



Mettha Bettha

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
11/08/2023	João Paulo da Silva	1.0	Criação do documento e atualização das seções 1, 2.1, 2.2, 2.3, 4.1.1, 4.1.2, 4.1.3, 4.1.4, 4.1.5, 4.1.6 e 4.1.7
27/08/2023	Heloisa Oliveira, João Paulo da Silva	2.0	Atualização da seção 4.2
06/09/2023	Enzo Boccia, Otto Lima	3.0	Atualização das seções 3 e 4.3
24/09/2023	Clara Mohammad, Ever Sousa, Pedro Faria	4.0	Atualização da seção 4.4
05/09/2023	Clara Mohammad, Heloisa Oliveira	5.0	Atualização da seção 4.5 e 5. Revisão e modificações finais em todo o documento.

Sumário

1. Introdução	4
2. Objetivos e Justificativa	4
2.1. Objetivos	4
2.2. Proposta de Solução	5
2.3. Justificativa	5
3. Metodologia	6
3.1. CRISP-DM	6
3.2. Ferramentas	6
3.3. Principais técnicas empregadas	7
4. Desenvolvimento e Resultados	8
4.1. Compreensão do Problema	8
4.1.1. Contexto da indústria	8
4.1.2. Análise SWOT	8
4.1.3. Análise das 5 forças de Porter	11
4.1.4. Planejamento Geral da Solução	12
4.1.5. Value Proposition Canvas	13
4.1.6. Matriz de Riscos	14
4.1.7. Personas	15
4.1.8. Jornadas do Usuário	19
4.1.9. Política de Privacidade e LGPD	20
4.2. Compreensão dos Dados	27
4.2.1. Descrição dos dados	27
4.2.2. Descrição estatística básica dos dados	28
4.2.3. Descrição da predição desejada	33
4.3. Preparação dos Dados	34
4.3.1. Aprendizado de máquina	34
4.3.2. Descrição das manipulações realizadas	34
4.3.3. Agregação e derivação de atributos	35
4.3.4. Tratamento de valores nulos	35
4.3.5. Features selecionadas	36
4.4. Modelagem	39
4.4.1 KNN + K-means	39
4.4.2. Random Forest + K-means	41
4.4.3. Naive Bayes + K-means	43
4.5. Avaliação	46
5. Conclusões e Recomendações	47
6. Referências	48
Anexos	49

1. Introdução

O parceiro de negócios desse projeto é a empresa Bettha, cujo objetivo é ajudar jovens no início de suas carreiras a encontrarem vagas de empregos, além de oferecer capacitações para aprimorarem suas habilidades. Ela atua de forma 100% digital, tendo como público alvo, brasileiros, no início de suas carreiras, e ajudando-os na construção de seus perfis, além de encontrar oportunidades de carreiras que combinam com suas características.

O projeto que o grupo Mettha desenvolve tem como proposta resolver a demanda da Bettha, que trouxe como problema a falta de conhecimento das pessoas, que buscam por emprego, sobre quais capacitações podem ser importantes ao aplicarem às vagas e quais oportunidades possuem um maior alinhamento com seus perfis.

2. Objetivos e Justificativa

Nesta seção, será apresentado os objetivos do projeto juntamente com suas justificativas. Além disso, também será citado a proposta de solução e sua relevância para os parceiros.

2.1. Objetivos

A empresa Bettha procura ajudar as pessoas a encontrar vagas de empregos condizentes com suas personalidades, perspectivas e formações. Para isso, oferece testes que analisam o perfil do usuário e que o guiam para uma escolha assertiva. Para alavancar a sua plataforma, a empresa pretende implementar um modelo preditivo que, de acordo com os resultados dos usuários, faça um pareamento entre eles e as empresas contratantes, de forma a auxiliar o candidato a encontrar seu emprego dos sonhos.

2.2. Proposta de Solução

O modelo preditivo propõe a indicação de vagas de emprego para o usuário com base nos resultados dos testes, como de *lifestyle* e *genius* (testes fornecidos pelo Bettha ao fazer cadastro na plataforma), propostos pela Bettha. Dessa forma, é possível conectar os valores e cultura de cada indivíduo com a empresa que mais se alinha com o seu perfil. Assim, os usuários terão potencialmente uma maior satisfação profissional, por estarem em um ambiente condizente com seus valores, além de garantir para a empresa contratante, que seus novos funcionários estejam engajados e que agregam valor para a companhia.

2.3. Justificativa

A proposta do projeto é reforçada por diversos fatores que, em conjunto, são capazes de alavancar os lucros da empresa e permitir uma maior relevância no mercado, além de garantir uma abordagem mais humanizada para os candidatos.

Em primeiro lugar, um dos principais fatores, o modelo preditivo é capaz de analisar os candidatos de maneira holística, ou seja, considerando o indivíduo como um todo, levando em conta tanto as competências técnicas quanto socioemocionais, permitindo uma abordagem mais humana ao selecionar pessoas não somente pela sua capacidade técnica, mas também pelas habilidades socioemocionais e comportamentais.

Em segundo lugar, o modelo é capaz de promover uma maior escalabilidade do site da Bettha por meio da automatização da tarefa de recomendação de vagas. Com o uso do machine learning, a plataforma conseguirá garantir uma maior eficiência e velocidade no momento de contratação, garantindo que sejam admitidos funcionários que realmente se alinham com os valores da empresa, o que potencialmente promove uma maior satisfação profissional e realização pessoal. Portanto, há uma maior chance de integrar essas pessoas na companhia e retê-las, diminuindo drasticamente o *turnover*.

Em suma, a nossa proposta é benéfica para ambos os lados, empresa e candidato, à medida que ajuda a construir um ambiente de trabalho mais engajado e fluido. Assim, pode garantir que a companhia aumente sua produtividade geral e, conseqüentemente, seus lucros, e possivelmente promovendo uma maior satisfação do funcionário, de forma a fomentar seu bem-estar geral.

3. Metodologia

Nesta seção, serão abordadas as principais metodologias utilizadas durante o desenvolvimento do projeto, em que o grupo Mettha trabalhou principalmente com a metodologia CRISP-DM. Além disso, serão detalhadas as ferramentas utilizadas e outras técnicas implementadas.

3.1. CRISP-DM

A metodologia *CRISP-DM* (em inglês: Cross Industry Standard Process for Data Mining) é utilizada para organização do processo de mineração e tratamento de dados. A partir deste modelo, a equipe Mettha traçou algumas etapas sequenciais, as quais podem ser revisitadas ao longo do processo.

A primeira etapa não está diretamente relacionada aos dados e se trata do Entendimento do Negócio, visando contextualizar a equipe sobre o problema e entender quais necessidades e expectativas os dados precisam atender. Em seguida, passamos pela etapa de Compreensão dos Dados, na qual devemos analisar todas as bases de dados e entender a correlação entre as variáveis e as colunas, conhecer os dados disponíveis e avaliar a sua qualidade. Já a terceira etapa, se trata da Preparação dos Dados e é durante ela que realizamos a seleção dos dados para a construção do modelo, assim como a limpeza de inconsistências, adequação e ajustes.

Posteriormente, iniciamos a fase de Modelagem, na qual se empregam técnicas de mineração e algoritmos, a partir dos quais realizam-se vários testes para calibração de parâmetros. Depois, na quinta etapa, a de Avaliação, analisam-se as saídas entregues pelo algoritmo e elas são avaliadas se estão de acordo com as expectativas do projeto. Por fim, existe a fase de Implantação, em que o projeto é de fato entregue e validado pelo cliente.

3.2. Ferramentas

A principal ferramenta utilizada pelo grupo no estudo dos dados foi a interface Google Colaboratory (Colab) que serve para auxiliar no processo de machine learning e exploração da inteligência artificial. O Google Colab é uma plataforma que trabalha com linguagens de programação (geralmente Python), com marcações de textos (conhecidas por markdown), permitindo com que os usuários possam rodar e executar os códigos com

mais flexibilidade e rapidez, já que ela contém seu próprio sistema de nuvem que armazena as informações.

Outro ponto importante, é o uso de bibliotecas do python como Pandas, NumPy e Matplotlib que foram utilizadas durante o pré-processamento dos dados e, portanto, são consideradas ferramentas para o desenvolvimento do projeto. Mais informações sobre como elas são usadas podem ser encontradas na seção 4.3.

3.3. Principais técnicas empregadas

Inicialmente, foi feito o *merge* das tabelas para que os dados possam ser analisados de forma conjunta. Em seguida, para o tratamento de dados, foram removidas informações duplicadas e dados null. Por fim, a denominação com label encoding para as colunas categóricas. É importante ressaltar que as classificações em churn e não-churn é binário.

Os algoritmos com melhor acurácia foram Random Forest e árvore de decisão, os seus métodos de decisão são respectivamente: uma grande “floresta” de árvores de decisão onde cada árvore parte de um dado diferente das outras árvores, e a árvore de decisão parte de um dado e vai se bifurcando a partir dos dados anteriores.

O Random Forest tem a vantagem de partir de vários dados diferentes, se tornando menos genérica e cada vez mais específicas, sendo geralmente mais precisa. A árvore de decisão parte de um dado e se torna mais precisa baseada nos resultados anteriores de sua análise.

A plataforma Google Colaboraty foi utilizada a fim de estruturar o modelo preditivo de forma que qualquer dataframe por ele selecionado seja funcional para a empresa.

4. Desenvolvimento e Resultados

Nesta seção, serão abordados todos os passos que o grupo Mettha realizou para desenvolver a solução, além do processo de análise e compreensão do problema proposto pelo Bettha, incluindo o estudo sobre os dados.

4.1. Compreensão do Problema

Primeiramente, é necessário entender a indústria e o mercado em que a empresa parceira Bettha está inserida. Para tal, os próximos tópicos irão contextualizar esse cenário, por meio de uma análise de negócios desenvolvida pelo grupo Mettha que inclui Análise SWOT e 5 Forças de Porter, além de um estudo sobre a experiência do usuário, passando brevemente pela solução geral desenvolvida e Matriz de Riscos do Projeto. Por fim, é apresentado a política de privacidade da empresa e sua aplicação no projeto.

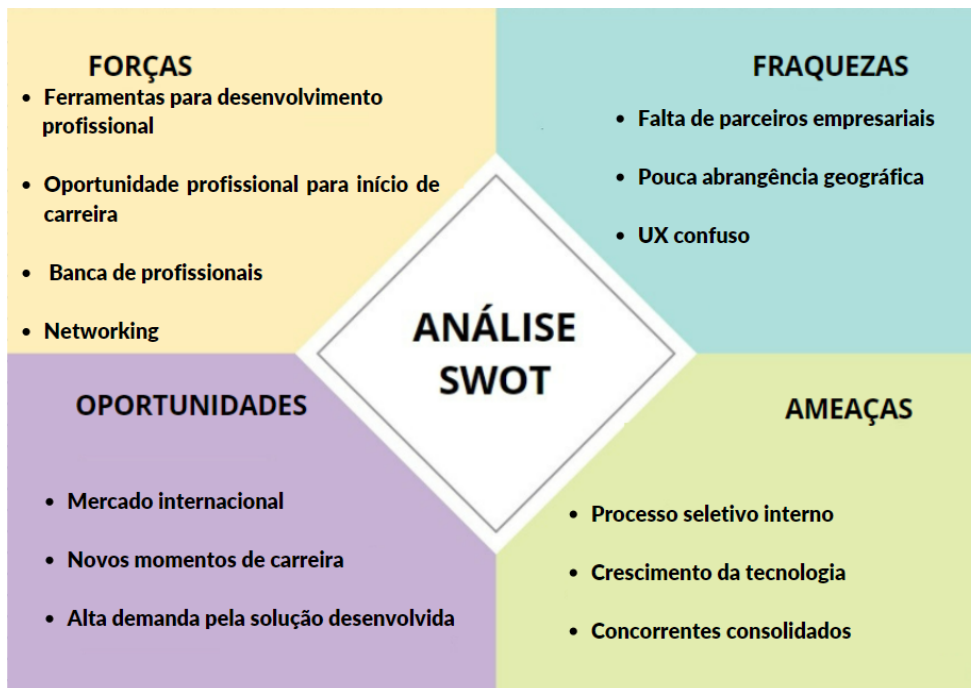
4.1.1. Contexto da indústria

A empresa parceira Bettha trabalha no setor de recrutamento e seleção. Dessa forma, é possível compará-la com outras empresas do mercado, que possuem objetivos parecidos, como o LinkedIn. Assim, o grupo Mettha fez uma análise mais detalhada acerca do setor que a empresa está inserida, que pode ser lida no tópico 4.1.3 desta seção.

4.1.2. Análise SWOT

A análise SWOT é amplamente utilizada para avaliar a posição estratégica de uma empresa em relação ao mercado e à concorrência. Essa ferramenta consiste em avaliar o ambiente interno (forças e fraquezas) e externo (oportunidades e ameaças). Esses ambientes, respectivamente, dizem sobre o que está no controle da empresa, que depende da mesma, e sobre os fatores que não dependem, considerando o contexto em que está inserida. Quando bem desenvolvida e interpretada, a análise SWOT oferece um diagnóstico confiável que demonstra as reais necessidades da empresa, permitindo a elaboração de planos de ação e planejamento mais seguros e eficazes a médio e longo prazo. Podemos analisar na imagem abaixo o nosso modelo de matriz swot:

Figura 1: Matriz SWOT da empresa Bettha



Fonte: Elaborado pelos autores

Fraquezas:

Pouca abrangência geográfica → A plataforma, por focar no público brasileiro, acaba limitando suas oportunidades para aqueles que buscam empregos no exterior, além de restringir um potencial crescimento do Bettha.

Falta de parceiros empresariais → O sucesso de um dos principais serviços oferecidos pelo Bettha está vinculado à qualidade e quantidade de empresas e empregadores parceiros. A falta de parceria pode ser causada por diversos fatores, como um marketing pouco atrativo da empresa. Essa dependência pode resultar em variações nas ofertas de vagas e prejudicar a consistência das oportunidades.

Ux confuso → A plataforma desenvolvida pelo Bettha oferece muitos recursos através de um único lugar, o que é uma ótima escolha para facilitar o acesso, entretanto, a navegação no site acaba ficando mais confusa por conter muita informação.

Forças:

Oportunidade profissional para início de carreira → O Bettha se destaca por oferecer um amplo espectro de oportunidades de empregos. A plataforma beneficia principalmente os jovens que buscam iniciar ou progredir em suas carreiras, abrindo portas para diversos setores e experiências.

Ferramentas para desenvolvimento profissional → A plataforma oferece acesso gratuito a conteúdos e ferramentas que ajudam a aprimorar habilidades e a se preparar para entrevistas. Isso demonstra um compromisso com o desenvolvimento profissional dos usuários, gerando fidelidade e confiabilidade.

Networking → A empresa oferece oportunidade dos candidatos se conectarem com outros profissionais, o que resulta não só em uma troca de conhecimentos e *insights* quanto desenvolvimento de relacionamentos valiosos para o mundo empresarial.

Banca de profissionais → A existência de um banco de currículos de jovens talentosos ajuda na comunicação entre empresas e candidatos, o que agiliza o processo de recrutamento e capacitação, beneficiando os dois lados.

Oportunidades:

Mercado Internacional → Países como Canadá e Austrália tem o mercado muito aquecido e com bastante demanda de estágio e trabalhos de início de carreira, portanto tendo um ambiente favorável para o Bettha.

Novos momentos de carreira → O mercado de TI brasileiro desenvolve pouco profissionais seniores, ou seja, em razão dessa falta de profissionais, o sistema do Bettha pode adentrar nesse mercado favorável e aquecido.

Alta demanda pela solução desenvolvida → Devido à instabilidade financeira no Brasil, muitos recém-formados possuem dificuldades de começar a carreira por não terem muita experiência e as empresas contratantes também não terem orçamento para treinar um jovem. Assim, o Bettha se beneficia uma vez que a demanda pela sua plataforma aumenta.

Ameaças:

Processos de seletividade interno → As empresas contratantes realizam o processo seletivo de forma interna, assim a Bettha perderá vagas de empregos para fornecer aos usuários, consequentemente diminuindo a utilização de seu sistema.

Crescimento da tecnologia → Com o avanço tecnológico, empregos e estágios estão sendo substituídos pela IA e outras tecnologias, assim diminuindo as vagas para iniciantes no mercado de trabalho.

Concorrentes consolidados → Existe uma grande quantidade de concorrentes, como LinkedIn, Gumpy e onze, nesse meio empresarial, assim existe uma grande possibilidade dos seus clientes, sendo empresas ou pessoas, escolherem outras empresas para aplicar esse sistema.

4.1.3. Análise das 5 forças de Porter

As 5 forças de Porter são utilizadas para avaliar a competitividade de uma indústria. A análise consiste em ajudar as empresas a entender as forças que influenciam os ambientes competitivos e a rentabilidade. Essas forças impactam a concorrência, o acesso a fornecedores, compradores, produtos substitutos e a entrada de novos concorrentes na indústria. A partir dessa análise, vantagens competitivas e ameaças no mercado ficam evidentes para as empresas, o que possibilita um entendimento melhor do mercado para que elas tomem as melhores decisões e se destaquem em relação aos seus competidores.

Abaixo segue o template com uma análise sobre as 5 forças atuantes na indústria de *HRTechs* (empresas do segmento de RH).

Figura 2: Tabela das 5 Forças de Porter

FORÇAS?	QUEM SÃO?	QUAIS AMEAÇAS?	QUAIS AS MINHAS POSSÍVEIS REAÇÕES
F1- Concorrentes Atuais	Empresas de recrutamento e educação como: 99jobs, Gupy, recruitAI, etc.	Devido ao marketshare dessas empresas e o tamanho de seus nomes, possíveis clientes podem acabar desviando sua atenção do Bettha.	Otimizar o sistema e regra de negócios para maior consolidação no nicho atual e, eventualmente, possibilidade de explorar outros.
F2- Concorrentes Potenciais	Empresas de recrutamento estrangeiras como: Jobright.	Devido à baixa barreira de entrada no mercado, empresas no exterior podem aproveitar a oportunidade de expandir para o Brasil	Inovar e otimizar o sistema existente além de construir contatos e relações sólidas com organizações nacionais e internacionais.
F3- Produtos Substitutos	Novas tecnologias(AI's) que possam substituir o aval humano no processo seletivo.	Ao automatizar todo processo, as empresas podem internalizar toda a seleção bem como a captação.	Estar atenta às tendências para construir uma estrutura e oferta-la no momento certo.
F4- Fornecedores	Serviços de Infraestrutura como AWS, Internet, serviços de gerenciamento de time e escritório.	Caso qualquer um destes entes falhe, a operação pode parar por algum tempo e gerar prejuízos.	Efetuar verificações periódicas e ter planos de ação para eventuais ocorrências.
F5- Clientes	Tantos as empresas quanto aos usuários do sites	As ameaças que uma empresa pode ter seria o poder de negociação, por parte dos usuários, o poder de sair do site e trocar para um candidato é muito forte.	Estudar e otimizar o sistema para a melhor UX possível além de fornecer valor para as empresas.

Fonte: Elaborado pelos autores

4.1.4. Planejamento Geral da Solução

Os dados utilizados neste projeto foram coletados pela empresa parceira Bettha e disponibilizados para os integrantes do grupo em tabelas do tipo csv. Eles são referentes a scores (desvio padrão das notas) de cada usuário, sendo ele candidato ou gestor, proveniente dos testes de Genius (genialidade) e Lifestyle/Workstyle (estilo de vida/trabalho), abordados mais profundamente na seção 4.2 deste documento.

O modelo preditivo busca usar os testes de estilo de vida e habilidades para sugerir vagas alinhadas com os valores dos usuários, aumentando a satisfação e o engajamento profissional, além de agregar valor às empresas contratantes. Diante disso, o projeto será um sistema de recomendação, então não se trata de um sistema que se utiliza unicamente de um método de regressão ou classificação. Pode-se utilizar um modelo de

regressão para prever um parâmetro numérico e um classificador para prever atributos categóricos como a satisfação de um usuário.

Em suma, ambos os métodos serão utilizados para melhorar a recomendação, porém como se trata de um problema de recomendação de vagas, será utilizada também a técnica de filtragem colaborativa, que consiste em recomendar com base na semelhança entre os dados. Assim, partindo do pressuposto de que a melhor vaga para um usuário é aquela que mais se assemelha ao perfil dele (semelhança entre gestor da área e candidato), esta técnica se torna uma das opções mais indicadas.

Ademais, essa solução será utilizada dentro da plataforma da Bettha, como complemento da função de indicação de vagas. Ela será responsável por indicar vagas personalizadas que mais se encaixem com o perfil do usuário. As melhores propostas serão sugeridas por meio de uma interface que, posteriormente, pode ser implementada pela Bettha em seu site.

Outro ponto importante, que vale a pena ser destacado, são os benefícios da solução, que contemplam a maior velocidade e eficiência no processo de recrutamento, no que tange os ganhos da empresa, e a facilidade de se encontrar propostas de emprego que se encaixam com o perfil do usuário e que potencialmente lhe garantirá maior satisfação pessoal.

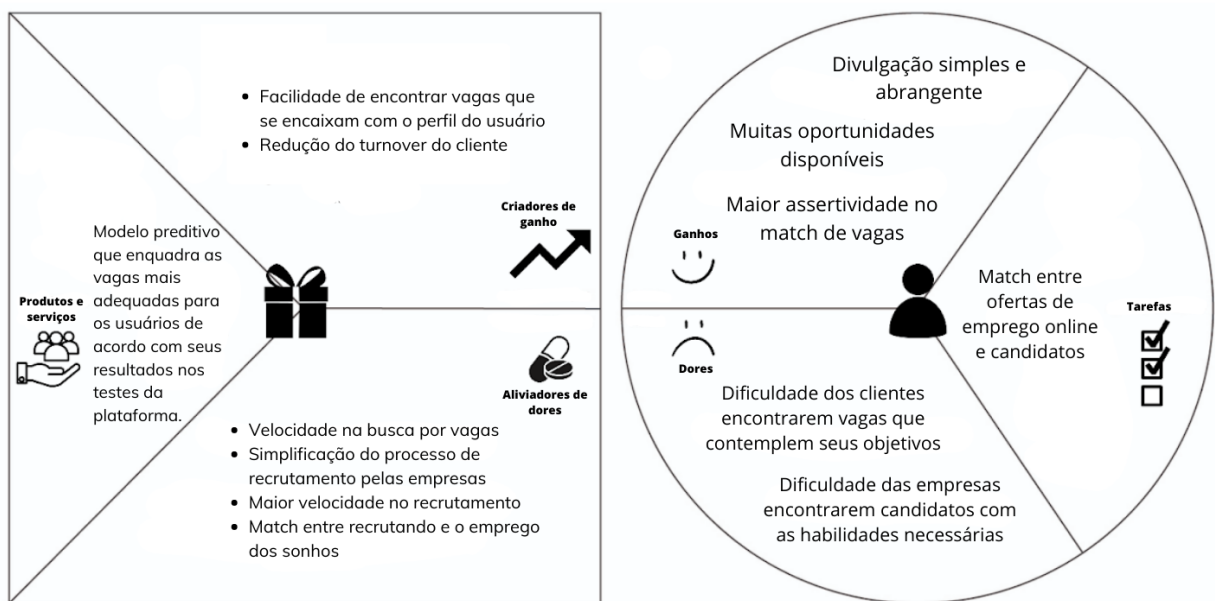
Por fim, o critério de avaliação será um modelo que consiga atingir, pelo menos, 80% de precisão, no que tange o acerto das 3 vagas mais indicadas para o usuário. Um critério utilizado será, além da semelhança entre os dados (sejam distâncias entre instâncias ou similaridade de cosseno, dependendo do método utilizado), a satisfação do usuário com a recomendação. Podem ser adotadas outras métricas, mas estas são de extrema importância para o diagnóstico da performance. É possível ver mais sobre as métricas utilizadas no projeto na seção 4.4 deste documento.

4.1.5. Value Proposition Canvas

O Value Proposition Canvas é um modelo visual que permite a visualização rápida e fácil da proposta de valor do projeto. Em suma, ela apresenta como o usuário executa suas tarefas usualmente, assim como os seus ganhos e dores com ela, e como principal

foco as vantagens geradas pelo nosso modelo, tanto os ganhos que ele cria, quantos as dores do cliente que ele sana.

Figura 3: Value Proposition Canvas do projeto Mettha



Fonte: Canvas elaborados pelos autores

4.1.6. Matriz de Riscos

A Matriz de Riscos é uma ferramenta visual que permite que a equipe tenha consciência dos riscos que o projeto apresenta e, dessa forma, tomar decisões para evitá-los. Além disso, pode-se acrescentar a Matriz de Oportunidades, em que são apresentadas as oportunidades que podem ocorrer no projeto, guiando o grupo a agir para que elas aconteçam. Segue a Matriz de Riscos elaborada pela equipe Mettha.

Figura 4: Matriz de Risco e Oportunidades do grupo Mettha

		Ameaças					Oportunidades				
Probabilidade	90%			VIA travar na hora da apresentação			Adquirir conhecimentos sobre o setor em que a empresa parceira está inserida	Adquirir conhecimentos sobre o módulo			
	70%						Desenvolver habilidades de colaboração				
	50%			Falta de conhecimentos acadêmicos	Atrasar alguma entrega de artefato	Falta de comunicação entre os membros da equipe			Ser o melhor projeto da nossa turma e se destacar		
	30%			Risco de vazamento de dados	Falha de comunicação com o cliente	Falta de engajamento por parte dos membros	Usar comunicação não-violenta no nosso grupo				
	10%	Problemas com Backup				Problemas de saúde de algum dos membros					
		Muito Baixo	Baixo	Moderado	Alto	Muito Alto	Muito Alto	Alto	Moderado	Baixo	Muito Baixo
Impacto											

Fonte: Elaborado pelos Autores

4.1.7. Personas

As personas são representações semi fictícias dos usuários reais que ajudam a alinhar as necessidades, objetivos e comportamentos dos potenciais clientes do produto ou serviço desenvolvido. De forma mais profunda, é possível afirmar que a persona é desenvolvida para refinar a comunicação com os usuários e a entender o posicionamento do produto no mercado. Portanto, elas desenvolvem um papel crucial na solução final, permitindo uma visualização mais ampla e clara do público-alvo, suas características demográficas, preferências e desafios, colaborando com um projeto mais eficiente que forneça a melhor experiência para o usuário.



Nome: Davi Coelho Cerqueira
Idade: 21 anos
Formação: Engenharia de Software
Background: Davi está no terceiro ano de Engenharia de Software na UNISANTOS. Ele ainda não conseguiu seu primeiro estágio e está começando a se preocupar com sua preparação para o mercado de trabalho. Davi tem o desejo de desenvolver soft skills que sejam valorizadas nas empresas, a fim de se tornar mais confiante para os processos seletivos.

Objetivo: Ele deseja participar de algum programa que ofereça treinamento de soft skills e liderança, a fim de aumentar suas chances de encontrar uma vaga na área que deseja.

User story -> Eu, Davi, como estudante, quero começar a me preparar melhor para iniciar uma carreira a fim de ser bem colocado no mercado e aumentar minhas chances de sucesso.

Figura 5: Persona Davi



Davi Coelho Cerqueira

Terceiro ano de Engenharia de Software na UNISANTOS. Não conseguiu o primeiro estágio e está preocupado com a carreira. Ele deseja desenvolver suas *softskills* para se sentir mais confiante nos processos seletivos.

Idade, Gênero
21 anos, masculino

Emprego
Estudante

Educação
Engenharia de Software

Local
Santos, SP

Estado civil
Solteiro

Personalidade
Intuitivo
Técnico
Introvertido

Objetivos
Conseguir um emprego
Se comunicar melhor
Aprender a trabalhar em equipe

Necessidades
Aprender novas *softskills*
Conseguir uma vaga de estágio obrigatória

Desafios
Falta de habilidades interpessoais
Dificuldade em se adaptar a diferentes ambientes
Dificuldade em trabalhar sob pressão

Hobbies/ Interesses
Conversar com os amigos online
Ler artigos científicos
Jogar *pickleball* aos sábados



Fonte: Elaborado pelos autores



Nome: Laura Sousa Magalhães

Idade: 26 anos

Formação: Ciências da Computação

Background: Laura concluiu a faculdade há quatro anos e está trabalhando como desenvolvedora web. No entanto, ela está

insatisfeita com a progressão da sua carreira e quer focar no mercado de inteligência artificial e deep learning. Nesse sentido, ela procura por vagas que lhe proporcionem a entrada neste mercado e que se alinhem com as suas expectativas.

Objetivo: Laura deseja encontrar vagas de emprego mais personalizadas na sua área de interesse. A fim de poupar tempo na participação de processos seletivos.

User story → Eu, Laura, como profissional formada em desenvolvimento web, quero encontrar ofertas de emprego em outras áreas de forma mais personalizada e otimizada para poupar tempo nos processos seletivos.

Figura 6: Persona Laura



Laura Sousa Magalhães
Concluiu a faculdade há quatro anos e trabalha como desenvolvedora web. No entanto, está insatisfeita com a carreira e quer ingressar no mercado de inteligência artificial. Ela procura por vagas que lhe permitam entrar nesse mercado.

Idade, Gênero
26 anos, feminino

Emprego
Desenvolvedora Web

Educação
Ciências da Computação

Local
Mauá, SP

Estado civil
Casado

Personalidade
Assertiva
Criativa
Extrovertida

Objetivos
Fazer a transição de carreira
Trabalhar na área de interesse
Conseguir um bom salário

Necessidades
Aprender novas *hardskills*
Se realocar no mercado e ser reconhecida

Desafios
Dificuldade de encontrar vagas que se alinhem com suas expectativas financeiras
Participar de diversos processos seletivos

Hobbies/ Interesses
Correr no parque
Assistir a documentários
Viajar com o marido



Fonte: Elaborado pelos autores



Nome: Eurico Mendes Nunes

Idade: 58 anos


Formação: Gestão de pessoas

Background: Eurico trabalha na empresa Negócios e Cia. como gestor do RH. Seu trabalho é preparar o processo seletivo de entrada na empresa e avaliar os candidatos. No entanto, muitas vezes ele se depara com o desafio de encontrar candidatos que possuam as habilidades necessárias para tais vagas, a fim de diminuir ao máximo o turnover.

Objetivo: Eurico deseja encontrar os candidatos que melhor se adaptem às vagas de emprego disponibilizadas, a fim de poupar recursos no processo seletivo e fazer escolhas mais assertivas.

User story -> Eu, Eurico, como gestor de RH de uma empresa, quero melhorar o processo de recrutamento da empresa que trabalho a fim de encontrar candidatos mais qualificados para as vagas de forma mais otimizada.

Figura 7: Persona Eurico



Eurico Mendes Nunes

Eurico trabalha na empresa Negócios e Cia como *Tech Recruiter*. Seu objetivo é elaborar o processo seletivo de entrada na empresa, assim como avaliar os candidatos que se inscrevem, a fim de encontrar os melhores candidatos para as vagas oferecidas.

Idade, Gênero
58 anos, masculino

Emprego
Tech Recruiter

Educação
Engenharia
Elétrica

Local
Itaquaquecetuba, SP

Estado civil
Divorciado


Personalidade
Observador
Julgador
Disciplinado

Objetivos
Encontrar a melhor relação candidato/vaga
Diminuir o *turn-over*

Necessidades
Contratar candidatos com as habilidades requisitadas
Gastar o menor tempo possível

Desafios
Dificuldade para encontrar candidatos que se alinhem às expectativas propostas pela empresa

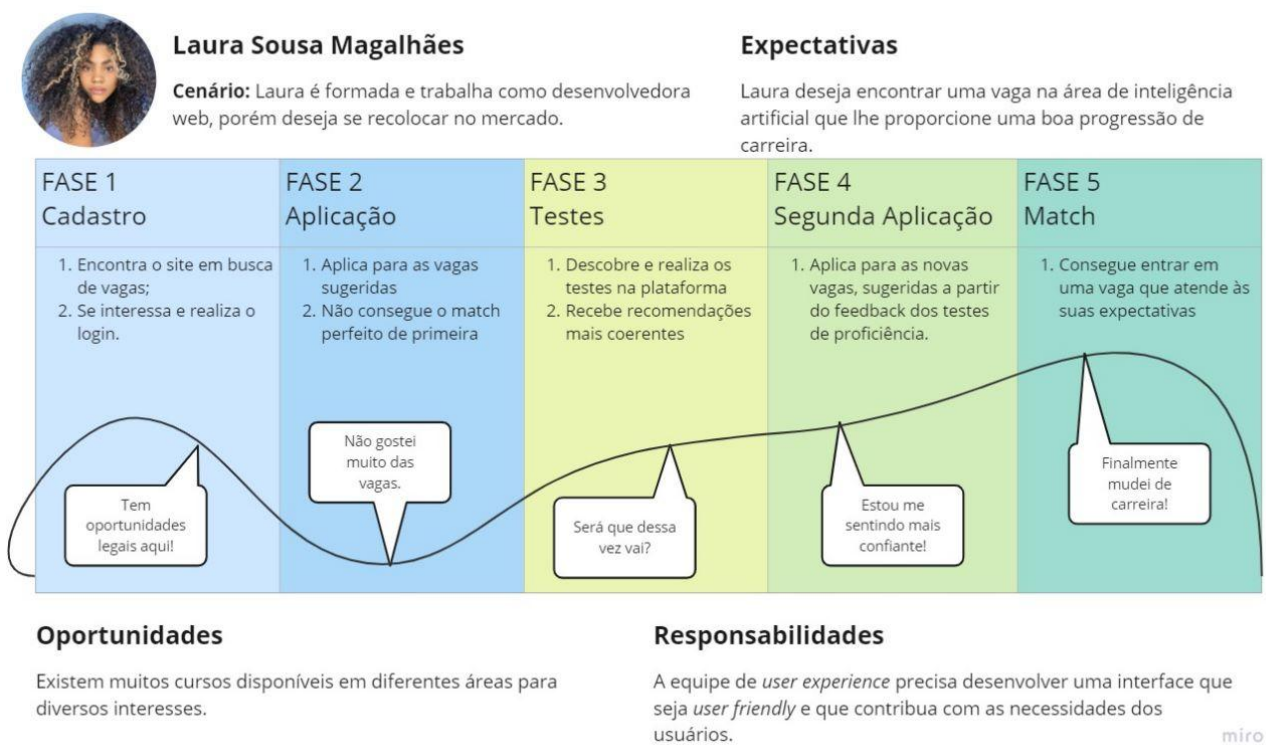
Hobbies/ Interesses
Assistir ao telejornal
Estudar sobre psicologia
Passar um tempo com a neta



4.1.8. Jornadas do Usuário

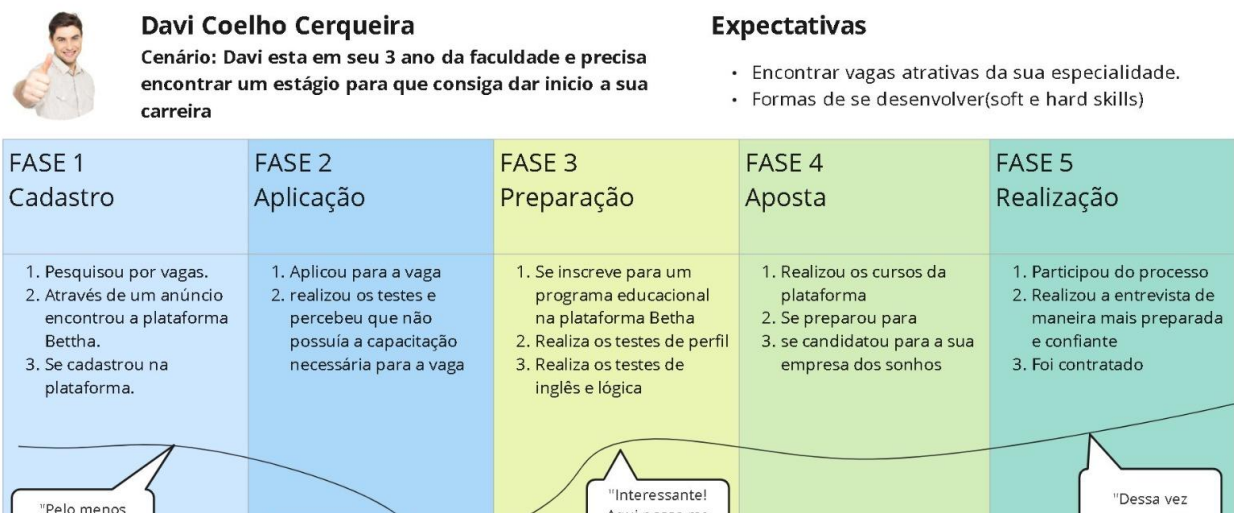
Neste tópico, serão apresentadas as Jornadas dos Usuários, que consiste em uma representação visual de todas as etapas de interação entre o usuário e a plataforma. Com ela, é possível observar em quais momentos os usuários têm mais dificuldades e suas dores. Além disso, ela destaca todas as emoções e ações envolvidas, sendo possível melhorar a aplicação a partir disso. A seguir, são apresentadas as Jornadas dos Usuários das personas apresentadas no subtópico anterior deste documento.

Figura 8: Jornada de Usuário da Laura



Fonte: Miro disponibilizado pelo professor de UX alterado pelos autores

Figura 9: Jornada de Usuário do Davi



4.1.9. Política de Privacidade e LGPD

Essa seção tem como objetivo apresentar a política de privacidade da empresa parceira e sua aplicação no projeto conforme a Lei Geral de Proteção de Dados (LGPD). Na seção 6, ao final deste documento, é possível encontrar a fonte dessas informações.

1. Informações gerais sobre a empresa/organização:

A empresa Bettha atua de forma 100% digital, tendo como público alvo brasileiros, no início de suas carreiras. Ela procura ajudá-los na construção de seus perfis, além de encontrar oportunidades de carreiras que combinam com suas características, oferecendo capacitações para aprimorarem suas habilidades.

2. Informações sobre o tratamento de dados:

O Bettha utiliza os dados adquiridos para fins de:

- (a) envio de convites para participação de processos de recrutamento e seleção conduzidos por quaisquer das empresas do GCT(Grupo Cia de Talentos), que - conforme as características do TITULAR (sendo TITULAR, nesse contexto, o usuário da plataforma) - poderão ser para vagas de aprendiz, estágio, trainee ou emprego;
- (b) realização de processo(s) seletivo(s) para o(s) qual(is) o TITULAR se candidatou à vaga;
- (c) realização de jornada de desenvolvimento profissional;
- (d) envio de convites para participação de pesquisas e eventos relacionados à carreira e mercado; que poderão ser acessados ou não pelo TITULAR;
- (e) envio de links de terceiros, relacionados a conteúdos para aprimoramento profissional, que podem ser cursos, pesquisas, eventos, artigos, dentre outros correlatos, que poderão ser acessados ou não pelo TITULAR;
- (f) envio de e-mail marketing sobre ações e atividades desenvolvidas pelo GCT ou com seu apoio, relacionados à carreira, mercado de trabalho e oportunidades de aprimoramento profissional.
- (g) comunicar, fornecer, manter, e melhorar todo conteúdo disponibilizado nos sites do GCT;
- (h) enviar ao TITULAR avisos técnicos e atualizações;

- (i) responder os comentários, dúvidas e solicitações feitas pelo TITULAR;
- (j) acompanhar e analisar tendências de uso.

3. Quais dados pessoais são coletados (inclusive os dados não informados pelo usuário, como IP, localização, etc)?

Para a coleta dos dados citados abaixo, o usuário deve concordar com o termo de consentimento para tratamento de dados pessoais citado no site da própria empresa “Em atenção ao disposto na Lei Geral de Proteção de Dados, registra-se, através deste 'Termo de Consentimento para Tratamento de Dados Pessoais(Termo)', a permissão livre, consciente e inequívoca dada pelo Usuário dos sites www.ciadetalentos.com.br e www.bettha.com, doravante designado 'TITULAR', para que as empresas integrantes do GRUPO CIA DE TALENTOS (GCT) tratem seus dados pessoais para a(s) finalidade(s) especificada(s) neste Termo.

Ao manifestar seu aceite a este Termo, o TITULAR consente e concorda que as empresas do GCT abaixo identificadas, doravante designadas - individual e em conjunto - 'Controladora', tomem decisões referentes ao tratamento de seus dados pessoais, que podem abranger operações de coleta, produção, recepção, classificação, utilização, acesso, reprodução, transmissão, distribuição, processamento, arquivamento, armazenamento, eliminação, avaliação ou controle da informação, modificação, comunicação, transferência, difusão ou extração”

Dados coletados pela empresa que são informados diretamente pelos usuários:

- Nome completo;
- Número de Cadastro de Pessoa Física ('CPF');
- Endereço de e-mail;
- Número de telefone celular;
- Grau de escolaridade;
- Áreas de interesse;
- Conhecimento de idiomas;
- Cidade, estado e país do usuário.

Dados coletados dos usuários por meio da utilização da plataforma:

- Informações sobre as interações do Usuário com a plataforma, incluindo a data e hora de quaisquer ações, como atualizações cadastrais. Pode igualmente incluir

informações sobre a utilização de aplicações de terceiros e publicidade que receber, tais como os LOGS dos sistemas;

- Conteúdos inseridos na plataforma pelo Usuário, incluindo as mensagens que enviar ou receber através da plataforma, as informações fornecidas nas inscrições para processos seletivos, além de suas interações com a Equipe de Apoio do GCT;
- Dados técnicos, que poderão incluir informações de URL, dados de cookies, o endereço IP do Usuário, os tipos de dispositivos por este utilizados para acessar a Plataforma, identificações exclusivas do dispositivo, atributos do dispositivo, tipo de ligação à rede (por exemplo, Wi-Fi, 3G, LTE, Bluetooth) e fornecedor de rede, desempenho da rede e do dispositivo, tipo de navegador, idioma e versão da Plataforma;
- Dados anonimizados .

Dados opcionais que o usuário pode ou não fornecer:

- As fotografias e vídeos do Usuário;
- A origem racial ou étnica;
- A orientação sexual;
- Pessoa com Deficiência;
- A localização precisa do dispositivo do Usuário;
- Microfone do dispositivo.

Os dados opcionais somente serão coletados com autorização prévia do usuário, como fotografias, vídeos, orientação sexual, dados referentes à saúde, localização ou dados de voz. O usuário poderá sempre mudar de ideia e retirar o seu consentimento a qualquer momento. Estes dados têm o intuito de fornecer-lhe funcionalidades, informações, publicidade ou outro conteúdo baseado em dados específicos do usuário.

4. Onde os dados são coletados (fonte)?

Os dados de usuários são coletados através do cadastro feito pelos usuários na plataforma, ou por meio de serviços de terceiros, como, por exemplo, o LinkedIn, Gmail ou Facebook, para efetuar seu cadastro/criar uma conta.

Além disso, os dados sobre a utilização da plataforma são coletados quando o usuário está utilizando a plataforma adotada pelo GCT (“Grupo Cia de Talentos”).

5. Para quais finalidades os dados são utilizados?

Os dados são utilizados com a finalidade de melhorar a experiência do usuário na plataforma e poder fornecer conteúdo personalizado relacionado aos temas da carreira, funcionalidades, informações e publicidades. Também, com esses dados, pode-se facilitar na criação de seu currículo e guardar registros dentro da plataforma, além de ajudar no sistema de recomendação de cursos e vagas para o usuário.

6. Onde os dados ficam armazenados?

Os dados ficam armazenados em um banco de dados que pode ser acessado por todas as empresas do GCT.

7. Qual o período de armazenamento dos dados (retenção)?

Os dados serão retidos pelo GCT durante o período exigido pela legislação aplicada.

8. Uso de cookies e/ou tecnologias semelhantes?

A empresa Bettha usa cookies para aprender como o usuário interage com o conteúdo da plataforma e para melhorar a experiência dele ao visitar o website. Por exemplo, alguns cookies lembram o idioma ou preferências para que o usuário não tenha que efetuar estas escolhas repetidamente sempre que visitar o website. Além disso, os cookies permitem que eles ofereçam conteúdos específicos, tais como vídeos. Ademais, o Bettha pode empregar o que aprende sobre o comportamento dos usuários para oferecer anúncios direcionados em website(s) de terceiros em um esforço para “reapresentar-lhes” os seus produtos e serviços.

Quais os tipos de Cookies que o Bettha utiliza?

- Cookies Primários e Cookies de Terceiros: Utilizam tanto cookies primários quanto cookies de terceiros no website.
- Cookies de sessão: Cookies da sessão são cookies temporários utilizados para lembrar de você durante o curso da sua visita ao website, e eles perdem a validade quando você fecha o navegador.
- Cookies Persistentes: Cookies persistentes são utilizados para lembrar suas preferências do website e permanecem no seu desktop ou dispositivo móvel

mesmo após você fechar o seu navegador, ou efetuar uma reinicialização. O Bettha os usa para analisar o comportamento do usuário e estabelecer padrões, de modo que a empresa possa melhorar a funcionalidade do website para os usuários. Estes cookies também permitem o oferecimento de anúncios segmentados e medir a eficácia do site e a funcionalidade de tais anúncios.

Como os Cookies são usados para o propósito de anúncios?

- Cookies e tecnologias de anúncios tais como 'web beacons', 'pixels' e 'tags' de redes de anúncios ajudam a oferecer anúncios relevantes de forma mais eficaz. Eles também ajudam a coletar dados consolidados para fins de auditorias, pesquisas e relatórios de desempenho para anunciantes. Os pixels permitem a compreensão e a melhora da oferta de anúncios para o usuário, e permitem ainda saber quando determinados anúncios já lhe foram apresentados. Como o navegador pode requisitar anúncios e 'web beacons' diretamente de servidores de rede de anúncios, estas redes podem visualizar, editar ou configurar seus próprios cookies, como se você tivesse acessado uma página web do site deles.

9. Com quem esses dados são compartilhados (parceiros, fornecedores, subcontratados)?

O usuário, ao se candidatar a vaga ofertada em um processo de recrutamento e seleção conduzido pelo Bettha, os seus dados e informações, no todo ou em parte, conforme o caso, serão disponibilizados e compartilhados com a empresa cliente, contratante do processo seletivo daquela vaga.

Os dados pessoais serão tratados pelo Bettha e, sempre que necessário, por ele compartilhados com terceiros, para realização das trilhas de desenvolvimento profissional disponibilizadas no site, para realização de processo seletivo do qual o usuário participe, para envio de convites para participar de pesquisas, eventos, cursos, conteúdos, e-mail marketing sobre o Bettha e sobre o GCT, dentre outros assuntos relacionados à carreira e mercado de trabalho e à comunidade GCT.

10. Informações sobre medidas de segurança adotadas pela empresa

A empresa Bettha implementa medidas técnicas e organizacionais para proteger a segurança dos dados pessoais do usuário. No entanto, o Bettha pede ciência do usuário de que nenhum sistema é completamente seguro. Ele também implementa várias

políticas, entre elas, anonimização, políticas de acesso e de retenção para prevenir contra o acesso não autorizado e a retenção desnecessária de dados pessoais nos nossos sistemas.

A senha protege a conta de usuário, por isso eles o incentivam a utilizar uma senha única e forte, a limitar o acesso ao computador e navegador e a encerrar a sessão após utilizar a Plataforma.

11. Orientações sobre como a empresa/organização atende aos direitos dos usuários:

A Lei Geral de Proteção de Dados (LGPD), Lei n.º 13.709/18, concede aos indivíduos certos direitos relacionados aos seus dados pessoais. Estes direitos incluem:

1. Confirmação da existência de tratamento de dados;
2. Acesso aos próprios dados;
3. Correção de dados pessoais incompletos, inexatos ou desatualizados;
4. Anonimização, bloqueio ou eliminação de dados desnecessários, excessivos ou tratados de forma não conforme com a LGPD;
5. Portabilidade dos dados para outro fornecedor de serviço ou produto, protegendo segredos comerciais e industriais;
6. Eliminação de dados pessoais tratados com consentimento, exceto em casos previstos na LGPD;
7. Informação sobre compartilhamento de dados com entidades públicas e privadas;
8. Conhecimento sobre a possibilidade de não conceder consentimento e suas consequências;
9. Revogação do consentimento, conforme estabelecido na LGPD.

Caso deseje exercer esses direitos ou entrar em contato, o Bettha oferece os seguintes e-mails:

- a) Para cadastros no site www.ciadetalentos.com.br: meajuda@grupociadetalentos.com.br.
- b) Para cadastros no site <https://www.bettha.com>: contato@bettha.com.

Se o usuário receber mensagens de publicidade eletrônica, ele pode retirar o consentimento ou recusar a qualquer momento sem custos adicionais. As mensagens de

publicidade eletrônica enviadas pelo GCT contém um mecanismo de recusa, como um link para cancelamento de inscrição nos e-mails enviados.

12. Informações sobre como o titular de dados pode solicitar e exercer os seus direitos.

Caso o titular de dados queira exercer o seu direito, o Bettha libera alguns dos direitos:

- Direito de confirmação da existência de tratamento;
- Direito de acesso aos dados;
- Direito de correção de dados incompletos, inexatos ou desatualizados;
- Direito à anonimização, bloqueio ou eliminação de dados desnecessários, excessivos ou tratados em desconformidade com a LGPD;
- Direito à portabilidade dos dados a outro fornecedor de serviço ou produto, resguardados os segredos comerciais e industriais do GCT;
- Direito à eliminação dos dados pessoais tratados com o consentimento do titular, exceto nas hipóteses de guarda legal e outras dispostas na LGPD;
- Direito à informação das entidades públicas e privadas com as quais o GCT realizou o uso compartilhado de dados;
- Direito à informação sobre a possibilidade de não fornecer o consentimento e sobre as consequências da negativa;
- Direito à revogação do consentimento, nos termos da LGPD.

13. Informações de contato do Encarregado de Proteção de Dados (DPO - Data Protection Officer) da organização.

Caso o usuário queira informações de contato do Encarregado de Proteção de Dados, o termo de uso disponibiliza no fim dele o acesso ao e-mail e ao contato das empresas e caso queira entrar em contato com a Bettha a informação de contato estará disponível no próprio termo também.

4.2. Compreensão dos Dados

Esta seção abordará toda a parte de descrição dos dados, bem como uma análise estatística das colunas das tabelas utilizadas pelo grupo Mettha e, por fim, a formulação de 3 hipóteses pelo grupo como resultado da exploração dos dados realizada.

4.2.1. Descrição dos dados

Os principais dados a serem utilizados são aqueles referentes aos testes de *superfit*, *lifestyle* e *workstyle*, os quais são avaliações disponíveis no site da Bettha, de onde os dados foram adquiridos. Os atributos avaliados e o público-alvo são diferentes em cada teste, sendo o *superfit* voltado tanto para os candidatos quanto para os gestores, e apresenta as seguintes competências: consistente, pragmático, original, colaborativo, engajado e resiliente. Além disso, são atribuídas notas para cada um deles em uma escala numérica de 0 a 90.

Já no que diz respeito ao *lifestyle*, que é voltado somente para os candidatos, e ao *workstyle*, focado somente nos gestores, as características são divididas em dois grupos. O primeiro contém os atributos: *classic*, *order*, *change*, *tireless* e *explorer*, avaliados de 0 a 9, enquanto o segundo contém o *specialist* e *generalist*, mensurados de 0 a 5. Os atributos citados dizem respeito sobre algum aspecto da personalidade ou do estilo de trabalho de quem respondeu a pesquisa, sendo uma forma de avaliar o perfil comportamental dos candidatos e possibilitar um melhor entendimento deles.

Além desses testes, há também as avaliações de proficiência em inglês e em excel, assim como um teste de lógica. Estes serão utilizados para filtrar as vagas que requerem tais habilidades mais específicas, e não como critérios principais. Todos esses dados serão mesclados por meio da união das tabelas utilizando-se dos *IDs* dos usuários, uma vez que a biblioteca Pandas oferece os recursos para tal operação, tornando essa mesclagem simples.

Outro ponto relevante é que os dados fornecidos são confiáveis, dado que são colhidos por meio das pesquisas disponíveis no site da Bettha. As informações não dependem da entrada do usuário, mas sim da seleção de opções pré-definidas. Dessa forma, é baixa a chance de ocorrer ruídos nos dados, como erros de digitação, sendo estes limitados a possíveis, mas improváveis, erros no código da plataforma. Mas no escopo dos dados disponibilizados, nenhum foi encontrado.

Nesse sentido, existem dados faltantes, já que nem todos os testes são obrigatórios. Nesse caso, eles serão tratados devidamente, sendo substituídos, provavelmente, pela média dos outros dados ou pela sua mediana, para que não haja a perda de uma grande quantidade de informações. Em relação aos dados com valores extremos (*outliers*), por serem naturais dos testes, já que haverão indivíduos com performances acima da média, não serão retirados.

No que tange a diversidade dos dados, não temos conhecimento das informações demográficas daqueles que responderam os testes. Entretanto, é possível inferir que a sua maioria esteja concentrada na região sudeste, já que essa área é o principal foco das vagas oferecidas na plataforma da Bettha. Nessa lógica, os dados serão utilizados em sua totalidade para garantir uma maior precisão do modelo preditivo. Além disso, não há necessidade da seleção de uma amostra a partir da população, já que o Python, aliado ao Pandas e as outras bibliotecas envolvidas, oferecem a potência necessária para analisar por completo as tabelas.

Outro fator que vale a pena ser ressaltado, é que não há nenhum dado sensível descrito nas tabelas fornecidas pelo Bettha e, portanto, não há nenhum risco de segurança atrelado ou de ferir as diretrizes da Lei Geral de Proteção de Dados (LGPD). Além disso, por não haver a inclusão de dados sensíveis, torna-se improvável o enviesamento do modelo preditivo. Entretanto, ainda é necessário ter cuidado com o tratamento dos dados, já que muitas vezes, mesmo sem esses dados presentes, é possível haver uma relação implícita entre os resultados dos testes e certas classes sociais, por exemplo, que podem causar um viés no algoritmo.

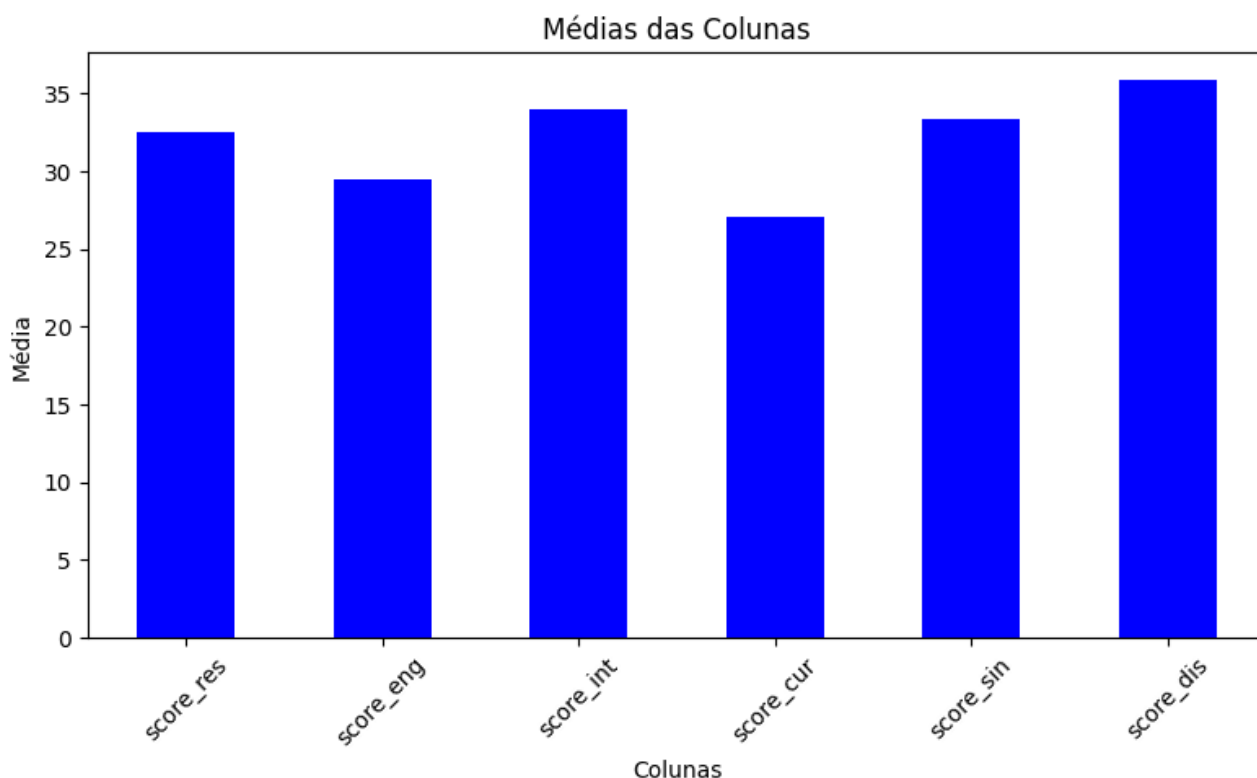
4.2.2. Descrição estatística básica dos dados

As variáveis numéricas presentes na base de dados da Bettha são aquelas referentes aos testes de *lifestyle*, *workstyle*, *superfit*, inglês, excel e lógica, que compreendem os atributos citados no subtópico 4.2.1, assim como os IDs dos usuários. Há também as variáveis que constam as notas das jornadas de capacitação na plataforma, assim como a contagem de avaliações. Já as variáveis categóricas cobrem a classificação dos testes de inglês, excel e lógica, que são atribuídas de acordo com a pontuação do usuário nessas avaliações por uma letra ou palavra que a representa. Nas

tabelas dos testes, há também o registro de quando a avaliação foi iniciada pelo candidato.

Para melhor entendimento das tabelas e dos dados, alguns gráficos foram criados para possibilitar a sua fácil e eficiente visualização, que são apresentados a seguir.

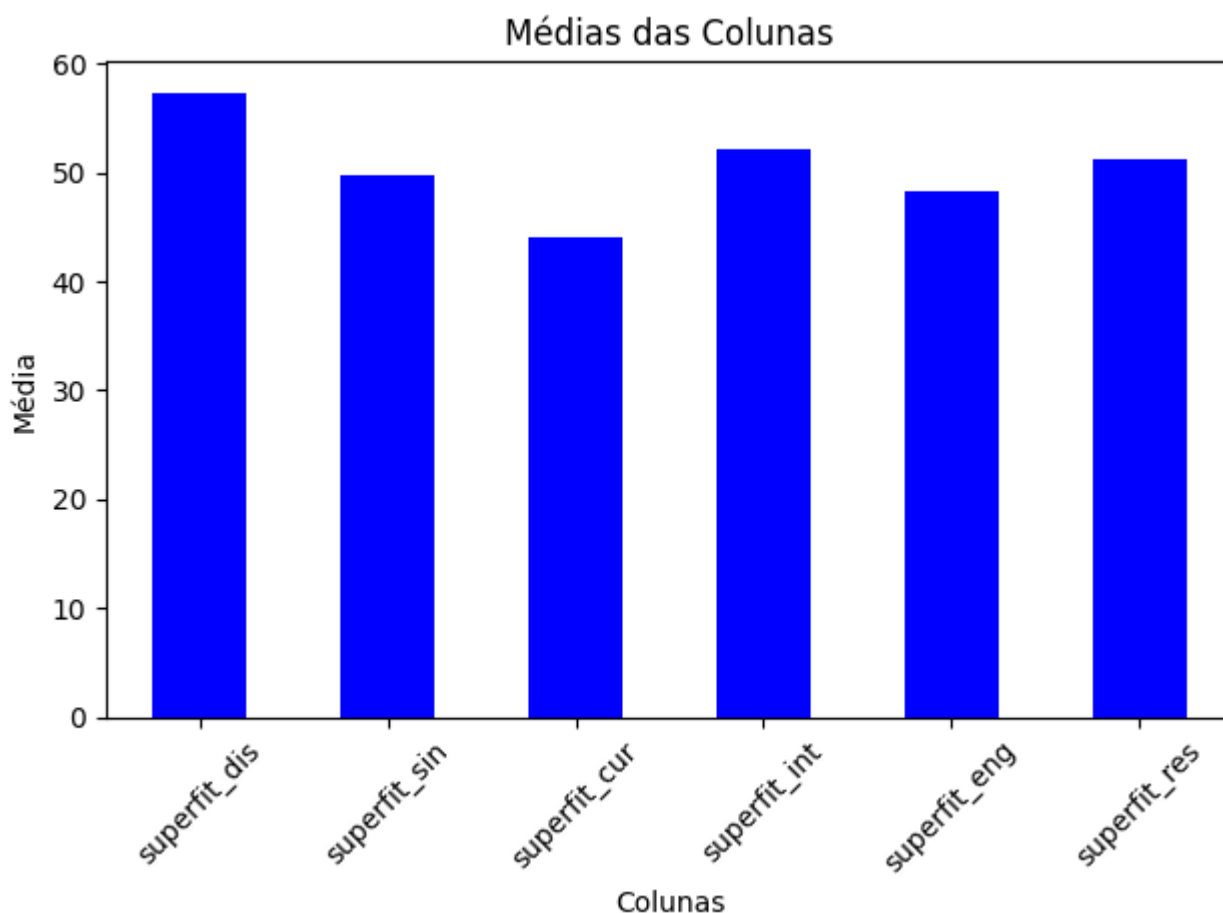
Figura 10: Média aritmética dos resultados do teste de *superfit* dos gestores.



Fonte: Colab notebook elaborado pelos autores

As variáveis presentes no eixo horizontal dizem respeito, respectivamente, às competências: resiliente, engajado, colaborativo, original, pragmático e consistente. Por meio desse gráfico, é possível compreender as características que mais aparecem nas vagas. Nesse caso, elas estão bem distribuídas entre as competências, o que nos leva a concluir que todos os usuários serão contemplados com uma vaga que se enquadra com a sua personalidade.

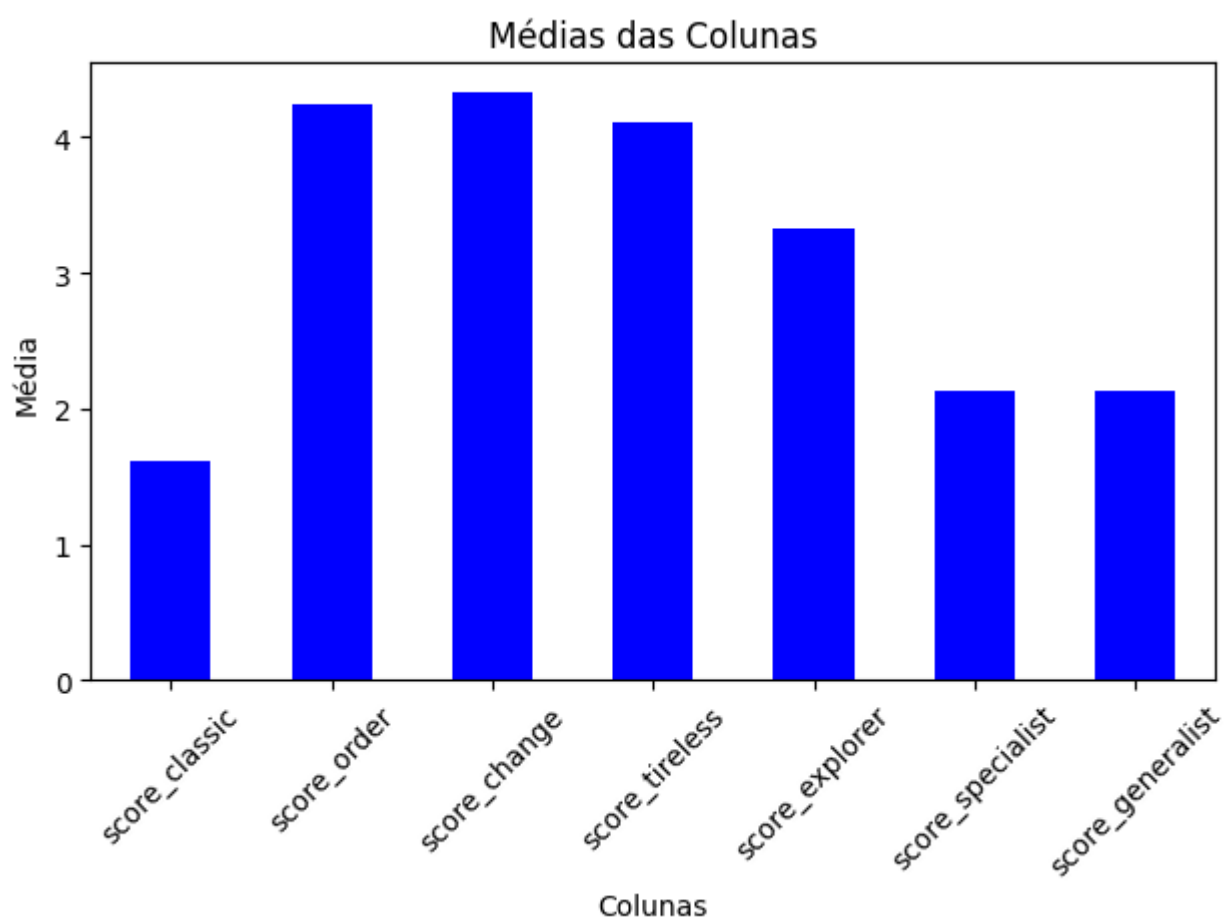
Figura 11: Média aritmética dos resultados do teste de *superfit* dos usuários.



Fonte: Colab notebook elaborado pelos autores

As variáveis do eixo horizontal correspondem, respectivamente, aos atributos: consistente, pragmático, original, colaborativo, engajado e resiliente. Com esse gráfico, é possível relacionar os dados dos testes de *superfit* tanto dos gestores quanto dos usuários. Assim, é possível concluir que há uma certa proximidade entre as proporções dos valores, por mais que os valores absolutos sejam diferentes. Portanto, haverá, provavelmente, uma harmonia entre as personalidades dos usuários e os perfis requisitados pelas vagas.

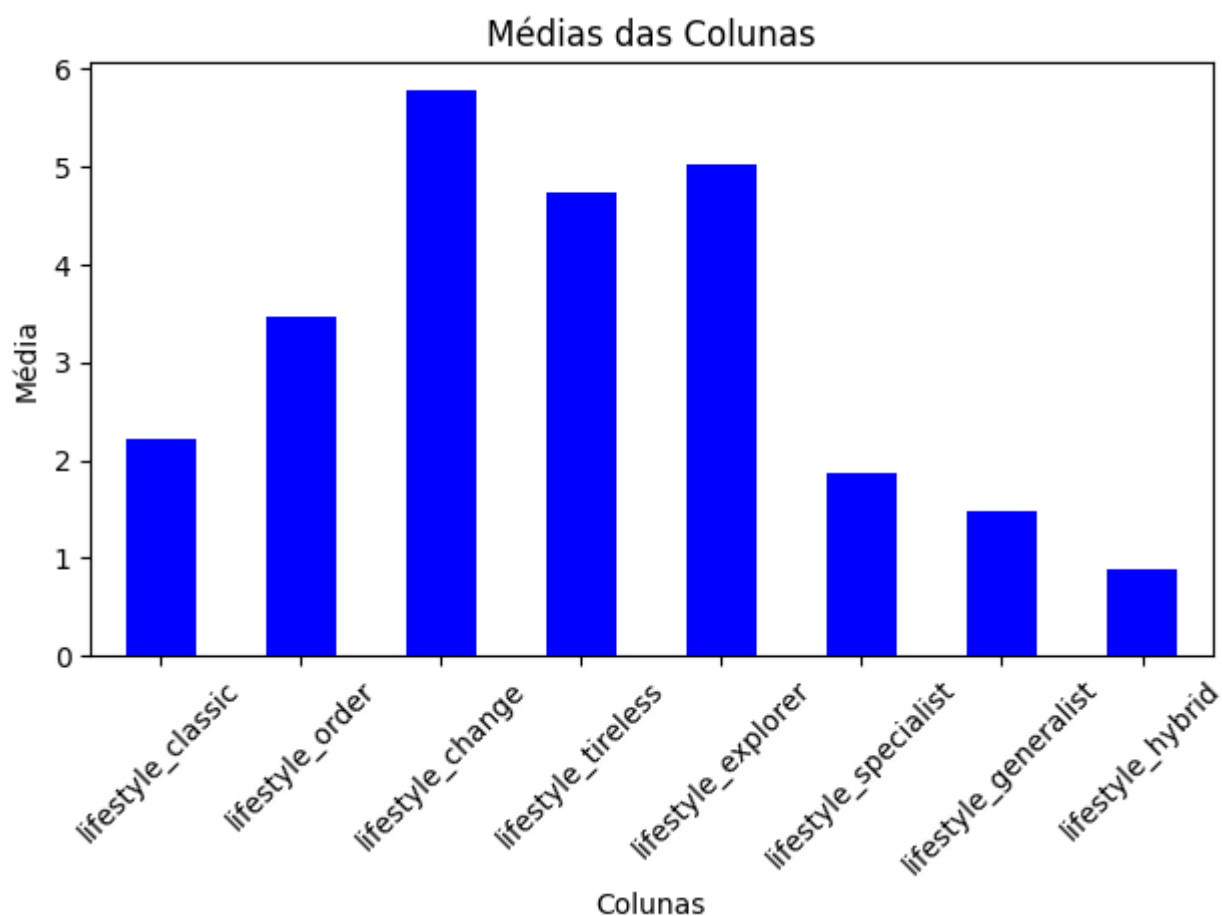
Figura 12: Média aritmética dos resultados do teste de *lifestyle* dos gestores.



Fonte: Colab notebook elaborado pelos autores

As variáveis do eixo horizontal dizem respeito a cada característica do teste de lifestyle, descritos no tópico 4.2.1. Por meio da sua análise, é possível concluir que há uma maior tendência dos atributos *order*, *change* e *tireless*, sendo eles os mais proeminentes nas vagas.

Figura 13: Média aritmética dos resultados do teste de *lifestyle* dos candidatos às vagas.



Fonte: Colab notebook elaborado pelos autores

Assim como no gráfico 3, as variáveis no eixo horizontal são descritas no tópico 4.2.1. Por meio da interligação entre os gráficos 3 e 4, é possível perceber que, por mais que os valores absolutos destoem entre si, há uma clara proporção entre eles. Logo, é possível concluir que há uma harmonia entre os resultados dos testes de *lifestyle* dos

candidatos e workstyle dos gestores, assim como foi observado nos testes de superfit supracitados.

4.2.3. Descrição da predição desejada

Para alcançarmos o modelo mais preciso possível, serão testadas 3 hipóteses distintas. Após sua modelagem, serão avaliadas a acurácia de cada uma, ou seja, o quanto o modelo é capaz de acertar as vagas ideais para cada candidato, e a melhor delas será escolhida para ser implementada.

1º modelo: a primeira abordagem será por meio da união de dois algoritmos de natureza binária, conhecidos como K-Means e KNN. O primeiro se baseia no agrupamento de dados por meio de características em comum. Ele reúne os dados que mais se assemelham entre si em volta de um dado central (cluster), e será utilizado para separar as informações em grupos. Já o segundo consiste na categorização de um dado desconhecido de acordo com os seus vizinhos mais próximos, ou seja, aqueles cuja diferença dos valores seja a menor possível. Dessa forma, aliando esses dois modelos, somos capazes de enquadrar as vagas que mais se encaixam com cada usuário.

2º modelo: a terceira e última abordagem será por meio do algoritmo de árvore de decisão, o qual se baseia na criação de nós, em que apresentam condições para a filtragem das vagas ideais para certo usuário. Por exemplo, em um nó, pode conter a condição de um resultado no teste ser maior do que 6. Em outro, poderia haver a condição de o setor ser da área de tecnologia. Por meio da utilização de nós como esses, é possível fazer uma seleção com base em critérios objetivos e transparentes, de forma a chegar ao melhor resultado possível.

3º modelo: a segunda abordagem será por meio do algoritmo de Rede Bayesiana, a qual se baseia na interligação de probabilidades condicionais, utilizando-se de eventos passados conhecidos para calcular a chance de um evento em questão acontecer. Assim, esse modelo pode ser utilizado para aferir a probabilidade de uma vaga se encaixar com o perfil de certo candidato com base em variáveis como os requisitos de habilidade e seu setor de atuação, os quais são disponibilizados pela base de dados da Bettha.

4.3. Preparação dos Dados

Nesta seção, será abordado o tipo de aprendizado de máquina utilizado, além de todo o processo de tratamento dos dados, incluindo todas as manipulações neles feitas como agregações e derivações de atributos, tratamento de dados nulos e as features selecionadas para a elaboração do modelo.

4.3.1. Aprendizado de máquina

O aprendizado não supervisionado ocorre quando se tem um conjunto de dados desconhecidos, ou seja, não há rótulos (labels) ou um alvo (target) a ser previsto. Nesse contexto, o objetivo é encontrar estruturas ou padrões nos dados, identificando semelhanças e agrupando-os de acordo com essas semelhanças. Nesse projeto, o algoritmo K-Means é utilizado para identificar grupos de semelhanças entre gestores e usuários mediante uma técnica de clusterização que agrupa dados em clusters com base na proximidade entre eles.

É importante notar que, mesmo em projetos de aprendizado não supervisionado, a análise e a compreensão dos resultados muitas vezes podem ser aprimoradas com a introdução de elementos do aprendizado supervisionado. O uso do algoritmo K-Nearest Neighbors (KNN) é um exemplo disso. Ele é utilizado para calcular a distância entre gestores e usuários. No contexto do KNN, a distância é uma métrica de semelhança, ou seja, quanto mais próximos estão os pontos de dados, mais similares eles são, permitindo identificar gestores que estejam mais próximos em termos de perfil ou preferências.

Portanto, o projeto é fundamentado no aprendizado não supervisionado, no qual o K-Means será utilizado para criar clusters de gestores e usuários com base em suas características semelhantes, enquanto o KNN será empregado para calcular a distância entre eles, ajudando assim na recomendação de vagas para os usuários com base na proximidade de seus perfis.

4.3.2. Descrição das manipulações realizadas

Durante o período de preparação dos dados, foi necessário realizar alguns processos de manipulação, a fim de alinhar as informações às expectativas e necessidades do projeto. Nesse sentido, conta-se com a Limpeza e Transformação de

Dados, assim como processos de agrupamento. Para prosseguir com a limpeza dos dados, é importante tratar os valores nulos, preenchendo ou removendo-os. Também são eliminadas as duplicatas e lida-se com os outliers conforme as necessidades.

Ademais, se inicia o processo de transformação dos dados. Ela contou com os seguintes processos: padronização dos dados, a fim de colocá-los na mesma escala; criação de novas features baseadas nas já existentes, de acordo com as necessidades surgidas durante o processo; e conversão de variáveis categóricas em numéricas, a fim de facilitar o treinamento do algoritmo. Por fim, realizam-se processos de agrupamento de variáveis a partir de técnicas de clusterização, com a finalidade de identificar padrões.

4.3.3. Agregação e derivação de atributos

A combinação de registros e a criação de novos atributos devem seguir os princípios da engenharia de características e do modelo de dados nítidos, a fim de preparar os dados de maneira adequada e torná-los mais compreensíveis para o algoritmo. Mais especificamente, a combinação de registros envolve a fusão de dados de outras tabelas para verificar uma característica, enquanto a criação de novos atributos se refere à geração de recursos adicionais para análise, como, por exemplo, unir duas colunas de uma tabela para criar uma nova coluna com dados derivados das duas últimas.

4.3.4. Tratamento de valores nulos

Para eliminar ou substituir valores faltantes, ou em branco, é crucial aplicar técnicas de engenharia de características. Isso permite uma limpeza eficaz desses valores ausentes. Esses valores em branco ou ausentes geralmente correspondem a recursos relevantes para o modelo preditivo. Assim, a biblioteca Pandas pode ser uma ferramenta útil para facilitar o tratamento dessas situações.

Este processo também envolve a detecção e remoção de outliers, garantindo que os dados restantes sejam mais precisos e representativos da realidade. Por exemplo, ao analisar uma tabela de cursos, pode haver entregadores que, de acordo com os dados, não concluíram nenhum curso. Nesses casos, é fundamental remover esses usuários do conjunto de dados para treinar o algoritmo com dados limpos e obter previsões mais precisas. A presença desses valores atípicos poderia afetar a precisão das previsões.

Para realizar essa tarefa, é possível usar estruturas de repetição, como um "for," com instruções condicionais "if/else." Isso permitirá identificar os usuários que não têm cursos concluídos. Com base nessa informação, é possível criar uma tabela contendo esses usuários e combiná-la com a tabela que contém todos os registros, garantindo assim a limpeza adequada dos dados.

4.3.5. Features selecionadas

A seguir, são apresentadas as features selecionadas pelo grupo para a elaboração do modelo.

Id: A feature "id" representa o id do usuário ou da vaga no sistema e todas as outras informações estão ligadas a ela.

score_Resiliente: O termo "score_Resiliente" refere-se a uma métrica relacionada à avaliação da resiliência de um usuário ou gestor no contexto profissional. Essa pontuação é obtida por meio do processo de mapeamento Superfit. A resiliência, neste contexto, se refere à capacidade de uma pessoa se recuperar de adversidades, superar desafios e lidar eficazmente com situações difíceis. Para os candidatos, essa pontuação pode ser usada como um indicador de sua capacidade de enfrentar desafios e adaptar-se a situações adversas. Para os gestores, essa pontuação pode fornecer informações valiosas sobre suas habilidades de liderança, como a capacidade de motivar e orientar suas equipes.

score_Engajado: O termo "score_Engajado" refere-se a uma métrica relacionada à avaliação do engajamento de um usuário ou gestor no contexto profissional. Essa pontuação é obtida por meio do processo de mapeamento Superfit. O engajamento aqui se refere à disposição, envolvimento e entusiasmo de um indivíduo em relação a suas responsabilidades, tarefas ou funções em um ambiente profissional. Colaboradores engajados tendem a ser mais produtivos, criativos e comprometidos com os objetivos da empresa. Portanto, ao avaliar o engajamento de um gestor ou candidato, a organização pode obter informações valiosas sobre suas competências em relação a aspectos como liderança, motivação da equipe e contribuição para o ambiente de trabalho.

score_Colaborativo: O termo "score_Colaborativo" refere-se a uma métrica relacionada à avaliação da colaboração de um usuário ou gestor no contexto profissional. Essa pontuação é obtida por meio do processo de mapeamento Superfit. A colaboração diz respeito à habilidade de uma pessoa trabalhar efetivamente com outras pessoas,

compartilhando conhecimento, ideias e recursos para alcançar objetivos comuns. Colaboradores que são eficazes na colaboração geralmente contribuem para um ambiente de trabalho mais produtivo e fortalecem a coesão da equipe. Portanto, ao avaliar a capacidade de colaboração de um gestor ou candidato, a organização pode obter informações valiosas sobre suas competências relacionadas à interação social, trabalho em equipe e habilidades de comunicação.

score_Original: O termo "score_Original" refere-se a uma métrica relacionada à avaliação da originalidade de um usuário ou gestor no contexto profissional. Essa pontuação é obtida por meio do processo de mapeamento Superfit. A originalidade se refere à capacidade de uma pessoa criar, inovar e pensar de maneira única, oferecendo soluções, ideias ou abordagens distintas em seu trabalho. Colaboradores que demonstram originalidade podem contribuir para a diferenciação da empresa, a resolução de problemas de maneira única e a geração de novas oportunidades de negócios. Essas habilidades podem ser particularmente relevantes em cargos de liderança e em setores que dependem da constante evolução e adaptação.

score_Pragmático: O termo "score_Pragmático" refere-se a uma métrica relacionada à capacidade de um usuário ou gestor adotar uma abordagem prática e orientada para resultados em um contexto profissional. Essa pontuação é obtida por meio do processo de mapeamento Superfit. Colaboradores pragmáticos tendem a ser eficazes na execução de tarefas, na resolução de problemas e no cumprimento de metas e prazos.

score_Consistente: O termo "score_Consistente" refere-se a uma métrica relacionada à avaliação da consistência de um usuário ou gestor no contexto profissional. Essa pontuação é obtida por meio do processo de mapeamento Superfit. A consistência refere-se à capacidade de um usuário ou gestor manter um desempenho estável e uniforme em um contexto profissional. Colaboradores consistentes tendem a ser confiáveis, capazes de manter padrões de qualidade e cumprir compromissos, o que é essencial para o sucesso a longo prazo de uma equipe ou organização.

score_classic: O termo "score_classic" refere-se a uma métrica relacionada à avaliação do comportamento de um usuário ou gestor no contexto profissional. Essa pontuação é obtida por meio do Teste de Lifestyle/Workstyle. Esse score é uma medida quantitativa que reflete a avaliação do grau de conformidade do candidato com comportamentos considerados mais convencionais em um ambiente profissional.

Candidatos que se encaixam bem na cultura organizacional e adotam comportamentos tradicionais podem contribuir para um ambiente de trabalho mais harmonioso e produtivo.

score_order: O termo "score_order" refere-se a uma métrica relacionada à avaliação do comportamento de um usuário ou gestor no contexto profissional. Essa pontuação é obtida por meio do Teste de Lifestyle/Work Style. Esse score é uma medida quantitativa que reflete a avaliação do grau de importância que um candidato atribui à ordem e à organização em suas atividades profissionais. Candidatos que valorizam e demonstram habilidades de organização podem contribuir para um ambiente de trabalho mais eficaz e livre de desordem.

score_change: O termo "score_change" refere-se a uma métrica relacionada à avaliação do comportamento de um usuário ou gestor no contexto profissional. Essa pontuação é obtida por meio do Teste de Lifestyle/Work Style. Essa métrica tem como objetivo medir a disposição de um candidato para lidar e adaptar-se a mudanças em seu ambiente de trabalho. Candidatos que demonstram ser abertos e capazes de se ajustar a mudanças podem contribuir para a resiliência e o sucesso da organização.

score_tireless: O termo "score_tireless" refere-se a uma métrica relacionada à avaliação do comportamento de um usuário ou gestor no contexto profissional. Essa pontuação é obtida por meio do Teste de Lifestyle/Work Style. Essa métrica tem como objetivo medir a disposição e a energia de um candidato para trabalhar de forma incansável e persistente em suas atividades profissionais. Candidatos que demonstram ser incansáveis podem ser fundamentais para atingir metas e objetivos organizacionais.

score_explorer: O termo "score_explorer" refere-se a uma métrica relacionada à avaliação do comportamento de um usuário ou gestor no contexto profissional. Essa pontuação é obtida por meio do Teste de Lifestyle/Work Style. Essa métrica tem como objetivo medir a disposição de um candidato para ser um explorador ou um buscador de novas ideias, abordagens e oportunidades em seu ambiente de trabalho. Isso pode ser particularmente relevante para cargos que envolvem liderança, empreendedorismo ou desenvolvimento de estratégias de negócios.

score_specialist: O termo "score_specialist" refere-se a uma métrica relacionada à avaliação do comportamento de um usuário ou gestor no contexto profissional. Essa pontuação é obtida por meio do Teste de Lifestyle/Work Style. Essa métrica pode ser

usada para avaliar a adequação de um candidato para uma função que valoriza a profundidade de conhecimento em uma área específica.

score_generalist: O termo "score_generalist" refere-se a uma métrica relacionada à avaliação do comportamento de um usuário ou gestor no contexto profissional. Essa pontuação é obtida por meio do Teste de Lifestyle/Work Style. Essa métrica é usada para avaliar a adequação de um candidato para uma função que não requer conhecimento especializado, mas sim uma ampla gama de habilidades e adaptabilidade.

mean_genius: O termo "mean_genius" refere-se a uma métrica que representa a média dos resultados obtidos no mapeamento da Superfit. Essa média pode ser usada para entender o desempenho global de um usuário, candidato ou gestor em várias competências ou áreas específicas.

mean_life: O termo "mean life" refere-se a uma métrica que representa a média dos resultados obtidos no mapeamento de Lifestyle/Work Style. Essa média pode ser útil para entender como a pessoa aborda o equilíbrio entre trabalho e vida pessoal, sua preferência por determinados estilos de trabalho e seu ajuste geral em relação à cultura e às demandas de uma organização.

media_geral: O termo "media_geral" refere-se a uma métrica que representa a média aritmética ponderada das métricas "mean_genius" e "mean_life", com pesos diferentes atribuídos a cada uma delas. O "mean_genius" recebe um peso de 2, enquanto o "mean_life" recebe um peso de 1.

type_instance: Como as informações de “usuário” e “gestor” estavam em tabelas diferentes, foi necessário criar uma variável que identifica a instância das informações. Ou seja, se ela se refere ao “usuário” ou ao “gestor”.

grupo: O termo "grupo" refere-se à divisão dos dados de uma tabela em grupos distintos por meio do processo de K-means. O K-means é um algoritmo de aprendizado de máquina não supervisionado que é frequentemente usado para agrupar dados em clusters ou grupos com base em suas características semelhantes. A importância da criação de grupos usando o K-means reside no fato de que isso permite a organização e a interpretação dos dados de uma forma mais significativa, uma vez que grupos bem definidos podem ajudar na identificação de tendências.

4.4. Modelagem

Nesta seção, serão abordados o processo de construção dos modelos preditivos elaborados pelo grupo Mettha que respondem às hipóteses levantadas no subtópico 4.2.3 deste documento.

4.4.1. Modelos Preditivos

Um modelo preditivo é uma ferramenta essencial na análise de dados, sendo uma função matemática que busca identificar padrões com base em informações previamente coletadas. Sua função primordial é prever eventos futuros ou comportamentos com base em dados históricos, oferecendo *insights* valiosos para tomada de decisões.

O seu desenvolvimento ocorre mediante algoritmos de Machine Learning e técnicas de inteligência artificial, como KNN (*K-nearest neighbor*), *K-means*, Random Forest e Naive Bayes, explicados nos textos abaixo, além de ferramentas estatísticas. Essas abordagens capacitam o sistema a calcular a probabilidade e a precisão de ocorrência de determinadas situações.

4.4.2. KNN + K-means

Durante os treinamentos e testes, foi utilizado o modelo KNN juntamente com K-means. Esses algoritmos visam fazer recomendações de vagas de emprego (Jobs) para um usuário com base na similaridade de perfis entre o usuário e as vagas.

A função do K-means nesse modelo é através dos dados disponibilizados, fazer uma clusterização entre os gestores e usuários da plataforma, ou seja, separar em grupos por meio da similaridade de cada usuário. Já o KNN terá como objetivo achar as empresas mais próximas ao usuário, ou melhor, através dos pontos no gráfico de classificação o modelo irá dizer quais os perfis com maiores similaridades baseado na distância entre eles, sendo o quanto mais próximo mais similar.

A seguir, são apresentadas as etapas da construção desse modelo.

1. Seleção de variáveis independentes para o KNN (features de Jobs):

Foram selecionadas as variáveis independentes para o KNN a partir dos dados de treino (Jobs). As features selecionadas foram armazenadas na variável `x_jobs`.

2. Seleção da Variável Dependente:

A classe ou variável dependente é o `job_id`. Essa variável é importante porque representa as vagas de emprego às quais as recomendações se referirão. Os `job_id` correspondentes aos dados de treino (Jobs) foram armazenados na variável `y_jobs`.

3. Seleção dos Dados do Usuário para Recomendação:

Os dados do usuário para os quais deseja-se fazer recomendações foram selecionados a partir dos dados de teste (Users) e armazenados na variável `x_user`.

4. Definição da Função de recomendação:

Foi definida uma função chamada “recomendar” que aceita como entrada o perfil do usuário, o número de recomendações desejadas (`n_recomend`), a métrica de similaridade (`metric`), e um limite mínimo de similaridade (`min_similarity`) para filtrar as recomendações.

A função cria uma instância do classificador KNN (`KNeighborsClassifier`) com o número de vizinhos definido como `n_recomend` e a métrica de similaridade especificada.

Em seguida, o KNN é treinado com os dados de Jobs (`x_jobs` e `y_jobs`). A função utiliza o KNN para encontrar os vizinhos mais próximos do usuário fornecido (`user`) e calcula a similaridade entre o usuário e esses vizinhos. Os `job_id` correspondentes aos vizinhos mais próximos são extraídos.

A função também calcula a diferença entre o perfil do usuário e os perfis dos Jobs recomendados com base nas diferenças nas features.

Por fim, os resultados são armazenados em um `DataFrame` que inclui informações sobre o `job_id`, a similaridade, e as diferenças médias em algumas features específicas (por exemplo, `dif_média_genius`, `dif_média_life`, `dif_média_geral`).

5. Realização do Teste de Recomendação:

Foi realizado um teste de recomendação para um usuário específico no índice 1000 da matriz `x_user`. Foram recomendados 100 jobs para esse usuário com base na similaridade de perfil.

6. Resultados e Ordenação das Recomendações:

Os resultados da recomendação foram apresentados em um DataFrame que inclui o `job_id`, a métrica de similaridade e as diferenças médias em algumas features. Observou-se que os jobs foram ordenados conforme a similaridade, do mais similar ao menos similar, inclusive com similaridades negativas (indicando dissimilaridade).

Figura 14: Resultado do modelo

```
recomendar(X_user_with_lat_long[0])
```

	job_id	semelhança	dif_média_genius	dif_média_life	dif_média_geral	lat_dif	long_dif	grupo
0	2848.0	0.999961	0.281768	0.133424	0.348480	0.0	0.0	0.0
1	2849.0	0.999957	0.149323	0.169159	0.233902	0.0	0.0	0.0
2	2851.0	0.999945	0.140337	0.375041	0.327857	0.0	0.0	0.0
3	2888.0	0.999944	0.320944	0.497823	0.569856	0.0	0.0	0.0
4	2748.0	0.999942	0.129292	0.262994	0.260789	0.0	0.0	0.0
5	2811.0	0.999937	0.463718	0.025686	0.450876	0.0	0.0	0.0
6	2752.0	0.999927	0.634261	0.157200	0.712860	0.0	0.0	0.0
7	2850.0	0.999926	0.530614	0.102279	0.581753	0.0	0.0	0.0
8	2727.0	0.999926	0.066439	0.389847	0.128485	0.0	0.0	0.0
9	2884.0	0.999921	0.266244	0.098101	0.315294	0.0	0.0	0.0

Fonte: Colab notebook elaborado pelos autores

4.4.3. Random Forest + K-means

A partir dos testes e treinamentos realizados, foram utilizados os modelos Random Forest e K-means, com o intuito de, através de árvores de decisões, fazer recomendações de vagas (Jobs) para os usuários (Users).

A função do K-means segue a mesma do modelo citado anteriormente na seção 4.4.2. Já o modelo Random Forest tem como sua funcionalidade várias árvores de

decisões aleatórias, que a cada decisão ela divide o conjunto de dados em segmentos menores com base em regras aleatórias até chegar em uma vaga com a menor impureza. No final, o modelo junta esses resultados e entrega a melhor combinação.

A seguir, são descritas as etapas da construção desse modelo.

1. Definição da Função de Recomendação (`recomendar_rf`):

A função `recomendar_rf` é definida para recomendar oportunidades de emprego (`jobs`) com base em um usuário fornecido como parâmetro.

2. Seleção de Oportunidades de Emprego Próximas:

A função começa selecionando oportunidades de emprego (`jobs`) que estão próximas à localização geográfica do usuário. Ela filtra o conjunto de dados `data_full` para encontrar oportunidades de emprego com a mesma latitude e longitude do usuário.

3. Extração de Características Relevantes:*

São extraídas características relevantes das oportunidades de emprego, incluindo pontuações em habilidades (como resiliência, engajamento, colaboração, etc.), bem como informações geográficas (latitude e longitude).

4. Seleção do Alvo (`Y_jobs`):

A variável de destino `Y_jobs` é definida para conter os identificadores únicos (IDs) das oportunidades de emprego correspondentes ao subconjunto de dados filtrados.

5. Treinamento do Modelo Random Forest:*

Um modelo Random Forest é inicializado com 80 estimadores (árvores de decisão) e uma semente aleatória definida como 42. O modelo é treinado usando as características extraídas das oportunidades de emprego e os IDs correspondentes como alvo.

6. Previsão de Classificação:

O modelo Random Forest é usado para calcular o nível de impureza e achar o melhor padrão para encaixar um usuário com as características da vaga de emprego com base nos atributos do usuário fornecidos como parâmetro. O nível de

impureza é uma medida que indica o quão compatíveis são os usuários com as oportunidades de emprego. Esse valor é armazenado em `y_pred_test`.

7. Recomendação Final:

A função retorna um dicionário contendo o ID da oportunidade de emprego recomendada e informações adicionais sobre essa oportunidade (como média de genialidade, média de vida, média geral, latitude e longitude). A recomendação é baseada na oportunidade de emprego com o menor nível de impureza (`np.argmin(y_pred_test)`), indicando a mais compatível com base nas características do usuário.

Figura 15: Resultado do modelo

```
recomendar_rf(X_user_with_lat_long[0],10)
```

(1, 101)

	job_id	impureza	dif_média_genius	dif_média_life	dif_média_geral	lat_dif	long_dif	grupo
0	2727.0	0.0	0.066439	0.389847	0.128485	0.0	0.0	0.0
1	2808.0	0.0	0.039265	0.143040	0.110785	0.0	0.0	0.0
2	2848.0	0.0	0.281768	0.133424	0.348480	0.0	0.0	0.0
3	2810.0	0.0	1.860094	0.563394	2.141791	0.0	0.0	0.0
4	2798.0	0.0	1.099711	0.170362	1.184892	0.0	0.0	0.0
5	2841.0	0.0	0.320570	0.137509	0.389324	0.0	0.0	0.0
6	2797.0	0.0	0.865217	0.109324	0.919879	0.0	0.0	0.0
7	2796.0	0.0	0.439612	0.367379	0.255923	0.0	0.0	0.0
8	2807.0	0.0	0.568620	0.155011	0.491114	0.0	0.0	0.0
9	2757.0	0.0	1.360581	0.523135	1.622149	0.0	0.0	0.0

Fonte: Colab notebook elaborado pelos autores

4.4.4. Naive Bayes + K-means

Naive Bayes é um método de machine learning que usa as frequências das ocorrências em uma base de dados para prever uma variável de interesse. Seu nome vem do modelo estatístico bayesiano, que diz que o grau com que devemos acreditar numa afirmação vai ser ligeiramente alterado por novas evidências. Ele é utilizado para classificar dados em categorias com base na probabilidade condicional.

A seguir, são apresentadas as etapas para a construção desse modelo.

1. Seleção de variáveis independentes para o Naive Bayes (feature de Jobs)

Foram selecionadas as variáveis independentes para o Naive Bayes a partir dos dados de treino (Jobs) com base nas coordenadas de latitude e longitude do usuário. O código filtra os dados de Jobs com a mesma latitude e longitude do usuário (`data_full['latitude'] == user[-2]` e `data_full['longitude'] == user[-1]`). Em seguida, seleciona as features relevantes para esse subconjunto de dados, que incluem pontuações em várias dimensões, médias e coordenadas de latitude e longitude. Essas features foram armazenadas na variável `X_jobs_with_LatLong`.

2. Seleção da variável dependente (Classe - job_id)

A classe ou variável dependente é o id dos jobs correspondentes ao subconjunto de dados filtrados com base na latitude e longitude do usuário. Os `job_id` correspondentes foram armazenados na variável `Y_jobs`.

3. Treinamento do classificador

Foi criada uma instância do classificador Naive Bayes Gaussiano (GaussianNB) chamada NB. O classificador foi treinado com as features de Jobs (`X_jobs_with_LatLong`) e seus `job_id` correspondentes (`Y_jobs`).

4. Predição para o usuário

A função `recomendar_nb` recebe como entrada o perfil do usuário (`user`) e o número de recomendações desejadas (`n`). Foi feita uma previsão de probabilidade para o usuário fornecido usando o método `predict_proba`. Isso retorna as probabilidades associadas a cada classe de `job_id` com base no perfil do usuário.

5. Ordenação das recomendações e criação do dataframe de resultados:

As probabilidades previstas foram usadas para ordenar os `job_id` em ordem decrescente de probabilidade. As melhores recomendações (com as maiores probabilidades) foram selecionadas para o usuário.

Além disso, a função calculou as diferenças entre as features dos jobs recomendados e as features do usuário em relação a métricas relevantes (por exemplo, diferenças nas médias de genialidade, vida e geral, bem como diferenças de latitude e longitude).

Os resultados foram armazenados em um DataFrame que inclui informações sobre o `job_id`, a probabilidade prevista e as diferenças em métricas relevantes, bem como as diferenças de latitude e longitude.

6. Resultados e Saída da Recomendação:

A função retorna o DataFrame com as recomendações para o usuário.

Figura 16: Resultado do modelo

```
recomendar_nb(X_user_with_lat_long[3],10)
```

	job_id	Probabilidade	dif_média_genius	dif_média_life	dif_média_geral	lat_dif	long_dif	grupo
0	2915.0	1.0	0.611030	0.018262	0.601899	0.0	0.0	0.0
1	2727.0	0.0	1.141561	0.294282	0.994420	0.0	0.0	0.0
2	2840.0	0.0	1.174759	0.048894	1.199206	0.0	0.0	0.0
3	2811.0	0.0	0.611404	0.121251	0.672029	0.0	0.0	0.0
4	2845.0	0.0	0.285494	0.661188	0.045100	0.0	0.0	0.0
5	2845.0	0.0	0.285494	0.661188	0.045100	0.0	0.0	0.0
6	2845.0	0.0	0.285494	0.661188	0.045100	0.0	0.0	0.0
7	2845.0	0.0	0.285494	0.661188	0.045100	0.0	0.0	0.0
8	2849.0	0.0	1.224445	0.264724	1.356807	0.0	0.0	0.0
9	2767.0	0.0	0.358375	2.987016	1.135133	0.0	0.0	0.0

Fonte: Colab notebook elaborado pelos autores

4.5. Avaliação

A solução final escolhida para o modelo foi o K-Nearest Neighbors (KNN). Esse algoritmo classifica um dado desconhecido com base na proximidade dos seus vizinhos mais próximos (como diz o próprio nome do algoritmo). Para mais detalhes, verificar a seção 4.4.2. deste documento.

Pela natureza do projeto, não é possível validar o modelo por meio de métricas convencionais, como acurácia e precisão. Por conta disso, é necessário utilizar outros critérios para decidir o modelo finalista. Entre eles, há o índice de diferença média entre usuário e gestor, o qual, por meio de métricas geométricas, define o quão próximo os dados de um certo usuário está de uma certa oportunidade de trabalho.

Para tal, se considera tanto o princípio do KNN, quanto os resultados apresentados para ele. O primeiro ponto diz respeito ao funcionamento do algoritmo em si, ou seja, da maneira com que ele classifica os dados. Como citado anteriormente, esse modelo classifica um dado conforme os dados mais próximos a ele. Então, por meio desse conceito, ele é capaz de aproximar o usuário do Bettha aos gestores que apresentam as características mais próximas a eles. O segundo ponto refere-se aos resultados

apresentados por esse modelo, ou seja, sua performance em si. Dentre os apresentados, ele foi o que obteve maiores índices de semelhança, e por isso é capaz de apresentar vagas melhores para cada usuário.

5. Conclusões e Recomendações

Depois de 10 semanas de estudos sobre o contexto da empresa parceira, de pré-processamento de dados, de implementações de modelos, e de testar hipóteses, o grupo Mettha avaliou o melhor modelo como sendo o KNN, por apresentar melhor performance, como descrito na seção 4.5. Ademais, é importante destacar alguns aspectos finais desse projeto.

Primeiramente, é importante ressaltar a importância dos dados para o desenvolvimento desse projeto. Por se tratar de um modelo preditivo, a qualidade e quantidade são muito importantes para a performance do modelo, principalmente para o KNN, já que esse algoritmo é sensível a *outliers*. Desde a coleta de dados até o pré-processamento, é necessário que todas as etapas sejam efetivadas para garantir uma maior acurácia do modelo, de forma que ele possa recomendar as melhores vagas possíveis para o usuário. Assim, quanto melhor forem essas duas características, melhor se tornará o desempenho do modelo.

Além disso, mesmo que os dados utilizados tenham sido coletados por meio de testes, como são pessoas que os respondem na plataforma do Bettha, pode haver vieses no modelo construído. Dessa forma, é importante testar outras abordagens também, não se atendo somente às descritas nessa documentação, para minimizar esses possíveis vieses, que podem comprometer a performance de um modelo.

Outra sugestão, é utilizar as abordagens aqui feitas para construção de outros modelos com objetivos diferentes, que podem se inspirar em algumas aproximações feitas nesse projeto. Nesse sentido, é possível considerar outras features também, não comentadas neste documento, como hobbies e histórico educacional. Por fim, a escolha do melhor modelo (KNN) pode não ser a melhor para essas outras abordagens, que levam em consideração outros fatores. Por isso, é sempre interessante testar outras hipóteses e modelos distintos.

Em resumo, a importância dos dados, a presença de vieses, utilização do mesmo modelo para outras abordagens e a relevância de se testar outras hipóteses e soluções são fundamentais para se descobrir a melhor aplicação que funcione para os objetivos estabelecidos.

6. Referências

ChatGPT, <https://openai.com/blog/chatgpt>.

Para a elaboração do subtópico 4.1.9, as informações foram retiradas da política de privacidade da própria empresa (<https://www.bettha.com/politica-de-privacidade>) acesso 25/09/2023