TURNOVER DE FUNCIONÁRIOS
EVERYMIND

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
09/08	Jean Lucas Rothstein Machado	1.0	Inserção de Compreensão dos Dados
09/08	Giovanna Furlan	1.1	Inserção dos artefatos 1: 1. Persona; 2. Análise da Indústria; 3. Matriz SWOT; 4. Proposta de valor; 5. Descrição da solução;
09/08	Patrick V Miranda	1.2	1- Introdução 2-Objetivos e justificativa
10/08	Giovanna Furlan	1.3	Introdução em cada item do tópico 4
12/08	Emanuele Morais, Lucas Britto	1.4	4.2.3 Inserção de gráficos e descrições
12/08	Emanuele Morais, Giovanna Furlan, Jean Lucas e Pedro Sant'Anna	1.4	4.2.1 Edição e inserção de tabelas
12/08	Giovanna Furlan	1.4	Revisão dos texto de Introdução, Objetivos e justificativa, descrição do conjunto de dados e predição
12/08	Patrick Victorino	1.4	Atualização do 4.1.3 e do 4.2.4 de modelo regressivo para classificativo, alterando grande parte do texto
15/08	Giovanna Furlan, Patrick Victorino e Emanuele Morais	1.5	Inserção e criação da terceira persona
18/08	Giovanna Furlan	1.6	1. Criação e inserção da 4 Persona 2. Inserção das Jornadas dos usuários
25/08	Giovanna Furlan, Patrick Victorino, Jean Lucas, Emanuele Morais, Lucas Brito	1.7	Inserção e descrição dos códigos no tópico 4.3
25/08	Jean Lucas Machado	1.8	Inserção e descrição dos novos gráficos
26/08	Giovanna Furlan	1.9	1. Finalização dos textos e inserção de todas as imagens do tópico 4.3 2.Correção dos artefatos entregues na última sprint com exceção da matriz de risco.
05/09	Pedro Henrique	2.0	Inserção Matriz de Correlação.

Sumário

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 em altos custos, entre eles, o onboarding.

2. Objetivos e Justificativa

2.1. Objetivos

Objetivos gerais:

- Diminuir o turnover de funcionários;
- Diminuir gastos com onboarding;
- Aumentar longevidade dos funcionários na empresa;

Objetivos específicos:

- Classificar o funcionário, a partir de um modelo preditivo, de acordo com a chance dele de sair da empresa ;
- Obter informações, através da análise de dados, sobre quais características mais influenciam a saída de um funcionário;
- Identificar períodos ou situações que levam a perda de funcionários;

2.2. Justificativa

A proposta de solução é através de um modelo preditivo, obter-se o potencial de entregar ao cliente informações sobre os funcionários mais propensos a sair da empresa. Isto será gerado pela identificação de padrões nos dados fornecidos, que possibilitam a identificação das características que levam um colaborador a estar sujeito a pedir demissão.

Assim, através da solução a Everymind poderá combater o turnover de funcionários, diminuir gastos com onboarding e aumentar a longevidade dos funcionários na empresa. Tudo isso através de estratégias que serão viabilizadas com a análise precisa das características e situações que levam os funcionários a sair.

3. Metodologia

Descreva as etapas metodológicas que foram utilizadas para o desenvolvimento, citando o referencial teórico. Você deve apenas enunciar os métodos, sem dizer ainda como ele foi aplicado e quais resultados obtidos.

3.1. CRISP-DM

Descreva brevemente a metodologia CRISP-DM e suas etapas de processo

3.2. Ferramentas

Descreva brevemente as ferramentas utilizadas e seus papéis (Google Colaboratory)

3.3. Principais técnicas empregadas

Descreva brevemente as principais técnicas empregadas, algoritmos e seus benefícios

4. Desenvolvimento e Resultados

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.

4.1. Compreensão do Problema

Nessa sessão é possível identificar as análises de mercado e produto previstas para o projeto, baseado na empresa Everymind.

4.1.1. Contexto 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.

A. Ameaça de novos entrantes:

- Outras empresas que dão suporte a programas de gerenciamento começarem a atender a Salesforce;
    - **Barreira:** Necessidade de elas obterem selo parceiro Salesforce.
- Empresas parceiras Salesforce em outros países, que podem expandir sua operação para território nacional;
    - **Barreira:** Alta taxa de burocracia e regulamentação dentro do país.
- A própria Salesforce (caso ela abra um setor onde as pessoas possam solicitar funções e suporte personalizado);
    - **Barreira:** Coloca em risco a relação com parceiros Salesforce;

B. Ameaça de produtos ou serviços substitutos:

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

- Programas de planilhas que ajudam na gestão das empresas;
- Outras plataformas de gestão empresarial (que não são a Salesforce) e as empresas que dão suporte a elas.

C. Poder de barganha dos consumidores:

- Exigência de alta qualidade de software, uma vez que a Everymind está associada a Salesforce, que tem essa característica associada a sua imagem;
- 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;
- Negociação de tempo de entrega, já que outras empresas podem oferecer o mesmo serviço em menor tempo;

D. Poder de barganha dos fornecedores:

- Hospedagem de programa (aumento de custos de funcionamento);
- Plano de internet (aumento de custos de funcionamento);
- Programas necessários para criação de ambiente de desenvolvimento (aumento nos custos de funcionamento);

E. Rivalidade entre concorrentes:

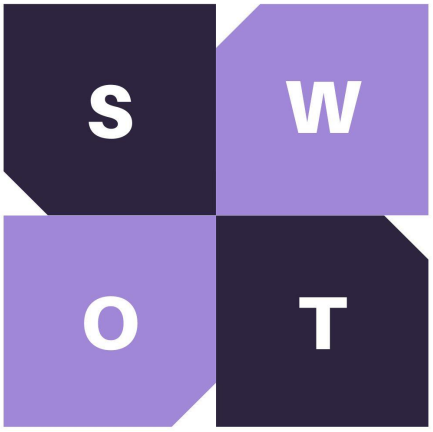
No Total são 137 parceiros Salesforce autorizados a atuar no Brasil, que podem oferecer produtos concorrentes a eles, alguns dos mais relevantes:

- 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);
- 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 prémios de excelência;

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

4.1.2. Análise 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 X , exibe uma imagem demonstrativa das quatro áreas que compõem a SWOT.



Fonte: Autoria própria

Pontos Fortes

1. A única empresa no Brasil que trabalha com todas as funções da plataforma SalesForce;
2. Aplicação da tecnologia de IA no gerenciamento de funcionários;
3. Horizontalidade da empresa;
4. Política de reconhecimento baseada no desempenho dos funcionários;

Pontos Fracos

1. Alta rotatividade de funcionários;
2. Alto gasto em recursos de onboarding;
3. Alto gasto de tempo e funcionários para o treinamento de novas pessoas;
4. Pouco diferenciais em relação às outras empresas líder de mercado

Oportunidade

1. Alto valor agregado em serviços na área de tecnologia/SalesForce;
2. Ausência de tecnologias que auxiliam na governança corporativa em outras empresas do mercado;
3. Demanda de mercado pela criação de um ambiente e funções personalizadas dentro da Salesforce;
4. Alto crescimento e preferência por serviços online, o que promove uma maior visibilidade da empresa no mercado.

Ameaças

1. Falta de profissionais na área de tecnologia;
2. Grande rotatividade na área da tecnologia;
3. Falta de profissionais adaptados à mentalidade Everymind;
4. Maior abrangência de empresas brasileiras no setor

4.1.3. Planejamento Geral da Solução

Abaixo encontra-se o detalhamento do problema a ser resolvido e a proposta de solução. Além dos principais benefícios esperados para a companhia.

A. Qual é o problema a ser resolvido ?

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.

B. Quais os dados disponíveis ?

- Dados de cadastramento de funcionários - Descrevendo informações pessoais e áreas que atuam;

- Dados dos reconhecimentos fornecidos aos colaboradores - Descrevendo quais ganharam uma promoção e mudaram de cargo e quais ganharam mérito e aumentaram o salário;

- 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.

C. Qual a solução proposta ?

A solução se baseia em uma ferramenta, que utiliza o aprendizado de máquina para realizar a previsão da taxa de rotatividade dos funcionários. 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.

D. Qual o tipo de tarefa?

Pode-se dizer que a tarefa a ser entregue é de método classificatório, apresenta 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 .

E. Como a solução proposta deverá ser utilizada?

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 devolutiva 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.

F. Quais os benefícios trazidos pela solução proposta

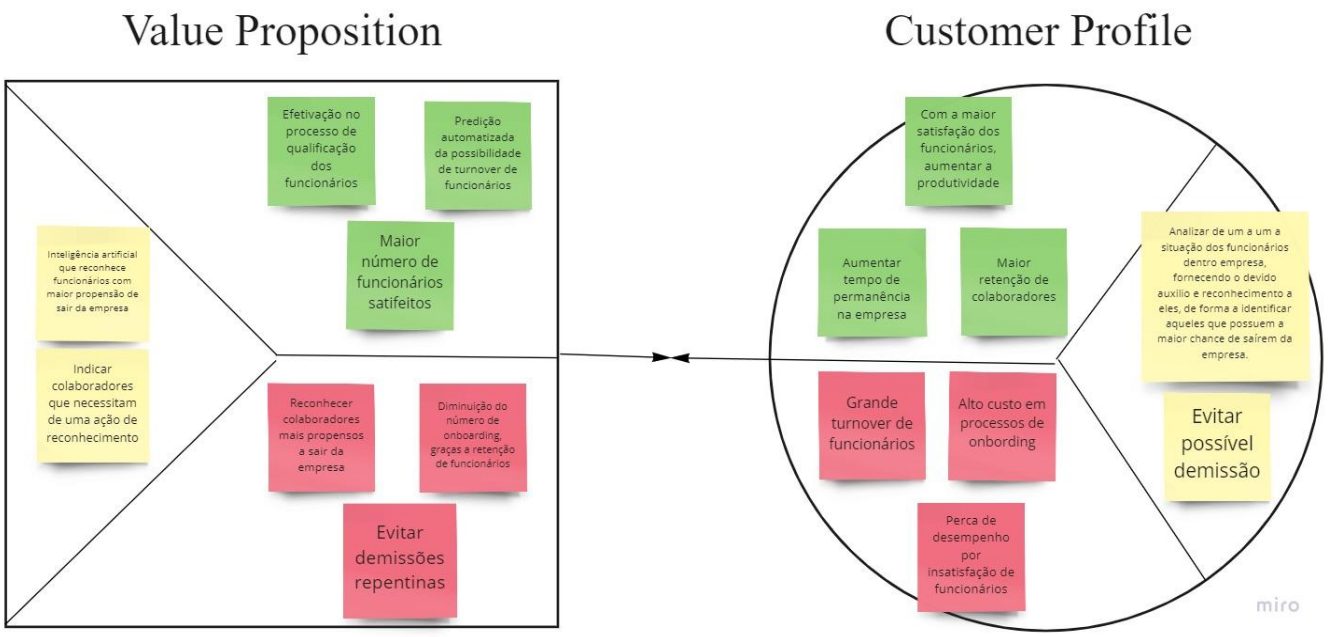
Com a entrega da solução proposta, espera-se de benefícios ao cliente: 1) Reduzir o percentual de colaboradores que desejam sair da companhia, não sendo necessário a substituição dos mesmos; 2) Auxílio à empresa na maneira de manter seus funcionários, mostrando quais parâmetros agregam mais para a sua 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.

G. Qual será o critério de sucesso e qual medida será utilizada para o avaliar?

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

4.1.4. Value Proposition Canvas

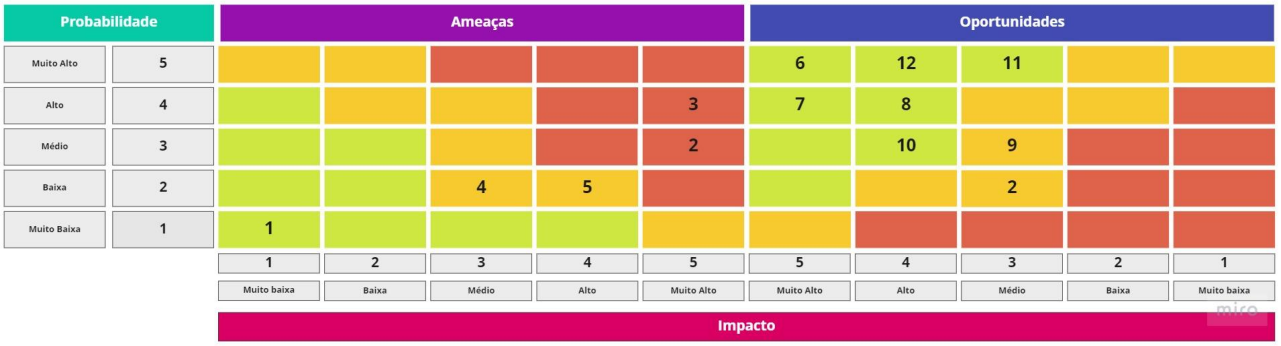
A principal vantagem apresentada pela proposta de valor é conseguir auxiliar a empresa a compreender melhor os seus clientes e funcionários. Na Figura X, é ilustrada a proposta construída para a Everymind.



Fonte: Autoria própria

4.1.5. Matriz de Riscos

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



Fonte: Autoria própria

Os números representam um risco ou oportunidade vista para o projeto e o impacto que ele ocasionará. Abaixo é encontrado a descrição de cada item:

Número 1 : Cliente não aprovar nenhuma parte do projeto

Número 2 : Modelo preditivo apontar resultados errôneos

Número 3 : Falta de dados suficientes

Número 4 : Poucas informações sobre o negócio

Número 5 : Não atender a necessidade do cliente

Número 6 : Reduzir a rotatividade de colaboradores

Número 7 : Reduzir os gastos com a contratação de novos funcionários

Número 8 : Evitar gastos contínuos com encargos trabalhistas

Número 9 : Aumentar a produtividade do negócio

Número 10 : Compreender as ineficiências da empresa

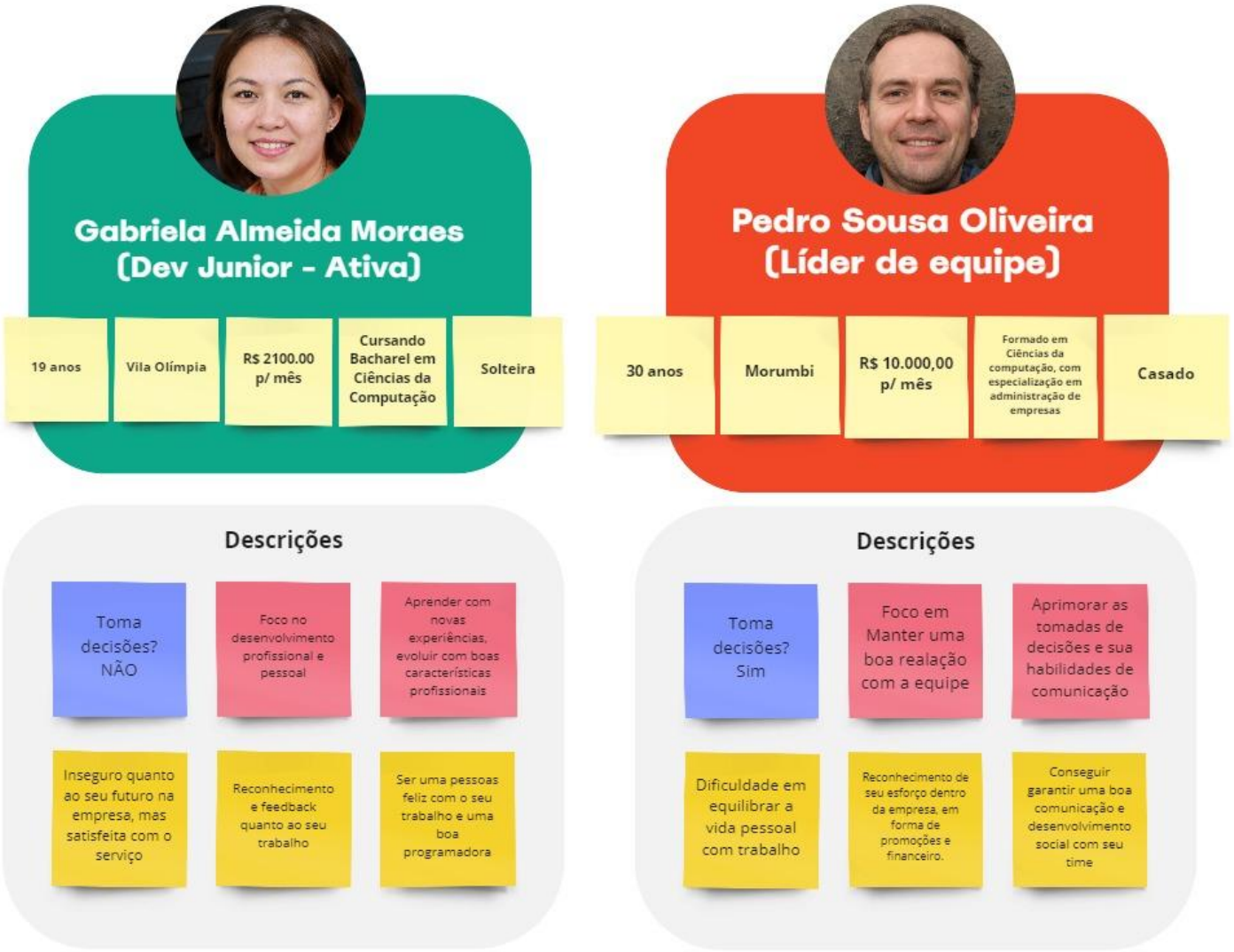
Número 11 : Melhora do clima organizacional

Número 12 : Reconhecer colaboradores que necessitam de reconhecimento

4.1.6. Personas

As personas do projeto são baseadas em dois setores principais, sendo eles, o colaborador da empresa e a gerente de RH. 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 X e X , exibem as personas construídas.

Fonte: Autoria própria



Fonte: Autoria própria

4.1.7. Jornadas 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 X,X,X e X, sendo elas respectivamente:

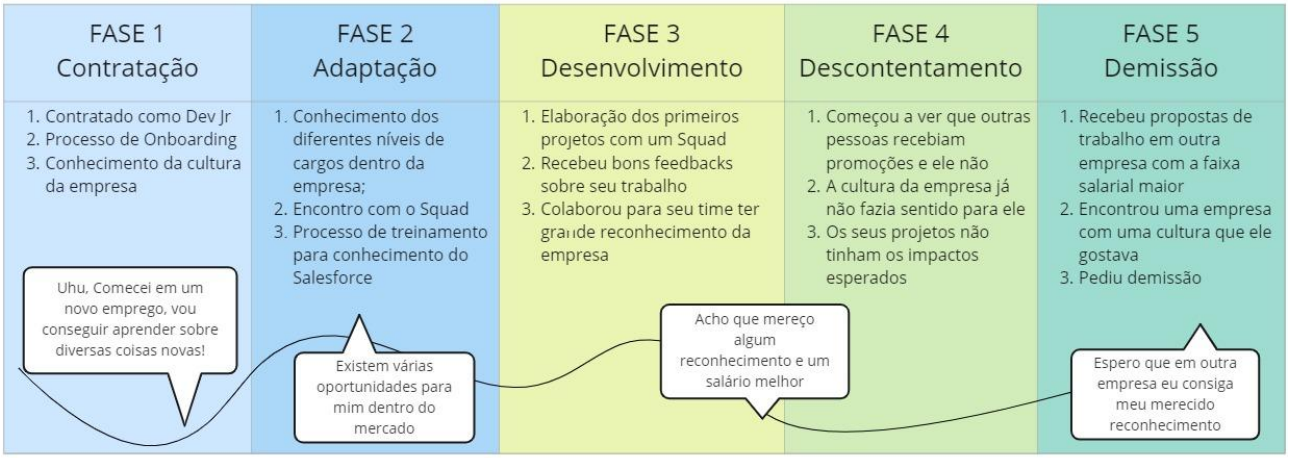
1. Dev Júnior que deseja sair da empresa;
2. Dev Júnior que deseja ficar na empresa;
3. Líder de equipe auxilia na decisão final;
4. Gerente de pessoas que toma a decisão final;

Allan Miranda Moura

Cenário: Funcionário quer encontrar empresa que esteja de acordo com seus valores e traga recompensas proporcionais ao seu nível profissional.

Expectativas

Obter reconhecimento profissional, tendo em vista a visão pessoal da sua qualificação dentro do mercado de trabalho.



Oportunidades

Devido ao seu conhecimento técnico e seu preparo profissional notou que havia outras oportunidades no mercado de trabalho que o levariam a ascensão profissional mais rapidamente do que em seu emprego atual, pois, embora estivesse se dedicando e trazendo bons resultados para a empresa, ele não era devidamente reconhecido.

Responsabilidades

Esse cenário só aconteceu pois o líder técnico não acompanhou o desenvolvimento de seus colaboradores, seus desejos pessoais e ambições. Além do seu descontentamento com a empresa em si.

miro

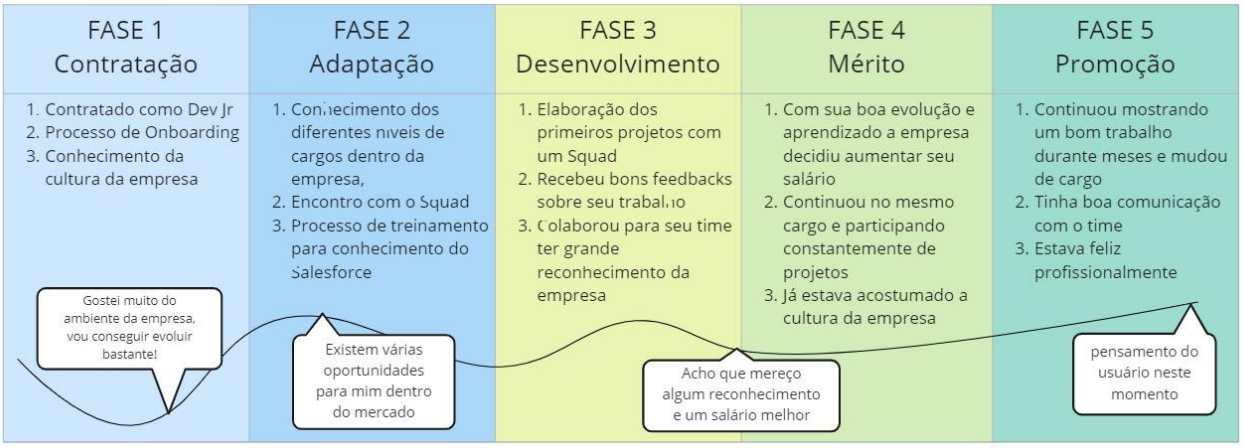
Fonte: Autoria própria

Gabriela Almeida Moraes

Cenário: Funcionário quer encontrar empresa que esteja de acordo com seus valores e traga recompensas proporcionais ao seu nível profissional.

Expectativas

Obter reconhecimento profissional, tendo em vista a visão pessoal da sua qualificação dentro do mercado de trabalho.



Oportunidades

O funcionário encontrou formas de avançar em sua carreira dentro da empresa, ser reconhecido pelo seu trabalho e alcançar seus desejos profissionais.

Responsabilidades

O líder técnico da equipe acompanhou o desenvolvimento do colaborador, suas ambições profissionais e deu o devido reconhecimento quando aplicável.

miro

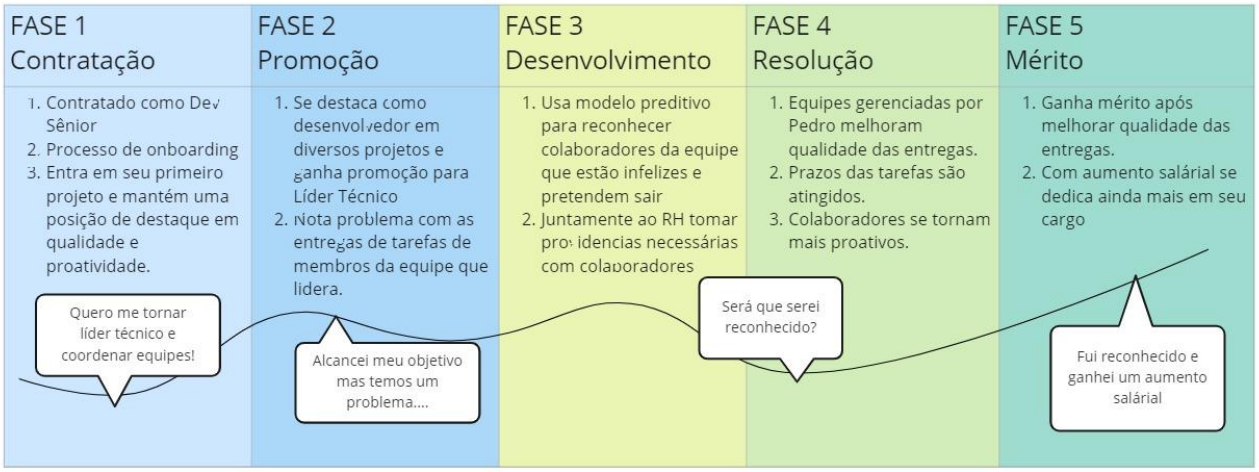
Fonte: Autoria própria

Pedro Costa Oliveira

Cenário: Contratado como desenvolvedor mas tem a ambição de se tornar um bom líder de equipe.

Expectativas

Ganhar reconhecimento de seu esforço dentro da empresa, em forma de promoções.



Oportunidades

Melhorar o desenvolvimento da equipe em relação a entregas para obter a oportunidade de se tornar Líder Tecnico

Responsabilidades

Responsabilidade de coordenar a equipe e melhorar entregas de tarefas, para isso usa do modelo preditivo.

miro

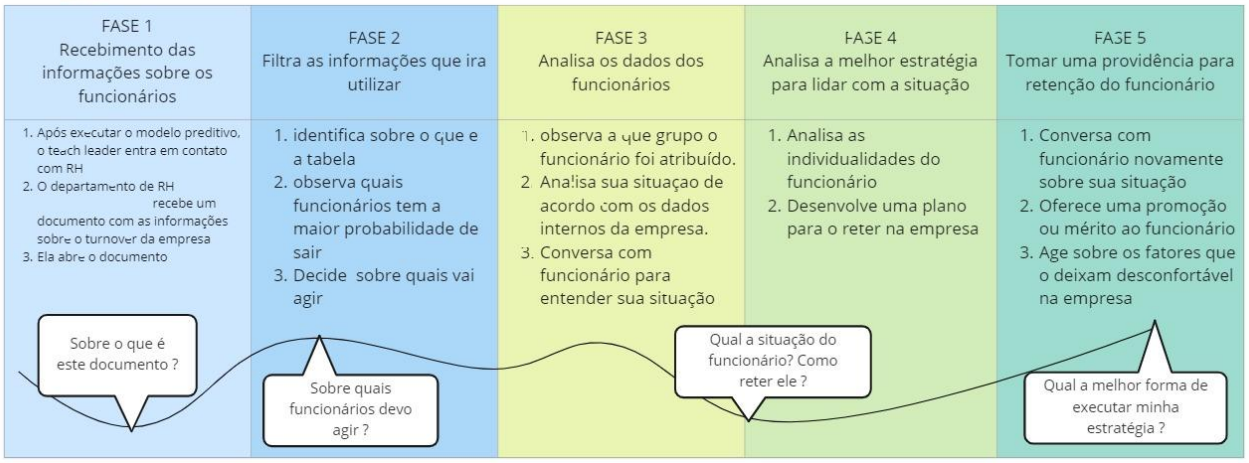
Fonte: Autoria própria

Andrea Silva Costa

Cenário: Retenção dos funcionarios a partir dos resultados obtidos pelo modelo preditivo

Expectativas

Redução do processo de onboarding na empresa, devido a retenção dos funcionários atuais.



Oportunidades

O modelo preditivo possibilita que ela converse com funcionários antes de que eles tomem a decisão de sair, possibilitando agir mais cedo sobre a situação.

Responsabilidades

Orientar os líderes técnicos sobre como podem melhorar a experiência e a jornada dos colaboradores e controlar promoções e salários da empresa.

miro

Fonte: Autoria própria

4.2. Compreensão dos Dados

4.2.1. Descrição dos dados utilizados

Nesta seção segue os dados disponibilizados na "BASE DE DADOS - Base Colaboradores Everymind_Inteli_2020 a 2022vModelo Preditivo.XLSX". A planilha foi disponibilizada pela empresa "Everymind". Base de análise de 475 funcionários.

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	Masculino ou feminino - 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

A planilha de reconhecimento 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.

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

Na planilha Ambiente de Trabalho, os atributos contemplados são relacionados a uma pesquisa de satisfação da empresa para os funcionários, medindo o quão agradável é conviver e trabalhar nesse ambiente.

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	funcionários podem pular, muito insatisfeito, insatisfeito, neutro, satisfeito, muito satisfeito	string
Taxa de confiabilidade	nível de veracidade das respostas	string

Abaixo encontra-se duas tabelas que descrevem os dados utilizados na três planilhas:

Dados Gerais - Utilizados nas 3 planilhas		
Etnia	**Estado Civil**	**Escolaridade**
Amarela	Casado	Ensino Médio

Branca	Divorciado	Ensino Médio Incompleto
Não Informada	Separado	Graduação
Parda	Solteiro	Mestrado
Preta	União Estável	Pós Graduação
-	-	Superior Incompleto
-	-	Técnico

Dados Gerais - Utilizados nas 3 planilhas	
Cargos	**Área de atuação**
Arquiteto	Agência Digital
Arquiteto Sr	AMS
Assistente I	Analytics
Assistente II	BAC
Auxiliar de Limpeza	Best Minds
Comercial IS	BPM
Comercial PI	Commerce
Consultor	Core & Indústrias
Dev Especialista	Core & Indústrias I
Dev Jr	Core & Indústrias II
Dev PI	CPG & Retail
Dev Sr	CPG & Retail I
Diretor	CPG & Retail II
Educação PI	Diretoria
Estagiária	Education
Financeiro Jr	Financeiro
Funcional Especialista	Infraestrutura
Funcional Jr	Integration
Funcional PI	Mkt Cloud
Funcional Sr	People
Gerente	Produtos
Gerente CS Sr	PS
Gerente PV	Vendas
Gerente Sr	-
Gerente Vendas I	-

Gerente Vendas II	-
Gerente Vendas III	-
Infraestrutura Jr	-
Marketing PI	-
Pessoas PI	-
Scrum Master Jr	-
Teste Jr	-
Teste Sr	-
Trainee - Dev	-
Trainee - Funcional	-
Vice Presidente	-

4.2.2. Descrição dos conjuntos de dados

- 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.

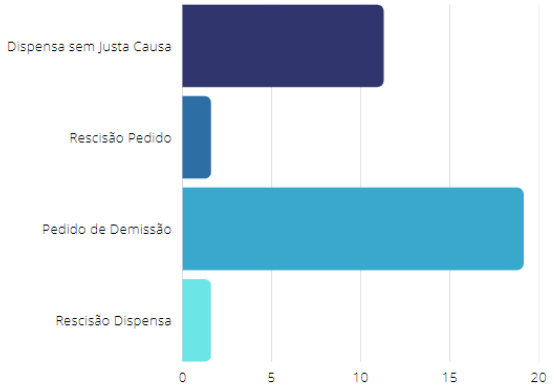
- Descrição dos riscos e contingências relacionados aos dados

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.

4.2.3. Descrição estatística básica dos dados

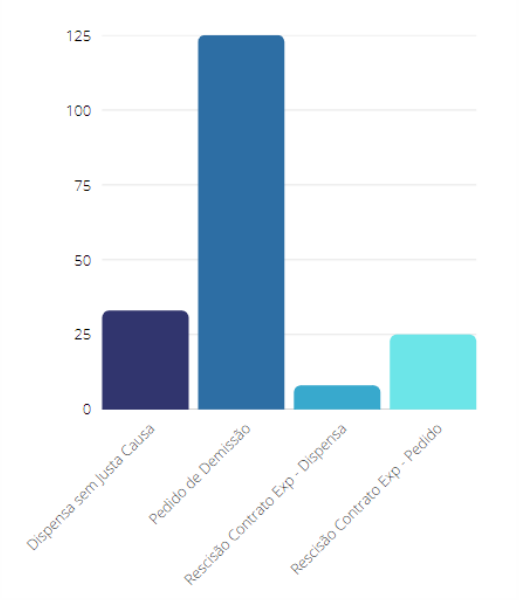
Tempo de permanência

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. EX: Com o gráfico observa-se que a média de permanência de pessoas que pedem demissão é de 1 ano e meio.



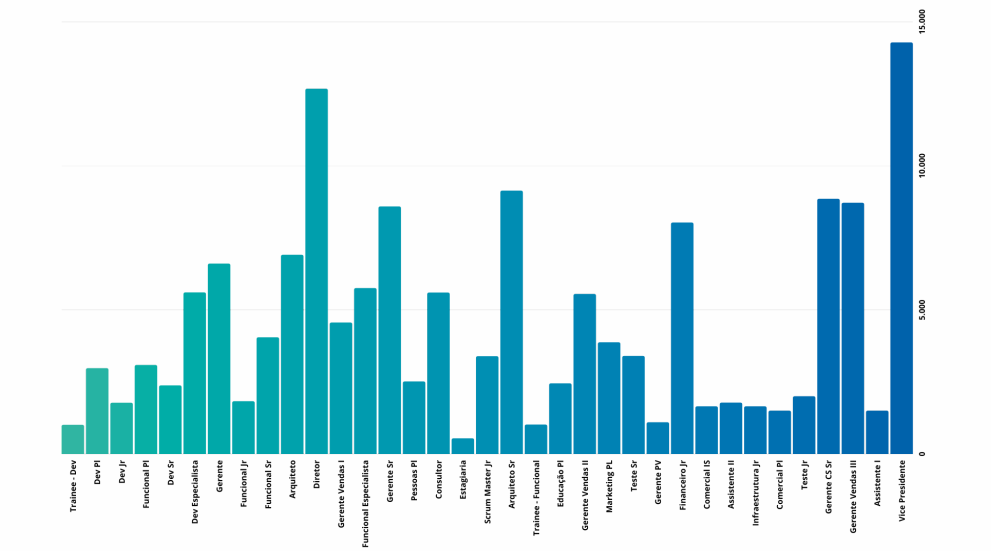
Demissões

Análise das demissões com base nos dados fornecidos, buscando padrões para guiarem o modelo preditivo. Usamos como filtro para buscar correlações com outros dados de funcionários.



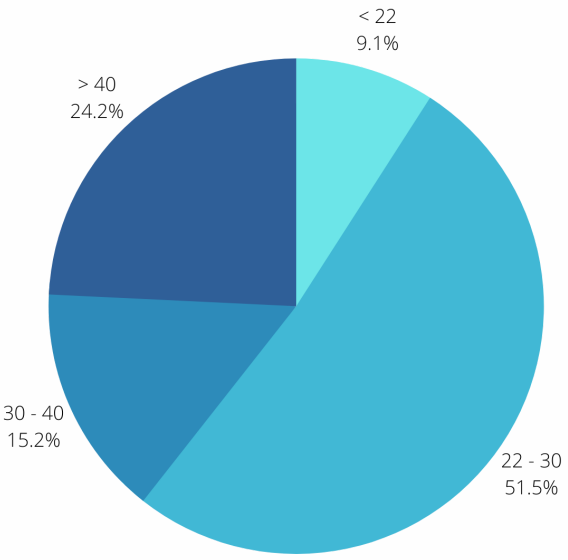
Salário médio dos cargos

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.



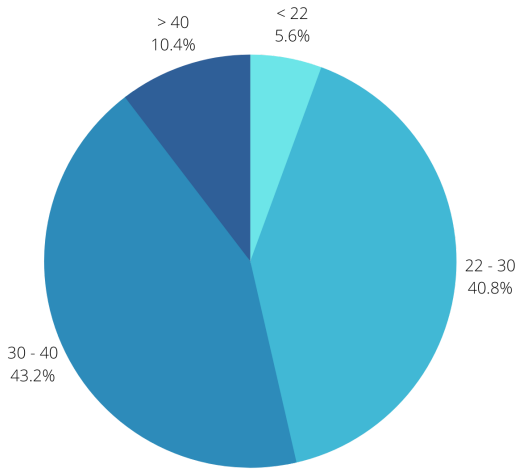
Relação de saída de demissões por justa causa

Esse gráfico apresenta a relação de demissões por justa causa com base na faixa etária do colaborador. Nessa análise é possível verificar que cerca de 50% das demissões se concentram na faixa etária de 22 a 30 anos. Tais dados serão utilizados para compreender quais faixas etárias possuem maior vazão de colaboradores e qual sua forma de saída.



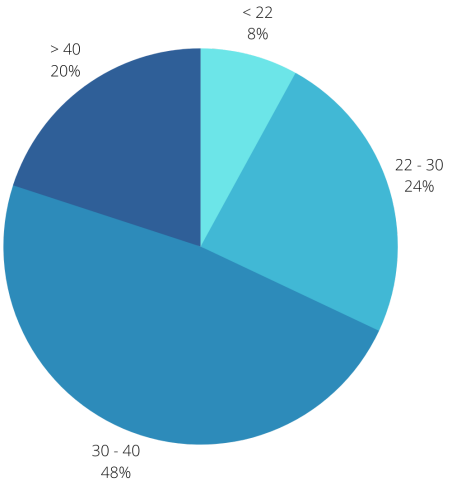
Relação de saída de pedidos de demissão

Esse gráfico apresenta a relação entre os pedidos de demissão com a faixa de idade dos funcionários. Nesse gráfico é possível observar que as faixas de 22-30 anos e 30-40 anos possuem as maiores taxas de pedido de demissão: 43.2% e 40.8%, respectivamente. Essa análise é importante para compreender as taxas de Turnover conforme as idades dos funcionários e poder identificar um padrão para levar à predição da IA.



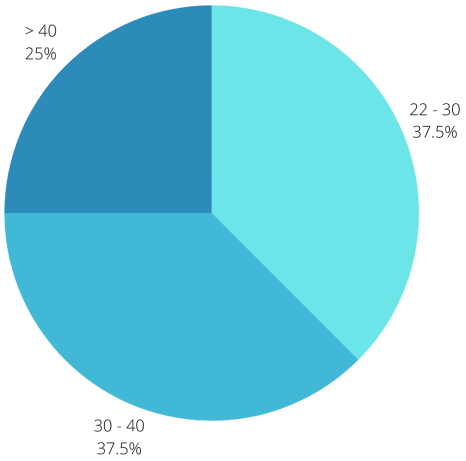
Relação de saída por pedidos de rescisão de contrato

Esse gráfico apresenta a relação de saída por pedido de rescisão de contrato com base na faixa etária do colaborador. Nessa análise é possível verificar que 48% dos pedidos de rescisão se concentram na faixa etária dos colaboradores de 30 - 40 anos. Tais dados serão utilizados para compreender quais faixas etárias possuem maior vazão de colaboradores e qual sua forma de saída



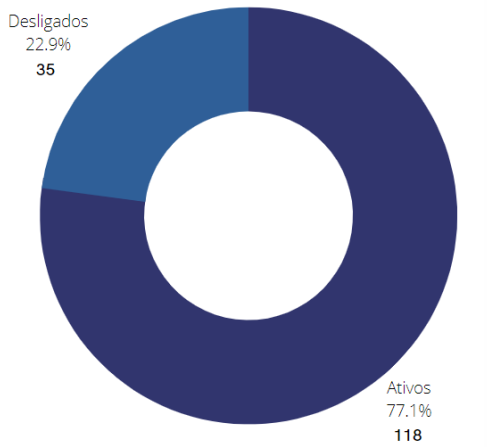
Relação de saída de dispensa por rescisão de contrato

A relação de rescisão de contrato Exp por dispensa com a idade dos funcionários, representa as idades com mais probabilidade de serem dispensadas pela empresa. O gráfico pode ser utilizado para uma comparação com as rescisões de contrato exp por pedido, assim podendo ilustrar como a idade impacta como a pessoa sai da corporação.



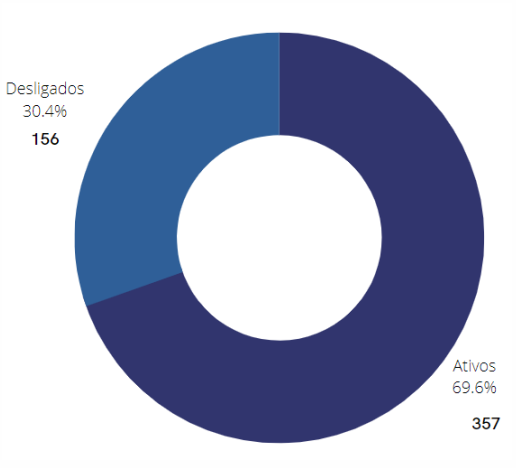
Modalidade (Home Office) X Demissões

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. 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.



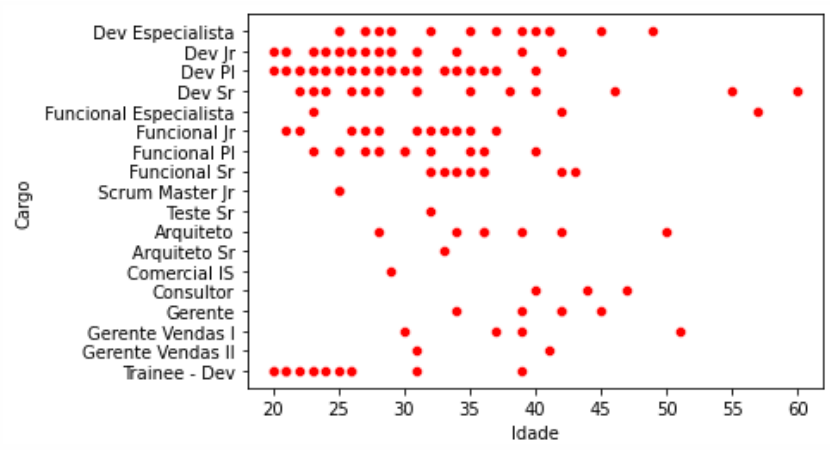
Modalidade (Híbrido) X Demissões

A análise da modalidade pela demissão de funcionários, se dá com base nos colaboradores de São Paulo, estes em modelo de trabalho híbrido. 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 home office, o híbrido ocasiona mais pedidos de demissões.



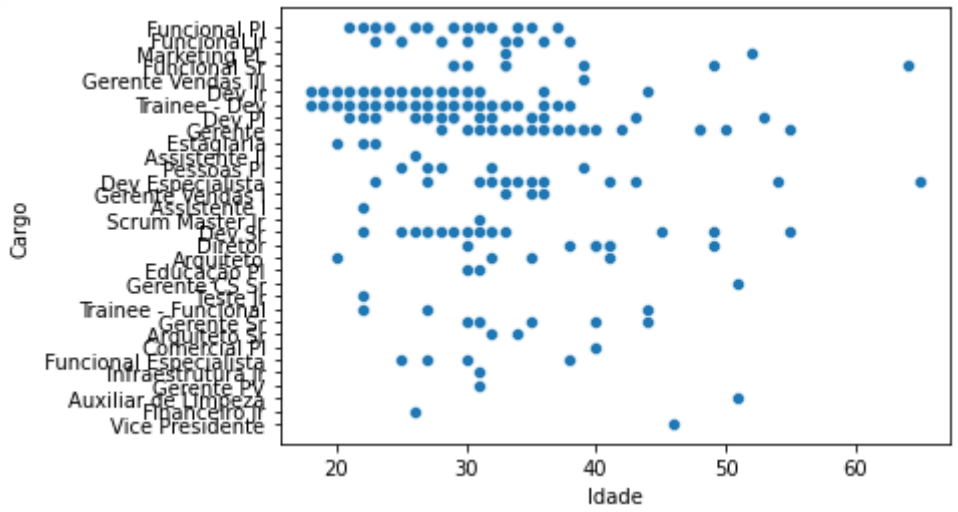
Dispersão de Cargos X Idade (desligados)

A relação da dispersão dos cargos é dada ao comparar quantitativamente os funcionários colaboradores, os quais já foram desligados da empresa, de acordo com suas idades e distribuindo-os pelos cargos que ocupavam. Com essa análise, espera-se visualizar uma tendência de desligamento de acordo com a ocupação de cargo ou nível de experiência profissional relacionada a idade da pessoa. Tal feature foi selecionada para compreender qual a faixa etária que possui maior probabilidade de sair.



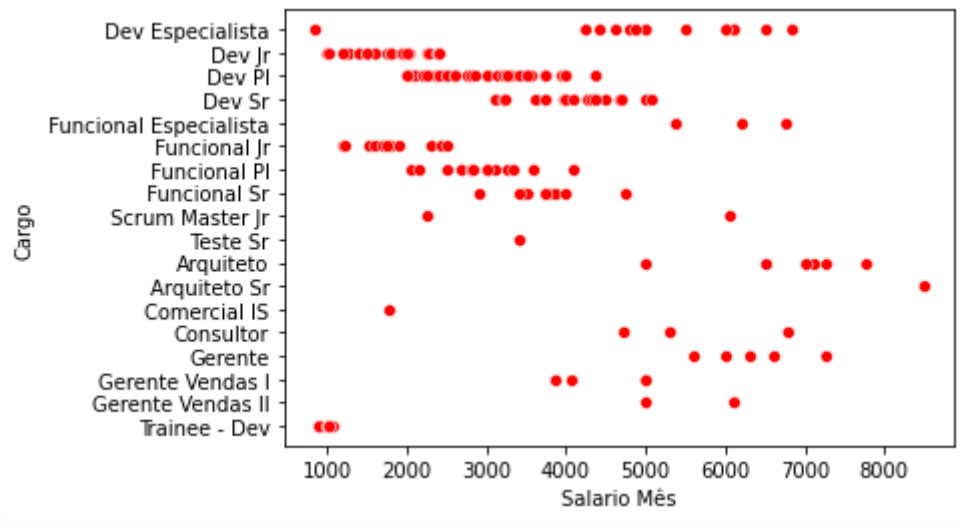
Dispersão de Cargos X Idade (Ativos)

A relação da dispersão dos cargos é dada ao comparar quantitativamente os funcionários colaboradores, os quais estão ativos na empresa, de acordo com suas idades e distribuindo-os pelos cargos que ocupam. Com essa análise, espera-se visualizar uma tendência de retenção de acordo com a ocupação de cargo ou nível de experiência profissional relacionada a idade da pessoa. A escolha desta feature foi feita após a leitura da reflexão e pesquisa de Mauro Wainstock, em um post do "LinkedIn" em que foi compreendido que funcionários mais velhos possuem a tendência de permanecer na mesma empresa onde trabalham. Sendo assim, é interessante investigar a faixa etária dos colaboradores.



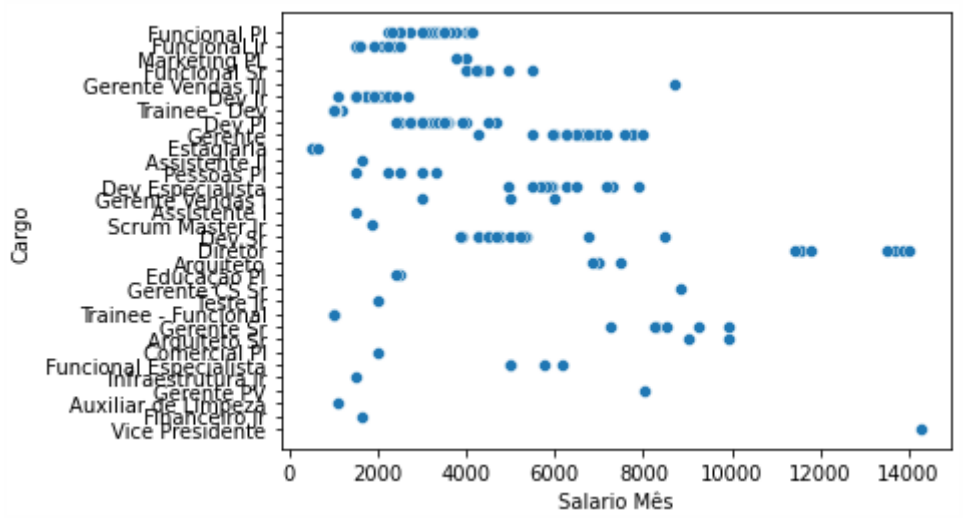
Dispersão de Cargos X Salário Mensal (desligados)

A relação da dispersão dos cargos é dada ao comparar quantitativamente os funcionários colaboradores, que foram desligados da empresa, de acordo com os salários que recebiam mensalmente em relação aos cargos que ocupavam. Com essa análise, espera-se verificar uma tendência de desligamento de acordo com o salário recebido relacionado à distribuição dos cargos. Tal feature foi selecionada com o objetivo de analisar a faixa salarial dos funcionários de cada cargo e entender se existem diferenças entre contribuidores que desempenham a mesma função e se isso é um fator de decisão para saída da empresa.



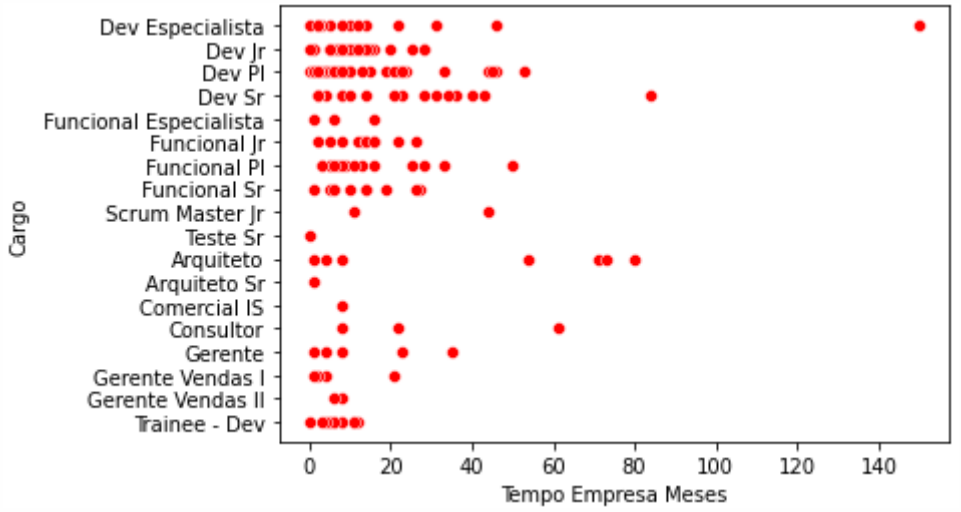
Dispersão de Cargos X Salário Mensal (ativos)

A relação da dispersão dos cargos é dada ao comparar quantitativamente os funcionários colaboradores, que estão ativos na empresa, de acordo com os salários que recebem mensalmente em relação aos cargos que ocupam. Com essa análise, espera-se verificar uma tendência de retenção de acordo com o salário recebido relacionado à distribuição dos cargos. A seleção dessa feature teve objetivo de analisar a faixa salarial dos funcionários de cada cargo e entender qual faixa etária tem maior propensão de permanecer na empresa.



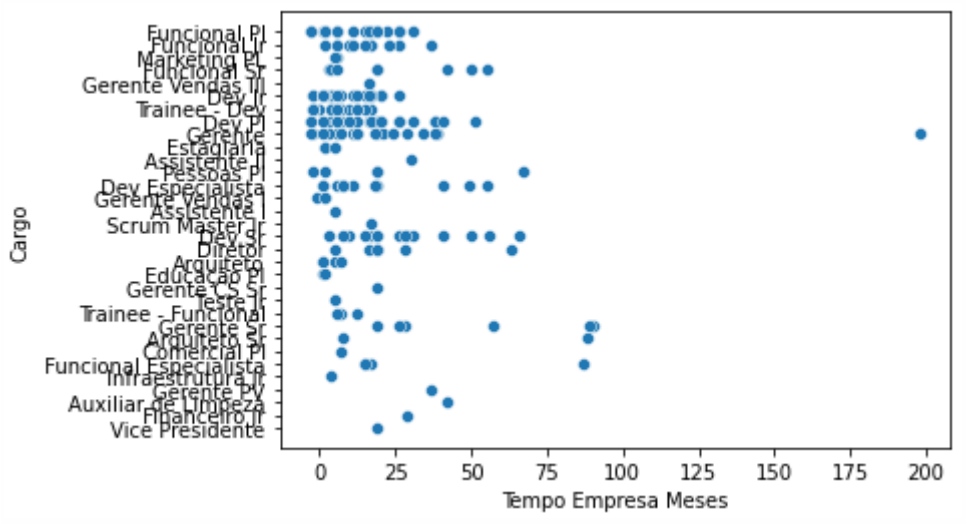
Dispersão de Cargos X Tempo Empresa Meses (desligados)

A relação de dispersão dos cargos é dada ao comparar quantitativamente os funcionários colaboradores, os quais foram desligados da empresa, de acordo com o tempo de permanência na empresa em meses relacionado aos cargos que esses ocupavam. Com essa análise, espera-se observar uma tendência de desligamento de acordo com o tempo de empresa dos funcionários ligados aos cargos que eram ocupados pelos colaboradores. A escolha desta feature foi objetivada para realizar a análise de tempo (em meses) em que os funcionários decidem pela saída da empresa.



Dispersão de Cargos X Tempo Empresa Mêses (ativos)

A relação de dispersão dos cargos é dada ao comparar quantitativamente os funcionários colaboradores, que estão ativos na empresa, de acordo com o tempo de permanência na empresa em meses relacionado aos cargos que esses ocupam. Com essa análise, espera-se observar uma tendência de retenção de acordo com o tempo de empresa dos funcionários ligados aos cargos que eram ocupados pelos colaboradores. A escolha desta feature foi objetivada para realizar a análise de retenção dos funcionários e tempo que permanecem nesta.



Mapa de Calor

Temos como objetivo predizer se o funcionário está propício a sair ou permanecer na empresa, utilizamos o mapa de calor para identificar a intensidade da relação entre variáveis que possivelmente afetam o treinamento e o teste de nosso modelo. Destacamos como mais importantes as correlações das variáveis a seguir:

1: "Numero Promocoes" x "Tempo Medio Promocao".

2: "Salario Mes" x "Tempo Empresa Meses".

3: "Idade" x "Tempo Empresa Meses".

4: "Matricula" x "Numero Promocoes".

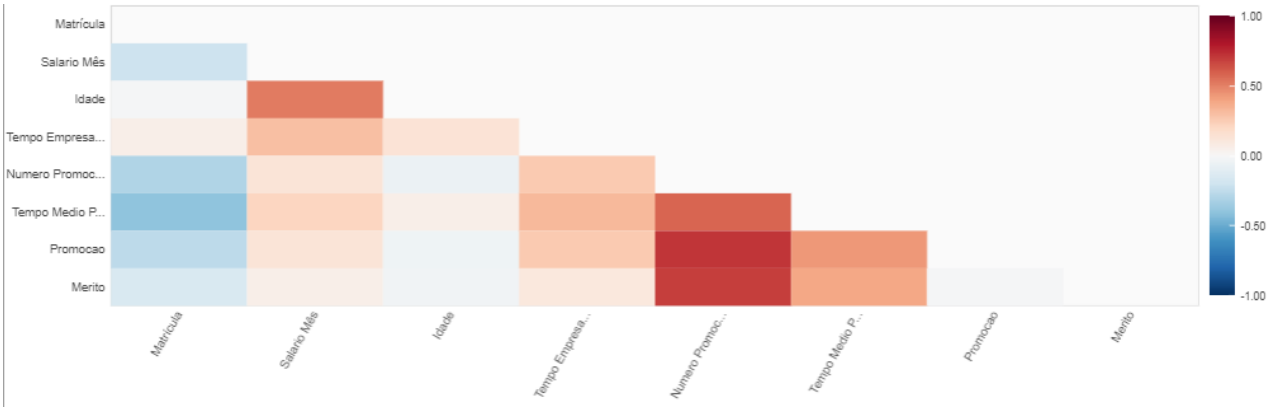
5: "Salario mês" x "Numero Promocoes".

6: "Tempo Empresa Meses" x "Numero Promocoes".

7: "Matricula" x "Tempo médio Promocao".

8: "Idade" x "Tempo Medio Promocoes".

9: "Numero Promocao" x "Tempo Medio Promocao".



Fonte: Autoria Própria

4.2.4. Descrição da predição

O tipo de predição desejada é a classificação. 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.

4.3. Preparação dos Dados - Feature Engineering

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.

1. **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. Nas figuras X e X é ilustrado um antes e depois da formatação com dois exemplos nas colunas "Escolaridade" e "Cidade".

- Antes da remoção dos espaços em branco:

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

Fonte: Autoria Própria

- Depois da remoção dos espaços em branco:

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

Fonte: Autoria Própria

2. 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 Saida, para indicar se o funcionário ainda está ativo na empresa. Nas Figuras X e X abaixo, pode-se notar como os valores foram adicionados, tendo como exemplo a coluna “Tipo de Saída”.

- Antes da adição de valores nos campos:

Dt Admissao	Dt Saida	Tipo Saida
06/06/2022	NaT	NaN
14/02/2022	NaT	NaN
02/03/2022	NaT	NaN
02/12/2019	NaT	NaN

Fonte: Autoria Própria

- Depois da adição de valores nos campos:

Dt Admissao	Dt Saida	Tipo Saida
2022-06-06	2022-08-26	Ativo
2022-02-14	2022-08-26	Ativo
2022-02-03	2022-08-26	Ativo
2019-02-12	2022-08-26	Ativo

Fonte: Autoria Própria

3. 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. As figuras X e X, ilustram o antes e o depois da formação, tendo como exemplo a coluna "Dt Admissao".

- Antes da formatação das datas:

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

Fonte: Autoria Própria

- Depois da formatação das datas:

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

4. 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. Nas Figuras X e X abaixo, pode-se notar como a coluna 'Idade' foi derivada.

- Antes da atribuição das idades na coluna:

Cidade	Area	Idade
Curitiba	CPG&RetailI	0
JoãoPessoa	Core&IndustriasII	0
SãoPaulo	AgenciaDigital	0
SãoPaulo	Core&IndustriasI	0

Fonte: Autoria Própria

- Depois da criação da coluna Idade:

Cidade	Area	Idade
Curitiba	CPG&RetailI	37
JoãoPessoa	Core&IndustriasII	23
SãoPaulo	AgenciaDigital	33
SãoPaulo	Core&IndustriasI	39

Fonte: Autoria Própria

5. 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 X, pode-se notar como a coluna 'Tempo Empresa Meses' foi derivada.

- Derivação do cálculo do 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

6. 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 X, pode-se notar como a coluna 'Tempo Ate Promocao Meses' foi derivada.

- Derivação do tempo de empresa até a promoção:

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

Fonte: Autoria Própria

7. 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. Nas Figuras X e X, é exemplificado o antes e depois da coluna "Nome Completo".

- Antes da criação da coluna Colaborador:

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

Fonte: Autoria Própria

- Depois da da criação da coluna Colaborador - Inicia Vazia:

Idade	Tempo Empresa Meses	Colaborador
37	2	
23	6	
33	6	
39	42	

Fonte: Autoria Própria

- Depois da da criação da coluna Colaborador - Após a separação e inserção dos dados:

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

Fonte: Autoria Própria

8. 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.

9. One Hot encoding

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. As figuras X e X, ilustram o exemplo da coluna “Estado Civil” antes e depois da formatação.

- Antes do One Hot Encoding:

Estado Civil
Casado
Solteiro
Solteiro
Divorciado

Fonte: Autoria Própria

- Depois do One Hot Encoding:

	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

10. Label Encoder

Determinados algoritmos de *Machine Learning* trabalham apenas com dados numéricos, contudo na base de dados da Everymind havia variáveis categóricas, como as de Escolaridade. Sendo assim, foi necessário converter esses dados para variáveis ordinárias, por exemplo:

Ensino Médio Incompleto foi convertido para o número 1, Ensino Médio para o número 2, e assim sucessivamente. Com isso, foi necessário utilizar o método *Label Encoder* para realizar essa conversão. A Figura X e X, exibe uma exemplificação de um antes e depois da coluna com a reformulação da tabela "Escolaridade".

- Antes do Label Encoder:

Escolaridade
Superiorincompleto
Superiorincompleto
Superiorincompleto
Graduação

Fonte: Autoria Própria

- Depois do Label Encoder:

	Escolaridade
0	4
1	4
2	4
3	3

Fonte: Autoria Própria

11. 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 mostra-se 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.

12. 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 X e X, exibe uma exemplificação da relação entre os ativos e os desligados com os 9 primeiros cargos.

- Relação do número de funcionários desligados por cargo:

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

Fonte: Autoria Própria

- Relação do número de funcionários ativos por cargo:

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

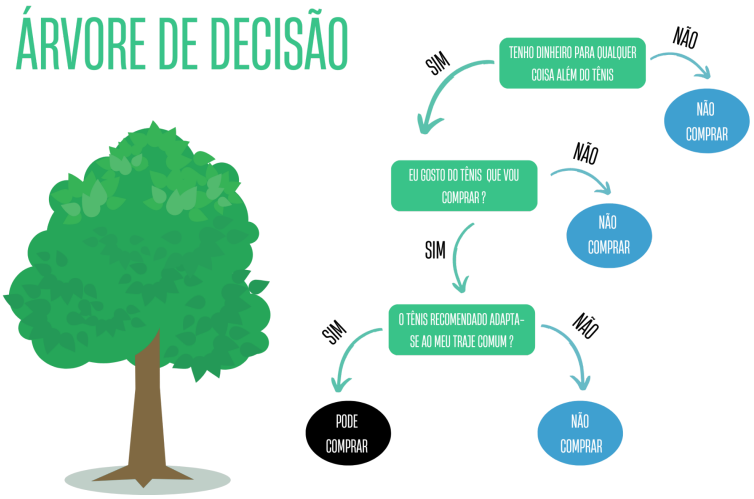
Fonte: Autoria Própria

4.4. Modelagem

4.4.1 Árvore de decisão

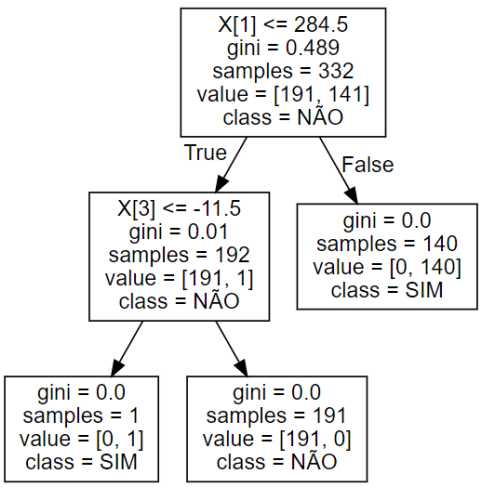
1. Funcionamento

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 X, é ilustrado um exemplo de uma árvore de decisão:



Fonte: Autoria própria

Exemplo de árvore de decisão criada pelo algoritmo:



Fonte: Autoria própria

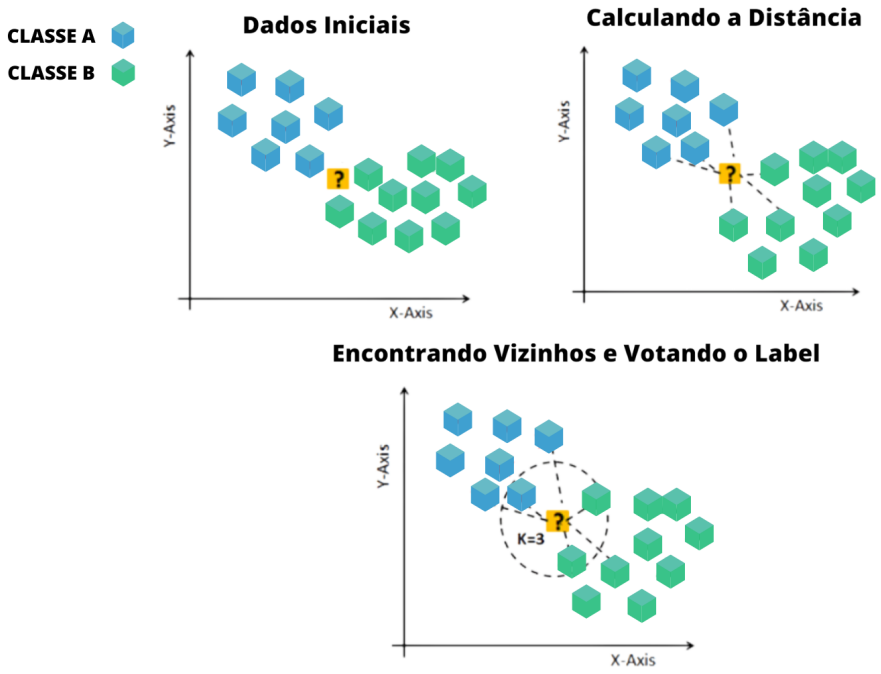
2. **Como adequa-se ao problema ?**

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.

4.4.2 KNN (K Nearest Neighbor)

1. **Funcionamento**

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 X abaixo, é ilustrado um exemplo do algoritmo KNN:



Fonte: Autoria Própria

2. **Como adequados ao problema?**

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.

4.4.3 Naive Bayes

1. **Funcionamento**

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 a fórmula utilizada para o cálculo do algoritmo.

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

Na qual, P(AIB) – representa a probabilidade do evento A acontecer sendo que b já ocorreu. e P(BIA) – 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.

2. **Como adequados ao problema?**

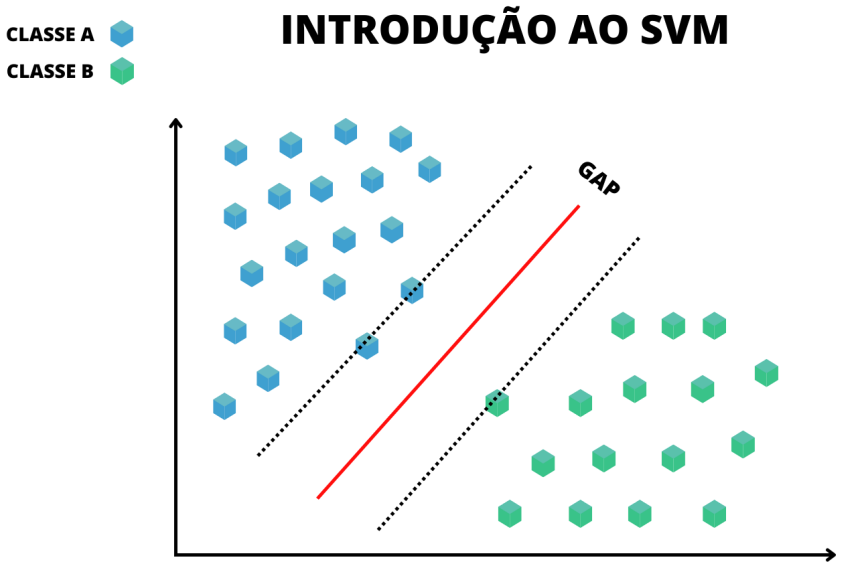
O algoritmo de Naive Bayes foi adequado ao problema que está sendo trabalhado. Isto se deve ao fato de ele ser recomendado para algoritmos classificatórios, que é o caso do projeto. Assim, levando em conta as classificações possíveis em nosso sistema e as variáveis relacionadas a elas , disponibilizadas 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.

4.4.4 Support Vector Machine(SVM)

1. Funcionamento do 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 X abaixo:



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.

2. Como o modelo é adequado ao problema?

O algoritmo SVM tem sua adequação a solução mediante a classificação dos dados escolhidos em categorias, sendo possível a visualização definida dos atributos referente 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.

4.5. Avaliação

4.5.1 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 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 eviesamento 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

4.5.2 Utilização de variáveis

Para os teste dos modelos foi utilizado o conjunto de dados apresentados na imagem abaixo, a variável "árvore" contempla os seguintes dados: Matrícula, Colaborador, Idade, Tempo de Empresa em Meses, 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.

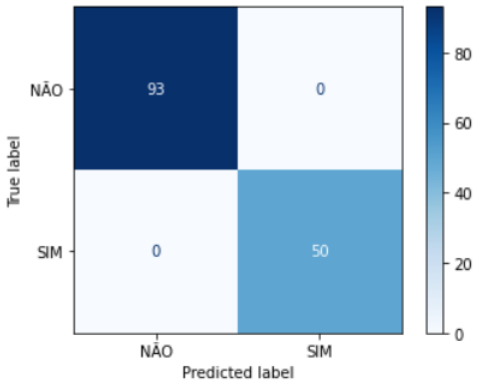
```
arvore = pd.DataFrame()
arvore = pd.concat([arvore ,df['Matrícula']], axis = 1)
arvore = pd.concat([arvore ,df['Colaborador']], axis = 1)
arvore = pd.concat([arvore ,df['Idade']], axis = 1)
arvore = pd.concat([arvore ,df['Tempo Empresa Meses']], axis = 1)
arvore = pd.concat([arvore ,df['Salario Mês']], axis = 1)
arvore = pd.concat([arvore ,dummie_cargos], axis = 1)
arvore = pd.concat([arvore ,dummie_genero], axis = 1)
arvore = pd.concat([arvore ,dummie_civil], axis = 1)
arvore = pd.concat([arvore ,dummie_estado], axis = 1)
arvore = pd.concat([arvore ,dummie_cidade], axis = 1)
arvore = pd.concat([arvore ,dummie_area], axis = 1)
```

Fonte: Autoria própria

4.5.3 Estratégia de avaliação do modelo - Árvore de decisão

- **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. Abaixo é ilustrado uma visualização gráfica da matriz de confusão obtida no modelo de árvore de decisão para a solução proposta:



Fonte: Autoria própria

- **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 Figura X abaixo é ilustrado a Acurácia obtida pelo modelo:

```
Acuracidade (treino):  1.0
Acuracidade (teste):  1.0
```

Fonte: Autoria própria

A acurácia preliminar obtida no modelo foi de 1.0 tanto para treino, quanto para o teste e isto significa que, nesse caso, para o modelo de árvore de decisão, o algoritmo obteve 100% de acerto.

- **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. A Figura X, ilustra a precisão preliminar obtida pelo modelo, demonstrado na tabela abaixo como "precision":

	precision
NÃO	1.00
SIM	1.00

Fonte: Autoria própria

A precisão obtida no modelo foi de 100% para os verdadeiros valores preditos de retenção e também de 100% para os verdadeiros valores preditos de saída, o que significa, que nesse modelo dentre os valores que queríamos prever, todos foram realmente previstos.

- **Recall**

O recall foi elencado pois tal método apresenta classe predita em relação ao que realmente espera-se de resultado. Sendo assim o Recall mostra dentre todos os casos classificados como Positivo, quanto está correto.

	precision	recall
NÃO	1.00	1.00
SIM	1.00	1.00

Fonte: Autoria própria

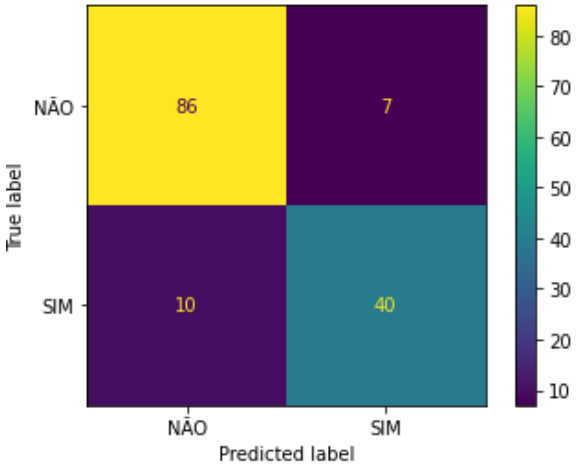
O recall obtido no modelo foi de 100% para os verdadeiros valores preditos de retenção e 100% para os verdadeiros valores preditos de saída, o que significa, que nesse modelo dentre os valores que prevemos, todos deveriam ser realmente previstos nessas classes.

4.5.4 Estratégia de avaliação do modelo - KNN

As estratégias de avaliação de modelo escolhidas foram as seguintes:

- **Matriz de confusão**

A matriz de confusão foi elencada, pois com a geração da matriz de confusão, anteriormente explicada, é possível observar a quantidade de erros e acertos da predição feita pelo modelo e, a partir dessa matriz é possível parametrizar outras métricas, como a acurácia, precisão, recall e f1-score. Abaixo segue a visualização gráfica da matriz de confusão obtida no modelo KNN:



Fonte: Autoria própria

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.

- **Acurácia**

A acurácia foi elencada, já que a partir desta pode-se observar o quanto o modelo conseguiu prever os valores que se quer observar em relação ao que se quer observar e o valor real obtido. Segue abaixo a Acurácia obtida pelo modelo:

```
Acuracidade (treino):  0.9337349397590361
Acuracidade (teste):  0.8811188811188811
```

Fonte: Autoria própria

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.

- **Precisão**

A precisão foi elencada, como se quer observar, dos valores, previstos quais de fato pertencem à classe que se quer obter. Segue abaixo a demonstrado na tabela “precision”, que apresenta a precisão obtida pelo modelo:

	precision
NÃO	0.90
SIM	0.85

Fonte: Autoria própria

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 queria-se prever, 90% e 85% foram realmente previstos.

- **Recall**

O recall foi elencado, pois a partir do que se quer observar como valores previstos, esse método apresenta o quanto observa-se da classe predita em relação ao que realmente deveria ser predito dessa classe. Abaixo é apresentado a demonstrado na tabela “recall”, com os resultado obtido pelo modelo:

	precision	recall
NÃO	0.90	0.92
SIM	0.85	0.80

Fonte: Autoria própria

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 previu-se 92% e 80% deveriam ser realmente previstos nessas classes.

4.5.5 Estratégia de avaliação do modelo - Naive Bayes

Durante o desenvolvimento do modelo utilizou-se diversas estratégias de avaliação do modelo. Sendo um deles a acurácia, proximidade de um resultado experimental com o seu valor

de referência real, é uma das estratégias de avaliação escolhida, pois com ela é possível identificar o grau de exatidão geral do modelo. O algoritmo de Naive Bayes obteve uma acurácia

```
Acuracidade (treino):  0.8373493975903614
Acuracidade (teste):  0.6153846153846154
```

de 0.62. Ou seja, acertou 62% das predições.

Fonte: Autoria própria

Outra das estratégias utilizadas foi a da precisão. Essa escolha se deve ao fato de a precisão informar a porcentagem de acertos referentes a cada classificação possível, assim sendo possível analisar os cenários individualmente. A precisão obtida com o modelo de Naive Bayes foi de 82% para não(funcionários que ficaram na empresa) e de 47% para sim(funcionários que saem da empresa).

	precision
NÃO	0.82
SIM	0.47

Fonte: Autoria própria

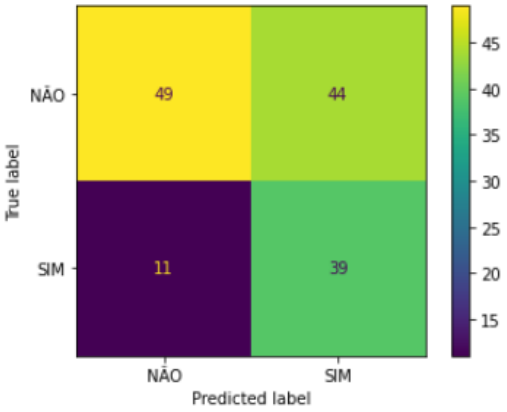
Por fim, outra das estratégias utilizadas foi o recall. Esta escolha se deve ao fato de o recall ser capaz de fornecer um resultado similar a precisão, mas com alta penalidade em falso negativos ou falso positivos, assim sendo possível visualizar se o modelo possui um alto índice de falso negativos ou positivos. O recall obtido no modelo de Naive Bayes testado foi de 53% para os funcionários que não saíram da empresa e de 78% para os que ficaram.

	precision	recall
NÃO	0.82	0.53
SIM	0.47	0.78

Fonte: Autoria própria

3. Resultados preliminares obtidos

Através da utilização do modelo de Naive Bayes. 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 49 verdadeiros negativos(funcionários preditos a ficar que realmente ficaram), 44 falso negativos(funcionários preditos a ficar que saíram), 11 falso 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.



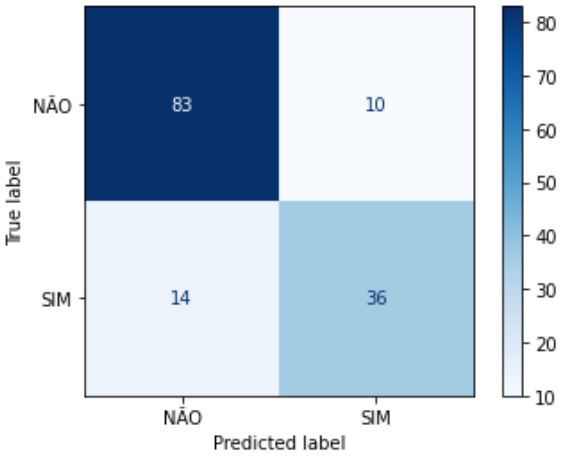
Fonte: Autoria própria

4.5.6 Estratégia de avaliação do modelo - SVM

Para a avaliação de modelo, utilizou-se das seguintes métricas, sendo elas: 1) Matriz de confusão; 2)Acurácia; 3) Precisão; e 4) Recall. Pode-se visualizar a descrição da utilização das mesmas nos textos abaixo.

- **Matriz de confusão**

A partir da matriz de confusão, explicada anteriormente, é possível observar a quantidade de erros e acertos da predição feita pelo modelo e, a partir dessa matriz é possível parametrizar outras métricas, como a acurácia, precisão, recall e f1-score. A Figura X, ilustra o resultado da matriz de confusão prevista para o modelo SVM.



Fonte: Autoria própria

- **Acurácia**

A acurácia, explicada anteriormente, utilizada para o modelo SVM, exibe o quanto o modelo conseguiu prever os valores corretos ao objetivo proposto, em relação ao que se quer observar e o valor real obtido. Na Figura X é exibido os resultados obtidos para o modelo SVM.

```
Acuracidade (treino):  0.8132530120481928
Acuracidade (teste):  0.8321678321678322
```

Fonte: Autoria própria

- **Precisão**

A precisão, explicada anteriormente, utilizada para o modelo SVM, apresenta dos valores previstos quais destes de fato pertencem à classe desejada. Na Figura X, é ilustrado o resultado obtido.

```
          precision

NÃO            0.86
SIM            0.78
```

Fonte: Autoria própria

- **Recall**

O recall, explicado anteriormente, utilizado para o modelo SVM, apresenta o quanto observamos da classe predita em relação ao que realmente deveria ser predito dessa classe.

	recall
NÃO	0.89
SIM	0.72

Fonte: Autoria própria

4.5.7 Resultado geral a partir das métricas de avaliação dos modelos

A partir da análise das métricas obtidas de cada modelo, pode-se inferir que o modelo ao qual apresentou um resultado mais satisfatório com os dados testados até o momento é o algoritmo de Árvore de Decisão. Por apresentar 100% de acurácia tanto no modelo de teste como de treino, pela base de dados trabalhada ser consideravelmente pequena, é possível o modelo acertar totalmente a saída, entretanto não é comum. Pretende-se posteriormente identificar o fator decisivo para este resultado e qual o protocolo a seguir quando acontece tal retorno.

5. Conclusões e Recomendações

Escreva, de forma resumida, sobre os principais resultados do seu projeto e faça recomendações formais ao seu parceiro de negócios em relação ao uso desse modelo. Você pode aproveitar este espaço para comentar sobre possíveis materiais extras, como um manual de usuário mais detalhado na seção “Anexos”.

Não se esqueça também das pessoas que serão potencialmente afetadas pelas decisões do modelo preditivo, e elabore recomendações que ajudem seu parceiro a tratá-las de maneira estratégica e ética.

6. Referências

Nesta seção você deve incluir as principais referências de seu projeto, para que seu parceiro possa consultar caso ele se interessar em aprofundar.

Utilize a norma ABNT NBR 6023 para regras específicas de referências. Um exemplo de referência de livro:

SOBRENOME, Nome. **Título do livro**: subtítulo do livro. Edição. Cidade de publicação: Nome da editora, Ano de publicação.

WAINSTOCK, Mauro.

Anexos

Utilize esta seção para anexar materiais como manuais de usuário, documentos complementares que ficaram grandes e não couberam no corpo do texto etc.