Grupo 3
Everymind

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
05/08/2022	Stefano Tinelli	0	Criação do documento
09/08/2022	Stefano Tinelli	1	Adição da Análise SWOT, Matriz de Risco na seção 4.
11/08/2022	Felipe Gomes, Stefano tinelli	1.1	Adição da Descrição de tabelas e Descrição dos Dados (dentro do tópico 4.2 Compreensão de dados - Descrição dos Dados)
12/08/2022	Felipe Gomes, Gustavo Monteiro, Mateus Neves	1.2	Preenchimento da classificação de dados e com demais dados, gráficos e análises realizadas. Adição de análises de indústrias e a descrição dos métodos utilizados para as análises.
25/08/2022	Stefano	1.3	Adicionado descrição de código de tratamento de datas.
26/08/2022	Gustavo	1.4	Ajuste nas personas, jornada do usuário e descrição das variáveis, com a descrição dos tratamentos da tabela “Reconhecimento”, com a análise prévia desses dados.
07/09/2022	Gustavo	1.5	Adicionado a descrição da metodologia utilizada na modelagem e na avaliação da modelagem do projeto.
09/09/2022	Gustavo, Alan, Stefano	1.6	Adicionada a descrição das modelagem preditivas utilizados, e a descrição dos resultados obtidos com os respectivos modelos.
12/09/2022	Gustavo	1.7	Acrescentada a Introdução e modificada a descrição dos objetivos gerais, específicos e

			justificativa
12/09/2022	Beny	1.8	Acrescentada a descrição da metodologia e das ferramentas utilizadas
21/09/2022	Alan e Felipe	1.9	Acrescentada a descrição do tratamento da tabela “Ambiente de Trabalho”
06/10/2022	Gustavo	2.0	Preenchida a seção 5 com os resultados obtidos, a conclusão do trabalho e recomendações.

Sumário

1. Introdução

O presente trabalho foi realizado por alunos do Inteli - Faculdade de Tecnologia e Liderança, sendo eles:

- Alan Rozensztajn Schipper;
- Beny Frid;
- Felipe Gomes Rodrigues dos Santos;
- Gustavo Monteiro;
- Mateus Guimarães Coelho Neves e,
- Stefano Matiotta Tinelli.

Com a parceria da empresa Everymind, a fim de desenvolver um trabalho acadêmico de um projeto com um cliente de mercado atuante, sob a orientação da Professora Ana Cristina dos Santos, durante o módulo 3, do 2º semestre.

A Everymind é uma empresa de médio porte, com cerca de 250 funcionários, com sede em São Paulo - SP, fundada em 2014, que faz parte do grupo Compass.uol. Também possui parceria com a plataforma Salesforce, atuando no setor de tecnologia e computação, prestando consultoria e desenvolvendo projetos de sistemas em ERP e CRM, utilizando produtos baseados em Salesforce, na gestão e implementação dos mesmos para indústrias de bens de consumo, varejo, saúde, empresas do mercado financeiro, serviços, energia, entre outras.

O intuito deste projeto é verificar se a modelagem preditiva, utilizando aprendizado de máquinas, pode ser útil para identificar a tendência de *turnover* de colaboradores da empresa parceira, com base na amostra de dados cedidos pela mesma, desta forma, para aprimorar o processo de gestão dos times pelos *Squad Leaders*, que atuam com os grupos de colaboradores dentro da empresa.

A rotatividade do quadro de colaboradores das empresas impacta diretamente nos custos de recrutamento, seleção, treinamento, e capacitação, além da perda de talentos e mão de obra, essenciais para a saúde da empresa e qualidade dos serviços prestados (Sansone e Vecchia, 2021).

Ainda é importante pontuar que a gestão de pessoas com a utilização da análise de dados de funcionários faz parte de uma metodologia de *People Analytics*, que ajuda o setor de

Pessoas e Cultura, e de RH de uma organização, a melhorar seus processos e otimizar estratégias de atração e retenção de talentos (Dias, 2022).

A utilização de modelos preditivos, em *machine learning*, neste projeto, é importante pela natureza do objetivo da análise, na identificação de padrões e para a previsão de tendências. Os avanços tecnológicos em *Machine Learning* ainda trazem o aspecto de melhoria e otimização dos processos utilizados pela modelagem preditiva e os algoritmos utilizados, tornando a metodologia aplicada uma solução adequada, já que a automatização da solução proposta depende dessa adaptação ao longo do tempo (Sansone e Vecchia, 2021).

2. Objetivos e Justificativa

2.1. Objetivos

2.1.1. Objetivo geral

Construção de um modelo preditivo, utilizando dados amostrais referentes aos últimos dois anos (referentes do ano 2020 ao ano 2022) de colaboradores da empresa, para classificação de funcionários que desejam se desligar da empresa. Decorre também deste projeto verificar o desempenho entre diferentes algoritmos de *Machine Learning*, em relação ao contexto da problemática levantada pelo parceiro de negócios, sobre a rotatividade de colaboradores na empresa.

2.1.2. Objetivos específicos

Foram estabelecidos os seguintes objetivos específicos, a fim de atingir o objetivo geral descrito acima:

- Tratamento dos dados da amostra, avaliando o desempenho dos modelos aplicados ao projeto;
- Identificar e testar modelos preditivos elegíveis para a base de dados amostral disponibilizada;
- Analisar os resultados obtidos através dos experimentos realizados, comparando a performance entre os modelos avaliados;

2.2. Justificativa

A qualidade dos serviços prestados pelas empresas depende diretamente da qualidade dos profissionais que contrata, do treinamento que disponibiliza aos seus colaboradores, e do clima organizacional (Sansone e Vecchia, 2021). Por isso é muito importante a utilização de ferramentas que ajudem a empresa na elaboração de estratégias para aumentar a retenção de talentos. Além disso, as aplicações de *Machine Learning* vem sendo cada vez mais utilizadas em diversos setores, portanto, a utilização destes recursos para ajudar a empresa a reduzir os custos, e a melhorar a sua gestão interna, passa a ser uma vantagem competitiva em relação aos seus concorrentes (Ortiz, 2021).

3. Metodologia

3.1. CRISP-DM

O CRISP-DM (*Cross Industry Standard Process for Data Mining*) é uma maneira comprovada de orientar seus esforços de mineração de dados. Como metodologia, inclui descrições das fases típicas de um projeto, as tarefas envolvidas em cada fase e uma explicação das relações entre essas tarefas. É um processo cíclico que utiliza-se de retorno de etapas quando se avalia necessário durante o desenvolvimento.

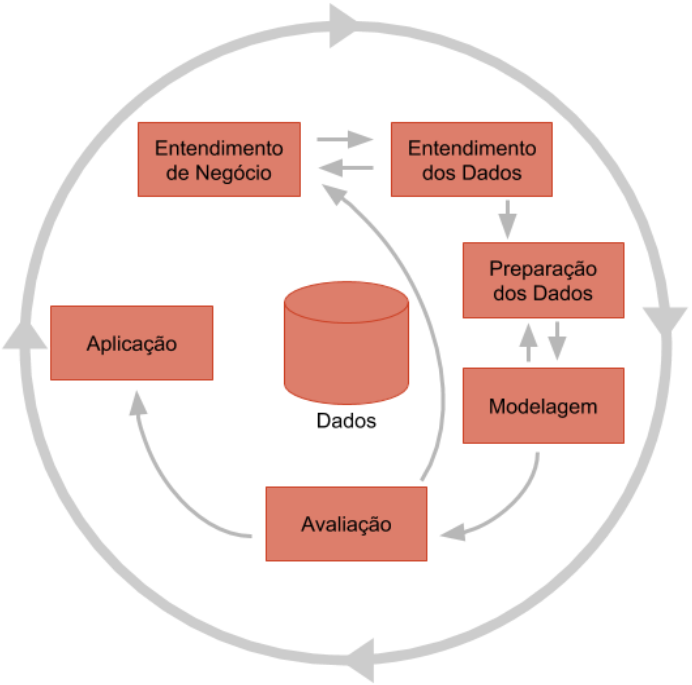


Figura 1. Diagrama da metodologia CRISP-DM.

3.1.1 Compreensão do negócio

Concentra-se na compreensão dos objetivos e requisitos do projeto a partir de uma perspectiva de negócios. O analista formula esse conhecimento como um problema de mineração de dados e desenvolve um plano preliminar. É nessa etapa que surge a compreensão da problemática e é de extrema importância para o desenvolvimento da solução de Ciência de Dados. Essa abordagem estende-se por todo o desenvolvimento.

3.1.2 Compreensão de dados

Começando com a coleta de dados inicial, o analista prossegue com as atividades para se familiarizar com os dados, identificar problemas de qualidade de dados e descobrir os primeiros *insights* sobre os dados.

3.1.3 Preparação de dados

A fase de preparação de dados abrange todas as atividades para construir o conjunto de dados final a partir dos dados brutos iniciais. Essa etapa compreende a seleção por atributos desses dados não tratados e a partir disso a limpeza dos mesmos. Assim, desenvolve-se o tratamento desses dados, alterando/retirando caracteres especiais, espaços, agregações, que podem acabar dificultando ou impossibilitando a utilização deles em um modelo.

3.1.4 Modelagem

O analista avalia, seleciona e aplica as técnicas de mineração de dados mais apropriadas, dependendo dos objetivos identificados na primeira fase. Assim é possível descobrir e analisar padrões nos dados.

3.1.5 Avaliação

O analista constrói e escolhe modelos que parecem ter alta qualidade. O analista os testa para garantir que eles possam generalizar os modelos em relação a dados não vistos. O analista também valida se os modelos cobrem suficientemente todas as áreas do negócio. O modelo é então estimado com base em seus resultados, utilizando métricas como o nível de acurácia. (*melhor descrição no tópico* ***4.5.2. Métricas utilizadas para avaliação dos modelos***)

3.1.6 Implantação

Geralmente, nesta fase final é hora de colocar o modelo em produção, para que possa ser usado. É importante salientar que uma vez entregue o modelo deve ser monitorado para que ele possa continuar útil e evoluí-lo quando necessário. Apesar de não ser o projeto final, o modelo gerado faz parte de seu desenvolvimento, sendo aplicado em uso real.

3.2. Ferramentas

3.2.1 Google Colab

Machine : O Colab permite que qualquer pessoa escreva e execute código Python por meio do navegador e é especialmente adequado para machine learning, análise de dados e educação.

3.2.2 Google Drive

Storage: Ele permite que os usuários armazenem e acessem arquivos online. O serviço sincroniza documentos armazenados, fotos e muito mais em todos os dispositivos do usuário, assim todos podem acessar os documentos e guardar eles de forma sistemática e organizada

3.2.3 GitHub

Repository: Ele é uma plataforma de desenvolvimento de software online usada para armazenar, rastrear e colaborar em projetos de software. Ele permite que os desenvolvedores carreguem seus próprios arquivos de código e colaborem com outros desenvolvedores em projetos de código aberto.

4. Desenvolvimento e Resultados

4.1. Compreensão do Problema

A partir de análises realizadas com diferentes metodologias e ferramentas, é possível perceber um setor competitivo, que além de competir por projetos e soluções para os clientes, também compete por profissionais para trabalhar nas empresas.

Existe uma tendência de turnover de profissionais de TI nas empresas, que podem apresentar o hábito de passar por empresas diferentes, dificultando uma retenção e encarecendo o treinamento que necessita para se especializar em produtos como os da Salesforce, que é o tipo de produto mais utilizado pela empresa Everymind.

A seguir serão apresentados as ferramentas, matrizes e análises realizadas, com uma breve descrição da importância de sua utilização. Porém é importante ressaltar que as análises e as ferramentas foram feitas no período do segundo semestre do ano de 2022, e foram feitas com base no contexto da empresa durante este período.

4.1.1. Contexto da indústria

A análise do microambiente é feita utilizando as 5 forças de Porter, possibilitando ambientar o setor em que a empresa está inserida, indicando potenciais competidores, e até mesmo competidores de outros setores que podem substituir o produto oferecido pela empresa, indo além de ver apenas os competidores diretos. Através deste método é possível realizar um planejamento estratégico de forma mais robusta, por analisar diferentes pontos, levando em consideração também o poder de negociação dos clientes e fornecedores (Bruijl, 2018).

4.1.1.1 Rivalidade entre Concorrentes

Dentro do setor de Tecnologia e computação, mais especificamente entre as empresas de desenvolvimento de sistemas, como CRM (*Customer Relationship Management*) e ERP (*Enterprise Resource Planning*), como a Hubspot CRM, Agendor, Pipedrive, e demais empresas de consultoria em tecnologia, desenvolvimento de softwares de gestão de vendas existe uma forte concorrência.

A empresa também compete com outras empresas que utilizam produtos da Salesforce para desenvolvimento de produtos, por oferecer o mesmo tipo de serviço especializado.

A própria Salesforce disponibiliza seu serviço para ser contratado de forma direta, sem passar pela Everymind, aumentando a chance de competição por se tornar uma alternativa para o desenvolvimento de sistemas de gestão.

4.1.1.2 Compradores

Os clientes, do setor de engenharia e construção, indústrias de bens de consumo, e serviços financeiros da empresa possuem uma variedade de possibilidades de escolha, de produtos mais baratos de sistemas de gerenciamento de vendas e gestão de clientes.

Pelo surgimento de produtos e serviços novos, os próprios clientes podem exigir o uso de alguma ferramenta em que a empresa não seja especializada, levando a uma curva de aprendizado para a empresa treinar e adaptar os seus funcionários.

4.1.1.3 Fornecedores

O fornecedor Salesforce pode decidir mudar ou encerrar sua parceria com a empresa, tirando uma vantagem competitiva em relação ao diferencial na especialização da empresa sobre a concorrência.

Pela valorização dos funcionários no mercado de trabalho atual, existe uma rotatividade grande de colaboradores na empresa, que possuem uma variedade de possibilidades de escolha de vagas de trabalho.

Funcionários novos requerem treinamento e adaptação até conseguir entender as ferramentas, principalmente as de Salesforce, utilizadas para os projetos, levando a uma curva de aprendizado até atingir um desempenho e performance, necessários para manter a qualidade dos serviços prestados.

4.1.1.4 Novos Entrantes

Com o desenvolvimento de novos sistemas e formas de implementação, novas empresas começam a entrar no setor de desenvolvimento de CRM e ERP, competindo com os produtos da Salesforce, onde a empresa é especializada.

Outro ponto é a facilitação no desenvolvimento desses sistemas por empresas das ferramentas e produtos, possibilitando novas empresas entrantes e o surgimento de novos tipos de profissionais que podem trabalhar com esse setor, como profissionais *freelancer,* que podem realizar consultorias em TI e desenvolvimento de sistemas de gerenciamento, bem como a terceirização de alguns setores da área de TI, que podem começar a atuar com esse setor em especial.

4.1.1.5 Substitutos

Clientes podem decidir executar as soluções e produtos por conta própria, com um setor dedicado à especialização em Salesforce, CRM e na implementação e gestão de projetos.

4.1.2. Análise SWOT

SWOT é a sigla em inglês para Forças (*Strengths*), Fraquezas (*Weaknesses*), Oportunidades (*Opportunities*) e Ameaças (*Threats*). A utilização de uma matriz SWOT é importante para considerar elementos internos e externos de uma empresa, e a posicionar de forma efetiva sobre os seus objetivos. Forças e fraquezas, que interferem diretamente no quão preparada a empresa está para atingir suas metas. Oportunidades e ameaças pontuam potenciais focos de atenção a partir do ambiente externo da empresa que podem interferir no negócio (Benzaghta et al., 2021).

Figura 2. Imagem da análise SWOT; Apresentação de forças, fraquezas, oportunidades e ameaças do cliente (Everymind - a compass.uol company).

A partir da matriz criada, torna-se possível avaliar alguns pontos vantajosos para a empresa, e pontos que necessitam ser melhor desenvolvidos e que precisam de atenção.

A Everymind possui uma vantagem pela especialização em ferramentas específicas, bem avaliadas, que são recorrentemente utilizadas para o desenvolvimento e implementação de projetos de sistemas de CRM e ERP. Cruzando este ponto com o aumento das empresas para a digitalização dos processos de gestão, traz uma oportunidade que a empresa pode usufruir. Porém com a rotatividade e falta de profissionais de tecnologia no país, a empresa

pode não conseguir atender a demanda que o mercado requer para prestar serviços com uma qualidade suficiente.

Portanto, um dos pontos a desenvolver é a questão da rotatividade e retenção de talentos, que é inclusive um dos pontos abordados por este trabalho em questão, para o desenvolvimento de um modelo de predição que auxilie a empresa a desenvolver estratégias de gestão de pessoas.

Cruzando os pontos que precisam ser desenvolvidos com as oportunidades, é possível analisar os passos importantes para o planejamento que a empresa pode desenvolver, e então se preparar para conseguir aproveitar a demanda que o mercado necessita, de adaptação tecnológica dos seus sistemas internos.

A partir dos pontos fortes, se cruzarmos com as ameaças, podemos também avaliar um risco que a empresa precisa mapear, para não se prejudicar, como os funcionários que foram treinados e são especializados em ferramentas de Salesforce, que se tornam um grande atrativo para empresas que querem contratar estes profissionais, levando em conta também o momento de alta competição entre as empresas como prestadoras de serviço e contratação e retenção de talentos.

Ainda cruzando os pontos fortes da empresa com as ameaças levantadas pela matriz S.W.O.T., é importante dizer que como um dos pontos fortes da empresa é ir desde o planejamento e desenvolvimento dos projetos até a sua implementação, a alta rotatividade dos profissionais de TI pode prejudicar o andamento dos projetos dentro da empresa, sendo importante o mapeamento deste risco e elaboração de uma estratégia para minimizar os impactos negativos.

4.1.3. Planejamento Geral da Solução

4.1.3.1. Os dados disponíveis foram dados em formato de tabelas, e incluem:

- Identificação dos candidatos; (Número, Data de nascimento, estado civil, etnia e gênero)
- Crescimento financeiro; (salário, cargo, promoção, data de admissão, Mudança de função, novo salário)
- Demissões; (data de admissão, data de saída, tipo de saída)

4.1.3.2. Solução Proposta:

Criar um sistema baseado em inteligência artificial. Que possui a capacidade de fazer relações entre os dados propostos, e prever o comportamento e intenções dos colaboradores..

4.1.3.3. tipo de tarefa (classificação):

O sistema realizará uma tarefa de classificação. Ao determinar quais funcionários são mais prováveis de se demitir, a inteligência classifica os mesmos.

4.1.3.4. Como a solução proposta deverá ser utilizada:

Ela será utilizada como uma ferramenta de gestão para facilitar a função de recursos humanos.

4.1.3.5. Quais os benefícios trazidos pela solução proposta:

A solução proposta facilita e amplifica o desempenho do RH. Criando uma visão objetiva, e mais precisa, para que decisões sejam tomadas.

4.1.3.6. Qual será o critério de sucesso e qual medida será utilizada para o avaliar:

O critério de sucesso será conseguir analisar os dados fornecidos e que o resultado seja o mais perto daquilo que tenha acontecido no passado. A medida para avaliar será em forma de gráficos e uma nota para os funcionários que indique qual a probabilidade dele sair.

4.1.4. *Value Proposition Canvas*

Value Proposition Canvas, do inglês, Canvas de proposta de valor, indica como uma empresa se diferencia para que o cliente a escolha e como ela pode beneficiá-lo. Visa entender os anseios do cliente com a empresa e como a mesma pode suprir tais anseios e resolver as dores de seus clientes (Pokorná, 2015).

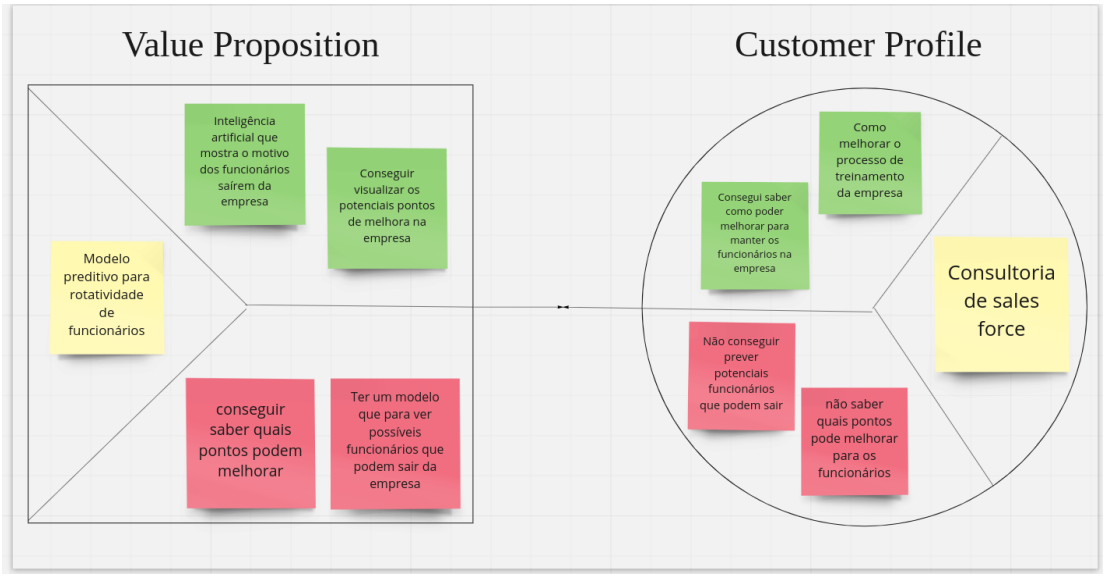


Figura 3. Imagem do Value Proposition; Apresentação do que cliente (Everymind - a compass.uol company) oferece para sanar a dor de seus clientes e o que os mesmos anseiam.

4.1.5. Matriz de Riscos

A partir de ameaças de fatores internos e externos, as empresas precisam mensurar esse risco e controlar estes elementos, já que podem trazer consequências negativas e assim prejudicar as metas da empresa. Através da matriz de riscos, é possível agregar no planejamento estratégico os elementos que irão exigir medidas preventivas e corretivas, que diminuam efeitos e reduzem incertezas (Alves e Tessmann, 2018).

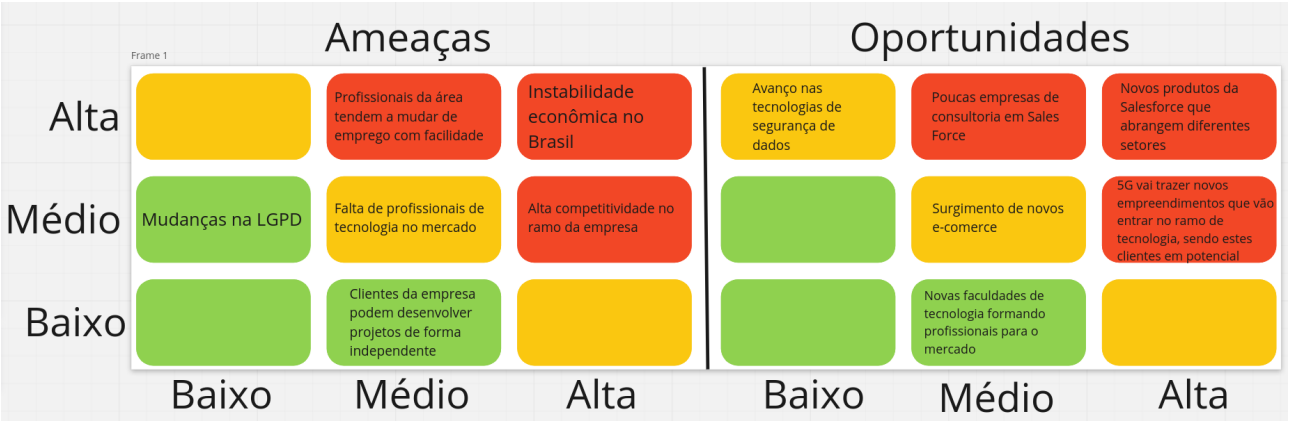


Figura 4. Imagem da Matriz Riscos; Apresentação dos principais riscos que envolvem o cliente (Everymind - a compass.uol company).

Seguindo a matriz de riscos desenvolvida, é perceptível a importância de se mapear os riscos e elaborar estratégias para mitigar os impactos negativos. Inclusive em relação às oportunidades, que a depender das estratégias adotadas, as oportunidades podem favorecer concorrentes e beneficiar outras empresas.

Os itens citados na Matriz de Riscos como ameaça são:

- Profissionais da área tendem a mudar de emprego com facilidade;
- Instabilidade econômica do Brasil;
- Mudanças na LGPD;
- Falta de profissionais de tecnologia no mercado;
- Alta competitividade no ramo da empresa;
- Clientes da empresa podem desenvolver projetos de forma independente.

E os itens citados na Matriz de Riscos como oportunidades são:

- Avanço nas tecnologias de segurança de dados;
- Poucas empresas de consultoria em Salesforce;
- Novos produtos da Salesforce abrangem diferentes setores;
- Surgimento de novos *e-commerce*;
- 5G vai trazer novos empreendimentos que vão entrar no ramo de tecnologia, sendo estes, clientes em potencial;
- Novas faculdades de tecnologia formando profissionais para o mercado.

É importante analisar a chance e o impacto, para que as estratégias possam ser melhor implementadas, em relação à frequência do item avaliado na ferramenta.

Uma análise interessante a ser feita em relação ao contexto estudado, é no aumento da busca das empresas na digitalização de processos de gestão e sistemas internos, que é uma oportunidade, mas pela alta demanda, favorece também novas empresas concorrentes que atuam no setor. Adicionalmente, pela falta de profissionais disponíveis, e pela alta rotatividade, através dessa oportunidade, pode agravar ainda mais a retenção de talentos e profissionais dentro da empresa que é objeto deste trabalho.

Isso se soma aos avanços tecnológicos, que tendem a aumentar ainda mais o volume de empresas, tanto que vão necessitar da integração desses sistemas, como de novos concorrentes.

4.1.6. Personas

Persona é a utilização de modelos para representação ideal de um cliente, ou usuário, de um produto. Esta ferramenta é utilizada para auxiliar a equipe de desenvolvimento a entender melhor a dor do usuário do produto e a elaborar a melhor solução para isso (Tu, 2010).

Para este trabalho foram elaboradas duas personas, para representar dois grupos importantes entre os colaboradores da empresa. A primeira persona representa um desenvolvedor, que trabalha em um dos grupos da empresa, e a segunda representa um Squad Leader, líder de uma das equipes da empresa, que é o usuário do produto deste trabalho.

A seguir cada persona será descrita mais detalhadamente:

Persona 1 - Raphael Cvaigman, Dev Junior

Raphael Cvaigman, 26 - Dev Jr

Passado

- Formado em ciências de computação na PUC
- Recém-casado
- Segunda empresa que esta trabalhando
- Curso técnico de análise de dados na Alemanha

Demografia

- Sexo Masculino
- Judeu
- Salario - 2000 Mensal
- Casado

Características

Calmo	Intelectual
Ágil	inteligente

Motivacoes

- Atingir liberdade financeira
- se sentir realizado no trabalho
- melhorar sua qualidade de vida
- Deseja resolver suas dores na empresa

Frase

"Faca o bem que voce gostaria de receber"

Figura 5. Imagem com informações referentes à persona número 1 desenvolvida para este projeto.

Persona 2 ,Leonardo Santorini, Squad Leade.

Principal usuário do nosso produto, a persona ajuda o time a se guiar para entender a importância de quem vai utilizar o produto.

Figura 6. Imagem com informações referentes à persona número 2 desenvolvida para este projeto.

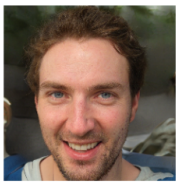
4.1.7. Jornadas do Usuário

Representação gráfica de etapas sobre o relacionamento do cliente com um produto ou serviço de determinada empresa. São descritos os passos que o consumidor toma antes, durante e depois da compra.

A importância da utilização desta ferramenta é pelo aspecto dinâmico que a jornada do usuário tem, de imbuir interações do usuário ao longo de um tempo, revelar pontos sensíveis das personas, e combinar o comportamento do usuário com uma storyline, de forma a entender melhor essa interação do usuário com o problema, ou com alguma situação (Aliari, 2018).

Foram elaboradas duas jornadas, uma para cada persona, a fim de contextualizar os eventos que se sucederam no cargo que desempenham, desde a entrada da persona na empresa

Jornada de usuário do Analista de negócio



Raphael Cvaigman

Cenário: Rafael acabou de transferir de empresa e está com dificuldades de se adaptar na nova empresa, além de sentir que seu trabalho não sendo reconhecido

Expectativas

Espera se acostumar com a empresa e que receba o reconhecimento que deseja , pois nao deseja sair da empresa

FASE 1 Raphael entra na empresa	FASE 2 Raphael está com dificuldades	FASE 3 Raphael está em busca de outra empresa	FASE 4 Solução
Raphael acaba de entrar na empresa nova como analista de negócio e está com muitas expectativas boas para essa nova fase de sua vida, porem ainda tem que se adaptar a nova metodologia da empresa e como as coisas funcionam	**Raphael não esta conseguindo se adaptar a empresa por questões metodológicas e por conta da alta rotatividade de pessoas ele esta com muita dificuldade de fazer amizades e trabalhar em equipe, o que está afetando a sua produtividade e satisfação na empresa. Ele entrou em contato com seu squad lider, porém não sentiu que fez diferença.**	**Raphael não esta contente, pois não se sentiu reconhecido, então ele foi reclamar para o seu chefe. Seu squad lider com o RH avaliou o trabalho dele durante seu tempo na empresa ao usar o modelo preditivo descobriram que ele estava propenso a sair, depois de toda análise, concluíram que o Raphael não merecia o reconhecimento que buscava.**	**Após sua insatisfação com a resposta de seu chefe e depois de muita procura de novas oportunidades de trabalho, Raphael encontra uma nova empresa, assim se demitindo.**

Oportunidades

Diante do mercado aquecido, muitos funcionários acabam superestimando seu trabalho, exigindo, às vezes, mais reconhecimento pelo seu desempenho na entrega.

Responsabilidades

Contratação de funcionários é algo que deve haver muita responsabilidade considerando que
o mercado de dev está muito valorizado, assim escolher o funcionário certo é essencial.

Figura 7. Imagem com informações referentes à jornada do usuário da persona número 1 desenvolvida para este projeto.

Jornada do usuário do Squad Leader

Leonardo Santorini

Cenário: Leonardo é squad lider de um time, porém está insatisfeito com o desempenho de dois membros da sua equipe

Expectativas

Leonardo tem altas expectativas para sua equipe, pois acredita que eles têm um grande potencial e espera que consiga resolver os problemas internos que está tendo com ajuda do modelo preditivo

FASE 1 Leonardo usa o modelo preditivo para ver os possíveis comportamentos dos funcionários	FASE 2 Leonardo conversa com os funcionarios	FASE 3 Leonardo analisa o que ira fazer em relacao ao problema	FASE 4 Leonardo toma uma decisão
Leonardo utiliza o modelo preditivo e um dos membros é classificado com 'tendencia a sair' e o outro não.	**Para tentar resolver a situação, Leonardo chama cada funcionário para conversar e entender os problemas que eles estão enfrentando. Um dos funcionários está com um desempenho ruim porque não está lidando muito bem com os membros da equipe e o outro funcionário não está sentindo que esta recebendo o reconhecimento que merece.**	**Após analisar a situação do primeiro funcionário Leonardo chegou a conclusão de que ele não estava com um bom desempenho por conta da equipe que estava, após analisar o segundo funcionário percebeu que ele estava entregando mais do que seu cargo exigia e que merecia uma promoção.**	**Após analisar os dois funcionários ele decidiu que o primeiro funcionário iria trocar de time para tentar melhorar seu rendimento e a respeito do segundo funcionário ele deu uma promoção para que ele se sentisse mais reconhecido e continuasse seu bom trabalho na empresa.**

Oportunidades

Diante do mercado aquecido, muitos funcionários acabam superestimando seu trabalho, exigindo, às vezes, mais reconhecimento pelo seu desempenho na entrega.

Responsabilidades

Contratação de funcionários é algo que deve haver muita responsabilidade considerando que
o mercado de dev está muito valorizado, assim escolher o funcionário certo é essencial.

Figura 8. Imagem com informações referentes à jornada do usuário da persona número 2 desenvolvida para este projeto.

4.2. Compreensão dos Dados

4.2.1 Descrição dos dados:

Os dados enviados pelo cliente (*Everymind - a compass.uol company)* consistem em um documento do Google Sheets em formato *.XLSX* contendo um total de três tabelas.

Tabela 1 – Everymind: Tabela que descreve a lista de funcionários da empresa, com seus respectivos salários, datas de ingresso e saída da empresa, motivo de saída e seus respectivos cargos. A referida tabela possui dados de admissão e demissão que vão do dia 1 de fevereiro de 2006 até o dia 27 de julho de 2022. A tabela tem 285 linhas e 17 colunas.

Campos:

- **Matrícula -** Identificação (ID) de colaborador
- **Nome Completo** - Nomes não revelados, identificação numérica de "Pessoa Colaboradora"
- **Dt Admissão** - Data na qual o colaborador foi admitido na empresa
- **Dt Said** - Data na qual o colaborador saiu da empresa
- **Tipo Saída -** Motivo da saída do colaborador
- **Cargo -** Cargo exercido pelo colaborador
- **Salário Mês** - Salário recebido pelo colaborador por mês
- **Dt Nascimento** - Data de nascimento do colaborador
- **Etnia** - Etnia do colaborador
- **Estado Civil -** Estado civil do colaborador
- **Escolaridade** - Nível de escolaridade do colaborador
- **Estado -** Estado em de onde o colaborador trabalha/trabalhava
- **Cidade** - Cidade em de onde o colaborador trabalha/trabalhava
- **Área** - Área de atuação do colaborador na empresa
- **Idade** - Idade do colaborador

Tabela 2 - Reconhecimento: Tabela que descreve a lista de funcionários da empresa que receberam promoção, com seus novos salários, datas de ingresso e data vigente à promoção e novo cargo. A referida tabela possui dados de admissão e demissão que vão do dia 1 de fevereiro de 2006 até o dia 1 de julho de 2022. A tabela tem 340 linhas e 10 colunas.

Campos:

- **Matrícula -** Identificação (ID) de colaborador
- **Codinome -** Nomes não revelados, identificação numérica de "Pessoa Colaboradora"
- **Situação -** Situação vigente do colaborador na empresa
- **Dt Admissão -** Data na qual o colaborador foi admitido na empresa
- **Dt Vigência -** Data na qual o colaborador recebeu sua respectiva promoção
- **Novo Cargo -** Novo cargo exercido pelo colaborador após a promoção
- **Novo Salário -** Novo salário recebido pelo colaborador por mês após a promoção
- **Motivo -** Motivo da promoção do colaborador
- **Alterou Função -** Indica se o colaborador alterou ou não de função

Tabela 3 - Ambiente de Trabalho 27.07:Tabela de pesquisa de satisfação e avaliação do ambiente de trabalho da empresa pelos colaboradores. As respostas são anônimas, impossibilitando o cruzamento de dados com as demais tabelas. A tabela tem 1695 linhas e 13 colunas.

Campos:

- **Divisão -** Respectiva divisão da empresa analisada na pesquisa
- **Pilar -** Campo ao qual a pontuação da pesquisa se refere
- **Pontuação¹ -** Pontuação do respectivo pilar
- **Fator -** Atributo ao qual se refere a respectiva pergunta
- **Pontuação² -** Pontuação do respectivo fator
- **Pergunta -** Respectiva pergunta feita em referente fator
- **Pulou -** Porcentagem de pessoas que pulou a pergunta
- **Muito Insatisfeito -** Porcentagem de pessoas que responderam com "muito insatisfeito"
- **Insatisfeito -** Porcentagem de pessoas que responderam com "insatisfeito"
- **Neutro -** Porcentagem de pessoas que responderam com "neutro"
- **Satisfeito -** Porcentagem de pessoas que responderam com "satisfeito"

- **Muito Satisfeito -** Porcentagem de pessoas que responderam com “muito satisfeito”
- **Taxa de Confiabilidade -** Quantidade respostas obtidas para cada pergunta

Mesmo que estejam em um formato simples, os dados estão desorganizados, os “colaboradores” estão fora de ordem e existem duas colunas em branco.

Subconjuntos a serem trabalhados:

- Gênero (Masculino/Feminino)
- Grupos Etários (18 - 25/ 26 - 35+)
- Estado Civil (Casado/Solteiro)
- Média salarial (500 - 1500/ 1501 - 2500/ 2501 - 3000+)
- Permanência (Anos Totais/ Salário)
- Incentivo (Anos Totais / Cargo / Novo Salário)

Restrições de segurança: Estes dados não possuem identificação direta dos candidatos, sendo utilizado números para identificação, mantendo o anonimato dos dados dos colaboradores. Porém, As tabelas possuem dados sensíveis, como faixa salarial da empresa, padrões de promoções de funcionários, entre outros.

4.2.2 Descrição estatística básica dos dados:

Com a primeira análise dos dados conseguimos identificar algumas inconsistências e alguns pequenos padrões que para serem confirmados necessitam de mais dados.

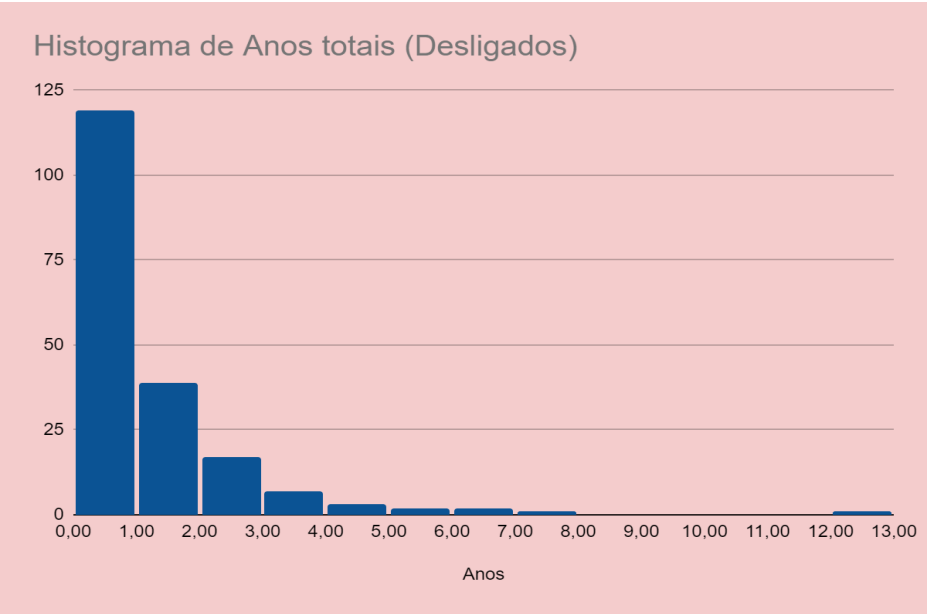


Gráfico 1. Mostrando a quantidade de funcionários que saíram em relação ao tempo de empresa. É possível ver que existe uma saída grande em menos de 1 ano de empresa

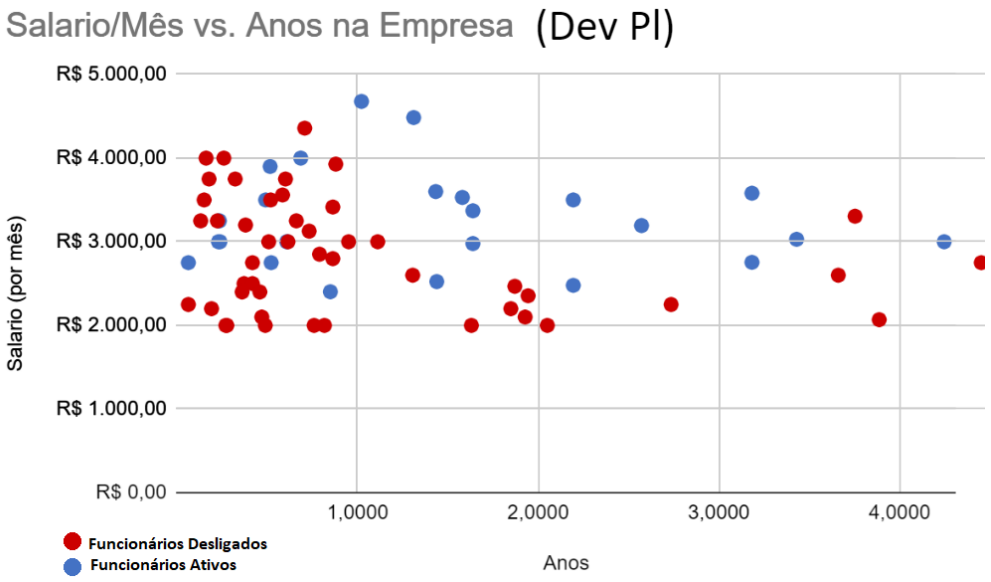


Gráfico 2. Em dispersão, mostrando os funcionários no cargo de Desenvolvedor Pleno, com a relação de salário e os anos na empresa. Através do gráfico é possível visualizar que o primeiro ano é um ano crítico para os funcionários, e após um ano de empresa começa a se formar um padrão de tendência da permanência na empresa.

4.2.3 Descrição da predição desejada:

O tipo de modelagem preditiva utilizada neste projeto é o de classificação binária, que consiste em classificar os elementos do conjunto de dados em dois tipos de classes, como "1" e "0".

Após essa classificação, os resultados podem ser avaliados de quatro formas:

- "1" pode ser classificado corretamente como o valor "1", sendo este o verdadeiro positivo (VP);
- "0" classificado erroneamente como "1", tornando-se um falso positivo (FP);
- "0" pode ser classificado corretamente como o valor "0", um verdadeiro negativo (VN);
- "1" classificado erroneamente como "0", sendo agora um valor falso negativo (FN);

4.3. Preparação dos Dados

4.3.1. Tabela “Reconhecimento”

Na tabela de Reconhecimento, existem muitos dados que indicam alterações nas colunas “Salario Mês”, repetindo valores de linhas, nas colunas de “Matricula” e “Codinome”, por se tratarem da mesma pessoa. Alguns dados também se repetem com dados presentes na tabela “Everymind”. Como esta tabela será integrada na tabela principal “Everymind”, os dados precisaram ser tratados. Os métodos de tratamento serão descritos a seguir:

- Remoção de todos os espaços, caracteres especiais e todos os acentos;
- Remoção de colunas desnecessárias, tais como a de “Pergunta”, “Fator”, “pontuação.1”(...);
- Cálculo da quantidade de tempo em dias da data de admissão até a data vigente da promoção;
- e alocação de tais valores em novas colunas;
- Remoção das colunas “Novo_Cargo”, “Novo_Salario”, ”Motivo” e “Alterou_Funcao”;
- Remoção de linhas duplicadas referente à coluna “Codinomes”;
- Criação de novas colunas para cada promoção e alocação de valores referentes à pessoa colaboradora;
- Cálculo da média de tempo de todos os funcionários para a primeira promoção;
- Criação de uma coluna indicando o número de promoções recebidas pela pessoa colaboradora;

4.3.1.1. Limpeza dos dados

O primeiro tratamento feito foi o de eliminar espaços nos valores dos campos, para normatizar e padronizar. Através de um código de substituição de caracteres, que faz uma varredura em todas as colunas da tabela, de forma automática.

Utilizando um código de substituição também foi tratado os dados de funcionários afastados, que ainda são considerados funcionários ativos, na coluna "Situação Atual”, para manter os valores de forma binária (Funcionários ativos e desligados).

Alguns valores do eixo “x” da tabela, referente a funcionários da tabela “Reconhecimento” não estão presentes na tabela “Everymind”, inviabilizando o uso desses valores, por apresentarem uma perda grande de dados, sendo assim limpos da tabela “Reconhecimento”.

4.3.1.2. Adição e Derivação de dados

A partir dos campos "Data de Admissão" e "Data Vigência", foram utilizados os valores para gerar uma coluna com valores referentes ao tempo, em meses, a partir da diferença entre os valores de datas.

4.3.1.3. Remoção de dados

Por existirem dados iguais entre as tabelas "Everymind" e "Reconhecimento", foi decidido apagar estes dados, para não provocar duplicidade de dados ao serem integrados na tabela "Everymind"

4.3.1.4. Análise prévia dos dados

A partir dos tratamentos é possível perceber uma proporção maior de funcionários ativos, em relação à quantidade de funcionários desligados que receberam uma atualização em seu salário e no cargo que atua por promoção, ou por mérito.

Também foi percebido que funcionários que recebem uma promoção tendem a receber outras atualizações em seu salário e na função que desempenham ao longo de sua carreira profissional.

Através de uma análise prévia dos dados presentes entre as tabelas "Everymind" e "Reconhecimento", foi notado que a proporção entre colaboradores que receberam algum tipo de alteração em seu salário e função através de uma promoção ou por mérito é menor do que a quantidade de funcionários desligados e ativos em geral.

É importante ressaltar que para uma análise mais profunda, e para validar as hipóteses levantadas por estas análises prévias, é importante utilizar um motor preditivo, e utilizar as variáveis indicadas para atestar através da taxa de erro se é uma hipótese válida ou não.

4.3.2. Tabela "Everymind"

Na tabela "Everymind" existem diferentes tipos de dados, que podem ser enriquecidos com variáveis da tabela "Reconhecimento". Mas antes foi feito o tratamento dos dados, que serão descritos a seguir:

- Retira os espaços e os acentos das letras, assim como substituir o "Ç" por "C".
- Na coluna nome completo substituímos "pessoa colaboradora" por "PC".
- Criação da coluna Género onde "1" representa o "masculino" e o "0" o "feminino".

- Criação da coluna “Estado Civil” onde cada número representa um estado, solteiro=1, casado=2, divorciado=3, União Estável=4 e separado=5.
- A partir da data do dia de hoje , padroniza as datas da coluna de data de admissão
- Cria uma nova coluna que calcula o período que o funcionário está na empresa
- Criação da Coluna Situação que indica se o funcionário ainda está ativo na empresa ou não, 1 ele está ativo e 0 ele não está mais na empresa.
- Modificação da coluna cargo onde todos os cargos agora são representados por um número.
- Modificação da coluna escolaridade onde todos os tipos de escolaridade agora são representados por um número.
- As colunas Matrícula, Estado, Cidade, dt_nascimento, dt saída e etnia foram removidas.

4.3.2.1. Limpeza dos dados

Para deixar a tabela limpa e clara, eliminamos espaços nos valores dos campos, para normatizar e padronizar, em todas as colunas da tabela, além disso para que não tenha problemas com as fórmulas. Ao procurar por um texto específico ele não será encontrado por conta desses espaços.

Simplificamos “PessoaColaboradora” para “PC” remove excesso de letras , e deixe a tabela e deixe o uso mais rápido e eficaz. Com essas mudanças a tabela fica clara, limpa e objetiva.

4.3.2.2. Adição e Derivação de dados

Transformação de dados, uma “subtração” entre as “Datas de Admissão” e “Datas de Saída”. Levando a criação de uma nova coluna com o tempo total de trabalho de um dos colaboradores, assim facilitando uma visão superficial , ajudando o algoritmo e ganhando tempo ao invés de calcular a diferença manualmente.

Também adicionamos a coluna de situação que indica se o funcionário ainda está na empresa ou não, essa é a variável que utilizamos como o nosso Y no modelo, ela é variável que estamos tentando prever.

A partir da proporção de dados entre os valores do *target* do projeto, foi realizado um *over sampling*, com a geração automática e *randômica* de valores, para aumentar o equivalente a proporção entre os resultados.

Por conta da alta variação entre valores mínimos e máximos dos valores das variáveis, foi estabelecido uma normalização dos dados utilizando o algoritmo MinMax Scaler, nas seguintes

variáveis: 'NumeroMeses', 'Promocao_1', 'Promocao_2', 'Promocao_3', 'Promocao_4', 'Promocao_5', 'Promocao_6', 'Numero_Promocoes', 'Tempo_Medio_Promocoes'.

Isso evita que o modelo apresente vícios por padrões pela alta variação entre os valores presentes no banco de dados fornecido.

4.3.2.3. Análise prévia dos dados

Após o tratamento dos dados, conseguimos observar previamente uma quantidade elevada de funcionários desligados da empresa, com um tempo de permanência menor que um ano de empresa.

O cargo do funcionário é um elemento crucial na pesquisa, mostrando que alguns apresentam mais problemas que outros. Entretanto possuem informações pouco relevantes, como a do estado civil, que torna impossível inferir algo pois é difícil diferenciar um divorciado sem filho e um solteiro, são muito similares neste aspecto, assim como um casado e um solteiro com filhos. Um fator que resolveria nosso problema seria informar se o funcionário possui filhos ou não , porém não possuímos esse dado.

Vale dizer que para uma análise mais profunda, é importante criar gráficos com as variáveis diversas da tabela a fim de decidir variáveis importantes para o motor do modelo preditivo do projeto, e utilizar as variáveis indicadas para atestar através da taxa de erro se é uma hipótese válida ou não.

4.3.3. Tabela “Ambiente de Trabalho”

A partir da análise feita nesta tabela, os dados presentes não se relacionam diretamente com os dados da tabela principal, dificultando a aplicação e integração desta tabela, mesmo com tratamento dos dados, resultando em uma perda significativa de dados, trazendo uma considerável quantidade de ruídos.

Foram feitos os seguintes tratamentos na tabela:

- Remoção de todos os espaços, caracteres especiais e todos os acentos;
- Remoção de colunas desnecessárias, tais como a de “Pergunta”, “Fator”, “pontuação.1”(...);
- Remoção de linhas com taxa de confiabilidade “Baixa” e “Muito Baixa”;
- Fusão de linhas de “Divisões” e “Pilares” referentes;
- Criação de novas colunas para cada item da coluna “Pilar”;
- Separação dos valores de pesquisas de cada linha para sua coluna de pilar referente;

4.4. Modelagem

Um modelo preditivo é constituído de funções matemáticas e algoritmos, e é utilizado para encontrar padrões, analisar tendências e calcular probabilidades a partir de dados de um banco de dados.

Atualmente existem diversos tipos de modelos preditivos, e a seguir serão descritos os tipos de modelagem experimentais utilizadas neste projeto para treino e teste das variáveis selecionadas:

4.4.1. Modelo SVM - Support Vector Machines

Este tipo de modelo preditivo supervisionado é muito utilizado em modelos de classificação, o SVM analisa os dados, e a partir de elementos críticos, encontra a melhor forma de divisão entre duas classes, classificando a partir do padrão encontrado.

Ideal para bases de dados menores, pelo tempo que leva para analisar os dados para o treinamento do modelo, o SVM é bastante preciso, e se adapta bem em diversos cenários. Porém não é tão efetivo em casos com conjuntos de dados ruidosos onde as classes ficam sobrepostas (Bambrick, 2016).

O SVM cria um limite de decisão entre duas classes no modelo de classificação que rotula a previsão de um ou mais vetores de características. O limite, conhecido como hiperplano, dispõe de uma orientação que calcula a maior distância entre os pontos os pontos mais próximos dos dados de cada uma das duas classes. Estes pontos são chamados de vetores de suporte (Huang, 2018).

A seguir é possível ver a fórmula que o modelo SVM utiliza:

$$(x_1, y_1), \ldots, (x_n, y_n), x_i \in R^d \text{ and } y_i \in (-1, +1)$$

Figura 9. Fórmula utilizada pela modelagem de SVM.

Onde o elemento x_i representa o vetor de características, y_i representa o rótulo de classe, podendo ser negativo ou positivo, de um composto de treinamento i.

O valor do hiperplano ótimo pode ser definido com a seguinte fórmula:

$$wx^T + b=0$$

Figura 10. Fórmula para definição de um hiperplano pelo modelo SVM.

O elemento w é o vetor de peso, x é o vetor de *input feature*, e b é a polarização.

Os elementos w e b podem satisfazer desigualdades de todos os elementos do conjunto de treino do modelo da fórmula a seguir:

$$wx_i^T + b \geq +1 \quad \text{if} \quad y_i=1$$
$$wx_i^T + b \leq -1 \quad \text{if} \quad y_i=-1$$

Figura 11. Fórmula para identificar as desigualdades entre os valores utilizados pela modelagem SVM.

Sendo assim, o objetivo do treino de um modelo SVM é achar os elementos w e b e criar o hiperplano que divide os dados e maximiza a margem entre as duas classes do modelo de classificação utilizado neste projeto.

4.4.2. Modelo Naïve Bayes

Este tipo de modelo preditivo sucede outro importante modelo de classificação em aprendizado supervisionado, a árvore de decisão é um dos modelos mais populares no aprendizado de máquina.

Sendo um modelo adequado para classificação de atributos discretos, o Naïve Bayes tem aplicações na análise de crédito, diagnósticos médicos ou busca por falhas em sistemas mecânicos. Tomando como premissa a suposição de independência entre as variáveis do problema, o modelo de Naïve Bayes realiza uma classificação probabilística de observações, caracterizando-as em classes pré-definidas.

A probabilidade condicional pode ser calculada usando a probabilidade conjunta, embora seja intratável. O Teorema de Bayes fornece uma maneira baseada em princípios para calcular a probabilidade condicional.

A forma simples de cálculo para o Teorema de Bayes é a seguinte:

- P(AIB) = P(BIA) * P(A) / P(B)

Onde a probabilidade de que estamos interessados em calcular P(AIB) é chamada de probabilidade posterior e a probabilidade marginal do evento P(A) é chamada de anterior.

Podemos enquadrar a classificação como um problema de classificação condicional com o Teorema de Bayes da seguinte forma:

- P(yi I x1, x2, ..., xn) = P(x1, x2, ..., xn I yi) * P(yi) / P(x1, x2, ..., xn)

O P(yi) anterior é fácil de estimar a partir de um conjunto de dados, mas a probabilidade condicional da observação baseada na classe P(x1, x2, ..., xn I yi) não é viável a menos que o número de exemplos seja extraordinariamente grande, por exemplo. grande o suficiente para estimar efetivamente a distribuição de probabilidade para todas as diferentes combinações possíveis de valores.

4.4.3. Modelo KNN - K-Nearest Neighbour

KNN(K — Nearest Neighbors) é um algoritmo de Machine Learning supervisionado. usado no campo de data mining e machine learning, ele é um classificador onde o aprendizado é baseado "no quão similar" é um dado (um vetor) do outro. O treinamento é formado por vetores de n dimensões.

As etapas de um algoritmo KNN são:

1. Recebe um dado não classificado;
2. Mede a distância (Euclidiana, Manhattan, Minkowski ou Ponderada) do novo; dado com todos os outros dados que já estão classificados;
3. Obtém as X(no caso essa variável X é o parâmetro K) menores distâncias;
4. Verifica a classe de cada da um dos dados que tiveram a menor distância e conta a quantidade de cada classe que aparece;
5. Toma como resultado a classe que mais apareceu dentre os dados que tiveram as menores distâncias;
6. Classifica o novo dado com a classe tomada como resultado da classificação

Com a fórmula a seguir será verificado a distância entre 1 ponto(sua amostra não classificada) e 1 outro ponto do seu dataset(1 outro dado já classificado) para então ver a similaridade dos dois, quanto menor é o resultado dessa fórmula, maior é a similaridade entre esses dois dados.

Distância entre duas instâncias $\mathbf{p}_i$ e $\mathbf{p}_j$ definida como:

$$d = \sqrt{\sum_{k=1}^{n} (p_{ik} - p_{jk})^2}$$

$\mathbf{p}_{ik}$ e $\mathbf{p}_{jk}$ para $k = 1, \ldots, \mathbf{n}$ são os $\mathbf{n}$ atributos que descrevem as instâncias $\mathbf{p}_i$ e $\mathbf{p}_j$, respectivamente

Figura 12. Representação da fórmula utilizada pelo algoritmo da modelagem KNN

4.4.4. Modelo AdaBoost - (Adaptive Boosting)

AdaBoost é um dos métodos mais populares de Boosting, ele foi um dos primeiros modelos de boosting desenvolvidos, ele se adapta e tenta se autocorrigir a cada iteração do processo de boosting ele inicialmente dá o mesmo peso para cada conjunto de dados. Em seguida, ajusta automaticamente os pesos dos pontos de dados após cada árvore de decisão. Ele dá mais peso aos itens classificados incorretamente, para corrigi-los para a próxima rodada. Ele repete o processo até que o erro residual, ou a diferença entre os valores reais e previstos.

As principais vantagens de se utilizar o modelo AdaBoost são:

- Lida muito bem com tipos de dados diversos
- Pode ser utilizado tanto para classificações quanto para regressões
- Tem uma precisão bem alta

Algumas desvantagens são:

- Tem o possível risco de overfiting
- Não funciona muito bem quando há correlação entre os recursos dos dados
- Caso tenha que ajustar os hiperparâmetros pode ser um processo bem demorado

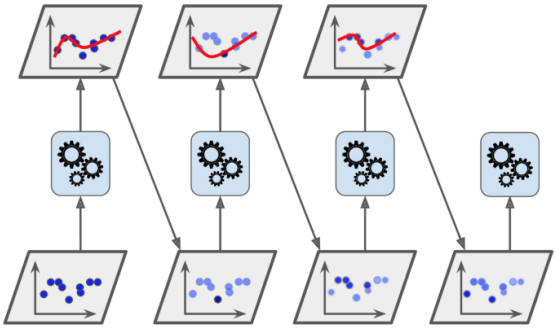


Figura 13. A figura acima mostra como o ilustra como o modelo AdaBoost funciona, onde em cada etapa o modelo aprende e ensina a próxima etapa o seu conhecimento anterior, resultando em um ótimo resultado.

4.4.5. Decision Tree

As Árvores de Decisão são um modelo preditivo supervisionado que não possui parâmetros, usado para classificação e regressão. Ele tem como objetivo criar um modelo que preveja o valor de uma variável de destino aprendendo regras de decisão simples inferidas a partir dos recursos de dados.

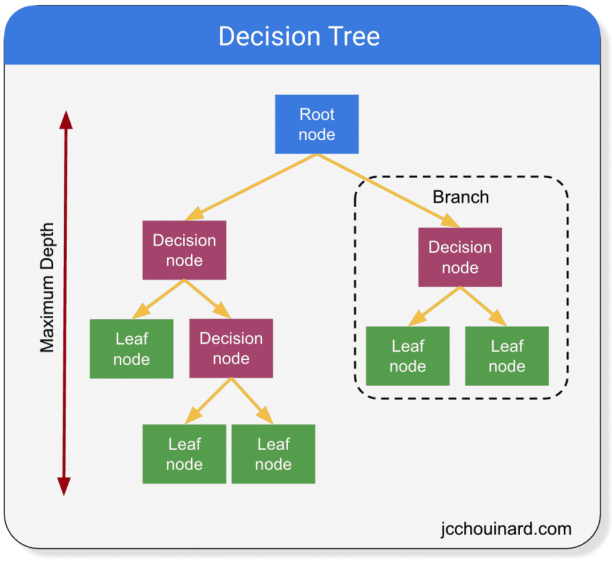


Figura 14. A figura acima representa um diagrama com a estrutura de uma árvore de decisão.

Elas possuem diversas vantagens como:

- Simples de entender e interpretar. As árvores podem ser visualizadas.
- Requer pouca preparação de dados.
- O custo de usar a árvore (ou seja, prever dados) é logarítmico no número de pontos de dados usados para treinar a árvore.
- Capaz de lidar com dados numéricos e categóricos.
- Capaz de lidar com problemas de várias saídas.
- Possível validar um modelo usando testes estatísticos. Isso torna possível explicar a confiabilidade do modelo.

Porém elas também possuem desvantagens como:

- Os aprendizes de árvores de decisão podem criar árvores super complexas que não generalizam bem os dados.
- podem ser instáveis porque pequenas variações nos dados podem resultar na geração de uma árvore completamente diferente.
- As previsões de árvores de decisão são aproximações constantes por partes, como visto na figura acima. Portanto, eles não são bons em extrapolação.
- Existem conceitos difíceis de aprender porque as árvores de decisão não os expressam facilmente, como problemas de XOR, paridade ou multiplexador.

4.4.6. Modelo Random Forest

Um classificador de floresta aleatória é um modelo preditivo, que irá criar muitas árvores de decisão, de maneira aleatória, que podemos enxergar como uma floresta, onde cada árvore será utilizada na escolha do resultado final, em uma espécie de votação. Ou seja, ajusta vários classificadores de árvore de decisão em várias sub amostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o ajuste excessivo.

Ele pode ser utilizado tanto para tarefas de classificação quanto para regressão e geralmente segue o modelo a seguir.

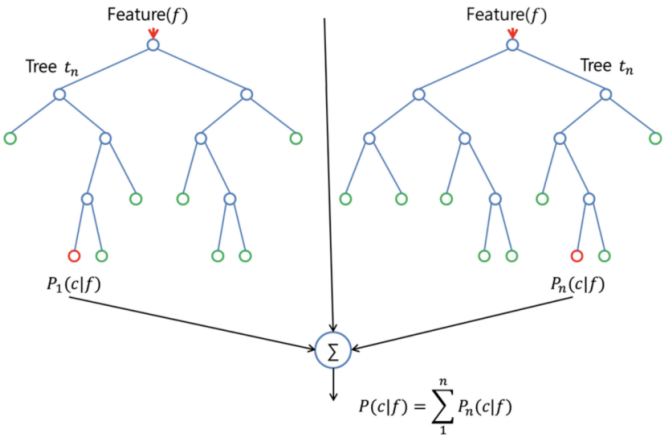


Figura 15. A figura acima mostra a representação da estrutura de uma Random Forest, e a fórmula utilizada por este modelo.

Outro fator importante das florestas aleatórias é a facilidade para se medir a importância de cada característica para a predição. Uma análise de quantos nodos das árvores, usam uma dada característica, pode reduzir a impureza geral da floresta.

4.4.7. Modelo XGBoost (Extreme Gradient Boosting)

XGBoost é um algoritmo de aprendizado de máquina, baseado em árvore de decisão que utiliza uma estrutura de Gradient boosting. Ele é uma implementação escalável e altamente precisa de aumento de gradiente que ultrapassa os limites do poder de computação para algoritmos de árvores aprimorados, construído na maioria para energizar o desempenho do modelo de aprendizado de máquina e a velocidade computacional. Com o XGBoost, as árvores são construídas em paralelo, em vez de sequencialmente como GBDT. O XGBoost funciona muito bem porque ele aprimora a estrutura básica do GBM por meio da otimização de sistemas e aprimoramentos algorítmicos.

Vantagens de usar o XGBoost :

- É mais rápido que Gradient Boosting
- Usa parallel processing o que melhora a precisão dos dados
- Funciona muito bem em banco de dados pequenos e médios

4.4.8. Hiperparâmetros utilizados nos modelos

A maioria dos algoritmos de *machine learning* são parametrizáveis, permitindo ajustes nas configurações de seus parâmetros, podendo melhorar o seu desempenho. Tais parâmetros, comumente chamados de hiperparâmetros, são definidos antes do treinamento para especificar como queremos que o treinamento do modelo aconteça. A partir do controle sobre as configurações de hiperparâmetros é possível controlar o processo de aprendizado da modelagem, evitando vícios de padronizações, como o *overfitting*, por exemplo (Barbosa, 2018).

Os algoritmos podem receber valores *default* de hiperparâmetros, mas com ajustes nessas configurações é possível melhorar o desempenho e performance dos modelos utilizados. Esse processo de buscar configurações que analisem os dados por parte dos modelos aplicados é chamado de otimização de hiperparâmetros, já que é treinado para analisar aquele conjunto de dados que o modelo irá receber.

4.4.9. *Grid Search*

Grid Search, utilizado para a otimização de hiperparâmetros, começa com a definição de uma grade de espaço de pesquisa. A grade consiste em nomes e valores de hiperparâmetros selecionados, e o Grid Search procura a melhor combinação desses valores fornecidos, treinando o modelo para indicar a melhor combinação (Kirchoff, 2019).

Atualmente, o *Grid Search* é considerado uma estratégia de encontrar hiperparâmetros bastante utilizada, por ser utilizado para treinar modelos rapidamente e gerar resultados precisos. porém é importante ressaltar que por conta da quantidade de hiperparâmetros cruzados e a quantidade de dados para o modelo predizer pode tornar o algoritmo de *Grid Search* ineficiente, fazendo com que o processo para otimização leve muito tempo para conclusão.

4.4.10. *Random Search*

Já no caso do *Random Search*, definimos distribuições para cada hiperparâmetro que podem ser definidas uniformemente ou com um método de amostragem. A principal diferença está na pesquisa aleatória, onde nem todos os valores são testados e os valores testados são selecionados aleatoriamente.

Como a pesquisa utilizada por essa estratégia é aleatória e não possui memória da otimização de hiperparâmetros utilizados, não existe uma garantia em relação ao alcance de um resultado ótimo para o problema proposto pelo projeto.

4.4.11. Considerações na escolha de algoritmos na otimização de hiperparâmetros

Estes algoritmos de otimização ajudam na configuração automática de hiperparâmetros, porém não garantem que sejam ideais para satisfazer os modelos utilizados, exigindo assim uma análise mais cautelosa por parte dos desenvolvedores, na escolha de variáveis e nos tipos de hiperparâmetros selecionados, para encontrar os conjuntos ideais para gerar resultados com melhores desempenhos dos modelos.

4.5. Avaliação

4.5.1. Conjuntos de dados utilizados para a modelagem de dados

Para o primeiro experimento realizado dos modelos deste projeto foram utilizados um mesmo conjunto de variáveis para aferir a acurácia dos testes a partir deste conjunto, bem como as mesmas métricas para a análise prévia dos resultados obtidos, sendo elas acurácia e matriz de confusão.

As variáveis escolhidas para o primeiro experimento de modelo preditivo foram das colunas “Salário_Mês”, “NumeroMeses”, “Genero” e “Cargo”, onde, respectivamente, se tratarem de um valor de salário recebido por cada colaborador por mês, o período em meses que o colaborador esteve ativo na empresa, o gênero do colaborador, e o cargo atual do colaborador.

Essa escolha se deu pela possibilidade de adaptação da modelagem aplicada em relação ao cruzamento de diferentes variáveis, enriquecendo o treino do modelo, e extraindo resultados prévios valiosos no teste experimental de diferentes modelos, e para uma avaliação preliminar da performance de cada um deles, onde a proporção dos dados para treino e teste foi de 80% e 20%, respectivamente, em cada modelagem que será descrita a seguir.

É importante ressaltar que a proporção 80:20 da separação dos dados para treino e teste, respectivamente, não segue a porção 1:3, que é a proporção comumente utilizada, já que corre-se o risco de separar os dados de forma desfavorável, pelo volume de amostras não ser tão significativo (Venturini, 2020).

Para o teste de modelos com hiperparâmetros foram utilizados o conjunto de variáveis citadas a seguir:

- Salário;
- Gênero;

- Período na empresa;
- Cargo atual;
- Área atual;
- Score do ambiente de trabalho da área;
- Tempo médio de promoções;
- Tempo das promoções recebidas;
- Estado civíl;
- Escolaridade;
- Estado.

E foram aplicadas as seguintes métricas de avaliação:

- Acurácia;
- Precisão;
- Revocação;
- F1 Score.

Utilizados outros atributos para enriquecer os modelos com hiperparâmetros, levando a uma análise mais profunda em relação ao desempenho e performance de cada modelo sobre os resultados obtidos.

4.5.2. Métricas utilizadas para avaliação dos modelos

A respeito das métricas utilizadas para avaliar os resultados dos experimentos, e para realizar uma análise preliminar de cada um dos modelos, foi a acurácia do treino de cada modelo, a acurácia do resultado do teste, e também foi aplicada uma matriz de confusão em cada uma das modelagens.

A acurácia é uma métrica simples de ser aplicada, já que é utilizada em sua fórmula a razão entre todos os acertos do modelo, em relação à quantidade total de elementos usados para a predição. Sendo assim, é importante pontuar também que apenas com o uso da acurácia, não é possível avaliar o real desempenho da modelagem, já que a fórmula apresenta valores sem um peso aplicado, o que significa que uma acurácia elevada pode não significar que o modelo é eficaz (Chen, et al, 2020).

$$Acurácia = \frac{Verdadeiros\ Positivos + Verdeiros\ Negativos}{Total}$$

Figura 16. Fórmula utilizada para o cálculo da acurácia.

Para complementar a análise da performance de cada modelo, foi utilizada uma matriz de confusão, que utiliza os acertos do modelo preditivo, e a "confusão", os erros, do modelo, para uma visualização qualitativa mais clara do resultado obtido da modelagem aplicada (Nascimento, 2019).

Um exemplo de matriz de confusão pode ser visto a seguir, onde os acertos do modelo são os quadrantes de Verdadeiro Positivo (VP) e Verdadeiro Negativo (VN), e os erros estão nos quadrantes Falso Positivo (FP) e Falso Negativo (FN):

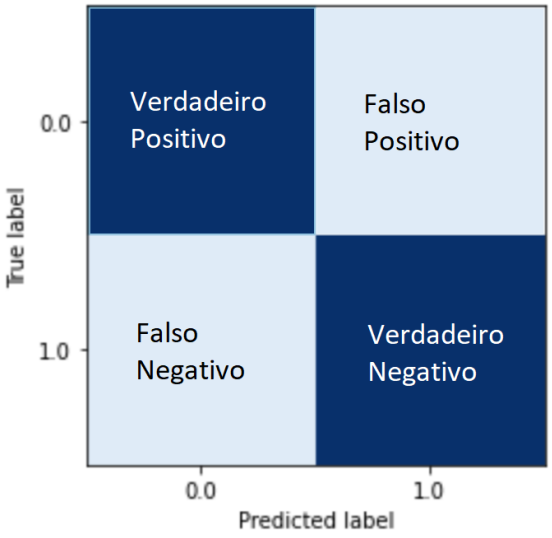


Figura 17. Representação de uma matriz de confusão.

4.5.2.1. Precisão

A precisão é uma métrica utilizada para identificar qual é a porcentagem de acertos para amostras com a relação entre positivos verdadeiros (VP) e falsos positivos (FP), onde a fórmula pode ser dada a seguir:

$$Precision = \frac{VP}{VP+FP}$$

Figura 18. Fórmula utilizada para o cálculo da precisão.

4.5.2.2. Revocação

No caso da revocação, é uma métrica semelhante à precisão, mas ela identifica a relação entre verdadeiro positivo (VP) com falso negativo (FN), com a fórmula a seguir:

$$Recall = \frac{VP}{VP+FN}$$

Figura 19. Fórmula utilizada para o cálculo da revocação.

4.5.2.3. F1 *Score*

F1 *Score* utiliza o resultado das métricas de revocação e precisão, sendo uma média harmônica entre as duas, com a seguinte fórmula:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Figura 20. Fórmula utilizada para o cálculo do F1 *Score*.

4.5.3. Modelo SVM - Support Vector Machines

O experimento realizado com este modelo foi com o conjunto de dados selecionados para a experimentação de todos os modelos, como indicado no tópico acima, e foi aplicada a métrica de acurácia do treino, do resultado do teste, e foi gerada uma matriz de confusão.

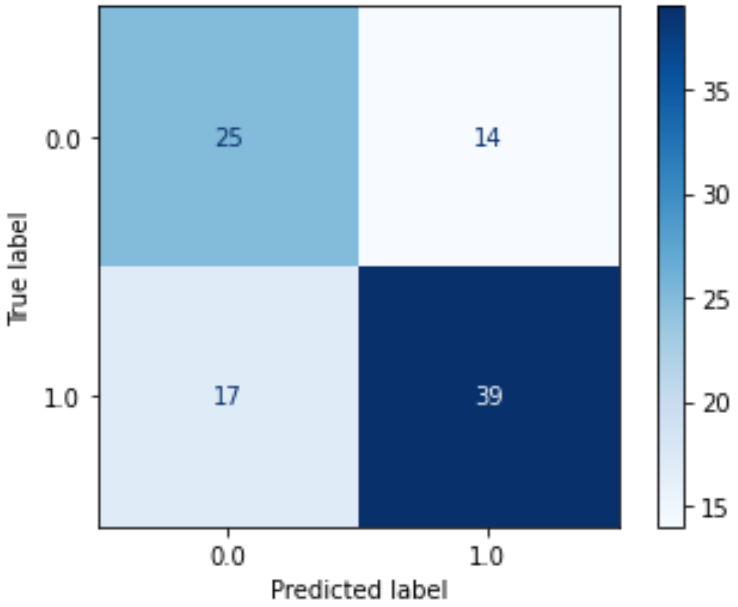


Figura 21. Matriz de confusão gerada pelo modelo SVM.

A partir da matriz de confusão gerada da modelagem de SVM, é possível notar que a maior parte dos acertos, representado pela cor mais escura, está no Verdadeiro Negativo (VN), que representa os funcionários ativos, corretamente preditos, e em seguida, o quadrante do Verdadeiro Positivo (VP), de funcionários desligados.

Com isso, a acurácia do acerto desta categoria de modelagem foi de 67,3%, indicando que o modelo nesse experimento acertou os valores acima da média, mas ainda apresenta um erro considerável, se levado em consideração os valores preditos de forma errônea.

A partir da acuracidade do treino, que foi de 67,1%, é importante destacar que as configurações dos dados que nutriram o motor do modelo precisam de ajustes para melhor separados para o treinamento.

Para finalizar a análise preliminar do modelo SVM, é possível ver com a matriz de confusão que a maior parte dos acertos está com os valores VN, porém os dados mais importantes para este projeto são os valores presentes no quadrante VP, mostrando que apenas a métrica absoluta da acurácia não revela o real estado do desempenho do modelo SVM.

4.5.4. Modelo Naïve Bayes

O experimento realizado com este modelo foi com o conjunto de dados selecionados para a experimentação de todos os modelos, como indicado no tópico 4.5.2., e foi aplicada a métrica de acurácia do treino, do resultado do teste, e foi gerada uma matriz de confusão.

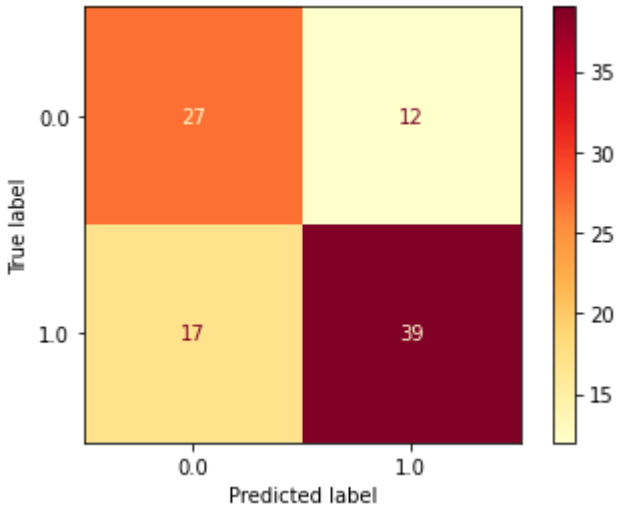


Figura 22. Matriz de confusão gerada pelo modelo Naive Bayes.

A partir da matriz de confusão gerada da modelagem de Naïve Bayes, é possível notar que a maior parte dos acertos, representado pela cor mais escura, está no Verdadeiro Negativo (VN), que representa os funcionários ativos, corretamente preditos, e em seguida, o quadrante do Verdadeiro Positivo (VP), de funcionários desligados.

Com isso, a acurácia do acerto desta categoria de modelagem foi de 69. 4%, indicando que o modelo nesse experimento acertou os valores acima da média, mas ainda apresenta um erro considerável, se levado em consideração os valores preditos de forma errônea.

A partir da acuracidade do treino, que foi de 61.8.%, é importante destacar que as configurações dos dados que nutriram o motor do modelo precisam de ajustes para melhor separados para o treinamento.

Para finalizar a análise preliminar do modelo Naïve Bayes, é possível ver com a matriz de confusão que a maior parte dos acertos está com os valores VN, porém os dados mais importantes para este projeto são os valores presentes no quadrante VP, mostrando que apenas a métrica absoluta da acurácia não revela o real estado do desempenho do modelo.

4.5.5. Modelo KNN - K-Nearest Neighbour

O experimento realizado com este modelo foi com o conjunto de dados selecionados para a experimentação de todos os modelos, como indicado no tópico 4.5.3. , e foi aplicada a métrica de acurácia do treino, do resultado do teste, e foi gerada uma matriz de confusão.

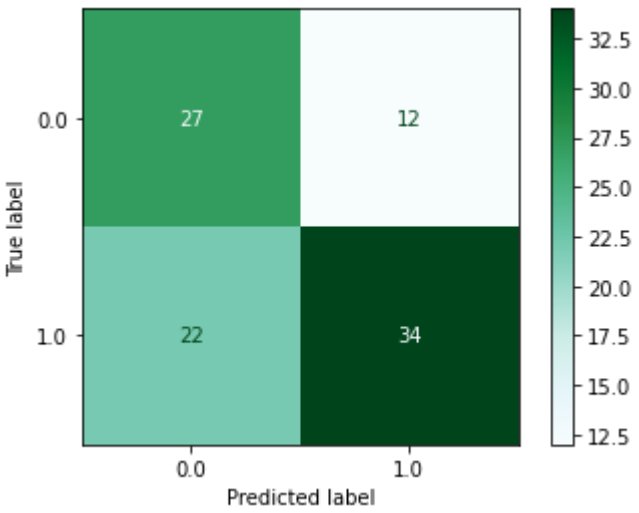


Figura 23. Matriz de confusão gerada pelo modelo KNN.

A partir da matriz de confusão gerada da modelagem de KNN, é possível notar que a maior parte dos acertos, representado pela cor mais escura, está no Verdadeiro Negativo (VN), que representa os funcionários ativos, corretamente preditos, e em seguida, o quadrante do Verdadeiro Positivo (VP), de funcionários desligados.

Com isso, a acurácia do acerto desta categoria de modelagem foi de 64.2%, indicando que o modelo nesse experimento acertou os valores acima da média, mas ainda apresenta um erro considerável, se levado em consideração os valores preditos de forma errônea.

A partir da acuracidade do treino, que foi de 73.3%, é importante destacar que as configurações dos dados que nutriram o motor do modelo precisam de ajustes para melhor separados para o treinamento.

Para finalizar a análise preliminar do modelo KNN, é possível ver com a matriz de confusão que a maior parte dos acertos está com os valores VN, porém os dados mais importantes para este projeto são os valores presentes no quadrante VP, mostrando que apenas a métrica absoluta da acurácia não revela o real estado do desempenho do modelo.

4.5.6. Modelo AdaBoost - (Adaptive Boosting)

Por ser um modelo que repete várias vezes o processo de análise e predição dos dados, através da matriz de confusão do modelo a seguir foi possível obter resultados diferentes entre os outros modelos aplicados na experimentação.

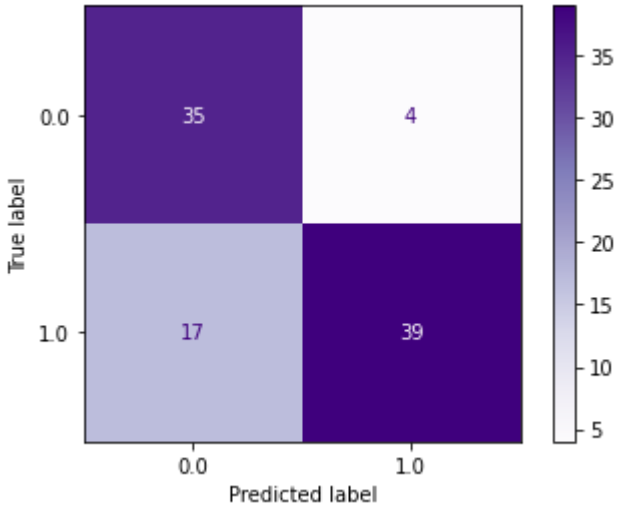


Figura 24. Matriz de confusão gerada pelo modelo AdaBoost.

Com uma acurácia de 77,8% em relação ao teste da predição do modelo, é possível ver na matriz que a proporção entre VP e VN foi a mais próxima, em relação aos outros modelos utilizados no experimento.

Outro ponto a destacar foi que o FP, considerado o ponto mais importante a evitar, já que apresenta o maior risco, por não identificar os colaboradores que apresentam a intenção de se desligar da empresa.

4.4.7. Modelo XGBoost (Extreme Gradient Boosting)

Já que esse modelo utiliza fórmulas semelhantes ao modelo de AdaBoost, foi possível avaliar que o resultado obtido com o experimento do modelo foi semelhante, a partir da matriz de confusão a seguir:

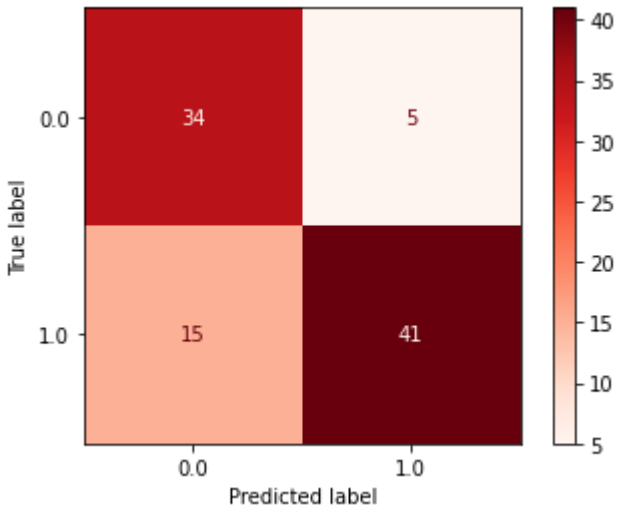


Figura 25. Matriz de confusão gerada pelo modelo XGBoost.

A principal diferença entre o resultado deste modelo com o AdaBoost foi a acurácia do teste, que foi de 78,9%, mas ao olhar a matriz de confusão, é possível aferir que o aumento dos acertos foi no quadrante VN, sendo que o quadrante VP seria o de mais valor para o intuito do projeto.

4.5.5. Modelo Decision tree

O experimento realizado com este modelo foi com o conjunto de dados selecionados para a experimentação de todos os modelos, como indicado no tópico 4.5.5. , e foi aplicada a métrica de acurácia do treino, do resultado do teste, e foi gerada uma matriz de confusão.

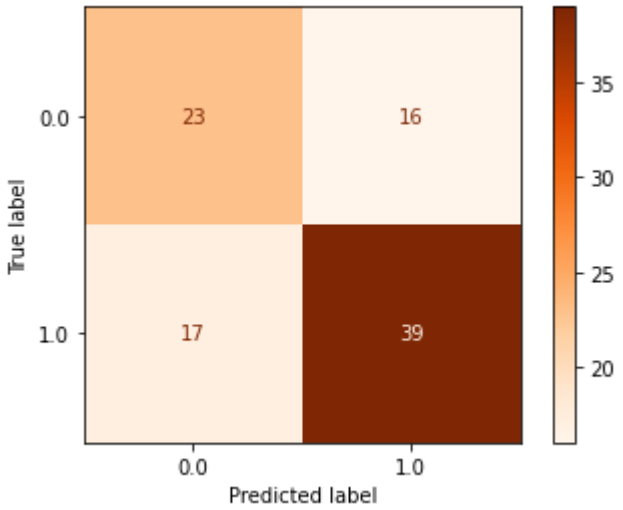


Figura 26. Matriz de confusão gerada pelo modelo Decision Tree.

A partir da matriz de confusão gerada da modelagem de Árvore de Decisão , é possível notar que a maior parte dos acertos, representado pela cor mais escura, está no Verdadeiro Negativo (VN), que representa os funcionários ativos, corretamente preditos, e em seguida, o quadrante do Verdadeiro Positivo (VP), de funcionários desligados.

Com isso, a acurácia do acerto desta categoria de modelagem foi de 65.2%, indicando que o modelo nesse experimento acertou os valores acima da média, mas ainda apresenta um erro considerável, se levado em consideração os valores preditos de forma errônea.

A partir da acuracidade do treino, que foi de 99.3%, é importante destacar que as configurações dos dados que nutriram o motor do modelo precisam de ajustes para melhor separados para o treinamento.

Para finalizar a análise preliminar do modelo Árvore de Decisão, é possível ver com a matriz de confusão que a maior parte dos acertos está com os valores VN, porém os dados mais importantes para este projeto são os valores presentes no quadrante VP, mostrando que apenas a métrica absoluta da acurácia não revela o real estado do desempenho do modelo.

4.5.5. Modelo Random Forest

O experimento realizado com este modelo foi com o conjunto de dados selecionados para a experimentação de todos os modelos, como indicado no tópico 4.5.6. , e foi aplicada a métrica de acurácia do treino, do resultado do teste, e foi gerada uma matriz de confusão.

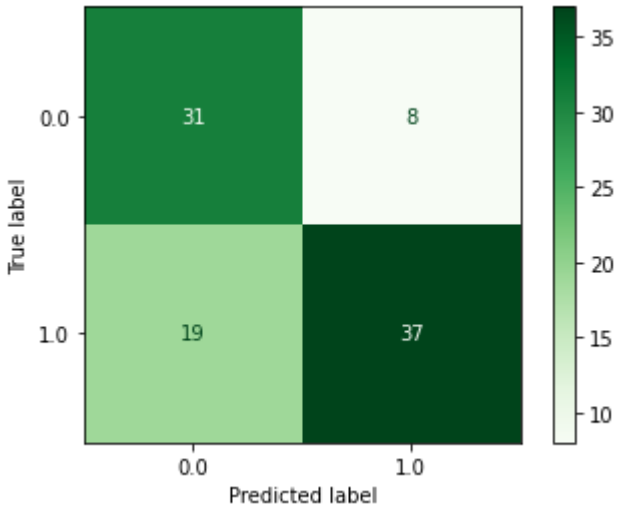


Figura 27. Matriz de confusão gerada pelo modelo Random Forest.

A partir da matriz de confusão gerada da modelagem de Random Forest, é possível notar que a maior parte dos acertos, representado pela cor mais escura, está no Verdadeiro Negativo (VN), que representa os funcionários ativos, corretamente preditos, e em seguida, o quadrante do Verdadeiro Positivo (VP), de funcionários desligados.

Com isso, a acurácia do acerto desta categoria de modelagem foi de 71.5%, indicando que o modelo nesse experimento acertou os valores acima da média, mas ainda apresenta um erro considerável, se levado em consideração os valores preditos de forma errônea.

A partir da acuracidade do treino, que foi de 99.3%, é importante destacar que as configurações dos dados que nutriram o motor do modelo precisam de ajustes para melhor separados para o treinamento.

Para finalizar a análise preliminar do modelo Random Forest, é possível ver com a matriz de confusão que a maior parte dos acertos está com os valores VN, porém os dados mais importantes para este projeto são os valores presentes no quadrante VP, mostrando que apenas a métrica absoluta da acurácia não revela o real estado do desempenho do modelo.

4.5.6. Modelo AdaBoost com Hiperparâmetros

Com os resultados obtidos a partir das métricas temos:

Métrica aplicada	Valor obtido
Acurácia	78%
Revocação	72%
Precisão	85%
F1 *Score*	78%

Com esses resultados podemos avaliar o desempenho do modelo sobre o valor do acerto da predição e o que o projeto avalia como pontos de atenção sobre a performance do modelo.

Com uma revocação mais baixa em relação às outras métricas, vemos que o modelo apresenta problemas para identificar resultados falsos negativos, em relação à precisão, que mostra uma melhor performance para evitar resultados falsos positivos.

Utilizando o F1 Score, podemos verificar nesse modelo, comparando com a acurácia, que os acertos do modelo apresentam valores próximos, que valida o valor da acurácia em termos dos acertos da predição.

4.5.7. Modelo XGBoost com Hiperparâmetros

Com os resultados obtidos a partir das métricas temos:

Métrica aplicada	Valor obtido
Acurácia	81%
Revocação	77%
Precisão	83%

F1 *Score*	80%

Com esses resultados podemos avaliar o desempenho do modelo sobre o valor do acerto da predição e o que o projeto avalia como pontos de atenção sobre a performance do modelo.

Com uma revocação mais baixa em relação às outras métricas, vemos que o modelo apresenta problemas para identificar resultados falsos negativos, em relação à precisão, que mostra uma melhor performance para evitar resultados falsos positivos.

Utilizando o F1 Score, podemos verificar nesse modelo, comparando com a acurácia, que os acertos do modelo apresentam valores próximos, que valida o valor da acurácia em termos dos acertos da predição

4.5.8. Modelo Random Forest com Hiperparâmetros

Com os resultados obtidos a partir das métricas temos:

Métrica aplicada	Valor obtido
Acurácia	85%
Revocação	80%
Precisão	88%
F1 *Score*	84%

Com esses resultados podemos avaliar o desempenho do modelo sobre o valor do acerto da predição e o que o projeto avalia como pontos de atenção sobre a performance do modelo.

Com uma revocação mais baixa em relação às outras métricas, vemos que o modelo apresenta problemas para identificar resultados falsos negativos, em relação à precisão, que mostra uma melhor performance para evitar resultados falsos positivos.

Utilizando o F1 Score, podemos verificar nesse modelo, comparando com a acurácia, que os acertos do modelo apresentam valores próximos, que valida o valor da acurácia em termos dos acertos da predição

4.5.9. Modelo Naive Bayes com Hiperparâmetros

A partir dos valores obtidos com o modelo utilizando hiperparâmetro temos:

Métrica aplicada	Valor obtido
Acurácia	53%
Revocação	13%
Precisão	100%
F1 *Score*	23%

Ao olhar o valor da acurácia já é possível perceber que mesmo com hiperparâmetros configurados, em relação ao teste realizado com o modelo anteriormente.

As demais métricas reforçam isso, com uma revocação muito baixa, e uma precisão em 100%, sugerindo que o modelo possui *overfitting*.

O F1 Score ainda revela a grande diferença entre uma revocação muito baixa com uma precisão tão alta, levando a uma proporção muito menor entre os acertos, em comparação com a acurácia apontada de 53,5%.

A baixa performance deste modelo pode vir da dificuldade que esse modelo tem de trabalhar com uma modelagem de classificação, sendo majoritariamente usado em modelagem regressiva, por exemplo.

5. Conclusões e Recomendações

Por conta da digitalização dos processos e dos sistemas das empresas, um grande volume de dados e informações digitais são gerados. Isso abre diversas possibilidades de usos sobre estes dados, que vão além de aperfeiçoar os processos internos, criando ferramentas que darão suporte a setores inteiros, que é o caso estudado neste trabalho.

Porém, por se tratar de uma grande quantidade de dados, estes requerem diversos tratamentos, processamentos e análises provenientes de uma estruturação bem definida, técnicas de visualização e manipulação de dados , metodologias de modelagem, entre outros.

Utilizar estes dados, transformando-os em informações relevantes, podendo servir para auxiliar tomadores de decisão, como gestores, a coordenarem estratégias entre os integrantes de sua equipe, representa um grande desafio, contínuo, que apresenta um grande valor para as empresas do setor de tecnologia, já que possui a proposta de reter profissionais do setor, em um momento de uma alta competição entre as empresas, por profissionais especializados em tecnologia da informação.

Também é importante destacar o trabalho recorrente que a predição de tendências com modelagem de dados necessita, por conta de avanços tecnológicos, atualizações legais sobre o uso de dados.

O desenvolvimento deste trabalho, de cunho acadêmico, utilizando uma amostra de dados disponibilizados pela empresa parceira, bem como orientação e suporte dos professores do Inteli - Instituto de Tecnologia e Liderança permitiu com este projeto:

- Compreender o processo de desenvolvimento de modelagem preditiva com a metodologia CRISP-DM;
- Analisar e desenvolver tratamentos necessários para manipular os dados adquiridos de forma ética, evitando viés no processamento desses dados;
- Identificar o tipo de algoritmo de aprendizado supervisionado de máquina adequado para predição e classificação dos dados presentes na amostra disponibilizada.

O modelo avaliado como mais consistente e equilibrado em termos de resultados obtidos, eleito então para manter como modelo de predição supervisionado escolhido foi utilizando o algoritmo de *Random Forest*, aplicando o hiperparâmetro de *Random Search*, onde o resultado final obtido pode ser verificado a seguir:

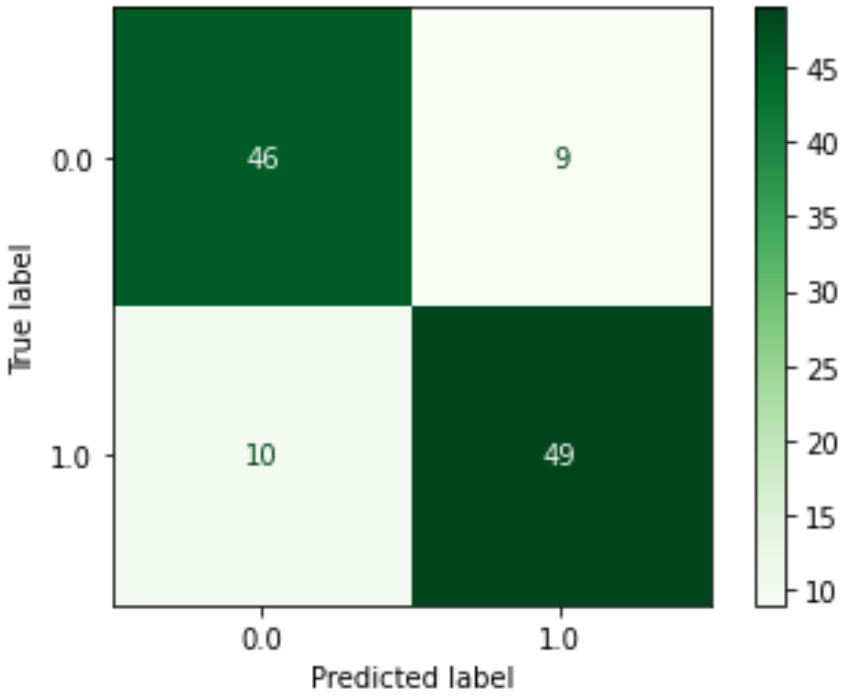


Figura 28. Matriz de confusão gerada pelo modelo Random Forest com a utilização do hiperparâmetro *Random Search*.

Métrica aplicada	Valor obtido
Acurácia	83%
Revocação	85%
Precisão	83%
F1 *Score*	84%

Com a matriz de confusão gerada após a predição do modelo de Random Forest, com as variáveis estabelecidas na seção e com o hiperparâmetro citado, sendo cruzada com os valores de Acurácia, Revocação, Precisão e F1 *Score* do modelo em questão, percebemos um equilíbrio entre os resultados, que variam entre 83% e 85%, indicando uma consistência no acerto da predição realizada com as configurações aplicadas.

Isso possibilita a utilização deste modelo final, como ferramenta, para a recomendação das seguintes etapas futuras, para dar sequência a este projeto:

- Teste com um novo conjunto de dados, com uma amostra diferente da utilizada, possibilitando a validação da modelagem desenvolvida, a partir da avaliação do desempenho e performance com esta nova amostra.
- A construção de um *front-end*, para permitir a usabilidade do modelo por alguns gestores, para testar o modelo com um cenário com dados reais, fora do ambiente com valores de teste, e avaliar o desempenho e performance dos acertos, com os valores de acurácia.

A utilização de modelos preditivos abrem a possibilidade de melhorar a eficiência, gestão, e aprimorar as estratégias aplicadas, no caso deste trabalho, com os colaboradores de tecnologia de uma empresa. Assim, é essencial destacar as implicações e responsabilidades éticas sobre sua utilização, e o impacto que essa ferramenta pode apresentar nos colaboradores da empresa.

Os dados presentes nas tabelas nutrem fórmulas matemáticas dos algoritmos da modelagem de predição, carregando consigo comportamentos e eventos mensuráveis por datas, valores de salário, idades, desempenhos e avaliações dos colaboradores, por exemplo.

Porém existem dados não coletados que pesam muito no comportamento e nas escolhas feitas por pessoas. Por isso, a recomendação final deste trabalho é utilizar esta ferramenta com responsabilidade, e com a finalidade de auxiliar os tomadores de decisão da empresa, gestores, e o setor de Pessoas e Cultura da empresa, levando em consideração a acuracidade do modelo, que acerta, no caso, cerca de 80% de suas predições, tornando esta ferramenta útil na elaboração de estratégias, de abordagem e aproximação em equipes que apresentarem integrantes em que o modelo considera que desejam se desligar da empresa, levando em conta a taxa de erro de 20% nas predições.

6. Referências

ALIARI, S.: Combination of persona and user journey map in service design process, University of Theran, Department of Industrial Design, 2018.

ALVES, N. H., & TESSMANN, L. G. dos S.: Matriz De Risco: Um Estudo Em Uma Empresa Calçadista Do Vale do Paranhana, Revista Eletrônica de Ciências Contábeis, 2018.

ARAÚJO NETO, A. P.: Governança de dados - 1. ed. Platos Soluções Educacionais S.A., 2021.

BARBOSA, F. R. M.: Otimizacao De Hiperparâmetros Em Algoritmos De ˆ Árvore De Decisão Utilizando Computação Evolutiva - Universidade Federal Do Tocantins, Campus Universitário De Palmas, Curso De Ciências Da Computação, 2018.

BENZAGHTA et al.: SWOT analysis applications: An integrative literature review - Journal of Global Business Insights, 2021.

BRUIJL, G. H. Th.: The Relevance of Porter's Five Forces in Today's Innovative and Changing Business Environment, 2018.

CHIAT , L. C., & PANATIK, S. A. . Perceptions of Employee Turnover Intention by Herzberg's Motivation-Hygiene Theory: A Systematic Literature Review . Journal of Research in Psychology, 2019.

FACELI, K. et al: Inteligência artificial: uma abordagem de aprendizado de máquina - 2. ed. Livros Técnicos e Científicos Editora Ltda, 2022.

HOLMSTRÖM, M., & SKOOG, E.: User Journey Map and development of a pantyliner, Presenting an approach for User Centro de Design in feminine care. Chalmers University Of Technology, Department of Product- and Production Development, 2017.

HUANG, S. et al.: Applications of Support Vector Machine (SVM) Learning in Cancer Genomics, Cancer Genomics & Proteomics, 2018.

KIRCHOFF, D. F.: Avaliação De Técnicas De Aprendizado De Máquina Para Previsão De Cargas De Trabalho Aplicadas Para Otimizar O Provisionamento De Recursos Em Nuvens Computacionais - Pontifícia Universidade Católica Do Rio Grande Do Sul, Escola Politécnica Programa De Pós-Graduação Em Ciência Da Computação, 2019.

LENZ, M. L. et al.: Fundamentos de aprendizagem de máquina, SAGAH, 2020.

NASCIMENTO, J. da C. .: Avaliação De Desempenho De Algoritmos De Classificação Em Mineração De Opinião Em Textos Em Português, Universidade Federal Do Acre - Centro De Ciências Exatas E Tecnológicas, 2019.

NETO, M. V. G.: O processo CRISP-DM aplicado na construção de uma solução para Análise de Risco de Crédito - Universidade Federal de Pernambuco - UFPE Centro de Informática, 2018

ORTIZ, O. A. G.: Aplicación del sistema de Machine Learning para aumentar la eficiencia de las organizaciones, Universidad Militar Nueva Granada - Facultad De Ciencias Economicas, 2021.

POKORNÁ et al.: Value Proposition Canvas: Identification of Pains, Gains and Customer Jobs at Farmers' Markets, AGRIS on-line Papers in Economics and Informatics, 2015.

SANSONE, V. T. B., VECCHIA, R. D.: Construção do modelo preditivo de desligamento de colaboradores, Revista Brasileira de Administração Científica, 2021.

SANTOS, T. M.; DEL VECHIO, G. H. A Gestão De Relacionamento com Clientes (CRM) Como Um Importante Recurso Para O Crescimento Empresarial. Revista Interface Tecnológica, [S. l.], v. 17, n. 1, p. 819–828, 2020.

SHARDA, R., DELEN, D., TURBAN, E.: Business intelligence e análise de dados para gestão do negócio - 4. ed. Bookman, 2019.

SILVA, L. A., PERES, S. M., BOSCARIOLI, C.: Introdução à mineração de dados: com aplicações em R - 1. ed. Elsevier Editora Ltda, 2016.

SKELTON, A. R., NATTRESS, D., DWYER, R. J.: Predicting manufacturing employee turnover intentions, Journal of Economics, Finance and Administrative Science, 2019.

TU, N., DONG, X., RAU, P. -L. P. and ZHANG, T.: Using cluster analysis in Persona development, 8th International Conference on Supply Chain Management and Information, 2010.

VENTURINI, F. C.: Uso de modelos preditivos na gestão de riscos da Fiscalização Tributária, Universidade de Brasília, Instituto de Ciências Exatas -Departamento de Ciência da Computação, 2020.

VILELA JUNIOR, G. B. et al: Métricas Utilizadas Para Avaliar A Eficiência De Classificadores Em Algoritmos Inteligentes, Revista CPAQV - Centro de Pesquisas Avançadas em Qualidade de Vida, 2022.

FERN FORT UNIVERSITY. Salesforce.com, inc. Porter Five Forces Analysis. Disponível em: <Salesforce.com, inc. Porter Five (5) Forces & Industry Analysis [Strategy] (fernfortuniversity.com)>. Acesso em: 08 de agosto de 2022.

HENRY, Zander. Salesforce Porter Five Forces Analysis - case48, 2018. Disponível em: <Salesforce Porter Five Forces Analysis (case48.com)>. Acesso em: 10 de agosto de 2022.

IBM SPSS Modeler CRISP-DM Guide. IBM Corporation, 2011. Disponível em: <public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf>. Acesso em: 1 de agosto de 2022.

BAMBRICK, N.: Support Vector Machines: A Simple Explanation - KDnuggets, 2016. Disponível em: <https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html>. Acesso em: 08 de setembro de 2022.

TCHILIAN, F.: Modelo Preditivo: o que é, para que serve e como aplicá-lo? - ClearSale, 2022. Disponível em: <https://blogbr.clear.sale/modelo-preditivo-saiba-como-aplica-lo#:~:text=Um%20modelo%20preditivo%20%C3%A9%2C%20de,matem%C3%A1tica%2C%20com%20probabilidade%20e%20estat%C3%ADstica.>. Acesso em: 08 de setembro de 2022.

CHEN, D., NIGRI, E., OLIVEIRA, G.,SEPULVENE, L., ALVES, T.: Métricas de Avaliação em Machine Learning: Classificação - Kunumi Blog, medium, 2020. Disponível em: <https://medium.com/kunumi/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-em-machine-learning-classifica%C3%A7%C3%A3o-49340dcdb198>. Acesso em: 09 de setembro de 2022.

DIAS, M.: People Analytics: o que é, benefícios e como aplicar no RH - Gupy Blog, 2022, Disponível em: <https://www.gupy.io/blog/people-analytics>. Acesso em: 12 de setembro de 2022.

John F. Magee: Decision Trees for Decision Making https://hbr.org/1964/07/decision-trees-for-decision-making Acesso em:12/09/2022.