Quantidade de empregos, anos trabalhados e tempo de permanência dos trabalhadores no próximo emprego



Fonte: Elaborado pelos autores.

Na Figura 1 é possível observar a distribuição dos 659 em relação às características e alvo utilizado no modelo de aprendizagem de máquina é perceptível a alta concentração de perfis com menos de 20 anos de carreira. A grande quantidade de pontos com cores mais próximas ao roxo levantam indícios de pouca variação de dados em relação ao tempo de permanência dos trabalhadores no próximo emprego. Ao analisar a Figura 2, é possível ver um histograma não uniforme que representa a pouca variação do tempo de permanência no próximo emprego dos perfis analisados na base de dados. Mais de 75% dos dados da base de dados são de perfis de trabalhadores que permaneceram por um período inferior a 3 anos no próximo emprego.

Figura 2 – Distribuição do Tempo de Permanência dos Trabalhadores no Próximo Emprego

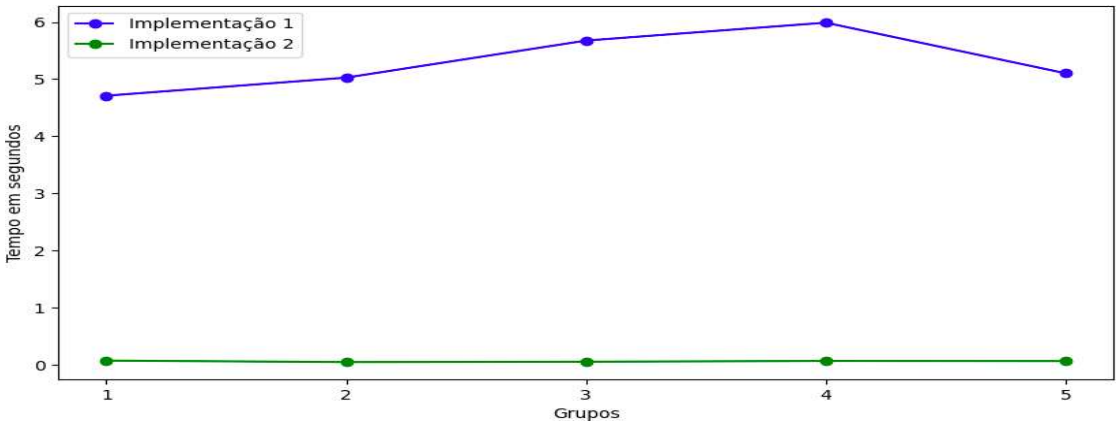


Fonte: Elaborado pelos autores.

Dos 659 perfis coletados 528 foram utilizados para treino dos modelos de predição em quando 131 foram utilizados para testar os modelos treinados.

Na Figura 3 é possível observar a disparidade entre os tempos de execução das duas implementações do SVM em cada um dos grupos da validação cruzada. O tempo médio de execução do treinamento da implementação 1 foi de aproximadamente 5,3 segundos, já o tempo médio de treinamento da implementação 2 foi de aproximadamente 0,06 segundos. A disparidade dos tempos de execução se dá pela implementação otimizada do modelo SVM da biblioteca sklearn que foi implementada com uma série de otimizações para garantir eficiência e rapidez quando aplicada em grandes bases de dados. A implementação manual não conta com otimizações de performance e eficiência e a utilização do método de ajuste interativo *Gradient Descent* torna a etapa de treinamento ainda mais ineficiente.

Figura 3 – Tempo de execução do treinamento por grupo da validação cruzada



Fonte: Elaborado pelos autores.

Para cada um dos modelos foram analisadas 3 métricas, a *Mean Squared Error*, *Mean Absolute Error* e o Coeficiente de Determinação. O MSE é uma média dos erros quadráticos entre os valores reais e os preditos pelo modelo, essa métrica é amplamente utilizada por penalizar erros maiores em comparação aos erros menores por conta da operação de elevação ao quadrado. Neste sentido, a implementação 1 obteve melhores resultados em relação à implementação 2.

Analisando o MAE a implementação 2 obteve maior precisão em relação à implementação 1 considerando o erro absoluto. Já o coeficiente de determinação da implementação 1 obteve um melhor resultado que a implementação 2.

Tabela 6 – Avaliação dos modelos SVM

Métrica	Implementação 1	Implementação 2
Mean Squared Error (MSE)	4.36301	4.63885
Mean Absolute Error (MAE)	1.35638	1.26590
Coeficiente de Determinação (R^2)	0.00221	-0.05871

Fonte: Elaborado pelos autores.

A nível de comparação entre as 2 implementações do modelo SVM a implementação 1 obteve resultados superiores a implementação 2 tendo como fator decisivo o coeficiente de determinação que foi negativo na implementação 2 o que indica que o modelo é pior do que uma simples média dos dados. Porém, por mais que a implementação 1 tenha atingido um coeficiente positivo este valor é muito próximo de zero, indicando que o modelo não encontrou uma função de regressão capaz de explicar a variância dos dados.

Levando em consideração que ambas implementações não conseguiram se aproximar de funções de regressão capazes de representar a relação entre os dados de base de treinamento é possível apontar duas supostas causas para a má performance dos modelos. A primeira causa seria o histograma desbalanceado apresentado na Figura 1, a pouca quantidade de perfis onde os trabalhadores permaneceram por períodos de tempo superiores a 3 anos dificultou a identificação de um padrão fazendo com que ambos os modelos estimassem curtos períodos de permanência no próximo emprego para perfis que permaneceram longos períodos de tempo. Os grupos que obtiveram maiores *Squared Error* foram os grupos com perfis de trabalhadores que permaneceram no próximo emprego por períodos longos.

A segunda possível causa é a pequena quantidade de características usadas para treinamento, há fortes indícios de que apenas a quantidade de empresas trabalhadas e o tempo total de trabalho não são características suficientes para determinar o tempo de permanência no próximo emprego.

Nenhum dos modelos implementados foi capaz de apresentar resultados significativamente superiores a uma simples média aritmética e a pouca variabilidade dos dados presentes na base de dados dificultou a criação de afirmações conclusivas sobre as

implementações uma vez que a capacidade dos modelos de aprenderem padrões significativos foi comprometida, fazendo com que suas previsões não desviasse substancialmente de uma simples média aritmética.

CONSIDERAÇÕES FINAIS

Este estudo apresentou uma abordagem inovadora ao utilizar dados extraídos do LinkedIn para treinar modelos de aprendizado de máquina, com o objetivo de prever o tempo de permanência de desenvolvedores de software em uma empresa. Porém a extração de dados em massa apresenta desafios em relação ao processo de extração e padronização dos dados abrindo caminhos para novos trabalhos com propostas de metodologias mais eficientes em relação a perda de dados padronizados.

As características utilizadas nos modelos de *machine learning* não foram suficientes para auxiliar os modelos na identificação da função de relação entre os dados. Em termos de projeções futuras, recomenda-se a incorporação de características adicionais como faixa salarial e modelo de trabalho, presencial ou remoto, são informações que auxiliam na predição do tempo de permanência dos trabalhadores.

Embora a primeira implementação tenha obtido resultados melhores, a segunda implementação se destaca pela eficiência, e tendo em vista a necessidade de aumentar a base de treinamento não apenas em quantidade de perfis mas também na quantidade de características a serem analisados a segunda implementação se apresenta como a mais adequada uma vez que a implementação manual do SVM apresentou um ineficiência com uma base de dados pequena.

Como a construção da base de dados foi um dos maiores desafios deste trabalho e impactou negativamente na performance dos modelos é de extrema importância que trabalhos futuros levem em consideração os resultados obtidos neste trabalho e desenvolvam métodos para criação de um base de dados que consiga equilibrar adequadamente a quantidade de dados e a variabilidade, permitindo assim que os modelos implementados consigam identificar padrões relevantes para as previsões. A utilização de outras redes sociais como o Glassdoor em conjunto com o LinkedIn abre possibilidades para trabalhos futuros analisarem e identificarem padrões que relacionam o tempo de permanência de um funcionário com no salário.

REFERÊNCIAS BIBLIOGRÁFICAS

AJIT, Pankaj. Prediction of employee turnover in organizations using machine learning algorithms. **algorithms**, v. 4, n. 5, p. C5, 2016.

AKASHEH, Mariam Al et al. A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review. **Expert Systems with Applications**, v. 238, p. 121794, 2024.

AKDUR, Gorkem; AYDIN, Mehmet N.; AKDUR, GIZDEM. Understanding Virtual Onboarding Dynamics and Developer Turnover Intention in the Era of Pandemic. **Journal of Systems and Software**, p. 112136, 2024.

ALVES, Bruna Ribeiro et al. Passivos trabalhistas, Turnover e a felicidade no trabalho nas empresas listadas na Brasil, Bolsa, Balcão–B3. **Revista Científica Hermes**, v. 34, p. 288-309, 2023.

ATEF, Markus; S ELZANFALY, Doaa; OUF, Shimaa. Early prediction of employee turnover using machine learning algorithms. **International journal of electrical and computer engineering systems**, v. 13, n. 2, p. 135-144, 2022.

CHAKRABORTY, Raj et al. Study and prediction analysis of the employee turnover using machine learning approaches. In: **2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)**. IEEE, 2021. p. 1-6.

CHO, Vincent; LAM, Wing. The power of LinkedIn: how LinkedIn enables professionals to leave their organizations for professional advancement. **Internet Research**, v. 31, n. 1, p. 262-286, 2021.

FUNG, Kit Fai et al. Improved SVR machine learning models for agricultural drought prediction at downstream of Langat River Basin, Malaysia. **Journal of Water and Climate Change**, v. 11, n. 4, p. 1383-1398, 2020.

MATOS, Nathalya Delfina Campos de. O papel do RH estratégico voltado para grande rotatividade dos funcionários nas organizações. **Repositório Institucional Unicambury**, v. 1, n. 1, 2022.

MTE, Ministério do Trabalho e Emprego. **Sondagem inédita feita pelo MTE aponta principais motivos para pedidos de demissão**. Disponível em: https://www.gov.br/trabalho-e-emprego/pt-br/noticias-e- conteudo/2024/Agosto/sondagem-inedita-feita-pelo-mte-aponta-principais-motivos-para-pedidoseemissao/copy_of_sondagemdesl igadosapedido080824.pdf. Acesso em: 17 nov. 2024.

RASCHKA, Sebastian; PATTERSON, Joshua; NOLET, Corey. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. **Information**, v. 11, n. 4, p. 193, 2020.

SALLABERRY, Jonatas Dutra et al. Características de Perfil dos Servidores do Ministério Público e sua Relação com a Intenção de Turnover. **Administração Pública e Gestão Social**, 2021.

SHARMA, Gaurav G.; STOL, Klaas-Jan. Exploring onboarding success, organizational fit, and turnover intention of software professionals. **Journal of Systems and Software**, v. 159, p. 110442, 2020.