Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

People Analytics - Um estudo de caso de ciência de dados na Gerência de Recursos Humanos da Apex-Brasil

César Antônio Ciuffo Moreira

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Gladston Luiz da Silva

Brasília
2025

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

MM838p

Moreira, César Antônio Ciuffo
People Analytics - Um estudo de caso de ciência de dados na Gerência de Recursos Humanos da Apex-Brasil / César Antônio Ciuffo Moreira; orientador Gladston Luiz da Silva. Brasília, 2025.
83 p.

Dissertação(Mestrado Profissional em Computação Aplicada) Universidade de Brasília, 2025.

1. Ciência de Dados. 2. People Analytics. 3. Agrupamento. 4. Classificação. 5. Séries Temporais. I. Silva, Gladston Luiz da, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

People Analytics - Um estudo de caso de ciência de dados na Gerência de Recursos Humanos da Apex-Brasil

César Antônio Ciuffo Moreira

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Gladston Luiz da Silva (Orientador)
PPCA/UnB

Prof. Dr. João Mello da Silva Prof. Dr. Filipe Alves Neto Verri

Prof. Dr. Marcelo Ladeira (Suplente)

Prof. Dr. Edna Dias Canedo
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 29 de Julho de 2025

Dedicatória

Dedico este trabalho à minha família: minha querida esposa Alessandra, meus maravilhosos filhos Sophia e Vicenzo. Agradeço a vocês o comprometimento com meus sonhos, a generosidade, o desapego e pelo apoio e carinho durante todo o Curso. Dedico, também, aos meus pais Cesar e Glória, que foram a primeira geração familiar a alcançar o nível superior, gerando oportunidade de crescimento e educação para mim e meu irmão. Todos me proporcionaram uma vida cheia de significado e propósito.

Audaces Fortuna Juvat (Virgílio, Eneida, X, 284)

Agradecimentos

Agradeço especialmente ao meu orientador Prof. Dr. Gladston Silva pela sabedoria, inspiração, parceria e incentivo e aos membros da minha banca Prof. Dr. Filipe Verri e Prof. Dr. João Mello, pelas correções e comprometimento.

Agradeço aos colegas João Batista Rodrigues Fonseca e João Vitor Rodrigues Batista pelo incentivo em todas as matérias do Curso que realizamos juntos e pela generosidade em compartilhar os conhecimentos e a paciência em todos os trabalhos que culminaram nesta Dissertação.

Agradeço à Apex-Brasil, na pessoa da minha líder e gestora Celene Boaventura, e a todo o time da Gerência de Recursos Humanos pelo apoio incansável, grandes esclarecimentos e pela paciência diante de tantas dúvidas.

Agradeço à Universidade de Brasília e à Coordenação do Programa de Pós-Graduação em Computação Aplicada (PPCA), que são fundamentais na construção do futuro do país, com abnegação e compromisso.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

O tema *People Analytics* ou *HR Analytics* ou *Workforce Analytics* é a implementação de técnicas de ciência de dados no ambiente de recursos humanos no âmbito das organizações, visando o melhor entendimento desta dinâmica e a predição de comportamentos. As práticas de *People Analytics* ainda não estão disseminadas no contexto brasileiro e, em especial, na Agência Brasileira de Promoção de Exportações e Investimentos (Apex-Brasil), que passa por forte automação de processos e por um processo de digitalização. Este ambiente se torna propício para o estudo e análise de diversos aspectos para entender a segmentação dos empregados, as possibilidades de sua classificação e a predição de absenteísmo. Este trabalho tem o propósito de estudar diversas técnicas de agrupamento, classificação e predição, para a prospecção e implantação de modelos de aprendizado de máquinas na Apex-Brasil. Em termos metodológicos, passará por todas as fases propostas, a saber: i) Entendimento do negócio e do contexto dos processos de recursos humanos na Apex-Brasil; ii) Entendimento, pré-processamento e análise exploratória dos dados; iii) Modelagem; iv) Avaliação de resultados e discussões; e, v) Proposição do plano de implantação dos modelos no ambiente da Apex-Brasil. Os resultados encontrados reforçam os exemplos da literatura e atendem aos objetivos propostos do trabalho. Sugere-se a implantação dos modelos selecionados na Agência e a medição de resultados para eventuais ajustes e melhor atuação.

Palavras-chave: People Analytics, HR Analytics, Working Analytics, Agrupamento, Classificação, Séries Temporais, Projeções

Abstract

The subject of People Analytics, HR Analytics, or Workforce Analytics is the implementation of data science techniques in the human resources (HR) environment within organizations, aiming at a better understanding of these dynamics and the prediction of behaviors.People Analytics practices are not yet widespread in the Brazilian context and, especially, in Agência Brasileira de Promoção de Exportações e Investimentos (Apex-Brasil), which is undergoing strong process automation and an accentuated digitalization process. This environment becomes conducive to the study and analysis of several aspects to understand the segmentation of employees, the possibilities of their classification, and the prediction of absenteeism. This work aims to study several grouping, classification, and prediction techniques for the prospecting and implementation of machine learning models at Apex-Brasil. In methodological terms, it will go through all the proposed phases, namely: i) understanding the business and the context of human resources processes at Apex-Brasil; ii) understanding, preprocessing, and exploratory analysis of data; iii) Modeling; iv) Evaluation of results and discussions; and v) proposal of the plan for implementing the models in the Apex-Brasil environment. The partial results found reinforce the examples in the literature and meet the proposed objectives of the work. It is suggested that the selected models be implemented in the Agency and that the results be measured for possible adjustments and better performance.

Keywords: People Analytics, HR Analytics, Working Analytics, Clustering, Classification, Time Series, Forecasts

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AED Análise Exploratória de Dados.

AIC *Akaike Information Criterion*.

Apex-Brasil Agência Brasileira de Promoção de Exportações e Investimentos.

ARIMA *Autoregressive Integrated Moving Average*.

CRISP-DM *Cross-Industry Standard Process for Data Mining*.

CX *Customer Experience*.

DBSCAN *Density-Based Spatial Clustering of Applications with Noise*.

ERP *Enterprise Resource Planning*.

ETL Extração, Transformação e Carga.

ETS *Error, Trend, Seasonality*.

EX *Employee Experience*.

GDPR *Europe General Data Protection Regulation*.

HR Human Resources.

IED Investimentos Estrangeiros Diretos.

KNN *K-Nearest Neighbors*.

LGPD Lei Geral de Proteção de Dados.

MAE *Mean Absolute Error*.

MDIC Ministério de Desenvolvimento, Indústria, Comércio e Serviços.

MSE *Mean Squared Error*.

RH Recursos Humanos.

RMSE *Root Mean Squared Error*.

SARIMA *Seasonal ARIMA*.

SES *Single Exponential Smoothing*.

SSE *Sum of Squared Errors*.

TCU Tribunal de Contas da União.

TEMAC Teoria do Enfoque Meta Analítico.

Capítulo 1

Introdução

O tema *People Analytics* ou *HR Analytics* ou *Workforce Analytics* é a implementação de técnicas de ciência de dados no ambiente de processos e dados de recursos humanos nas organizações, visando o melhor entendimento desta dinâmica e a predição de comportamentos. Este tema se torna relevante no ambiente corporativo com as mudanças e transformações, cada vez mais rápidas, que a sociedade vem passando, tendo em vista os aspectos políticos, econômicos, sociais, tecnológicos, ambientais e legais. Entender estes movimentos diversos e complexos e dar respostas objetivas e práticas que atendam aos aspectos de equilíbrio entre o bem-estar individual e a busca de produtividade e resultados dos empregados se torna um desafio. A criação de valor a partir da gestão de pessoas é crítica para a resiliência organizacional, é um diferencial estratégico e um fator crítico de sucesso para qualquer iniciativa.

A Apex-Brasil passa por forte automação de processos e por um processo de digitalização acentuado, onde o ambiente é propício para a aplicação de práticas de *People Analytics*, entendendo as variáveis disponíveis e mediante estudo e análise de modelagem. A aplicação de técnicas de ciências de dados permitirá o melhor agrupamento, classificação e predição em dados de absenteísmo, buscando maior precisão, acurácia e confiança nos aspectos mais importantes de recursos humanos e no processo decisório da Agência. Importante destacar que a ética e a privacidade no uso dos dados são fundamentais para a proteção do empregado no manuseio das informações e a integridade na aplicação dos modelos. A partir dos resultados dessa pesquisa, busca-se a implantação de modelos de aprendizado de máquina que auxiliem o processo decisório dos gestores da Apex-Brasil, com a segmentação e classificação dos empregados, bem como a predição de comportamentos.

A segmentação permite a identificação de características comuns de um determinado grupo estudado, entendendo aspectos sociais que podem determinar o melhor aproveitamento e adequação de diversos fatores para atender a estas necessidades identificadas.

A classificação em ciência de dados fornece uma base sólida para a tomada de decisões estratégicas na gestão de pessoas, promovendo um ambiente de trabalho mais produtivo, satisfatório e inclusivo. Enfim, o planejamento de recursos humanos, baseado em projeções de séries temporais, é fundamental para adequar a previsibilidade comportamental e financeira, adequando as ações de RH e o orçamento na busca de eficiência operacional.

Na revisão bibliográfica verifica-se que a busca de efetividade de práticas está totalmente alinhada à performance organizacional. Observa-se que o tema acaba por combinar temas de gestão e de ciência de dados, não sendo oportuno a separação. Contudo, apresenta-se num cenário de informações que precisam de validação a cada contexto, onde cabe o estudo de caso de como implementar as práticas de ciências de dados no contexto de recursos humanos com êxito, ainda que se disponha de mais informações com a adoção de sistemas integrados, tais como os *Enterprise Resource Planning* (ERP).

Este trabalho propõe o estudo de diversas técnicas de agrupamento, classificação e predição em dados de absenteísmo, visando ser prospectivo na implantação de um ambiente de *People Analytics* na Apex-Brasil, com ganho de maturidade e melhor adequação, bem como na visão integrada dos dados oriundos de diversos sistemas e de diversos processos de gestão de pessoas. Espera-se analisar cada técnica, testar os diversos algoritmos existentes e adequados ao perfil de dados existente, para que se possa prospectar tecnologias e análises que suportem o processo decisório da Agência.

Serão adotados os preceitos do Modelo *Cross-Industry Standard Process for Data Mining* (CRISP-DM), no que tange às atividades de Mineração de Dados. A diversidade de técnicas e modelos também é um desafio para este trabalho, em face do incremento de diversas linguagens de programação, modelos computacionais, modelos estatísticos e matemáticos e o advento da Inteligência Artificial. Será necessário entender o contexto da Apex-Brasil para a escolha do modelo que melhor represente as necessidades da agência.

Na aplicação das técnicas de agrupamento ou clusterização, deverão ser analisadas as vantagens, as desvantagens e sua melhor aplicação, tais como o K-means, a Clusterização Hierárquica e o método *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN). Já nas técnicas de classificadores, serão testadas metodologias de Regressão Logística Multivariada, Árvores de Decisão, Random Forest e *K-Nearest Neighbors* (KNN). E para ajuste da predição de Séries Temporais, serão utilizadas as técnicas: *Single Exponential Smoothing* (SES), Holt, Holt-Winters, *Error, Trend, Seasonality* (ETS), *Autoregressive Integrated Moving Average* (ARIMA) e *Seasonal ARIMA* (SARIMA). A análise e extração de dados dos sistemas de recursos humanos existentes representam um desafio, especialmente quanto ao entendimento de cada contexto e à diversidade dos atributos disponíveis.

1.1 Justificativa

A Agência de Promoção de Exportações do Brasil – Apex-Brasil é um Serviço Social Autônomo, criado pelo Decreto Presidencial nº 4.584, de 5 de fevereiro de 2003, alterado pelo Decreto nº 8.788, de 21 de junho de 2016, cuja instituição foi autorizada pela Medida Provisória nº 106, de 22 de janeiro de 2003, posteriormente convertida na Lei nº 10.668, em 14 de maio do mesmo ano.

É uma entidade sem fins lucrativos, de interesse coletivo e de utilidade pública, que tem por competência a execução das políticas de promoção de exportações e investimentos em cooperação com o poder público e em conformidade com as políticas nacionais de desenvolvimento, particularmente aquelas relativas às áreas industrial, comercial, de serviços e tecnológica. A partir de 2023, passou a ser supervisionada pelo Ministério de Desenvolvimento, Indústria, Comércio e Serviços (MDIC), em consonância com a Lei Nº 14.600, de 19 de junho de 2023, e com o Decreto Nº 11.427, de 02 de março de 2023, através de aditivo ao contrato de gestão.

O objetivo da Apex-Brasil é, em cooperação com o Poder Público, executar as políticas de promoção das exportações brasileiras e dos investimentos e a internacionalização de empresas públicas e privadas brasileiras, por meio da pesquisa, da formação e capacitação, do desenvolvimento institucional, dentre outras ações, observadas as políticas nacionais de desenvolvimento, mormente no que tange aos setores da indústria, comércio, serviços, tecnologia e agricultura, com atenção especial às ações estratégicas que promovam a inserção competitiva das empresas brasileiras nas cadeias globais de valor, a atração de investimentos e a geração de empregos, e apoio às empresas de pequeno porte.

A sua missão é desenvolver a competitividade das empresas brasileiras, promovendo a internacionalização dos seus negócios e a atração de Investimentos Estrangeiros Diretos (IED). A Apex-Brasil atua de diversas formas para promover a competitividade das empresas brasileiras em seus processos de internacionalização, oferecendo inteligência de mercado, qualificação empresarial, estratégia para internacionalização, promoção de negócios e imagem e atração de investimentos estrangeiros para empresas brasileiras, sem que se tenha o ânimo de lucro.

A Agência apoia, atualmente, mais de 14 mil empresas em mais de 100 setores produtivos da economia brasileira, que exportam para mais de 200 mercados. São realizadas amplas ações de promoção comercial, como missões empresariais, rodadas de negócios, apoio à participação de empresas brasileiras em grandes feiras internacionais, visitas de compradores estrangeiros ao Brasil, entre outras.

Gera conhecimento em diversos formatos (estudos, painéis, alertas, *insights*, informes, vídeos, webinários, capacitações, *podcasts*, metodologias diversas, entre outras) e atendimentos aos seus clientes (empresas brasileiras, compradores internacionais e investidores

estrangeiros), de forma própria ou em parceria, necessitando do desenvolvimento e da retenção constantes de seus colaboradores para atender os desafios colocados pelo mercado internacional. Para atender a este desafio, conta com quadro técnico altamente especializado, onde 0,954 (noventa e cinco por cento e quatro décimos) do quadro de pessoal possui ensino superior e 0,551 (cinquenta e cinco por cento e um décimo) possui educação avançada, tais como especialização, mestrado e doutorado.

Em termos de governança organizacional, a Apex-Brasil submete-se ao controle externo do Tribunal de Contas da União (TCU). Segundo o Referencial Básico de Governança do Tribunal[1], apresenta-se a seguinte necessidade de planejamento de pessoas que deve abranger todas as funções/subsistemas operacionais de gestão de pessoas, tais como recrutamento e seleção, treinamento e desenvolvimento, promoção da qualidade de vida no trabalho, avaliação de desempenho, concessão de benefícios e vantagens. É crucial a constante avaliação e manutenção desses processos por meio de seus dados e a interpretação clara que norteiem as ações da Agência. A Apex-Brasil apoia e pretende o aumento de maturidade em gestão de pessoas e seu resultado nos Índices de Monitoramento do referido órgão de controle.

Em termos estratégicos, a Apex-Brasil tem dois objetivos, vigentes no período de 2024 a 2027, que envolvem a participação direta da Gerência de Recursos Humanos: Ser uma agência digital, com modelos de negócios transversais, escaláveis, que entreguem a melhor experiência e valor para clientes e colaboradores; e desenvolver uma cultura organizacional de transversalidade, sinergia e resultados.

Estes objetivos corroboram com a necessidade de ações de *People Analytics*, dada a necessidade de diagnósticos precisos que embasem as ações da unidade organizacional. É prioritária a automação de processos internos de gestão de pessoas; para isso, conta com a Solução de Administração de Pessoal: Sistema ERP "Corpore RM" da empresa TOTVS – Módulo de Gestão de Pessoas. Bem como, existe a necessidade de integração de solução com outros dados dos empregados e de administração de pessoal, carecendo de uma análise mais detalhada, bem como a integração de técnicas que permitam a visão única dos empregados por diversos prismas.

A motivação do estudo se dá pela necessidade de contribuir para a manutenção das políticas, normativos e processos de recursos humanos atualizados e referenciados nas melhores práticas de mercado, direcionados ao Planejamento Estratégico da Agência, com o quadro de pessoal próprio preparado, adequado, bem remunerado e com as perspectivas de desenvolvimento adequadas às competências necessárias para atender aos desafios da Agência em sua missão institucional.

A relevância dos capitais sociais e intelectuais e das competências na geração de valor organizacional encontra respaldo em Nahapiet e Ghoshal [2], bem como nas abordagens

clássicas de Edvinsson e Malone [3], que destacam como ativos intangíveis, como o capital humano, organizacional e relacional, são fontes críticas de vantagem competitiva para a organização.

A finalidade da Gerência de Recursos Humanos, conforme Regimento Interno da Agência, consiste em desenvolver, implementar, executar, avaliar e aprimorar as políticas e diretrizes corporativas de Recursos Humanos, para promover condições de trabalho que propiciem o equilíbrio entre as esferas profissional, pessoal e familiar da vida de seus colaboradores, segundo o Código de Ética da Apex-Brasil. E a Agência, em sua Política de Qualidade, mantém compromisso com o desenvolvimento dos seus Colaboradores. Considerando-se o aspecto de governança, estratégico, tecnológico e normativo, este estudo se justifica e visa contribuir para a gestão dos recursos humanos da Apex-Brasil.

1.2 Objetivos

Este trabalho tem como objetivo geral propor a construção e seleção de modelos mais adequados ao agrupamento e classificação de empregados e de predição de séries históricas para a Apex-Brasil, a partir dos dados primários de diversos processos de gestão de pessoas existentes, no que tange à demografia dos empregados, informações de salários, encargos e atributos, de avaliação de desempenho, de absenteísmo, de clima organizacional e de educação corporativa.

Para isto, serão necessários alcançar os objetivos específicos relacionados a seguir:

- Analisar o contexto dos processos de recursos humanos e extrair, transformar e carregar os dados de absenteísmo para a pesquisa;
- Realizar a seleção sistêmica de variáveis relacionadas a processos de recursos humanos, para suportar a pesquisa e os modelos;
- Construir, medir e avaliar modelos de agrupamento, classificação e previsão;
- Avaliar os resultados obtidos em conjunto com as equipes de atuação da Gerência de Recursos Humanos da Apex-Brasil; e,
- Propor a automação desses processos no ambiente tecnológico da Agência.

Considera-se como critério de sucesso, a avaliação dos modelos propostos e sua comparação aplicados aos dados da Apex-Brasil, indicando a(s) melhor(es) técnica(s) aplicada(s), servindo de referência para a implantação dos algoritmos indicados nos ambientes computacionais da Agência, contribuindo para o processo decisório e análise das Políticas de Recursos Humanos.

1.3 Metodologia

Tendo em vista os objetivos gerais e específicos colocados, será utilizado como referência a estrutura proposta pelo *Cross-Industry Standard Process for Data Mining* (CRISP-DM) [4] e [5], passando por todas as fases propostas, a saber:

1. **Entendimento do negócio e dos processos de recursos humanos na Apex-Brasil**, onde será realizada a definição do contexto de cada processo analisado, entendimento do negócio e das diretrizes de RH e definição do escopo de pesquisa a ser estudado;

2. **Entendimento, pré-processamento e análise exploratória dos dados**, onde será realizado o levantamento, transformação e carga (ETL) dos dados, com as seguintes fases: Identificação e coleta dos dados; checagem de limpeza e integridade; busca da convergência dos dados; e carga e consolidação das tabelas. Após a consolidação das tabelas será realizada a análise exploratória dos dados, com a validação dos dados a serem trabalhados, utilizando técnicas e algoritmos em Linguagem e uso de Bibliotecas em R para conhecer melhor o contexto, com as seguintes fases: análise de variáveis numéricas e categóricas; sumarização, identificação de valores errados, omissos e *outliers*; visualização gráfica; identificação de ruídos, se for o caso; definição de funções de seleção, filtro, junção e agrupamento; e análise estatística, correlação e distribuição (Histogramas), quando for o caso;

3. **Modelagem**, utilizando técnicas de aprendizado de máquina descritivas ou preditivas, onde pretende-se: analisar e determinar técnica(s) mais pertinente(s) de agrupamento ou clusterização; analisar e determinar técnica(s) mais pertinente(s) de classificação; e analisar e determinar técnica(s) mais pertinente(s) de projeção de séries temporais; e demonstrar a visualização dos resultados;

4. **Discussões e Avaliação de Resultados**, onde a partir dos resultados encontrados, demonstrar a adequação dos modelos utilizados e propostos e discutir a melhor aplicação para o contexto de negócio e suas consequências para os objetivos da pesquisa;e

5. **Conclusão, Trabalhos Futuros e Proposição de Implantação no ambiente da Apex-Brasil**, definindo a implantação proposta, operação e manutenção evolutiva e corretiva dos modelos selecionados e ajustados, bem como a visualização dos resultados e as referências de análise, refinando todo o escopo de trabalho.

A figura abaixo demonstra a interação entre as fases, conforme proposto no modelo original [4] e [5]:

Figura 1.1: Estrutura Referencial CRISP-DM

1.4 Estrutura do Trabalho

Essa dissertação será estruturada em cinco capítulos, sendo que o Capítulo 1 contém a Introdução, a Justificativa da Necessidade do estudo para a Apex-Brasil, a Metodologia e os Objetivos Principal e Específicos. O Capítulo 2 apresenta o Referencial Teórico contextualizando o que é *People Analytics*, as técnicas de Mineração de Dados e trabalhos relacionados. O Capítulo 3 apresenta o Estudo de Caso, onde são apresentados os entendimentos do negócio, a preparação dos dados e a modelagem proposta. O Capítulo 4 apresenta a discussão e avaliação das técnicas aplicadas. Por fim, apresenta as Conclusões, a proposta de trabalhos futuros e as recomendações para implantação dos modelos na Apex-Brasil.

Capítulo 2

Referencial Teórico

Neste capítulo são apresentados os referenciais teóricos e artigos, tendo em vista os temas afeitos à gestão de pessoas ou de recursos humanos, sua importância e relação direta para a estratégia e impactos para a organização, bem como os objetos de estudo de ciência de dados, tais quais agrupamento, classificação e predição, que apoiaram o presente estudo.

2.1 Alinhamento estratégico e práticas de RH

As práticas de RH estão alinhadas com a estratégia da organização, conforme a transformação digital e a inteligência artificial aceleram a digitalização das organizações, as práticas de recursos humanos também passam por intensa mudança, impactando diretamente na performance corporativa. A experiência do empregado (EX), passa a ser totalmente integrada, alinhada com a experiência do cliente (CX), tendo a mesma velocidade e resposta da organização. A mensuração ainda constitui um desafio, que precisa ser enfrentado para reposicionar o custo de empregado como investimento alinhado à estratégia corporativa[6].

2.1.1 Estratégia de RH Digital e Resultados Organizacionais

Em pesquisas recentes, mostra-se que a Estratégia Digital de Recursos Humanos tem impacto direto na performance corporativa, dada a acelerada transformação digital das organizações e a digitalização dos processos de RH, gerando uma relação sinérgica direta [7]. Esta relação reduz custos operacionais e geram impactos positivos em indicadores de performance no longo prazo, ao tratar o empregado em diversos níveis e parâmetros, o modelo abaixo traduz a relação direta de impacto:

Percebe-se a complexidade dos temas da estratégia digital de RH e a necessidade de uma visão integrada para que se possa relacionar aos resultados estratégicos organizaci-

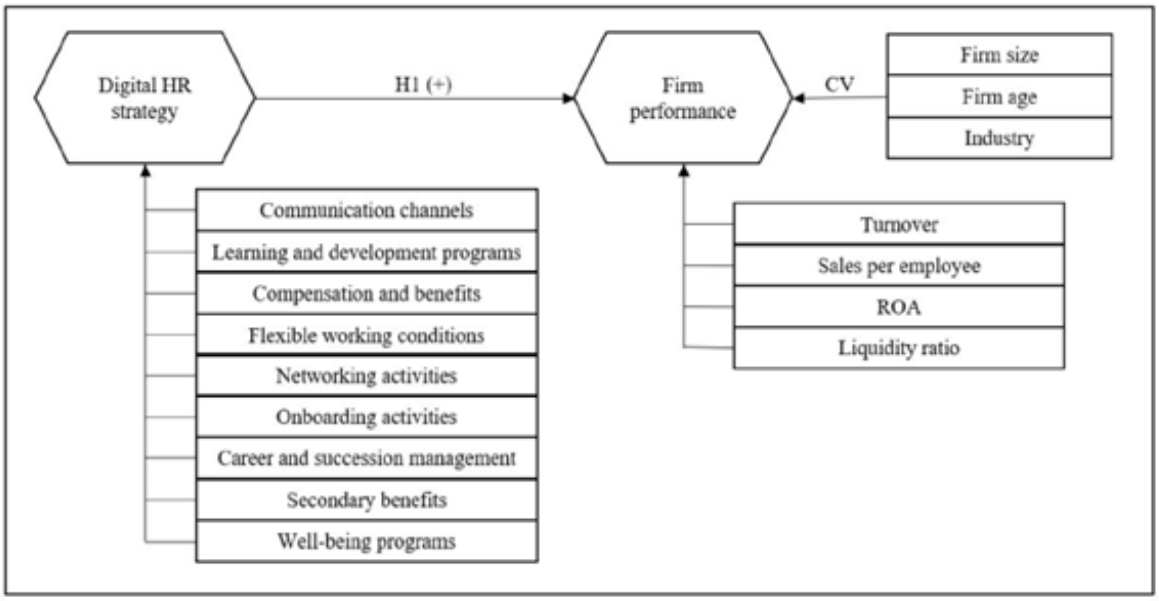

Fig. 1. The proposed research model (CV: Control variables).

Figura 2.1: Estratégia de RH e o impacto corporativo

onais. Além da infraestrutura digital, deve-se buscar elevar a liderança da organização para acompanhar as mudanças de culturas desejadas, bem como observar que a inovação, qualidade de vida e a performance da organização estão correlacionadas, comprovando vínculo entre estas práticas [8].

Entre os impactos de RH que possuem impacto direto na estratégia corporativa, está o alinhamento da força de trabalho à estratégia [9]. Devem ser levadas em consideração as demandas de análise de força de trabalho, tais como: a competitividade da organização, a relação hierárquica, as demandas legais e regulatórias, a eficiência operacional, a gestão de custos de recursos humanos, as pautas de diversidade, equidade e inclusão e as reflexões dos processos de RH.

2.1.2 Experiência Digital do Empregado

A experiência do empregado (*Employee Experience* (EX)) deve acompanhar a experiência compartilhada com os clientes da organização (*Customer Experience* (CX)), integrando as necessidades, desenvolvimento e impactos propostos, conforme a figura abaixo [10]:

A implantação do conceito de experiência do empregado nas organizações auxilia a estratégia de negócios de uma organização, aumentando a maturidade neste ambiente e melhorando a relação com os clientes. Os dados dos clientes são capazes de gerar *insights* descritivos, prescritivos e acionáveis, o mesmo acontece com os dados dos empregados

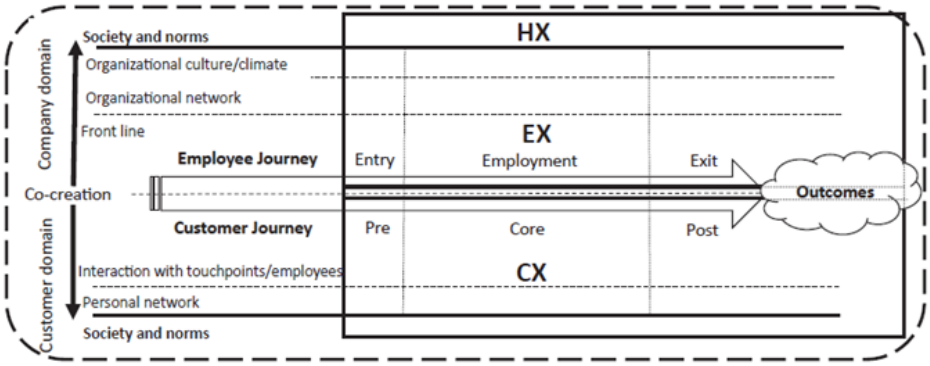

Figure 1.
Conceptual Model of CX, EX and HX (figure by authors)

Source(s): Figure by authors

Figura 2.2: Relação da Experiência do Empregado e do Cliente

[11].E, ainda, que boas políticas de ambiente de trabalho digital aumentam o bem-estar do empregado e a inovação corporativa, que junto com a atuação das lideranças no ambiente digital, também influenciam a performance organizacional [12].

Paralelo às estratégias de mercado, faz-se necessária a segmentação, classificação e previsibilidade dos empregados, sendo necessária a implementação de ferramentas que auxiliem os decisores em suas práticas, aumentando a efetividade dos empregados, sua performance e seu bem-estar de forma equilibrada. As métricas de RH permitem às organizações uma abordagem holística de diversas soluções possíveis[13].

2.1.3 Métricas de RH

Nesse contexto é importante analisar as métricas de RH e seus resultados para a organização, onde o modelo conhecido como de Harvard ainda permanece atual e considerado uma vantagem competitiva[14]:

Observa-se que os impactos de RH, como compromisso, competências, congruência e efetividade de custo, estão diretamente relacionados à efetividade organizacional e bem-estar individual e corporativo. Serão detalhadas as principais métricas de RH para a Apex-Brasil, bem como será dado tratamento estatístico nas variáveis numéricas e categóricas. E, tendo em vista o aspecto estratégico da iniciativa, complementa-se a justificativa do trabalho.

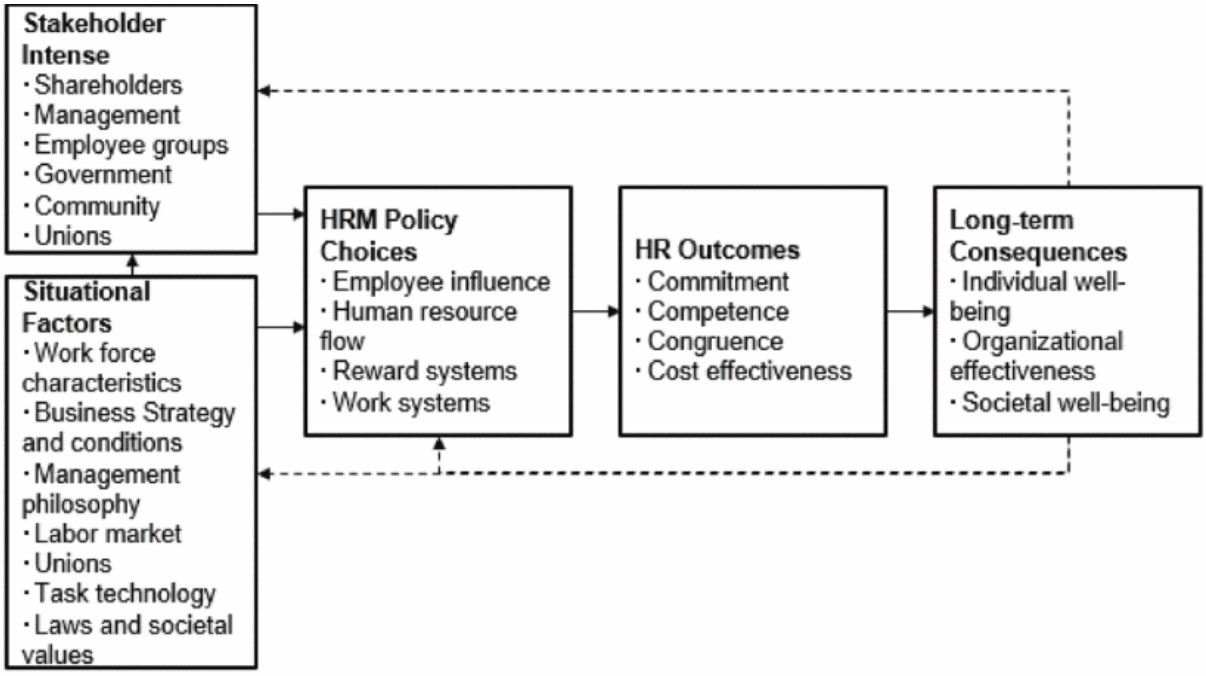

Figura 2.3: Modelo de Harvard - Impactos de RH e as consequências no longo prazo

2.2 *People Analytics*

People Analytics é uma disciplina emergente dentro da ciência de dados que se concentra na aplicação de técnicas analíticas avançadas para entender, gerenciar e melhorar a força de trabalho. Essa abordagem envolve a análise de grandes volumes de dados sobre funcionários e candidatos para fornecer *insights* que possam guiar decisões estratégicas e operacionais nas organizações. Serão explorados os fatores críticos de sucesso para a modelagem em *People Analytics*, além das melhores técnicas de agrupamento, classificação e predição (*forecasting*).

Os principais aspectos da área apontam que para a aplicação das técnicas de People Analytics serem efetivas, devem ser observados os seguintes temas [15] : i) entregas; ii) gestão de partes interessadas; iii) habilitadores; e iv) governança. Todas estas áreas de conhecimento devem ser conjugadas no interesse de buscar referência, as melhores práticas de mercado e análises preditivas [16].

Tendo em vista a natureza competitiva da informação, que trata de tema estratégico das organizações, a publicação acadêmica é restrita[17]. Em artigos que relacionam diversas fontes, fica clara a natureza genérica das informações e a carência de exploração de bancos de dados de empregados e discussão de modelos de ciência de dados.

A amplitude de temas e as possibilidades de entregas são variadas e complexas. Dado o contexto social da área de estudo. A abordagem acadêmica ainda é voltada para aspectos

genéricos e necessita de maior aprofundamento e aplicação técnica [18], portanto serão propostos, analisados e discutidos modelos de agrupamento, classificação e previsão para progredir na pesquisa.

2.2.1 Principais Entregas da Disciplina

Entre as principais entregas da disciplina[15] está a pesquisa organizacional, o monitoramento dos diversos aspectos do empregado e uma cultura baseada em resultados e evidências. A pesquisa organizacional é uma oportunidade para a utilização de algoritmos para o melhor entendimento e a possibilidade de um processo decisório mais equilibrado (*data-driven decision-making process*) [19], minimizando o viés decisório e maior rigor estatístico.

Entre os objetivos mais comuns [20] estão o planejamento da força de trabalho, o desenvolvimento de recursos humanos, a melhoria da assertividade do processo de recrutamento e seleção, a melhoria da performance dos empregados, conforme a tabela a seguir:

No contexto brasileiro e da necessidade da Apex-Brasil, levantado por meio de entrevista com os gestores, os seguintes aspectos de RH serão tratados nesta Dissertação, que deverão ser detalhados na fase do entendimento do negócio e dos processos de recursos humanos na Apex-Brasil:

- Remuneração Estratégica: tratamento das informações de salários, encargos e benefícios para projeções financeiras, tendo em vista a adequação de longo prazo do orçamento de pessoal da Agência;
- Bem-estar e engajamento: gestão do clima organizacional e da qualidade de vida no trabalho, para aumentar o engajamento e bem-estar dos empregados e a respectiva performance;
- Rotatividade dos Empregados ou Turnover[1]: melhoria do processo de recrutamento e seleção, das movimentações internas e de demissões, buscando a melhor adequação de perfil ao cargo;
- Gestão de Performance: entender os fatores que impactam na avaliação de desempenho, nos eixos comportamental e de resultados, na busca da predição de melhores entregas; e

[1]Turnover: refere-se ao número ou à proporção de empregados que deixam uma organização e são substituídos por novos funcionários em um determinado período de tempo. É uma métrica crítica para as áreas de recursos humanos, pois uma alta taxa de turnover pode indicar problemas como baixa satisfação no trabalho [21]

	Analysis Objective	Types of Data Used	Resulting Insights
Workforce Planning	▪ Identify skill gaps ▪ Verify whether the right people are placed in the right position ▪ Retain top performers	▪ Labor market data ▪ Work portal documents ▪ Individual profile data (employment information, personal detail, educational background, work history, performance reviews) ▪ Employee behavior and collaboration data collected by sensors	▪ Essential skills for performance ▪ Long-term workforce supply and demand plan
HR Development	▪ Identify training needs ▪ Check the effectiveness of current training programs	▪ Training content ▪ Individual training history ▪ Performance appraisal data	▪ Automated and individualized training recommendation ▪ Correlation between training and performance
Recruitment and Selection	▪ Recruit talented applicants in efficient and timely manner ▪ Nudge desired manager behaviors ▪ Expanding the candidate pool	▪ CV data ▪ Social media data ▪ AI-assisted interview data ▪ Job market data ▪ Past candidates' recruitment information ▪ E-mail records	▪ Person-job fit of the applicant ▪ Performance prediction ▪ Hiring process improvement ▪ Desired manager competencies and traits
Performance Improvement	▪ Analyze HR factors influencing organizational outcomes ▪ Detect employee sentiment related to performance	▪ Administrative data ▪ Turnover data ▪ Employee surveys ▪ Performance appraisal data ▪ Financial statistics ▪ Pop-up questions data ▪ External/internal social media data ▪ Anonymous bulletin board	▪ Performance-inducing HR factors ▪ Negative elements affecting employee satisfaction

Figura 2.4: Modelos de Aplicação de Analytics

- Absenteísmo: melhor predição das faltas dos empregados, tendo em vista as licenças e atestados, antecipando ações de saúde e segurança do trabalho para amenizar este impacto.

Em resumo, as técnicas de *People Analytics* devem ajudar a antecipar processos de perda de talentas, necessidades de melhoria na gestão de competências, a rotatividade de movimentações internas, bem como impactos no clima organizacional ou no bem-estar corporativo.

Dentre as necessidades, será priorizado para o trabalho a utilização das técnicas de agrupamento e classificação para as variáveis de recursos humanos para atender o estudo do absenteísmo, buscando a melhor técnica para o contexto da Agência.

Adicionalmente, serão utilizados modelos de predição de séries temporais tendo em vista somente os dados de licenças e atestados e, na sequência, será utilizado a partição dos dados pelo estudo de segmentação utilizando o resultado do melhor algoritmo de agrupamento.

Em termos da disciplina de *People Analytics* é de crucial importância a conjugação de diversos métodos com diferentes pontos de vista (visão holística) para a visão de gestão de pessoas. Acredita-se que o exercício de diversas técnicas demonstrará a complexidade de estudo e a necessidade de constante experimentação.

O sistema de *People Analytics*, cumpre uma função central de apoio à gestão estratégica de pessoas, viabilizando decisões orientadas por dados e fornecendo inteligência sobre padrões, riscos e oportunidades no quadro de empregados.

Sua estrutura é composta por uma integração entre a Gerência de Recursos Humanos, a área de Tecnologia da Informação e unidades responsáveis por inteligência de negócios, que operam sobre dados provenientes de diversos sistemas internos (folha de pagamento, ponto eletrônico, eSocial, treinamento e desenvolvimento).

O processo é composto por etapas bem definidas, que incluem coleta, tratamento, modelagem estatística, análise preditiva e entrega de informações, seja por meio de visualização interativa, relatórios executivos ou análises ad hoc. Este sistema está inserido em um contexto organizacional de crescente demanda por eficiência, transformação digital e accountability, com ênfase na adoção de práticas modernas de gestão de pessoas, capazes de gerar valor tanto operacional quanto estratégico para a Agência.

2.2.2 Gestão de Partes Interessadas

A gestão de partes interessadas, disciplina inicialmente da área de estratégia e de gestão de projeto, é um fator crítico de sucesso para a implementação de um modelo de *People Analytics*. A complexidade dos relacionamentos, sua rede de atuação e a natureza de cada expectativa são fundamentais para o contexto de análise. Abaixo, segue modelo coligido de diversos autores sobre o tema [16]:

No ambiente interno, a governança das políticas de gestão de pessoas focada em subsidiar a alta administração e realizar o *accountability* das iniciativas, buscando direcionar, monitorar e avaliar as ações de RH [1]. Suportar o processo decisório dos gestores, para que possam adequar da melhor maneira o time aos seus desafios. Em relação aos empregados, implementar a experiência do empregado, com a busca de equilíbrio entre performance e bem-estar. E com os profissionais de recursos humanos, melhorar a implementação de políticas e ações de RH, ajustando o propósito e a criação de valor.

No ambiente externo, por intermédio dos empregados, garantir relacionamento de longo prazo com os clientes, melhorar a imagem e a reputação da organização no ecossistema em que está inserida, além de prestar contas dos recursos investidos na missão institucional, garantindo retorno e transparência dos investimentos.

O entendimento das expectativas e o seu gerenciamento é fundamental, todas as organizações estão inseridas no tecido social, ou seja, possuem reputação e buscam o cum-

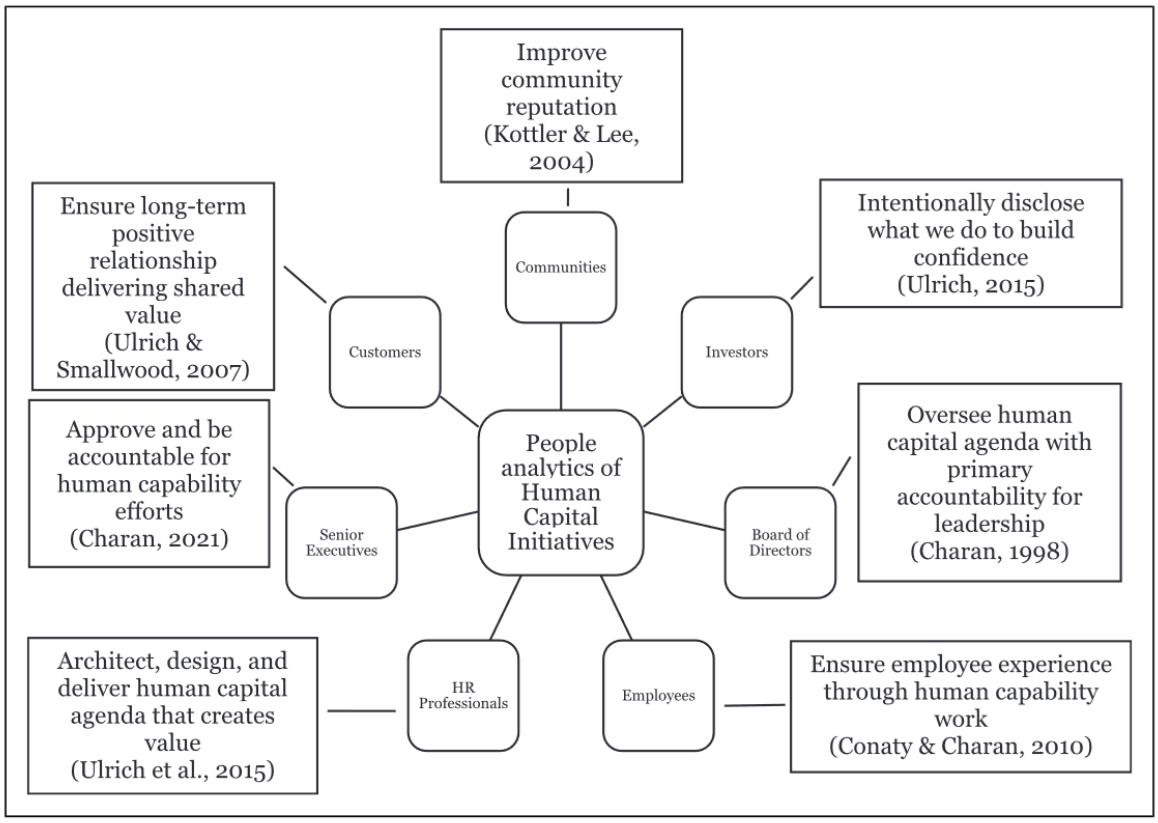

Figura 2.5: Modelo de Partes Interessadas em People Analytics

primento de sua missão institucional. Nesse aspecto, busca-se no ambiente interno e externo, o equilíbrio das iniciativas de RH e o atingimento do resultado corporativo. Portanto, ainda na fase do entendimento do negócio e dos processos de recursos humanos na Apex-Brasil, deverão ser alinhadas as expectativas e necessidades.

2.2.3 Habilitadores e infraestrutura

Entre os principais habilitadores para a implantação de um modelo de *People Analytics* estão o time de analistas de dados, a infraestrutura de dados e tecnológica e patrocínio da alta administração [15]. Para tal exercício, faz-se necessário o entendimento das condições necessárias para a adoção das melhores práticas, conforme abaixo[20]:

A integração dos dados de diversos sistemas e visões sobre os empregados é um desafio, que deve ser objeto de cautela durante a fase de entendimento, pré-processamento e análise exploratória dos dados, pois a compreensão dos dados e seu compartilhamento ainda são uma oportunidade de melhoria no contexto empresarial, dado os critérios de privacidade de dados dada pela Lei Geral de Proteção de Dados (LGPD) no Brasil e pela *Europe*

	IT Infrastructure	Culture	Institution
Data Management	▪ Build the integrated database ▪ Connect the data system for interoperability ▪ Establish the analytics tools for big data processing	▪ Reach consensus on the necessity of data collection and integration ▪ Cultivate positive attitude for data sharing	▪ Formulate the guideline for data format standardization ▪ Establish regulations for data quality management and data sharing
Staff Capabilities	▪ Design the training program for improving the digital literacy ▪ Position data experts in each division as well as the HR department	▪ Give autonomy to data analysis staff ▪ Tolerate trial and error/failure	▪ Define the analysis competency ▪ Implement the hiring process for the right talent ▪ Offer reasonable compensation package for data experts
Acceptance	▪ Disclose the analysis process and outcome in a transparent manner ▪ Form task force as HR analytics agent	▪ Reflect insights on decision-making ▪ Leverage top management support ▪ Cultivate collaboration between silos	▪ Institutionalize data engagement ▪ Share best practices ▪ Give incentives for analytics adoption

Figura 2.6: Condições Necessárias para a adoção de *HR Analytics*

General Data Protection Regulation (GDPR) na Europa[22], bem como já explicado os fatores competitivos[19].

A escolha por técnicas de análise de dados aplicadas à gestão de pessoas, no contexto de People Analytics, está alinhada com a abordagem defendida que destaca a competitividade orientada por analytics [23], e que detalham metodologias específicas para geração de valor em gestão de pessoas baseada em dados [9].

A capacitação e a formação do time de analistas de dados, ainda requer atenção no contexto, pois as competências de analistas de dados são fundamentais para o sucesso da iniciativa. Outro ponto crucial, é a cultura de tentativa e erro, que somente através do empirismo e a tentativa de diversos modelos, poderá se chegar ao melhor para os dados de determinada organização. A necessidade de compartilhamento de conhecimento e de práticas de sucesso é preponderante para a organização.

E, por fim, a disponibilidade de apoio e patrocínio para estabelecer as tecnologias necessárias, bem como a incorporação dos novos modelos no processo decisório. Conclui-se que a disponibilização das condições acima listadas também são fatores críticos de sucesso para a pesquisa.

2.2.4 Governança dos dados de RH

O direcionamento, monitoramento e avaliação constantes das práticas de *People Analytics* são importante fator de implantação do modelo, na fase de Proposição do plano de implantação dos modelos no ambiente da Apex-Brasil deverão ser contempladas práticas

que permitam estas ações. São fatores que incentivam a comunicação, colaboração e a mudança de cultura [19] e [20]. Devendo ser praticadas ao longo da implantação dos modelos e a melhoria do ambiente tecnológico e do processo decisório.

É esperado que, ainda na gestão de partes interessadas, esta frente melhore a relação com a alta administração e com os principais impactos pelas entregas e resultados corporativos [16]. A gestão dos times de usuários de RH e de tecnologia da informação também é fundamental para a implantação dos modelos, dada a complexidade e a curva de aprendizado necessária.

2.2.5 Ética na gestão de dados de RH

Tendo em vista o crescimento do campo de aplicação de *People Analytics* e seu impacto social, o tema de ética na governança e aplicação dos dados dos colaboradores é aspecto crítico que deve ser observado durante a pesquisa. Alguns fatores já são evidenciados na literatura acadêmica quanto ao tema[24]:

1. A opinião pública está mais crítica com práticas corporativas desleais;
2. Leis e regulações para a proteção de dados pessoais estão em evidência e preveem punição aos desvios;
3. No contexto internacional, a atuação em alguns países que não respeitem este preceito, podem impactar nas cadeias globais de suprimentos;
4. A comunicação digital expõe as organizações que não lidam com os riscos de informação;
5. O impacto de imagem e reputação corporativa é mais evidente com a propagação de histórias que expõe os empregados e as organizações; e
6. Processos de empregados estão em crescimento, caso haja abuso dos dados pessoais, principalmente em questões de gênero e raça.

Dado o cenário acima, os bancos de dados dos empregados serão mantidos em sigilo para manter a privacidade e o *compliance* da Apex-Brasil, não sendo divulgados ao final dos trabalhos. Somente serão publicados os achados e os algoritmos utilizados.

2.3 Mineração de Dados

A análise e extração de dados dos diversos sistemas de recursos humanos existentes é um desafio significativo, devido ao entendimento necessário de cada contexto, à pequena

quantidade de colaboradores e à diversidade de atributos existentes. Este trabalho se propõe a explorar técnicas de agrupamento, classificação e predição de séries temporais para enfrentar esses desafios.

Na fase de Entendimento, pré-processamento e análise exploratória dos dados, pretende-se analisar diversos modelos de agrupamento, classificadores e predição de séries temporais na busca dos modelos mais adequados ao banco de dados da Apex-Brasil, no que tange o Absenteísmo. Neste momento serão observados as vantagens, as desvantagens, os usos recomendados e os critérios de avaliação de cada técnica, aceitando ou rejeitando a sua implementação. Quando houver mais de uma técnica relevante, as mesmas serão comparadas para verificar a melhor adequação ao contexto da Agência.

Retomando, as técnicas de agrupamento ou clusterização são importantes para a segmentação dos empregados e entendimento das caraterísticas de cada agrupamento, de forma hierárquica ou não hierárquica, definindo referencias de atuação diferentes para diferente necessidades. As técnicas de classificação são cruciais para a predição de comportamentos, performances ou outras necessidades, buscando aplicar a segmentação em perfis de empregados novos. Já as técnicas de predição de séries temporais são importantes para predizer, considerando a sazonalidade, a previsão de valores ou faltas, possibilitando a atuação antes do fato [22][25].

Em termos de agrupamento, serão analisados as técnicas K-means[26], a Clusterização Hierárquica[26] e o métodos *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN). Para classificadores, serão testados as técnicas: Regressão Logística[22], Árvores de Decisão[27], Random Forest[27] e *K-Nearest Neighbors* (KNN)[27]. E para a predição de séries temporais, serão utilizados as técnicas[28]: *Single Exponential Smoothing* (SES), Holt, Holt-Winters, *Error, Trend, Seasonality* (ETS), *Autoregressive Integrated Moving Average* (ARIMA) e *Seasonal ARIMA* (SARIMA). Ao final, será ajustada cada técnica, com a análise de resultados, discussão de resultados e proposta de implantação na Apex-Brasil.

A seguir segue o detalhamento técnico de cada técnica.

2.3.1 Técnicas de Agrupamento

As técnicas de agrupamento são essenciais para identificar padrões e estruturas ocultas nos dados. As técnicas permitem separar grupos de colaboradores em função dos dados analisados, possibilitando o desenvolvimento de políticas e planos de ação customizados para cada público-alvo. É necessário estudar diferentes técnicas para verificar qual é a mais adequada para a Agência e seus desafios.

Agrupamento é uma técnica de *machine learning* não supervisionada usada para agrupar dados em subconjuntos, onde os dados em cada subconjunto (ou *cluster*) são mais

semelhantes entre si do que aos dados em outros agrupamentos. É usado para segmentação de clientes, onde os comportamentos de uso podem ser mapeados, criando ações voltadas às necessidades dos clientes.

Pode facilitar a visualização de grupos de clientes, bem como identificar os maiores demandantes. Testar o melhor algoritmo é essencial para identificar sua adequação e os resultados propostos. Em RH, será usado para mapear o perfil dos funcionários e identificar características comuns, facilitando a segmentação de grupos sociais e a possibilidade de respostas personalizadas para as políticas de RH.

Para todos os algoritmos selecionados, temos a estrutura de vantagens, desvantagens, usos recomendados e critérios de avaliação de sucesso. Tais informações serão essenciais para a comparabilidade dos resultados e a seleção de qual(ais) técnica(s) serão mais benéficas para o contexto da Apex-Brasil no domínio analisado.

Técnica K-means

O K-means é um algoritmo de agrupamento, com um esquema de aglomeração não hierárquico, que particiona os dados em k agrupamentos, minimizando a variância dentro de cada cluster [29].

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} \|x_j^{(i)} - \mu_i\|^2 \tag{2.1}$$

onde μ_i é o centróide do i-ésimo cluster e $x_j^{(i)}$ são os pontos pertencentes ao cluster i [29].

$$N_\epsilon(p) = \{q \in D | \|p - q\| \leq \epsilon\} \tag{2.2}$$

onde $N_\epsilon(p)$ é o conjunto de pontos dentro da distância ϵ de p

Análise de Pertinência [30][29][31]:

- Vantagens: Simplicidade na implementação; eficiência em grandes bancos de dados; e rapidez na convergência;
- Desvantagens: Sensível a *outliers*; necessidade de escolher k previamente; e forma esférica dos agrupamentos pode não refletir a realidade dos dados;
- Usos Recomendados: Agrupamento de dados homogêneos e esféricos; segmentação de mercado; e redução dimensional em dados esparsos.
- Critérios de Avaliação:

Índice de Silhueta: Mede a qualidade dos agrupamentos com base na proximidade entre pontos dentro do mesmo cluster e entre agrupamentos diferentes.

Coeficiente de Variação Intracluster: Mede a homogeneidade dentro dos agrupamentos.

Sum of Squared Errors (SSE): Soma dos quadrados das distâncias dos pontos à centróide do cluster.

Técnica de Clusterização Hierárquica

A Clusterização Hierárquica cria uma árvore de agrupamentos, facilitando a visualização das relações entre diferentes grupos de dados [32]. A Clusterização Hierárquica cria uma hierarquia de agrupamentos, agrupando dados de forma aglomerativa ou divisiva. Não requer a definição prévia de K e demonstra a hierarquia dos agrupamentos. Permite explorar diferentes granularidades de clusterização. Pode ser computacionalmente caro para grandes conjuntos de dados e a interpretação da hierarquia pode ser complexa.

O dendrograma de Clusterização Hierárquica representa a estrutura de agrupamento dos dados, utilizando o método hierárquico aglomerativo. As alturas das linhas indicam a distância entre os agrupamentos quando eles são unidos. O eixo vertical ("Height") mostra a similaridade entre os grupos, com valores mais altos representando uniões entre agrupamentos mais distintos.

Análise de Pertinência [33][34][35]:

- Vantagens: Fácil visualização das relações entre agrupamentos; não requer número de agrupamentos pré-definido; e adequado para dados de pequeno a médio porte;
- Desvantagens: Não escalável para grandes conjuntos de dados; pode ser influenciado por *outliers*; e sensível à escolha da métrica de distância;
- Usos Recomendados: Análise exploratória de bancos de dados pequenos a médios; construção de árvores filogenéticas; e segmentação de clientes;
- Critérios de Avaliação:

  Índice de Silhueta: Mede a qualidade dos agrupamentos com base na proximidade entre pontos dentro do mesmo cluster e entre agrupamentos diferentes;

  Índice de Davies-Bouldin: Mede a compacidade e separação dos agrupamentos; e

  Dendrogramas: Visualização hierárquica dos agrupamentos.

Técnica DBSCAN

O DBSCAN é um método baseado em densidade que identifica agrupamentos de forma arbitrária, detectando também outliers [36]. O DBSCAN identifica agrupamentos com base na densidade de pontos, descobrindo áreas densas de dados e ignorando pontos esparsos (ruído). Ele depende fortemente dos parâmetros de dois pontos a serem considerados no mesmo cluster (Eps) e do número mínimo de pontos necessários para formar um cluster (MinPts).

$$N_\varepsilon(p) = \{q \in D \mid \|p - q\| \leq \varepsilon\} \tag{2.3}$$

Onde:

- $N_\varepsilon(p)$ é o conjunto de vizinhos de p dentro de um raio ε;
- D representa o conjunto de dados;
- $\|p - q\|$ corresponde à distância entre os pontos p e q, geralmente calculada pela métrica euclidiana;
- ε é um parâmetro que define o raio de vizinhança.

O conjunto $N_\varepsilon(p)$ contém todos os pontos q que satisfazem a condição de estarem a uma distância menor ou igual a ε do ponto p.

Análise de Pertinência [36][37][38]:

- Vantagens: Detecta agrupamentos de forma arbitrária; identifica *outliers*; e não requer número de agrupamentos pré-definido;
- Desvantagens: Desempenho depende de MinPts; dificuldade em lidar com agrupamentos de densidade variável; e sensível à escolha da métrica de distância;
- Usos Recomendados: Dados com agrupamentos de forma arbitrária; detecção de *outliers* em grandes datasets; e análise de imagens e reconhecimento de padrões;
- Critérios de Avaliação:

  Índice de Silhueta: Mede a qualidade dos agrupamentos com base na proximidade entre pontos dentro do mesmo cluster e entre agrupamentos diferentes;

  Número de agrupamentos Detectados: Conta o número de agrupamentos formados pelo algoritmo; e

  Razão de Variação Intracluster: Mede a homogeneidade dentro dos agrupamentos.

A utilização deste algoritmo ajuda a tratar conjuntos de dados com ruído com mais acurácia e facilidade [36]. Portanto, ao analisar a aplicação do algoritmo deve ser estudado o aspecto do ruído e o tratamento de dados que fiquem fora do clusters indicados[27].

2.3.2 Técnicas de Classificação

A classificação é uma técnica de aprendizado de máquina supervisionado usada para prever a categoria a que uma nova observação pertence, com base em um conjunto de dados de treinamento e teste contendo observações cujas categorias são conhecidas e dentro do domínio do pesquisador. Ela permite a previsão de comportamentos esperados, facilitando o processo de tomada de decisão e o tempo de resposta de uma organização.

No contexto de Recursos Humanos, elas devem ser usadas com a devida cautela considerando o perigo de rotular uma pessoa. Elas devem servir para ajudar a identificar situações em que o bem-estar e o desempenho individual podem ser melhorados. Os classificadores são usados para prever a classe de novas observações com base em dados de treinamento. Os modelos considerados incluem:

Técnica de Regressão Logística

A Regressão Logística é um método estatístico para modelar a probabilidade de uma classe binária [39].

$$p(y = 1|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+...+\beta_p x_p)}} \tag{2.4}$$

onde β são os coeficientes do modelo.

Análise de Pertinência [39][40][41]:

- Vantagens: Interpretação fácil; bom para classes lineares separáveis; e estimativas probabilísticas das classes;
- Desvantagens: Não captura relações não lineares; sensível a multicolinearidade; e pode não funcionar bem com grandes bancos de dados desequilibrados;
- Usos Recomendados: Classificação binária ou multiclasse; modelagem de risco de crédito; e análise de sobrevivência;
- Critérios de Avaliação:

  Acurácia: Proporção de previsões corretas;

  Precisão: Proporção de verdadeiros positivos entre os exemplos classificados como positivos;

Revocação: Proporção de verdadeiros positivos entre todos os exemplos positivos;e

F1-Score: Média harmônica entre precisão e revocação.

Técnica de Árvores de Decisão

As Árvores de Decisão segmentam os dados em regiões homogêneas através de um processo recursivo, fácil de interpretar. A impureza do nó pode ser medida pela entropia ou pelo índice de Gini[42] [43].:

$$Gini(p) = 1 - \sum_{i=1}^{k} p_i^2 \tag{2.5}$$

Onde:

- $Gini(p)$ representa o índice de Gini do nó p;
- k é o número de classes distintas no nó;
- p_i é a proporção de observações da classe i no nó p.

O valor do índice de Gini varia de 0 (nó puro, contendo apenas uma classe) até um valor máximo que depende da quantidade de classes. Quanto maior o valor, maior a impureza do nó.

Análise de Pertinência [42][43][44]:

- Vantagens: Fácil interpretação; lida com dados categóricos; e não requer normalização de dados.
- Desvantagens: Propenso a overfitting; sensível a pequenas variações nos dados; e pode ser tendencioso para classes mais frequentes;
- Usos Recomendados: Problemas de classificação com baixa dimensionalidade; análise exploratória de dados; e análise de decisão em negócios;
- Critérios de Avaliação:

  Acurácia: Proporção de previsões corretas;

  Matriz de Confusão: Tabela que mostra as previsões corretas e incorretas;e

  Impureza de Gini: Mede a impureza ou pureza de um nó na árvore.

Técnica Random Forest ou Floresta Aleatória

O Random Forest é um conjunto de árvores de decisão que melhora a precisão e reduz o *overfitting*. A previsão final é feita pela média das previsões de todas as árvores:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i \tag{2.6}$$

onde $\hat{y}_i$ é a previsão da i-ésima árvore [45].

Análise de Pertinência [45][46]:

- Vantagens: Reduz *overfitting*; lida com grandes bancos de dados; e pode ser usado para classificação e regressão;
- Desvantagens: Complexidade e tempo de treinamento; menor interpretabilidade comparada a árvores de decisão individuais; e pode ser computacionalmente intensivo;
- Usos Recomendados: Classificação em problemas complexos e de alta dimensionalidade; análise de dados genômicos; e predição de perda de clientes.;
- Critérios de Avaliação:

  Acurácia: Proporção de previsões corretas;

  Importância das Variáveis: Mede a importância de cada variável na predição; e

  Out-of-bag Error: Estimativa do erro de predição utilizando exemplos não presentes no treinamento de cada árvore.

Técnica KNN

O *K-Nearest Neighbors* (KNN) classifica uma nova observação com base nas k observações mais próximas no espaço dos atributos. A classe é atribuída pela maioria das classes dos vizinhos:

$$\hat{y} = \arg\max_{c} \sum_{i=1}^{k} \mathbb{I}(y_i = c) \tag{2.7}$$

onde $\mathbb{I}$ é a função indicadora [47].

Análise de Pertinência [47][48][49]:

- Vantagens: Simplicidade; não faz suposições sobre a distribuição dos dados; e fácil de entender e implementar;

- Desvantagens: Tempo de computação alto; sensível à escala dos dados; e requer armazenamento de todo o conjunto de dados;
- Usos Recomendados: Classificação com poucos dados e estrutura simples; reconhecimento de padrões; e sistemas de recomendação;
- Critérios de Avaliação:

  Acurácia: Proporção de previsões corretas;

  Matriz de Confusão: Tabela que mostra as previsões corretas e incorretas;

  Curva ROC: Gráfico que mostra a taxa de verdadeiros positivos versus a taxa de falsos positivos.

2.3.3 Técnicas de Predição de Séries Temporais

Para a predição de séries temporais, utilizamos técnicas que capturam padrões temporais nos dados:

Técnica *Single Exponential Smoothing* (SES)

O *Single Exponential Smoothing* (SES) é um método simples que suaviza os dados históricos para fazer previsões. A fórmula de suavização é:

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha)\hat{y}_t \tag{2.8}$$

onde α é o coeficiente de suavização [50].

Análise de Pertinência [50][51]:

- Vantagens: Simplicidade; fácil de implementar; e requer poucos dados históricos;
- Desvantagens: Não captura tendência; não captura sazonalidade; e eficácia limitada para séries temporais complexas;
- Usos Recomendados: Séries temporais sem tendência ou sazonalidade; previsão de vendas a curto prazo; e previsão de demanda para produtos estáveis;
- Critérios de Avaliação:

  Mean Absolute Error (MAE): Média dos valores absolutos dos erros;

  Mean Squared Error (MSE): Média dos quadrados dos erros; e

  Root Mean Squared Error (RMSE): Raiz quadrada da média dos quadrados dos erros.

Técnica Holt-Winters

O modelo de Holt-Winters considera sazonalidade além das tendências:

$$\hat{y}_{t+m} = l_t + mb_t + s_{t+m-L} \tag{2.9}$$

$$l_t = \alpha \frac{y_t}{s_{t-L}} + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{2.10}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \tag{2.11}$$

$$s_t = \gamma \frac{y_t}{l_t} + (1-\gamma)s_{t-L} \tag{2.12}$$

onde s_t é o componente sazonal e L é o período sazonal [52].

Análise de Pertinência [52][50][28]:

- Vantagens: Captura tendências e sazonalidade; adequado para séries temporais sazonais; e flexível para diferentes padrões sazonais;
- Desvantagens: Complexidade aumentada; requer ajuste de mais parâmetros; e pode ser sensível a *outliers*;
- Usos Recomendados: Séries temporais com padrões sazonais fortes; previsão de vendas sazonais; e planejamento de capacidade;
- Critérios de Avaliação:

  Mean Absolute Error (MAE): Média dos valores absolutos dos erros;

  Mean Squared Error (MSE): Média dos quadrados dos erros; e

  Root Mean Squared Error (RMSE): Raiz quadrada da média dos quadrados dos erros.

Técnica *Autoregressive Integrated Moving Average* (ARIMA)

O *Autoregressive Integrated Moving Average* (ARIMA) é um método flexível que combina componentes autorregressivos, médias móveis e integração:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{2.13}$$

onde ϕ são os coeficientes autorregressivos, θ são os coeficientes de médias móveis e ε_t é o erro [28].

Análise de Pertinência [53][54][28]:

- Vantagens: Modelagem flexível de componentes temporais; captura padrões sazonais e de tendência; e ampla aplicabilidade em diferentes domínios;
- Desvantagens: Determinação dos parâmetros pode ser complexa; requer séries temporais estacionárias; e sensível a valores atípicos;
- Usos Recomendados: Séries temporais com dependência temporal forte; previsão de vendas; e análise econômica;
- Critérios de Avaliação:

  Mean Absolute Error (MAE): Média dos valores absolutos dos erros;

  Mean Squared Error (MSE): Média dos quadrados dos erros; e

  Akaike Information Criterion (AIC): Penaliza modelos com muitos parâmetros.

Técnica *Seasonal ARIMA* (SARIMA)

O *Seasonal ARIMA* (SARIMA) estende o ARIMA para capturar sazonalidade nos dados:

$$y_t = c + \phi(B)y_t + \theta(B)\varepsilon_t \tag{2.14}$$

onde $\phi(B)$ e $\theta(B)$ são polinômios de atraso que incluem componentes sazonais [28].

Análise de Pertinência [53][54][28]:

- Vantagens: Captura sazonalidade; extensão do ARIMA para padrões sazonais; e adequado para séries temporais sazonais complexas;
- Desvantagens: Complexidade aumentada; requer ajuste de muitos parâmetros; e sensível a valores atípicos;
- Usos Recomendados: Séries temporais com padrões sazonais regulares; previsão de vendas sazonais; e análise de dados meteorológicos;
- Critérios de Avaliação:

  Mean Absolute Error (MAE): Média dos valores absolutos dos erros;

  Mean Squared Error (MSE): Média dos quadrados dos erros; e

  Akaike Information Criterion (AIC): Penaliza modelos com muitos parâmetros.

2.4 Trabalhos Relacionados

Nesta seção são apresentados trabalhos relacionados ao objeto desta dissertação, identificados mediante a realização de pesquisa bibliográfica nas plataformas *Web of Science* e SCOPUS. Foram utilizadas, sozinhas ou conjugadas, as seguintes palavras-chave: "*People Analytics*", "*HR Analytics*"; "*Workforce Analytics*"; "LGPD", "GDPR", "*data science*", "*clustering*", "*classification*", "*forecast*" e "*time series*".

A partir da pesquisa, auxiliada pelo Método Teoria do Enfoque Meta Analítico (TEMAC) [55] e pela ferramenta *VOS Viewer* [56], software visualizador de *agrupamentos* de citação autores da Universidade de Leiden a partir de dados extraídos das plataformas de trabalhos científicos, foram revisados os referenciais e conceitos teóricos que suportaram este trabalho.

Garg [57] apresenta uma revisão da literatura que identifica contribuições relativas a funções de Recursos Humanos e objetivos de aprendizado de Máquina, neste artigo observa-se que os temas de Classificação e Clusterização estão presentes em grandes partes das funções, possibilitando a formatação de uma análise de *People Analytics*. A revisão sugere que os aplicativos de Aprendizado de Máquina são mais fortes nas áreas de recrutamento e gerenciamento de desempenho, e o uso de árvores de decisão e algoritmos de mineração de texto para classificação dominam todas as funções da Gestão de Recursos Humanos. Para processos complexos, os aplicativos ainda estão em um estágio inicial.

Em artigo de que analisa a crescente introdução de algoritmos de aprendizagem no ambiente corporativo, Giermindl et al. [58] destacam que cada vez mais a análise de pessoas é utilizada para tomar decisões baseadas em evidências e gerenciar sua força de trabalho em uma abordagem orientada por dados. Com base em uma ampla revisão teórica das expectativas organizacionais associadas ao uso de algoritmos de aprendizagem e IA, identificaram perigos da análise de pessoas. Conclui que O uso de *People Analytics* transformará o futuro do trabalho e da tomada de decisão humana.

2.4.1 Agrupamento

No artigo de Berkhin[59], apresenta a técnica de Agrupamento, com a divisão de dados em grupos de objetos similares. No Agrupamento, alguns detalhes são desconsiderados em troca da simplificação dos dados. O Agrupamento pode ser visto como uma técnica de modelagem de dados que fornece resumos concisos dos dados. O Agrupamento está, portanto, relacionado a muitas disciplinas e desempenha um papel importante em uma ampla gama de aplicações.

Andriyani [60] propõe um modelo de avaliação e comparação de resultados de diferentes algoritmos de clusterização, ressaltando a aplicabilidade de diversos modelos e a definição

de qual seria a melhor aplicação. No artigo, são testados os modelos de K-means e DBSCAN.

Em outro trabalho, Gupta [61] define que, na visão dos recursos humanos, o dever mais importante de uma grande empresa é garantir que seus funcionários tenham equilíbrio entre vida pessoal e profissional, mantendo sua imagem de marca preservada, sendo necessário o uso de algoritmos que possam segmentar o empregado.

Kakulapati et al. [62] conduziram estudos para analisar o desempenho dos funcionários aplicando técnicas de Agrupamento com base na similaridade de métricas de desempenho para analisar o desempenho dos funcionários. Além disso, foram utilizadas técnicas de classificação, tais como *Random Forest*, que facilitam a classificação de funcionários com base em sua renda mensal e forma informal de executar análises em dados.

Em proposta de modelo de análise de maturidade de métrica de RH, Lismont et al. [63] realizaram pesquisa descritiva com diversas organizações que utilizam *People Analytics*, entre os modelos pesquisados encontraram uma prevalência de técnicas compreensíveis, como regressão linear e árvores de decisão, técnicas de regressão, análise de sobrevivência, etc. Entre as técnicas mais utilizadas no universo pesquisado, estão: Regressão Logística, Árvores de Decisão, Regressão Linear, Predição de Séries Temporais, Agrupamento Hierárquico e o modelo K-means.

2.4.2 Classificação

Alsaadi et al.[64] define o uso de modelos de Classificação, com a aplicação de algoritmos de Regressão Logística, Árvore de Decisão e Random Forest de forma comparativa para analisar com sucesso práticas de Recursos Humanos, sendo utilizados como métricas de sucesso: Precisão e Acurácia. Yin [65] analisa modelos de classificação para aplicação no contexto de Recursos Humanos, utilizando diferentes modelos e comparando os melhores resultados de forma análoga.

Em estudo de Padama et al. [66], foram utilizados diversos experimentos com conjunto de dados utilizados em algoritmos de classificação, tais como Regressão Logística, KNN, Árvore de Decisão, *Random Forest* e AdaBoost para análise de suas métricas de desempenho.

No estudo *Predicting employee absenteeism for cost effective interventions* de Lawrance et al. [67], foram utilizados métodos de classificação para avaliação de custo de absenteísmo e definição de métricas de suporte ao processo decisório, baseados em algoritmos de classificação.

Fallucchi et al. [68] propõe análise de modelo de classificadores, comparando Acurácia, Precisão, *Recall* e o *F1 Score*. Alcançando sucesso na predição de rotatividade de funcionários usando técnicas de aprendizado de máquina.

2.4.3 Séries Temporais

A utilização de séries temporais para melhorar o desempenho de previsão dos modelos ideais confirma que a abordagem de modelagem de séries temporais tem a capacidade de prever a rotatividade de funcionários para o cenário específico observado neste trabalho, conforme indicam Zhu et al. [69]. Adicionalmente, Bandeira [70] demonstra que a combinação de métodos de previsão demonstra ser valiosa para um esquema de ponderação baseado no desempenho de cada modelo.

Margherita [71] faz revisão de literatura sobre aplicações (descritivas e diagnósticas/prescritivas) e valor (valor do funcionário e valor organizacional). Também especulamos sobre uma visão "exponencial"da análise de RH possibilitada pela afirmação da inteligência artificial e das tecnologias cognitivas, destacando as técnicas de *forecasting* como ferramentas de predição de temas de Recursos Humanos.

No campo de aplicação de técnicas de *forecasting* no tema de *people analytics* Xu et al. [72] realizaram análise de fluxo de talentos no campo de planejamento de recursos humanos, monitoramento de fuga de cérebros e previsão de força de trabalho futura, com a abordagem de uso de previsão de séries temporais de várias etapas. Os resultados também indicaram que o modelo proposto pode fornecer desempenho razoável mesmo se os dados históricos de fluxo de talentos não estiverem completamente disponíveis.

Capítulo 3

People Analytics: Estudo de Caso de Absenteísmo na Apex-Brasil

Nesta seção são apresentados resultados preliminares obtidos a partir de análises realizadas nas bases de dados identificadas a seguir, cujos códigos encontram-se disponíveis no GitHub [1].

3.1 Entendimento do negócio e do contexto dos processos de recursos humanos na Apex-Brasil

O escopo deste estudo foi definido a partir de reuniões realizadas com as equipes de Recursos Humanos (RH) e Tecnologia da Informação (TI) da Apex-Brasil. Ficou acordado que seriam utilizados dados dos colaboradores ativos em 31/12/2023, no período entre 01/2019 e 12/2023.

A gestão do absenteísmo, sob a perspectiva da Teoria Geral dos Sistemas [73], deve ser compreendida como um subsistema aberto que interage com o ambiente organizacional e social. Sua estrutura abrange os elementos institucionais que viabilizam a coleta e análise dos dados. A função está atrelada ao suporte gerencial e estratégico por meio da identificação de padrões e impactos. Os processos descrevem o fluxo de informações e a retroalimentação de decisões.

Por fim, o contexto contempla os fatores internos e externos que influenciam a ausência dos colaboradores, como condições de trabalho, clima organizacional e fatores socioeconômicos. Essa abordagem sistêmica permite uma compreensão mais profunda e integrada do fenômeno, contribuindo para ações mais eficazes e sustentáveis no âmbito da gestão de pessoas.

[1] `https://bit.ly/3Mys63c`

A segmentação de absenteísmo por grupos, dada a estrutura de dados disponível, e sua classificação aos novos empregados, propicia o entendimento de fatores correlatos que suportem a prevenção e o tratamento de desvios, suportando o processo decisório e as políticas de RH da Agência. A predição no tempo favorece a antecipação de possíveis fenômenos climáticos ou endemias ou epidemias, podendo-se aplicar a política de vacinação antecipada.

Além disso, definiu-se exercícios de clusterização e classificação, buscando os modelos mais adequados para o contexto da Agência, além de realizar uma previsão de absenteísmo e licenças para os anos de 2024 e 2025, de forma que a técnica mais adequada possa ser verificada para este propósito e futura comparação com os resultados obtidos.

Portanto, no contexto de *People Analytics*, a Apex-Brasil, com base neste estudo, passará a contar com as técnicas de ciência de dados mais adequadas ao seu ambiente interno, subsidiando políticas de RH e planos de ação [25] [22].

Pode ser enquadrado dentro da lógica da cadeia de valor, onde o sistema atua na intermediação e captura de dados transacionais provenientes de múltiplas fontes internas e externas, como folha de pagamento, controle de ponto e bases públicas como o eSocial. Na etapa de processamento interno dos dados, inclui-se a limpeza, padronização, transformação e aplicação de modelos analíticos e estatísticos, como agrupamento, classificação, predição e análise descritiva. Na fase de solução, o sistema entrega produtos informacionais, como *dashboards*, relatórios analíticos e perfis preditivos, que representam soluções diretamente aplicáveis às necessidades da gestão de pessoas. Finalmente, esses *insights* são disponibilizados às lideranças e áreas demandantes, transformando-se em suporte efetivo para a tomada de decisão, fechando o ciclo de valor e convertendo dados em inteligência organizacional aplicável.

3.2 Entendimento, pré-processamento e análise exploratória dos dados

Nesta etapa, os dados foram extraídos, transformados e carregados (ETL). Após o carregamento dos dados no RStudio, foi realizada a Análise Exploratória de Dados (AED), cujos resultados preliminares são apresentados a seguir.

3.2.1 Extração de Dados

Utilizando o modelo de relatórios do Sistema Integrado Totvs RM - Gestão de Pessoas, foram realizadas consultas no banco de dados, com a extração da amostra de dados, caracterizada na seção anterior. Foram gerados dois arquivos, com a seguinte configuração:

- **Conjunto de Dados 1** - Informações dos Colaboradores ("BD-APEX-SUM"): com 58 atributos e 336 linhas, contendo as principais informações dos colaboradores ativos, abrangendo: dados pessoais (tratados com privacidade e removidos na fase de transformação), dados de absenteísmo e remuneração. Após a transformação dos dados, restaram 10 atributos e 335 linhas;
- **Conjunto de Dados 2** - Informações de absenteísmo ("ATESTADOS-2019-2023"): com 18 atributos e 7689 linhas, contendo informações de séries temporais de absenteísmo, com motivos e histórico de 2019 a 2023. Após a análise de transformação, restaram 3 atributos e 7.664 linhas.

3.2.2 Transformação dos Dados

Na fase de transformação dos dados, no banco de informações dos colaboradores, foram geradas novas colunas para facilitar a análise, como informações consolidadas de cônjuge e companheiro, transformação dos dados de gênero e controle de entrada e saída do trabalho como fator. Foram criadas faixas informativas para os atributos salário, idade e total de horas de absenteísmo, para facilitar a visualização e categorização dos colaboradores.

A seguir, estão as variáveis transformadas com as respectivas unidades de medida do conjunto de dados 1:

Conjunto de Dados 1 - Tipos de Licença

- ***Idade***: convertida para faixa etária (em anos)
- ***Dependentes***: total de dependentes (em unidades)
- ***Tempo de Casa***: tempo de serviço na organização (em anos)
- ***Salário Mensal***: salário mensal (em Reais)
- ***Total Horas***: total de horas de absenteísmo (em horas)
- ***Gênero***: gênero (Masculino/Feminino)
- ***Cargo***: tipo de cargo ou função
- ***Faixa Etária***: faixa etária (agrupada em categorias)
- ***Faixa de Horas***: faixa de horas de absenteísmo
- ***Cônjuge***: tipo consolidado de união civil (em unidades)

A seguir, estão as variáveis transformadas com as respectivas unidades de medida do conjunto de dados 2:

Conjunto de Dados 2 - Informações de Séries Temporais de Absenteísmo

- ***Data de Início***: data de início do atestado médico
- ***Data de Término***: data de término do atestado médico
- ***Total de Horas de Absenteísmo***: Detalhamento do total de horas de absenteísmo por dia

3.2.3 Carga de Dados

Para carregamento e paralelização do processamento das informações foi utilizada a Linguagem **R**, com as bibliotecas ***Sparklyr*** e ***Dyplir*** [74] no ***RStudio***, possibilitando o uso do sistema ***Spark*** e o respectivo processamento dos dados. Ambos os conjuntos de dados foram tratados para lidar com dados ausentes e/ou faltantes. Após a transformação dos dados, finalizou-se o processo de ETL, com a classificação das variáveis em numéricas e categóricas.

3.2.4 Análise Exploratória de Dados (AED)

Nesta seção realizou-se a análise exploratória de dados (AED). Foram classificadas como numéricas, as seguintes variáveis: Idade, Dependentes, Tempo de Casa, Salário Mensal, Horas de Absenteísmo. Em termos de análise das variáveis numéricas, serão analisadas de forma descritiva as seguintes informações:

- Medidas de Tendência Central: Média, mediana, moda, valor mínimo, valor máximo;
- Medidas de Dispersão: Desvio-padrão, amplitude, erro padrão e variância;
- Mapa de Calor da Matriz de Correlação de Variáveis Numéricas;
- Distribuição de Frequências: Histogramas, boxplots, gráficos de densidade.

E, de forma análoga, foram classificadas como categóricas as seguintes variáveis: Gênero, Cargo, Faixa Etária, Faixa de Minutos e Cônjuge. Serão analisadas as seguintes informações:

- Tabelas de Frequência: Frequência absoluta e relativa de cada categoria;

- Gráficos: Gráficos de barras;e
- Teste de Associação ou Qui-Quadrado para verificação de independência nas variáveis categóricas.

Dados os resultados, foram propostas técnicas de análise bivariada, bem como modelagem e inferência estatística, quando possível.

Análise Descritiva de Variáveis Numéricas

Abaixo apresenta-se a Tabela [3.1] que demonstra as estatísticas descritivas das variáveis numéricas em um conjunto de dados com 335 observações.

Tabela 3.1: Estatísticas Descritivas das Variáveis Numéricas

Estatísticas	**Idade**	**Dependentes**	**Tempo de Casa**	**Salário Mensal**	**Horas de Absenteísmo**
Nº de Observações	335	335	335	335	335
Média	42,8	0,8	8,5	19451,7	231,9
Desvio-Padrão	10,0	1,3	5,7	11314,0	301,5
Mediana	42,0	0,0	8,0	17804,0	159,9
Valor Mínimo	25,0	0,0	0,0	3259,9	0,0
Valor Máximo	79,0	6,0	23,0	71740,1	2485,6
Erro Padrão	0,5	0,1	0,3	618,1	16,5

Pode-se observar na Figura [3.1], que existem níveis de correlação moderada entre a variável *Idade* com o *Tempo de Casa* ($r = 0.41$) e *Idade* e *Salário Mensal* ($r = 0.47$), sugerindo que, em média, quanto maior a idade dos indivíduos, maior o tempo de casa e maior o salário mensal. A correlação moderada entre *Salário Mensal* e *Tempo de Casa* de $r = 0.40$, o que pode estar relacionada à progressão na carreira ou aumentos salariais com o tempo.

A correlação entre *Dependentes* e outras variáveis denota correlação baixa, o que sugere que o número de dependentes não está fortemente associado à *Idade*, *Tempo de Casa* ou *Salário Mensal*. Enquanto existe correlação negativa baixa com *Horas de Absenteísmo* com *Salário Mensal* ($r = -0.18$), sugerindo que funcionários com salários mais altos tendem a ter menos minutos de absenteísmo.

A análise de correlação entre *Tempo de Casa* e outras variáveis apresenta-se moderada com *Idade* e *Salário Mensal*, mas uma correlação muito baixa com *Horas de Absenteísmo* e *Dependentes*, sugerindo que o tempo de permanência na empresa está mais relacionado ao perfil do funcionário (*Idade* e *Salário Mensal*) do que ao comportamento de ausência ou número de dependentes.

Figura 3.1: Mapa de Calor da Matriz de Correlação

Em resumo, as correlações mais significativas são entre *Idade* e *Salário Mensal*, *Idade* e *Tempo de Casa* e *Salário Mensal* e *Tempo de Casa*. Isso indica que essas características estão inter-relacionadas e podem ser exploradas em análises mais profundas para entender melhor o perfil dos funcionários em relação à idade, tempo na empresa e salário.

Conforme Tabela [3.1], a Apex-Brasil apresenta o seguinte perfil de colaboradores por "Idade", que a idade média é de 42,8 anos com um desvio-padrão de 10,0 anos. A amplitude é de 54 anos, com valores entre 25 e 79 anos. Observa-se na Figura [3.2] a distribuição da variável, bem como sua curva de densidade.

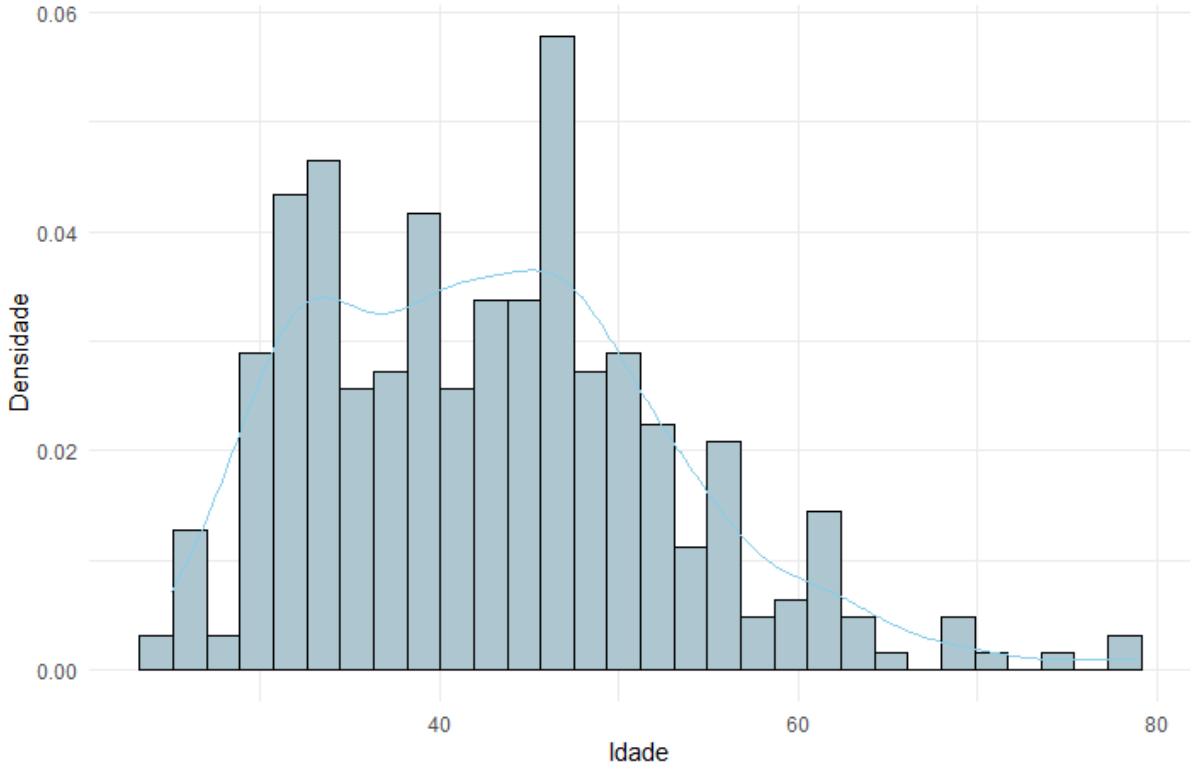

Figura 3.2: Gráfico de Distribuição e Densidade de Idade

Pela Figura [3.3], verifica-se que a mediana está em torno de 40 anos, o que indica o ponto central dos dados. Há alguns valores extremos, indicando que há alguns funcionários com idades significativamente mais altas do que os demais, próximas de 80 anos. A ausência de valores extremos na parte inferior sugere uma assimetria nos dados, com uma cauda à direita.

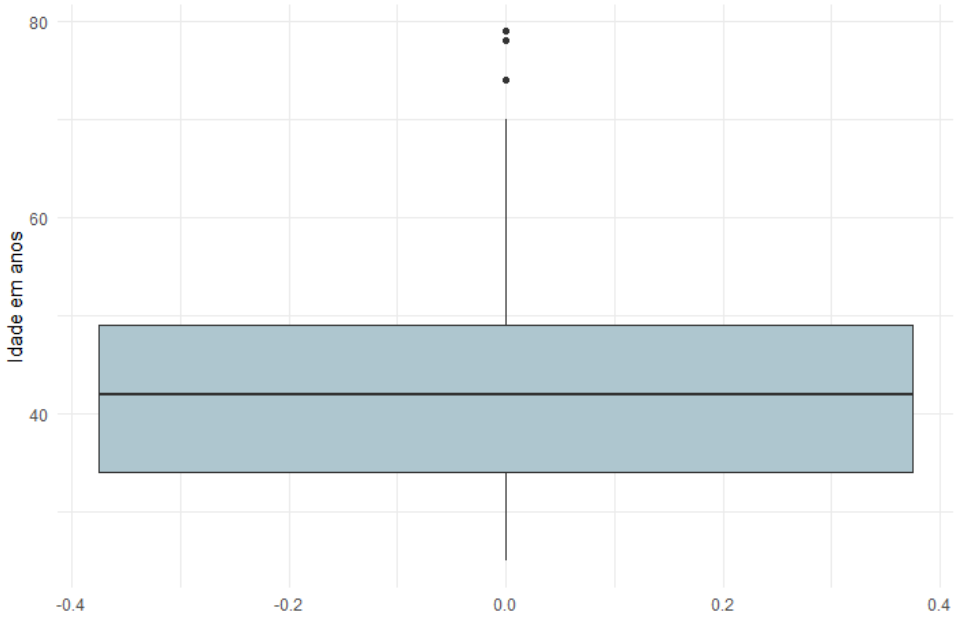

Figura 3.3: Análise Boxplot de Idade

Conforme Tabela [3.1], verifica-se que os colaboradores da Apex-Brasil possuem, em média, 0,8 dependentes, com um desvio-padrão de 1,3, sugerindo que a maioria das pessoas possui poucos dependentes. O valor máximo é de 6, mostrando que algumas pessoas possuem um número mais elevado de dependentes, mas a mediana de 0 indica que pelo menos metade das pessoas não possui dependentes. Observa-se na Figura [3.4] a distribuição da variável, bem como sua curva de densidade.

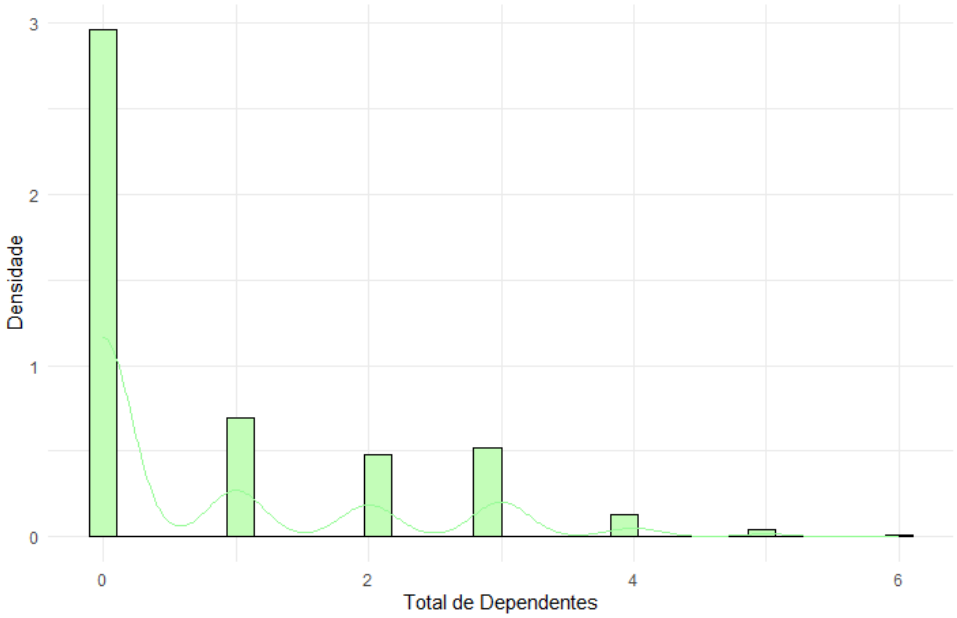

Figura 3.4: Gráfico de Distribuição e Densidade de Dependentes

Pela Figura [3.5], verifica-se uma concentração de dados na parte inferior da escala, com a maioria das observações próximas entre 0 e 1. Vale destacar que, sob o ponto de vista estatístico, possuir 3 ou mais filhos é discrepante em relação ao perfil dos colaboradores da Apex-Brasil.

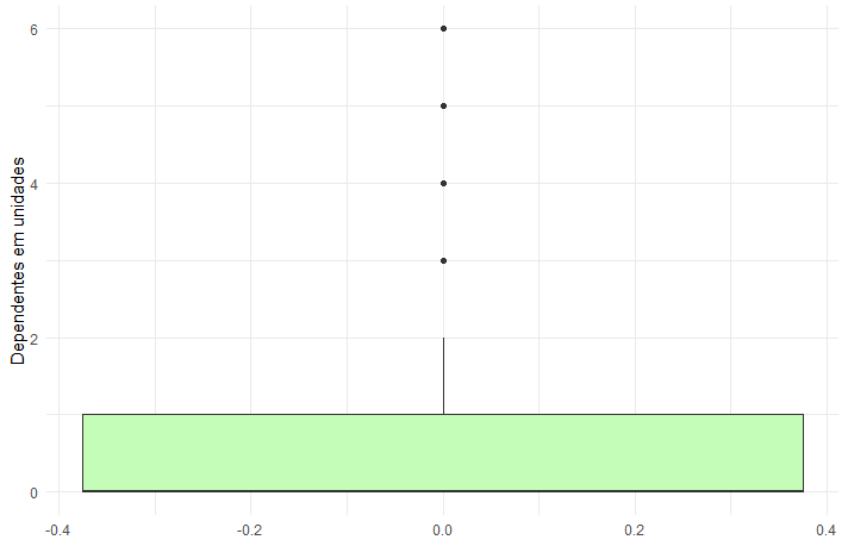

Figura 3.5: Análise Boxplot de Dependentes

Conforme Tabela [3.1], destaca-se que o tempo médio de permanência na Apex-Brasil é de 8,5 anos, com um desvio-padrão de 5,7 anos. A amplitude é de 23 anos, variando entre 0 e 23 anos, sugerindo um grupo misto de funcionários mais novos e antigos. Observa-se na Figura [3.6] a distribuição da variável, bem como sua curva de densidade.

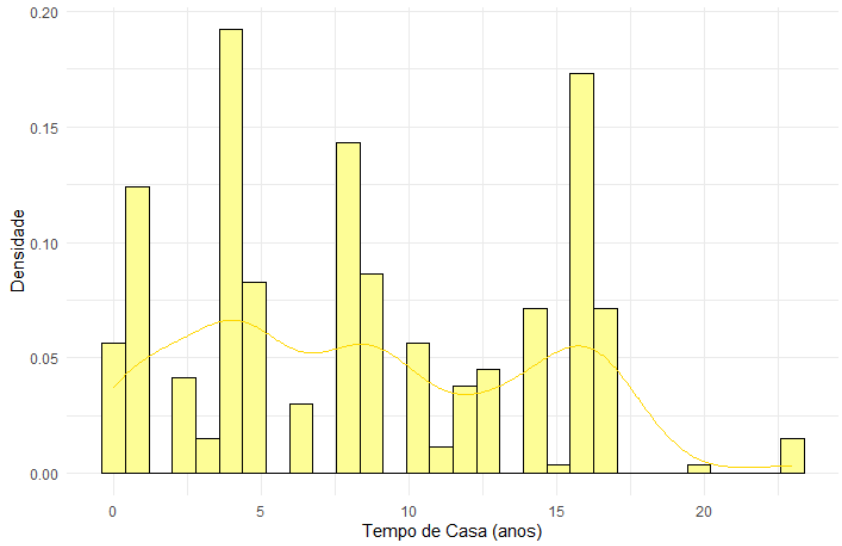

Figura 3.6: Gráfico de Distribuição e Densidade de Tempo de Casa

A Figura [3.7] apresenta o boxplot da variável *Tempo de Casa* na Agência. A mediana está próxima de 5 anos, indicando que metade dos colaboradores trabalha há menos de cinco anos na Apex-Brasil.

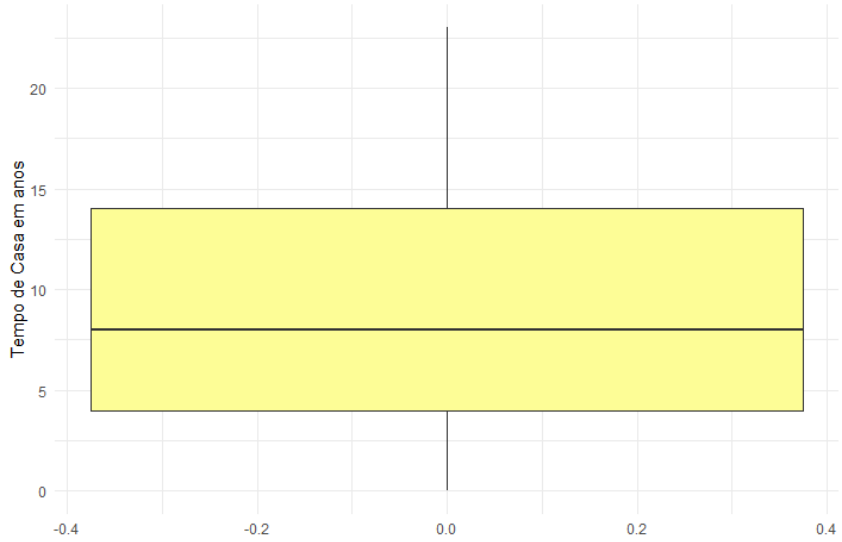

Figura 3.7: Análise Boxplot de Tempo de Casa

Conforme Tabela [3.1], a Apex-Brasil apresenta média salarial de 19.451,70 reais, com uma ampla variação (desvio-padrão de 11.314,00 reais), o que indica uma distribuição salarial desigual. A amplitude de 68.480,20 reais sugere uma grande disparidade salarial, desde o mínimo de 3.259,90 reais até o máximo de 71.740,10 reais. Observa-se na Figura [3.8] a distribuição da variável, bem como sua curva de densidade.

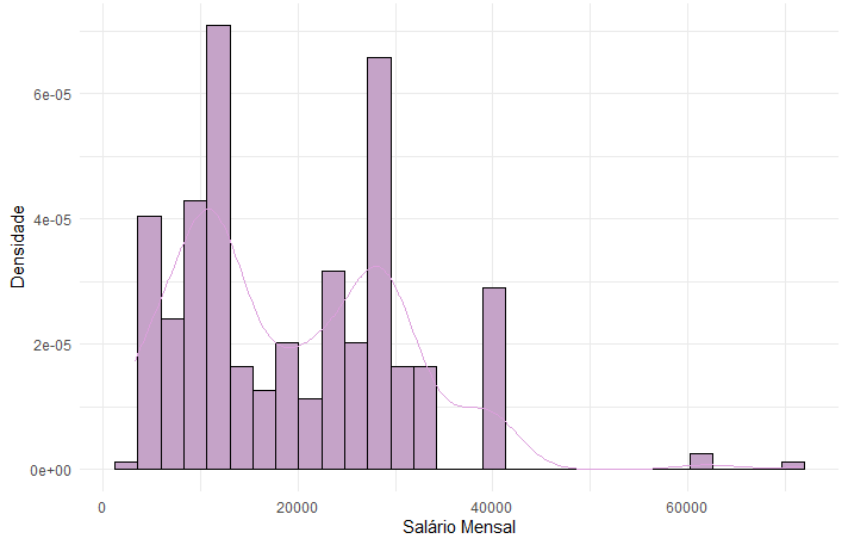

Figura 3.8: Gráfico de Distribuição e Densidade de Salário Mensal

Abaixo no Gráfico Boxplot na Figura [3.3], revela algumas características importantes sobre a distribuição do Salário Mensal pago na Apex-Brasil. Cerca de 25% dos colaboradores ganham entre R$ 10 mil e R$ 17,8 mil. A distribuição assimétrica e a presença de *outliers* sugerem que o uso de técnicas robustas ou transformações nos dados pode ser necessário para análises adicionais que requerem normalidade, apresentando uma cauda à direita.

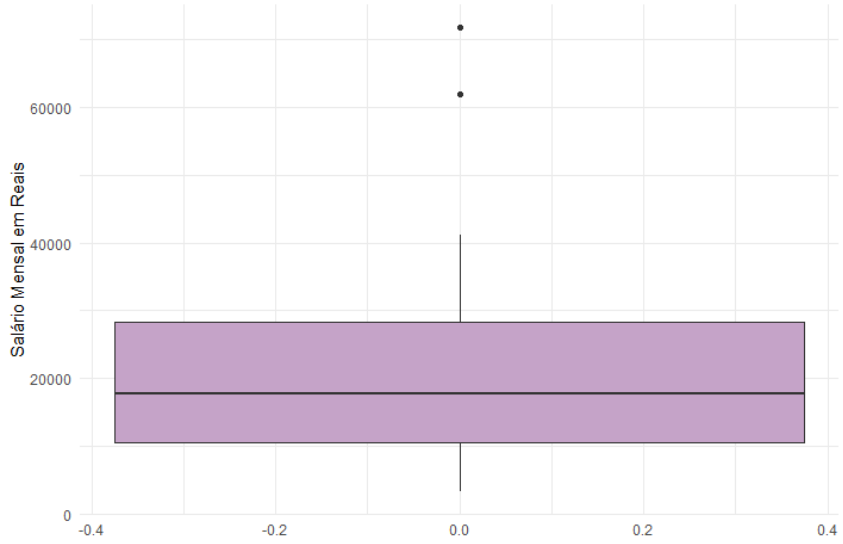

Figura 3.9: Análise Boxplot de Salário

Conforme Tabela [3.1], a Apex-Brasil apresenta o absenteísmo médio de 231,9 horas ou 9,7 dias, com um elevado desvio-padrão de 301,5 horas ou 12,6 dias. A amplitude é extremamente grande (103,6 dias), indicando que alguns funcionários apresentam níveis de absenteísmo muito altos. Esses dados mostram que a distribuição das variáveis apresenta considerável variação em todas as categorias, sugerindo grupos diversos, tanto em idade e salário quanto em tempo de casa e absenteísmo. Isso pode refletir diferenças nas práticas de contratação e perfis de funcionários dentro da organização, além de possíveis variáveis externas que influenciam o absenteísmo e a variação salarial. Observa-se na Figura [3.10] a distribuição da variável, bem como sua curva de densidade.

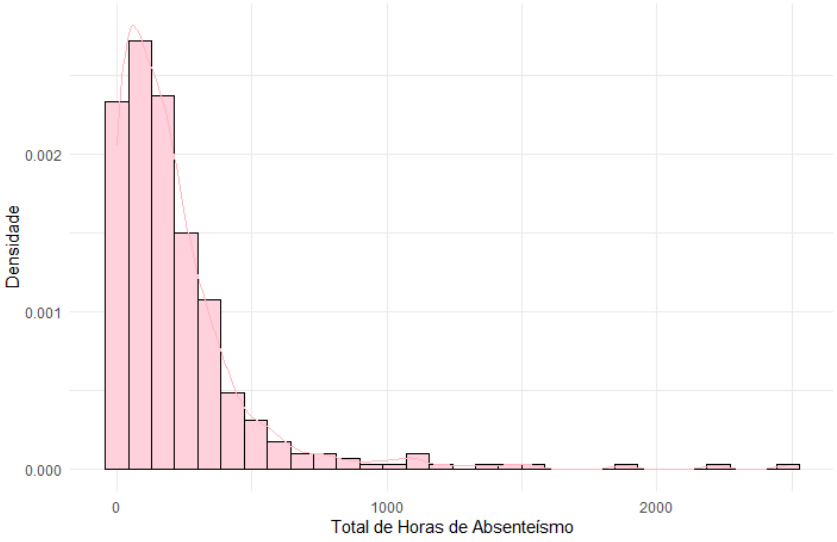

Figura 3.10: Gráfico de Distribuição e Densidade de Horas de Absenteísmo

Abaixo, no Gráfico Boxplot na Figura [3.11], revela uma distribuição com muitos valores extremos (*outliers*) acima do limite superior da caixa. A maioria dos valores está concentrada dentro do intervalo interquartil, próximo da base da caixa, sugerindo que a maior parte das observações de absenteísmo é relativamente baixa. No entanto, existem diversos pontos bem acima da caixa, indicando altos níveis de absenteísmo para algumas observações. Esses valores extremos elevam significativamente a variabilidade dos dados e sugerem uma distribuição assimétrica com uma cauda longa à direita.

Figura 3.11: Análise Boxplot de Horas de Absenteísmo

Análise Descritiva de Variáveis Categóricas

A tabela [3.1] apresenta as frequências das variáveis categóricas em um conjunto de dados com 335 observações.

Quanto ao *Gênero*, a Figura [3.12] indica uma leve predominância do gênero masculino (53,4%).

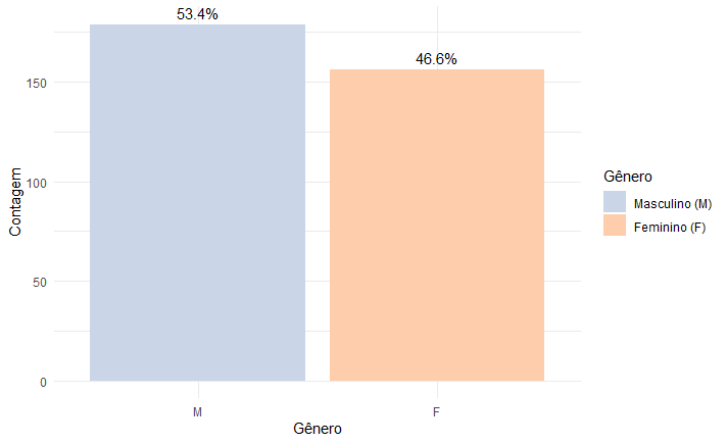

Figura 3.12: Distribuição por Gênero

A Agência apresenta uma distribuição dos tipos de cargos, conforme a Figura [3.13], que revela uma predominância do cargo de *Analista*, que representa 56% do total. Em seguida, temos os cargos de *Assistente* e *Coordenador*, com 17% e 12%, respectivamente, sugerindo um quadro de pessoal com uma proporção considerável de funções de suporte e supervisão.

A camada intermediária inclui *Gerente* e *Assessor*, com 7% e 5%, indicando uma presença menor de cargos de confiança. Já os cargos de *General Manager*, *Diretor*, *Secretária*, *Officer* e *Presidente* possuem uma representação mínima, cada um com menos de 1%, o que é típico em posições de alta liderança e executivas. Esses dados sugerem uma estrutura organizacional piramidal clássica, com foco maior em funções técnicas e de suporte e menor concentração em níveis de liderança, característica comum em operações que demandam um grande número de colaboradores na base operacional.

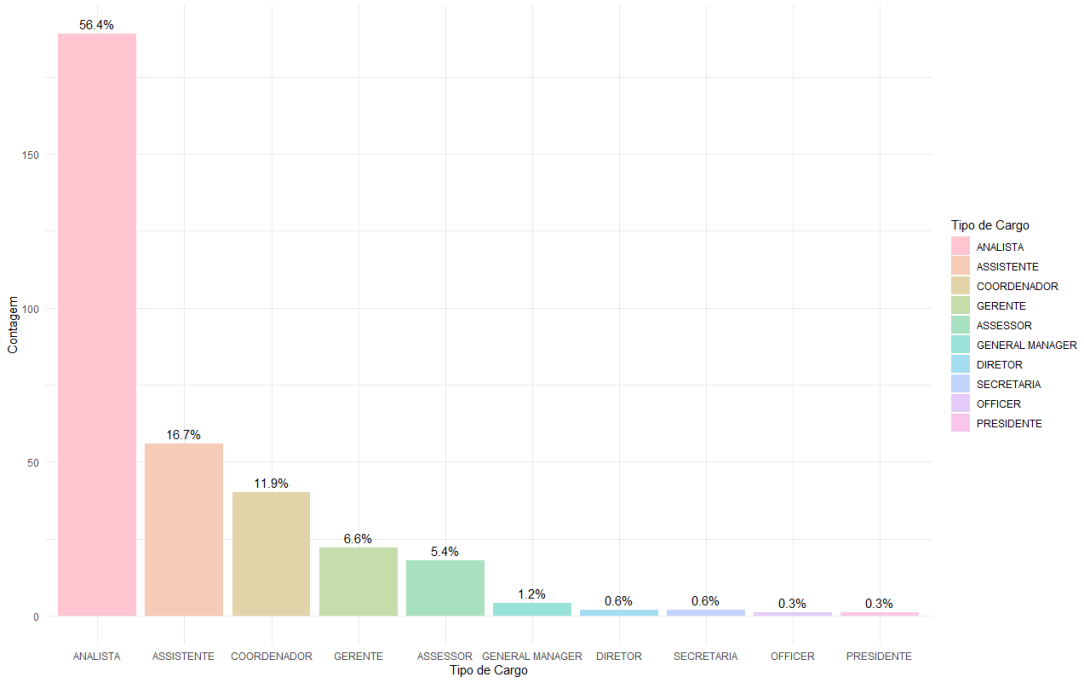

Figura 3.13: Distribuição por Cargo

Pela Figura [3.14], verifica-se que a maior parte dos indivíduos está concentrada nas faixas de *31-40* e *41-50 anos*, que representam 34% e 36% do total, respectivamente. Isso sugere que a organização possui uma força de trabalho majoritariamente composta por indivíduos de meia-idade, o que pode refletir uma base de funcionários experientes e em estágios de carreira mais avançados. A faixa etária "51-60 anos"também tem uma presença significativa, representando 0,17 do total, indicando um grupo adicional de colaboradores em idade madura. Faixas etárias mais jovens, como *21-30 anos* (6%), e as mais avançadas, como *61-70 anos* (5%) e *71-80 anos* (1%), têm uma representação muito menor. Essa distribuição etária sugere uma organização com experiência consolidada, mas com menor proporção de novos talentos ou colaboradores em idade próxima da aposentadoria.

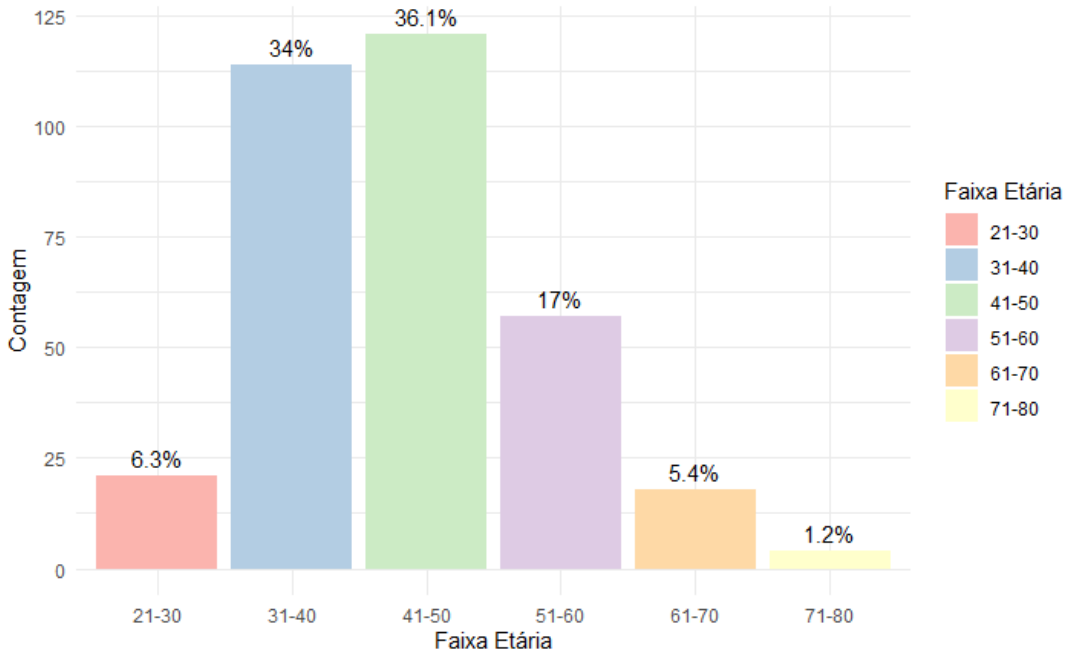

Figura 3.14: Distribuição por Faixa Etária

A análise da distribuição por faixa de horas de absenteísmo, vide a Figura [3.15], indica que a maior concentração está na faixa de *0-100 horas*, representando 36,1% do total. Em seguida, temos a faixa *100,1-200 horas* com 25,1% e a faixa *200,1-300 horas* com 15,2%, o que mostra que mais de 76% dos casos de absenteísmo estão concentrados nessas três faixas. Faixas de absenteísmo mais altas, como *500,1-1000 horas* e *1000,1-2500 horas*, têm uma representação menor, com 6,3% e 3,3%, respectivamente.

Há também um percentual significativo de 19% na categoria "NA", indicando que uma parcela dos registros não possui valores atribuídos para minutos de absenteísmo, ou seja, são colaboradores que não se ausentaram ou não registraram suas faltas. Esses dados sugerem uma tendência de absenteísmo mais concentrada em faixas menores de minutos, mas com uma probabilidade de que parte considerável dos dados pode estar faltante, o que pode influenciar a análise se não forem investigados ou tratados adequadamente.

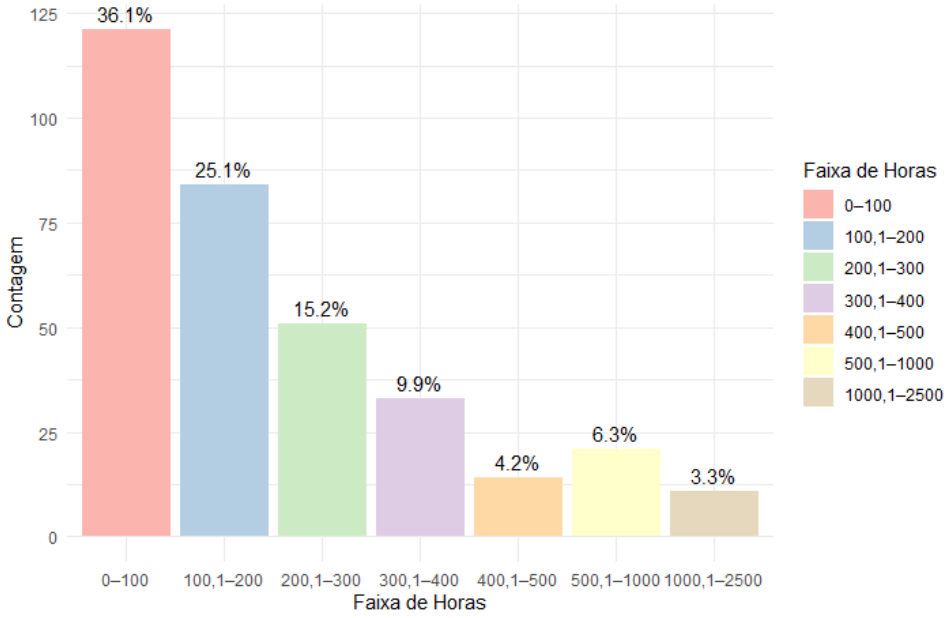

Figura 3.15: Distribuição por Faixa de Horas

Quanto a análise da distribuição de colaboradores com cônjuge, a Figura [3.16], indica que 68% dos indivíduos não possuem cônjuge, enquanto 32% têm cônjuge, esposo(a) ou companheiro(a). Essa diferença sugere que a maioria dos colaboradores está solteira, ou sem companheiro registrado. Essa informação pode ser útil para entender o perfil demográfico dos colaboradores, o que pode influenciar decisões organizacionais em áreas como benefícios, horários de trabalho e políticas de bem-estar, especialmente para atender às necessidades de diferentes grupos familiares.

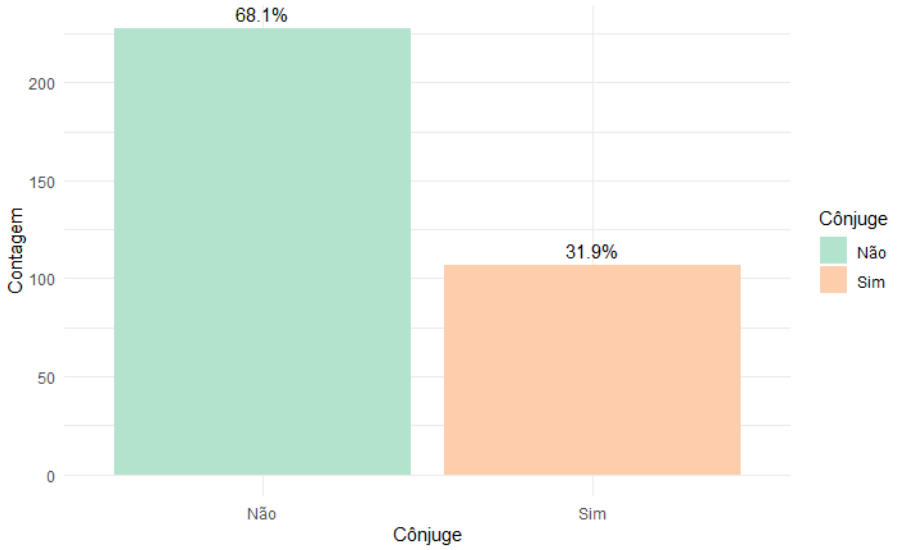

Figura 3.16: Distribuição por Cônjuge

A Tabela [3.2] apresenta resultados dos Testes de Associação entre as variáveis categóricas[75], que indicaram que somente as variáveis Gênero e Faixas de Horas apresentaram associação significativa (p-valor = 0,001), sugerindo uma associação (ou dependência) entre essas variáveis. As demais combinações de variáveis, como *Tipo de Cargo*, *Cônjuge* e *Faixa Etária* em diferentes pares, não mostraram associações significativas (valores-p > 0,05). Além disso, alguns testes resultaram em valores indisponíveis devido a dados insuficientes, especialmente para combinações envolvendo "Faixa Etária".

Tabela 3.2: Resultados dos Testes de Associação

Variável X-squared	**Variável 1**	**Variável 2**	**Valor χ^2**	**Graus de Liberdade**	**Valor-p**	**Interpretação**
X-squared2	Gênero	Faixas de Horas	18.3941	4	0.0010	Significativo
X-squared5	Tipo de Cargo	Faixas de Horas	54.8061	44	0.1274	Não significativo
X-squared9	Faixas de Horas	Cônjuge	5.7454	4	0.2190	Não significativo
X-squared6	Tipo de Cargo	Cônjuge	12.1120	13	0.5185	Não significativo
X-squared	Gênero	Tipo de Cargo	8.8005	13	0.7878	Não significativo
X-squared3	Gênero	Cônjuge	0.0250	1	0.8744	Não significativo
X-squared1	Gênero	Faixa Etária	NaN	7	NaN	<NA>
X-squared4	Tipo de Cargo	Faixa Etária	NaN	91	NaN	<NA>
X-squared7	Faixa Etária	Faixas de Horas	NaN	28	NaN	<NA>
X-squared8	Faixa Etária	Cônjuge	NaN	7	NaN	<NA>

Análise Bivariada

Concluída a análise das variáveis categóricas e numéricas, passamos ao estudo da análise bivariada[75], principalmente aquelas variáveis que possuem maior relação no contexto do Absenteísmo, sendo resultado da Análise de Correlação, nas variáveis numéricas, ou resultante do Teste de Associação, nas variáveis categóricas, ou da análise de variáveis numéricas e categóricas.

Dada a Figura 3.1, no campo das variáveis numéricas, apurou-se que as variáveis que apresentam correlação moderada são entre *Idade* e *Salário Mensal*, *Idade* e *Tempo de Casa* e entre *Salário Mensal* e *Tempo de Casa*, anteriormente analisadas. Em termos de variáveis categóricas, apenas a relação entre *Gênero* e *Faixa de Horas* apresentou associação estatisticamente significativa (valor-p de 0,001). Essa fase é fundamental para investigar a relação dos fatores e possíveis *insights* na gestão de pessoas. O objetivo é aprofundar a visão voltada para dados e resultados.

Análise de *Idade* e *Salário Mensal*

A análise do gráfico de dispersão entre *Idade* e *Salário Mensal* revela alguns aspectos interessantes sobre a relação entre essas variáveis, conforme Figura [3.17]. Primeiramente, nota-se que a maioria dos salários está concentrada abaixo de 40.000, independentemente da idade, com uma faixa ainda mais densa entre 10.000 e 30.000. Esse padrão sugere uma

faixa salarial comum para a maioria das pessoas representadas no gráfico, indicando que a maioria dos salários fica dentro de um intervalo relativamente estreito, independentemente da idade.

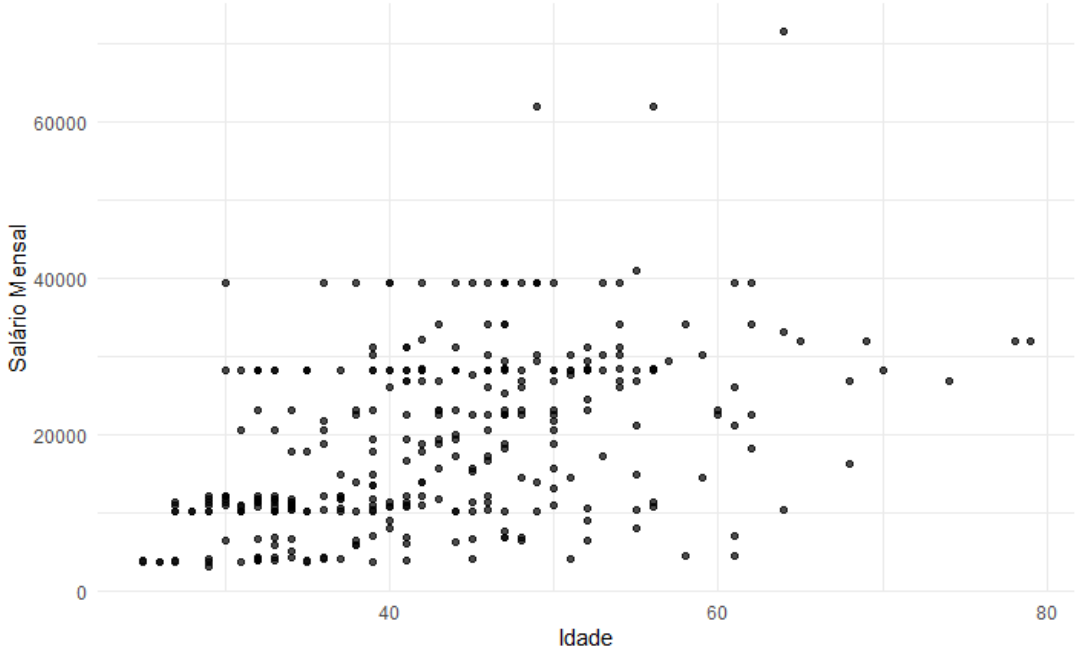

Figura 3.17: Análise de Idade e Salário Mensal

Em relação à variação salarial com a idade, há uma leve tendência de aumento nos salários conforme a idade avança, mas essa relação não é claramente linear. Em idades mais avançadas, os salários não parecem aumentar significativamente. Esse fenômeno sugere que, embora idade e salário possam ter alguma relação, outros fatores como experiência, nível educacional ou cargo podem ter um papel determinante na definição do salário.

Outra observação relevante é a presença de alguns *outliers* com salários acima de 60.000. Esses pontos indicam indivíduos com salários bem acima da média, em cargos de alto nível ou setores com remunerações mais elevadas; esses casos podem revelar características especiais associadas a esses indivíduos.

Além disso, a maioria dos dados está concentrada entre as idades de 25 a 55 anos, representando a força de trabalho mais ativa. Após os 55 anos, a dispersão dos salários diminui, o que pode indicar uma redução na quantidade de pessoas em funções de alta remuneração ou uma transição para a aposentadoria parcial.

Outro ponto interessante é que o salário máximo parece estabilizar-se nas faixas etárias mais altas, sugerindo que, após uma certa idade (em torno de 50 anos), há uma redução no crescimento salarial. Essa estabilização pode ser um reflexo da dinâmica do mercado de trabalho, onde o potencial de crescimento salarial atinge um limite com o avanço da idade.

Por fim, há uma ampla variação salarial entre 30 e 50 anos, indicando que os salários podem variar bastante nessa faixa etária. Isso pode refletir o desenvolvimento de carreira e diferentes trajetórias profissionais, que se tornam mais diversas ao longo do tempo.

Em resumo, o gráfico sugere uma relação moderada entre idade e salário, mas também indica que há fatores adicionais influenciando a variação salarial em cada faixa etária. Análises complementares, considerando variáveis como experiência, nível educacional e cargo, podem ajudar a esclarecer as razões para a dispersão observada e a identificar padrões mais específicos dentro dos dados.

Análise de *Idade* e *Salário Mensal*

A análise do gráfico de dispersão entre *Idade* e *Tempo de Casa* mostra a relação entre a idade dos indivíduos e o tempo que cada um passou na organização, conforme Figura [3.18]. Observa-se que o tempo de casa atinge uma concentração significativa entre 5 e 20 anos para a maioria das idades, com uma concentração mais alta entre 10 e 15 anos de tempo de casa. Essa concentração sugere que, independentemente da idade, muitos indivíduos tendem a acumular um tempo considerável de serviço, com a maioria se estabilizando nesse intervalo.

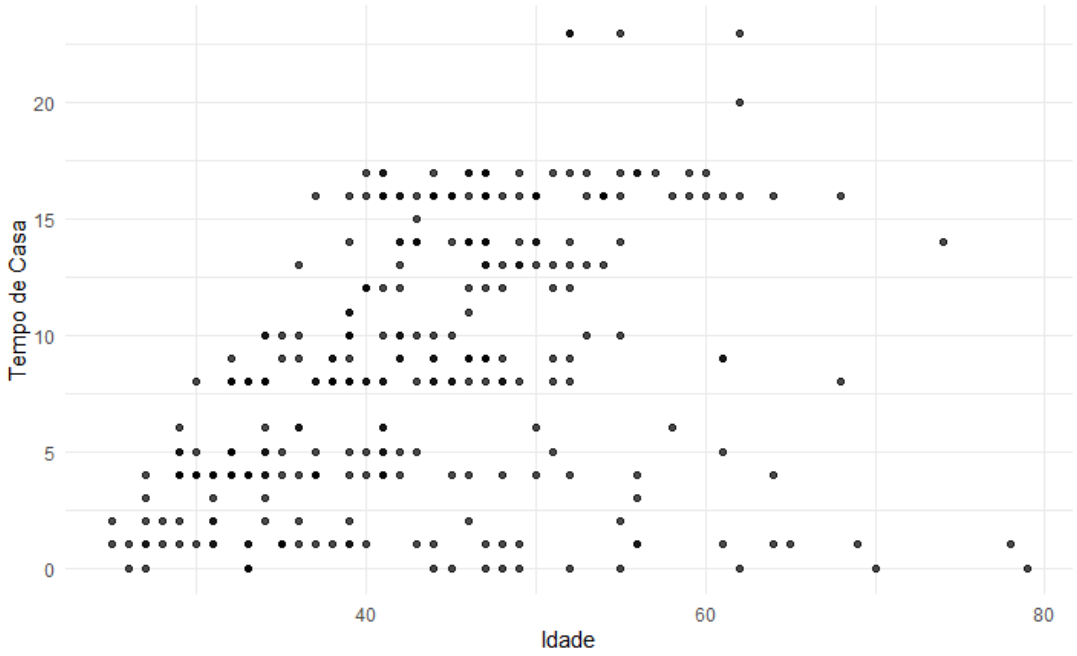

Figura 3.18: Análise de Idade e Tempo de Casa

Em termos de variação com a idade, não parece haver uma relação linear clara. Indivíduos mais jovens, abaixo dos 40 anos, apresentam uma maior diversidade em relação ao tempo de casa, variando de poucos anos até cerca de 15 anos. Essa dispersão sugere que,

para idades mais jovens, existem tanto novos contratados quanto pessoas com experiência significativa na organização.

Outro ponto interessante é que, acima dos 40 anos, o tempo de casa dos indivíduos tende a se estabilizar, com muitos funcionários mantendo entre 10 e 20 anos de casa. Isso pode indicar uma tendência de permanência entre os funcionários mais experientes, que se estabilizam na organização por longos períodos.

Acima dos 60 anos, a dispersão no tempo de casa reduz-se consideravelmente, com poucos indivíduos apresentando tempos longos. Isso pode estar relacionado a aposentadorias ou à transição para papéis de menor tempo de serviço nessa faixa etária.

Além disso, nota-se a presença de *outliers* em algumas idades, como indivíduos mais jovens com tempos de casa incomumente longos para a média de sua faixa etária. Esses pontos podem representar pessoas que ingressaram na organização muito cedo ou que tiveram uma rápida progressão em anos de serviço.

Em resumo, o gráfico sugere uma variação no tempo de casa com a idade, onde a estabilidade no emprego aumenta a partir dos 40 anos, enquanto faixas etárias mais jovens apresentam maior dispersão. A análise de outros fatores, como cargo ou progressão na carreira, pode ajudar a explicar melhor essas variações no tempo de casa em diferentes faixas etárias.

Análise de *Tempo de Casa* e *Salário Mensal*

A análise do gráfico de dispersão entre *Tempo de Casa* e *Salário Mensal* explora a relação entre o tempo que os indivíduos permanecem na organização e a remuneração que recebem, conforme Figura [3.19]. Primeiramente, observa-se que a maioria dos salários se concentra abaixo de 40.000, independentemente do tempo de casa, com uma faixa ainda mais comum entre 10.000 e 30.000. Isso indica que o salário médio de grande parte dos funcionários permanece em um intervalo similar, sem grandes aumentos associados ao tempo de casa.

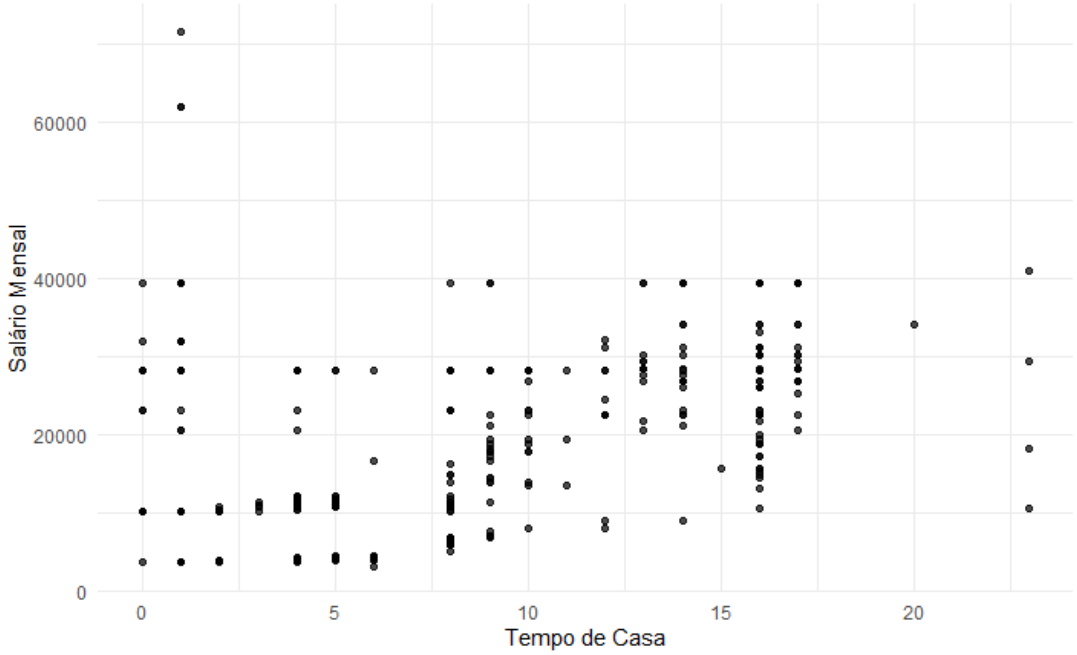

Figura 3.19: Análise de Tempo de Casa e Salário Mensal

Em relação ao tempo de casa, a dispersão dos salários se distribui de forma relativamente uniforme ao longo dos primeiros 20 anos. No entanto, não há uma tendência clara de aumento de salário conforme o tempo de casa aumenta. Esse fato pode indicar que, na organização, o tempo de casa por si só não é um forte determinante para aumentos salariais significativos, sugerindo que outros fatores, como promoções, qualificações ou cargos específicos, possam ser mais influentes na determinação do salário.

Observa-se também a presença de *outliers*, com alguns funcionários recebendo salários bem acima da média (acima de 60.000) mesmo com diferentes tempos de casa. Esses *outliers* provavelmente representam cargos ou posições de maior nível, onde o salário não está diretamente correlacionado com o tempo de casa, mas sim com outras qualificações ou responsabilidades.

Outro ponto interessante é que, embora existam funcionários com mais de 15 anos de casa, o salário máximo não parece crescer substancialmente com o aumento do tempo de serviço. Isso sugere que, após certo ponto, os aumentos salariais podem se estabilizar, possivelmente devido a limites salariais dentro dos cargos ou a uma estrutura de progressão salarial que não depende diretamente da longevidade.

Em resumo, o gráfico indica uma relação limitada entre tempo de casa e salário, sugerindo que o salário dos funcionários não aumenta diretamente em função do tempo que passam na empresa. A análise de outros fatores, como promoções, cargos específicos e qualificações adicionais, poderia ajudar a compreender melhor as variações salariais observadas.

Análise de Gênero e Faixa de Horas

Dada a relação entre "Gênero"e "Faixa de Horas"apresentou associação estatisticamente significativa (valor-p de 0,001) no Teste de Associação, temos a Figura [3.20]. O gráfico de barras empilhadas mostra que nas faixas de horas acumuladas mais baixas (0-100), há uma predominância masculina, com a proporção de homens superando a das mulheres. Conforme aumenta o tempo de horas acumuladas, a participação feminina também cresce, especialmente nas faixas superiores. Nas faixas mais altas de horas, a partir de 400 horas, observa-se uma inversão, onde as mulheres passam a ser maioria. Esse padrão sugere que as mulheres tendem a acumular mais horas de absenteísmo ao longo do tempo em comparação aos homens, o que pode refletir dinâmicas específicas dentro da organização, como diferenças no tempo de serviço, tipos de cargo ou áreas de atuação, ou ainda atribuições domésticas.

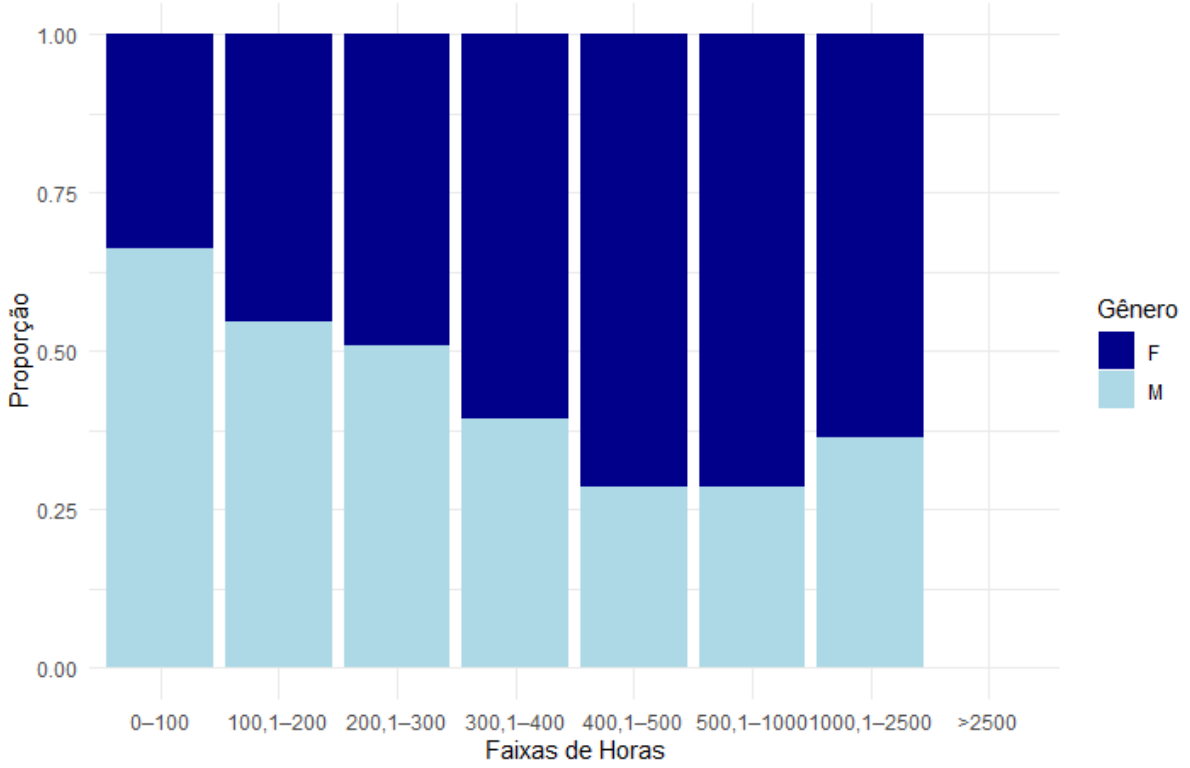

Figura 3.20: Análise de Gênero e Faixa de Minutos

3.3 Modelagem

A modelagem de dados, tendo como referência a Estrutura CRISP-DM, para agrupamento, classificação e predição de séries temporais oferece uma variedade de ferramentas

e técnicas que podem ser aplicadas para resolver diferentes problemas de negócios. A escolha do método apropriado permite que as empresas tomem decisões mais informadas, melhorem a eficiência operacional e aumentem a capacidade de resposta às mudanças de mercado. Neste estudo, testaremos três técnicas para cada eixo, buscando definir a mais adequada para o contexto de negócios da Apex-Brasil.

3.3.1 Agrupamento

Conforme demonstrado no Capítulo 2, serão utilizadas técnicas de agrupamento, buscando segmentar o público da Apex-Brasil, tendo em vista a necessidade de segmentação a partir de dados demográficos e absenteísmo. O esforço se justifica para escolher a técnica mais adequada ao contexto da Agência, comparando os modelos a partir das modelagens propostas. Todos os dados foram normalizados para a correta aplicação das técnicas.

Para utilizar as técnicas de agrupamento, é fundamental definir quantos clusters são necessários; para isso, foram utilizadas a Técnica do Cotovelo, conforme Figura [3.21] e o Índice de Silhueta, demonstrado na Figura [3.22] abaixo:

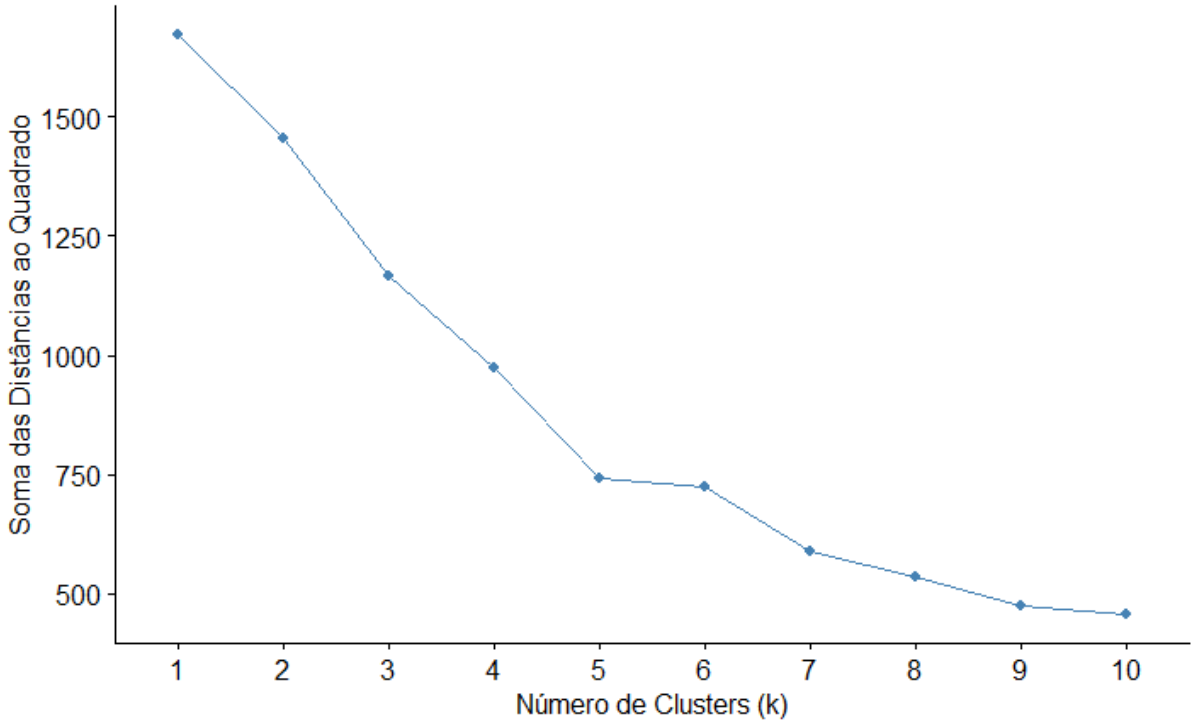

Figura 3.21: Método do Cotovelo para a Determinação do Número de Clusters

O gráfico mostra o Método do Cotovelo aplicado para determinar o número ideal de clusters, com o eixo X representando o número de clusters (k) e o eixo Y mostrando a

soma das distâncias ao quadrado dentro dos clusters. Observa-se uma queda acentuada, indicando um *cotovelo*, na soma das distâncias ao quadrado até o ponto onde o número de clusters é aproximadamente k=5 e nova queda a partir de k=7 e , por fim, com k=8 clusters. Faz-se necessária a comparação dos Índices de Silhueta dos três valores de clusters (k = 5, 7 ou 8) para definir a melhor técnica e mais apropriada. Abaixo, na Figura 3.22, ainda que o valor indicado seja de k=7, ainda vale a investigação e comparação de resultados.

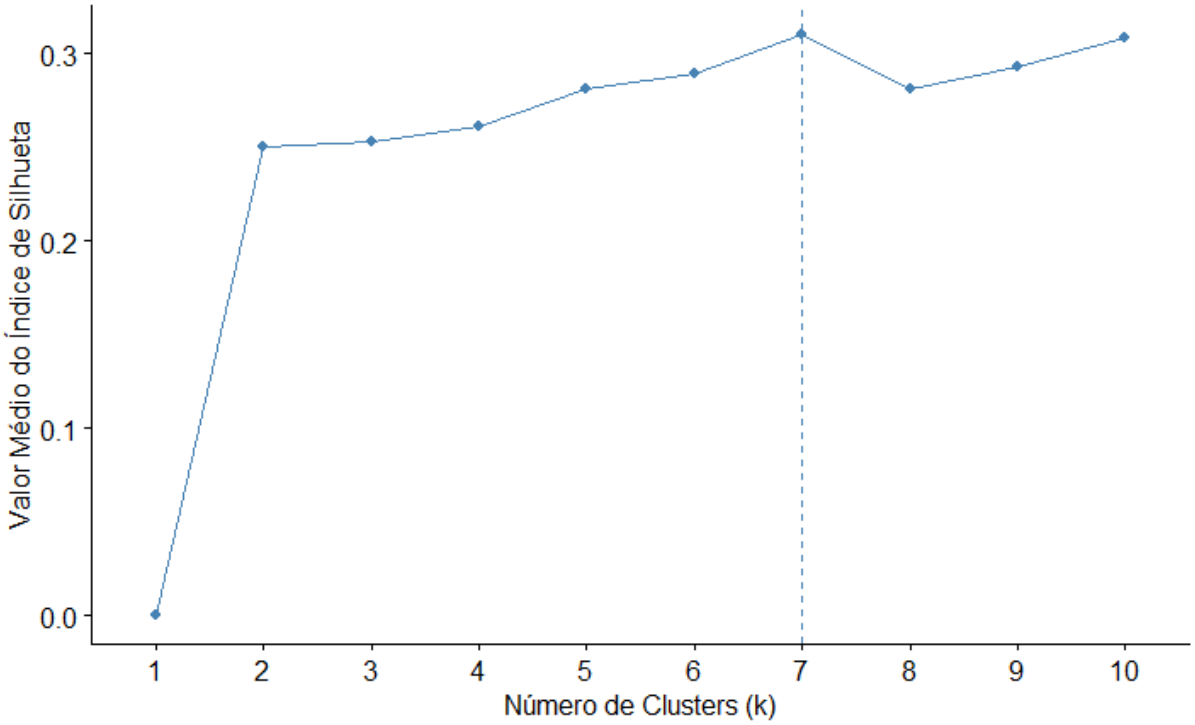

Figura 3.22: Índice de Silhueta para Determinação do Número de Clusters

Análise de comparação dos resultados do Índice de Silhueta para diferentes técnicas de agrupamento

Na Tabela [3.3], calculou-se os Índices de Silhueta para cada método de agrupamento, a seguir:

Tabela 3.3: Índice de Silhueta por Técnica de Agrupamento e Número de Clusters (K)

Técnica	**K = 5**	**K = 7**	**K = 8**
K-means	0,30	0,32	0,34
Hierarchical	0,28	0,30	0,31
DBSCAN (ε = 0,3; MinPts = 5)	0,62	0,63	0,37

O Índice de Silhueta mede a qualidade dos clusters formados, com valores que variam entre -1 e 1 [75]. Valores próximos de 1 indicam que os pontos estão bem agrupados e os clusters estão claramente separados, o que representa uma boa coesão interna e distinção entre os clusters. Valores próximos de 0 mostram que os pontos estão nos limites dos clusters, indicando uma separação pouco clara entre eles e sugerindo que os pontos podem estar igualmente próximos de mais de um cluster. Por outro lado, valores negativos indicam que os pontos podem estar mal alocados e potencialmente atribuídos ao cluster errado, o que reflete uma baixa qualidade do agrupamento.

Com base na Tabela 3.3 apresentada, que compara o índice de silhueta para as técnicas de K-means, Hierarchical Clustering e DBSCAN para diferentes valores de clusters $k = 5, 7, 8$, temos que a melhor técnica é a DBSCAN, pois apresenta os melhores resultados absolutos, com índices 0,62 (K=5) e 0,63 (K=7), muito superiores aos obtidos por K-means e Hierarchical Clustering. Mesmo para $k = 8$, a técnica mantém uma performance comparável (0,37), ainda superior às demais.

A técnica DBSCAN [36] é especialmente eficaz para dados com formas arbitrárias de cluster e ruído (outliers), o que pode explicar seu desempenho superior.Os bons índices sugerem que a estrutura dos dados se ajusta melhor a densidade do que à partição esférica assumida por K-means.

Embora DBSCAN não use diretamente o parâmetro K, os testes foram alinhados com diferentes valores de ε para manter a comparabilidade. Caso o dataset contenha *outliers* ou clusters de tamanhos/desvios variados, a vantagem do DBSCAN tende a aumentar.

A técnica de agrupamento DBSCAN apresenta o melhor desempenho com base no índice de silhueta para todos os valores testados, especialmente com $\varepsilon = 0{,}3$ e 'MinPts = 5', sendo a recomendada para segmentação dos dados analisados.

Conforme apontado no Capítulo 2, ao estudar a aplicação da técnica de DBSCAN, deve-se analisar o ruído do conjunto de dados e sua correta aplicação e separação nos agrupamentos. Passa-se a esta análise detalhada de ruído no conjunto de dados da Apex-Brasil.

Ao realizar o cruzamento de ε com os respectivos Índices de Silhueta e ao fazer a medição da intensidade de ruído, temos os seguintes resultados na Tabela 3.4 abaixo:

Pode-se observar que muitas observações ficam fora dos agrupamentos, conforme Gráfico 3.23 abaixo, a análise dos resultados obtidos por meio da aplicação do algoritmo DBSCAN revelou um comportamento característico de troca entre a qualidade dos agrupamentos, medida pelo índice de silhueta, e a quantidade de dados classificados como ruído (cluster 0).

Observou-se que os maiores valores de silhueta foram obtidos com parâmetros mais restritivos. Por exemplo, com eps = 0,20 e MinPts = 4, o índice de silhueta atingiu

Tabela 3.4: Resultados da Análise de Índice de Silhueta com DBSCAN

ε	**MinPts**	**Silhueta**	**Ruído**	**% de Ruído**
0,20	4	0,92	325	97,0
0,25	5	0,86	318	94,9
0,25	4	0,82	307	91,6
0,40	6	0,72	266	79,4
0,35	5	0,68	275	82,1
0,30	4	0,66	274	81,8
0,40	5	0,63	256	76,4
0,30	5	0,62	290	86,6
0,30	6	0,58	297	88,7
0,45	6	0,58	241	71,9

0,923, indicando excelente separação e coesão entre os clusters. No entanto, este modelo classificou 97% das observações como ruído, agrupando apenas uma fração mínima dos dados.

À medida que o valor de eps aumentou, houve uma redução gradual do índice de silhueta, indicando uma menor qualidade dos agrupamentos. Por outro lado, observou-se também uma redução no percentual de ruído, ou seja, mais dados foram efetivamente alocados em clusters. Essa relação é típica do DBSCAN, que privilegia a densidade dos agrupamentos e tende a descartar como ruído os pontos que não satisfazem os critérios de densidade.

O melhor compromisso entre qualidade dos clusters e abrangência do modelo foi identificado em configurações como $\varepsilon = 0{,}35$ com MinPts = 5, e $\varepsilon = 0{,}40$ com MinPts = 6. Nessas combinações, o índice de silhueta variou entre 0,676 e 0,716, e o percentual de ruído foi reduzido para a faixa de 76% a 82% — ainda elevado, mas consideravelmente inferior aos modelos com silhueta máxima.

Finalmente, o modelo com o menor percentual de ruído foi aquele com $\varepsilon = 0{,}45$ e MinPts = 6, que apresentou um índice de silhueta de 0,58 e classificou apenas 71,9% dos dados como ruído, o que representa a melhor cobertura entre os cenários avaliados.

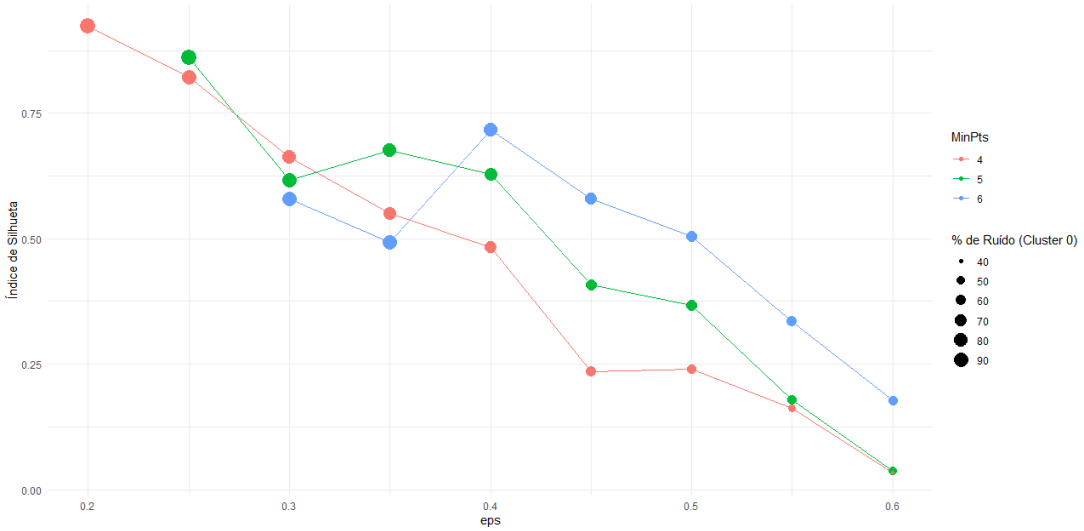

Figura 3.23: Análise do Índice de Silhueta vs. Eps (DBSCAN)

Como a escolha da técnica DBSCAN depende diretamente do objetivo da análise, que é o máximo de agrupamento dos dados e, tendo em vista a segmentação dos funcionários da Apex-Brasil, não apresentou boa performance dado o ruído apresentado e, apesar dos melhores Índices de Silhueta, não obteve sucesso dada a quantidade de empregados que não participaram de nenhum cluster. A visualização apresentada na Figura 3.24 demonstra o ruído pela quantidade de observações no Cluster *0*, que foi o mais abrangente:

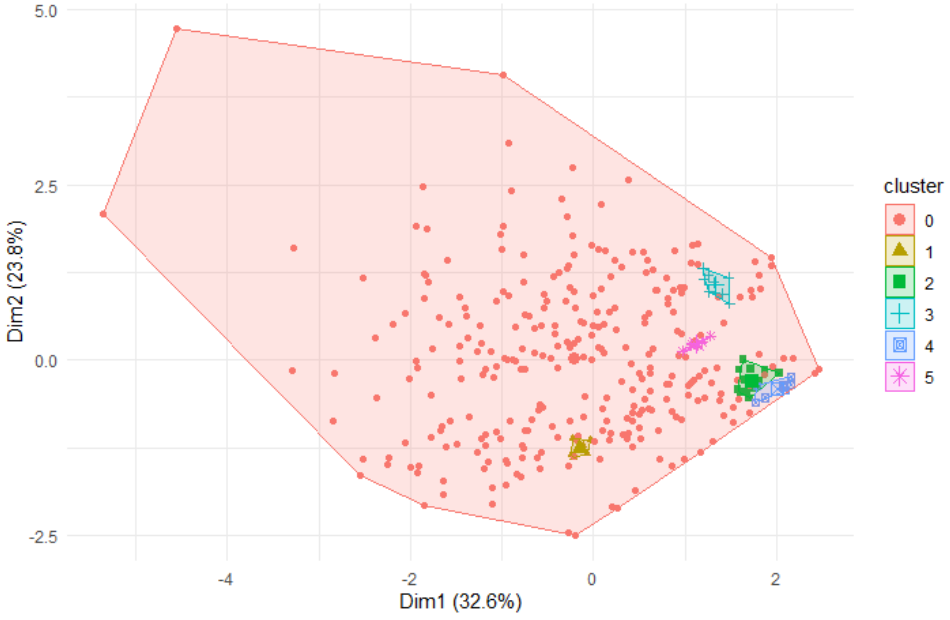

Figura 3.24: Agrupamento por método DBSCAN

Tendo em vista os resultados da Tabela 3.3 , temos que o segundo melhor resultado é o da Técnica K-means com a indicação de Clusters. A qual será detalhada na próxima seção.

Técnica K-means

O gráfico na Figura [3.25] abaixo mostra a visualização de clusters obtidos por meio do método K-means em um conjunto de dados normalizado, com k=8. Cada grupo ou cluster está representado por uma cor distinta e uma forma geométrica específica, facilitando a distinção visual entre os diferentes clusters. Os eixos Dim1 e Dim2, que explicam juntos 0,56 da variância dos dados, são componentes principais que ajudam a reduzir a dimensionalidade e a projetar os dados em um espaço bidimensional. A sobreposição entre alguns clusters sugere que certos grupos possuem características semelhantes, enquanto outros se destacam de forma mais isolada, indicando padrões mais diferenciados nos dados. Esse tipo de visualização ajuda a avaliar a qualidade dos clusters e a identificar possíveis áreas de interseção entre grupos.

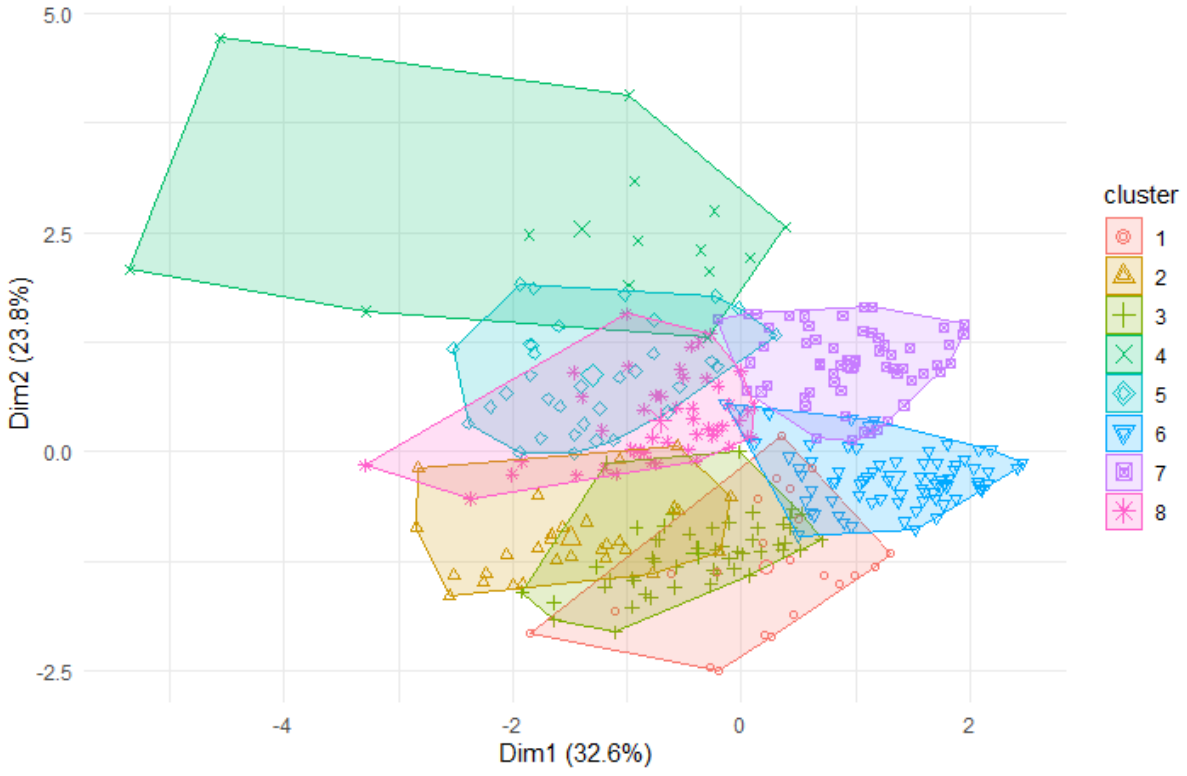

Figura 3.25: Método K-means para Visualização de Clusters

O gráfico de silhueta abaixo, demonstrado na Figura [3.26], para o modelo K-means mostra a qualidade da separação dos clusters. Cada barra representa um ponto de dados

dentro de um cluster, com a largura da silhueta (Si) indicando o quão bem esse ponto está associado ao seu cluster em comparação com os outros. Os valores próximos de 1 indicam que os pontos estão bem agrupados em seus respectivos clusters, enquanto valores próximos de 0 ou negativos indicam pontos mal agrupados ou que poderiam pertencer a outro cluster.

Observa-se que a maioria dos clusters tem uma largura de silhueta média positiva, o que sugere uma separação razoável entre eles. No entanto, há variabilidade na largura das barras dentro de cada cluster, indicando que alguns pontos estão menos bem definidos dentro de seus grupos. A linha pontilhada vermelha representa a largura média geral da silhueta, que é um indicador da qualidade geral do agrupamento. Clusters com pontos negativos ou com barras muito estreitas mostram que esses grupos têm sobreposição ou uma fraca distinção com outros clusters.

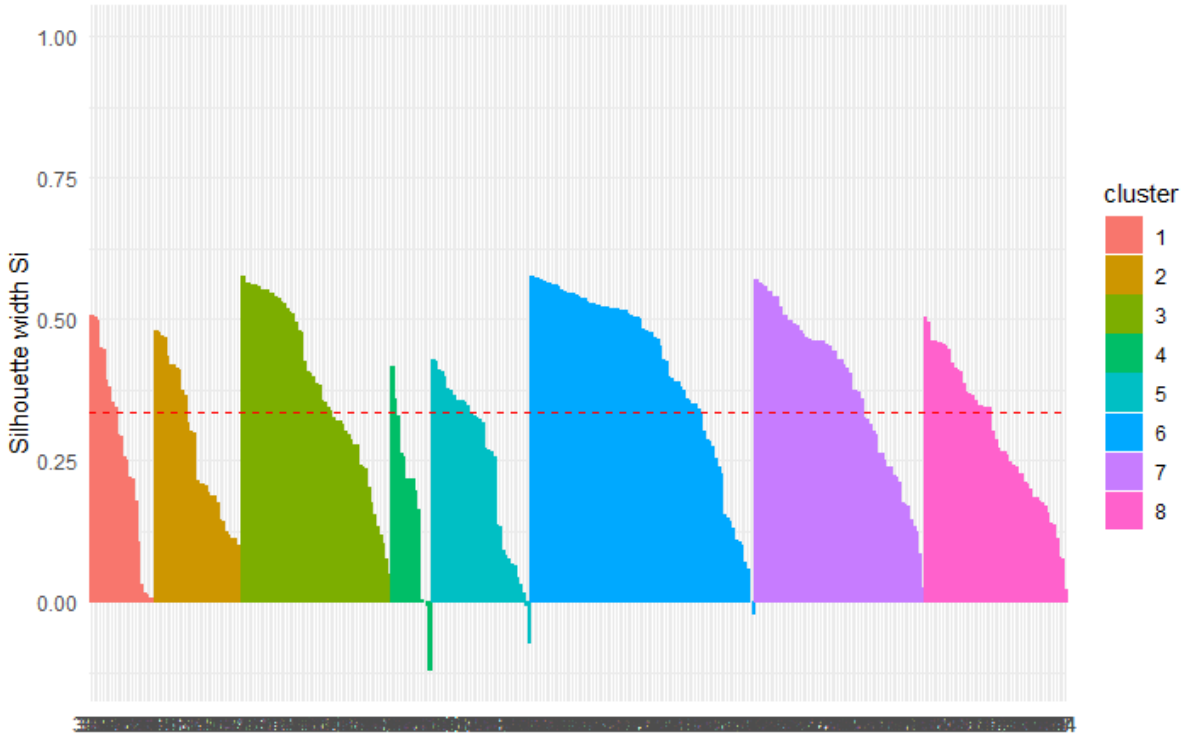

Figura 3.26: Gráfico de Silhueta - Método K-means

Segmentação dos Empregados da Apex-Brasil, considerando o Absenteísmo

Tendo em vista o resultado acima, os empregados da Agência foram segmentados, conforme a Técnica K-means, obtendo uma proposta de segmentação dos clientes internos (funcionários) da Gerência de Recursos Humanos da Apex-Brasil. A análise dos oito clus-

ters gerados a partir das variáveis demográficas revelou perfis distintos entre os grupos de profissionais.

Perfil do Empregado no Cluster 1

É composto por indivíduos com idade média de 61,5 anos, os mais velhos da amostra, com tempo médio de casa de apenas 2,7 anos e praticamente nenhum dependente. A predominância do gênero masculino sugere um perfil de homens mais velhos e recém-contratados em liderança técnica sênior.

Perfil do Empregado no Cluster 2

Os empregados deste cluster reúnem profissionais com idade média de 50,4 anos, tempo de casa de 13,2 anos e o maior número médio de dependentes (2,9). Com representação exclusivamente masculina, esse grupo parece corresponder a homens com longa trajetória institucional e alta carga familiar, provavelmente ocupando cargos de confiança ou chefia estável.

Perfil do Empregado no Cluster 3

Este grupo inclui empregados que possuem características similares ao anterior quanto à senioridade, com idade média de 46,9 anos e tempo de casa de 13,9 anos, o maior da amostra. Entretanto, apresenta um número muito reduzido de dependentes (0,2), indicando profissionais experientes com baixa carga familiar. O grupo também é exclusivamente masculino.

Perfil do Empregado no Cluster 4

Este grupo se destaca por uma composição de gênero mais equilibrada, com idade média de 40,3 anos, tempo de casa de 10 anos e um número intermediário de dependentes (1,0). Este perfil pode representar profissionais em cargos técnicos ou de coordenação, com estabilidade moderada e equilíbrio entre vida pessoal e profissional.

Perfil do Empregado no Cluster 5

Os empregados deste grupo são formados exclusivamente por mulheres, com idade média de 43,8 anos, tempo de casa de 11,1 anos e o maior número de dependentes da amostra (3,0). Trata-se de um grupo de mulheres maduras, estáveis institucionalmente e com alta carga familiar, potencialmente vinculadas a funções técnicas e administrativas de longa permanência.

Perfil do Empregado no Cluster 6

Este cluster é composto por profissionais jovens (idade média de 34,0 anos), com tempo de casa de 3,7 anos e poucos dependentes (0,4). Exclusivamente masculino, esse cluster aparenta ser formado por homens recém-ingressos na organização, possivelmente ocupando cargos operacionais ou em início de carreira.

Perfil do Empregado no Cluster 7

Os empregados deste grupo apresentam um perfil semelhante ao Cluster 6, porém exclu-

sivamente feminino. A idade média é de 35,5 anos, com tempo de casa de 4,5 anos e um número médio de dependentes de 0,3. Esse grupo pode ser caracterizado como mulheres jovens, com baixa carga familiar e trajetória institucional recente.

Perfil do Empregado no Cluster 8

Por fim, o Cluster 8 reúne mulheres com idade média de 48,2 anos, tempo de casa de 12,8 anos e poucos dependentes (0,3). Trata-se de um grupo de profissionais seniores do sexo feminino, com longa permanência na instituição e, provavelmente, estabilidade funcional, mas com baixa dependência familiar.

3.3.2 Classificação

A partir dos dados de treinamento e teste dos clusters definidos no Modelo de agrupamento, foram gerados os dados necessários para a realização do Modelo de Classificação. Para os propósitos desta pesquisa e a necessidade de processamento do modelo KNN, esse processamento foi paralelizado usando a Biblioteca R (doParallel)[76].

Para a realização do treinamento e teste dos modelos de classificação, foi utilizado 70 por cento da massa de dados para treinamento dos modelos e 30 por cento para teste. Para todos os modelos, foram calculadas Acurácia, Precisão, Recall e Pontuação F-1. Os gráficos das matrizes de confusão, que fornecem *insights*sobre o desempenho desses modelos, para os quatro modelos de classificação são detalhados, a saber: Regressão Logística Multinomial, Árvore de Decisão, Random Forest e KNN.

Análise de comparação dos resultados das Técnicas de Classificação

A partir da Tabela [3.5] abaixo, apresenta-se a análise Com base nas métricas de avaliação, resumida para cada técnica. Demonstra que a Regressão Logística Multinomial apresenta o melhor desempenho geral, com uma alta acurácia de 1,00, precisão de 1,00, recall de 1,00 e pontuação F1 de 1,00, o que indica um modelo bem equilibrado e eficaz na classificação.

O Modelo *Random Forest* também apresenta um desempenho elevado, com acurácia de 0,96, precisão de 0,96, recall de 0,96 e pontuação F1 de 0,93, ficando próximo ao desempenho da Regressão Logística Multinomial, o que o torna uma opção robusta para esse conjunto de dados.

Modelo	**Acurácia**	**Precisão**	**Recall**	**Pontuação F1**
Regressão Logística Multinomial	1,00	1,00	1,00	1,00
Random Forest	0,96	0,96	0,93	0,94
KNN	0,49	0,63	0,52	0,51
Árvore de Decisão	0,49	0,58	0,34	0,65

Tabela 3.5: Métricas de avaliação para diferentes modelos

Em contraste, a técnica de Árvore de Decisão tem uma acurácia muito mais baixa, de 0,49, com precisão de 0,58 e recall de 0,34, mas uma pontuação F1 moderada de 0,65. Essa pontuação F1 elevada em comparação com o recall sugere que a técnica é mais precisa nas previsões corretas, mas não consegue capturar bem todas as classes, resultando em um desempenho menos eficaz na classificação global.

Por fim, a técnica KKNN apresenta o pior desempenho, com uma acurácia de apenas 0,49, precisão de 0,63, recall de 0,52 e pontuação F1 de 0,51. Esses resultados indicam que a técnica KNN tem dificuldades significativas em capturar os padrões do conjunto de dados, tornando-a a menos adequada entre as técnicas avaliadas para essa tarefa de classificação.

Com base na comparação de resultados, a técnica de Regressão Logística Multinomial é ligeiramente superior à técnica *Random Forest*. Apesar de ambas serem confiáveis, fornecendo alta precisão e previsões balanceadas em todas as classes, para o conjunto de dados será utilizada a técnica de Regressão Logística Multinomial, conforme abaixo.

Técnica de Regressão Logística Multinomial

A matriz de confusão para o modelo de Regressão Logística Multinomial, conforme Figura [3.27], apresenta o desempenho do modelo em classificar corretamente as observações em cada uma das categorias. As linhas representam as classes de referência (ou classes reais), enquanto as colunas representam as predições feitas pelo modelo. As células na diagonal indicam as previsões corretas, onde a classe predita coincide com a classe real, e a frequência de cada previsão correta é mostrada.

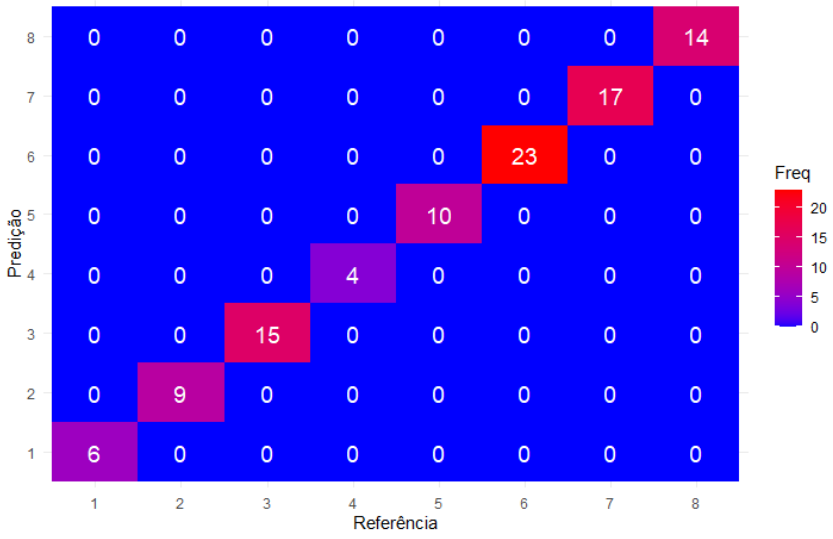

Figura 3.27: Matriz de Confusão de Regressão Logística Multinomial

A coloração da matriz ajuda a identificar visualmente as áreas com maior frequência de acertos (células vermelhas) e erros (células azuis). Observando a matriz de confusão, nota-se que o modelo tem bom desempenho em várias classes, como o cluster 6, com 23 predições corretas, e a classe 7, com 17 predições corretas.

A matriz apresentada indica um ótimo desempenho do modelo, com uma forte correspondência entre as classes previstas (eixo Y: predição) e as classes reais (eixo X: referência). Cada um dos oito clusters (de 1 a 8) foi corretamente identificado com alta precisão.

Todos os valores relevantes estão distribuídos ao longo da diagonal principal, o que demonstra que o modelo está classificando corretamente quase todos os exemplos. Não há registros significativos fora da diagonal, ou seja, não há confusões entre classes diferentes, o que indica baixa taxa de erro.

Especificamente, os clusters 1 a 8 apresentam quantidades como 6, 9, 15, 4, 10, 23, 17 e 14 exemplos corretamente classificados, respectivamente. Não há qualquer previsão equivocada (zeros fora da diagonal), o que sugere ausência de *overfitting* visível neste caso — a menos que o mesmo comportamento não se mantenha nos dados de teste. Por ser a matriz do teste, o modelo apresenta excelente generalização.

3.3.3 Predição de Séries Temporais

A predição de séries temporais envolve analisar dados de uma sequência temporal para fazer previsões futuras. Elas ajudam a organização a se preparar para desafios futuros, com possibilidade de antecipação e maior sucesso no contexto empresarial. Nas práticas de Recursos Humanos, servem para definir sazonalidade e prever orçamentos ou comportamentos das pessoas como o absenteísmo, que tem suas atividades projetadas na linha do tempo, ambos de fundamental importância para os resultados e produtividade corporativos.

Análise de comparação dos resultados das Técnicas de Predição de Série Temporal

A Tabela [3.6], que compara o resultado dos diferentes modelos de previsão usando três métricas: RMSE (erro quadrático médio), MAE (erro médio absoluto) e MAPE (erro percentual médio absoluto).

A comparação das técnicas de previsão revela que a Suavização Exponencial Simples (SES) oferece um bom ajuste aos dados, apresentando baixos valores de RMSE e MAE, além de um MAPE relativamente baixo, o que indica precisão na modelagem. A técnica Aditiva Holt-Winters, por outro lado, apresenta erros significativamente maiores que o

SES e o Auto.ARIMA, sugerindo que ela tem dificuldade em capturar adequadamente a sazonalidade e a tendência dos dados.

Tabela 3.6: Comparação de Modelos com Métricas de Erro

Modelo	**RMSE**	**MAE**	**MAPE**
Auto.ARIMA	**885.80**	**709.77**	**0.71**
Single Exp. Smoothing - SES	**885.80**	**709.77**	**0.71**
Holt-Winters Aditivo	1965.08	1811.31	2.23
Holt-Winters Multiplicativo	2021.63	1864.89	2.25
Custom SARIMA	2114.44	1593.80	1.59

A técnica Multiplicativa Holt-Winters mostra ainda mais erros, indicando que não é adequada para este conjunto de dados. O Auto.ARIMA, com desempenho quase idêntico ao SES, destaca-se como uma das melhores opções para previsão, enquanto o modelo SARIMA, embora superior aos modelos Holt-Winters, ainda apresenta erros maiores que o SES e o Auto.ARIMA, sugerindo que não é a melhor escolha para este caso.

Recomenda-se considerar ajustes adicionais ou a integração de fatores externos para aumentar a precisão dos modelos. Além disso, uma análise dos períodos específicos com altos resíduos pode oferecer *insights* valiosos para aprimorar a modelagem e a previsão. A análise detalhada aponta a superioridade das técnicas SES e Auto.ARIMA, com o Auto.ARIMA emergindo como a melhor escolha para previsões de ausência neste conjunto de dados.

Análise de Série Temporal no Contexto de Absenteísmo na Apex-Brasil

De acordo com o Gráfico na Figura [3.28], a análise dos dados de Horas de Absenteísmo revela que não há uma tendência clara de aumento ou diminuição ao longo do período analisado, indicando uma certa estabilidade com picos esporádicos. Os valores mostram uma certa estabilidade, com flutuações constantes ao longo do tempo, sugerindo que o absenteísmo não está seguindo um padrão linear de crescimento ou declínio.

Observam-se vários picos significativos em diferentes momentos, com destaque para um grande pico em 2023, possivelmente ligado a eventos sazonais, com fatores que causaram ausências em massa. O padrão dos dados sugere a presença de sazonalidade, com repetições em determinados períodos, o que justifica uma análise mais detalhada para identificar possíveis causas dessas ausências. Além disso, a alta variabilidade nos dados, com grandes flutuações mensais, indica que as ausências não são uniformes ao longo do tempo e podem ser influenciadas por diferentes fatores externos.

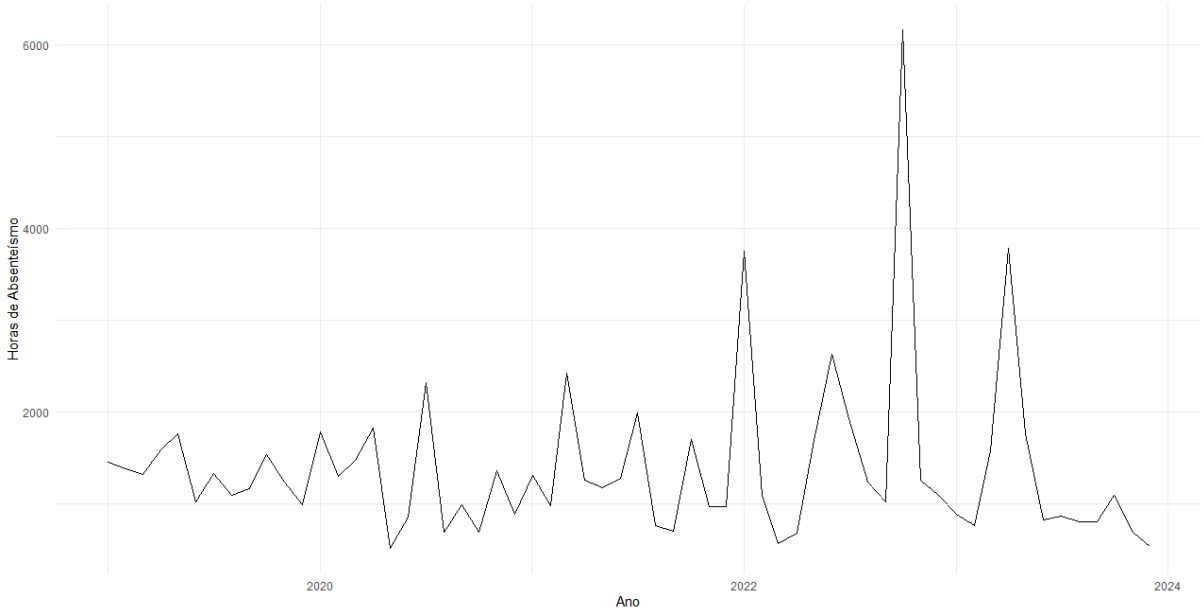

Figura 3.28: Total de Horas de Absenteísmo por Ano

A análise dos resultados indica que os modelos Single Exponential Smoothing (SES) e Auto.ARIMA apresentaram o melhor desempenho entre os métodos avaliados, com valores idênticos e mínimos nas três métricas: RMSE de 885,80, MAE de 709,77 e MAPE de 0,71%. Esses resultados evidenciam uma excelente capacidade preditiva, com baixa variabilidade nos erros absolutos e relativos, posicionando ambos os modelos como as melhores opções para previsão do absenteísmo em horas.

Em contraste, os modelos Holt-Winters, tanto na versão aditiva quanto multiplicativa, apresentaram desempenhos inferiores, com RMSE acima de 1,960 e MAPE em torno de 2,23% a 2,25%, indicando maior dispersão nos erros e menor precisão nas previsões. O modelo Custom SARIMA, embora tenha se saído melhor do que os modelos de Holt-Winters em termos de MAPE (1,59%), ainda demonstrou erro absoluto e quadrático significativamente mais elevados que os modelos SES e Auto.ARIMA.

Diante disso, recomenda-se a adoção de Auto.ARIMA ou SES como modelos preferenciais para aplicações práticas, mas com vantagem adicional para o Auto.ARIMA por sua flexibilidade em capturar componentes sazonais e tendências complexas de forma automatizada. Portanto, serão analisados os resultados do Modelo Auto.ARIMA.

Análise do Gráfico de Sazonalidade

O gráfico de sazonalidade, conforme Figura [3.29] mostra o total de horas de ausência por mês, comparando diferentes anos de 2019 a 2023. Este tipo de gráfico é útil para identificar padrões sazonais que se repetem ao longo dos anos.

A análise dos dados de ausência revela a presença de padrões sazonais, com aumento claro de ausências em determinados meses do ano, especialmente em outubro, que apresentou um pico significativo em 2023. Além disso, os meses de abril, maio e outubro mostram variações consideráveis de um ano para o outro, o que sugere possíveis influências sazonais ou eventos específicos que ocorrem nesses períodos. Em relação às variações anuais, observa-se que a magnitude das ausências varia substancialmente entre os anos. Por exemplo, o ano de 2023 se destaca com um pico muito alto em outubro, enquanto outros anos apresentam picos menores e mais distribuídos ao longo do ano. Anos como 2020 e 2022 parecem apresentar menos variabilidade no total de horas de ausência ao longo do ano em comparação com 2023.

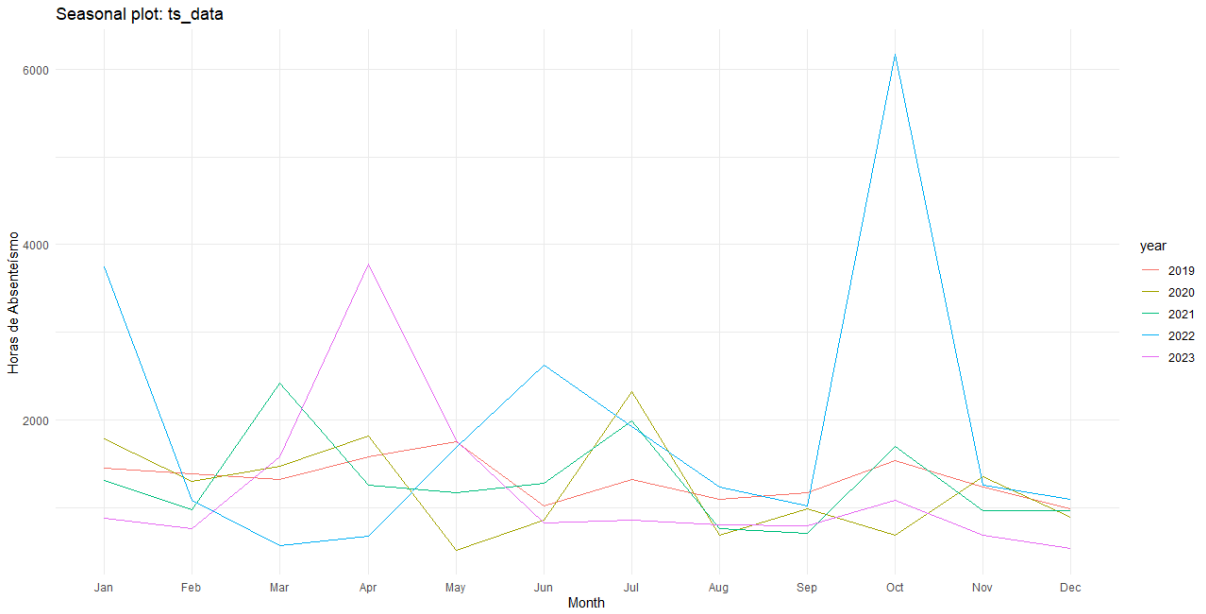

Figura 3.29: Gráfico sazonal: Total de Horas de Absenteísmo por Mês

Além disso, alguns meses, como fevereiro e março, tendem a ter menos minutos de ausência na maioria dos anos, embora haja algumas variações em anos específicos. Na interpretação dos resultados, destaca-se o pico extremo em outubro de 2023, como já visto acima, que pode indicar um evento atípico, como mudanças organizacionais que levaram a um aumento massivo nas ausências. A presença de picos em meses específicos sugere a influência de fatores sazonais, que podem incluir períodos de férias, doenças sazonais ou outros eventos recorrentes. A comparação entre diferentes anos revela que, apesar de haver uma tendência sazonal, a intensidade das ausências varia consideravelmente de ano para ano, indicando que, além da sazonalidade, há outros fatores contextuais que influenciam esses dados.

Análise do Gráfico de Defasagem - LAG

O gráfico de defasagem (ou LAG em inglês) na Figura [3.30], demonstra a relação entre as séries temporais em diferentes defasagens e como o total de horas de ausência varia ao longo do tempo. Este tipo de gráfico é útil para identificar autocorrelações e padrões de repetição ao longo do tempo. A relação entre o valor da série temporal e suas defasagens, variando de 1 até 16 períodos, com as cores indicando os diferentes meses, representa a sazonalidade ao longo do ano.

Observa-se que as defasagens menores, como Lag 1 e Lag 2, apresentam uma relação mais próxima e consistente com o valor da série temporal. Isso sugere que os valores recentes da série estão fortemente correlacionados com os períodos subsequentes. No entanto, à medida que a defasagem aumenta, a correlação entre os valores passados e os atuais parece diminuir, indicando que a dependência com os valores passados mais distantes é menor.

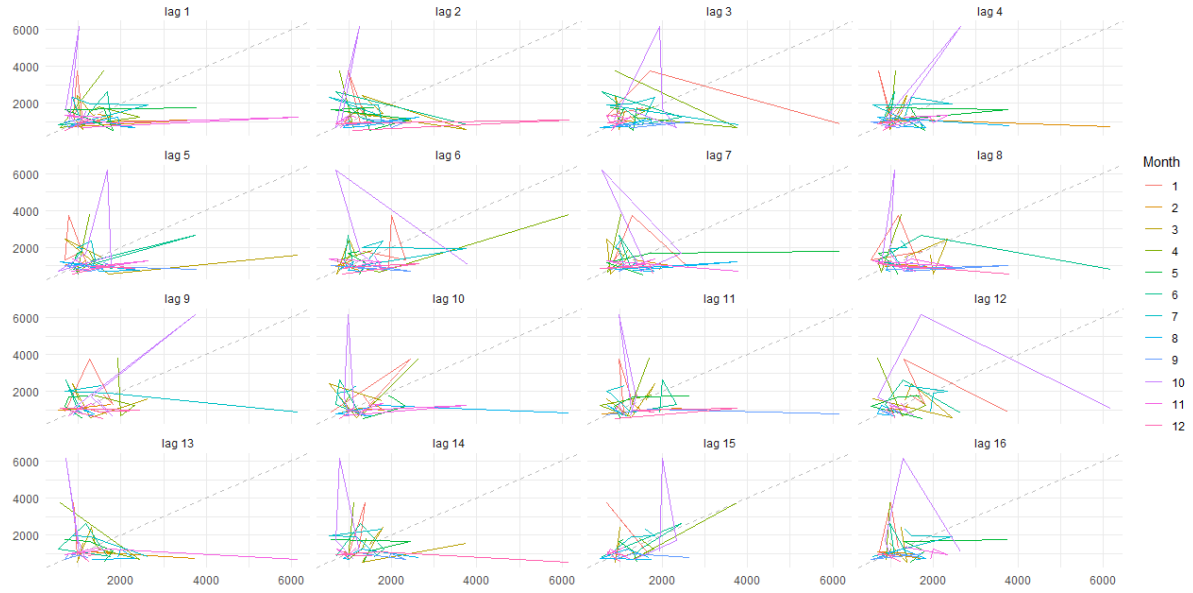

Figura 3.30: Gráfico de Defasagem - LAG

Além disso, a presença de sazonalidade mensal é visível pelas cores que representam diferentes meses. Em algumas defasagens, há uma separação clara das cores, sugerindo que certos meses influenciam de forma mais evidente os valores subsequentes. Alguns meses apresentam padrões mais altos, enquanto outros apresentam padrões mais baixos, o que pode ser um reflexo de fatores sazonais ou de eventos recorrentes específicos para cada período do ano.

À medida que o atraso aumenta para o médio prazo (5-8), a dispersão dos pontos também aumenta, indicando que a correlação começa a diminuir. Esse comportamento é esperado, pois eventos que ocorreram mais distantes no tempo tendem a ter menos influência direta uns sobre os outros, enfraquecendo a correlação com o aumento do atraso.

Nos atrasos de longo prazo (9-16), a dispersão dos pontos continua a aumentar e os padrões de correlação tornam-se menos claros. Isso sugere que os valores de ausência em períodos muito distantes no tempo apresentam pouca ou nenhuma correlação direta, indicando que a influência de um valor de ausência é temporária e diminui ao longo do tempo.

A presença de padrões repetitivos em certos meses ao longo dos atrasos pode indicar a existência de sazonalidade na série temporal. Por exemplo, se a ausência em um determinado mês tende a se repetir após 12 meses (defasagem 12), isso reforça a hipótese de um componente sazonal anual, que se manifesta em intervalos regulares ao longo do tempo.

A interpretação desses resultados aponta para uma forte autocorrelação em atrasos curtos, o que sugere que os valores de ausência são significativamente influenciados por valores dos meses imediatamente anteriores. Isso pode ser causado por fatores contínuos, como a progressão de uma epidemia ou políticas organizacionais internas. Além disso, a diminuição da correlação com o aumento do atraso é uma característica comum em muitas séries temporais e indica que eventos mais distantes no tempo têm uma influência cada vez menor entre si. Por fim, os padrões repetitivos observados nos atrasos múltiplos de 12 reforçam a hipótese de sazonalidade, sugerindo uma repetição anual desses eventos de ausência.

Análise de Gráfico de Função de Autocorrelação (ACF)

O Gráfico de Autocorrelação (ACF), demonstrado na Figura [3.31] é uma ferramenta estatística usada para medir a correlação entre valores de uma série temporal em diferentes defasagens. Ele ajuda a identificar a presença de padrões sazonais e a estrutura de dependência temporal nos dados.

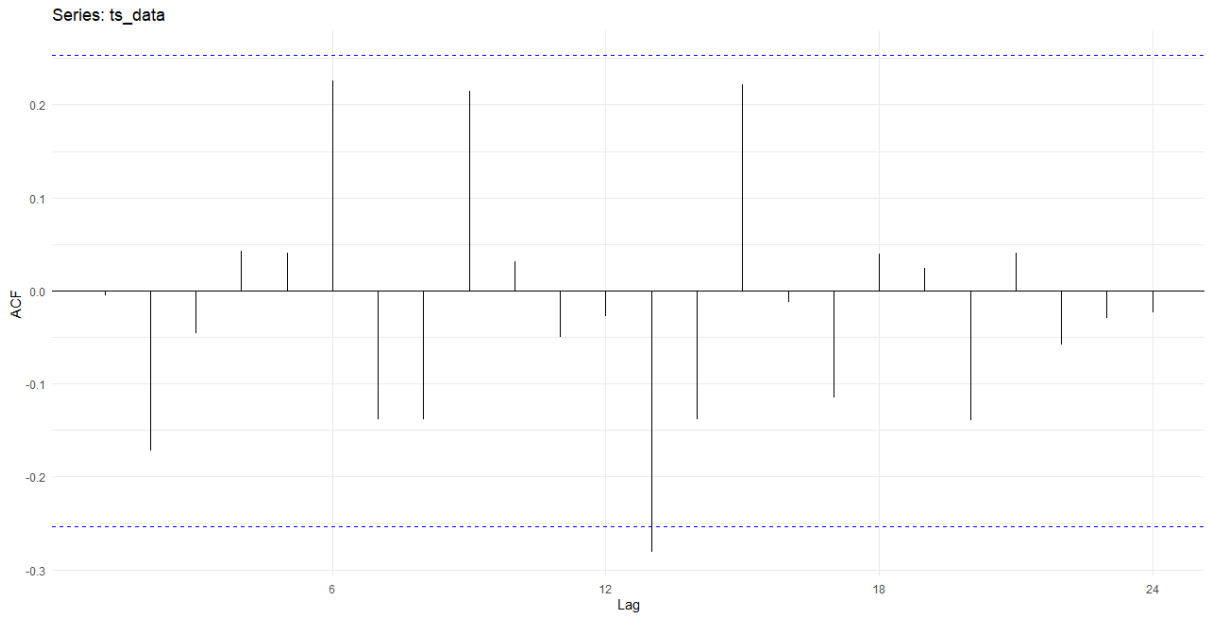

Figura 3.31: Análise de Gráfico de Função de Autocorrelação (ACF)

Há picos significativos de autocorrelação nas defasagens 1, 2, 12 e 16; a autocorrelação positiva nos lags 1 e 2 indica que há uma dependência entre valores próximos no tempo, ou seja, os valores de ausência de um mês são influenciados pelos valores dos meses imediatamente anteriores; o pico significativo no lag 12 sugere a presença de uma sazonalidade anual, significando que os padrões de ausência tendem a se repetir anualmente; e outro pico significativo no lag 16 pode indicar alguma influência de um ciclo diferente ou eventos periódicos específicos.

Existe autocorrelações negativas em algumas defasagens (como 6, 13 e 18) que indicam que há uma reversão na tendência dos dados nesses pontos. Isso pode ser interpretado como períodos de recuperação ou compensação após eventos de alta ausência. Muitos lags não apresentam autocorrelação significativa, sugerindo que, fora dos lags identificados, não há forte dependência temporal nos dados.

A presença de picos em defasagens curtas (1 e 2) indica que os valores de ausência são altamente dependentes dos valores dos meses imediatamente anteriores. O pico significativo na defasagem 12 confirma a sazonalidade anual observada anteriormente, indicando que eventos que ocorrem a cada ano têm um impacto significativo nas ausências. Correlações negativas em defasagens específicas sugerem que, após períodos de alta ausência, há uma tendência de valores subsequentes se recuperarem ou normalizarem.

Análise Combinada do Gráfico de Série Temporal, ACF e PACF:

Os gráficos da Figura [3.32] incluem a série temporal de Minutos de Absenteísmo por mês, composto pelo gráfico de autocorrelação (ACF) e pelo gráfico de autocorrelação

parcial (PACF). Esses gráficos juntos são extremamente úteis para entender a estrutura temporal dos dados e para selecionar modelos apropriados para previsão.

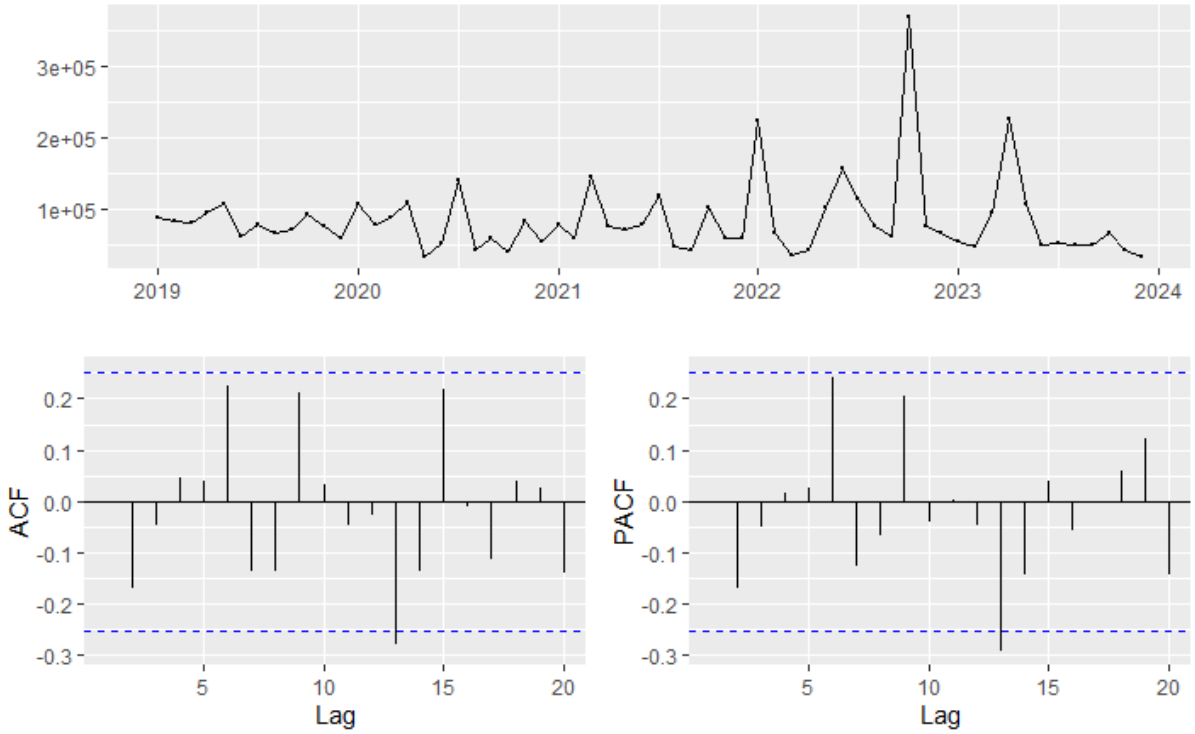

Figura 3.32: Análise Combinada do Gráfico de Série Temporal, ACF e PACF

O Gráfico de Autocorrelação Parcial (PACF) apresenta picos em atrasos curtos (1, 2) que confirmam a dependência de curto prazo observada no ACF; o Lag 12 no PACF também é significativo, reforçando a presença de sazonalidade anual. A correlação parcial diminui rapidamente após os primeiros atrasos, indicando que a maior parte da dependência temporal é capturada pelos primeiros atrasos.

A dependência de curto prazo, por indicação da presença de picos em atrasos curtos tanto no ACF quanto no PACF sugere que valores próximos no tempo são fortemente correlacionados. Isso pode ser modelado usando termos autorregressivos (AR) de baixa ordem. Os picos significativos no atraso 12 tanto no ACF quanto no PACF confirmam a sazonalidade anual, que pode ser modelada com termos sazonais. As autocorrelações negativas em certos atrasos indicam a presença de ciclos de recuperação após picos, o que pode ser importante ao ajustar modelos de previsão.

Técnica *Autoregressive Integrated Moving Average* (ARIMA)

O gráfico da Figura [3.33] aponta as previsões de Minutos de Absenteísmo do modelo Auto-ARIMA(0,0,0) com média diferente de zero. A análise da previsão versus os dados reais utilizando o modelo Auto-ARIMA(0,0,0) com média diferente de zero revela uma tentativa de ajuste aos dados históricos de ausência. No gráfico, a linha preta representa os dados reais do total de horas de ausência, enquanto a linha vermelha mostra os valores ajustados pelo modelo Auto-ARIMA.

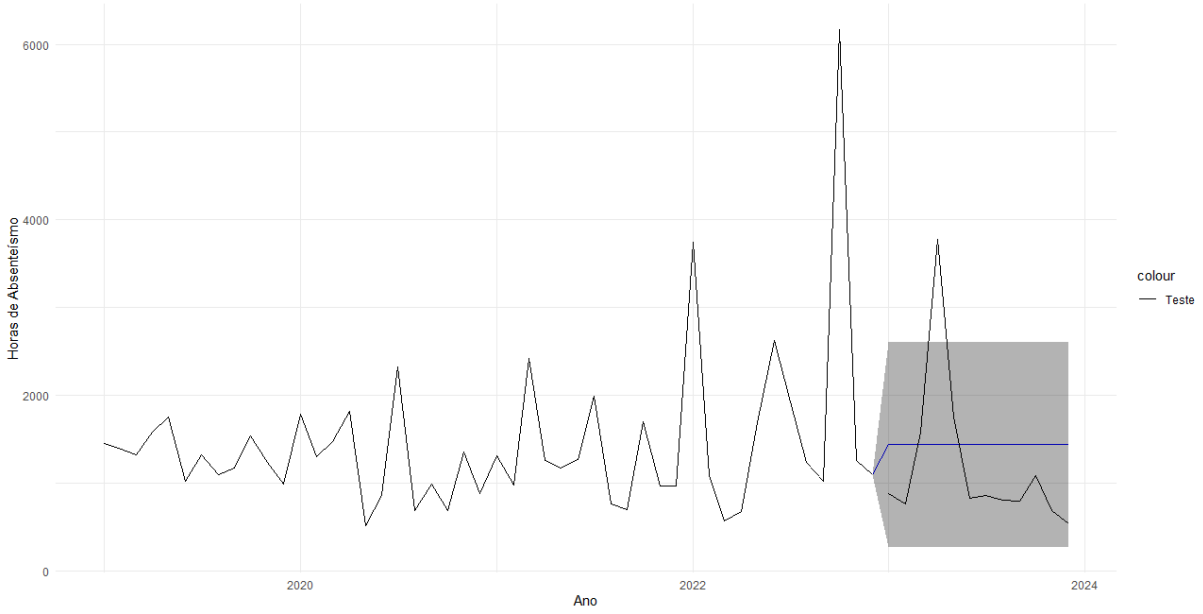

Figura 3.33: Predição da Técnica Auto-Arima de Absenteísmo

A área sombreada em azul ao redor da previsão indica o intervalo de confiança, proporcionando uma estimativa de incerteza para os valores previstos. Em termos de desempenho, embora a técnica Auto-ARIMA(0,0,0) consiga se alinhar aos dados históricos de forma básica, ela subestima significativamente os picos de ausência, especialmente o grande pico observado em 2023. A previsão permanece em um nível relativamente constante, incapaz de capturar a variabilidade presente nos dados reais, o que sugere que esse modelo não é ideal para séries com picos de alta variabilidade.

A análise dos resíduos da técnica ARIMA, conforme Figura [**??**] ao longo do tempo, revela limitações em capturar adequadamente os picos de ausência. O gráfico superior mostra os resíduos, definidos como as diferenças entre os valores reais e os valores ajustados, com grandes variações, especialmente em períodos de pico, indicando que a técnica não representa bem esses eventos extremos.

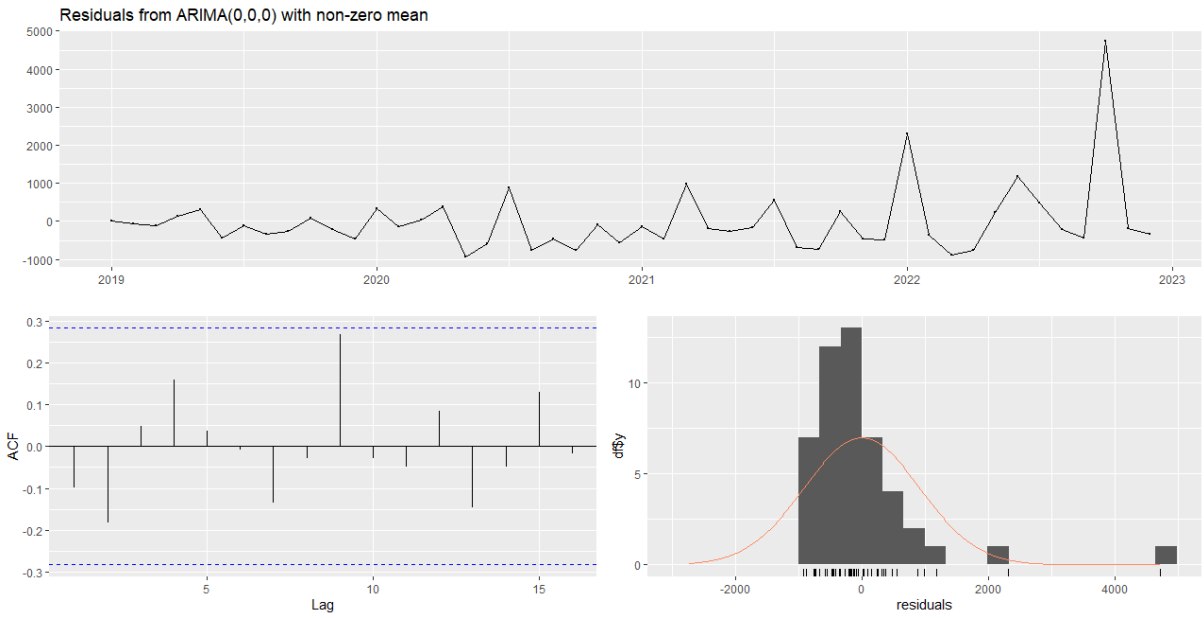

Figura 3.34: Resíduos Auto-Arima

O gráfico de autocorrelação residual (ACF) exibe a correlação dos resíduos em diferentes defasagens e apresenta picos significativos em algumas delas, sugerindo que os resíduos não são puramente aleatórios e que há padrões temporais que a técnica ARIMA não consegue captar.

Além disso, o histograma dos resíduos revela uma distribuição assimétrica, com uma longa cauda à direita, o que indica a presença de grandes valores de resíduos positivos. Esse padrão sugere que o modelo tende a subestimar os valores reais em determinados pontos, reforçando a necessidade de técnica mais robusta para capturar adequadamente a variabilidade nos dados de ausência.

A interpretação dos resultados da técnica ARIMA(0,0,0) revela limitações significativas na captura dos picos de ausência, com uma subestimação acentuada dos valores reais em períodos de alta ausência.

Além disso, a assimetria na distribuição dos resíduos, marcada por discrepâncias entre os valores reais e ajustados em determinados momentos, sugere a ocorrência de eventos atípicos ou mudanças de regime que a técnica ARIMA não consegue captar.

Em resumo, a técnica ARIMA(0,0,0) com média diferente de zero oferece uma base inicial, mas não é suficiente para capturar a complexidade e variabilidade dos dados de ausência. Uma análise aprofundada dos eventos atípicos pode contribuir para melhorar a precisão das previsões e facilitar um gerenciamento mais eficaz das ausências na Apex-Brasil.

3.4 Discussão e Principais Resultados

3.4.1 Avaliação das Variáveis

A análise exploratória de dados (AED) incluiu variáveis numéricas (como idade, dependentes, tempo de casa, salário mensal e horas de absenteísmo), analisadas por medidas de tendência central e dispersão, além de histogramas, boxplots e gráficos de densidade.

A matriz de correlação indicou que a idade tem correlação moderada com salário e tempo de casa, sugerindo que trabalhadores mais experientes e com mais tempo na empresa tendem a ter salários mais altos. As variáveis categóricas (como gênero, cargo, faixa etária e faixa de horas) foram analisadas com tabelas de frequência, gráficos de barras e testes de associação, revelando associação entre gênero e faixas de minutos, apontando diferenças de absenteísmo entre homens e mulheres.

A análise de variáveis numéricas, conforme estatísticas descritivas, demonstrou variação moderada entre as idades e grande disparidade salarial, enquanto o tempo de casa e minutos de absenteísmo apresentaram ampla variabilidade e valores extremos. Os gráficos de boxplot confirmaram a presença de *outliers* e não normalidade nas variáveis. A análise bivariada evidenciou correlações entre idade, tempo de casa e salário, sugerindo que fatores como experiência e progressão na carreira influenciam na remuneração.

A relação significativa entre gênero e minutos de absenteísmo reforça a necessidade de investigar fatores internos e externos que possam impactar o absenteísmo, como políticas organizacionais, diferenciação de cargos e responsabilidades. Em conjunto, os resultados destacam o valor de integrar variáveis adicionais para aprimorar a previsão e gestão de ausências na organização.

A visualização dos dados dos funcionários da Apex-Brasil permitiu o melhor entendimento da dinâmica de distribuição de idade, tempo de casa e salário, evidenciada nas correlações moderadas. Em termos de políticas de recursos humanos, a utilização de análises descritivas e a técnica de correlação de variáveis permitem entender a dinâmica dos empregados, condicionando as políticas.

Dada a existência da Apex-Brasil, que é relativamente nova, tendo a sua fundação oficial em 14 de maio de 2003, o perfil dos colaboradores, com uma idade média de 42,8 anos e com tempo médio de carreira de 8,5 anos, é mais que coerente. O entendimento desta variável influencia nas políticas de longo prazo, como previdência privada, carreira, cargos e salários ou, ainda, em questões de saúde e medicina do trabalho.

Outro acompanhamento importante é o de dependentes, que se demonstrou com uma média de 0,8 dependentes por funcionário. Esta variável é fundamental nas políticas de sustentabilidade de Recursos Humanos, onde a Agência apresenta uma série de benefícios e licenças compreensivas com relação ao apoio aos dependentes dos empregados. Outro

fato que denota o possível aumento de dependentes é que 68,1% dos empregados são solteiros ou não possuem cônjuge, o que indica (aliado à faixa etária) que é possível o crescimento de dependentes, companheiros ou filhos. Recomenda-se o monitoramento constante da variável, uma vez que impacta em termos orçamentários e no bem-estar do colaborador.

Um resultado que preocupa é a média de 231,9 horas de absenteísmo no período de 2019 a 2023, um valor considerado excessivo de ausências e licenças no ambiente do trabalho. Esta afirmativa é ainda impactada quando confrontada com o fato de uma idade média relativamente baixa. Este fato indica a necessidade de pesquisa adicional. No estudo de série temporal dessa pesquisa, fica claro o aumento após a pandemia, fato que também deve ser observado.

Algumas reflexões já são possíveis a partir das variáveis quantitativas, ao verificar que a política de cargos, carreiras e salários da Agência está refletida diretamente nos achados. A Agência, que é uma entidade paraestatal, tem a obrigação de realizar processo seletivo público, análogo aos concursos públicos, mas com um pouco mais de flexibilidade.

Esta política fica evidente ao demonstrar relação na análise do perfil dos funcionários em relação à idade, tempo na empresa e salário. Pois denota que o corpo técnico mais antigo apresenta maior salário, conforme previsto nas políticas de progressões e promoções.

Em termos das variáveis categóricas, a Agência apresenta uma relação entre homens e mulheres muito interessante em seu quadro, tendo um equilíbrio de gênero que atende às melhores políticas de sustentabilidade de gênero. As políticas de gênero têm um impacto direto nas projeções da área de Recursos Humanos, tendo em vista as características de cada gênero, bem como no complemento de conhecimentos, habilidades e atitudes complementares.

Ao se analisar a distribuição de cargo, a mesma está com mais da metade do quadro de pessoal concentrada na carreira de analista (nível superior), oriundos do processo seletivo público, fato que demonstra a força da carreira e sua capacidade de operacionalização da missão institucional. Em termos de ações de RH, é importante a previsão de Programa de Educação Corporativa que suporte o desenvolvimento de competências destes colaboradores, bem como dos assistentes (nível médio), tendo em vista a reciclagem de temas e o aprendizado contínuo.

Ao se separar o quadro de pessoal por faixas etárias, fica evidente a visualização de que pelo menos 70,1% das pessoas estão com idades entre 30 e 50 anos, o que leva à proposta de entender melhor este grupo e suas necessidades. Nos próximos anos, o envelhecimento da força de trabalho pode levar a necessidades de renovação por novos processos seletivos públicos e políticas de demissão voluntária. Outro aspecto a ser considerado é a gestão do conhecimento, pois a Agência fica vulnerável, caso não faça esta gestão no futuro, perto

da idade de aposentadoria deste grupo.

Ao realizar o Teste de Associação e Análise Bivariada, ficou evidenciada a relação entre gênero feminino e as maiores faixas de horas de absenteísmo, indicando que um gênero acaba por ter maiores ausências do ambiente de trabalho. Recomenda-se pesquisa junto a este grupo para evidenciar possíveis causas qualitativas para que se possa atuar na redução do absenteísmo.

Portanto, tendo em vista a análise de variáveis, existem diversos insumos e inspirações para investigações posteriores quanto aos dados e achados desta pesquisa. Recomenda-se a análise periódica e a implantação de painéis de dados, para que se implemente um modelo de processo decisório baseado em dados, bem como o compartilhamento com os gestores diretos dos funcionários.

3.4.2 Agrupamento

A análise de clusters revela grupos distintos de indivíduos com diferentes perfis demográficos e de emprego. Essas reflexões podem ser usadas para personalizar estratégias de engajamento, desenvolver programas específicos para diferentes segmentos de funcionários e otimizar políticas de retenção e benefícios. O melhor resultado das técnicas atualizadas é a técnica K-means testada com 8 (oito) agrupamentos e Índice de Silhueta com o valor de 0,34. O uso da análise PCA é recomendado em estudos subsequentes. A técnica selecionada foi capaz de propor uma segmentação baseada em dados de absenteísmo para funcionários na Apex-Brasil.

Pela técnica DBSCAN ter apresentado o melhor índice de silhueta, mas não ter feito o melhor agrupamento, conforme demonstrado na fase de Modelagem, confirma-se a necessidade de sempre analisar as técnicas empregadas e seu uso empírico. Outro ponto importante é o fato de comparar quantos clusters ou agrupamentos seriam necessários para aumentar a acurácia dos resultados, que também diferiu das indicações dos Gráficos do Método de Cotovelo e do Índice de Silhueta.

O agrupamento e segmentação dos empregados utilizando a técnica K-means permitiu uma separação clara de grupos que poderão inspirar a segmentação de atuação da Gerência de Recursos Humanos, permitindo uma abordagem diferenciada de diferentes soluções para públicos diversos.

Este fato fica evidenciado na análise do Perfil dos Empregados nos diversos clusters apresentados. O Cluster 1 (**Alta Administração com Mandato Fixo**)concentrou a Alta Administração da Agência, que tem tratamento diferenciado, mas pontual. Este nível tem um mandato de 4 (quatro) anos. Em tese, são pessoas de mais idade e não possuem dependentes. Apesar de receberem um tratamento pontual diferenciado devido à sua posição hierárquica, este grupo pode se beneficiar de ações voltadas à sucessão, à transição

de liderança e ao alinhamento estratégico, como *coaching* executivo e participação em conselhos ou funções de representação institucional.

O Cluster 2 (**Profissionais de Meia-Idade, Longa Permanência e Responsabilidades Familiares**)apresentou empregados de meia idade, com maior tempo de casa e do gênero masculino, considerando que a Agência tem 22 anos e quadro próprio a partir do primeiro processo seletivo público em 2007, ou seja, há 18 anos. Este grupo ainda apresenta uma quantidade média de 2,9 dependentes. A atuação de RH deve incluir políticas de equilíbrio entre vida pessoal e profissional, benefícios voltados à saúde familiar, programas de reconhecimento à trajetória e iniciativas de planejamento de carreira. Programas de sucessão e transferência de conhecimento também são recomendados.

O Cluster 3 (**Profissionais de Perfil Semelhante ao Cluster 2, com Menores Obrigações Familiares**) se assemelha em perfil ao Cluster 2, em termos de idade e gênero, mas o número de dependentes é inferior. Ou seja, permite a diferenciação de tratamento e a necessidade de segmentação prova-se eficaz. Essa diferenciação reforça a importância da segmentação para políticas mais refinadas. Aqui, recomenda-se o incentivo a trajetórias de desenvolvimento mais flexíveis, participação em projetos estratégicos e ações que estimulem a autonomia e mobilidade profissional.

O Cluster 4 (**Profissionais em Meio de Carreira com Equilíbrio de Gênero**) demonstra um equilíbrio maior de gênero e número médio de 1 dependente, que pode ser o cônjuge ou filho; portanto, também é notória a efetividade da segmentação ao diferenciar este grupo dos demais, por necessitarem de tratamento diverso. As ações de RH devem priorizar modelos híbridos de trabalho, programas de capacitação técnica e liderança intermediária, estratégias de engajamento que conciliem metas de desempenho com bem-estar, além da identificação de talentos para planejamento sucessório.

O Cluster 5 (**Mulheres Jovens com Maior Número de Dependentes**) já traz o grupo de gênero feminino e com maior média de dependentes, que pode se diferenciar dada a sobrecarga, conforme indicado na seção acima. Este grupo, junto com os agrupamentos com maior número de dependentes, deve ser analisado no estudo aprofundado do absenteísmo, buscando analisar possíveis causas. Recomenda-se que as políticas de RH promovam equidade de gênero, programas de liderança inclusiva, mentorias femininas e treinamentos para preparação de lideranças. A flexibilidade no trabalho e o suporte à expansão familiar também são fundamentais. Este Cluster, assim como outros com alta média de dependentes, deve ser priorizado em estudos aprofundados sobre absenteísmo.

Os Clusters 6 e 7 são formados por pessoas mais jovens, com tempo de carreira reduzido e baixa média de dependentes. A separação entre os Clusters 6 e 7 é coerente, sobretudo pela diferença no tempo de casa.faz a separação ser coerente e indicam grupos de pesquisa que podem ter ações diferenciadas do ponto de vista de carreira, mas similares nos demais

itens.

No Cluster 6 (**Jovens Recém-Admitidos**) e pessoas com média de 4,5 anos de carreira, onde as ações ideais incluem programas estruturados de integração, trilhas de desenvolvimento inicial, estímulo à participação em projetos transversais e criação de vínculos por meio de mentorias com profissionais mais experientes. A formação de identidade organizacional é prioridade para este grupo.

No Cluster 7 (**Jovens com Estabilidade Inicial de Carreira**), apesar da similaridade etária com o Cluster 6, este grupo já possui, em média, 4,5 anos de empresa, representando o início da consolidação profissional. O RH deve atuar com planejamento de carreira, participação em ações interáreas, reforço ao aprendizado contínuo e reconhecimento por desempenho.

Enfim, o Cluster 8 (**Profissionais Seniores com Poucos Dependentes**) apresenta pessoas com idade média maior, ou seja, pessoas mais maduras e com maior tempo de casa, mas com número médio de dependentes menor. Este Grupo é fundamental para o estudo da progressão da carreira e da retenção de conhecimento. As ações de RH devem incluir políticas de longevidade na carreira, papéis consultivos, preparação para a aposentadoria com transição gradativa, reconhecimento simbólico e programas de mentoria para novos talentos.

Portanto, a análise dos agrupamentos gerados por técnicas quantitativas traz reflexões valiosas para o desenho de políticas de recursos humanos mais justas, eficazes e personalizadas. Os perfis identificados permitem aprofundar estudos específicos, como o do absenteísmo, e viabilizam abordagens diferenciadas que fortalecem a resiliência organizacional, o bem-estar dos colaboradores e o alinhamento com os objetivos estratégicos da Apex-Brasil.

3.4.3 Classificação

No experimento de classificação, a Regressão Logística Multinomial é a técnica com alta precisão com todas as previsões ao longo da diagonal. Ela não apresenta classificações errôneas e é eficaz na distinção entre as classes.

A árvore de decisão mostra precisão moderada com um número maior de classificações errôneas em comparação com a regressão logística multinomial e a floresta aleatória. Ela apresenta problemas para distinguir entre as classes 1, 3 e 5.

A técnica KNN exibe baixa precisão com muitos elementos fora da diagonal e tem dificuldades significativas na classificação correta das instâncias, especialmente entre as classes 1, 3, 5 e 7.

O resultado a partir da segmentação, que foi enriquecida no banco de dados, e a eficácia do algoritmo realizar a classificação a partir dos resultados obtidos na técnica de

Regressão Logística Multinomial é impressionante, sendo o *overfitting* afastado a partir da análise da visualização na matriz de confusão.

Este resultado é fundamental, pois existe a necessidade de classificação dos empregados a partir do seu ingresso na Apex-Brasil. Ou seja, será muito fácil a análise de perfil e a correta classificação para incluir os novos perfis na segmentação proposta para a Agência, tornando mais fácil a adaptação e tratamento das necessidades destes empregados.

3.4.4 Predição de Séries Temporais

No experimento de predição de séries temporais, as técnicas SES e Auto.ARIMA são as mais eficazes para o conjunto de dados usado, ainda que tenham restrições, com o Auto.ARIMA melhor, dado que trata a sazonalidade, dados os valores mais baixos de RMSE, MAE e MAPE. As técnicas Holt-Winters (aditivos e multiplicativos) e SARIMA têm erros significativamente maiores e, portanto, são menos recomendadas para esta análise.

O tratamento de predição de séries temporais de dados de absenteísmo pode ser uma ferramenta essencial no aspecto de projeção de compras de vacinas, conscientização, orçamento de contratação de pessoal e negociação de acordos coletivos de trabalho.

Tendo em vista todos os estudos realizados, a predição junto com a segmentação proposta e a classificação realizada, podem dar direcionamento à gestão de recursos humanos e ao assessoramento à Alta Administração da Agência. Ao prever os picos de ausência e sua sazonalidade, ações de medicina e saúde no trabalho podem diminuir o absenteísmo.

Outro ponto importante, que pode influenciar, é o estudo de dimensionamento da força de trabalho, que impacta diretamente na definição de vagas e em processos de recrutamento e seleção, com a realização de processos seletivos públicos.

A quantidade de horas média apresentada na análise de variáveis mostra um gap médio de 231,9 horas, que apresenta um empregado a menos no quadro de pessoal. No estudo da série temporal, em especial no ano de 2022, a falta pode ter chegado a 10 pessoas. Isto demonstra o impacto nas rotinas e processos, bem como na sobrecarga dos empregados que não faltaram.

Em termos de negociação de abonos e licenças adicionais ou seu uso, estes dados podem gerar maior conscientização dos empregados ou direcionar a negociação dos acordos coletivos de trabalho futuros, uma vez que a organização apresenta seus ganhos e perdas e pode subsidiar as negociações no melhor direcionamento e à justa medida da utilização das licenças e abonos trabalhistas.

Conclusões, Limitações Intrínsecas e Trabalhos Futuros

Conclusões

Este trabalho teve como objetivo geral propor a construção e seleção de modelos mais adequados ao agrupamento e classificação de empregados e à predição de séries históricas para a Apex-Brasil, utilizando os dados primários oriundos de processos de gestão de pessoas. O escopo abrangeu informações sobre demografia dos empregados, salários, encargos, atributos funcionais e absenteísmo, com vistas a ampliar a capacidade analítica da área de Recursos Humanos por meio da adoção de técnicas de *People Analytics*.

A execução metodológica seguiu as etapas propostas pelo modelo CRISP-DM, cobrindo desde o entendimento do negócio até a proposição de implantação dos modelos em ambiente real. Os resultados obtidos demonstraram aderência aos objetivos definidos, sendo possível realizar com sucesso a clusterização dos empregados, a classificação preditiva de perfis e a projeção de padrões de absenteísmo com modelos robustos e métricas consistentes.

Foram analisados o contexto dos processos de recursos humanos onde foi realizada com sucesso a extração, transformação e carga dos dados de pessoas e de absenteísmo para a pesquisa em dois conjuntos de dados distintos.

Realizou-se a seleção sistêmica de variáveis relacionadas a processos de recursos humanos, para suportar a pesquisa e os modelos, adaptando-se às necessidades de formatação para processamento dos algoritmos selecionados. Foram pesquisadas e selecionadas técnicas de agrupamento, classificação e predição, onde foram construídos, comparados e avaliados com sucesso.

Em avaliação dos resultados obtidos em conjunto com as equipes de atuação da Gerência de Recursos Humanos da Apex-Brasil, os resultados denotam consistência e facilitam reflexões e indicações de negócio importantes. Por fim, recomendou-se a implantação das soluções no ambiente da Agência, por meio de projeto específico.

No tocante ao agrupamento, o algoritmo K-means se destacou como técnica mais eficaz, mesmo que a Técnica DBSCAN apresentasse índices melhores, dado ruído e erro de agrupamento dos empregados da Agência. A técnica K-means apresentou bons índices de silhueta e separação clara dos grupos de empregados.

No campo da classificação, a Regressão Logística Multinomial apresentou os melhores resultados em Acurácia, Precisão, Recall e F1-score, chegando a uma classificação sem erros, demonstrando que perfis previamente segmentados podem ser corretamente identificados com base em atributos individuais, sendo descartado o caso de *overfitting*. A técnica mostra alta precisão com todas as previsões ao longo da diagonal.

Para a previsão de séries temporais, a técnica Auto-ARIMA obteve os menores erros médios de previsão (RMSE e MAE), sendo altamente indicada para o planejamento de ações e orçamentos em períodos críticos de absenteísmo.

A avaliação conjunta dos modelos, realizada em colaboração com a Gerência de Recursos Humanos da Apex-Brasil, indicou a viabilidade prática da aplicação dos algoritmos no apoio à tomada de decisão. Estas discussões foram evidenciadas na discussão e apresentação dos principais resultados.

Todas as técnicas selecionadas obtiveram sucesso ao apontar indicações de pesquisas adicionais para investigar e qualificar os achados no futuro. Outro ponto importante foi verificar que a segmentação e a classificação se mostraram efetivas no aspecto de gestão de pessoas na Apex-Brasil, bem como a predição de séries temporais no caso de absenteísmo foi eficaz ao propor ações de curto, médio e longo prazo.

Com isso, considera-se como critério de sucesso alcançado a capacidade dos modelos propostos de gerar valor informacional e estratégico, indicando as técnicas mais eficazes para posterior automação e integração nos ambientes computacionais da Agência.

Além de reforçar a relevância da abordagem baseada em dados para a gestão de pessoas, os achados desta pesquisa revelam a maturidade crescente da Apex-Brasil na adoção de soluções analíticas, alinhando-se aos seus objetivos estratégicos de digitalização, transversalidade e foco em resultados.

Em termos de uso responsável, foi tratado o aspecto ético e transparente de dados sensíveis dos colaboradores, como um princípio basilar de toda a modelagem proposta, respeitando integralmente os marcos legais como a LGPD, onde os dados foram anonimizados.

Foi definida a recomendação para a adoção dos algoritmos baseados em aprendizado de máquina, devendo ser alocado orçamento, pessoas e fornecedores para este fim. A pesquisa alcançou seu objetivo de recomendações de implantação.

O trabalho desenvolvido na dissertação posiciona-se claramente como um vetor de fortalecimento dos capitais sociais da Apex-Brasil. Atua na realização e na potencializa-

ção dos capitais humano, organizacional e relacional, por meio da implementação de um sistema de *People Analytics* que promove transformação digital na gestão de pessoas.

Além disso, facilita a obtenção de impactos econômicos ao otimizar processos e reduzir custos associados ao absenteísmo. Consolida-se como uma intervenção de alto valor estratégico, que transcende a dimensão técnica e passa a integrar os ativos intangíveis da organização.

Limitações Intrínsecas

O agrupamento realizado nesta Dissertação foi conduzido com base em uma metodologia interna, utilizando exclusivamente os dados disponíveis da própria organização. Embora tenha sido adotada métrica robusta de avaliação interna, como o índice de Silhueta, não foi realizada validação externa dos agrupamentos.

Consequentemente, não é possível afirmar com segurança que os perfis identificados sejam generalizáveis para outros contextos organizacionais ou períodos temporais distintos. A ausência de replicação em bases externas, bem como de confrontação com modelos preexistentes ou referenciais, constitui uma limitação intrínseca do estudo, inerente à própria natureza dos dados e ao escopo metodológico adotado.

Essa limitação não compromete a validade interna dos achados, mas impõe restrições quanto à sua aplicabilidade em outros contextos, indicando a necessidade de que futuras pesquisas realizem validações externas para fortalecer a robustez dos modelos propostos.

Este estudo baseia-se fundamentalmente em dados administrativos oriundos dos sistemas de gestão de pessoas da organização. Embora esses dados sejam consistentes, estruturados e confiáveis para análises quantitativas, eles refletem apenas as informações registradas formalmente, o que implica em uma limitação intrínseca quanto à ausência de variáveis qualitativas ou subjetivas.

Aspectos como clima organizacional, percepção de justiça, bem-estar psicológico, engajamento, satisfação no trabalho e fatores contextuais externos não estão capturados, o que restringe uma compreensão mais ampla e holística dos fenômenos analisados.

Os métodos de agrupamento utilizados, especialmente o algoritmo DBSCAN, apresentam sensibilidade aos parâmetros de entrada, notadamente o raio de vizinhança ε e o número mínimo de pontos (MinPts). A definição desses parâmetros impacta diretamente a formação dos agrupamentos, podendo levar à identificação de um número maior ou menor de clusters, à inclusão de ruídos ou à junção indevida de grupos distintos.

Embora tenham sido realizadas análises de sensibilidade e validação interna para suportar a escolha dos parâmetros, essa característica é uma limitação intrínseca da técnica, sujeita à subjetividade de análise e à especificidade dos dados.

Para as técnicas de predição de séries temporais, os dados dos colaboradores por cluster não eram suficientes para o processamento dos algoritmos, limitando a capacidade das técnicas de representar completamente as dinâmicas comportamentais dos empregados.

A análise foi realizada sobre um conjunto de dados correspondente a um período específico da organização. Os perfis identificados refletem as características demográficas, organizacionais e comportamentais dos empregados nesse intervalo temporal. Mudanças organizacionais, políticas internas, contextos econômicos, avanços tecnológicos ou transformações culturais que ocorram após esse período podem alterar significativamente os padrões observados.

Os resultados possuem validade temporal limitada, não sendo garantido que as segmentações ou predições permaneçam aplicáveis no futuro sem atualizações e revalidações periódicas dos modelos. Restringe-se a generalização dos resultados para outros contextos organizacionais ou temporais.

As conclusões aqui apresentadas devem ser interpretadas à luz dessas limitações intrínsecas, o que não compromete sua validade interna, mas indica caminhos para futuras pesquisas que possam complementar e aprofundar os achados.

Trabalhos Futuros

Para os trabalhos futuros, propõe-se expandir a análise para outras frentes de RH, como clima organizacional, desempenho e rotatividade, considerando o enriquecimento dos conjuntos de dados e a proposta de novas segmentações e abordagens. Sugere-se revisar variáveis, técnicas e métricas com a equipe de Tecnologia da Informação (TI) e de Recursos Humanos (RH).

Pode-se aplicar técnicas de correlação não linear para incrementar a avaliação das variáveis, onde nesses casos, a associação ocorre por meio de padrões curvilíneos, exponenciais, logarítmicos ou outros comportamentos complexos, sendo necessário empregar métricas e modelos específicos para capturá-la.

Em termos de trabalhos futuros, deve-se implementar o desenvolvimento de dashboards preditivos para gestores na ferramenta *Power BI* da Microsoft, que é a oficial da Agência, para suportar o processo decisório da Agência, bem como proporcionar reflexões importantes na condução das políticas e estratégias de recursos humanos. Além disso, pode-se realizar testes com técnicas de redes neurais e *deep learning*, aumentando a quantidade de dados e observações. Em termos de infraestrutura, recomenda-se utilizar *RStudio Server*, *Power BI Gateway* e integração ao *ERP TOTVS*.

É possível, ainda, a aplicação de técnicas de Processamento de Linguagem Natural para análise de sentimentos em fichas de *feedbacks*. E, recomenda-se o monitoramento contínuo com pipelines automatizados integrados ao ERP da Agência.

Em seguida, recomenda-se a instituição de governança de dados específica para *People Analytics* no âmbito da Apex-Brasil. Deve-se estabelecer política de dados sensíveis para dados de RH, anonimização e acesso seguro.

Em termos de recomendações institucionais, sugere-se manter o alinhamento com objetivos estratégicos 2024–2027 da Agência, buscar garantir patrocínio da alta liderança para continuidade e investimentos, a partir dos resultados e ganhos demonstrados.

Deve-se fomentar uma cultura organizacional baseada em dados e evidências. Todos os resultados apresentam potencial de melhoria no planejamento estratégico da Agência, onde as análises podem apoiar decisões relativas à gestão de pessoas, bem-estar, orçamento e dimensionamento de pessoal.

Deve-se promover a disseminação de conhecimento, para capacitar continuadamente a equipe de RH em ciência de dados e em técnicas orientadas em processo decisório orientado a dados (*data driven*). Em termos de treinamento de equipe, aconselha-se capacitar os especialistas de RH em ciência de dados e visualização de dados, com especialização em *People Analytics*. Por fim, revisar periodicamente o desempenho das técnicas e recalibrar sempre que possível.

Conclui-se que a implantação dos modelos de *People Analytics* sugeridos nesta dissertação pode ocasionar ganhos significativos em termos de eficiência, bem-estar dos colaboradores e assertividade nas decisões. Este trabalho representa um avanço concreto no uso de ciência de dados para políticas públicas e reforça o papel da Apex-Brasil como instituição inovadora, estratégica e orientada por evidências.

Referências

[1] Brasil e Tribunal de Contas da União: *Referencial Básico de Governança do TCU - 3ª Edição*, 2020. https://portal.tcu.gov.br/lumis/portal/file/fileDownload.jsp?fileId=8A81881F7AB5B041017BABE767F6467E, acesso em 2024-07-13. 4, 14

[2] Nahapiet, Janine e Sumantra Ghoshal: *Social Capital, Intellectual Capital, and the Organizational Advantage*. The Academy of Management Review, 23(2):242, abril 1998, ISSN 03637425. http://www.jstor.org/stable/259373?origin=crossref, acesso em 2025-06-22. 4

[3] Edvinsson, Leif e Michael S. Malone: *Intellectual capital: realizing your company's true value by finding its hidden brainpower*. HarperBusiness, New York, 1. ed., [nachdr.] edição, 1999, ISBN 978-0-88730-841-3. 5

[4] Chapman, P. e et al.: *CRISP-DM 1.0 Step-by-step data mining guide*, 2000. 6

[5] Verri, Filipe Alves Neto: *Data Science Project: An Inductive Learning Approach*, dezembro 2024. https://zenodo.org/doi/10.5281/zenodo.14498011, acesso em 2025-06-22, Version Number: v0.1.0. 6

[6] Becker, Brian E. e Mark A. Huselid: *Strategic Human Resources Management: Where Do We Go From Here?* Journal of Management, 32(6):898–925, dezembro 2006, ISSN 0149-2063, 1557-1211. http://journals.sagepub.com/doi/10.1177/0149206306293668, acesso em 2024-07-06. 8

[7] Ruiz, Laura, Jose Benitez, Ana Castillo e Jessica Braojos: *Digital human resource strategy: Conceptualization, theoretical development, and an empirical examination of its impact on firm performance*. Information & Management, 61(4):103966, junho 2024, ISSN 03787206. https://linkinghub.elsevier.com/retrieve/pii/S037872062400048X, acesso em 2024-07-13. 8

[8] Chatterjee, Sheshadri, Ranjan Chaudhuri, Demetris Vrontis e Guido Giovando: *Digital workplace and organization performance: Moderating role of digital leadership capability*. Journal of Innovation & Knowledge, 8(1):100334, janeiro 2023, ISSN 2444569X. https://linkinghub.elsevier.com/retrieve/pii/S2444569X23000306, acesso em 2024-07-13. 9

[9] Ferrar, Jonathan e David Richard Green: *Excellence in people analytics: how to use workforce data to create business value*. Kogan Page Limited, London, United Kingdom New York, NY, USA, 2021, ISBN 978-0-7494-9829-0 978-0-7494-9830-6. 9, 16

[10] Gustafsson, Anders, Delphine Caruelle e David E. Bowen: *Customer experience (CX), employee experience (EX) and human experience (HX): introductions, interactions and interdisciplinary implications*. Journal of Service Management, 35(3):333–356, maio 2024, ISSN 1757-5818. https://www.emerald.com/insight/content/doi/10.1108/JOSM-02-2024-0072/full/html, acesso em 2024-07-13. 9

[11] Tursunbayeva, Aizhan: *Human resource technology disruptions and their implications for human resources management in healthcare organizations*. BMC Health Services Research, 19(1):268, dezembro 2019, ISSN 1472-6963. https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-019-4068-3, acesso em 2024-06-20. 10

[12] Kambur, Emine e Tulay Yildirim: *From traditional to smart human resources management*. International Journal of Manpower, 44(3):422–452, maio 2023, ISSN 0143-7720. https://www.emerald.com/insight/content/doi/10.1108/IJM-10-2021-0622/full/html, acesso em 2024-07-13. 10

[13] Aswale, Neerja e Kavya Mukul: *Role of Data Analytics in Human Resource Management for Prediction of Attrition Using Job Satisfaction*. Em Sharma, Neha, Amlan Chakrabarti e Valentina Emilia Balas (editores): *Data Management, Analytics and Innovation*, volume 1042, páginas 57–67. Springer Singapore, Singapore, 2020, ISBN 978-981-329-948-1 978-981-329-949-8. http://link.springer.com/10.1007/978-981-32-9949-8_5, acesso em 2024-07-13, Series Title: Advances in Intelligent Systems and Computing. 10

[14] Sato, Yusuke, Nobuyuki Kobayashi e Seiko Shirasaka: *A Proposal of HR System's Visualization Based on Harvard Model, Life Cycle, and Organization Strategy and Management Type*. Em *2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI)*, páginas 740–745, Toyama, Japan, julho 2019. IEEE, ISBN 978-1-72812-627-2. https://ieeexplore.ieee.org/document/8992816/, acesso em 2024-07-13. 10

[15] Peeters, Tina, Jaap Paauwe e Karina Van De Voorde: *People analytics effectiveness: developing a framework*. Journal of Organizational Effectiveness: People and Performance, 7(2):203–219, julho 2020, ISSN 2051-6614. https://www.emerald.com/insight/content/doi/10.1108/JOEPP-04-2020-0071/full/html, acesso em 2024-06-20. 11, 12, 15

[16] Rasmussen, Thomas H., Mike Ulrich e Dave Ulrich: *Moving People Analytics From Insight to Impact*. Human Resource Development Review, 23(1):11–29, março 2024, ISSN 1534-4843, 1552-6712. http://journals.sagepub.com/doi/10.1177/15344843231207220, acesso em 2024-06-20. 11, 14, 17

[17] Marler, Janet H. e John W. Boudreau: *An evidence-based review of HR Analytics*. The International Journal of Human Resource Management, 28(1):3–26, janeiro 2017, ISSN 0958-5192, 1466-4399. https://www.tandfonline.com/doi/full/10.1080/09585192.2016.1244699, acesso em 2024-06-20. 11

[18] Tursunbayeva, Aizhan, Stefano Di Lauro e Claudia Pagliari: *People analytics—A scoping review of conceptual boundaries and value propositions*. International Journal of Information Management, 43:224–247, dezembro 2018, ISSN 02684012. https://linkinghub.elsevier.com/retrieve/pii/S0268401218301750, acesso em 2024-06-20. 12

[19] Polzer, Jeffrey T.: *The rise of people analytics and the future of organizational research*. Research in Organizational Behavior, 42:100181, dezembro 2022, ISSN 01913085. https://linkinghub.elsevier.com/retrieve/pii/S0191308523000011, acesso em 2024-07-06. 12, 16, 17

[20] Cho, Wonhyuk, Seeyoung Choi e Hemin Choi: *Human Resources Analytics for Public Personnel Management: Concepts, Cases, and Caveats*. Administrative Sciences, 13(2):41, janeiro 2023, ISSN 2076-3387. https://www.mdpi.com/2076-3387/13/2/41, acesso em 2024-06-20. 12, 15, 17

[21] Hom, Peter W. e Rodger W. Griffeth: *Employee turnover*. South-Western series in human resources management. South-Western College Pub, Cincinnati, Ohio, 1994, ISBN 978-0-538-80873-6. 12

[22] Edwards, Martin R. e Kirsten Edwards: *Predictive HR analytics: mastering the HR metric*. KoganPage, London New York New Delhi, second edition edição, 2019, ISBN 978-0-7494-8444-6 978-0-7494-9802-3. 16, 18, 32

[23] Frederiksen, Anders: *Competing on analytics: The new science of winning*. Total Quality Management & Business Excellence, 20(5):583–583, maio 2009, ISSN 1478-3363, 1478-3371. http://www.tandfonline.com/doi/abs/10.1080/14783360902925454, acesso em 2025-06-22. 16

[24] Tursunbayeva, Aizhan, Claudia Pagliari, Stefano Di Lauro e Gilda Antonelli: *The ethics of people analytics: risks, opportunities and recommendations*. Personnel Review, 51(3):900–921, abril 2022, ISSN 0048-3486. https://www.emerald.com/insight/content/doi/10.1108/PR-12-2019-0680/full/html, acesso em 2024-06-20. 17

[25] Waters, Shonna D., Valerie N. Streets, Lindsay McFarlane e Rachael Johnson-Murray: *The practical guide to HR analytics: using data to inform, transform, and empower HR decisions*. SHRM, Society for Human Resource Management, Alexandria, Virginia, first edition edição, 2018, ISBN 978-1-58644-535-5 978-1-58644-532-4. 18, 32

[26] Fávero, Luiz: *Manual de análise de dados: estatística e modelagem multivariada com excel, SPSS e stata*. Elsevier, fevereiro 2017, ISBN 978-85-352-7087-7. 18

[27] Han, Jiawei, Micheline Kamber e Jian Pei: *Data mining: concepts and techniques*. Morgan Kaufmann series in data management systems. Elsevier/Morgan Kaufmann, Amsterdam Boston, 3rd ed edição, 2012, ISBN 978-0-12-381479-1. 18, 22

[28] Hyndman, Rob J. e George Athanasopoulos: *Forecasting: principles and practice*. Otexts, Online Open-Access Textbooks, Melbourne, Australia, third print edition edição, 2021, ISBN 978-0-9875071-3-6. 18, 26, 27

[29] Lloyd, S.: *Least squares quantization in PCM*. IEEE Transactions on Information Theory, 28(2):129–137, março 1982, ISSN 0018-9448. http://ieeexplore.ieee.org/document/1056489/, acesso em 2024-07-14. 19

[30] Macqueen, J: *SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS*. MULTIVARIATE OBSERVATIONS, páginas 281–297. 19

[31] Jain, Anil K.: *Data clustering: 50 years beyond K-means*. Pattern Recognition Letters, 31(8):651–666, junho 2010, ISSN 01678655. https://linkinghub.elsevier.com/retrieve/pii/S0167865509002323, acesso em 2024-07-14. 19

[32] Murtagh, Fionn e Pedro Contreras: *Algorithms for hierarchical clustering: an overview*. WIREs Data Mining and Knowledge Discovery, 2(1):86–97, janeiro 2012, ISSN 1942-4787, 1942-4795. https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.53, acesso em 2024-07-06. 20

[33] Murtagh, Fionn e Pierre Legendre: *Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?* Journal of Classification, 31(3):274–295, outubro 2014, ISSN 0176-4268, 1432-1343. http://link.springer.com/10.1007/s00357-014-9161-z, acesso em 2024-07-14. 20

[34] Rokach, Lior e Oded Maimon: *Clustering Methods*. Em Maimon, Oded e Lior Rokach (editores): *Data Mining and Knowledge Discovery Handbook*, páginas 321–352. Springer-Verlag, New York, 2005, ISBN 978-0-387-24435-8. http://link.springer.com/10.1007/0-387-25465-X_15, acesso em 2024-07-14. 20

[35] Xu, R. e D. WunschII: *Survey of Clustering Algorithms*. IEEE Transactions on Neural Networks, 16(3):645–678, maio 2005, ISSN 1045-9227. http://ieeexplore.ieee.org/document/1427769/, acesso em 2024-07-14. 20

[36] Ester, Martin, Hans Peter Kriegel e Xiaowei Xu: *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. 21, 22, 55

[37] Campello, Ricardo J. G. B., Davoud Moulavi e Joerg Sander: *Density-Based Clustering Based on Hierarchical Density Estimates*. Em Hutchison, David, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda e Guandong Xu (editores): *Advances in Knowledge Discovery and Data Mining*, volume 7819, páginas 160–172. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, ISBN 978-3-642-37455-5 978-3-642-37456-2. http://link.springer.com/10.1007/978-3-642-37456-2_14, acesso em 2024-07-14, Series Title: Lecture Notes in Computer Science. 21

[38] Schubert, Erich, Jörg Sander, Martin Ester, Hans Peter Kriegel e Xiaowei Xu: *DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN*. ACM Transactions on Database Systems, 42(3):1–21, setembro 2017, ISSN 0362-5915, 1557-4644. https://dl.acm.org/doi/10.1145/3068335, acesso em 2024-07-14. 21

[39] Hosmer, David W., Stanley Lemeshow e Rodney X. Sturdivant: *Applied Logistic Regression*. Wiley Series in Probability and Statistics. Wiley, 1ª edição, março 2013, ISBN 978-0-470-58247-3 978-1-118-54838-7. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387, acesso em 2024-07-06. 22

[40] Cox, D. R.: *The Regression Analysis of Binary Sequences*. Journal of the Royal Statistical Society Series B: Statistical Methodology, 20(2):215–232, julho 1958, ISSN 1369-7412, 1467-9868. https://academic.oup.com/jrsssb/article/20/2/215/7027376, acesso em 2024-07-14. 22

[41] Menard, Scott W.: *Applied logistic regression analysis*. Número 106 em *Quantitative applications in the social sciences*. Sage Publ, Thousand Oaks, Calif., 2. ed., [nachdr.] edição, 2008, ISBN 978-0-7619-2208-7. 22

[42] Breiman, Leo, Jerome H. Friedman, Richard A. Olshen e Charles J. Stone: *Classification And Regression Trees*. Routledge, 5-32, 1ª edição, outubro 2017, ISBN 978-1-315-13947-0. https://www.taylorfrancis.com/books/9781351460491, acesso em 2024-07-06. 23

[43] Quinlan, J. R.: *Induction of decision trees*. Machine Learning, 1(1):81–106, março 1986, ISSN 0885-6125, 1573-0565. http://link.springer.com/10.1007/BF00116251, acesso em 2024-07-14. 23

[44] Safavian, S.R. e D. Landgrebe: *A survey of decision tree classifier methodology*. IEEE Transactions on Systems, Man, and Cybernetics, 21(3):660–674, junho 1991, ISSN 00189472. http://ieeexplore.ieee.org/document/97458/, acesso em 2024-07-14. 23

[45] Breiman, Leo: *Random Forests. Machine Learning*. Machine Learning, 45(1):5–32, 2001, ISSN 08856125. http://link.springer.com/10.1023/A:1010933404324, acesso em 2024-07-14. 24

[46] Cutler, D. Richard, Thomas C. Edwards, Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson e Joshua J. Lawler: *RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY*. Ecology, 88(11):2783–2792, novembro 2007, ISSN 0012-9658. http://doi.wiley.com/10.1890/07-0539.1, acesso em 2024-07-14. 24

[47] Cover, T. e P. Hart: *Nearest neighbor pattern classification*. IEEE Transactions on Information Theory, 13(1):21–27, janeiro 1967, ISSN 0018-9448, 1557-9654. http://ieeexplore.ieee.org/document/1053964/, acesso em 2024-07-14. 24

[48] Altman, N. S.: *An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression*. The American Statistician, 46(3):175–185, agosto 1992,

ISSN 0003-1305, 1537-2731. http://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879, acesso em 2024-07-06. 24

[49] Duda, Richard O., Peter E. Hart e David G. Stork: *Pattern classification*. Wiley, New York, 2nd ed edição, 2001, ISBN 978-0-471-05669-0. 24

[50] Gardner, Everette S.: *Exponential smoothing: The state of the art*. Journal of Forecasting, 4(1):1–28, janeiro 1985, ISSN 0277-6693, 1099-131X. https://onlinelibrary.wiley.com/doi/10.1002/for.3980040103, acesso em 2024-07-14. 25, 26

[51] Chatfield, C.: *The Holt-Winters Forecasting Procedure*. Applied Statistics, 27(3):264, 1978, ISSN 00359254. https://www.jstor.org/stable/10.2307/2347162?origin=crossref, acesso em 2024-07-14. 25

[52] Winters, Peter R.: *Forecasting Sales by Exponentially Weighted Moving Averages*. Management Science, 6(3):324–342, abril 1960, ISSN 0025-1909, 1526-5501. https://pubsonline.informs.org/doi/10.1287/mnsc.6.3.324, acesso em 2024-07-06. 26

[53] Box, George E. P., Gwilym M. Jenkins, Gregory C. Reinsel e Greta M. Ljung: *Time series analysis: forecasting and control*. Wiley series in probability and statistics. John Wiley & Sons, Inc, Hoboken, New Jersey, fifth edition edição, 2016, ISBN 978-1-118-67502-1. 27

[54] Hamilton, James D.: *Time series analysis*. Princeton University Press, Princeton, N.J, 1994, ISBN 978-0-691-04289-3. 27

[55] Mariano, Ari Melo e Maíra ROCHA Santos: *Revisão da Literatura: Apresentação de uma Abordagem Integradora.* 2017. 28

[56] Perianes-Rodriguez, Antonio, Ludo Waltman e Nees Jan Van Eck: *Constructing bibliometric networks: A comparison between full and fractional counting*. Journal of Informetrics, 10(4):1178–1195, novembro 2016, ISSN 17511577. https://linkinghub.elsevier.com/retrieve/pii/S1751157716302036, acesso em 2024-06-20. 28

[57] Garg, Swati, Shuchi Sinha, Arpan Kumar Kar e Mauricio Mani: *A review of machine learning applications in human resource management*. International Journal of Productivity and Performance Management, 71(5):1590–1610, maio 2022, ISSN 1741-0401. https://www.emerald.com/insight/content/doi/10.1108/IJPPM-08-2020-0427/full/html, acesso em 2024-09-08. 28

[58] Giermindl, Lisa Marie, Franz Strich, Oliver Christ, Ulrich Leicht-Deobald e Abdullah Redzepi: *The dark sides of people analytics: reviewing the perils for organisations and employees*. European Journal of Information Systems, 31(3):410–435, maio 2022, ISSN 0960-085X, 1476-9344. https://www.tandfonline.com/doi/full/10.1080/0960085X.2021.1927213, acesso em 2024-11-08. 28

[59] Berkhin, P.: *A Survey of Clustering Data Mining Techniques*. Em Kogan, Jacob, Charles Nicholas e Marc Teboulle (editores): *Grouping Multidimensional Data*, páginas 25–71. Springer-Verlag, Berlin/Heidelberg, 2006, ISBN 978-3-540-28348-5. http://link.springer.com/10.1007/3-540-28349-8_2, acesso em 2024-09-08. 28

[60] Andriyani, Fitri e Yan Puspitarani: *Performance Comparison of K-Means and DB-Scan Algorithms for Text Clustering Product Reviews*. SinkrOn, 7(3):944–949, julho 2022, ISSN 2541-2019, 2541-044X. https://jurnal.polgan.ac.id/index.php/sinkron/article/view/11569, acesso em 2024-09-08. 28

[61] Gupta, Sonal e R.R.K. Sharma: *Types of HR Analytics Used for the Prediction of Employee Turnover in Different Strategic Firms with the use of Enterprise Social Media*. Em *Proceedings of the International Conference on Industrial Engineering and Operations Management*, páginas 1977–1994, Rome, Europe, julho 2022. IEOM Society International, ISBN 978-1-79239-161-3. https://index.ieomsociety.org/index.cfm/article/view/ID/10854, acesso em 2024-09-08. 29

[62] Kakulapati, V., Kalluri Krishna Chaitanya, Kolli Vamsi Guru Chaitanya e Ponugoti Akshay: *Predictive analytics of HR - A machine learning approach*. Journal of Statistics and Management Systems, 23(6):959–969, agosto 2020, ISSN 0972-0510, 2169-0014. https://www.tandfonline.com/doi/full/10.1080/09720510.2020.1799497, acesso em 2024-11-08. 29

[63] Lismont, Jasmien, Jan Vanthienen, Bart Baesens e Wilfried Lemahieu: *Defining analytics maturity indicators: A survey approach*. International Journal of Information Management, 37(3):114–124, junho 2017, ISSN 02684012. https://linkinghub.elsevier.com/retrieve/pii/S0268401216305655, acesso em 2024-11-08. 29

[64] Alsaadi, Elham Mohammed Thabit A., Sameerah Faris Khlebus e Ashwak Alabaichi: *Identification of human resource analytics using machine learning algorithms*. TELKOMNIKA (Telecommunication Computing Electronics and Control), 20(5):1004, outubro 2022, ISSN 2302-9293, 1693-6930. http://telkomnika.uad.ac.id/index.php/TELKOMNIKA/article/view/21818, acesso em 2024-09-08. 29

[65] Yin, Qing: *Comparison of Machine Learning Models for Employee Turnover Prediction*. Applied and Computational Engineering, 8(1):228–232, agosto 2023, ISSN 2755-2721, 2755-273X. https://www.ewadirect.com/proceedings/ace/article/view/2533, acesso em 2024-09-08. 29

[66] P, Juhi Padmaja, Vinoodhini D e Uma K. V: *Effective Classification Of Ibm Hr Analytics Employee Attrition Using Sampling Techniques*. Em *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, páginas 1–6, Bhilai, India, abril 2022. IEEE, ISBN 978-1-66541-120-2. https://ieeexplore.ieee.org/document/9808057/, acesso em 2024-11-08. 29

[67] Lawrance, Natalie, George Petrides e Marie Anne Guerry: *Predicting employee absenteeism for cost effective interventions*. Decision Support Systems, 147:113539,

agosto 2021, ISSN 01679236. https://linkinghub.elsevier.com/retrieve/pii/S016792362100049X, acesso em 2024-11-08. 29

[68] Fallucchi, Francesca, Marco Coladangelo, Romeo Giuliano e Ernesto William De Luca: *Predicting Employee Attrition Using Machine Learning Techniques*. Computers, 9(4):86, novembro 2020, ISSN 2073-431X. https://www.mdpi.com/2073-431X/9/4/86, acesso em 2024-11-08. 29

[69] Zhu, Xiaojuan, William Seaver, Rapinder Sawhney, Shuguang Ji, Bruce Holt, Gurudatt Bhaskar Sanil e Girish Upreti: *Employee turnover forecasting for human resource management based on time series analysis*. Journal of Applied Statistics, 44(8):1421–1440, junho 2017, ISSN 0266-4763, 1360-0532. https://www.tandfonline.com/doi/full/10.1080/02664763.2016.1214242, acesso em 2025-02-02. 30

[70] Bandeira, Saymon Galvão, Symone Gomes Soares Alcalá, Roberto Oliveira Vita e Talles Marcelo Gonçalves De Andrade Barbosa: *Comparison of selection and combination strategies for demand forecasting methods*. Production, 30:e20200009, 2020, ISSN 1980-5411, 0103-6513. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-65132020000100210&tlng=en, acesso em 2024-09-08. 30

[71] Margherita, Alessandro: *Human resources analytics: A systematization of research topics and directions for future research*. Human Resource Management Review, 32(2):100795, junho 2022, ISSN 10534822. https://linkinghub.elsevier.com/retrieve/pii/S1053482220300681, acesso em 2024-11-08. 30

[72] Xu, Huang, Zhiwen Yu, Jingyuan Yang, Hui Xiong e Hengshu Zhu: *Dynamic Talent Flow Analysis with Deep Sequence Prediction Modeling*. IEEE Transactions on Knowledge and Data Engineering, 31(10):1926–1939, outubro 2019, ISSN 1041-4347, 1558-2191, 2326-3865. https://ieeexplore.ieee.org/document/8478343/, acesso em 2024-11-08. 30

[73] Bertalanffy, L. von.: *Teoria Geral Dos Sistemas: Fundamentos, Desenvolvimentos E Aplicações*. Editora Vozes, 2008, ISBN 978-85-326-3690-4. 31

[74] Alcoforado, Luciane: *Utilizando a linguagem R*. Alta Books, outubro 2021, ISBN 978-85-508-1442-1. 34

[75] Devore, Jay L.: *Probabilidade e estatística para engenharia e ciências*. Cengage Learning, julho 2021, ISBN 978-85-221-2803-7. 47, 55

[76] Wickham, Hadley e Garrett Grolemund: *R para Data Science: importe, arrume, transforme, visualize e modele dados*. Alta Books, outubro 2021, ISBN 978-85-508-0324-1. 61