INSTITUTO DE TECNOLOGIA E LIDERANÇA – INTELI

MODELO PREDITIVO – EVERYMIND

IN & OUT DEVS

SÃO PAULO – SP
2022

INSTITUTO DE TECNOLOGIA E LIDERANÇA – INTELI

MODELO PREDITIVO – EVERYMIND

IN & OUT DEVS

Autores: Emanuele Lacerda Morais Martins

Giovanna Furlan Torres

Jean Lucas Rothstein Machado

Lucas de Britto Vieira

Patrick Victorino Miranda

Pedro Henrique Sant’Anna Oliveira

Data de criação: 09 de Agosto de 2022

SÃO PAULO – SP

2022

Sumário

Índice de figuras

Índice de tabelas

1. Introdução

A Everymind é uma das maiores parceiras Salesforce na América Latina com escritório no Brasil, além de atuações em implementações nas Américas, Japão e Europa. Oferecendo suporte técnico e gestão empresarial da Salesforce e o desenvolvimento de novas funcionalidades para a plataforma. A empresa possui um perfil consultivo, com centenas de profissionais qualificados para o desenvolvimento do ecossistema Salesforce, diversos projetos concluídos e um nome já consolidado. Além de toda a estrutura, a companhia possui interesse em entender o que retém seus funcionários dentro da empresa. Assim, a construção de um modelo preditivo para a alta taxa de turnover de funcionários, auxilia a Everymind a ter um direcionamento a respeito da longevidade dos colaboradores na empresa que implica altos custos, entre eles, o onboarding.

2. Objetivos

2.1 Objetivos Gerais

A empresa Everymind objetiva, em termos gerais, diminuir o turnover de funcionários, fazendo com que aumente a longevidade destes na organização. Para alcançar esse fim, a companhia espera conseguir prever a tendência de um funcionário sair e se antecipar em relação a melhoria de possíveis fatores que o retenham e melhorem sua convivência e bem-estar, diminuindo inclusive o gasto atual com onboarding de colaboradores.

2.2 Objetivos Específicos

Deseja-se conseguir classificar um funcionário, a partir de um modelo preditivo, de acordo com a chance dele de sair da empresa, obtendo informações, através da análise de dados, sobre quais características mais influenciam a saída deste, identificando quais são os períodos ou situações que levam a perda de funcionários na empresa e assim conseguir tomar providências acerca da permanência do mesmo. Tais mudanças podem estar relacionadas ao bem-estar pessoal, benefícios adquiridos, salário, relacionamento com os integrantes dos times e gestores, além de identificação com a cultura da organização.

3. Descritivo da Solução

A proposta de solução visa a criação de um modelo preditivo de classificação para propriedade da empresa Everymind. Um modelo preditivo são funções matemáticas que aplicadas a uma base de dados selecionada, seja capaz de identificar padrões e prever com eficiência a tendência de algo acontecer, baseando-se em sua variável alvo. Em relação a tal modelo oferecido, a variável alvo é a tendência de um funcionário ficar ou sair da empresa (sendo representada de forma binária “sim” ou “não”).A principal funcionalidade desse modelo preditivo é a de servir como auxílio na tomada de decisão para prever o turnover de funcionários na Everymind, entendendo quais são as pessoas mais propensas a saírem da empresa ou que necessitem de uma ação de reconhecimento.

3.1 Justificativa

Atualmente as empresas vêm sendo afetadas pela intensa rotatividade dos seus colaboradores. Esse problema atinge a companhia de diversas maneiras, como: 1) Os gastos contínuos com contratação; e 2) Treinamento e desenvolvimento de novos funcionários. Além disso, essa situação interfere em toda dinâmica do negócio, desde a produção, criação, desenvolvimento até a entrega final para os consumidores.

Esse modelo de predição irá fornecer a área de RH da Everymind quais colaboradores são mais propenso a saírem da empresa, contribuindo para que eles encontrem maneiras de reduzir a taxa de turnover e que melhorem a experiência dos seus colaboradores, através de um “Lock in”, sendo esse uma forma de beneficiar os funcionários que apresentam características que condizem com a cultura da empresa, fornecendo incentivos de permanência na instituição.

3.2 Partes Interessadas

A empresa Everymind exerce o papel de auxiliar os desenvolvedores com o desenvolvimento do modelo preditivo, seja por meio do fornecimento dos dados a respeito dos colaboradores, quanto da empresa, constatando quais conteúdos são ou não restritos para o compartilhamento.

Com a entrega da solução proposta, espera-se de benefícios ao cliente: 1) Reduzir o percentual de colaboradores que desejam sair da companhia, não sendo necessário a substituição dos mesmos; 2) Auxílio à empresa na maneira de manter seus funcionários, mostrando quais parâmetros agregam mais para a sua permanência; 3) Destacar os trabalhadores que mediante as características estabelecidas, alcançaram o nível de receberem um reconhecimento, mérito ou promoção; e 4) Tornar todo o processo mais dinâmico, ágil e eficaz, com cores separando cada alocação de pessoas e suas probabilidades de permanência, dashboard de visualização com gráficos e ambiente de inserção de dados dos colaboradores.

3.3 Descrição da predição

O tipo de predição desejada é o método classificatório, apresentando um modelo que estuda as relações entre duas variáveis numéricas. Em que, todo valor da variável independente (x) é associada com um valor da variável dependente (y), por exemplo: uma variável independente seria a profissão e a dependente o salário, através da associação das duas é possível identificar a maneira como elas se relacionam e interferem na saída do funcionário. Estimando esse resultado através de classificações atribuídas ao funcionário, como propenso a sair, não propenso a sair, calculadas a partir do impacto positivo ou negativo na permanência do funcionário gerado pela associação entre dados.

A escolha se deve ao fato de que ao classificar os funcionários em faixas, ela consegue identificar se uma pessoa é mais propensa a sair do que outra, e atribuir um significado a esse resultado. Assim, fornece uma classificação que indica se o funcionário possui ou não a possibilidade de sair, ajudando a empresa a saber em quem ela deve focar primeiro, e o nível de atenção que deve fornecer. Assim, a escolha possui caráter não binário, pois o modelo possui mais de duas faixas de classificação.

Pode-se utilizar a solução proposta para, ao inserir dados dos colaboradores no sistema, através de um excel importado pelo Google Drive ao notebook do Google Collaboratory, as células de código do modelo preditivo será rodada e através das análises de padrões de dados encontrados, tem-se como devolutivas quais funcionários estão mais propensos a sair ou permanecer na empresa. Além de um Wireframe com um dashboard previsto para futuramente ser integrado ao modelo, produzindo um retorno visual, em forma de gráfico, porcentagem e texto, destacando através de cores, quais

usuários estão propensos a abandonar a empresa e os que querem permanecer neste ambiente corporativo.

3.4 Dados utilizados na solução

Abaixo se apresenta as bases de dados utilizadas durante o desenvolvimento da solução, no decorrer do documento será exemplificado e contextualizado quais foram suas atuações e importâncias no modelo.

1. Dados de cadastramento de funcionários – Descrevendo informações pessoais e áreas que atuam;
2. Dados dos reconhecimentos fornecidos aos colaboradores – Descrevendo quais ganharam uma promoção e mudaram de cargo e quais ganharam mérito e aumentaram o salário;
3. Dados de ambiente de trabalho – Descreve os meios que a empresa é avaliada pelas áreas da instituição e como cada categoria é afetada pelo modo que o bem-estar dos trabalhadores é gerido.

Para realizar a avaliação do critério de sucesso do modelo, após a preparação da base de dados disponibilizada, será medido a partir da divisão entre os dados do modelo, aplicando em primeiro instante somente metade destes para treinamento, e o restante para validação de acerto. Utilizando como medida de avaliação o cálculo da porcentagem de erro ou acerto do modelo.

4. Metodologia

Abaixo é apresentado as metodologias utilizadas como base para a criação do modelo preditivo como um todo.

4.1 CRISP-DM

Nesta sessão é exibido as etapas que correspondem a metodologia do CRISP-DM.

4.1.1 Entendimento do negócio

Busca-se ter uma visão clara do problema que se precisa resolver, é nesta fase que se deve traçar os objetivos do negócio, buscar mais detalhes do problema, listar os recursos disponíveis e o impacto esperado. Tem como características estabelecer métricas e os critérios quantitativos para os possíveis resultados. Priorizando aqueles que influenciam sua meta e também criar uma análise da vantagem do projeto, além do custo-benefício. Define-se os modelos, relatórios, apresentações e os dados.

4.1.2 Entendimento dos dados

Nesta segunda fase, se obtêm os dados e verifica-se se eles são adequados às suas necessidades. É importante ter feito uma boa fase 1, para que nesta fase não tenha que revisar o entendimento do negócio, nem repensar metas e planos.

Os objetivos desta fase são coletar os dados, descrevê-los, explorá-los e verificar a qualidade dos mesmos. Estabelecer formato para esses dados, é possível que seja necessário reunir novos dados, enfrentar limitações de software ou hardware. E encontrar imperfeições nos dados.

Na parte da documentação é importante estabelecer o *feature selection*, especificar os campos relevantes e criar uma descrição geral dos dados que possui, assim como os formatos, variáveis, técnicas estatísticas e qualquer informação que possa ser relevante.

É o lugar para criar, testar e documentar hipóteses geradas após a exploração dos dados.

4.1.3 Preparação dos dados

Nesta fase que os cientistas de dados passam a maior parte do seu tempo, agora que a maioria dos dados usados já foram coletados, necessita de refinamento antes de ser usado na modelagem. Esta fase possui cinco principais tarefas:

1. **Selecionar os dados:** É o momento de justificar quais dados serão ou não utilizados, documentar a relevância desses para seu objetivo, os problemas técnicos,
2. **Limpar esses dados:** Corrigir alguns dados específicos, excluir ou substituir por valores padrões para uma técnica de modelagem mais sofisticada.
3. **Documentar** bem detalhadamente os processos utilizados nesta etapa e o possível impacto gerado por essa escolha
4. **Construção dos dados:** Criar campos e documentá-los explicando os motivos.
5. **Integração dos dados:** Agora você provavelmente terá diversos conjuntos de dados e precisará mesclá-los e prepará-los para a fase de modelagem. Formatar os dados para o formato mais conveniente para o projeto.

4.1.4 Modelagem

Nesta fase serão escolhidas as técnicas mais adequadas para modelagem, ou seja, está etapa envolve a seleção e a utilização de técnicas e algoritmos que atendam as necessidades do negócio. Geralmente os dados são divididos em duas partes: um de treino (que são gerados os modelos) e um de teste (que se refere a validação do modelo). Com base nisso, é definido se continua o desenvolvimento da modelagem (avaliação) ou se retorna para a fase de preparação de dados.

4.1.5 Avaliação

Nesta fase será avaliada a qualidade e a segurança dos resultados obtidos na etapa anterior. De modo que seja possível verificar se esse resultado corresponde às expectativas do projeto. Caso não atenda, devem ser realizadas as modificações

necessárias (como correção na entrada de dados, correção no tratamento dos atributos, entre outros).

4.1.6 Implementação

Nesta fase é realizada o desenvolvimento dos modelos criados e avaliados. Durante essa etapa são realizadas tarefas, como: implantação da solução, monitoramento e manutenção, geração de relatórios e avaliação os resultados finais. Vale ressaltar que essa forma de implementação depende do tipo de modelo e projeto. Além disso, é preciso que o usuário final consiga interpretar e operar o produto com facilidade.

4.2 Ferramentas

As ferramentas utilizadas para a construção da solução, consiste em aquelas utilizadas para o desenvolvimento, organização e compartilhamento de arquivos. Primeiramente, definiu-se uma ferramenta para a organização, tendo como base o aplicativo Notion, que permite organizar, através de cards, todas as tarefas da equipe, sendo possível visualizar o que está sendo feito pelos integrantes e gerenciar as entregas já concluídas. Em paralelo a isso, tem-se a ferramenta de desenvolvimento. Para isso, utilizou-se o Google Collaboratory, onde criou-se o notebook do projeto, o qual é utilizado para criação, organização e execução do código.

As ferramentas de compartilhamento de arquivos. Para os arquivos de desenvolvimento do trabalho, é utilizado o Google Drive, que possui integração com o Google Collaboratory. Assim, sendo possível compartilhar em tempo real os arquivos referentes ao desenvolvimento. E por fim, é utilizado o Github, que possibilita compartilhar todos os arquivos do projeto, referente a descrição, organização e desenvolvimento em um ambiente que será possível ter uma visão ampla do que foi desenvolvido.

5. Compreensão do Problema

Apresenta-se nessa sessão as descrições das análises voltadas ao desenvolvimento de resultados do projeto, para a Everymind, a respeito da construção de um modelo preditivo para turnover de funcionários. Sendo exibido as identificações do mercado e produtos em comparação a solução prevista.

5.1 Análise da Indústria (5 Forças)

O contexto da indústria é utilizado para a empresa visualizar seu posicionamento no mercado, independente do seu tamanho e nicho de atuação. Abaixo encontra-se a análise prevista para a companhia Everymind.

I. Ameaça de novos entrantes:

Pode-se identificar como novos entrantes para a solução, os seguintes casos:

A) Outras empresas que dão suporte a programas de gerenciamento começarem a atender a Salesforce, possuindo como barreira a necessidade de obterem selo parceiro Salesforce;

B) Empresas parceiras Salesforce em outros países, que podem expandir sua operação para território nacional, possuindo como barreira a alta taxa de burocracia e regulamentação dentro do país;

C) A própria Salesforce (caso ela abra um setor onde as pessoas possam solicitar funções e suporte personalizado), possuindo como barreira colocar em risco a relação com parceiros Salesforce.

II. Serviços substitutos:

Pode-se identificar como serviços substitutos para a solução, os seguintes casos:

A) A própria Salesforce (caso ela adicione ao programa base serviços ou funcionalidades que a Everymind desenvolve);

B) Programas de planilhas que ajudam na gestão das empresas;

C) Outras plataformas de gestão empresarial (que não são a Salesforce) e as empresas que dão suporte a elas.

III. Poder de barganha dos consumidores:

Pode-se identificar como Poder de barganha dos consumidores para a solução, os seguintes casos:

A) Exigência de alta qualidade de software, uma vez que a Everymind está associada a Salesforce, que tem essa característica associada a sua imagem;

B) Negociação de preço, principalmente pelo fato de os serviços serem personalizados, ou seja, diferente para cada cliente, e várias outras empresas oferecem os mesmos serviços;

C) Negociação de tempo de entrega, já que outras empresas podem oferecer o mesmo serviço em menor tempo.

IV. Poder de barganha dos fornecedores:

Pode-se identificar como Poder de barganha dos fornecedores para a solução, os seguintes casos:

A) Hospedagem de programa (aumento de custos de funcionamento);

B) Plano de internet (aumento de custos de funcionamento);

C) Programas necessários para criação de ambiente de desenvolvimento (aumento nos custos de funcionamento).

V. Rivalidade entre concorrentes:

No Total são 137 parceiros Salesforce autorizados a atuar no Brasil, que podem oferecer produtos concorrentes a eles, alguns dos mais relevantes. Pode-se identificar como Poder de barganha dos concorrentes para a solução, os seguintes casos:

A) Deloitte, uma das maiores parceiras Salesforce do mundo, principalmente na área financeira e bancaria, que presta alguns dos serviços da Everymind. Ela possui suporte e estrutura no Brasil;

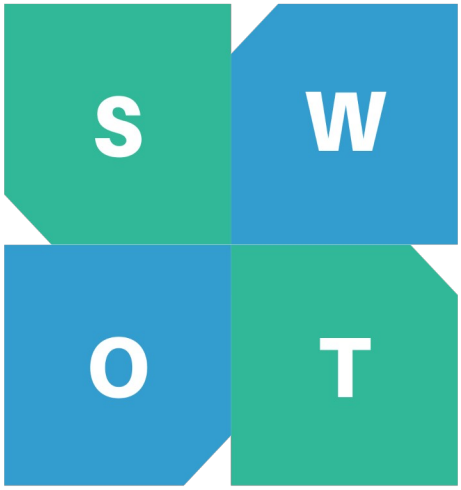
B) IBM, outra grande empresa parceira Salesforce com a permissão de atuar no Brasil, possui uma grande quantidade de serviços ofertados, que concorrem com a Everymind, e diversos prêmios de excelência;

C) DaspeWeb, também oferece consultoria para utilização da Salesforce e a implementação de novas funcionalidades, e assim como a Everymind e originária do Brasil.

5.2 Análise de cenário: Matriz SWOT

A análise SWOT é uma ferramenta que possibilita a empresa a realizar análises de cenário ou de ambiente, sejam eles internos ou externos. Assim, é demonstrado as formas como ela atua no setor, suas fraquezas, forças, oportunidades e ameaças. A Figura 1, exibe uma imagem demonstrativa das quatro áreas que compõem a SWOT.

Figure 1: Representação dos quatro pilares da matriz SWOT



Fonte: Autoria Própria

I. Pontos Fortes:

Pode-se identificar como pontos fortes para a solução, os seguintes casos:

A) A única empresa no Brasil que trabalha com todas as funções da plataforma SalesForce;

B) Aplicação da tecnologia de IA no gerenciamento de funcionários;

C) Horizontalidade da empresa;

D) Política de reconhecimento baseada no desempenho dos funcionários.

II. Pontos Fracos:

Pode-se identificar como pontos fracos para a solução, os seguintes casos:

A) Alta rotatividade de funcionários;

B) Alto gasto em recursos de onboarding;

C) Alto gasto de tempo e funcionários para o treinamento de novas pessoas;

D) Pouco diferenciais em relação às outras empresas líderes de mercado.

III. Oportunidades:

Pode-se identificar como oportunidades para a solução, os seguintes casos:

A) Alto valor agregado em serviços na área de tecnologia/SalesForce;

B) Ausência de tecnologias que auxiliam na governança corporativa em outras empresas do mercado;

C) Demanda de mercado pela criação de um ambiente e funções personalizadas dentro da Salesforce;

D) Alto crescimento e preferência por serviços online, o que promove uma maior visibilidade da empresa no mercado.

IV. Ameaças:

Pode-se identificar como ameaças para a solução, os seguintes casos:

A) Falta de profissionais na área de tecnologia;

B) Grande rotatividade na área da tecnologia;

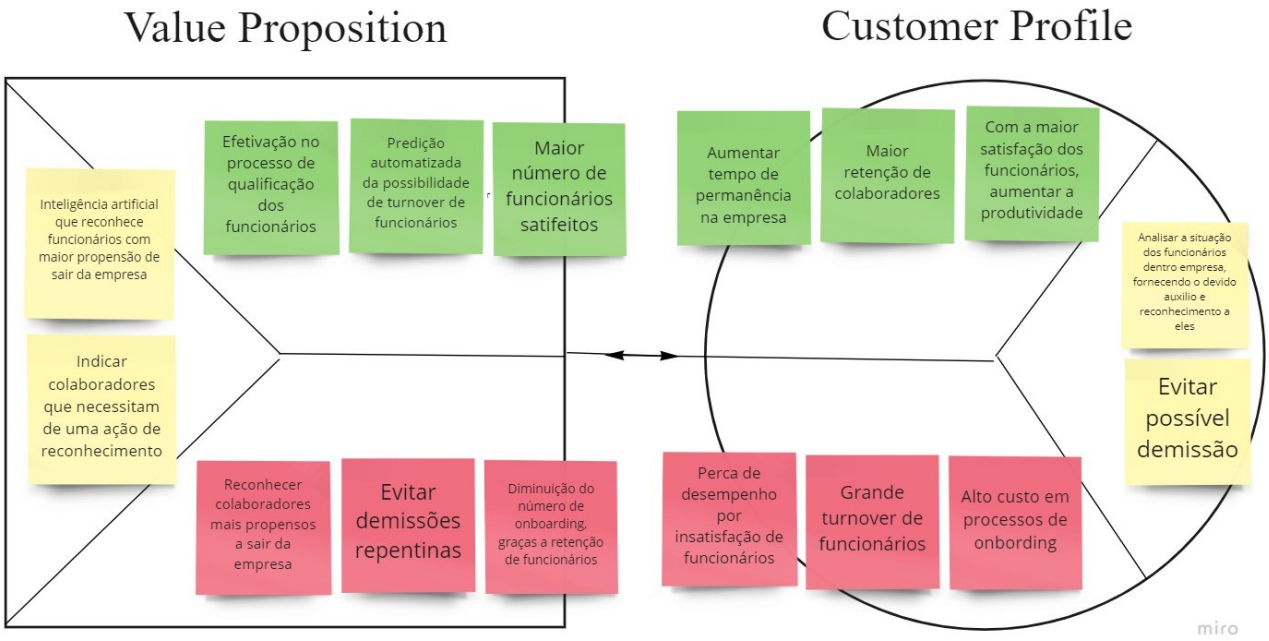
C) Falta de profissionais adaptados à mentalidade Everymind;

D) Maior abrangência de empresas brasileiras no setor.

5.3 Proposta de Valor

A principal vantagem apresentada pela proposta de valor é conseguir auxiliar a empresa a compreender melhor os seus clientes e funcionários. Na Figura 2, é ilustrada a proposta construída para a Everymind.

Figure 2: Proposta de valor prevista para a solução

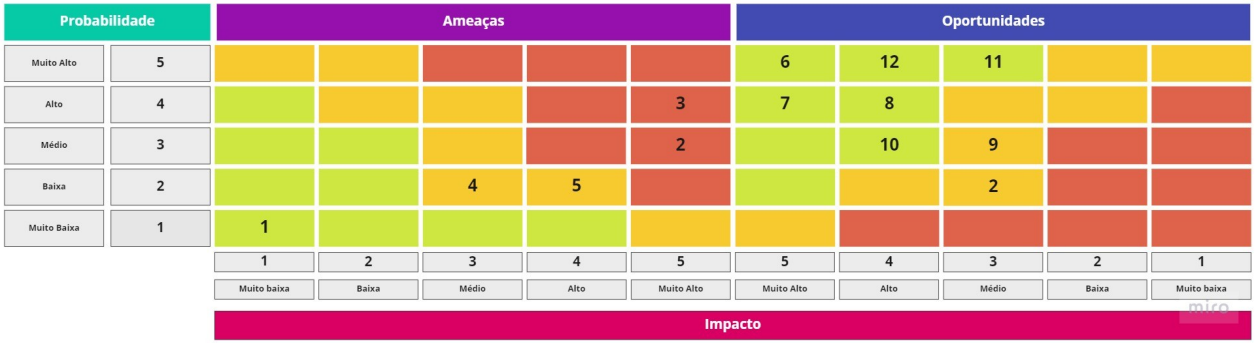


Fonte: Autoria Própria

5.4 Matriz de Risco

É uma das principais ferramentas na análise de negócios, utilizada para o gerenciamento de riscos operacionais existentes na empresa. A Figura 3, ilustra a construção da matriz de risco para o projeto.

Figure 3: Matriz de risco prevista para o projeto



Fonte: Autoria Própria

Cada número exposto na imagem acima, representa um risco ou oportunidade vista para o projeto e o impacto que ele ocasionará. Na tabela 1 abaixo, é disponibilizado a descrição de cada item:

Table 1: Descrição dos números apresentados na matriz de risco

Números	Descrições do risco
1	Cliente não aprovar nenhuma parte do projeto;
2	Modelo preditivo apontar resultados errôneos;
3	Falta de dados suficientes;
4	Poucas informações sobre o negócio;
5	Não atender a necessidade do cliente;
6	Reduzir a rotatividade de colaboradores;
7	Reduzir os gastos com a contratação de novos funcionários;
8	Evitar gastos contínuos com encargos trabalhistas;
9	Aumentar a produtividade do negócio;
10	Compreender as ineficiências da empresa;
11	Melhora do clima organizacional;
12	Reconhecer colaboradores que necessitam de reconhecimento;

Fonte: Autoria Própria

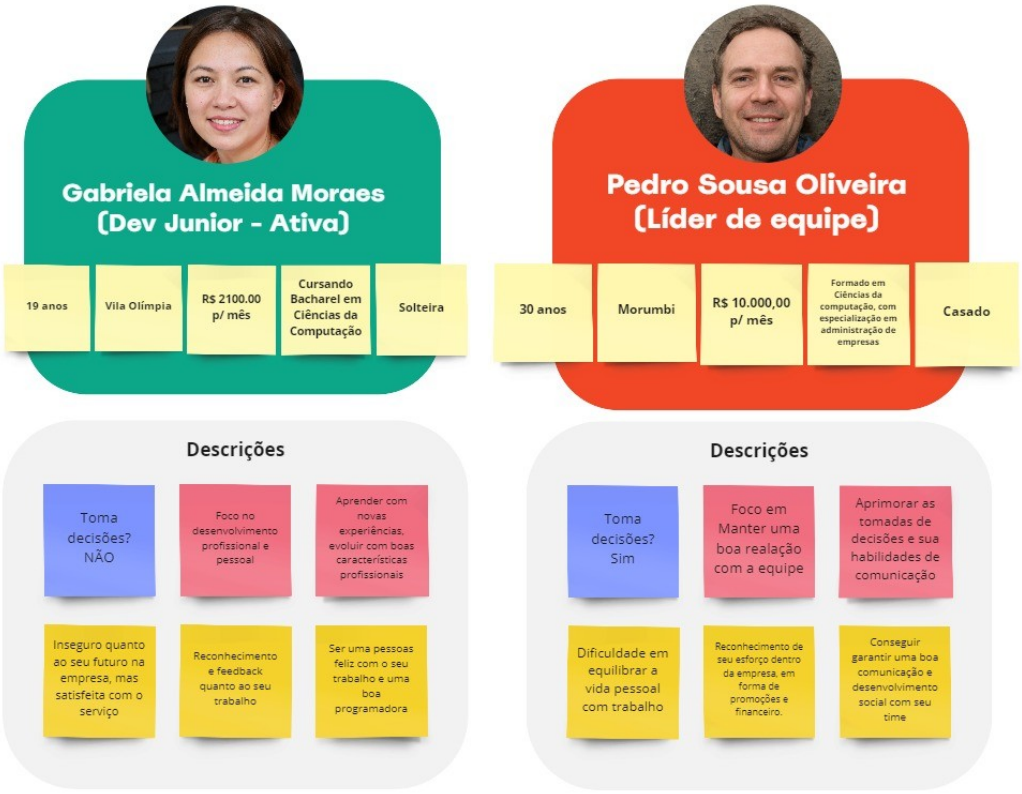
5.5 Personas

As personas do projeto são baseadas em dois setores principais, sendo eles, dois colaboradores da empresa, gerente de *‘people’* e o líder de equipe. Estes representam a ideia de cliente ideal, porém fictícia, e os dados apresentados (comportamentos e características), são equivalentes a o contexto em que a empresa se encontra. As Figuras 4 e 5, exibem as personas construídas.

Figure 4: Personas - Colaborador 1 e Gerente de 'people'

Fonte: Autoria Própria

Figure 5: Personas - Colaborador 2 e Líder de equipe



Fonte: Autoria Própria

5.6 Jornada do usuário

A jornada do usuário construída consiste na representação das etapas principais que envolvem o relacionamento entre os colaboradores, chefes de equipe e gestores de pessoas, dentro da empresa. Nesse sentido, encontra-se detalhado possíveis motivos que levam as pessoas a saírem ou ficarem dentro da corporação em questão. São divididas em quatro estruturas, exibidas nas figuras 6, 7, 8 e 9, sendo elas respectivamente:

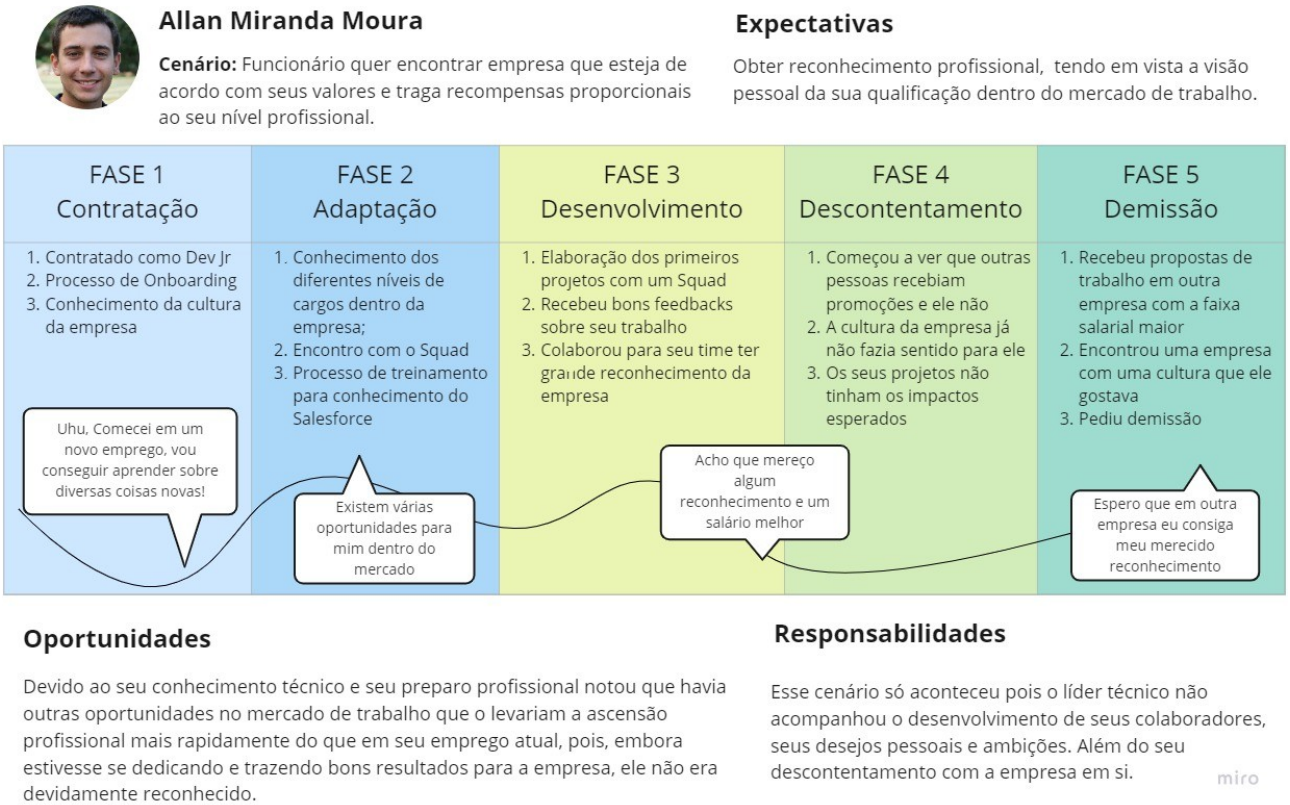
I. Dev Júnior que deseja sair da empresa;

II. Dev Júnior que deseja ficar na empresa;

III. Líder de equipe auxilia na decisão final;

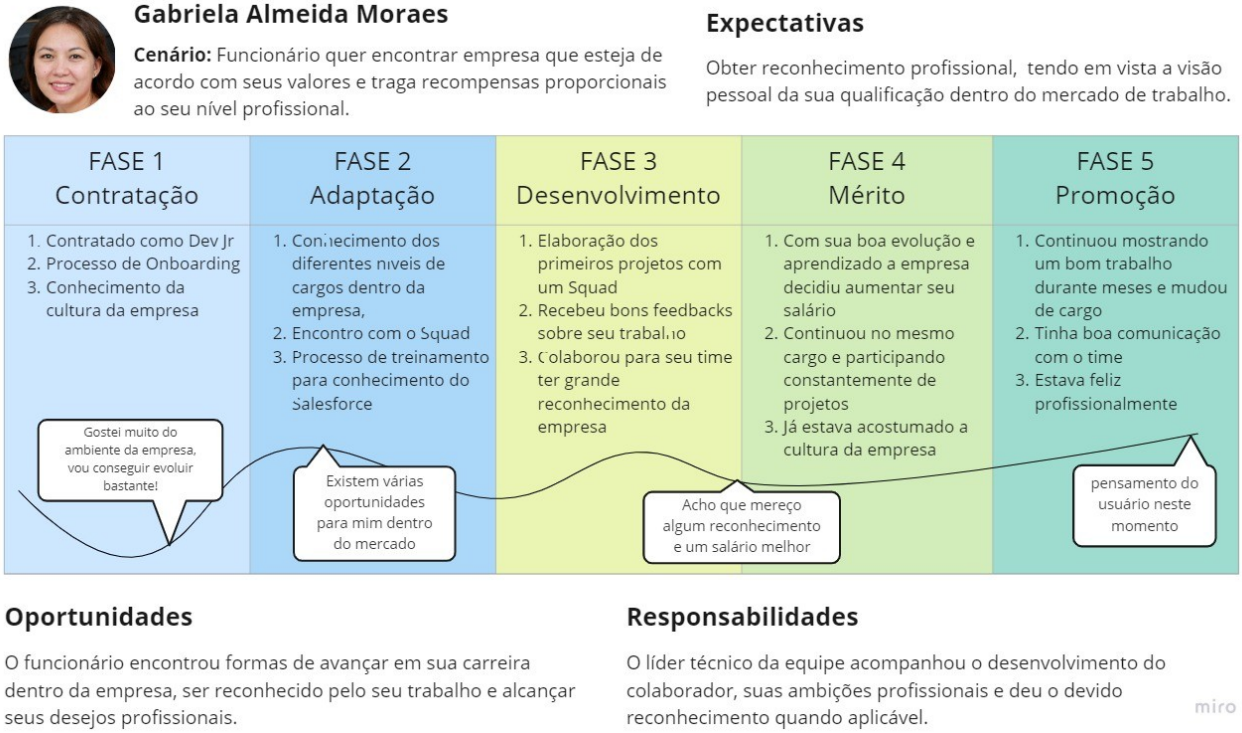
IV. Gerente de *'people'* que toma a decisão final;

Figure 6: Jornada do Usuário - Dev Júnior que deseja sair da empresa



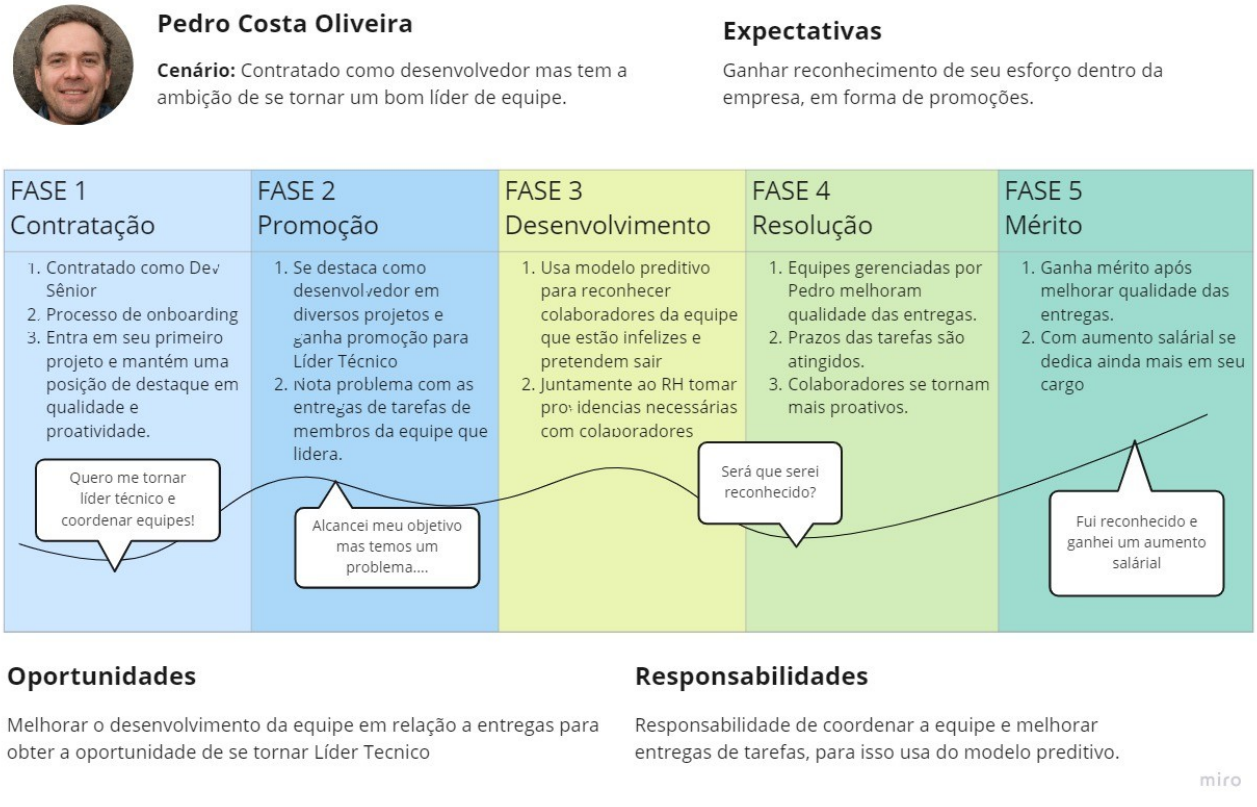
Fonte: Autoria Própria

Figure 7: Jornada do Usuário - Dev Júnior que deseja ficar na empresa



Fonte: Autoria Própria

Figure 8: Jornada do Usuário - Líder de equipe que auxilia na decisão final



Fonte: Autoria Própria

Figure 9: Jornada do Usuário - Gerente de 'people' que toma a decisão final

Fonte: Autoria Própria

6. Compreensão dos dados

As sessões abaixo apresenta o conjunto de dados trabalhado, seus principais atributos, descrições e análises esteatíticas.

6.1 Descrição dos dados utilizados

Neste tópico apresenta-se os dados disponibilizados na *"Base Colaboradores Everymind Inteli 2020 a 2022 Modelo Preditivo"*. A base de dados a ser trabalhada foi disponibilizada pela empresa "Everymind", tendo uma base de análise de 475 funcionários. A tabela 2 abaixo, é descrita os principais atributos, suas descrições e tipos da planilha "Geral Everymind".

Table 2: Descrição dos atributos - Planilha Geral Everymind

Planilha Geral - Funcionários		
Atributo	**Descrição**	**Tipo**
Matrícula	Registro do funcionário na empresa	Número - inteiro
Nome Completo	Nome do Funcionário colaborador	String + inteiro
Dt Admissão	Data de admissão do funcionário	dd/mm/yy - data
Dt Saída	Data de saída do funcionário	dd/mm/yy - data
Tipo Saída	Descrição do desligamento do colaborador	Strings
Gênero	Identidade de gênero dos funcionários	String ou bool
Salário Mês	Salário que o funcionário ganha mensalmente	Float
Dt Nascimento	Data de Nascimento do funcionário	dd/mm/yy - data
Etnia	Identificação étnica dos funcionários	String
Estado Civil	Estado civil dos funcionários	String
Escolaridade	Nível de ensino mais recente dos funcionários	String
Estado	Estado que o funcionário reside atualmente	String
Cidade	Cidade que o funcionário reside atualmente	String
Área	Área de atuação no mercado de trabalho	String

Fonte: Autoria Própria

A seguir, apresenta-se os principais dados acerca da planilha de reconhecimento, a qual é utilizada para visualizar quais colaboradores receberam promoção ou mérito no período de 2020 a 2022. Podendo se correlacionar, como uma alteração de cargo ou salário, afeta a permanência do funcionário na empresa. Na tabela 3 abaixo, é descrita os principais atributos, suas descrições e tipos da planilha “Reconhecimento”.

Table 3: Descrição dos atributos - Planilha Reconhecimento

Reconhecimento		
Atributo	**Descrição**	**Tipo**
Matrícula	Registro do funcionário na empresa	Número (Int)
Codinome	Nome do Funcionário colaborador	Número(Int) - string
Situação - Afastado	Situação do funcionário na empresa	String
Situação - Ativo	Situação do funcionário na empresa	String
Situação - Desligado	Situação do funcionário na empresa	String
Data de Admissão	Data de admissão dos funcionários	dd/mm/yy - Data
Data de vigência	Data de promoção do funcionário	dd/mm/yy - Data
Novo cargo	Cargo de promoção do funcionário	String

Fonte: Autoria Própria

Abaixo se apresenta a planilha Ambiente de Trabalho, com os principais atributos contemplados, estes são relacionados a uma pesquisa de satisfação da empresa para os funcionários, medindo o quão agradável é conviver e trabalhar nesse ambiente. Na tabela 4 abaixo, é descrita os principais atributos, suas descrições e tipos.

Table 4: Descrição dos dados - Planilha Ambiente de trabalho

Ambiente de Trabalho		
Atributo	**Descrição**	**Tipo**
Divisão	área do funcionário na empresa	string
Pilar	Tópico da pergunta da pesquisa	string
Pontuação	Nota que o funcionário fornece para a empresa	float
Fator	Subtópico da pergunta da pesquisa	string
Pontuação	Nota que o funcionário fornece para a empresa	float
Pergunta	Pergunta que é feita ao funcionário	string
Níveis de satisfação	Muito insatisfeito, insatisfeito, neutro, satisfeito, muito satisfeito	string
Taxa de confiabilidade	nível de veracidade das respostas	string

Fonte: Autoria Própria

Na tabela 5 abaixo se encontra a descrição dos principais atributos, suas descrições e tipos, presentes nas três planilhas.

Table 5: Descrição dos atributos - Três Planilhas

Dados Gerais - Utilizados nas 3 planilhas		
Atributo	**Descrição**	**Tipo**
Escolaridade	1) Ensino médio completo ou incompleto; 2) Técnico; 3) Graduação; 4) Ensino Superior; 5) Mestrado; e 6) Pós-Graduação.	String
Estado Cívil	1) Casado; 2) Divorciado; 3) Separado; 4) Solteiro; e 5) União Estável.	String
Etnia	1) Amarela; 2) Branca; 3) Parda; 4) Preta; e 5) Não Informada.	String
Cargos	Refere-se a todos os 36 possíveis cargos a se ter dentro da empresa;	String
Área de Atuação	Refere-se a todos os 23 possíveis áreas de atuação a se ter dentro da empresa;	String

Fonte: Autoria Própria

6.2 Descrição dos conjuntos de dados

I. Descrição de como os dados serão agregados/mesclados:

Foram disponibilizados dois conjuntos de dados, sendo eles: 1) As informações sobre os funcionários ativos demitidos; 2) Reconhecimento de promoções e méritos de cargo. Uma das possíveis mesclagens de dados, podem ser feitas através da junção de como as promoções afetam a saída dos funcionários na empresa.

II. Descrição dos riscos e contingências relacionados aos dados:

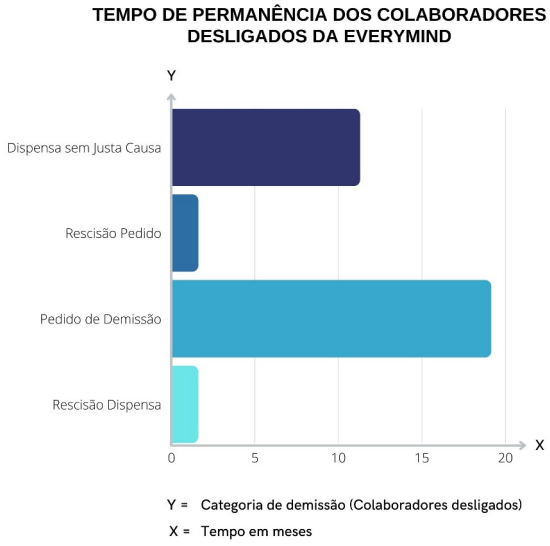
Os dados não oferecem grandes riscos de falta de confiabilidade, isto se deve ao fato de todos os dados serem coletados e disponibilizados pela própria empresa. Ou seja, a chance de serem falsos ou imprecisos é extremamente baixa, portanto de alta qualidade. Eles cobrem todos os aspectos que o parceiro considerou importante para o desenvolvimento do projeto, já que eles selecionaram os dados a serem repassados. Em quesito diversidade os dados são referentes às informações sobre cada funcionário, a única limitação é a pesquisa de satisfação que não é informado a resposta individual de cada colaborador, somente ao percentual geral.

6.3 Descrição estatística básica dos dados

Tempo de permanência

A análise da média de tempo em que o funcionário permanece na empresa, se dá mediante a data de saída, menos a data de admissão do colaborador. Resultando em uma média em meses de quando uma demissão é feita. Este dado será utilizado para verificar quanto o tempo de permanência do funcionário impacta na forma de saída dele. Na Figura 10 é ilustrado a construção do gráfico com estes dados. EX: Com o gráfico observa-se que a média de permanência de pessoas que pedem demissão é de 1 ano e meio.

Figure 10: Gráfico - Tempo de permanência

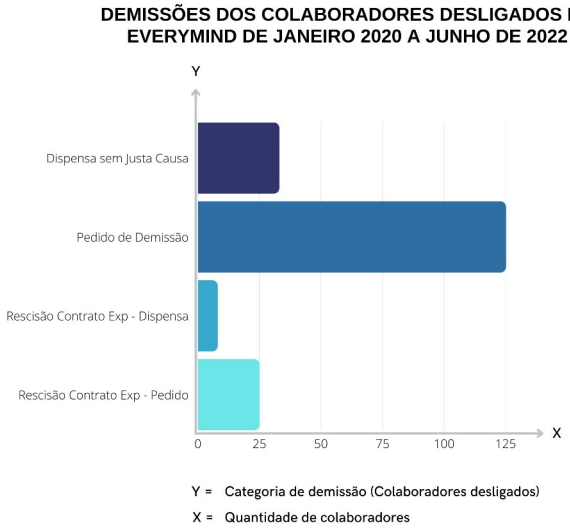


Fonte: Autoria Própria

Demissões

O gráfico abaixo, representado na figura 11, é responsável pela análise das demissões dos colaboradores com base nos dados fornecidos, durante o período de janeiro de 2020 a julho de 2022. Utilizado para buscar padrões para guiarem o modelo preditivo, além de mostrar se a empresa possui maior pedidos de demissões ou demissões por justa causa, levando a observar uma maior correlação entre tal atributo e outras varianças nos dados da companhia.

Figure 11: Demissões dos colaboradores desligados em relação ao tempo

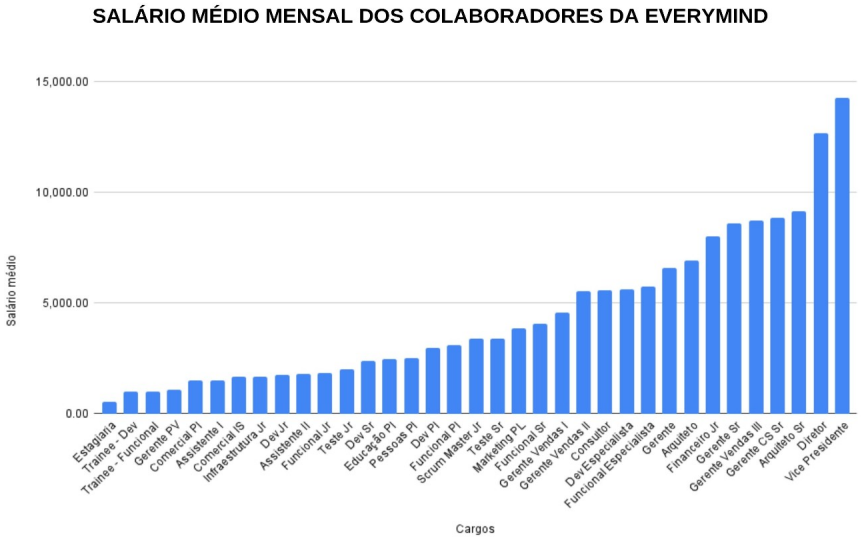


Fonte: Autoria Própria

Salário médio dos cargos

Esse gráfico apresenta o salário médio mensal dos colaboradores da Everymind por cargo. A média salarial dos arquitetos é de R$6.908,24, e um determinado arquiteto que ganhava, R$5.000,00 foi desligado da empresa. Esses dados serão utilizados para entender como esse aspecto influencia na permanência de um colaborador, comparando um determinado salário com a média do cargo. Na Figura 12, ilustra a criação do gráfico.

Figure 12: Relação de salário com cargo dos colaboradores

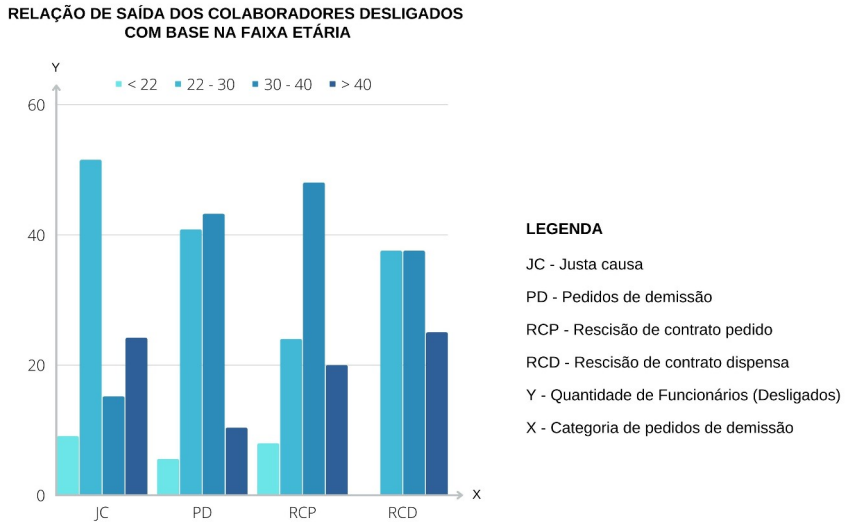


Fonte: Autoria Própria

Relação de saída dos colaboradores com base na faixa etária do colaborador:

Abaixo, apresenta-se na Figura 13, o gráfico entre a relação de saída dos colaboradores, baseadas em: 1) Justa causa; 2) Pedidos de demissão; 3) Pedidos de rescisão de contrato; e 4) Rescisão de contrato em si dos funcionários. Nas primeiras quatro barras à esquerda é possível observar que cerca de 50% dos desligamentos por Justa Causa concentram-se nas faixas etárias de 22-30 anos. Em seguida, são representados os pedidos de demissão onde é possível verificar que as faixas de 22-30 anos e 30-40 anos possuem as maiores taxas de pedidos de demissão, com cerca de 43.2% e 40.8% dos pedidos respectivamente. Logo após, os pedidos de rescisão de contrato são apresentados em que é possível observar que cerca de 48% dos pedidos de rescisão concentram-se na faixa etária dos colaboradores de 30-40 anos. Por último, tem-se representado as dispensas por rescisão de contrato, em que é possível verificar, que tal desligamento concentra-se nas faixas etárias de 22-30 anos e 30-40 anos. A partir da análise dos dados dispostos nesse gráfico, pode-se encontrar padrões para se compreender melhor as taxas de Turnover de acordo com as idades dos funcionários e o tipo de desligamento para poder, por fim, levar a um modelo preditivo de aprendizado da máquina.

Figure 13: Relação saída e causa

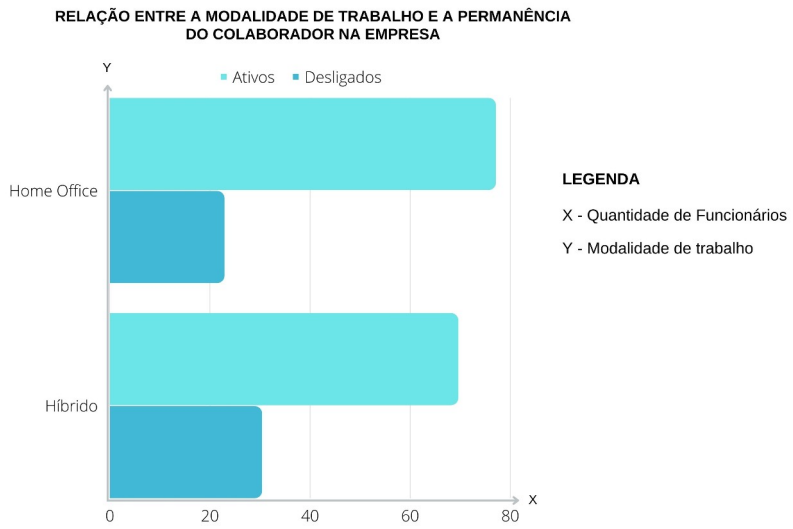


Fonte: Autoria Própria

Modalidade (Home Office, Híbrido) X Demissões

A análise da modalidade pela demissão de funcionários, se dá com base nos colaboradores que residem fora da cidade de São Paulo, com a modalidade de trabalho em Home Office, e nos que residem na cidade, trabalhando na modalidade Híbrida. Exibindo quais foram desligados da empresa e quais permanecem ativos. Com esses dados, espera-se visualizar qual modelo de trabalho retém mais funcionários na empresa. Exemplo: Em comparação com o modelo híbrido, o home office ocasiona menos pedidos de demissão, como apresentado na Figura 14.

Figure 14: Relação modalidade de trabalho e pedido de demissão



Fonte: Autoria própria

6.4 Preparação dos dados

Nesta seção, vamos colocar em prática a análise exploratória dos dados fornecidos pela empresa, de acordo com a execução proposta pelo modelo CRISP-DM, a fim de tornar nossa base de dados mais adequada para o treinamento do modelo de predição do Aprendizado de Máquina. Assim, vamos tratar os dados a partir de funções que os modificam, após encontrar tendências e padrões qualitativos relacionados a esses, a fim de adequá-los para inserção em equações presentes no modelo preditivo e possibilitar o treinamento desses dados. Inicialmente, alguns dos dados que estão sendo preparados são relacionados a datas, tempo, nomes e categorias.

Exclusão de espaços em branco:

Para substituir os espaços em branco, todas as colunas do tipo *'object'* foram alteradas, tendo seus espaços (“ “) retirados (“”). Essa *Feature* foi selecionada para possibilitar o *One Hot Encoding* e *Label Encoder*, que necessita que todos os dados estejam padronizados para um bom funcionamento, prevenindo possíveis erros de digitações e sendo alocado para todas as strings das tabelas. Na Figura 15, é ilustrado um antes e depois da formatação com dois exemplos nas colunas “Escolaridade” e “Cidade”.

Figure 15: Comparação - Remoção de espaços em branco

ANTES

Escolaridade	Estado	Cidade
Superior incompleto	PR	Curitiba
Superior incompleto	PB	João Pessoa
Superior incompleto	SP	São Paulo
Graduação	SP	São Paulo

DEPOIS

Escolaridade	Estado	Cidade
Superiorincompleto	PR	Curitiba
Superiorincompleto	PB	JoãoPessoa
Superiorincompleto	SP	SãoPaulo
Graduação	SP	SãoPaulo

Fonte: Autoria Própria

Adição de valores nos campos sem informações:

A seleção dessa Feature foi selecionada pois os campos em branco em uma base de dados, resulta em problemas de qualidade dos dados apresentados. Esse quesito pode impactar no treinamento do modelo e nas conclusões a serem tiradas das análises realizadas. Já que algumas colunas apresentavam valores faltantes, a solução utilizada foi contemplar o tipo de saída esperada, e calcular as entradas prováveis das variáveis a serem manipuladas.

Neste caso, o método escolhido para tratar os valores ausentes/em branco foi a substituição, que foi realizada nas linhas com dados vazios através do método *'replace'*, para trocar o dado faltante pela data de hoje na coluna 'Dt Saída', valor esse que será utilizado para o cálculo do 'Tempo Empresa Meses' e 'Idade' nas Features (4 e 5), e por ativo na coluna 'Tipo Saída', para indicar se o funcionário ainda está ativo na empresa. Na Figura 16, pode-se notar como os valores foram adicionados, tendo como exemplo a coluna "Tipo de Saída".

Figure 16: Comparação - Adição de valores nas colunas em branco

ANTES

	Dt Admissao	Dt Saida	Tipo Saida
0	06/06/2022	NaT	NaN
1	14/02/2022	NaT	NaN
2	02/03/2022	NaT	NaN
3	02/12/2019	NaT	NaN

DEPOIS

	Dt Admissao	Dt Saida	Tipo Saida
0	2022-06-06	2022-08-26	Ativo
1	2022-02-14	2022-08-26	Ativo
2	2022-02-03	2022-08-26	Ativo
3	2019-02-12	2022-08-26	Ativo

Fonte: Autoria Própria

Formatação de datas:

Para a manipulação correta das datas e horários na base de dados, todas precisam estar no mesmo formato, sendo o modelo escolhido yyyy/mm/dd (Exemplo: 2003/05/30). As tabelas que tiveram seus campos alterados foram as Planilhas "Everymind" e "Reconhecimento". As colunas afetadas pela formatação são: 1) "Dt Admissao"; 2) "Dt

Nascimento”; 3) “Dt Saida”; 4) “Data de Admissão”; e 5) “Data Vigência”. Essa Feature foi selecionada pois sem a formatação das datas resultaria em um difícil manuseio dos dados. A Figura 17, ilustra o antes e o depois da formação, tendo como exemplo a coluna “Dt Admissao”.

Figure 17: Comparação - Formatação de datas

ANTES

Matrícula	Nome Completo	Dt Admissao
476.0	Pessoa Colaboradora 1	06/06/2022
373.0	Pessoa Colaboradora 10	14/02/2022
392.0	Pessoa Colaboradora 100	02/03/2022
110.0	Pessoa Colaboradora 101	02/12/2019

DEPOIS

Matrícula	Nome Completo	Dt Admissao
476.0	PessoaColaboradora1	2022-06-06
373.0	PessoaColaboradora10	2022-02-14
392.0	PessoaColaboradora100	2022-02-03
110.0	PessoaColaboradora101	2019-02-12

Fonte: Autoria Própria

Manipulação das idades:

Para obter dados mais significativos e em um melhor formato para serem analisados, foi derivado um novo atributo, ‘Idade’, a partir da coluna ‘Dt Nascimento’, assim convertendo a data de nascimento para a idade da pessoa. Esta Feature foi selecionada pelo fato de a idade dos funcionários ser um dado de extrema importância para a análise dos dados. Assim, sendo de grande impacto da idade na saída pretendida.

Neste caso, foi realizada a derivação de um novo atributo. Esta derivação foi realizada através da criação de uma nova coluna ‘Idade’, que no caso referente aos funcionários que ainda estão na empresa, esta coluna foi calculada através do cálculo da diferença entre a data de agora, obtida através do método *‘data.today()’* e a data de nascimento do funcionário. Já no caso referente aos funcionários que saíram da empresa, foi calculada a idade com que a pessoa saiu da empresa, através do cálculo da diferença entre a data de saída e a data de nascimento do funcionário. Na Figura 18, pode-se notar como a coluna ‘Idade’ foi derivada.

Fonte: Autoria Própria

Cálculo do Tempo de Empresa:

Para obter dados mais significativos e em um melhor formato para serem analisados, foi derivado um novo atributo, ‘Tempo Empresa’, a partir da coluna ‘Dt Admissao’ e ‘Dt Saida’, assim utilizando esses atributos para calcular o tempo de empresa desse funcionário. Esta Feature foi selecionada pelo fato de o tempo de empresa dos funcionários ser um dado de extrema importância para a análise dos dados. Assim, sendo possível utilizar esta informação para o cálculo da classificação do funcionário.

Neste caso, foi realizada a derivação de um novo atributo. Esta derivação foi realizada através da criação de uma nova coluna ‘Tempo De Empresa Meses’, no caso referente aos funcionários que ainda estão na empresa, esta coluna foi calculada através do cálculo da diferença entre a data de agora, obtida através do método ‘*data.today()’* e a data de admissão do funcionário. Já no caso referente aos funcionários que saíram da empresa, foi calculado o tempo de empresa até o momento que a pessoa sai dela, cálculo esse feito a partir da diferença entre a data de saída e a data de admissão do funcionário. Na Figura 19, pode-se notar como a coluna ‘Tempo Empresa Meses’ foi derivada.

Figure 19: Derivação - Tempo de empresa

Area	Idade	Tempo Empresa Meses
CPG&RetailI	37	2
Core&IndustriasII	23	6
AgenciaDigital	33	6
Core&IndustriasI	39	42

Fonte: Autoria Própria

Tempo Reconhecimento

Para obter um que dado útil para o sistema preditivo, o tempo foi derivado um novo atributo, ‘Tempo Ate Promocao Meses’, a partir da coluna ‘Data de Admissão'' e ‘Data de Vigência'', utilizando dessas colunas para calcular o tempo até o funcionário receber o reconhecimento ou promoção. Esse Feature foi selecionada pelo fato de ela ser capaz de

derivar um atributo do tempo até a promoção, dado este que pode ser utilizado para chegar na classificação desejada.

Neste caso, foi realizada a derivação de um novo atributo. Esta derivação foi realizada através da criação de uma nova coluna ‘Tempo Ate Promocao Meses’, que é referente ao tempo que o funcionário demorou para conseguir a promoção ou reconhecimento desde que ele entrou, esta coluna foi calculada através do cálculo da diferença entre a data da promoção e a data de admissão do funcionário. Na Figura 20, pode-se notar como a coluna ‘Tempo Ate Promocao Meses’ foi derivada.

Figure 20: Derivação - Tempo de promoção (em meses)

Novo Cargo	Tempo Ate Promocao Meses
Gerente Sr	81
Arquiteto	78
Arquiteto	69
Arquiteto	65

Fonte: Autoria Própria

Criação de novo atributo (Separação do número com o nome do colaborador)

A seleção dessa Feature foi necessária pois a coluna “Nome Completo” na planilha “Everymind” e a “Codinome” na planilha “Reconhecimento”, apresentavam formatos inutilizáveis já que o nome de um colaborador não é um fator relevante para ele ser mandado embora da empresa. Assim, cria-se um novo atributo chamado “Colaborador”, que divide o texto da célula entre a palavra “Pessoa colaboradora” e o número que a acompanha. Resultando somente em números que são responsáveis pela identificação desses funcionários. Na Figura 21, é exemplificado o antes e depois da coluna “Nome Completo” e “Colaborador”.

Figure 21: Separação - Tipos String e Number

ANTES

	Matrícula	Nome Completo
0	476.0	PessoaColaboradora1
1	373.0	PessoaColaboradora10
2	392.0	PessoaColaboradora100
3	110.0	PessoaColaboradora101

DEPOIS

Idade	Tempo Empresa Meses	Colaborador
37	2	1
23	6	10
33	6	100
39	42	101

Fonte: Autoria Própria

Exclusão de Colunas não utilizadas

A partir da análise dos dados foi decidido pela retirada de algumas colunas da "Base Colaboradores Everymind", sendo elas as colunas "Etnia", "Nome Completo" e "Codinome". Essa Feature foi selecionada pois, em primeiro lugar a coluna "Etnia" foi motivada pela sensibilidade dos dados e ser antiético a análise da permanência de colaboradores a partir da etnia destes. Nesse prisma, a continuidade dessa coluna criará um modelo com resultados enviesados. Já a retirada da coluna "Nome Completo" e "Codinome" ocorreu por esta não contribuir de forma alguma com a construção do modelo, uma vez que, um nome não pode ser um fator de decisão.

One Hot Encoder

A Feature foi selecionada pois para utilizarmos as variáveis categóricas é necessário realizar uma transformação nos dados, que resultam em formas binárias (não ordenada), as quais serão aplicadas em futuras equações matemáticas no modelo de aprendizado de máquina. Nesse aspecto, fez-se necessário a criação de um data frame, que seleciona a coluna especificada que corresponde às propriedades (campos) da base de dados e suas linhas são identificadas como um registro. A Figura 22, ilustra o exemplo da coluna “Estado Civil” antes e depois da formatação.

Figure 22: Comparação - One Hot Encoder

ANTES

Estado Civil
Casado
Solteiro
Solteiro
Divorciado

DEPOIS

	Casado	Divorciado	Separado	Solteiro	UniãoEstável
0	1	0	0	0	0
1	0	0	0	1	0
2	0	0	0	1	0
3	0	1	0	0	0

Fonte: Autoria Própria

Criação novo Database

Para trabalhar melhor cruzando variáveis de dados de funcionários que ainda estão presentes na empresa e que já não estão mais presentes, houve a criação de duas bases de dados com a divisão dessas entre atributos focados nos funcionários ativos e desligados. Essa feature se mostra fundamental para melhor manipulação e análise exploratória dos dados. Com a organização dos DataFrames a partir dessas bases, haverá uma maior compreensão dos dados a fim de explicitar tendências para futuramente levá-las à modelagem preditiva que revelará a possibilidade de desligamento e retenção dos funcionários por meio do aprendizado de máquina.

Análise de colunas

Com a feature da "Criação nova Database" a Database original foi separada em funcionários ativos e desligados. A partir disso foi feita uma nova feature da análise da quantidade de funcionários ativos e desligados. Tal análise foi fundamental pois assim não trabalhamos com dados enviesados, dessa maneira conseguimos relacionar a quantidade de se existem dados desproporcionais como, por exemplo, mais funcionários desligados do que ativos em determinado cargo. A Figura 23, exibe uma exemplificação da relação entre os ativos e os desligados com os 9 primeiros cargos.

Figure 23: Relação - Funcionários Ativos e Desligados

ATIVOS

Trainee-Dev	98
DevPl	72
DevJr	60
FuncionalPl	39
DevSr	32
DevEspecialista	27
Gerente	27
FuncionalJr	23
FuncionalSr	18
Arquiteto	11

DESLIGADOS

DevPl	47
DevJr	29
DevSr	17
Trainee-Dev	16
FuncionalPl	15
DevEspecialista	15
FuncionalJr	12
FuncionalSr	10
Arquiteto	7

Fonte: Autoria Própria

7. Modelagem

As sessões abaixo apresenta os algoritmos escolhidos para teste do modelo preditivo, suas descrições, principais funções e exemplificações.

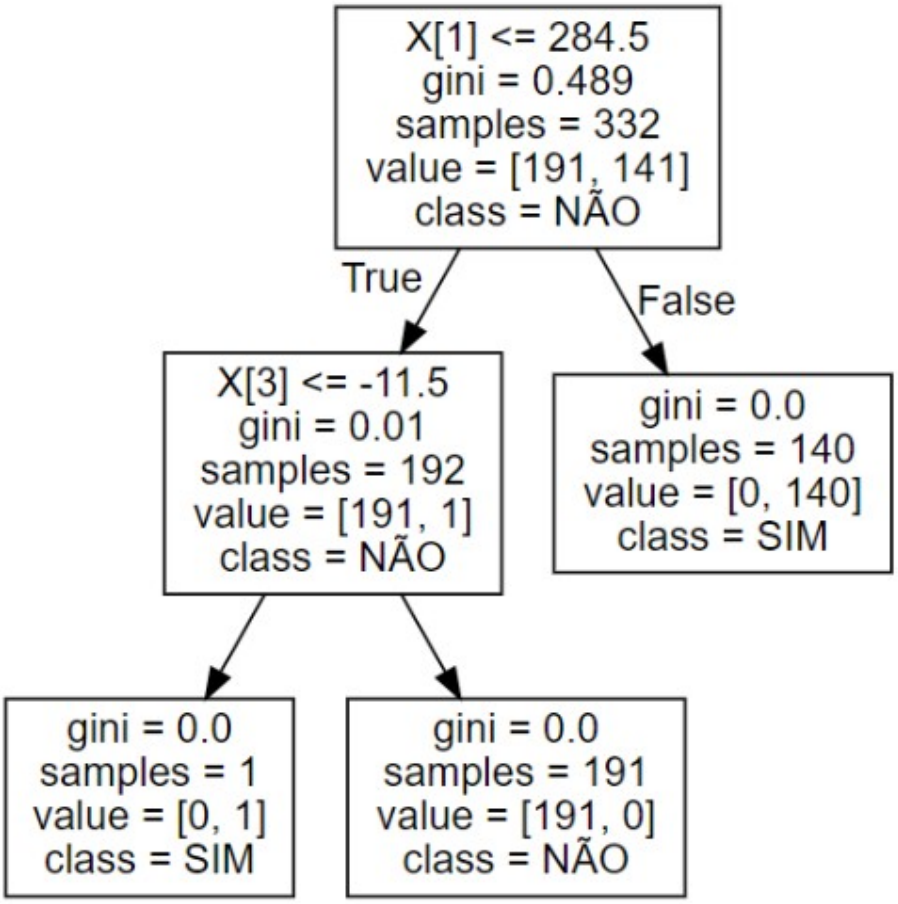
7.1 Árvore de decisão

Como diz o nome, o algoritmo de árvore de decisão cria vários pontos de decisão para encontrar a solução do problema. Os pontos são conhecidos como “nós” e cada um deles possui decisões a serem tomadas. Os caminhos existentes na árvore de decisão são conhecidos como “ramos”. O objetivo do algoritmo é aprender as regras básicas e assim conseguir obter o resultado. Na Figura 24, é ilustrado um exemplo de uma árvore de decisão e na Figura 25, a ilustração da árvore criada para o modelo preditivo:

Figure 24: Exemplo - Árvore de decisão

Fonte: Autoria Própria

Figure 25: Ilustração - Árvore de decisão do modelo preditivo



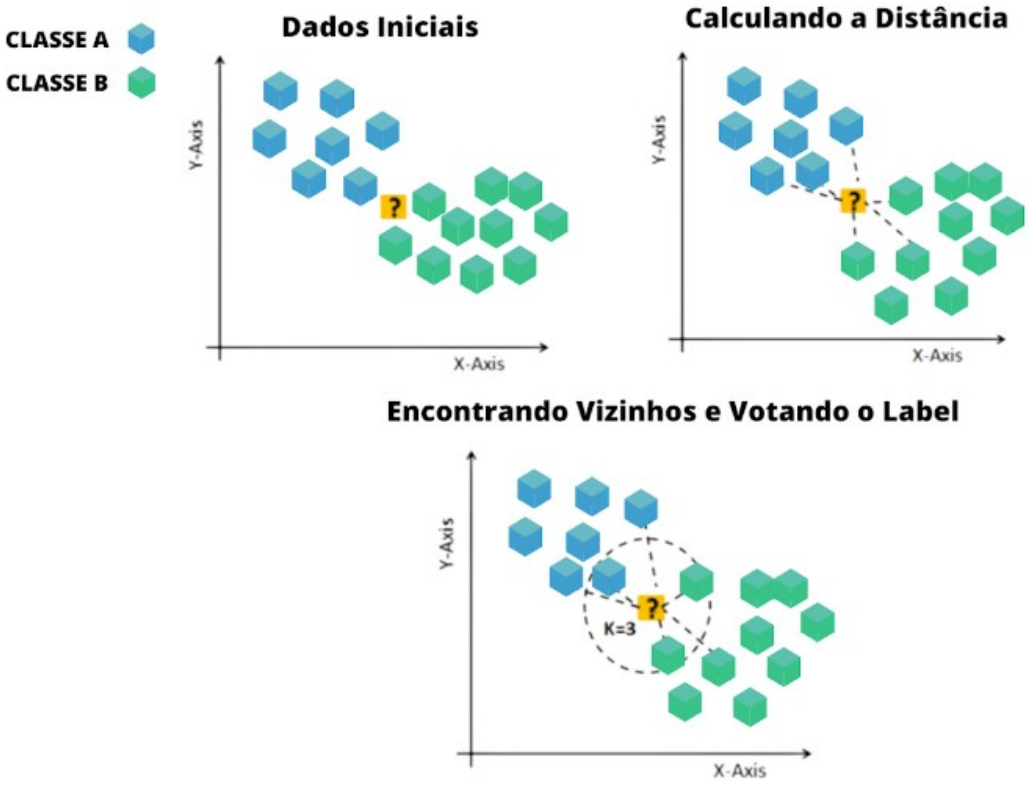
Fonte: Autoria Própria

O algoritmo de Árvore de decisão adequa-se a solução desenvolvida pois com ela é possível realizar a classificação dos dados, atribuindo-os os rótulos propostos pela variável alvo, “Sim” para probabilidade de sair e “Não” para pouca probabilidade de sair, através de diversos nós de decisão que levam a resposta final.

7.2 K Nearest Neighbor (KNN)

O KNN (K Nearest Neighbor ou K-ésimo Vizinho Mais Próximo), também conhecido como algoritmo de aprendizado lento, não precisa necessariamente de dados de treinamento para a criação do algoritmo, o que gera um treinamento mais rápido dos dados, mas, em contrapartida, possui teste e validação lentos. Nesse algoritmo, temos um parâmetro K, o qual direcionará a quantidade de dados vizinhos mais próximos, e então, classificará a nova variável de acordo com a classe da maioria dos vizinhos mais próximos determinados por K. Na Figura 26 abaixo, é ilustrado um exemplo do algoritmo KNN:

Figure 26: KNN - Exemplo de algoritmo KNN



Fonte: Autoria Própria

O algoritmo KNN adequa-se ao problema proposto pela “Everymind”, pois como seu funcionamento é baseado na classificação de dados, atribuindo-os rótulos a partir de dados já classificados mais próximos e em maior quantidade, esse possui a capacidade de classificar novas entradas, no caso, novos colaboradores ou colaboradores ainda não avaliados, entre os rótulos propostos pelas variáveis alvo, as quais são representadas por “Sim” para a maior probabilidade de sair e “Não” para a maior probabilidade de retenção.

7.3 Naive Bayes

O algoritmo de Naive Bayes, utilizado no projeto, parte da premissa de calcular a probabilidade de algo ocorrer, sendo que outro evento já aconteceu. O algoritmo de Naive Bayes calcula a classificação mais adequada a partir da fórmula de Bayes, segue abaixo na figura 27, a fórmula utilizada para o cálculo do algoritmo.

Figure 27: Equação de Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Fonte: Autoria Própria

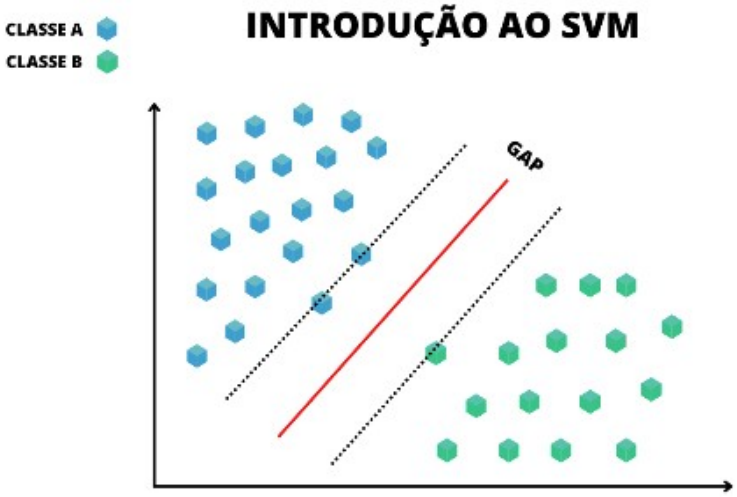
Na qual, P(A|B) – representa a probabilidade do evento A acontecer sendo que b já ocorreu. e P(B|A) – representa a probabilidade do evento B acontecer, sendo que A já ocorreu. Assim, sendo possível fornecer classificações com base na probabilidade de elas ocorrerem em decorrência de uma variável que já ocorreu.

O algoritmo de Naive Bayes foi adequado ao problema que está sendo trabalhado. Isto se deve ao fato de ele ser recomendado para algoritmos classificatórios, que é o caso do projeto. Assim, levando em conta as classificações possíveis em nosso sistema e as variáveis relacionadas a elas, disponibilizadas no treino, e a probabilidade da classificação em decorrência da variável. Foi possível a utilização do algoritmo de Naive Bayes para estimar a classificação dos funcionários.

7.4 Support Vector Machine (SVM)

A definição do Support Vector Machine(SVM), pode ser dada por um algoritmo que visa encontrar o hiperplano de separação ideal para os dados propostos, sendo o seu maior objetivo a maximização das distâncias das variáveis deixando-as o mais definidas possível. Este tende a ser mais complexo que o KNN e apresentar resultados mais estruturados. O hiperplano de separação utilizado para as análises pode ser descrito como uma linha, que passa entre os dados, tentando delimitar uma separação dos atributos selecionados, como visto na Figura 28 abaixo:

Figure 28: SVM - Exemplo de modelo



Fonte: Autoria Própria

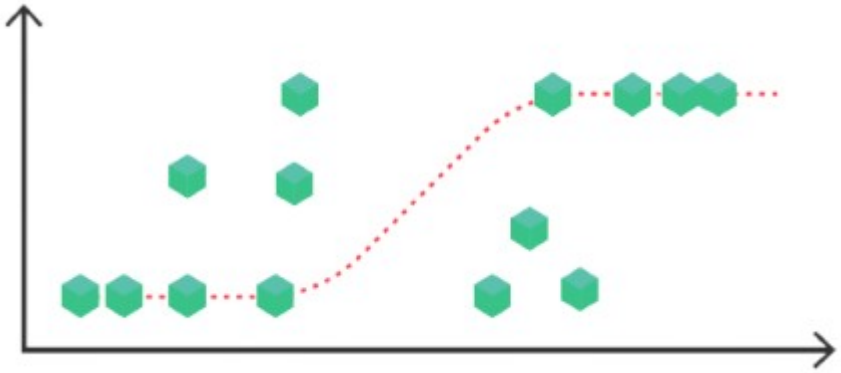
O hiperplano utilizado é basicamente a generalização de um plano qualquer, com mais de três dimensões. Visto isso, o objetivo primordial do SVM é conseguir traçar mediante os dados manipulados o hiperplano de separação ideal, visando a classificação de maneira correta dos atributos.

O algoritmo SVM tem sua adequação a solução mediante a classificação dos dados escolhidos em categorias, sendo possível a visualização definida dos atributos referentes a variável alvo, resultando em “Sim” para probabilidade de sair e “Não” para pouca probabilidade de sair, a partir da análise do hiperplano citada anteriormente.

7.5 Regressão Logística

É um algoritmo estatístico, de aprendizado supervisionado que é usado para a classificação e análise preditiva. Ela estima a probabilidade de um evento ocorrer se apoiando em um conjunto de dados. Na figura 29, abaixo é apresentado um exemplo gráfico de regressão logística.

Figure 29: Exemplo de regressão logística



Existem três tipos de algoritmos para regressão logística, eles são definidos com base no resultado.

1. Regressão logística Binária: A variável tem apenas dois resultados possíveis, "0 e 1". Este é o algoritmo mais comumente usado.
2. Regressão Logística multinomial: Nesse algoritmo a variável possui três resultados possíveis, entretanto não possui uma ordem desses valores.

3. Regressão Logística Ordinal: este algoritmo é aplicado quando a variável possui 3 ou mais resultados possíveis, porém o resultado tem uma ordem já definida. Por exemplo, de "A-E" ou escalas de "1-5"

Como ela mede a relação entre uma variável alvo e outras variáveis independentes, utilizamos o algoritmo binário para identificar fatores importantes que impactam a nossa variável alvo: "Saiu da empresa" nos retornando "sim” ou “não" significando a saída ou permanência do funcionário na empresa.

8. Avaliação do modelo

A sessão a seguir é responsável por apresentar os testes realizados no modelo preditivo e seus respectivos resultados.

8.1 Divisão dos dados

Antes de modelar os algoritmos para predição das classes das variáveis alvo, é necessário organizar os atributos escolhidos, para levar ao Aprendizado de Máquina, entre variáveis de teste e variáveis de treino, que estão explicadas abaixo:

- **Dados de Treino:**

Os dados de Treino são, como o nome sugere, dados selecionados de uma base de dados que representam cerca de 70% da totalidade do conjunto da base e são levados para o treinamento do algoritmo de predição do Machine Learning;

- **Dados de Teste:**

Os dados de Teste são, como o nome sugere, dados levantados de uma base de dados que representam em torno de 30% do conjunto completo da base e servem para testar o algoritmo preditivo criado pelo aprendizado de máquina.

É importante ressaltar que haja a separação desses dados de maneira aleatória, para que não ocorra enviesamento dos dados por meio do aprendizado de padrões que limitam a probabilidade de predição, e a separação é necessária também para que não haja casos de overfitting, ou seja, um ajuste desproporcional aos dados apresentados.

8.2 Variáveis Utilizadas

Para a realização dos testes dos modelos utilizou-se o seguinte conjunto de dados: Idade, Tempo de Empresa em Meses, Número Promoções, Tempo médio – Promoção, Promoção, Mérito, Salário Mês, Cargos, Gênero, Estado Civil, Estado, Cidade, Área de atuação. Parte dos dados utilizados foram fornecidos pelo parceiro e outros derivados dos dados originais, como por exemplo a “Tempo Empresa Meses”. O conjunto de variáveis escolhidas foram selecionadas a partir da análise e percepção daqueles que fazem maior sentido para a solução.

8.3 Estratégia de Avaliação do modelo

Nesta sessão é apresentado todas as avaliações dos algoritmos utilizados para a construção do modelo preditivo e seus respectivos resultados obtidos.

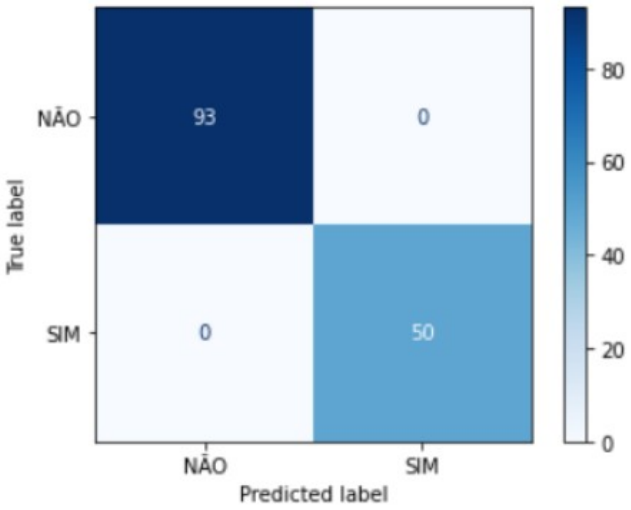
8.3.1 Matriz de Confusão

Pode-se definir matriz de confusão como, uma tabela que representa a frequência de classificação para as variáveis declaradas no modelo. O uso dessa ferramenta de avaliação é de grande importância pois é possível realizar a análise de como o modelo se saiu nas previsões, verificando erros e acertos do modelo preditivo.

8.3.1.1 Árvore de Decisão

Na figura 30 abaixo é ilustrado uma visualização gráfica da matriz de risco para o algoritmo de árvore de decisão. Apresentando uma previsão onde de 143 pessoas analisadas, 93 delas não sairão da empresa e 50 delas irão, tendo 100% de acerto mediante a base de dados analisada.

Figure 30: Matriz de risco - Árvore de decisão



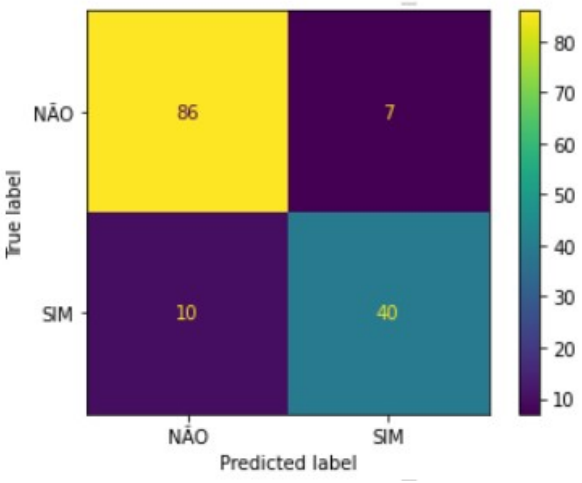
Fonte: Autoria própria

8.3.1.2 KNN

Na figura 31 abaixo é ilustrado uma visualização gráfica da matriz de risco para o algoritmo de árvore de decisão. A partir da análise da Matriz de Confusão criada, pode-se

perceber que, nos resultados parciais, o modelo conseguiu prever 86 dos 93 funcionários que permaneceram na empresa e 40 dos 50 dos funcionários que saíram da empresa.

Figure 31: Matriz de confusão - KNN

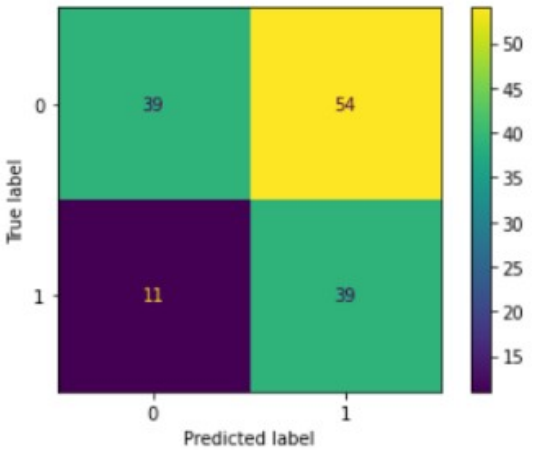


Fonte: Autoria própria

8.3.1.3 Naive Bayes

Na figura 32 abaixo é ilustrado uma visualização gráfica da matriz de risco para o algoritmo de árvore de decisão. Com os resultados obtidos em relação à variável alvo, “Saiu da Empresa”, que pode ser definida em sim ou não, foi gerada uma matriz de confusão de 39 verdadeiros negativos (funcionários preditos a ficar que realmente ficaram), 54 falsos negativos (funcionários preditos a ficar que saíram), 11 falsos positivos (funcionários preditos a ficar que saíram) e 39 verdadeiros positivos(funcionários preditos a sair que saíram). Assim sendo possível visualizar o resultado obtido pelo modelo.

Figure 32: Matriz de risco - Naive Bayes

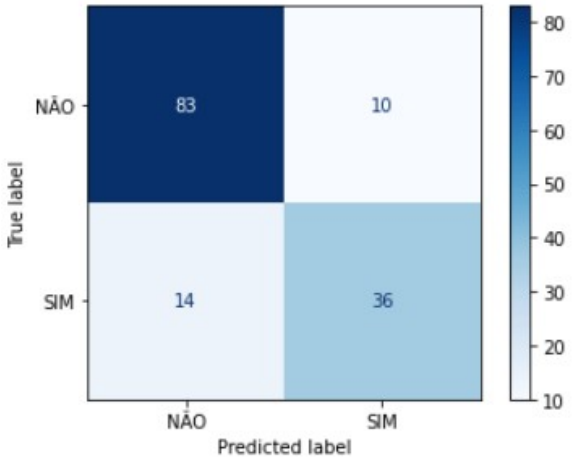


Fonte: Autoria própria

8.3.1.4 SVM

Na figura 33 abaixo é ilustrado uma visualização gráfica da matriz de risco para o algoritmo de árvore de decisão. A partir da análise da Matriz de Confusão criada, pode-se perceber que, nos resultados parciais, o modelo conseguiu prever 83 dos 93 funcionários que permaneceram na empresa e 36 dos 50 dos funcionários que saíram da empresa.

Figure 33: Matriz de risco - SVM

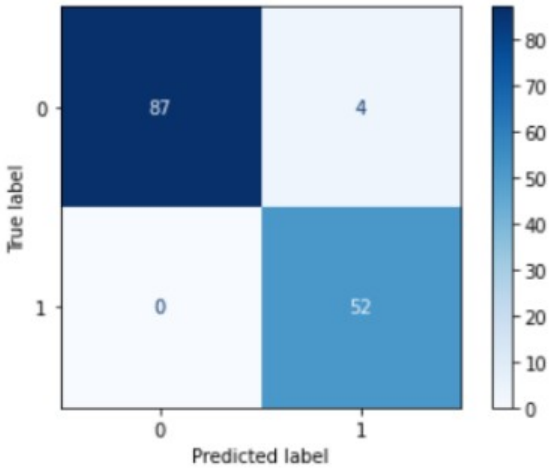


Fonte: Autoria própria

8.3.1.5 Regressão Logística

Na figura 34 abaixo é ilustrado uma visualização gráfica da matriz de risco para o algoritmo de árvore de decisão. A partir da análise da Matriz de Confusão criada, pode-se perceber que, nos resultados parciais, o modelo conseguiu prever 87 dos 93 funcionários que permaneceram na empresa e 52 dos 52 dos funcionários que saíram da empresa.

Figure 34: Matriz de risco - Regressão Logística



Fonte: Autoria própria

8.3.2 Acurácia

A acurácia diz respeito à proximidade entre o valor obtido experimentalmente e o valor verdadeiro. A importância dessa estratégia de avaliação se dá pelo fato de determinar a confiabilidade e grau de exatidão do modelo. Na tabela 6 abaixo, exibe-se os resultados obtidos em relação a acurácia de treino e teste para os cinco algoritmos testados.

Table 6: Acurácia dos algoritmos escolhidos

Algoritmos	Acurácia	
	Treino	Teste
Árvore de decisão	1	1
KNN	0.93	0.88
Naive Bayes	0.78	0.54
SVM	0.81	0.83
Regressão Logística	1	0.97

Exemplo de interpretação prevista para a análise da tabela, utilizada e adaptada para qualquer um dos algoritmos ilustrados: A acurácia preliminar obtida no modelo foi de 93% para o treino, e 88% para o teste e isto significa que, nesse caso, para o modelo KNN, o algoritmo obteve 88% de acerto.

8.3.3 Precisão

A precisão foi elencada pois ela observa se os valores previstos de fato pertencem à classe que se quer obter. A precisão demonstra dentre todas as classificações positivas, quais são as verdadeiras. Na tabela 7 abaixo, exibe-se os resultados obtidos em relação a precisão de treino e teste para os cinco algoritmos testados.

Table 7: Precisão dos algoritmos escolhidos

Algoritmos	Precisão	
	Treino	Teste
Árvore de decisão	1	1
KNN	0.9	0.85
Naive Bayes	0.78	0.42
SVM	0.86	0.78
Regressão Logística	1	0.94

Exemplo de interpretação prevista para a análise da tabela, utilizada e adaptada para qualquer um dos algoritmos ilustrados: A precisão obtida no modelo foi de 90% para os verdadeiros valores preditos de retenção e 85% para os verdadeiros valores preditos de saída, o que significa, que nesse modelo dentre os valores que se queria prever, 90% e 85% foram realmente previstos.

8.3.4 Recall

O Recall foi elencado pois tal método apresenta classe predita em relação ao que realmente se espera de resultado. Sendo assim o Recall mostra dentre todos os casos classificados como Positivo, quanto está correto. Na tabela 8 abaixo, exibe-se os resultados obtidos em relação ao Recall de treino e teste para os cinco algoritmos testados.

Table 8: Recall dos algoritmos escolhidos

Algoritmos	Precisão	
	Treino	Teste
Árvore de decisão	1	1
KNN	0.9	0.85
Naive Bayes	0.78	0.42
SVM	0.86	0.78
Regressão Logística	1	0.94

Exemplo de interpretação prevista para a análise da tabela, utilizada e adaptada para qualquer um dos algoritmos ilustrados: O recall obtido no modelo foi de 92% para os verdadeiros valores preditos de retenção e 80% para os verdadeiros valores preditos de saída, o que significa, que nesse modelo dentre os valores que se previu 92% e 80% deveriam ser realmente previstos nessas classes.

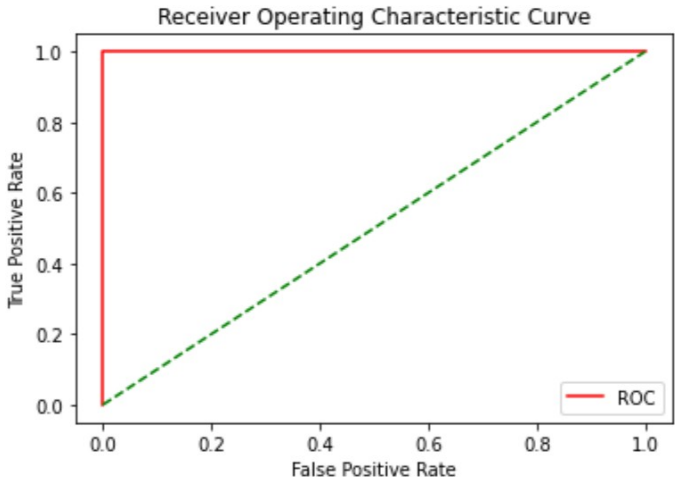
8.3.5 Curva ROC (Receiver Operating Characteristic)

A curva ROC é capaz de demonstrar o desempenho de um modelo de Machine Learning, que fará uma classificação binária, por meio da relação entre Taxa de Verdadeiro Positivo e da Taxa de Falso Positivo, variando assim os pontos de corte da probabilidade estimada (threshold). Na curva é capaz de determinar a AUC onde observa-se o desempenho do algoritmo por meio do valor obtido a partir da área abaixo da curva ROC que une os limiares dos pontos de corte da probabilidade estimada.

8.3.5.1 Árvore de Decisão

Na figura 35 abaixo é ilustrado uma visualização gráfica da curva ROC para o algoritmo de árvore de decisão. Na linha onde é apresentado o valor “1.000”, diz respeito à área abaixo da curva ROC. Para tal algoritmo implementado o *‘Random Prediction’* apresentou 1.0 mediante a AUROC.

Figure 35: Curva ROC - Árvore de decisão

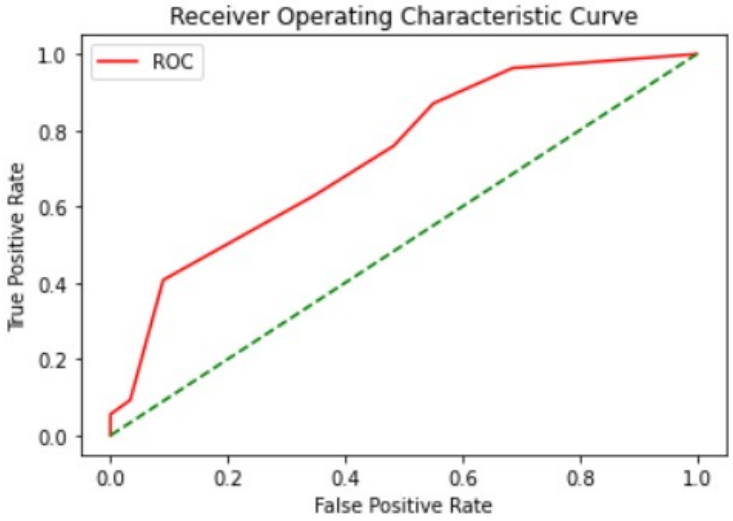


Fonte: Autoria Própria

8.3.5.2 KNN

Na figura 36 abaixo é ilustrado uma visualização gráfica da curva ROC para o algoritmo KNN. A partir do valor apresentado, é possível observar que a taxa de valores verdadeiros positivos em relação à taxa de valores falsos positivos é de 73,1%. Para tal algoritmo implementado o *‘Random Prediction’* apresentou 0.73 mediante a AUROC.

Figure 36: Curva ROC - KNN

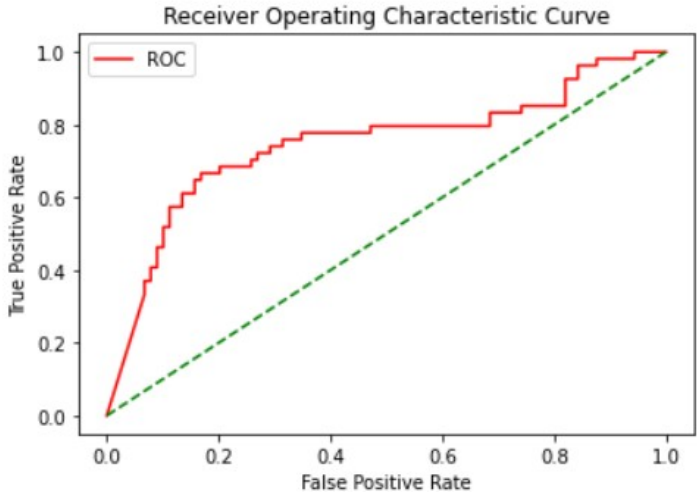


Fonte: Autoria Própria

8.3.5.3 Naive Bayes

Na figura 37 abaixo é ilustrado uma visualização gráfica da curva ROC para o algoritmo KNN. A partir do valor apresentado, é possível observar que a taxa de valores verdadeiros positivos em relação à taxa de valores falsos positivos é de 74,8%. Para tal algoritmo implementado o *'Random Prediction'* apresentou 0.74 mediante a AUROC.

Figure 37: Curva ROC - Naive Bayes

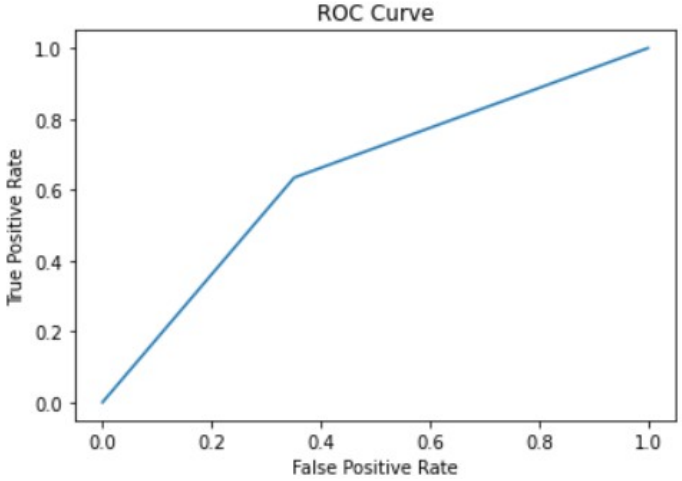


Fonte: Autoria própria

8.3.5.4 SVM

Na figura 38 abaixo é ilustrado uma visualização gráfica da curva ROC para o algoritmo KNN. A partir do valor apresentado, é possível observar que a taxa de valores verdadeiros positivos em relação à taxa de valores falsos positivos é de 64,1%. Para tal algoritmo implementado o *'Random Prediction'* apresentou 0.64 mediante a AUROC.

Figure 38: Curva ROC - SVM

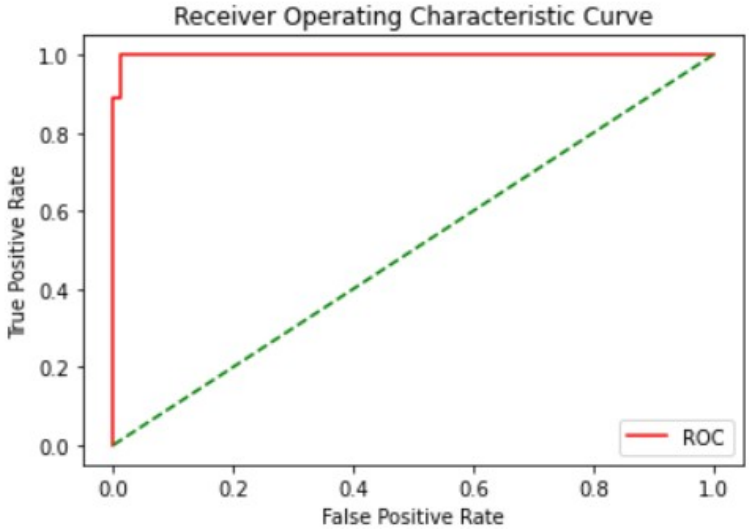


Fonte: Autoria própria

8.3.5.5 Regressão Logística

Na figura 39 abaixo é ilustrado uma visualização gráfica da curva ROC para o algoritmo KNN. A partir do valor apresentado, é possível observar que a taxa de valores verdadeiros positivos em relação à taxa de valores falsos positivos é de 99,9%. Para tal algoritmo implementado o *'Random Prediction'* apresentou 0.99 mediante a AUROC.

Figure 39: Curva ROC - Regressão Logística



Fonte: Autoria própria

8.3.6 Taxa de erro

Na sessão abaixo é apresentado a taxa de erro calculada para os algoritmos, sendo ela resultante de uma regra de três para o número de erros em relação ao número total de casos, de forma a transformar esse valor em porcentagem.

8.3.6.1 Árvore de Decisão

No caso da árvore de decisão a taxa de erro encontrada foi de 0% de erro, uma vez que, houve 100% de acerto desse modelo. Tal resultado pode ser comprovado pelas imagens das métricas de avaliação descritas acima que apontam o acerto total pelo modelo. Após diversas análises do modelo foi constatado que este não apresenta nenhum erro em sua criação. Acredita-se que sua grande precisão vem do fato de que para a sua execução foram fornecidas um número restrito de dados para o treino e teste do modelo. Dessa maneira, não é possível deduzir o quão incontestável é tal resultado.

8.3.6.2 KNN

No modelo KNN houve um total de 7 falsos sim e 10 falsos não em seus dados teste, significando que, de um total de 11,9% de erro. Tal informação pode ser deduzida a partir das métricas de avaliação demonstradas acima.

8.3.6.3 Naive Bayes

No modelo de Naive Bayes houve um total de 11 falsos positivos e 54 falsos negativos em seus dados teste, significando que, de um total de 45,4% de erro. Tal informação pode ser deduzida a partir das métricas de avaliação demonstradas acima. Dessa maneira conclui-se que, em comparação com os demais modelos, Naive Bayes teve menor precisão.

8.3.6.4 SVM

No modelo de SVM houve um total de 14 falsos positivos e 10 falsos negativos em seus dados teste, significando que, de um total de 16,8% de erro. Tal informação pode ser deduzida a partir das métricas de avaliação demonstradas acima.

8.3.6.5 Regressão Logística

No modelo de Regressão Logística houve um total de 0, falsos positivos e 3 falsos negativos em seus dados teste, significando que, de um total de 2,09% de erro. Tal informação pode ser deduzida a partir das métricas de avaliação demonstradas acima.

8.4 Hiperparâmetros

Hiperparâmetros são parâmetros selecionados em que o valor é usado para melhorar o aprendizado de máquina. O objetivo de sua utilização é aperfeiçoar a precisão dos modelos manipulados. Os algoritmos de pesquisa dos hiperparâmetros foram o *"Grid Search"* em que a busca acontece em grade simples dentro do espaço viável e o *"Random Search"* em que é feita uma busca aleatória dentro do espaço viável.

8.4.1 Árvore de Decisão

Na Figura 40 abaixo são apresentados os parâmetros utilizados para o algoritmo de árvore de decisão, visando apresentar quais foram selecionados para melhorar o desempenho do modelo e sua taxa de acerto.

Fonte: Autoria própria

Na tabela 9 abaixo é apresentado a descrição do significado de cada parâmetro utilizado para o algoritmo de árvore de decisão.

Table 9: Descrição dos parâmetros - Árvore de decisão

Parâmetro	Descrição
“criterion”	Mede a qualidade de uma divisão.
“max_depth”	Diz qual será o tamanho máximo da árvore.
“min_samples_split”	Descreve qual será o mínimo de amostras necessárias para dividir um nó.
“min_samples_leaf”	Número mínimo de amostras para estar em uma folha.

8.4.2 KNN

Na Figura 41 abaixo são apresentados os parâmetros utilizados para o algoritmo KNN, visando apresentar quais foram selecionados para melhorar o desempenho do modelo e sua taxa de acerto.

Figure 41: Hiperparâmetros - KNN

```
parametros = {'n_neighbors': [3, 5, 7, 9, 13, 17, 21, 29],
              'weights': ['uniform', 'distance'],
              'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
              'leaf_size': [5, 10, 15, 30, 45, 60]}
```

Fonte: Autoria própria

Na tabela 10 abaixo é apresentado a descrição do significado de cada parâmetro utilizado para o algoritmo KNN.

Table 10: Descrição dos parâmetros - KNN

Parâmetro	Descrição
n_neighbors”	Número de amostras vizinhas que serão analisadas
“weights”	Peso usado na previsão
“algorithm”	Algoritmo usado para calcular os vizinhos mais próximos
“leaf_size”	Tamanho da folha utilizado para “ball_tree” e “kd_tree”

8.4.3 Naive Bayes

Na Figura 42 abaixo são apresentados os parâmetros utilizados para o algoritmo de Naive Bayes, visando apresentar quais foram selecionados para melhorar o desempenho do modelo e sua taxa de acerto.

Figure 42: Hiperparâmetro - Naive Bayes

```
parametros = {'var_smoothing': [1.0, 0.00009, 0.0000009]}
```

Fonte: Autoria própria

Na tabela 11 abaixo é apresentado a descrição do significado de cada parâmetro utilizado para o algoritmo de Naive Bayes.

Table 11: Descrição dos parâmetros - Naive Bayes

Parâmetro	Descrição
“var_smoothing”	Parte da maior variância de todos os recursos que é adicionada às variâncias para estabilidade de cálculo.

8.4.4 SVM

Na Figura 43 abaixo são apresentados os parâmetros utilizados para o algoritmo SVM, visando apresentar quais foram selecionados para melhorar o desempenho do modelo e sua taxa de acerto.

Figure 43: Hiperparâmetros - SVM

```
parameters = {'C':[1.0, 1.5, 0.5],
              'kernel' : ['linear', 'poly'],
              'gamma' : ['scale', 'auto']}
```

Fonte: Autoria própria

Na tabela 12 abaixo é apresentado a descrição do significado de cada parâmetro utilizado para o algoritmo de SVM.

Table 12: Descrição de parâmetros - SVM

Parâmetro	Descrição
“C”	Parâmetro de regularização.
“kernel”	Específica tipo de kernel utilizado pelo algoritmo.
“gamma”	Coeficiente de kernel

8.4.5 Regressão Logística

Na Figura 44 abaixo são apresentados os parâmetros utilizados para o algoritmo de Regressão Logística, visando apresentar quais foram selecionados para melhorar o desempenho do modelo e sua taxa de acerto.

Figure 44: Hiperparâmetros - Regressão Logística

```
parametros = {'C':[0.1, 0.3, 0.5, 0.7, 0.9, 1, 10, 100],
              'penalty':['l1', 'l2']
             }
```

Fonte: Autoria própria

Na tabela 13 abaixo é apresentado a descrição do significado de cada parâmetro utilizado para o algoritmo de Regressão Logística.

Table 13: Descrição dos parâmetros - Regressão Logística

Parâmetro	Descrição
“C”	Inverso da força de regularização.
“penalty”	Especifica a norma de uma penalidade.

8.4.6 Grid Search

No *Grid Search*, tentamos combinações de valores pré definidos e avaliamos os modelos para cada. Os valores são colocados em forma de matriz. Anota-se a precisão dos modelos e com base nos resultados são definidos os melhores. A desvantagem é que quando se trata de dimensionalidade o número de hiperparâmetros cresce exponencialmente. Mediante as alterações realizadas com a implementação do método de *Grid Search*, a tabela 14 abaixo ilustra os novos valores de acurácia de treino, acurácia de teste, revocação, precisão e f1_score, para os cinco algoritmos analisados.

Table 14: Descrição parâmetros - Grid Search

Algoritmo	Acc* treino	Acc* teste	Revocação	Precisão	F1_score
Árvore de decisão	0.98	0.95	1	0.89	0.94
KNN	1	0.81	0.73	0.74	0.73
Naive Bayes	0.93	0.93	0.98	0.85	0.91
SVM	0.99	0.96	1	0.91	0.95
Regressão Logística	0.98	0.94	0.96	0.88	0.92

**Acc é a acurácia do algoritmo.*

8.4.7 Random Search

Uma técnica que usa combinações aleatórias de hiperparâmetros para encontrar a melhor solução. É semelhante ao *Grid Search* e produz resultados melhores pois encontra o resultado de forma mais eficiente. A desvantagem é que ela produz uma alta variação devido à aleatoriedade. Mediante as alterações realizadas com a implementação do método de *Random Search*, a tabela 15 abaixo ilustra os novos valores de acurácia de treino, acurácia de teste, revocação, precisão e f1_score, para os cinco algoritmos analisados.

Table 15: Descrição de parâmetros - Random Search

Algoritmo	Acc* treino	Acc* teste	Revocação	Precisão	F1_score
Árvore de decisão	0.98	0.96	1	0.91	0.95
KNN	0.86	0.76	0.67	0.68	0.67

Naive Bayes	0.93	0.93	0.98	0.85	0.91
SVM	0.99	0.96	1	0.91	0.95
Regressão Logística	1	0.74	0.70	0.62	0.66

8.4 Estabilidade de dados (conjunto de treino e teste)

Em relação a mudança no *'Random State'* do conjunto de teste e treino do modelo, analisou-se que os valores mesmo com alterações de parâmetros, permanecem com as mesmas proporções de métricas de avaliação. As variações que foram submetidas, estão relacionadas aos cinco algoritmos utilizados, tendo como base os números de *'Random State'*, 42, 43, 44 e 45. Abaixo nas tabelas 6, 7, 8 e 9, é ilustrado o resultado das métricas após a aplicação dos novos parâmetros.

Table 16: Random State – 42

Algoritmos	Precisão	Recall	Acurácia
Resultado de Treino			
Árvore de decisão	0.98	0.95	0.97
KNN	0.76	0.63	0.63
SVM	0.76	0.61	0.62
Naive Bayes	0.78	0.42	0.55
Regressão Logística	0.99	0.94	0.95
Resultado de Teste			
Árvore de decisão	0.92	0.98	0.97
KNN	0.48	0.62	0.63
SVM	0.47	0.64	0.62
Naive Bayes	0.42	0.78	0.55
Regressão Logística	0.89	0.98	0.95

Table 17: Random State – 43

Algoritmos	Precisão	Recall	Acurácia
Resultado de Treino			
Árvore de decisão	1	0.95	0.94
KNN	0.76	0.63	0.63
SVM	0.76	0.61	0.62
Naive Bayes	0.78	0.42	0.55

Regressão Logística	0.99	0.94	0.95
Resultado de Teste			
Árvore de decisão	0.92	0.92	0.94
KNN	0.48	0.62	0.63
SVM	0.47	0.64	0.62
Naive Bayes	0.42	0.78	0.55
Regressão Logística	0.89	0.98	0.95

Table 18: Random State – 44

Algoritmos	Precisão	Recall	Acurácia
Resultado de Treino			
Árvore de decisão	0.97	0.96	0.97
KNN	0.76	0.63	0.63
SVM	0.76	0.61	0.62
Naive Bayes	0.78	0.42	0.55
Regressão Logística	0.99	0.94	0.95
Resultado de Teste			
Árvore de decisão	0.92	0.96	0.97
KNN	0.48	0.62	0.63
SVM	0.47	0.64	0.62
Naive Bayes	0.42	0.78	0.55
Regressão Logística	0.89	0.98	0.95

Table 19: Random State – 45

Algoritmos	Precisão	Recall	Acurácia
Resultado de Treino			
Árvore de decisão	1	0.95	0.94
KNN	0.76	0.63	0.63
SVM	0.76	0.61	0.62
Naive Bayes	0.78	0.42	0.55
Regressão Logística	0.99	0.94	0.95
Resultado de Teste			
Árvore de decisão	0.89	0.96	0.94
KNN	0.48	0.62	0.63

SVM	0.47	0.64	0.62
Naive Bayes	0.42	0.78	0.55
Regressão Logística	0.89	0.98	0.95

Analisando as tabelas com as diferentes métricas, tendo como base os algoritmos mais promissores, árvore de decisão e regressão logística, os parâmetros de Random State considerado com baixa variação, foram os com números 42 e 44, tendo uma variabilidade de 1% a 2%. Nesse sentido, utilizaremos como base na construção do modelo, o parâmetro 44, este apresenta um maior balanceamento de recall, analisando os dados de treino (com média de 0.96) e teste (com média de 0.97) do modelo.

8.5 Comparação modelos com Hiperparâmetros

Após a execução e análise dos hiperparâmetros, selecionou-se os algoritmos com a menor taxa de variação, tendo como base os métodos, GridSearch, Random e Sem Hiperparâmetros. Sendo eles, 1) Árvore de decisão; 2) Regressão Logística; e 3) SVM, respectivamente. Na tabela 20, ilustra-se os resultados obtidos para o primeiro algoritmo.

Table 20: Árvore de decisão - Comparação com Hiperparâmetros

Métricas	GridSearch	Random	Sem Hiperparâmetros
Resultado de Treino			
Precisão	1	1	1
Recall	0.95	0.92	0.96
Acurácia	0.97	0.95	0.97
Resultado de Teste			
Precisão	0.91	0.88	0.93
Recall	1	1	1
Acurácia	0.97	0.95	0.97

Com a análise dos diferentes parâmetros para o algoritmo da árvore de decisão, constatou-se que os melhores resultados foram com a árvore base, ou seja, sem os hiperparâmetros. Uma vez que os números obtidos, em todas as três categorias analisadas (precisão, recall e acurácia), foram maiores em 3% em relação aos com a implementação dos hiperparâmetros.

Na tabela 21, ilustra-se os resultados obtidos para o segundo algoritmo.

Table 21: Regressão Logística - Comparação com Hiperparâmetros

Métricas	GridSearch	Random	Sem Hiperparâmetros
Resultado de Treino			
Precisão	1	1	1
Recall	0.96	0.95	0.96
Acurácia	0.97	0.97	0.97
Resultado de Teste			
Precisão	0.93	0.91	0.93
Recall	1	1	1
Acurácia	0.97	0.97	0.97

Com a análise dos diferentes parâmetros para o algoritmo da regressão logística, constatou-se que os melhores resultados foram com a regressão base, ou seja, sem os hiperparâmetros. Uma vez que os números obtidos, na precisão, foram maiores em 2% ou iguais em relação aos com a implementação dos hiperparâmetros.

Na tabela 22, ilustra-se os resultados obtidos para o segundo algoritmo.

Table 22: SVM – Comparação com Hiperparâmetros

Métricas	GridSearch	Random	Sem Hiperparâmetros
Resultado de Treino			
Precisão	1	1	0.76
Recall	0.95	0.95	0.65
Acurácia	0.97	0.97	0.64
Resultado de Teste			
Precisão	0.91	0.91	0.51
Recall	1	1	0.63
Acurácia	0.97	0.97	0.64

Com a análise dos diferentes parâmetros para o algoritmo SVM, constatou-se que os melhores resultados foram com o SVM com hiperparâmetros, seja o GridSearch ou Random. Uma vez que os números obtidos, em todas as três categorias analisadas

(precisão, recall e acurácia), foram maiores em cerca de 30% em relação aos com a implementação sem hiperparâmetros.

Na tabela 23, ilustra-se os resultados obtidos para o segundo algoritmo.

Table 23: KNN – Comparação com Hiperparâmetros

Métricas	GridSearch	Random	Sem Hiperparâmetros
Resultado de Treino			
Precisão	0.85	0.85	0.81
Recall	0.86	0.86	0.78
Acurácia	0.81	0.81	0.74
Resultado de Teste			
Precisão	0.75	0.75	0.64
Recall	0.73	0.73	0.67
Acurácia	0.81	0.81	0.74

Com a análise dos diferentes parâmetros para o algoritmo KNN, constatou-se que os melhores resultados foram com o KNN com hiperparâmetros, seja o GridSearch ou Random. Uma vez que os números obtidos, em todas as três categorias analisadas (precisão, recall e acurácia), foram maiores em cerca de 10% em relação aos com a implementação sem hiperparâmetros.

Na tabela 24, ilustra-se os resultados obtidos para o segundo algoritmo.

Table 24: Naive Bayes – Comparação com Hiperparâmetros

Métricas	GridSearch	Random	Sem Hiperparâmetros
Resultado de Treino			
Precisão	0.99	0.99	0.83
Recall	0.9	0.9	0.48
Acurácia	0.93	0.93	0.61
Resultado de Teste			
Precisão	0.85	0.85	0.48
Recall	0.98	0.98	0.83
Acurácia	0.93	0.93	0.61

Com a análise dos diferentes parâmetros para o algoritmo Naive Bayes, constatou-se que os melhores resultados foram com o Naive Bayes com hiperparâmetros, seja o GridSearch ou Random. Uma vez que os números obtidos, em todas as três categorias analisadas (precisão, recall e acurácia), foram maiores em cerca de 30% na acurácia e no recall e 10% na precisão em relação aos com a implementação sem hiperparâmetros.

8.6 Métricas

Na sessão abaixo se encontra as novas métricas utilizadas para teste do modelo.

8.6.1 Especificidade

Tal métrica avalia a capacidade do método de detectar com sucesso, resultados classificados como negativos. Ela utiliza uma porção de VN (verdadeiro negativo nesse caso o dado era negativo e foi previsto como negativo) em relação ao total de negativos (que é se refere a soma do verdadeiro negativo com o falso positivo). Essa avaliação pode ser realizada utilizando a equação ilustrada na Figura 45.

Figure 45: Formula - Método de Especificidade

$$\frac{VN}{(FP+VN)}$$

Fonte: Autoria própria

8.6.2 Sensibilidade

Tal métrica avalia a capacidade do método de detectar com sucesso, resultados classificados como positivos. Ela utiliza uma porção de VP (verdadeiro positivo, nesse caso o dado era positivo e foi previsto como positivo) em relação ao total de positivos (que é se refere a soma do verdadeiro positivo com o falso negativo). Essa avaliação pode ser realizada utilizando a equação ilustrada na Figura 46.

Figure 46: Formula - Método de sensibilidade

Fonte: Autoria própria

Com a aplicação de tais novas métricas nos modelos obteve-se tais resultados, ilustrados na tabela 25 abaixo.

Table 25: Avaliação - Métricas de Especificidade e Sensibilidade

Algoritmo	Especificidade	Sensibilidade
Sem Hiperparâmetros		
Árvore de decisão	1	0.97
KNN	1	0.96
Naive Bayes	0.82	0.48
SVM	0.63	0.65
Regressão Logística	1	0.96
Com Hiperparâmetros		
Árvore de decisão	1	0.95
KNN	0.73	0.86
Naive Bayes	0.98	0.9
SVM	0.73	0.75
Regressão Logística	1	0.95

8.7 Algoritmo Escolhido

Com base nos resultados analisados tanto na implementação dos algoritmos bases quanto dos com hiperparâmetros (GridSearch ou Random), constatou-se que o modelo que apresentava melhores desempenho foi o da árvore de decisão. Uma vez que, seus números mostraram valores superiores ou iguais aos maiores obtidos em todas as

métricas e parâmetros impostos. Além de comparado aos outros algoritmos utilizados, seu tempo de execução é relativamente melhor que os outros.

9 Referências

Phase 1 of the CRISP-DM Process Model: Business Understanding. Disponível em: <https://www.dummies.com/article/technology/information-technology/data-science/general-data-science/phase-1-of-the-crisp-dm-process-model-business-understanding-148161/>. Acesso em: 21 set. 2022.

Phase 2 of the CRISP-DM Process Model: Data Understanding. Disponível em: <https://www.dummies.com/article/technology/information-technology/data-science/general-data-science/phase-2-of-the-crisp-dm-process-model-data-understanding-148213/>. Acesso em: 21 set. 2022.

Phase 3 of the CRISP-DM Process Model: Data Preparation. Disponível em: <https://www.dummies.com/article/technology/information-technology/data-science/general-data-science/phase-3-of-the-crisp-dm-process-model-data-preparation-148177/>. Acesso em: 21 set. 2022.