Firing Machine
Everymind

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
09/08/2022	Kil Mateus	1.1	atualização da secção 4 (4.1)
12/08/2022	Kil Mateus	1.2	atualização das secções 1, 2, 3 e 4 (4.2)
22/08/2022	Gabriel Nascimento	2.1	atualização da secção 4 (4.1.6)
26/08/2022	Gabriel Nascimento	2.2	atualização da secção 4 (4.1.6, 4.1.7, 4.2,4.3)
08/09/2022	Ana Clara Zaidan	3.1	atualização da secção 4 (4.4)
09/09/2022	Ana Clara Zaidan	3.2	atualização da secção 4 (4.5 e 4.3)
12/09/2022	Ana Clara Zaidan	4.1	atualização da secção 4 (4.3- a, b, c, d) (correção)
12/09/2022	Mariana Paula	4.2	atualização da secção 4.1 (correção)
15/09/2022	Daniel Dávila	4.3	correção geral
23/09/2022	Vitória	4.4	atualização das secções 4.3, 4.4, 4.5
25/09/2022	Vitória	4.5	atualização das secções 4.3, 4.4, 4.5
03/09/2022	Mariana Paula	5.1	revisão geral do documento, focada nas seções 4.3, 4.4 e 4.5
10/06/2022	Mariana Paula	5.2	revisão geral do documento inteiro e atualização do sumário

Sumário

1. Introdução

O parceiro de negócio é a Everymind. Uma das empresas de consultoria estratégica para Salesforce com maior market share do país consoante seu próprio site[1], a Everymind é reconhecida pelo ISG Provider Lens como o Salesforce Ecosystem Partner brasileiro líder nas esferas de serviços de implementação para soluções analíticas em Salesforce, em Marketing Cloud Midmarket, e em Core Clouds Midmarket. Nesse mesmo relatório, a Everymind também recebeu o título de líder brasileira para 3 outras áreas: "Multicloud implementation & Integration Services for Large Enterprises", e "Managed Application Services" para Large Enterprises e para Midmarket. Isso é demonstrador da grande força exercida continuamente pela Everymind no mercado de Salesforce do Brasil.

Focando em comercializar soluções utilizantes de tecnologias da análise de dados - tecnologias cuja origem é a Salesforce - a Everymind, inspirada pelo conceito de boutique, alfaiata esse processo conforme a demanda de cada cliente. Tais clientes são, em maioria, grandes empresas. Por consequência de um modelo de negócio tão polido, os produtos oferecidos pela Everymind apresentam alto grau de eficiência. E o fato de que a Everymind é composta por 250 indivíduos altamente competentes em suas respectivas áreas vastamente acrescenta para que o processo aqui delineado ocorra com extrema suavidade.[2]

Além disso, é importante ressaltar o ERP (*Enterprise Resource Planning* [Sistema Integrado de Gestão Empresarial]) entre a Everymind e a Salesforce. Tendo em vista que essa ferramenta é importante para a gestão empresarial, já que ela garante a integração de todos os dados e funções da empresa em um único sistema. Como isso é uma forma de aprimoramento de negócios, ambas empresas vendem esse tipo de serviço.

---

[1] https://mcjb15vjp4x3shyj9vwqlqvwnky1.pub.sfmc-content.com/vczccluo15c, acesso em 6/10/2022.
[2] https://www.everymind.com.br/sobre-nos/, acesso em 6/10/2022.

2. Objetivos e Justificativa

2.1. Objetivos

Os objetivos gerais do projeto consistem em diminuir o turnover de funcionários, visando diminuir a perda contínua de colaboradores e a necessidade de contratação, que reduziria investimentos monetários e operacionais em onboarding para novos funcionários.

Os objetivos específicos do projeto, por outro lado, consistem em obter dados sobre tendências de saída da empresa para cada funcionário, e, a partir disso, entender os principais motivos que levam os colaboradores a quererem deixar a empresa. Após obter tal informação, será possível criar estratégias para reduzir a perda de funcionários, o que terá um impacto direto nos investimentos monetários e operacionais realizados pela empresa para com a chegada de novos funcionários.

2.2. Justificativa

A proposta de solução é a construção de um modelo preditivo (algoritmo de machine learning) que, após identificar padrões nos dados relacionados ao contexto da saída de funcionários da empresa, possibilitará ação imediata sobre eles.

O sucesso do modelo que propomos engendra, dentre outros benefícios, a redução do turnover, maior alinhamento dos funcionários à cultura da empresa, e maior orientação ao possível impacto de mudanças na governança corporativa.

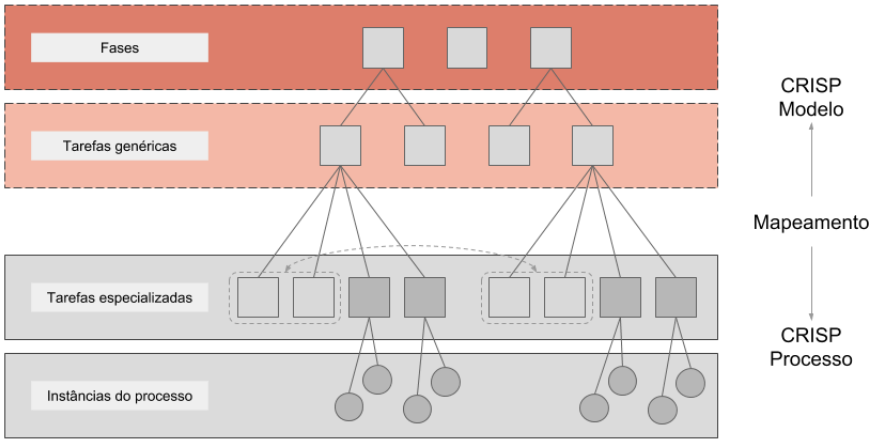
O método usado em nossa solução possui destaque sobre o dos competidores por sua natureza inerentemente ágil, flexível, e rizomática. Com a capacidade de realizar análises sem intervenção humana; com potencial de processamento de grande quantidade de dados; e eficaz em identificar padrões em um período de tempo extremamente curto e com grande precisão, indubitavelmente, o algoritmo que desenvolvemos não pode ser subestimado.

3. Metodologia

3.1. CRISP-DM

O "Cross Industry Standard [for] Data Mining", ou CRISP-DM, é, em síntese, a norma universal para realização de mineração de dados, e possui como protocolo um processo que segue uma hierarquia de crescentes níveis de abstração.

3.1.1. Hierarquia



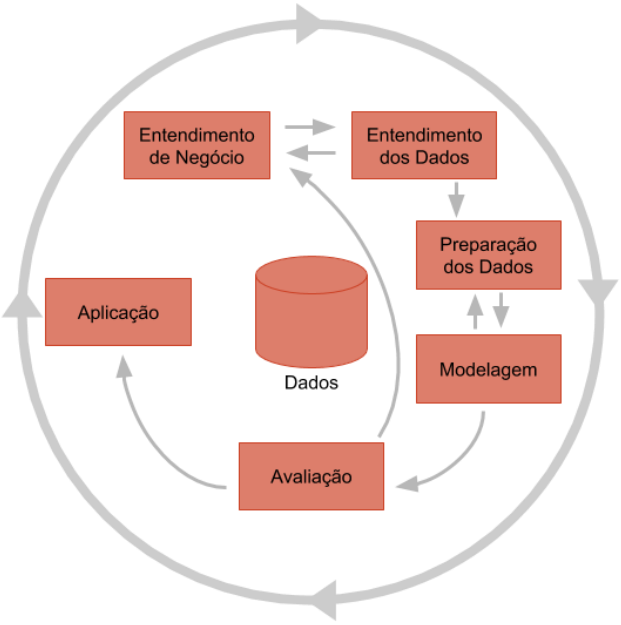
O primeiro nível dessa hierarquia, denominado "phase", consiste na ordenação do processo de mineração em determinado número de fases, sendo cada uma dessas fases um conjunto de determinada quantidade de tarefas. O segundo nível é o conjunto de tarefas em questão, e é denominado "generic", posto que tais tarefas são classificadas como genéricas.

Já o terceiro nível, "specialized task", é onde determina-se como serão realizadas as tarefas em cada contexto específico. O nome do quarto nível aptamente resume a sua função: "process instance" - o local em que são gravadas as decisões, ações, e resultados de cada engajamento que ocorre em cada um dos demais níveis durante o processo de mineração.

Além disso, é importante mencionar que, segundo especialistas, existem dois pré-

requisitos para aplicar mineração de dados em um empreendimento: o primeiro é o entendimento do negócio; o segundo, o entendimento dos dados sobre os quais ocorrerá a mineração.

3.1.2 Processo de modelagem



Fonte: https://medium.com/@kvmoura/crisp-dm-79580b0d3ac4

Diagrama de funcionamento do modelo CRISP-DM (Adaptada de Wirth, 2000). | Download Scientific Diagram

O processo de modelagem pelo CRISP-DM é dividido em 6 fases, em que, como mostrado na figura acima, as setas indicam as dependências mais importantes e frequentes entre as fases, mas elas não constituem uma ordem fixa e imutável, pois esse é um processo iterativo e não engessado.

Entendimento de Negócio: Define o entendimento do ponto de vista do negócio, qual é o produto final, o que o meu cliente quer, qual é o problema a ser resolvido- relacionado a Ciência de Dados. Essa é uma das fases mais importantes do método CRISP-DM, pois, quanto mais claro e nítido estiverem essas informações, mais precisa será a solução desenvolvida.

Entendimento dos Dados: É a compreensão e levantamento de hipótese de como aquela base de dados que foi fornecido fará sentido para o desenvolvimento da solução final, se possui qualidade e/ou quantidade suficientes de informações, se tem relação com o problema e principalmente se ela está relacionada com o entendimento do negócio, pois só é possível progredir se esses insights forem validados.

Preparação dos dados: Essa é a parte onde os dados serão tratados. Informações em formato de String por exemplo, necessitam de um tratamento devido ao funcionamento dos modelos de predição citados nesse documento, e entre outros, como normalização/padronização, Oversampling, eliminação de variáveis outliers, etc. Geralmente, essa etapa se repete várias vezes até que o modelo retorne resultados satisfatórios.

Modelagem: Aqui começa a parte de Mineração dos Dados. Essa parte acontece quando os dados tratados passam por diversos processos de predição a fim de descobrir padrões/sequências que fazem um determinado evento acontecer. O tratamento de dados aqui tem um enorme peso nos resultados, por isso, geralmente a etapa anterior é revisada e refinada até obter uma predição com um bom nível de confiabilidade. Vale ressaltar que existem diversos modelos de predição, sendo assim, tende-se encontrar aquele de melhor performance e atende a proposta.

Avaliação: Depois da modelagem, a partir de diferentes métricas de avaliação (como acurácia, precisão, f1, matriz de confusão e curva roc) são escolhidos os modelos que conseguiram obter as melhores performances, com uma boa confiabilidade sobre a base de dados tratadas. Logo depois, os requisitos são validados com os stakeholders, a fim de passar para a aplicação, e, caso esses requisitos não tenham sido todos atendidos, voltamos para a etapa inicial de entendimento de negócio, visto que este é um processo iterativo e cíclico, em que todos os processo precisam ser revisados até uma solução concreta e que atenda ao critério de sucesso seja encontrada.

Aplicação: Quando o projeto for validado e atender os requisitos, ele é entregue para o cliente, para, enfim, ser colocado em prática.

3.2. Ferramentas

Abaixo encontram-se listadas as principais ferramentas utilizadas pelo grupo para realizar o projeto, seguidas de especificação de seus respectivos usos dentro desse mesmo contexto.

- Google Colaboratory - desenvolvimento do código (escrito na linguagem Python).
- GitHub - setor de entregas do código e da documentação.
- Google Docs - documentação.
- Google Drive - repositório da base de dados.

4. Desenvolvimento e Resultados

4.1. Compreensão do Problema

A Everymind é uma empresa de tecnologia que presta serviços de consultoria utilizando Salesforce, e está inserida em um mercado muito competitivo e aquecido, onde a disputa por profissionais de tecnologia é muito alta. Com isso, a alta taxa de rotatividade de colaboradores também é um problema que ela enfrenta, corroborando com um fenômeno de todo o mercado de tecnologia. Dessa forma, o problema apresentado se refere à exposição a esse fenômeno de alto turnover no mercado de trabalho atual, que pode acabar por afetar os funcionários da EveryMind - que deseja prever e impedir que isso aconteça.

4.1.1. Contexto da indústria

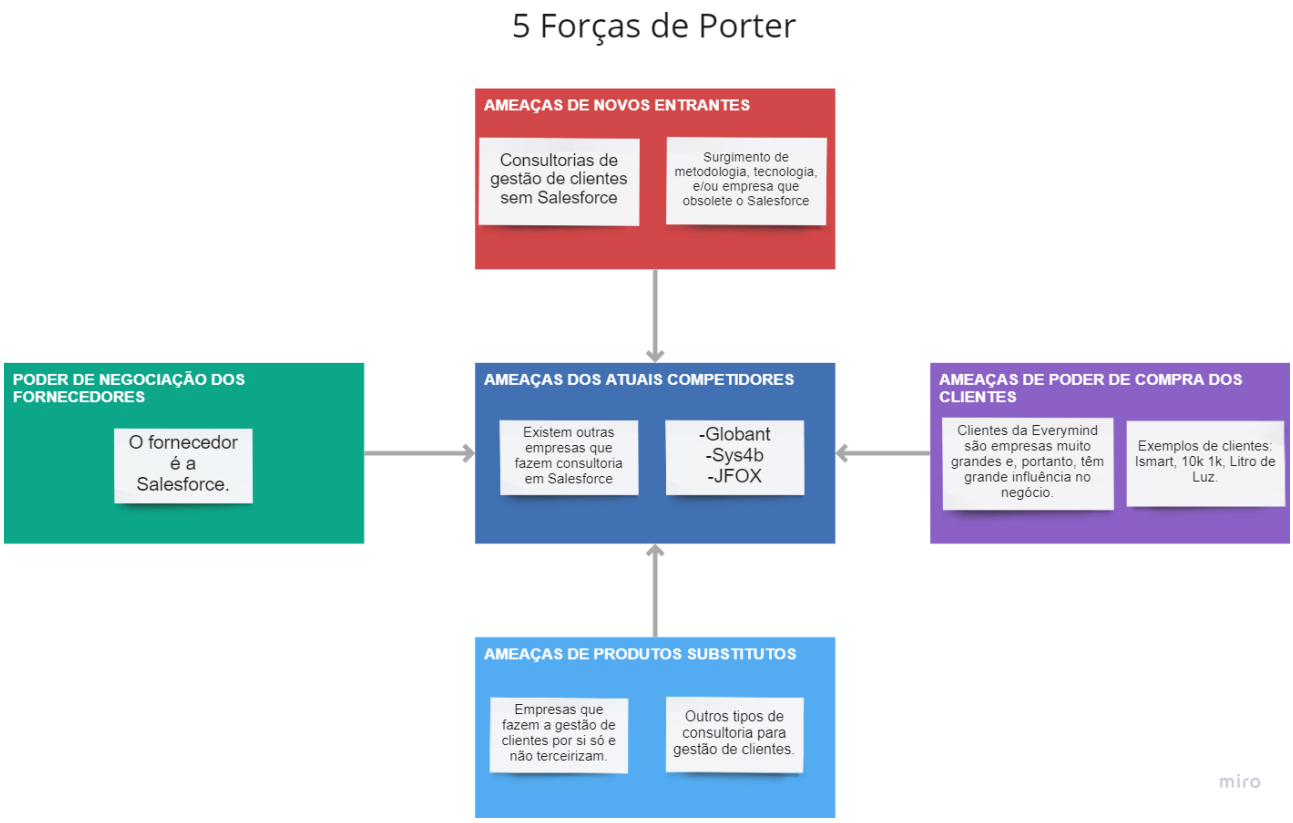
Para compreender o contexto de indústria no qual se encontra a Everymind, deve-se de início entender o modelo de negócio da Everymind, que é, essencialmente, o comércio de soluções utilizantes de tecnologias da análise de dados - tecnologias cuja origem é a Salesforce. Esse processo, como dita o conceito de boutique que muito inspira a Everymind, é alfaiatado conforme a demanda de cada cliente, que são, em maioria, grandes empresas. Por consequência de um modelo de negócio tão polido, os produtos oferecidos pela Everymind apresentam alto grau de eficiência, especialmente se comparados com aqueles de seus principais competidores: Sys4b, Globant, e JFOX - consultorias de Salesforce prestadoras de serviço semelhante àquele oferecido pela Everymind.

Já para compreender o contexto de indústria de inteligências artificiais no eixo do mercado relevante para este documento, é, em síntese, necessário entender onde são majoritariamente utilizadas. Mais comumente, o mercado de consultorias as utiliza para melhorar a experiência de usuários; para conseguir identificar potenciais compradores; para analisar o comportamento de clientes; para monitorar o marketplace; e como alavanca para o início do uso de Salesforce. Além disso, os ativos de TI recorrem à inteligência artificial para anteciparem problemas de

desempenho, automatizando as devidas correções antes que tais problemas sejam detrimentosos à performance.

Além disso, é importante ressaltar que, não só a Everymind sofre com a alta rotatividade, segundo o site ITforum, a taxa de turnover nas empresas de tecnologia podem chegar a 13% , o que nos faz pensar em uma solução para que a empresa se destaque no contexto onde está inserida.

Nesse sentido, o diagrama abaixo “5 Forças de Porter” organiza e ilustra os principais atores do Contexto de Indústria da EveryMind:



https://miro.com/app/board/uXjVOgCyebI=/?share_link_id=925777377534

4.1.2. Análise SWOT

Um pré-requisito para a compreensão da análise SWOT é a compreensão do respectivo acrônimo. "S" representa *strengths;* significando os pontos fortes do empreendimento se analisado com relação ao contexto de mercado; "W" representa *weakness*, significando, analogamente, os pontos fracos do empreendimento se analisado com relação ao contexto de mercado. "O" representa *opportunities*, significando possíveis maneiras em que o mercado do empreendimento pode ser melhor explorado. "T", *threats*, significa possíveis ameaças a tal exploração.

Com esses conceitos em mente, lista-se o que é demandado por cada inicial em um plano XY: "S" localiza-se no canto superior esquerdo e o "W" no direito; "O" localiza-se no canto inferior esquerdo e o "T" no direito. Dessa maneira é construída a matriz SWOT, que permite fácil visualização de uma síntese do contexto de mercado em que é situado o projeto, e, por consequência, melhor direcionamento da equipe dentro dos objetivos de tal projeto.

Matriz SWOT

https://miro.com/app/board/uXjVOgCyebI=/?share_link_id=925777377534

4.1.3. Planejamento Geral da Solução

4.1.3.1. Problema a ser resolvido

O problema a ser resolvido é o alto índice de rotatividade de funcionários. Tal problema ocasiona outros problemas, dentre eles: desconhecimento do motivo de saída de muitos funcionários; gastos com contratação de novos colaboradores (que envolve custos financeiros e operacionais); e dificuldade dos líderes de projeto em identificar quais funcionários têm mais chance de sair.

4.1.3.2. Solução proposta

Construção de algoritmo de machine learning que, após identificar padrões nos dados relacionados ao contexto da saída de funcionários da empresa, possibilitará ação imediata sobre eles. Ou como aptamente articulou a demanda oficial, "Propor um modelo preditivo que possibilite ter a visibilidade de tendência de risco de saída dos colaboradores e desta forma contribua para ações de retenção e redução de taxa de turnover, tanto como revisitar os demais processos de carreira e de desenvolvimento".

4.1.3.3. Tipo de tarefa de predição

O método de classificação mostra-se como o mais adequado para o desenvolvimento da AI requisitada, posto que os rótulos* pertencerão a um conjunto discreto e finito de categorias - nomeadamente, duas: “tem tendência de sair” ou “não tem tendência de sair”.

*Coluna-alvo que, no caso, possui a nomenclatura "Status" e é responsável por informar se o indivíduo analisado continuará na empresa ou não.

4.1.3.4. Utilização da solução

Quando os gestores ou outros líderes da Everymind quiserem uma opinião quantitativa em relação à tendência de algum funcionário sair, ele deverá rodar o modelo. Os resultados do

algoritmo apuram o poder de decisão desses líderes em analisar as possíveis causas da maior chance de saída por parte de cada funcionário específico.

4.1.3.5. Benefícios

O projeto possui como objetivo apurar o poder de decisão dos gestores e, como consequência, reduzir o turnover; os gastos com contratação de novos colaboradores (que envolve investimento em fatores desde treinamento até tempo de adaptação); e a dificuldade dos líderes de projeto em identificar quais funcionários têm mais chance de sair. Tais benefícios permitirão que sejam arquitetadas estratégias para segurar uma porcentagem maior de funcionários por uma quantidade de tempo mais extensa, o que também engendra o benefício de mantê-los superiormente alinhados à cultura da empresa.

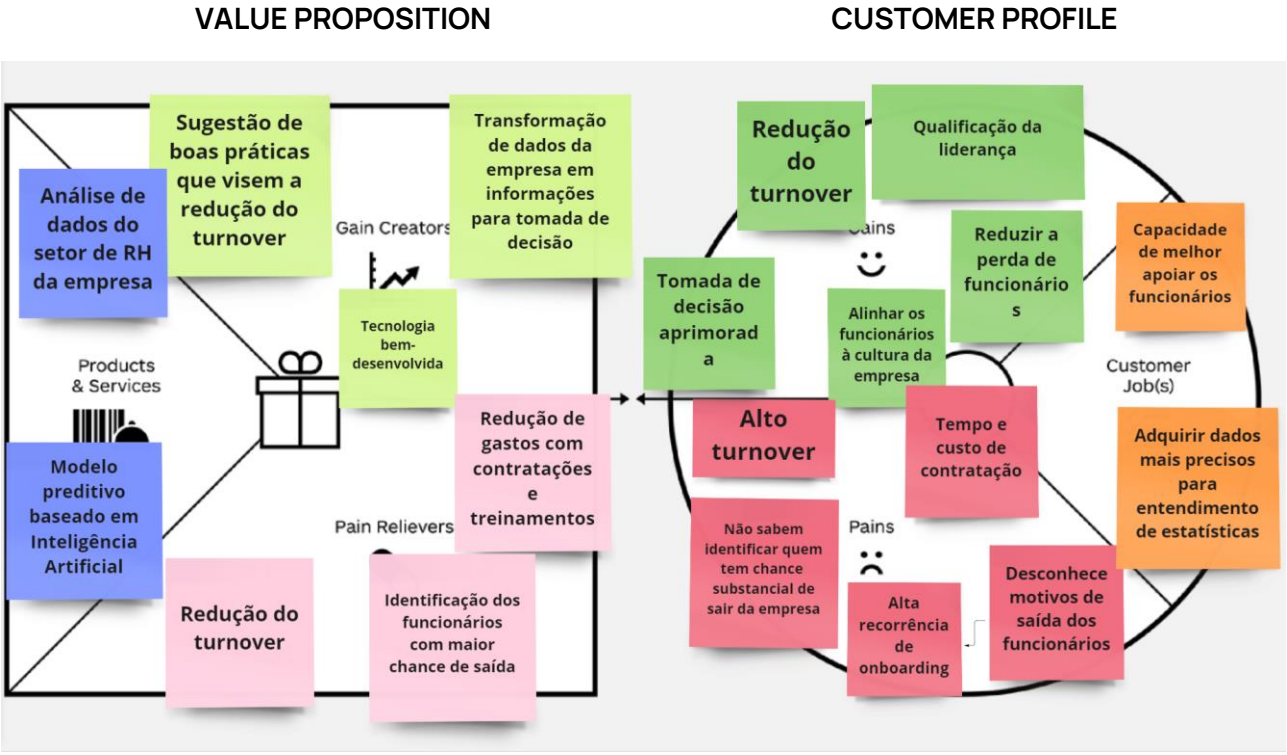
Tendo em mente os fatos mencionados, os benefícios propriamente ditos serão consolidados se, e somente se, os padrões de qualidade do modelo satisfazerem o Critério de Sucesso estabelecido pelos desenvolvedores, que envolve uma acurácia de pelo menos 80% do modelo, e uma amostragem de outras métricas de avaliação - matrizes de confusão e curva ROC.

4.1.4. Value Proposition Canvas

Posto que a percepção espaço-temporal humana é predominantemente visão-configurada, não pode ser subestimado o valor do uso de ferramentas que posicionam informação de maneira facilmente visualizável. Uma dessas ferramentas é o Value Proposition Canvas.

Value Proposition Canvas consiste em um framework que objetiva certificar a compatibilidade do produto em desenvolvimento para com o mercado. Isso é feito por meio da modelagem da relação entre o valor agregado a tal produto e as expectativas inerentes ao público alvo - que por sua vez permite certificar qual o valor criado pelo produto, e qual o público alvo para tal produto. Para ilustrar essa relação, lista-se, para o produto, após o produto em-si ("*Products & Services*"), os fatores geradores de ganho ("*Gain Creators*"), e os fatores redutores

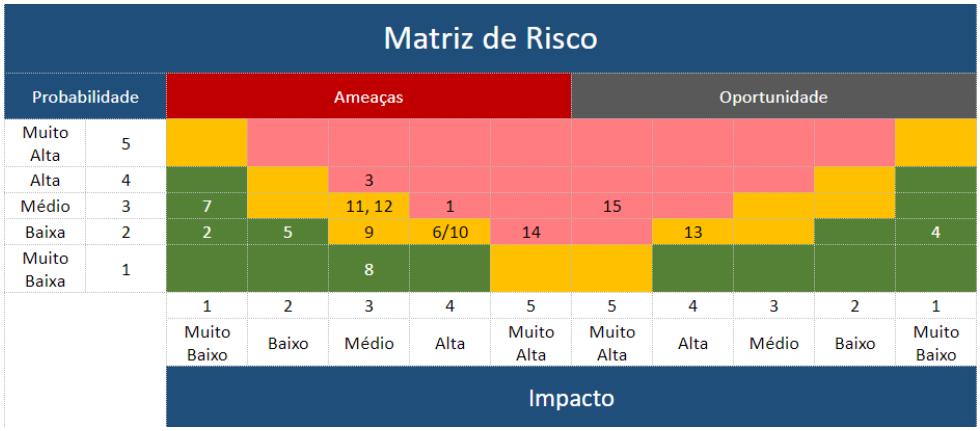
de danos ("*Pain Relievers*"). E para o público alvo, ganhos consequentes do uso do produto ("*Gains*"), dores consequentes da ausência do produto ("*Pains*"), e, por fim, funcionalidades criadas pela presença do produto ("*Customer Jobs*").



https://miro.com/app/board/uXjVOgCyebI=/?share_link_id=925777377534

4.1.5. Matriz de Risco

A matriz de riscos visa prever e analisar os riscos que podem afetar um negócio/projeto, ver o quanto cada um deles o afeta, e também a probabilidade de cada um acontecer. Os números na matriz representam cada item da lista, e a relação entre a probabilidade de um risco acontecer com seu impacto, possibilita o posicionamento desse risco na matriz, fazendo com que os que estão posicionados nos espaços vermelhos mereçam maior atenção - por ser uma grande oportunidade ou uma grande ameaça, e os no verde merecem menos atenção.



https://miro.com/app/board/uXjVOgCyebI=/?share_link_id=925777377534

Lista de riscos:

1- Variáveis pouco claras

2- Falta de dados necessários

3- Resposta pouco específica/subjetiva

4- Falta de experiência do time ao utilizar as ferramentas novas

5- Não alcançar expectativas do cliente

6- Falta de organização e gestão de tempo

7- Mau entendimento sobre o contexto da indústria de Sales force

8- Tecnologias pouco eficientes

9- Problemas com o GitHub ser open source

10- Falta de comunicação entre o grupo

11- Falta de proatividade dos integrantes

12- Má divisão de tarefas, sobrecarregando poucos

13- Complexidade alta demais do projeto

14- Perda/roubo do código e/ou banco de dados

15- Mudança de escopos constantes

4.1.6. Personas

1ª persona

(funcionário do RH - **utiliza o modelo**)

Nome: Luisa

- **Idade:** 27 anos
- **Ocupação**: Funcionária do setor de Pessoas e Cultura do EveryMind
- **Biografia:** gosta muito do seu trabalho, e acredita nos valores da empresa; trabalha com a parte de recrutamento, inclusão...
- **Características (personalidade, conhecimentos, interesses, habilidades):** Racional, com habilidades de resolução de problemas e link entre parte lógica e humana de processos
- **Motivações com modelos preditivos:** resposta de dados bem visual (com diferentes gráficos), e, mesmo não sendo profissional em tech, tem facilidade em entender a lógica
- **Dores com modelos preditivos:** às vezes as respostas são subjetivas e pouco claras em relação à projetos que ela já presenciou.
- **Motivações/necessidades com o problema:** Deseja analisar as respostas do modelo preditivo, podendo, assim, desenvolver um plano de ação juntamente aos líderes para diminuir o turnover
- **Dores com o problema:** como trabalha com recrutamento, ela tem que organizar eventos de onboarding constantemente, tendo um cenário de funcionários pouco alinhados aos valores da empresa; quer melhorar a imagem da empresa

2ª persona

(squad lider de projeto da empresa - **utiliza o modelo**)

Nome: Janice

- **Idade:** 30 anos
- **Ocupação:** squad líder de projeto do setor de desenvolvimento do Everymind
- **Biografia**: estuda constantemente sobre novas metodologias e busca sempre se aprimorar pessoal e profissionalmente
- **Características (personalidade, conhecimentos, interesses, habilidades):** mente inovadora e sempre aberta a novas soluções.
- **Motivações com modelos preditivos:** Deseja analisar as respostas do modelo preditivo, podendo, assim, desenvolver um plano de ação para evitar o turnover
- **Dores com modelos preditivos:** é uma tecnologia nova, e tem medo de não receber instruções suficientes para como lidar com as respostas do modelo
- **Motivações/necessidades com o problema:** quer melhorar a gestão de sua equipe, através da utilização das respostas do modelo, para, assim, melhorar o rendimento do time e aumentar a qualidade da experiência dos funcionários na empresa
- **Dores com o problema:** não consegue identificar as necessidades especificas de cada funcionário da equipe, para ter um tratamento e reconhecimento personalizado com cada um; rendimento da equipe está baixo

3ª persona

(funcionário da empresa - **afetado pelo modelo**)

Nome: Jonas

- **Idade:** 23 anos
- **Ocupação:** funcionário Pleno do setor de desenvolvimento do Everymind
- **Biografia**: foi contratado a menos de 6 meses e ainda está se adaptando à empresa
- **Características (personalidade, conhecimentos, interesses, habilidades):** grande habilidade em lógica e programação, se enxergando em um mercado de trabalho muito movimentado e com muitas oportunidades
- **Motivações com modelos preditivos:** gosta de usar a alexa em casa para despertadores e lembretes, mas não aproveita de features mais complexas
- **Dores com modelos preditivos:** não tem total confiança nos resultados desses modelos; se são assertivos o suficiente
- **Motivações/necessidades com o problema:** reconhecerem sua insatisfação e a alta chance de se demitir, fará com que a empresa enxergue suas dores e invista em sua trajetória na empresa, mostrando o porque de ela estar ali (valores da empresa)
- **Dores com o problema:** ainda se sente meio perdido na empresa e não está muito satisfeito, pois não se sente visto pelos superiores e acha que não tem o reconhecimento que merece

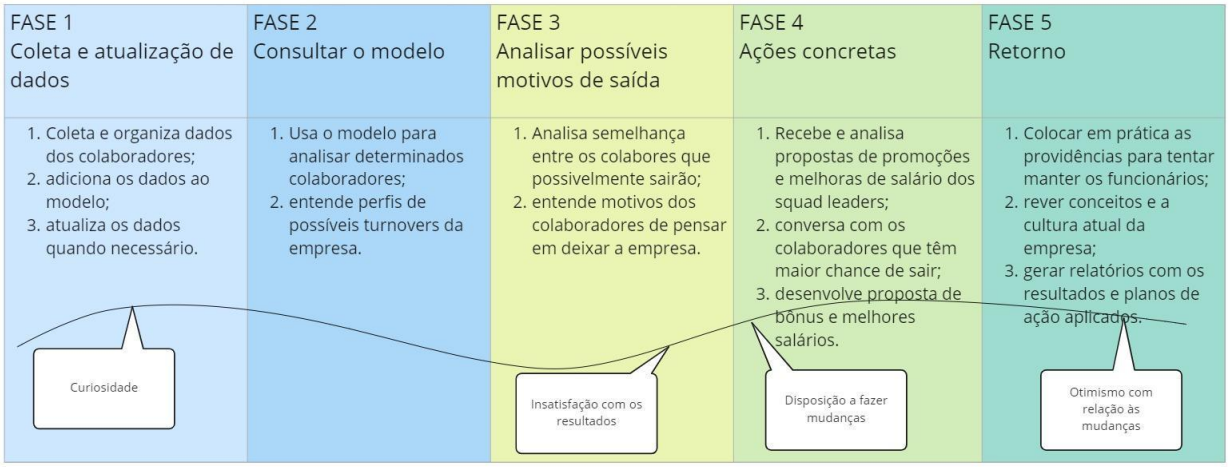
4.1.7. Jornadas do Usuário

Luísa

Cenário: Funcionária do setor de Pessoas e Cultura da Everymind, responsável pelo recrutamento de novos funcionários. Identifica um cenário de colaboradores com valores pouco alinhados.

Expectativas

Desenvolver um plano de ação para reduzir o turnover, melhorando a imagem da empresa.



Oportunidades

Utilizar o modelo para entender o alto índice de turnover, ao trabalhar contribuindo com o squad líder, existe as oportunidades de promover planos de ação concretos e personalizados, diminuindo as demissões.

Responsabilidades

Tem, Como funcionária do setor de Pessoas e Cultura, a responsabilidade de coletar e alimentar os dados, e garantir que sejam atualizados para que o modelo seja o mais preciso possível.

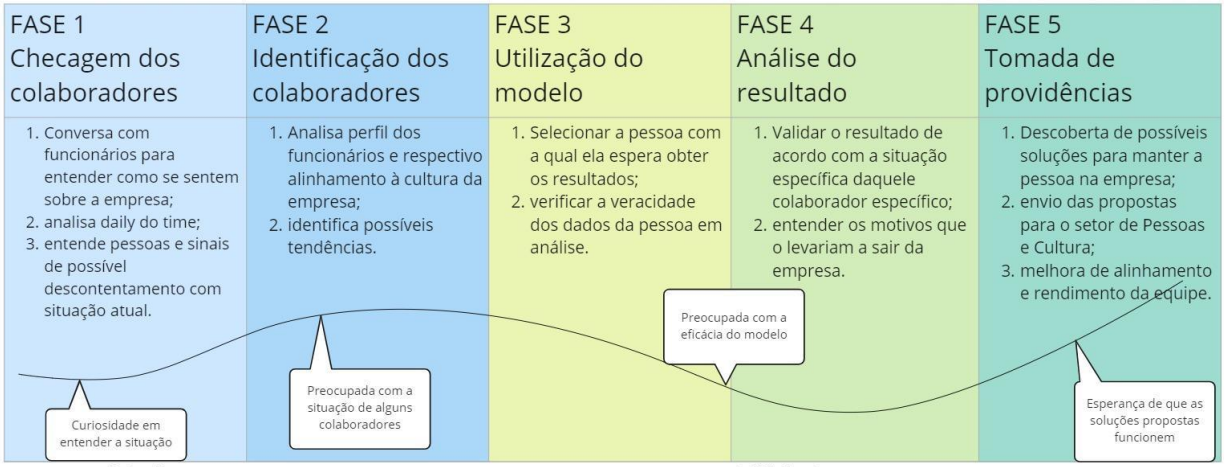
miro

Janice

Cenário: Squad leader do setor de desenvolvimento. Possui dificuldade em identificar quais funcionários têm maior potencial de deixar a empresa e por quê.

Expectativas

Pretende, através do uso do modelo, conseguir entender o que leva os funcionários a saírem, passar a saber quais são os colaboradores com maior potencial de deixar a empresa, e encontrar formas de prevenir isso.



Oportunidades

Melhorar a gestão da sua equipe, melhorar o rendimento do time, e aumentar a qualidade da experiência dos funcionários da empresa, tornando-a mais personalizada.

Responsabilidades

Analisar a resposta do modelo em relação a cada colaborador do seu squad, e partir para a ação, identificando as necessidades específicas de cada funcionário

miro

4.2. Compreensão dos Dados

4.2.1. Descrição dos dados

Os dados disponíveis envolvem um arquivo com três planilhas XLSX com as informações dos colaboradores que saíram e que foram contratados. A primeira planilha - 'EveryMind'- tem 475 linhas que representam, cada uma, um colaborador que está ou não na empresa, e 15 colunas; A segunda - 'Reconhecimento' - possui 339 linhas e 9 colunas, e a terceira - 'Ambiente de Trabalho - 27.07' - 1687 linhas e 13 colunas.

4.2.2. Dados disponíveis

Os dados disponibilizados pelo parceiros de negócio consistem em 3 planilhas em um arquivo excel. Para explicitá-los mais eficazmente, serão utilizados os seguintes critérios:

-> escritos entre colchetes [] representam o tipo do dado não-tratado;

-> escritos entre parênteses () representam, caso necessário, a explicação ou exceção da respectiva coluna.

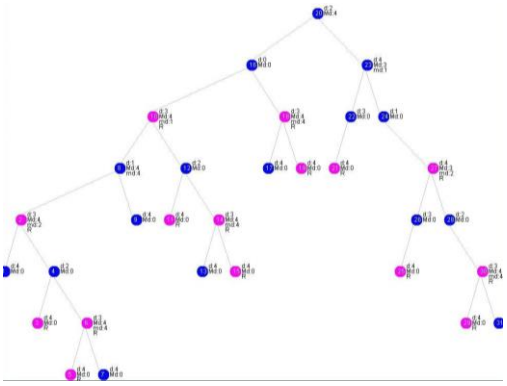
- A primeira planilha ('EveryMind') possui os seguintes dados básicos sobre os funcionários:
    - Nome completo [string] , data de admissão [string], data de saída (se estiver desligado) [data], tipo de saída (dispensa, demissão, etc.)[string], cargo[string], salário mensal[float], data de nascimento [data], gênero[string], etnia[string], estado civil [string], grau de escolaridade[string], área [string], Estado [string], cidade[string].
- A segunda planilha ('Reconhecimento'), por outro lado, envolve dados sobre a quantidade de reconhecimento que cada colaborador recebeu, como méritos (aumento de salário sem mudança de cargo) e promoções (mudança de cargo e, consequentemente, aumento de salário). São eles, em colunas:
    - Situação (ativo, afastado, ou desligado)[string], data de admissão [data], data de vigência [data], novo cargo [string], novo salário [float], motivo ("promoção" ou "mérito")[string], "alterou função" ("sim" ou "não")[string].
- A terceira planilha ('Ambiente de Trabalho - 27.07'), por sua vez, traz dados sobre uma pesquisa de ambiente de trabalho, com features relacionadas a setores e a perguntas, com a porcentagem de pessoas que responderam em cada nível de satisfação. Suas colunas englobam:
    - Divisão (área em que foram computadas as perguntas)[string], Pilar (campos mais genéricos relacionados ao ambiente de trabalho, como bem-estar, relacionamento com gestor e etc.)[string], Pontuação em relação ao Pilar[float], Fator (ramificações do Pilar, como saúde mental ou física dentro do pilar de bem-estar)[string], Pontuação em relação ao Fator[float], Pergunta[string], Pulou (pessoas que não responderam à pergunta em questão)[string], 5 níveis de

satisfação (muito insatisfeito, insatisfeito, neutro, satisfeito, muito satisfeito)[string], taxa de confiabilidade[string].

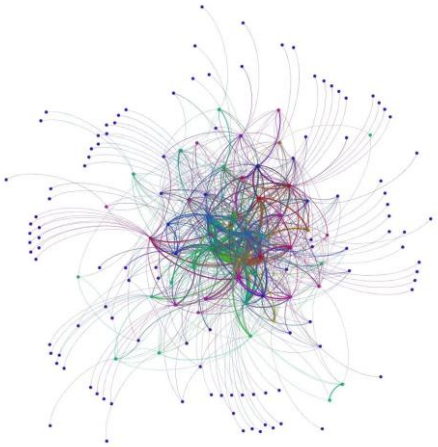
- Dados da pesquisa de satisfação incluem perguntas que abordam fatores como: colaboração, compensação, comunicação, confiança, Diversidade e Responsabilidade Social, qualidade e frequência do reconhecimento, saúde pessoal, propósito e direcionamento, estresse, frequência, saúde mental, valores, desenvolvimento profissional, confiança, comunicação e colaboração com o gestor, autonomia, qualidade, promotor, equilíbrio entre vida profissional e pessoal, ambiente de trabalho, felicidade no trabalho, função dentro da empresa, orgulho e sugestões.

4.2.2.1. Agregação de conjuntos de dados

A interpretação dos spreadsheets será feita a partir das instruções dadas pelo Cliente, e, no momento de escrita (fase incipiente do projeto), mostra-se ideal que sua agregação seja feita de maneira rizomática (em oposição à maneira arborescente).



A imagem acima retrata uma estrutura arborescente: caracterizada por sua orientação por princípios totalizantes, binarismo, e dualismo. Progresso unidirecional, sem a possibilidade de retroatividade e de cortes binários contínuos.



Rizomas, representados pela imagem acima, ao contrário de árvores, são pontos de entrada e saída não-hierárquicos na representação e na interpretação de dados. Isto é, uma concepção horizontal e não-hierárquica em que qualquer coisa pode estar ligada a qualquer outra, sem priorização por espécies. De acordo com Deleuze & Guattari, os princípios de um rizoma são:

- 1 e 2. Princípios de conexão e heterogeneidade: qualquer ponto de um rizoma pode ser conectado a qualquer outro, e assim deve ser.
- 3. Princípio da multiplicidade: só quando o múltiplo é efetivamente tratado como substantivo, "multiplicidade", deixa de ter qualquer relação com o Um;
- 4. Princípio da ruptura significante: um rizoma pode ser rompido, mas recomeçará em uma de suas velhas linhas, ou em novas linhas;
- 5 e 6. Princípios de cartografia e decalcomania: um rizoma não é passível de nenhum modelo estrutural ou generativo; é um mapa, e não um traçado. O que distingue o mapa do traçado é que ele é inteiramente orientado para uma experimentação em contato com o real.

Parafraseando Nick Land, "Schizoanalysis works differently. It avoids Ideas, and sticks to diagrams: networking software for accessing bodies without organs. BWOs, machinic singularities, or tractor fields emerge through the combination of parts with (rather than into) their whole; arranging composite individuations in a virtual/ actual circuit. They are additive

rather than substitutive, and immanent rather than transcendent: executed by functional complexes of currents, switches, and loops, caught in scaling reverberations, and fleeing through intercommunications, from the level of the integrated planetary system to that of atomic assemblages. Multiplicities captured by singularities interconnect as desiring-machines; dissipating entropy by dissociating flows, and recycling their machinism as self-assembling chronogenic circuitry".

Tendo em mente os fatos mencionados, pode-se concluir que interpretar os dados de acordo com inferências originadas pelas falas do Cliente, e que agregá-los de acordo com uma lógica rizomática, mostra-se, no momento, a decisão mais adequada.

4.2.2.2. Riscos e contingências

A qualidade dos dados pode ser um risco, visto que, para construir uma AI que providencie resultados mais precisos, é necessário uma base de dados com uma boa quantidade de dados e de qualidade, com o mínimo de dados em branco ou nulos . Isto é, fatores que seriam de extrema importância para melhorar a precisão do modelo e não foram fornecidos. Quanto ao acesso, pode ser classificado como de boa qualidade, pois o spreadsheet e as imagens são, pela própria natureza de seus respectivos formatos, facilmente acessíveis.

Outro risco, é que, é possível que os dados estejam enviesados, tendo em vista que a coluna “Matrícula”, a qual representa o nome do colaborador em questão na primeira planilha, possui uma taxa de correlação com o Status de ativo e desativo do colaborador. Ou seja, uma variável que não pode ser usada por não fazer sentido no contexto da análise de dados traz uma grande correlação com nosso alvo.

4.2.2.3. Seleção do subconjunto para análises iniciais

Inicialmente, nas análises iniciais, antes do tratamento dos dados, para selecionar o subconjunto para análises, nos atentamos a apenas hipóteses baseadas em pesquisas sobre turnover e gráficos básicos. Dessa forma, as análises iniciais foram feitas apenas em cima da tabela 1 (‘EveryMind’), que compuseram nosso subconjunto inicial.

4.2.2.4. Restrições de segurança

A base de dados disponibilizada para a criação da solução possui informações pessoais e confidenciais, por isso, a própria Everymind fez o mascaramento de dados antes de nos fornecê-la. Além disso, essa base não pode ser postada em nenhum repositório público, já que os dados são confidenciais e isso iria contra as normas LGPD, as quais são seguidas pelo projeto.

4.2.3. Predição desejada - "target"

O 'target' do modelo de predição classificatório desenvolvido é o 'Status', que tem como valores de natureza binária: 'Ativo' e 'Desligado', ou seja, se o funcionário tem tendência a sair ou não da empresa. Nesse sentido, é importante apontar que a utilização do modelo não tem como objetivo a tomada de uma decisão avulsa, mas apenas apontar uma tendência, que deve ser analisada juntamente a um ser humano com análises além do algoritmo.

4.3. Feature engineering

Feature engineering é uma técnica que aproveita os dados para **criar novas variáveis**, através de derivações e cruzamentos, por exemplo. Com essa técnica é possível produzir novas características, com intuito de simplificar e acelerar as transformações de dados. Serve também para preparar os dados para mineração, com seleção e limpeza.

Logo, objetivando gerência de maior qualidade sobre o modelo, as manipulações de dados realizadas a seguir têm como objetivo torná-los adequados para o uso no modelo.

4.3.1. Variáveis Categóricas

4.3.1.1. Label Encoding

Inicialmente decidimos tratar as variáveis categóricas com o método label encoding que é uma excelente ferramenta para converter variáveis categóricas que possuem alguma relação de ordem, no entanto não é indicado para variáveis que não possuem tal relação devido a possibilidade de introduzir problemas no modelo. Basicamente esse método atribui cada valor

único de uma coluna a um número , porém após uma segunda análise dos dados e dos valores das variáveis categóricas apresentadas optamos por não utilizar mais o label encoding na nossa preparação de dados utilizando no lugar dele, o método one hot encoding, que será apresentado na seção 4.3.2.2.

Porém, para fins de documentação e para deixar registrado as técnicas testadas ao decorrer do projeto, fica nessa seção a documentação da tentativa de aplicação do label encoding, mesmo que não mais utilizado.

(tab 1) representa a primeira tabela (nosso dataframe principal, derivado da planilha Everymind)
(tab 2) representa a segunda tabela (o dataframe derivado da planilha de reconhecimento)
(tab 3) representa a terceira tabela (dataframe derivado da planilha de ambiente de trabalho)

1. “Cidade” (tab 1):
    a. conversão dos dados dessa coluna em dados numéricos, através do label encoding.
2. Criação da coluna “Idade” (tab 1) a partir da coluna “Dt Nascimento”.
3. Criação da coluna “faixa_etária” (tab 1), que agrupa, de 4 em 4 anos, os dados da coluna “idade”(tab 1).
4. Criação da coluna “Status” na primeira tabela:
    a. A criação foi realizada com base na coluna “Dt Saida” (tab 1), na qual os colaboradores dividem-se entre “desligados”, e os sem data de saída como “ativos”. Os "desligados" são aqueles com data de saída especificada; os "ativos", aqueles sem. Tais classificações foram transformadas, respectivamente, para 0's e para 1's.
5. Transformação de colunas das tabelas 1 e 2 com dados em string para colunas numéricas, por meio do método de label encoding. Abaixo encontra-se tabela detalhando tais mudanças.

Tabela	Coluna Original	Coluna Numérica Derivada
1	Genero	Genero_Numerico
1	Tipo Saida	Tipo_Saida_Numerico
1	Estado	Estado_Numerico
1	Regiao	Regiao_Numerico
1	Cargo	Cargo_Numerico
1	Estado Civil	ECivil_Numerico
2	Situação	Situacao_Numerico
1	Dt Saida	Status
1	Idade	Faixa_etaria

6. "**Dt admissão" (tabela 1)**: Por virtude do fato de que os dados de data estavam em ordem diferentes de mês e de dia, foi necessária a realização de um tratamento para que esses dados pudessem ser utilizados no modelo. Tal tratamento consistiu na padronização da exibição de datas por meio da aplicação da estrutura dia/mês/ano.
7. ECivil_Numerico: criamos a variável 'ECivil_numerico', a partir da coluna 'Estado Civil', na df1. Decidimos agrupar os valores 'UniãoEstável', 'Divorciado' e 'Separado' ao valor 'Casado', visto que esses 3 primeiros valores têm uma frequência muito menor em comparação aos outros.

   O critério de agrupamento parte da hipótese de que colaboradores com pessoas financeiramente dependentes a elas (pessoas como esposos e filhos, por exemplo), tendem a querer maior estabilidade profissional, e por isso apresentam menores

tendências de saída. Com isso, acreditamos que pessoas nas situações citadas acima têm mais chance de terem filhos, por exemplo.

Entre os anexos (seção 6 deste documento), consta um dicionário de todas as features com Label Encoder aplicado.

4.3.1.2. One Hot Encoding

Técnica de tratamento das variáveis com valores categóricos, irá gerar colunas para cada valor das variáveis categóricas, atribuindo valor 0 ou 1, dependendo da presença ou ausência da característica, respectivamente.

-> Esse método foi utilizado nas colunas: 'Cargo', 'Area', 'Estado', 'Genero', 'Cidade', 'Regiao', 'Escolaridade' e 'Estado Civil', gerando um **novo dataframe 'cat_df'**.

4.3.2. Eliminação de features

No conjunto de dados disponibilizados pela Everymind, não existem valores ausentes/em branco no que diz respeito à qualidade de dados. Ou seja, as únicas colunas com valores ausentes são quando aquele atributo não se aplica ao colaborador daquela linha (quando o atributo diz respeito à saída e o colaborador ainda está ativo). Exemplos dessa situação são as colunas: "Dt Saida" (tab 1) e "Tipo Saida"(tab 1).

Nesse sentido, em relação à coluna 'Tipo Saida", optamos por substituir esses valores em branco por "0", diretamente na coluna numérica derivada "Tipo_Saida_Numerico". Essa substituição foi realizada a partir da técnica label encoding, através do método ".replace" e é demonstrada a seguir:

Tabela	Coluna Original	Coluna Numérica Derivada
1	Vazio	0

Tabela	Coluna Original	Coluna Numérica Derivada
1	'Rescisão Contrato Exp-dispensa'	1
1	'Rescisão ContratoExp-Pedido'	2
1	Dispensa sem Justa Causa	3
1	Pedido de Demissão	4

Quanto à coluna “Dt Saida”, optamos por substituir, por meio do método replace(), os valores em branco pela data hodierna, representando a Dt Saida como se o colaborador ativo pudesse sair hoje. Seus valores, por outro lado, foram, por meio de cruzamentos e da criação de novas colunas a partir dela. Uma dessas colunas é a de nome “Tempo de Trabalho”. Tendo em mente tal contexto, o grupo acabou por optar a não substituir e/ou alterar valores vazios da coluna “Dt Saida”, alavancando a situação para criar outras colunas, adaptando a coluna original para cada objetivo específico.

Ainda em relação a valores faltantes, deve-se acrescentar que não existem valores em branco dentre os dados da coluna “etnia”. Mas, posto que existe o valor “NãoInformada”, que é utilitariamente equivalente a um valor em branco, o grupo optou por não utilizar essa coluna, por virtude do fato de que, pelos motivos supramencionados, ela não possui informações suficientes, e a sua utilização, portanto, poderia originar um viés de dados.

Etnia	Quantidade
Branca	195
Não Informada	169
Parda	84

Preta	17
Amarela	10

4.3.3. Variáveis Numéricas

Contendo todas as variáveis numéricas ja existentes e derivadas durante o processo de pré-modelagem, foi criado um **dataframe 'numeric_df',** que contém as features: 'Salario Mês', 'Media_Salarial', 'Tempo_de_Trabalho', 'Idade', 'Faixa_Etaria','Estagnação', 'Reconhecimento_Numerico', 'Reconhecimento_Medio' e 'Feedback'.

Nesse sentido, os tratamentos descritos a seguir foram aplicados nesse dataframe **'numeric_df'.**

Apesar de serem duas formas diferentes de tratamento, a **Normalização e a Padronização** têm o mesmo objetivo: manipular os dados numéricos, a fim de deixá-los com a mesma ordem e grandeza. O objetivo é evitar que essas informações enviesassem o modelo por features desequilibradas.

4.3.3.1. Normalização

A normalização é uma técnica de Min e Max, onde trata o dado deixando entre uma faixa de -1 a 1 ou de 0 a 1. Ela é recomendada para quando temos dados em uma distribuição não Gaussiana (conhecida também como distribuição normal), ou um desvio padrão muito pequeno. É recomendável também quando os limites de valores de atributos distintos são muito diferentes.

Dados outliers que são valores que se diferenciam drasticamente de todos os outros, pode ser por um erro humano, de medição, contaminação e assim por diante. Nesses casos, quando temos muitos dados com essa característica, a normalização só é viável mediante tratamento dos dados ou até mesmo a exclusão dos mesmos.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Min-Max fórmula

Fonte: https://medium.com/data-hackers/normalizar-ou-padronizar-as-variáveis-3b619876ccc9

4.3.3.2. Padronização

A padronização é uma técnica que utiliza a fórmula de z-score, e é uma adequação bastante usada quando lidamos com features numéricas é mapear os valores de uma distribuição para valores de uma distribuição normal padrão para que, independentemente dos valores que temos na distribuição, tenhamos a mesma **grandeza** de valores.

Diante o contexto e a partir de inúmeras testagens com as métricas de avaliação da Acurácia dos modelos, foi escolhida a técnica de Padronização com o método StandardScaler() da biblioteca sklearn.

$$z = \frac{x - \mu}{\sigma}$$

z-score fórmula

Fonte: https://medium.com/data-hackers/normalizar-ou-padronizar-as-variáveis-3b619876ccc9
Fonte: https://acervolima.com/normalizacao-vs-padronizacao/

4.3.4. Oversampling e Undersampling no modelo

Oversampling e Undersampling são técnicas dentro da análise de dados que ajustam a distribuição de classes em um conjunto de dados. A referência de dados que foi utilizada para o ajuste é se o colaborador está ou não na empresa. O **Undersampling** é feito com a retirada de

dados excedentes, nesse caso, se existem mais pessoas que saíram do que estão na empresa, alguns desses dados são retirados. A técnica de **Oversampling** funciona de maneira contrária, essa técnica adiciona dados, ou seja, se existe um dado em menor quantidade, essa técnica adiciona mais desse tipo de dado para que haja balanceamento.

Diante o contexto da quantidade de dados relativamente pequena, e a partir de testes com métricas de avaliação do modelo com as duas técnicas, no modelo da solução proposta, foi escolhida a aplicação do **Oversampling** sobre a base de dados, contribuindo para o balanceamento dos dados e para a precisão do modelo.

Nesse sentido, dentre as técnicas mais comuns de oversampling a que mais se encaixa no contexto do nosso modelo, e que apresentou melhor desempenho foi o 'SMOTE', que é uma técnica de oversampling que se mostra mais precisa e adequada que o 'Random Oversampler', visto que ajuda a resolver o problema de overfitting que muitas vezes é gerado pelo Random.

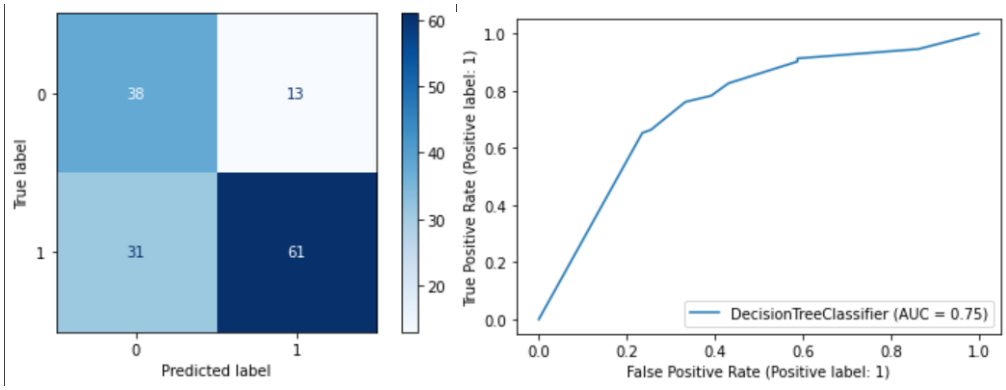
4.3.5. Desempenho das técnicas de preparação

A partir de uma análise dos resultados das métricas de avaliação pré e pós técnicas de preparação - OverSampling e Padronização - pode-se comparar os resultados e perceber um impacto muito positivo a partir do uso dessas ferramentas de balanceamento dos modelos, que serão melhor desenvolvidos na seção 4.4 deste documento.

É importante frisar que as métricas de avaliação aqui citadas estão descritas na seção 4.5 deste documento.

Nesse sentido, os resultados do modelo com algoritmo de Árvore de Decisão pré técnicas de balanceamento são os seguintes:

Acurácia de Treino: 0.89
Acurácia de Teste: 0.69
Matriz de confusão e curva roc, respectivamente:

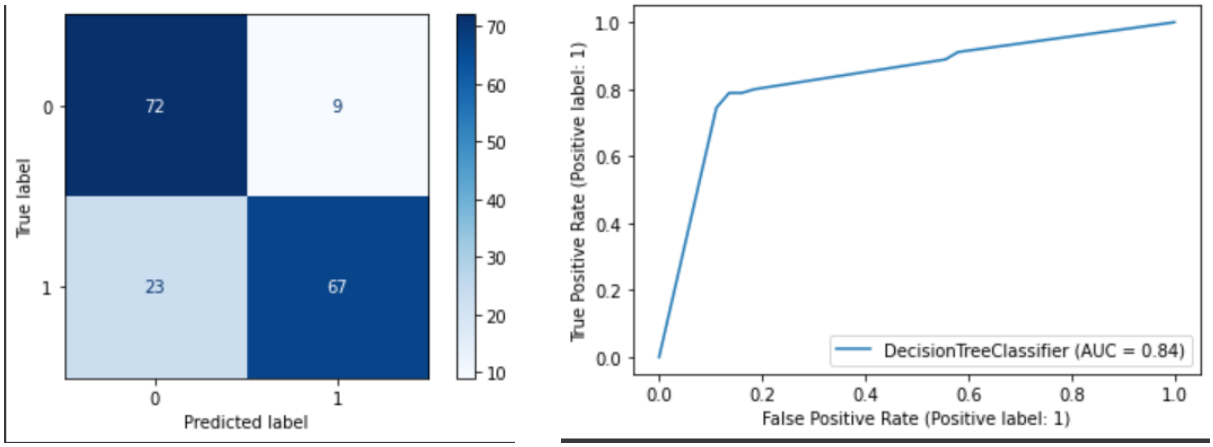


E os resultados pós técnicas de balanceamento foram:

Acurácia de Treino: 0.91

Acurácia de Teste: 0.81

Matriz de confusão e curva ROC, respectivamente:



4.3.6. Features Derivadas ou Cruzadas

Para criar novas features, foram feitas as seguintes manipulações e cruzamentos de dados, a fim de aumentar os campos de análise dos dados, tornando-a mais completa e profunda:

1) **Jornada de Trabalho**

tabela 1: Calculamos o tempo que um colaborador trabalha/trabalhou na empresa, a partir dos dados das colunas 'Dt Saída' (tab 1) e 'Dt Admissão' (tab 1), onde geramos uma nova coluna '**Tempo de Trabalho**' (tab 1), que retorna, em dias, o tempo entre a data de admissão e a data de saída (se a pessoa está desligada), ou o tempo entre a data de admissão e o dia de hoje (se a pessoa está ativa).

```
#Função pega a data de admissão do colaborador e a data do seu desligamento, e encontra o perído entre elas.
#.dropna() Ela dropa todos os dados que possuem valores Na...
Jornada = (pd.to_datetime(planilha['Dt Saida']) - pd.to_datetime(planilha['Dt Admissao'])).dropna()
```

```
[5] Jornada

0    4594 days
```

2) Idade

tabela 1: Calculamos a idade dos colaboradores a partir dos dados da coluna 'Dt Nascimento' (tab 1) de cada colaborador, calculando, em anos, a diferença com a data atual.

```
[1] #Ele pega a data de hoje e subtrai da data de nascimento, retornando a idade, np.timedelta64 retorna a da data em ano.
    df1['Idade'] = ((pd.to_datetime('today')-pd.to_datetime(df1['Dt Nascimento']))/ np.timedelta64(1, 'Y')).astype(int)
```

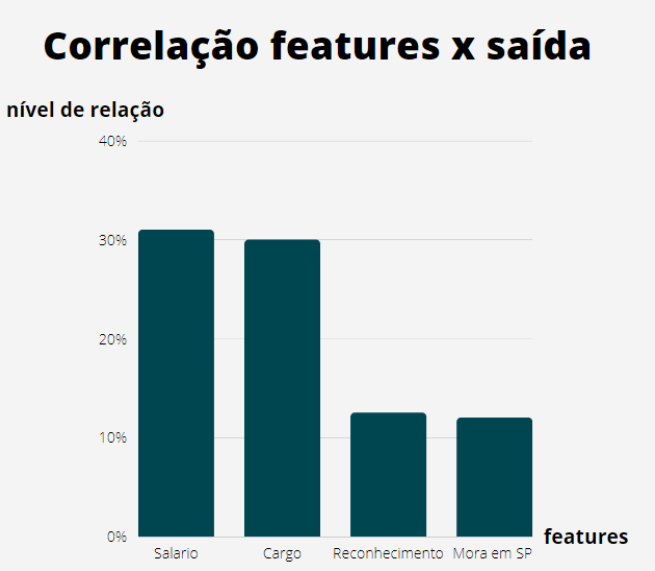
Além disso, a partir da ideia de cruzamento de dados e criação de novos atributos, foram criadas as seguintes features:

3) "**Media_salarial**" - tabela 1: a partir de um cruzamento entre os dados de "Salario Mês" e de "Cargos", calculamos o salário médio de cada cargo.
4) "**Salario_Comparado**" - tabela 1: compara o salário médio do cargo (coluna "Media_Salarial" criada) com o salário de cada colaborador, classificando o salário mensal do colaborador como "acima" (0) ou "abaixo" (1) da média de seu cargo.
5) "**Estagnação**" - tabelas 1 e 2: calcula, em dias, o tempo entre a última promoção recebida por um colaborador e a data de hoje;
6) "**Reconhecimento Num**" - tabela 1 e 2: soma a quantidade de promoções e méritos que um colaborador recebeu. Quando o colaborador não possui reconhecimento, o valor (0) é adicionado à coluna. A interseção das tabelas foi feita através do número de matrícula usado como chave estrangeira.

7) “**Regioes_Numerico**” - tabela 1: agrupamento dos dados da coluna, resultando na criação de uma nova coluna, denominada "Regiões" (tab 1);
8) “**estadoSP**” - tabela 1: a partir da coluna “Estados_Numerico”, separa os colaboradores localizados no Estado de São Paulo daqueles dos demais estados. Isso materializa-se por meio da classificação de tais colaboradores em 0 ou em 1, sendo 1 aqueles pertencentes a São Paulo, e 0, aos demais estados.

4.3.7. Seleção de Features

	Status
Status	1.000000
Salario_Comparado	0.316975
Trainee-Dev	0.248352
DevPl	0.216092
Core&Industrias	0.171740
Reconhecimento_Numerico	0.125633
CPG&Retail	0.124058
SP	0.123696
estadoSP	0.123696
MktCloud	0.119437
Everymind	0.114358
People&Culture	0.113969
Gerente	0.108608



A partir de inúmeras testagens de diferentes features (situadas no tópico 4.5 deste documento), e a partir dos resultados da matriz de correlação acima, que relaciona as features com o ‘target’, tornou-se possível chegar às seguintes features para o modelo final:

Nesse sentido, criamos um **dataframe oficial**, que será utilizado para a modelagem**: ‘df’,** que contém as features do dataframe ‘numeric_df’, as do ‘cat_df’, e a feature alvo ‘Status’, que vem da df1.

-> Dessa forma todas as features que serão utilizadas no modelo, e que serão descritas a seguir, estão situadas na tabela final **"df"** e presentes na seção 4.3.

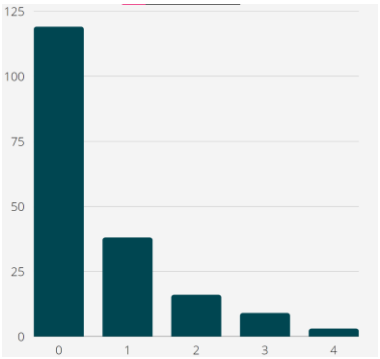
Após uma multiplicidade de experimentações, avaliações, e análises, as features selecionadas foram:

1. **'Estagnação'**: exibe a quantia de tempo que um colaborador está sem receber promoções e/ou méritos (i.e. "reconhecimento"). Esse tempo pode tanto fundamentar sentimentos de insatisfação por parte do funcionário, quanto pode refletir algo que contribua para um pedido de demissão (possivelmente vindo do funcionário, possivelmente da empresa), ou qualquer coisa entre esses dois pólos.
2. **'Salario_Comparado':** compara o salário de um colaborador com o salário médio de indivíduos do mesmo cargo. Mostra-se relevante pois, de acordo com o levantamento de nome "FIA Employee Experience (FEEx)", a insatisfação salarial é uma das frustrações de maior expressividade para pedidos de demissão na idade contemporânea.
3. **'Media_Salarial'**: representa o salário mensal médio do cargo de um funcionário, e, também de acordo com os resultados da pesquisa FEEx, mostra-se relevante.
4. **'faixa_etaria'**: exibe a faixa etária de um colaborador, dessa maneira demonstrando a possível influência da idade de um colaborador na tomada de decisão de abandonar a empresa. Como aponta reportagem feita pela EBC, listada na seção "Referências", idade é um fator de importância.
5. **'Região':** mostra em qual região do Brasil um colaborador mora. É importante pois aponta possível padrão de saída relacionado à localização do colaborador. Outro fator representado por esse dado é se um colaborador trabalha presencialmente ou não. -> engloba todas as colunas geradas pelo one hot encoding
6. **'Estado':** mostra em qual Estado do Brasil um colaborador mora. É importante por apontar possíveis padrões de saída relacionados à localização do colaborador, assim como o atributo Região. Além disso, é importante para saber se o

colaborador trabalha pessoalmente ou não. -> engloba todas as colunas geradas pelo one hot encoding

7. **'Gênero':** Identidade de gênero do colaborador. Esse pode ser um fator importante já que a quantidade de homens presentes na empresa é muito superior à quantidade de mulheres.
8. **"Área":** Área na qual o colaborador está inserido e exerce seu cargo. Isso reflete diversos padrões generativos de hipóteses e, principalmente, importa muito para a interação com a terceira tabela, em que criamos a feature 'feedback'. A Área também indica tendências de saída visando pelas áreas que já não existem mais ou que estão indo muito bem dentro da cultura organizacional da empresa.
9. **'Cargo'**: identifica qual o cargo de um colaborador. Isso reflete diversos padrões generativos de hipóteses, a exemplo dos índices de satisfação dos funcionários variando de acordo com os diferentes cargos. Mostra-se relevante para o "modelo.estado". -> engloba todas as colunas geradas pelo one hot encoding
10. **'Estado Civil'**: a escolha dessa feature parte da hipótese de que pessoas com pessoas dependentes financeiramente (esposos ou filhos), tendem a querer maior estabilidade no emprego, tendo menor tendência a sair. Com isso, acreditamos que essas pessoas são as que se identificam nas situações 'Casado', 'Separado', 'União Estável' e 'Separado'.-> engloba todas as colunas geradas pelo one hot encoding
11. "**Reconhecimento_Medio**": Em teoria, deve-se calcular o reconhecimento médio por meio da divisão do reconhecimento numérico pelo tempo de trabalho. Porém, na prática, para que sejam obtidos valores inteiros, deve-se realizar o cálculo por meio da divisão do tempo de trabalho pelo reconhecimento numérico.
12. **"Quantidade de Reconhecimento":** A quantidade de reconhecimentos que a pessoa recebeu ao longo de toda sua jornada de trabalho na Everymind. Isso foi importante para derivar outro atributo que recebeu uma grande taxa de correlação, o reconhecimento médio, e, também, foi importante para o modelo em si.

13. **“Tempo de Trabalho”:** Tempo de trabalho representa a quantidade padronizada de dias que o colaborador está na empresa, independentemente do recebimento de promoções. Isso é importante principalmente pensando na grande quantidade de pessoas que saem com poucos meses de trabalho, enquanto a retenção de funcionários sempre aumenta ao longo do tempo de trabalho da pessoa.



14. **“Feedback”:** Representa a média das pontuações dos pilares de ambiente de trabalho da empresa. Isso é relevante para entender como variações nas pontuações dos pilares são importantes para a tendência de saída dos funcionários.
15. **“Salário Mensal”:** Variável básica da base de dados entregue pela Everymind.

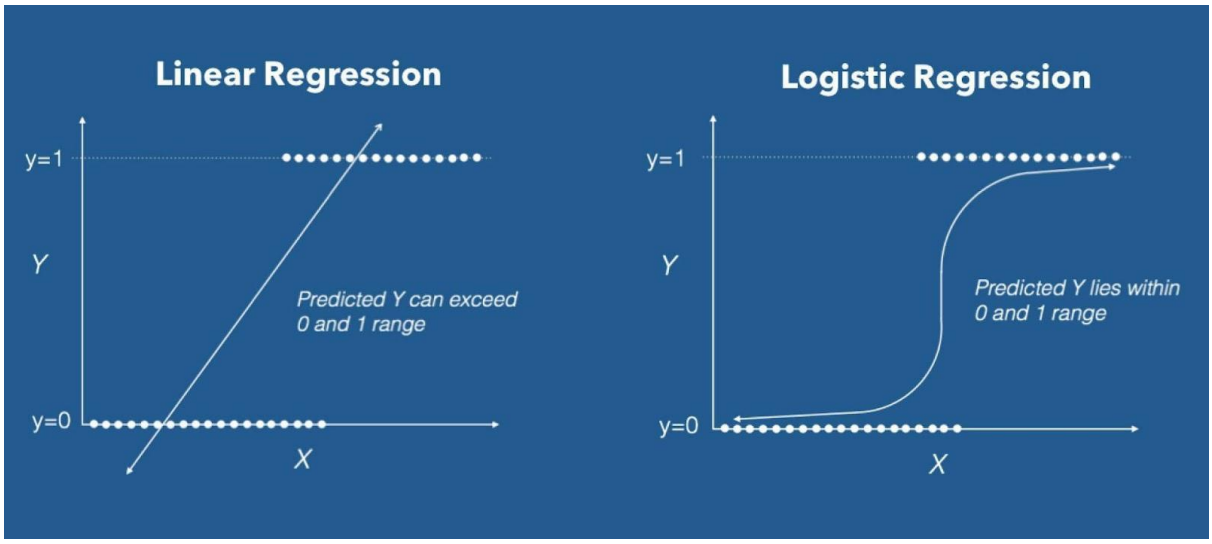
4.4. Modelagem

Foram experimentados 6 diferentes tipos de algoritmos de classificação. Tais algoritmos são: **Regressão Logística, KNN, Árvore de decisão, SVM, Naive Bayes e Redes Neurais.**

4.4.1 Regressão Logística

O Modelo de Regressão Logística cria uma função matemática que visa predizer valores em relação às variáveis categóricas. Sendo muito parecido a uma função de regressão linear, ele se diferencia pela utilização do valor Y: ao invés de assumir um valor específico, ele retorna um valor binário (0 ou 1). Dessa forma, esse modelo utiliza operações estatísticas, o que significa que ele é mais indicado para situações nas quais a resposta é binária.

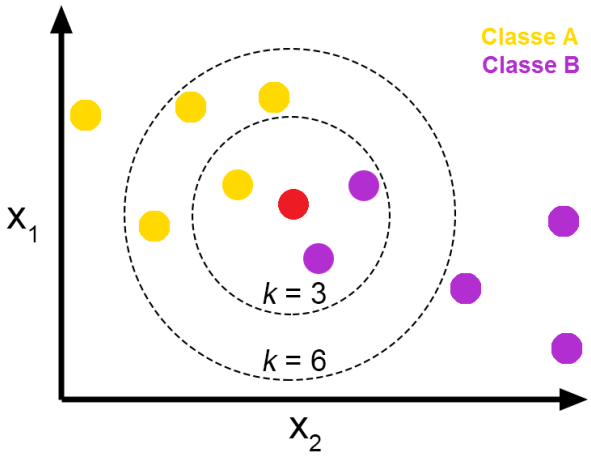
Algumas das características desse modelo são: requer pequeno número de suposições; fornece resultados em termos de probabilidade; classifica os elementos em categorias; tem facilidade com variáveis independentes categóricas; não precisa de escala de recursos de entrada; e não necessita de grandes quantidades de recursos computacionais.



4.4.2. KNN (K-Nearest Neighbors)

KNN (K-Nearest Neighbors, em tradução literal “K-vizinhos mais próximos”) é um algoritmo para reconhecimento de padrões utilizando de método não-paramétrico que, basicamente, classifica a base de dados em dados para treino e em dados para testes. A distância entre os pontos de treino e os pontos de testes é avaliada, e o ponto com a menor distância é classificado como o nearest neighbor. O algoritmo KNN prevê o resultado com base na maioria, como demonstra a imagem abaixo.

A principal desvantagem do KNN, se comparado a outros modelos, resume-se na possibilidade de que o período de processamento seja desconfortavelmente longo, posto que a quantidade de tempo de processamento aumenta de acordo com o tamanho da base de dados.



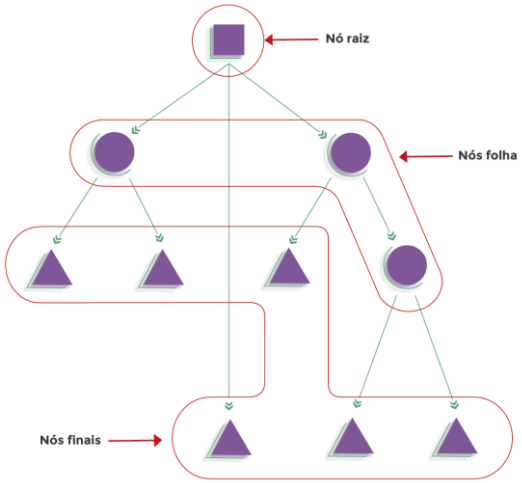
Na prática, o processo de experimentação de diferentes features foi dividido em 4 etapas: a divisão do dataset, o dimensionamento da feature, a definição do “k” e a avaliação do modelo.

1. A divisão do dataset para treino e teste foi feita através do “train_test_split”, importado da biblioteca “sklearn.model_selection”. Para todos os testes, ele foi definido com o padrão 0.3 do dataset para teste.

2. Para definir o parâmetro “k”, foi utilizado o padrão em que k é igual à raiz quadrada do tamanho do conjunto de testes.
3. Para a avaliação dos resultados do modelo KNN, utilizou-se a taxa de acurácia e a matriz de confusão (explicadas detalhadamente na secção 4.5).

4.4.3. Árvore de Decisão

A árvore de decisão é uma das técnicas mais populares de mineração, principalmente para a tarefa de classificação, e consiste em uma coleção de nós internos e nós folhas, organizados em um modelo hierárquico.



Nesse sentido, assim como apresentado na imagem acima, os nós internos representam os atributos descritivos do modelo, as subárvores saem de cada nó interno, e contém os possíveis valores do seu nó raiz, que é a feature mais importante e mais influente no modelo, e os nós folha representam uma decisão sugerida pelo modelo (rótulos). Basicamente, uma árvore de decisão se ramifica em diversas escolhas que podem ser tomadas, que levam a outras escolhas, até chegar em um rótulo.

Essencialmente, a aplicação desse modelo no projeto ocorre no eixo em que são verificadas as características de cada colaborador e, com base nelas, calculadas respectivas probabilidades de abandono da empresa.

Pseudocódigo:

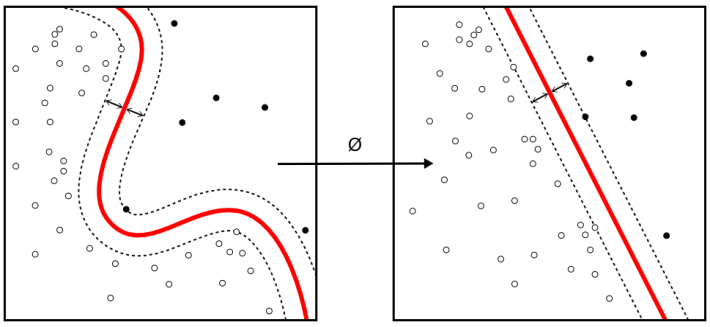
- para cada depth-first search assignment
    - para cada variável V_i considerada
        - para cada valor X_i dentro do domínio D_i
            - para cada constraint C entre (X_i ; X_j) onde X_j é um elemento do domínio D_j [$X_j \in D_j$]
                - se "ali" não existir, X_j tal que a constraint entre (X_i ; X_j) é satisfeita [1]
                    - remove X_i de D_i.

[1] se não existir um valor em uma variável adjacente de maneira que a constraint seja satisfeita, tem-se um problema; é necessário livrar-se daquele valor. Então X_i é removido de D_i.

4.4.4. SVM (Support Vector Machine)

O modelo SVM (Support Vector Machine) é um algoritmo de aprendizado supervisionado que contribui muito para tarefas de classificação e categorização. O algoritmo busca uma linha de separação entre duas classes distintas analisando um ponto de cada grupo que mais estão próximos um do outro. Ou seja, o SVM escolhe uma reta entre dois grupos que está equidistante de ambos. Um detalhe importante: também existe o SVM não linear, que separa os grupos sem necessariamente uma reta. Isso acontece por meio de uma transformação não-linear do espaço.

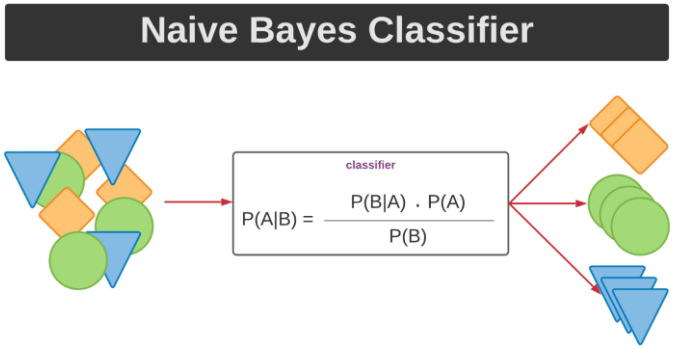
Para o nosso contexto, hipoteticamente, o modelo SVM faz sentido por classificarmos binariamente a tendência de turnover, nesse sentido, o modelo foi testado para verificarmos essa hipótese.



Fonte: Wikimedia Commons.

Contudo, após realizarmos os testes, verificamos que a acurácia média do modelo SVM em relação à nossa tabela de testes foi de 64%. Objetivamente, 64% não é uma boa acurácia, especialmente se comparada àquelas dos outros modelos testados. Eis os resultados:

4.4.5. Naïve Bayes



Naive Bayes é um algoritmo de aprendizado de máquina supervisionado que funciona a partir de uma técnica de classificação de dados, e, basicamente, gera uma tabela de probabilidades. Esse algoritmo é uma aplicação do Teorema de Bayes, que é uma fórmula de probabilidade que calcula a possibilidade de um evento ocorrer, e seu funcionamento pode ser descrito nos seguintes termos estatísticos: para fazer a predição, o algoritmo define uma tabela de probabilidades, em que consta a frequência dos preditores com relação aos rótulos.

A questão do teorema é que é preciso ter alguma informação anterior, ou seja, é preciso saber que um determinado evento já ocorreu e qual a probabilidade desse evento.

Para o cálculo da probabilidade de um evento A dado que um evento B ocorreu, “P(AIB)”, pelo teorema temos:

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

- P(BIA): probabilidade de B ocorrer dado que A ocorreu
- P(A): probabilidade de A ocorrer
- P(B): probabilidade de B ocorrer

Então, o cálculo final leva em conta a probabilidade maior para oferecer uma solução.

Algumas características desse modelo são que ele funciona melhor para situações em que todas as features tem a mesma importância, e só precisa de um pequeno número de dados para concluir classificações com uma boa precisão, o que é ótimo para o contexto do projeto.

4.4.6. Redes Neurais

O modelo de redes neurais é estruturado sobre o conceito de que a lógica de funcionamento subjacente ao processo cognitivo humano é praticamente superior às demais. Isto manifesta-se fisicamente, em síntese, no traduzir da biologia presente na arquitetura de um neurônio e de conjuntos de neurônios em fórmulas matemáticas que podem ser interpretadas por um computador de maneira que é originada AI.

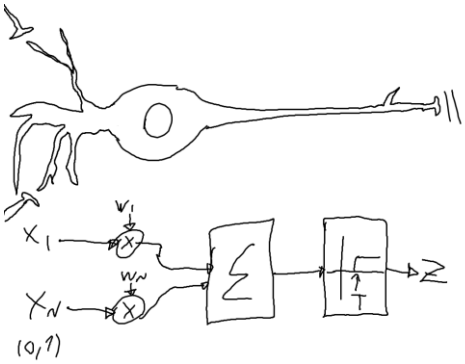


Imagem autoral

Na imagem:

- X1 e Xn: input value - 1 ou 0.
- W1 e Wn: weights ("pesos") - pode ser mais ou pode ser menos forte; se for mais, w1 aumenta; se for menos, Wn diminui. Isso reflete a influência da sinapse na decisão do axon de ser estimulado por inteiro.
- X: multiplicado por peso.
- E: executa inputs pelo somatório para agrupá-los e adquirir força coletiva.
- T: para decidir se tal força coletiva de todos esses inputs é suficiente para fazer o neurônio disparar, executa o somatório via um threshold box que exibe a relação entre o input e o output. Nada acontece até que o input exceda um threshold 'T'. Se exceder ->
- Output Z é um 1; se não, é um 0.
- Isto é, binário entra, binário sai. Os pesos sinápticos são modelados a partir desses multiplicadores; os efeitos cumulativos de todo esse input são modelados até o neurônio por um somatório, e é decidido se será um tudo-ou-nada ao executá-lo por um somatório via o threshold box e verificar se a soma dos produtos são maiores do que o threshold. Se sim, recebemos um 1.

E o que faz um conjunto desses neurônios? Para mais fácil explicação, consideremos um crânio - uma grande caixa, repleta de neurônios, que, por sua vez, são repletos de weights e de thresholds.

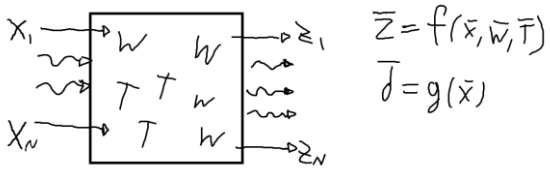


Imagem autoral

Nessa caixa, entram uma variedade de inputs, de X1 até Xn, que são capazes de orientar-se dentro da entropia. E do outro lado saem uma variedade de outputs, de Z1 até Zn. Ou seja, por meio da influência dos weights e dos thresholds, esses inputs saem como outputs. Matematicamente, isso equivale a "z = f(x, w, t)", onde "z" é uma função do vetor input ("x"), do vetor weight ("w"), e do vetor threshold ("t").

E isso é tudo o que é uma rede neural. E quando treina-se uma rede neural, tudo o que é possível fazer trata-se de ajustar tais weights e tais thresholds de maneira que aquilo que sai é aquilo que queremos. Em síntese, uma rede neural é um aproximador de funções.

d = g(x): Por exemplo, talvez tenhamos sample data que nos cede um output vector que é desejado como outra função do input, descartando weights e thresholds. E é isso o que queremos que saia.

4.5. Avaliação - Árvore de Decisão

Utilizamos os mesmos métodos de avaliação para todos os modelos, para tornar possível uma comparação mais clara entre eles. As formas de avaliação foram: Matriz de confusão, acurácia, curva ROC e a avaliação de combinações diferentes de features. Essas avaliações foram feitas em todos os modelos citados neste documento, e o que apresentou resultados mais satisfatórios foi a Árvore de Decisão.

4.5.1. Matriz de confusão

A matriz de confusão consiste em uma matriz - que pode ser transformada em uma tabela - que exibe as frequências de classificação para cada classe de um modelo (mostra o número de predições corretas e incorretas em cada classe), permitindo uma análise mais visual do desempenho de um algoritmo.

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte: https://diegonogare.net/2020/04/performance-de-machine-learning-matriz-de-confusao/

No nosso contexto, a tabela possui, em coluna, o que o modelo respondeu (rótulo), ou seja, se o colaborador possui tendência a sair ou não; e em linha, o que realmente aconteceu - se o colaborador saiu ou não, de fato.

Nesse sentido, nossas frequências seguem a seguinte regra:datas

- TP - true positive:
    - Quando o modelo diz que há tendência de sair e o colaborador realmente sai.
- FP - false positive:
    - Quando o modelo diz que há tendência de sair e o colaborador não sai.
- TN - true negative:
    - Quando o modelo diz que não há tendência de sair e o colaborador não sai.
- FN - false negative:
    - Quando o modelo diz que não há tendência de sair e o colaborador sai.

4.5.2. Acurácia

A partir dos conceitos apresentados na matriz de confusão, é possível calcular a taxa de acertos/acurácia dos modelos.

O cálculo da acurácia consiste na fórmula "**acurácia =** $\frac{VN+VP}{VP+FN+VN+FP}$", em que: VP = Verdadeiros Positivos, VN = Verdadeiros Negativos, FP = Falso Positivos e FN = Falso Negativo. Para esse cálculo, utilizamos "dataset.score" no código dos modelos.

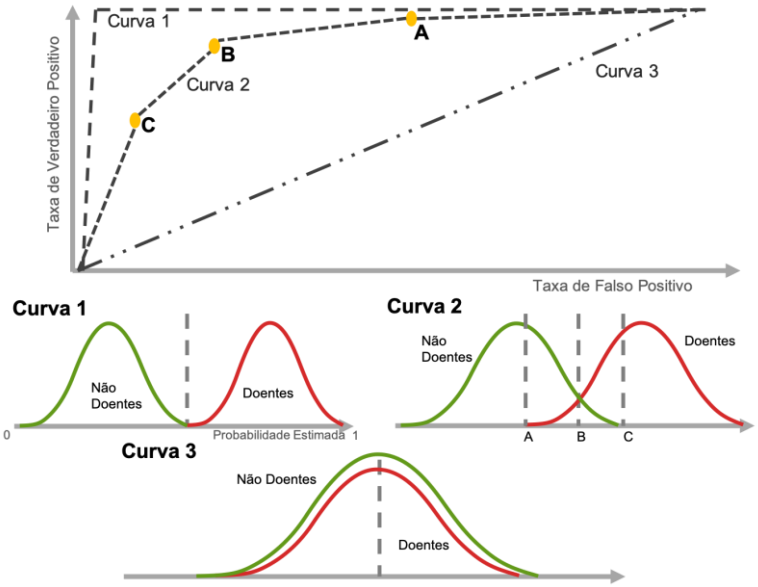
4.5.3. Curva ROC

As **curvas ROC** (receiver operator characteristic **curve**) são uma forma de representar a relação, normalmente antagónica, entre a sensibilidade e a especificidade de um teste diagnóstico quantitativo, ao longo de um contínuo de valores de "cutoff point".

É uma curva de probabilidade. Criada ao traçar a taxa de verdadeiro-positivo (TPR - true positive rate) contra a taxa de falsos-positivos (FPR - false positive rate).

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{TN + FP}$$

Ou seja, o número de vezes que o classificador acertou a predição contra o número de vezes que errou. AUC (area under the curve) representa a área da ROC considera-se como o grau ou medida de separabilidade. Quanto maior o valor, melhor é o modelo em prever ou (por exemplo) distinguir entre os funcionários que ficam ou saem da empresa.


Curva 1
A
B
Curva 2
C
Curva 3
Taxa de Verdadeiro Positivo
Taxa de Falso Positivo
Curva 1
Não Doentes
Doentes
0
Probabilidade Estimada 1
Curva 2
Não Doentes
Doentes
A
B
C
Curva 3
Não Doentes
Doentes

4.5.4. Avaliação de features

A partir de hipóteses sobre as features mais relevantes no modelo, realizamos testagens com diferentes combinações de features, comparando-as apenas em relação à acurácia de teste e de treino, para, depois, realizarmos testagens mais completas com mais métodos de avaliação. Dessa forma, as testagens foram organizadas da seguinte maneira: a primeira coluna é composta por 11 features iniciais selecionadas, e em cada coluna seguinte retiramos apenas uma das features de cada vez para saber qual é seu impacto individual no modelo, como vemos a seguir na tabela 1 :

['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico']	['Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico']	['Idade','Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico']	['Idade', 'Regiao_Numerico', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico']	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico'], ESTAGNAÇÃO	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico'], SALARIO MES
0.80	0.80	0.80	0.80	0.76	0.79
0.84	0.84	0.84	0.84	0.68	0.78

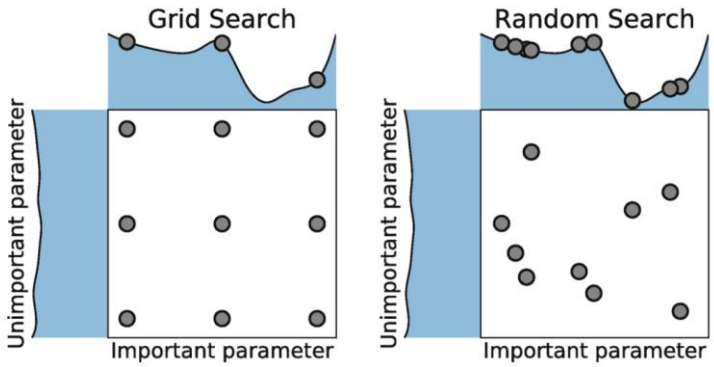
tabela 1 - 1ºvalor: acurácia (treino) e 2ºvalor: acurácia (teste)

continuação tabela 1:

['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico'], RECONHECIMENTO MÉDIO	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'Tempo_de_Trabalho', 'Genero_Numerico']	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Genero_Numerico'] TEMPO DE TRABALHO	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho']
0.82	0.80	0.79	0.80
0.82	0.84	0.77	0.84

A partir de análises sob essas tabelas, é possível perceber que as features com maior impacto são a de Estagnação, Salário mês, Reconhecimento Médio e Tempo de trabalho, visto que quando elas foram removidas, a acurácia de teste diminuiu. Além disso, retirar as demais features só manteve a acurácia estável, ou seja, nenhuma das outras diminui a acurácia do modelo, levando-nos a mantê-las no modelo.

4.5.4. Grid Search e Random Search



Utilizamos das funções de "grid search" e "random search", a fim de ter um direcionamento em relação aos valores dos hiperparâmetros do modelo, para não deixá-los no default. Dessa forma, determinamos os que fazem mais sentido para o contexto do projeto: min_samples_split, min_samples_leaf, max_depth e criterion.

Os hiperparâmetros citados têm as respectivas funções dentro do modelo:

- **Min_samples_split**: o menor número de amostras para dividir um nó interno;
- **Min_samples_leaf: o menor número de amostras para estar em uma folha;**
- **max_depth**: A profundidade da árvore; se mostrou essencial a definição de um valor baixo, visto que valores altos levavam ao overfitting;
- **criterion:** mede a qualidade de uma subdivisão da árvore

```
parameters_arv = { 'criterion':['gini', 'entropy', 'log_loss'],
                   'splitter':['best', 'random'],
                   'max_depth':range(2,10),
                   'min_samples_split':range(1,20),
                   'min_samples_leaf':range(1,20)}
```

Além dos hiperparâmetros destacados acima, foram testados também o “splitter”, “max_features”, min_weight_fraction_leaf, max_leaf_nodes, min_impurity_decrease, porém, a definição desses hiperparâmetros fora do default diminuiu a acurácia do modelo, ou não encaixavam no nosso contexto

Esses hiperparâmetros foram escolhidos pois facilitam a compreensão da árvore, aumentam a acurácia e a profundidade da análise. Dessa forma, após a definição dos valores dos hiperparâmetros, foi atestada um aumento de 5% na acurácia dos testes.

ps: uma tabela de testagens de diferentes valores de hiperparametros é apresentada no tópico 4.5.6 deste documento.

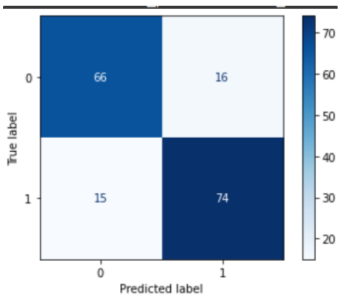
4.5.6. Avaliação de hiperparâmetros

4.5.6.1. No inicio da modelagem

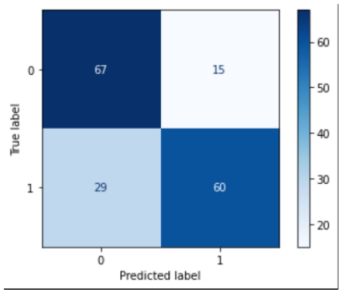
A partir da utilização das **features**: 'Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho' e 'Genero_Numerico’, foram realizados testes com os hiperparâmetros da Árvore de decisão, levando em conta os resultados do **Random Search e do Grid Search** (apresentados na secção 4.4.1.3.4 deste documento), que mostraram certo direcionamento para escolhermos os valores, mesmo que eles mesmo não tenham apontado a melhor combinação possível de valores de hiperparâmetros. Dessa forma, segue na tabela 2 os resultados desses experimentos:

tabela2:

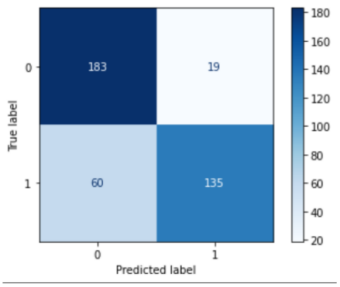
	sem hiperparametros	pelo **RANDOM SEARCH:** 'splitter': 'random', 'min_samples_split': 9, 'min_samples_leaf': 19, 'max_depth': 3, 'criterion': 'gini'	pelo **GRID SEARCH:** 'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 5, 'min_samples_split': 22, 'splitter': 'best'	min_samples_split=14, min_samples_leaf=10, max_depth=4, criterion= 'gini'	min_samples_split=17, min_samples_leaf=5, max_depth=4, criterion= 'gini'	min_samples_split=17, min_samples_leaf=5, max_depth=4, criterion= 'entropy'
id	1	2	3	4	5	6
random_state	42	83	42	42	42	42
acurácia (teste)	0.80	0.74	0.82	0.80	0.84	0.82
acurácia (treino)	1.0	0.75	0.80	0.79	0.81	0.80
curva ROC (AUC)	0.82	0.77	0.88	0.84	0.88	0.88



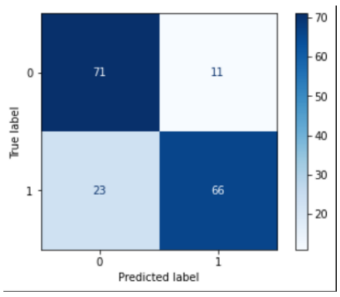
id=1



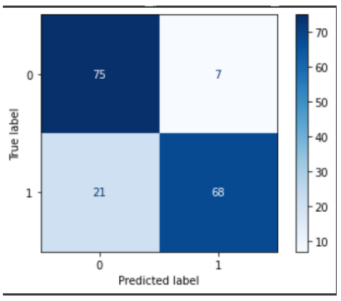
id=2



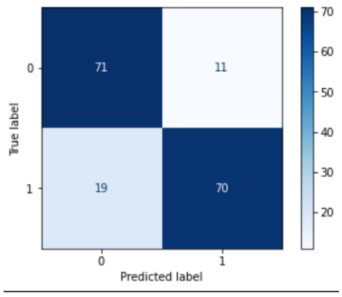
id=3



id=4



id=5



id=6

A partir de uma análise da tabela 2 e das matrizes de confusão abaixo da tabela, é possível perceber que os melhores resultados foram alcançados com os valores dos hiperparâmetros definidos da coluna com id=5 (destacada em verde), de acordo com as métricas de avaliação da acurácia de teste, acurácia de treino, curva ROC, e matrizes de confusão (apresentadas nas seções 4.5.1, 4.5.2 e 4.5.3 deste documento).

4.5.6.2. Hiperparâmetros escolhidos

Tendo em vista os testes realizados na seção 4.5.6.1., concluímos que os seguintes hiperparâmetros eram os que melhor se encaixavam no algoritmo da árvore de decisão:

- `random_state=32,`
  - ➢ O random_state é usado para que os valores fiquem aleatórios, isso permite variabilidade e um modelo com menos viés.
- `criterion = 'gini',`
  - ➢ O critério Gini mede a impureza de cada comparação, ou seja, ele calcula, em todas as decisões, a frequência com que cada conjunto vai ser erroneamente classificado SE for classificado aleatoriamente.
- `max_depth = 8,`
  - ➢ Max depth mede a profundidade máxima da árvore, ou seja, quantas decisões e comparações o modelo faz até chegar nos nós finais.
- `min_samples_leaf = 2,`
  - ➢ Esse hiperparâmetro determina o número mínimo de observações e informações que cada folha deve ter.
- `min_samples_split = 3,`
  - ➢ Esse hiperparâmetro garante o mínimo de nós que um corte na árvore faça, ou seja, é a mínima quantidade de nós folhas a mais que um nó raiz gera.

4.5.7. Avaliação de Estabilidade

Ainda com as features usadas nos testes do tópico 4.5.5., foram realizados testes com variações do hiperparâmetro “random state” na Árvore de decisão. Esse hiperparâmetro controla a aleatoriedade envolvida no aprendizado da máquina de árvore de decisão. Dessa forma, segue na tabela 3 os resultados desses experimentos:

A fim de avaliar a estabilidade do nosso modelo hiper-parametrizado de árvore de decisão, fizemos testes mantendo as features usadas nos testes do tópico 4.5.5 variando o hiperparâmetro “random state” na árvore. Esse hiperparâmetro controla a aleatoriedade envolvida no aprendizado do modelo preditivo, tendo influência relevante na acurácia e precisão

dele. Nesse sentido, pudemos perceber os seguintes resultados: a acurácia do modelo altera menos de 2% para todos os random_states testados de 0 a 500.

4.5.8 Comparação de Modelos

4.5.8.1 Comparação do inico da fase de modelagem

A partir da avaliação dos modelos, tornou-se possível comparar os resultados das experimentações de cada modelo, através da criação de tabelas e ferramentas visuais. Dessa forma, conclusões foram alcançadas em relação tanto à escolha das feature engineerings quanto à escolha dos modelos mais precisos para o objetivo central do projeto, isto é, classificar os funcionários para saber se eles têm ou não chance de saírem da empresa.

Uma etapa importante do processo de comparação dos modelos foi a comparação das **taxas de acurácia** de teste que cada um apresentou para cada combinação de possíveis variáveis que se mostraram mais relevantes. Dessa forma, foi gerada uma tabela (tabela 1), em que as linhas representam cada modelo testado, e as colunas representam cada combinação de variáveis. Além disso, a coluna “Média de acurácia” apresenta a acurácia média de cada modelo diante das experimentações feitas.

tabela 1:

	Média de acurácia	tamanho do teste	Idade, Cargo, Regiao, SP, Salario Comparado	faixa etaria, Cargo, Regiao, EstadoSP, Salario Mês, Salario Comparado	faixa etaria, Salario Comparado, SP	faixa etaria, Salario Comparado, SP, Estagnação	faixa etaria, Salario Comparado, SP, Estagnação, Cargo	Idade, Cargo, Região, Salario Comparado, Estagnação	Idade, Salario Comparado, SP, Estagnação	faixa etaria, Cargo, Regiao, Salario Comparado, Estagnação
KNN	67%	0.3	0.65 - 65%	0.63 - 63%	0.69 - 69%	0.69 - 69%	0.67 - 67%	0.69 - 69%	0.69 - 69%	0.67 - 67%
Árvore de decisão	72%	0.3	0.6783 67%	0.6783 67%	0.7343 73%	0.7482 74%	0.7622 76%	0.7203 72%	0.7063 70%	0.7762 77%
SVM	64%	0.3	0.678321 - 68%	0.608391 - 60.8%	0.699300 - 70%	0.62937 - 63%	0.62937 - 63%	0.62937 - 63%	0.62937 - 63%	0.62937 - 63%
Naive Bayers	67%	0.3	0.650349 - 65%	0.601398 - 60%	0.678321 - 67%	0.67832 - 67%	0.67132 - 67%	0.69930 - aprox 70%	0.67132 - 67%	0.72027 - 72%
Regressão logística	71%	0.3	0.699300 - aprox 70%	0.748251 - 75%	0.727272 - 72%	0.748251 - 74%	0.720279 - 72%	0.706293 - 70%	0.741258 - 74%	0.692307 - 69%

Nesse contexto, é possível analisar que os modelos de Árvore de decisão e de Regressão Logística foram os que apresentaram os melhores desempenhos em relação aos outros, com média de acurácia de 72% e 71%, respectivamente, e atingiram as maiores taxas individuais da tabela (a Árvore de decisão com 77%). A partir disso,vamos para a análise dos conjuntos de

variáveis, especialmente nesses dois modelos, e percebe-se que o uso das variáveis “Salário Mês” e “idade” diminuem a acurácia, levando à desconsideração delas nos próximos experimentos. Além delas, anterior a essas testagens, foi percebida uma ineficácia da variável “Gênero”, por isso descartada antes mesmo da rodada oficial de testes.

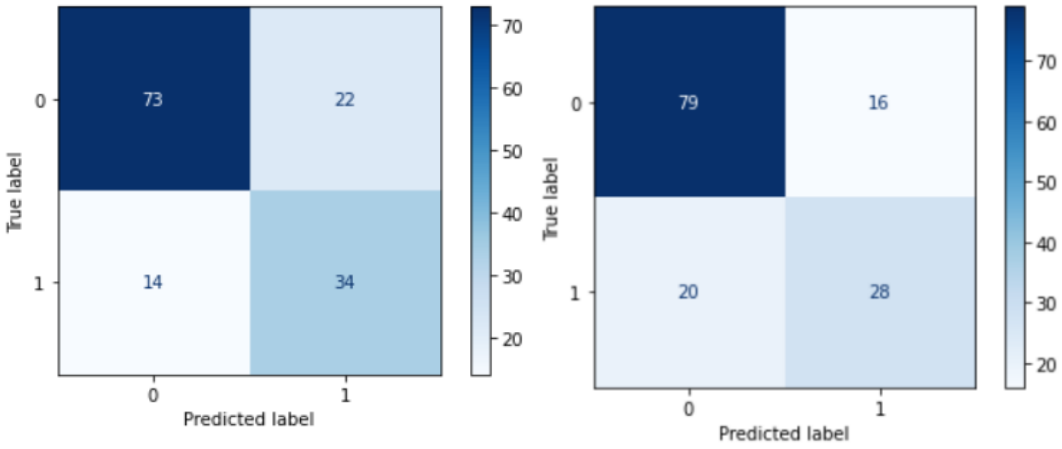
Dessa forma, as 3 experimentações de melhor desempenho estão em destaque em azul na tabela 1, e, para melhor visualização, foram separadas em uma nova tabela (tabela 2).

tabela 2:

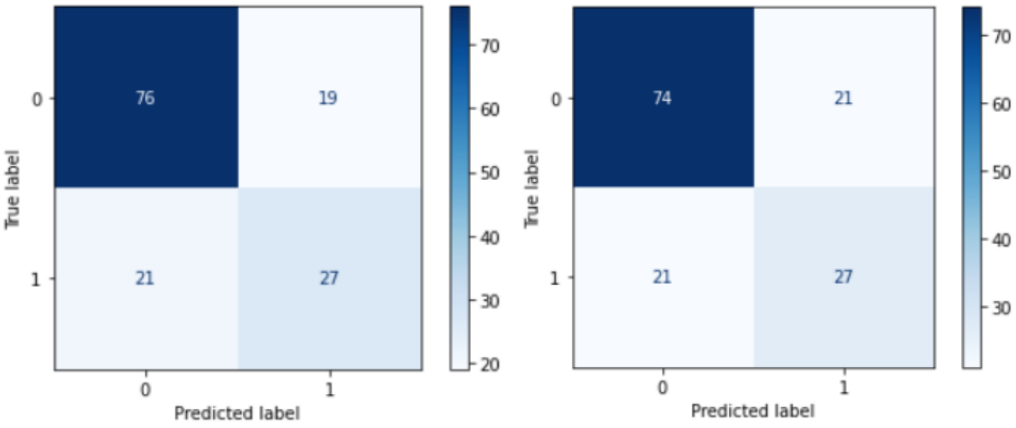
	faixa etaria, Salario Comparado, SP, Estagnação	faixa etaria, Salario Comparado, SP, Estagnação, Cargo	faixa etaria, Salario Comparado, Região, Estagnação, Cargo
Árvore de decisão	0.7482 74%	0.7622 76%	0.7762 77%
Regressão logistica	0.7482 - 74%	0.7202 - 72%	0.6923 - 69%

Pode-se perceber que o conjunto de variáveis referentes à “faixa etária, salário comparado, SP e estagnação” gerou uma acurácia de aproximadamente 74% para os dois modelos (1º experimento), e, com a adição da variável “cargo”, a acurácia da Árvore de decisão aumentou em aproximadamente 2%, mas a acurácia da regressão logística diminuiu em 2%, o que ilustra como, em termos de acurácia, as variáveis diferem de comportamento entre modelos diferentes, e é por isso que utilizamos outras formas de avaliação para complementar a anterior, como a matriz de confusão. Seguindo na análise, o mesmo acontece com a substituição da variável “SP”, pela variável “Região”, que aumenta a acurácia da árvore de decisão, mas diminui a da regressão logística.

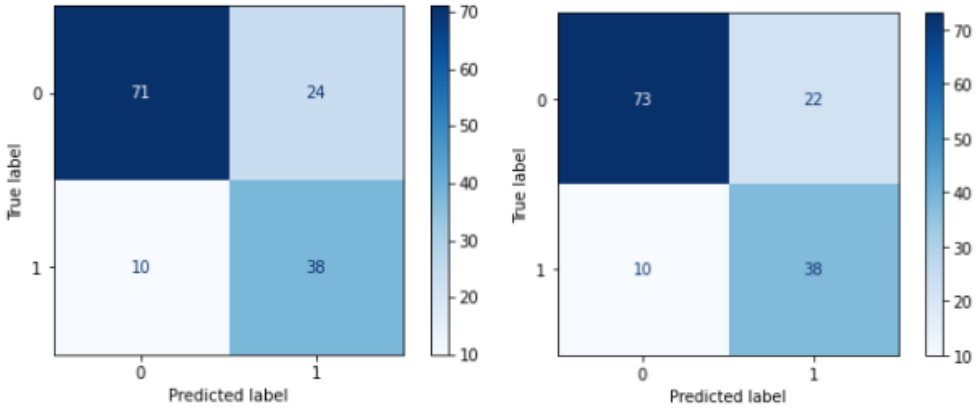
Logo, diante das análises feitas através da acurácia, seguimos para as análises pela **matriz de confusão**, para decisões mais profundas sobre as ocorrências dos erros e acertos do modelo.



Ambas matrizes de confusão acima apresentam acurácia de 74,8%. A matriz de confusão da direita é resultado da regressão logística, e a da esquerda é resultado da árvore de decisão. Apesar de ambas possuírem a mesma acurácia, classifica-se como superior a matriz da esquerda, por virtude do fato de que ela apresenta menor taxa de erro no quadrante inferior esquerdo.



Acima, a matriz de confusão do lado esquerdo apresenta 72,2% de acurácia, e a do lado direito, 69%. Ambas são resultado de regressão logística, e possuem quase as mesmas taxas de erro. No entanto, as duas combinações de variáveis que formaram essas matrizes foram descartadas para uso em regressão logística, pois possuem uma acurácia menor do que as demais combinações para esse modelo.



As duas matrizes acima são resultados de árvore de decisão. A matriz da direita possui 76% de acurácia; a da esquerda, 77%. Ambas possuem a mesma taxa de erro na métrica do quadrante inferior esquerdo, a mais importante. Deve-se analisar, portanto, o canto superior direito: posto que a matriz da esquerda (77%) possui uma menor taxa de erro nesse quadrante, ela mostra-se como a mais adequada dentre todas as demais.

4.5.8.2 Comparação ao final da modelagem

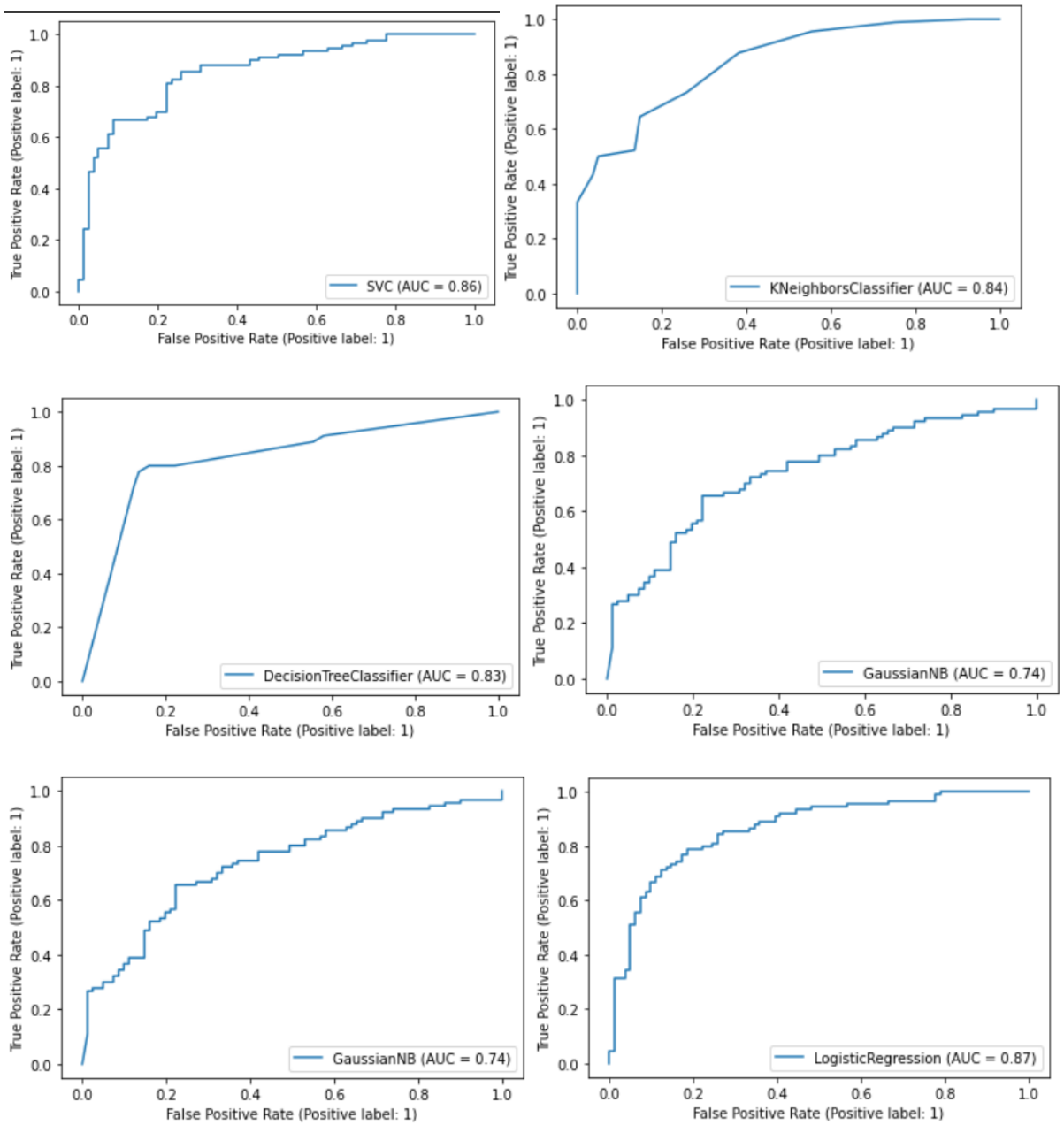
-> Resultados em relação à taxa de **Acurácia de treino e teste** de cada modelo:

```
Acurácias de treino:
SVM : 0.7732997481108312
KNN : 0.7128463476070529
DecTree : 0.906801007556675
Bayes : 0.6120906801007556
LogReg : 0.8211586901763224
NeuralNet : 0.9345088161209067
------------------------------------------------------------
Acurácias de teste:
SVM : 0.7602339181286549
KNN : 0.6842105263157895
DecTree : 0.7953216374269005
Bayes : 0.5263157894736842
LogReg : 0.7953216374269005
NeuralNet : 0.7719298245614035
```

-> Resultados em relação à **matriz de confusão** de cada modelo:

```
SVM :       KNN :       DecTree :   Bayes :      LogReg :    NeuralNet :
[[70 11]    [[70 11]    [[71 10]    [[80  1]     [[66 15]    [[59 22]
 [30 60]]    [43 47]]    [25 65]]    [80 10]]     [20 70]]    [17 73]]
```

-> Resultados em relação à **CURVA ROC** de cada modelo:



E analisando todas as medidas de precisão apresentadas acima, como acurácia, matriz de confusão e curva ROC, é possível perceber o motivo da escolha da árvore de decisão como

modelo definitivo, ela apresenta a melhor taxa de acurácia em treino e teste, com 90% e 79%, a sua matriz de confusão apresenta a menor taxa de erro, e a sua curva ROC apresenta uma das melhores taxas de precisão, cerca de 86%. E portanto, foi levando todos esses fatores em consideração que decidimos escolher a árvore de decisão.

5. Conclusões e Recomendações

Com o desenvolvimento desta solução, pode-se concluir que sua implementação irá contribuir para muitos fatores internos da Everymind. Um deles é a diminuição da taxa de turnover, uma vez implementado, o modelo pode fazer com que se encontre o motivo de tendência de saída dos colaboradores. Depois de descobrir esses motivos, um funcionário que esteja em uma posição de liderança pode decidir como proceder com os colaboradores que têm a tendência de desligamento da empresa.

Vale ressaltar que não é o algoritmo de predição que deve tomar as decisões, sempre cabe aos líderes tomar a decisão final, já que esses são responsáveis pelas pessoas do seu setor. Além disso, esses líderes devem sempre tentar entender o motivo da tendência de desligamento de um colaborador com o próprio funcionário, mesmo depois de fazer a análise com o modelo.

O modelo escolhido para implementação foi a Árvore de Decisão, já que este modelo é acessível e visual, e teve a maior acurácia que os outros modelos testados. Foram feitas diversas avaliações , como a matriz de confusão, curva ROC, avaliação de features, avaliação com hiperparâmetros e a comparação com outros modelos, para aumentar a acurácia que tínhamos no início. Depois de aplicar essas técnicas, foi obtida a acurácia de 81% nos testes.

6. Anexos

- Dicionário das colunas finalizadas com “_Numerico”:

- Gênero Numérico (df1):
  - 0 representa "masculino";
  - 1 representa "feminino".
- Tipo Saída Numérico (df1):
  - 0 representa ativo;
  - 1 representa rescisão de contrato por pedido de demissão;
  - 2 representa rescisão de contrato por demissão;
  - 3 representa demissão;
  - 4 representa pedido de demissão.
- Região Numérico (df1):
  - 1 representa a região Norte;
  - 2 representa a região Nordeste;
  - 3 representa a região Centro-Oeste;
  - 4 representa a região Sudeste;
  - 5 representa a região Sul.
- Situação (df2):
  - 0 representa "desativado";
  - 1 representa "ativo" ou "afastado".
- Status (df1):
  - 0 representa "desativado";
  - 1 representa "ativo".
- Salário Comparado (df1):
  - 0 indica que o salário do colaborador referido está igual ou maior à média salarial do cargo;

- 1 indica que o salário do colaborador referido é menor do que a média salarial do cargo.

- Faixa etária (df1):
  - 0 representa idades entre 18 e 21;
  - 1 representa idades entre 22 e 25;
  - 2 representa idades entre 26 e 29;
  - 3 representa idades entre 30 e 33;
  - 4 representa idades entre 34 e 37;
  - 5 representa idades entre 38 e 41;
  - 6 representa idades entre 42 e 45;
  - 7 representa idades entre 46 e 49;
  - 8 representa idades entre 50 e 65.
- Estado SP (df1):
  - 0 representa aqueles que não moram em SP;
  - 1 representa aqueles que moram em SP.
- EC Numérico (df1):.
  - 0 representa os solteiros (incluindo divorciados e separados)
  - 1 representa casados (incluindo união estável)
  - Nós partimos da hipótese que os divorciados e separados, assim como os casados, podem ter dependentes ou querer possuir uma maior estabilidade.
- Cidade numérica(df1):
  - A cidade de cada um dos colaboradores foi transformado em um número

- Área (df1):

14 - Education 15 - Financeiro 16 - Infraestrutura 17 - Integration 18 - MktCloud 19 - PS 20 - People 21 - Produtos	6 - CPG&Retail 7 - CPG&RetailI 8 - CPG&RetailII 9 - Commerce 10 - Core&Industrias 11 - Core&IndustriasI 12 - Core&IndustriasII 13 - Diretoria	0 - AMS 1 - AgenciaDigital 2 - Analytics 3 - BAC 4 - BPM 5 - BestMinds 22 - Vendas

- Estado Civil (df1):
    - 0 - solteiros (incluindo divorciados e separados)
    - 1 - Casados (incluindo união estável)
- Cidades (df1)

8 = BelaVistadeGoiás 6 = BalneárioCamboriú 5 = Bacabal 3 = Anápolis 2 = Ananindeua 34 = Embu-Guaçu 36 = Eusébio 72 = NovaLima 52 = Itanhaém 71 = Novalguaçu 69 = Navegantes 67 = MogiMirim 65 = Maringá 64 = Maricá 63 = Mairiporã 10 = Belém 37 = Florianópolis	99 = SantaMariadoPará 0 = Abaetetuba 73 = NovoHamburgo 19 = CampoLimpoPaulista 33 = DoisVizinhos 28 = Contagem 27 = Congonhas 25 =Colombo 24 = Charqueadas 23 = Caucaia 22 = Catanduvas 20 = Canoas 18 = CampoGrande 35 = Erechim 14 = Cabreúva 13 = Cabedelo 11 = BragançaPaulista	58 = JuazeirodoNorte 56 = Jandira 51 = Itaiópolis 40 = FranciscoBeltrão 50 = Ipanema 49 = Indaiatuba 48 = Igarassu 47 = Igarapé 44 = Guaiçara 43 = Guaimbê 42 = Garanhuns 41 = FranciscoMorato 118 = Vinhedo 108 = SãoCaetanodoSul 17 = Campinas 53 = ItapecericadaSerra 54 = Itapevi	97 = SantaBárbaraD'Oeste 75 = Palmital 110 = SãoJosé 79 = PedroLeopoldo 117 = Unaí 103 = Sertãozinho 81 = Pindamonhangaba 115 = Taubaté 83 = PortoAlegre 104 = Sobral 95 = Salgado 111 = SãoLourenço 92 = RioGrande 109 = SãoCristóvão 98 = SantaMariadaVitória 91 = RibeirãoPreto 87 = Promissão

86 = PresidentePrudente 77 = Paulista 78 = Paulínia 70 = Niterói 26 = Concórdia 4 = Atibaia 60 = Limeira 1 = Alfenas 84 = Poá 57 = JoãoPessoa 113 = SãoVicente 106 = Suzano	9 = BeloHorizonte 89 = Recife 66 = Mauá 30 = Curitiba 21 = Carapicuíba 55 = Itaquaquecetuba 94 = RiodeJaneiro 38 = Fortaleza 31 = Diadema 29 = Cotia 102 = Santos 59 = Jundiaí 116 = Uberlândia 114 = TaboãodaSerra	16 = Camaragibe 15 = Caieiras 93 = RioNegro 39 = Franca 32 = Divinópolis 90 = RibeirãoPires 82 = Piracicaba 76 = Patos 100 = SantanadeParnaíba 80 = Pelotas 46 = Hortolândia 96 = Salvador	88 =RafaelFernandes 112 = SãoPaulo 74 = Osasco 107 = SãoBernardodoCampo 45 = Guarulhos 101 = SantoAndré 105 = Sorocaba 68 = MogidasCruzes 61 = Londrina 85 = PraiaGrande 62 = Mafra 7 = Barueri 12 = Brasília

7. Referências

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.

[falta colocar cada referência para cada secção em ordem crescente. o 1 já está feito, usá-lo como modelo]

1. Introdução
    a. https://mcjb15vjp4x3shyj9vwqlqvwnky1.pub.sfmc-content.com/vczccluo15c - Acessado em 04/10/2022
    b. Sobre Nós – Everymind - Acessado em 04/10/2022
2. CHAPMAN, Pete; CLINTON, Julian; KERBER, Randy;KHABAZA, Thomas; REINARTZ Thomas; SHEARER, Colin; WIRTH, Rüdiger. CRISP-DM 1.0: Step-by-step Data Mining Guide. SPSS, 2000
3. Imagem arborescente por Do not want - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=14947263 - Acessado em 04/10/2022
4. DELEUZE, Gilles; GUATTARI, Felix. (1987) [1980]. A Thousand Plateaus. Translated by Massumi, Brian. University of Minnesota Press.
5. LAND, Nick. Fanged Noumena: Collected Writings 1987-2007, ed. Robin Mackay and Ray Brassier (Urbanomic, 2011). ISBN 978-0955308789
6. Turnover de funcionários: os fatores que aumentam a demissão - Acessado em 04/10/2022
7. Idosos estão adiando cada vez mais saída do mercado de trabalho | Agência Brasil - Acessado em 04/10/2022
8. Entenda o que é AUC e ROC nos modelos de Machine Learning | by Vinícius Rodrigues | bio-data-blog | Medium - Acessado em 04/10/2022
9. https://www.organicadigital.com/blog/algoritmo-de-classificacao-naive-bayes/
10. https://www.digitalhouse.com/br/blog/naive-bayes/
11. https://www.voitto.com.br/blog/artigo/teorema-de-bayes
12. https://www.suno.com.br/artigos/teorema-de-bayes/
13. Curvas ROC - Acessado em 04/10/2022
14. https://itforum.com.br/colunas/turnover-em-ti-como-analisar-e-criar-estrategias-para-evita-lo/ - Acessado em 06/10/2022

15. https://neilpatel.com/br/blog/erp-o-que-e/ - Acessado em 6/10/2022.
16. https://medium.com/@kvmoura/crisp-dm-79580b0d3ac4 - acessado em 10/06/2022