

# Datathon Fase 05

Empresa Decision

# Sobre o Projeto

Nosso propósito é **sugerir soluções de produto** capazes de aprimorar significativamente o processo de contratação da Decision. Para isso, realizamos um **estudo aprofundado** das bases e sugestões em relação ao processo atual, incorporando as **principais dores** mencionadas no briefing do projeto.






# Base de vagas

## Base de Vagas

O documento é uma extensa lista de títulos de vagas de emprego, predominantemente na área de tecnologia da informação (TI), mas também abrangendo áreas como Administração, Finanças e Recursos Humanos. As posições variam em nível de experiência, incluindo Júnior, Pleno, Sênior e Especialista. As descrições das vagas destacam diversas especializações em TI, como desenvolvimento (ABAP, Java, .NET, FullStack), administração de sistemas (Linux, Windows, Oracle, SAP), suporte, segurança da informação, análise de dados (BI, Power BI) e metodologias ágeis (Agile Coach, Scrum Master). Há também menções a setores específicos, como financeiro, manufatura e varejo, e a localidades, indicando que são oportunidades em diferentes regiões.



# Detalhamento sobre a diversidade de vagas

As vagas podem ser clusterizadas em 8 macro temas

## 1. Papéis de Análise e Desenvolvimento de Software e Sistemas:

*Desenvolvimento Geral e Linguagens:*

*DevOps e Automação:*

## 2. Papéis de Infraestrutura e Suporte de TI:

*Administração de Sistemas e Redes*

*Suporte Técnico*

## 3. Papéis de Análise de Dados e Business Intelligence (BI):

*Focados na coleta, processamento, análise e visualização de dados para suporte à decisão:*

## 4. Papéis de Análise de Negócios e Processos:

*Focados na ponte entre a tecnologia e as necessidades do negócio*

## 5. Papéis de Qualidade e Testes:

*Essenciais para garantir a robustez e funcionalidade das soluções:*

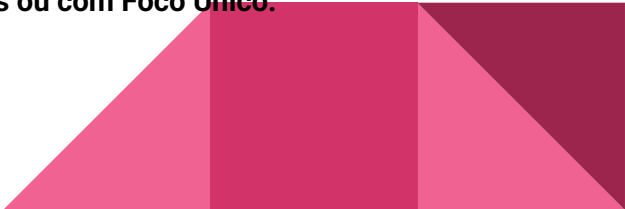
## 6. Papéis Administrativos e Corporativos:

*Abrangem funções de suporte à gestão e operações da empresa:*

## 7. Papéis de Governança, Risco e Conformidade (GRC):

*Focados na segurança da informação, auditoria e gestão de riscos:*

## 8. Papéis Específicos ou com Foco Único:



## Parecer Técnico da Base de Dados – dados\_vagas

A base `dados\_vagas` contém informações detalhadas sobre solicitações de vagas, perfis desejados, condições de contratação e requisitos técnicos. Apesar de seu potencial para análises preditivas e automatização de processos seletivos, ela apresenta problemas estruturais, semânticos e de qualidade que prejudicam sua utilização direta em projetos de dados e machine learning. Este diagnóstico propõe um plano de saneamento e modelagem para viabilizar análises robustas e escaláveis.



# Diagnóstico Técnico da base\_vagas


Sintoma	Exemplos / Evidências	Impacto Direto
Muitos valores nulos	<code>telefone</code> (≈100%), <code>valor_compra_2</code> (≈99,7%)	Métricas incompletas, esparsidade em modelos
Dados sensíveis desnecessários	<code>nome</code> , <code>telefone</code> , <code>nome substituto</code>	Risco à privacidade (LGPD)
Colunas com nomes inconsistentes	<code>nivel profissional</code> , <code>competencia_tecnicas_e_comportamentais</code>	Dificulta padronização e análise
Valores financeiros como texto	<code>valor_venda</code> , <code>valor_compra_1</code> com símbolos e texto	Impede cálculos e análise financeira
Datas em formato textual	<code>data_requisicao</code> , <code>data_final</code>	Erros em ordenações e filtros
Campos compostos em texto livre	<code>titulo_vaga</code> , <code>principais atividades</code>	Dificulta extração de atributos estruturados
Categóricos com variações textuais	<code>tipo_contratacao</code> , <code>origem_vaga</code> com grafias distintas	Quebra de agrupamentos, inconsistência em relatórios

# Proposta de Modelo Alvo base\_dimenssão e base\_fato

Dimensão	Atributos (Dimenssão)
<b>dim_empresa</b>	<code>id_empresa, nome_cliente, solicitante, analista_responsavel, empresa_divisao</code>
<b>dim_vaga</b>	<code>id_vaga, titulo_padronizado, area_atuacao, nivel, tipo_contratacao, prioridade, objetivo, pcd, local, regioao</code>
<b>dim_tempo</b>	<code>id_tempo, data_requisicao, limite_contratacao, data_inicio, data_final</code>
<b>dim_requisitos</b>	<code>id_requisito, competencias, habilidades_comportamentais, idiomas, nivel_academico, faixa_etaria</code>

## Atributos (Modelo Fato)

`id_vaga, id_empresa, id_tempo, id_requisito, valor_venda, valor_compra, exige_viagens, equipamentos, status_vaga`





# Etapas de Saneamento e Padronização e Benefícios

- Remoção de Dados Sensíveis (excluir `nome`, `telefone`, `nome\_substituto`)
- Conversão de Datas com `pd.to\_datetime`
- Tratamento de Valores Monetários para `float`
- Padronização de Colunas Categóricas (clientes, analistas, tipo de contrato)
- Renomear colunas para snake\_case
- Aplicar técnicas de NLP nos campos de texto livre
- Imputar ou descartar colunas com muitos nulos

## Ganhos Esperados

Comparação entre a situação atual e os benefícios após a aplicação das melhorias.



# Próximos Passos Recomendados

## ✓ Curto Prazo (0–30 dias)

- Padronizar nomes de colunas e tipos de dados
- Criar dicionários para campos categóricos
- Aplicar limpeza básica e salvar versão tratada

## 🔧 Médio Prazo (30–90 dias)

- Implementar pipeline de transformação com validação
- Desenvolver esquema dimensional para análises e dashboards
- Iniciar extração de atributos textuais com NLP

## 📈 Longo Prazo (90+ dias)

- Modelagem preditiva
- Uso da base em motor de recomendação
- Conectar com dados de performance

# Base de Prospeção

## Diagnóstico Técnico da Base de Dados – dados\_prospect

A base de dados `dados\_prospect` contém informações importantes sobre candidatos, vagas e interações com recrutadores. No entanto, ela apresenta inconsistências estruturais e semânticas que prejudicam análises confiáveis e a criação de indicadores automatizados. Este diagnóstico propõe um plano de padronização e estruturação com vistas a um modelo robusto e escalável, aderente às boas práticas de engenharia de dados.



# Diagnóstico Técnico da base\_vagas


Sintoma	Exemplos / Evidências	Impacto Direto
Muitos valores nulos	<code>`modalidade`</code> (~97% nula), <code>`comentario`</code> (~74%), outras com 5–12% ausentes	Relatórios incompletos e métricas enviesadas
IDs não únicos	<code>`codigo`</code> com mais de 27 mil duplicados (≈ 48%)	Dificuldade em utilizar como chave primária
Granularidade conflitante	Mesmo candidato aparecendo em múltiplas linhas sem um identificador único	Complexidade para mapear históricos individuais
Coluna <code>`titulo`</code> mal estruturada	Mescla informações de ID da vaga, nível (JR/PL/SR), função e tecnologia	Impossibilidade de filtros e agrupamentos automatizados
Recrutadores com grafias diversas	77 formas diferentes de representar o mesmo recrutador	Quebra de agrupamentos e métricas de performance
Datas não padronizadas	<code>`data_candidatura`</code> e <code>`ultima_atualizacao`</code> em formatos variados	<code>`data_candidatura`</code> e <code>`ultima_atualizacao`</code> em formatos variados

# Proposta de Modelo Alvo base\_dimenssão e base\_fato

Dimensão	Atributos (Dimenssão)
<b>dim_recrutador</b>	id_recrutador, matricula, nome_padronizado, email_corporativo, area
<b>dim_candidato</b>	id_candidato, nome_limpo, cpf/email/telefone, data_primeiro_contato
<b>dim_vaga</b>	id_vaga, codigo_externo, titulo_padronizado, nivel, tecnologia_base, modalidade, status

## Atributos (Moldeo Fato)

fato\_candidatura: sk\_candidatura, id\_candidato, id\_vaga, id\_recrutador, data\_candidatura, data\_ultimo\_movimento, situacao, observacoes



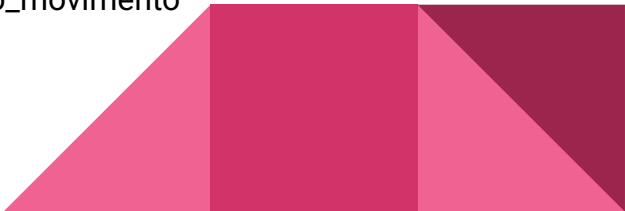
# Etapas de Saneamento e Padronização

- Remoção de Linhas vazias ou com informações incompletas
- Quebra do Campo `titulo` (id\_vaga, nivel, titulo\_padronizado)
- Criação de Dicionário de Recrutadores
- Normalização de Datas
- Deduplicação de Candidatos
- Controle de Valores (Enums)
- Governança e Boas Práticas de Nomeação



# Ganhos Esperados

Item	Antes	Depois
Joins confiáveis	Quebras por duplicidade de código	FKs consistentes entre dimensões e fato
Tempo de análise	Necessidade de limpeza manual	Querys diretas em camadas "silver/gold"
Métricas por nível/vaga	Regex manual e impreciso	Coluna `nivel` estruturada
Indicadores por recrutador	Fragmentados por grafias	Métricas sólidas via `id_recrutador`
Auditoria	Rastreamento difícil	Histórico detalhado por `data_ultimo_movimento`





# Próximos Passos Recomendados

## ✓ Curto Prazo (até 30 dias)

- Criar playbook de ingestão limpa com validações
- Gerar versão estruturada da base histórica
- Implementar testes automatizados

## 🔧 Médio Prazo (30–90 dias)

- Criar pipelines agendados de ingestão e limpeza
- Implantar repositório de código versionado (Git)
- Desenvolver dashboards conectados às novas tabelas dimensionais

## 📈 Longo Prazo (90+ dias)

- Treinar time de recrutamento em uso correto de códigos e níveis
- Iniciar modelagens preditivas com dados confiáveis
- Expandir modelagem para incluir pipeline completo de seleção

# Base de Candidatos

## Base de Candidatos

A base de dados de candidatos fornecida, embora contenha informações valiosas para recrutamento e seleção, apresenta problemas significativos de consistência, padronização e completude que comprometem a qualidade de análises de dados e a eficiência de processos automatizados. Há uma notável duplicação de colunas, valores ausentes generalizados, formatos inconsistentes (especialmente em datas e telefones) e dados textuais não estruturados que dificultam a extração de insights. Um plano de saneamento e padronização é essencial para transformar esta base em um ativo estratégico.



# Diagnóstico Técnico da base\_applicants

Sintoma	Exemplos / Evidências	Impacto Direto
<b>Duplicação de Colunas</b>	nome (aparece 2x), email (aparece 2x), telefone_recado (aparece 2x)	Redundância de dados, aumento do risco de inconsistência, maior complexidade na manipulação e armazenamento.
<b>Valores Nulos/Ausentes Significativos</b>	cpf, email_secundario, telefone_recado (muitos vazios), objetivo_profissional (frequentemente vazio)	Métricas incompletas, dificuldade em realizar análises demográficas ou de contato, falha em modelos de Machine Learning.
<b>Inconsistência na Formatação de Datas</b>	0000-00-00 em data_nascimento, DD-MM-YYYY HH:MM:SS em data_criacao, DD/MM/YYYY HH:MM em data_aceite	Impede ordenação cronológica, cálculos de idade precisos, filtros e análises temporais.
<b>Inconsistência na Formatação de Telefones</b>	(XX) XXXXX-XXXX vs XX XXXXX-XXXX	Dificulta a busca, padronização de contatos e validação.
<b>Campos de Texto Livre Não Estruturados</b>	objetivo_profissional (descrições longas e variadas), fonte_indicacao (Site de Empregos: Infojobs, Outros: Contato do RH)	Dificuldade em categorizar, extrair insights, criar filtros consistentes ou usar em modelos preditivos sem NLP.

Sintoma	Exemplos / Evidências	Impacto Direto
Inconsistência em Dados Categóricos	sexo (Feminino, Masculino, mas ausente em vários casos), estado_civil (ausente, variação de capitalização esperada)	Análise demográfica e de perfil comprometida, erros em agrupamentos. pcd também apresenta ausência.
Dados Derivados ou Redundantes na fonte_indicacao	Contém a fonte e, em alguns casos, um detalhe adicional (Site de Empregos: Infojobs).	Mistura de informações, dificultando a análise direta da fonte principal.
Valores Inválidos/Placeholder	0000-00-00 em data_nascimento, Cadastro anterior ao registro de aceite em data_aceite	Impede cálculos de idade, sinaliza falta de informação real ou erro na coleta/migração de dados.
Nomes de Colunas com Caracteres Especiais/Espaços	sabendo_de_nos_por (espaços), fonte_indicacao (acentuação pode ser um problema)	Dificulta manipulação em ferramentas de análise e linguagens de programação (Python, R, SQL).



# Proposta de Modelo Alvo base\_dimenssão e base\_fato

Dimensão	Atributos (Dimenssão)
<b>Candidato</b>	ID, Nome, CPF, Data Nascimento, Sexo, Estado Civil, PCD, Endereço (separar em Rua, Número, Bairro, Cidade, Estado, CEP), Telefone Celular, Telefone Recado, E-mail Principal, E-mail Secundário
<b>Origem</b>	ID, Fonte Principal, Detalhe da Fonte
<b>Profissional</b>	ID, Objetivo Profissional (categorizado ou limpo), Código Profissional
<b>Tempo</b>	ID, Data Criação, Data Atualização, Data Aceite
<b>Responsável</b>	ID, Nome do Responsável (inserido_por)

*\*Ausência de Chaves Primárias e Estrangeiras Claras: Embora codigo\_profissional possa atuar como um ID de candidato, não há chaves estrangeiras que liguem claramente os dados de contato, profissionais e de origem, aumentando a dificuldade de consultas complexas e garantindo a integridade dos dados*

# Próximos Passos Recomendados

## ✓ Curto Prazo (0-30 dias): Foco na Limpeza e Padronização Básica

### 1. Identificar e Unificar Colunas Duplicadas:

- Consolidar `nome`, `email`, `telefone_recado`. Decidir qual é a coluna "oficial" ou criar uma lógica para combinar.

### 2. Renomear Colunas para `snake_case` e Padronizar Nomes:

- Ex: `telefone_recado` (já está ok), `data_criacao`, `sabendo_de_nos_por` para `origem_informacao` ou `como_soube_de_nos`.

### 3. Tratamento de Datas:

- Converter `data_criacao`, `data_atualizacao`, `data_aceite`, `data_nascimento` para o tipo `datetime`.
- Substituir valores inválidos (0000-00-00, texto como `Cadastro anterior...`) por `NaN` (nulos).
- Padronizar o formato de exibição das datas (ex: `YYYY-MM-DD`).

### 4. Tratamento de Telefones:

- Remover caracteres especiais ( , , - , espaços para obter apenas os dígitos.
- Separar DDD e número em colunas distintas, se necessário para análise regional.
- Identificar e corrigir formatos inconsistentes (ex: `71 99870-9516` vs `(11) 97048-2708`).



Curto Prazo (0-30 dias): Foco na Limpeza e Padronização Básica

## 5. Padronização de Colunas Categóricas:

- **sexo**: Corrigir inconsistências (ex: "Feminino" para nomes masculinos) e padronizar valores (ex: "M", "F", "Não Informado"). Imputar nulos ou categorizá-los.
- **estado\_civil**: Padronizar (ex: "Solteiro", "Casado", "União Estável", "Divorciado", "Viúvo", "Não Informado").
- **pcd**: Padronizar (ex: "Sim", "Não", "Não Informado").

## 6. Tratamento de Nulos/Valores Ausentes:


- Analisar a porcentagem de nulos em cada coluna. Para colunas com muitos nulos (**cpf**, **email\_secundario**, **telefone\_recado** secundário), decidir se devem ser descartadas, imputadas ou mantidas como nulas, entendendo o impacto na análise.



## 7. Extração de Informações da **fonte\_indicacao**:

- Dividir **fonte\_indicacao** em **tipo\_fonte** (ex: "Site de Empregos", "Outros", "Anúncio", "Indicação de colaborador") e **detalhe\_fonte** (ex: "Infojobs", "Contato do RH", "Indeed", "Glassdoor").

## 8. Limpeza de Campos de Texto Livre (**objetivo\_profissional**):

- Remover espaços extras, caracteres especiais desnecessários.
  - Considerar aplicar NLP para extrair termos-chave ou categorizar automaticamente, mesmo que em um nível básico.
- 

## Médio Prazo (30-90 dias): Foco na Estrutura e Enriquecimento

### 1. Refatoração da Estrutura (Normalização):

- Criar tabelas separadas para Dimensões (Candidato, Origem, Tempo, Responsável, Profissional) e uma Tabela Fato que relacione essas dimensões, se houver eventos (ex: "inscrição em vaga", "entrevista"). Se a base é apenas de cadastro de candidatos, uma boa normalização de tabelas (**candidatos**, **contatos**, **enderecos**, **historico\_profissionais**) já seria um grande ganho.

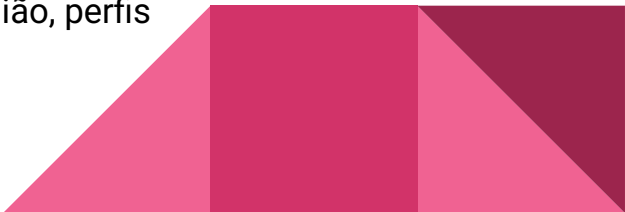
### 2. Definição e Aplicação de Chaves:

- Garantir chaves primárias únicas para cada entidade (ex: **id\_candidato**).
- Estabelecer chaves estrangeiras para ligar as tabelas, garantindo a integridade referencial.

### 3. Padronização Geográfica:

- Separar **local** em **cidade** e **estado**.
- Criar uma tabela de dimensão de localização, se a granularidade de análise for maior.
- Padronizar nomes de cidades e estados (ex: "São Paulo" vs "são paulo", "SP" vs "São Paulo").

### 4. Enriquecimento de Dados (Opcional, mas valioso):

- Se possível, integrar dados externos (ex: dados de renda média por região, perfis de vagas mais procuradas) para enriquecer a análise.
- 

## **Longo Prazo (> 90 dias): Foco em Análise Avançada e Automação**

### **1. Construção de Dicionário de Dados:**

- Documentar todas as colunas, seus tipos, valores esperados, restrições e propósito.


### **2. Implementação de Pipeline de Dados:**

- Automatizar o processo de limpeza e padronização (ETL/ELT).

### **3. Análises Preditivas e Modelos de ML:**

- Com os dados limpos e estruturados, será possível desenvolver modelos para prever sucesso de contratação, identificar perfis mais adequados, otimizar a busca por candidatos, etc.

### **4. Dashboards e Relatórios:**

- Construir dashboards interativos para monitorar métricas de recrutamento, perfis de candidatos e fontes de aquisição.
- 

# Proposta para Decision

Entrevista de Engajamento dos candidatos

# Principais dores mencionadas

Analisando o briefing da proposta de trabalho baseado nas principais dores, entendemos que seria uma oportunidade bastante relevante apoiar com uma dor:

*“Hoje a Decision possui algumas dores como:*

- *Falta de padronização em entrevistas, o que pode gerar perda de informações valiosas.*
- *Dificuldade em identificar o real engajamento dos candidatos.”*

Fatores que tornam a entrevista parte essencial do processo de match:

- Fit Cultural: avaliamos se o candidato se encaixa nos valores e cultura da empresa contratante.
- Engajamento e Motivação: identificamos se o candidato está realmente interessado na vaga e motivado para assumir o desafio.



# Objetivos da entrevista

Avaliar o nível de engajamento do candidato com o trabalho, propósito da empresa, aprendizado contínuo, e cultura organizacional, para facilitar ranking e comparação entre candidatos.

A entrevista consiste em 5 blocos:

**Bloco 1 – Propósito e Motivação Pessoal**

**Bloco 2 – Proatividade e Entrega de Valor**

**Bloco 3 – Aprendizado e Evolução Contínua**

**Bloco 4 – Cultura, Relacionamento e Sentimento de Pertencimento**

**Bloco 5 – Comprometimento e Visão de Futuro**

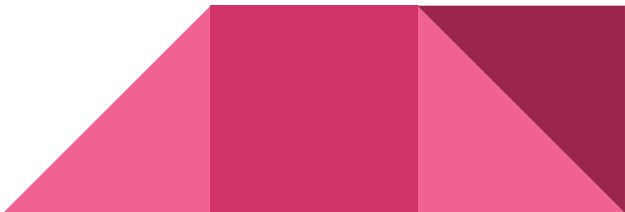


# Sugestão de Escala de Avaliação (Ranking)

Para ranquear candidatos com base no engajamento, você pode usar uma **escala de 1 a 5**, sendo:

- **1 – Muito baixo engajamento**
- **2 – Baixo engajamento**
- **3 – Engajamento moderado**
- **4 – Engajamento alto**
- **5 – Altamente engajado**

Cada pergunta pode ser avaliada individualmente com base em critérios como:

- Clareza na resposta
  - Consistência com valores da empresa
  - Sinais de motivação intrínseca
  - Evidências de ação e proatividade
  - Interesse genuíno na vaga/empresa
- 

# Premissas de cálculo da nota

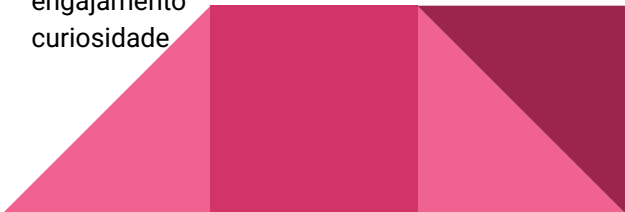
A nota é calculada levando em consideração tamanho da resposta e palavras chave, criamos a pesquisa com 60 palavras chave, mas as perguntas e as palavras podem ser reajustada conforme necessidade, é apenas uma proposta:

cultura  
evolução  
aprendizado  
autonomia  
desenvolvimento  
proatividade  
pertencimento  
colaboração  
desafio  
feedback  
responsabilidade  
valores  
melhoria contínua  
engajamento  
motivação  
inovação  
reconhecimento  
comunicação

responsabilidade  
valores  
melhoria contínua  
engajamento  
motivação  
liderança  
resolução de problemas  
curiosidade  
autoconhecimento  
comunicação  
liderança  
responsabilidade  
valores  
melhoria contínua  
responsabilidade  
autoconhecimento  
visão de futuro  
relacionamento interpessoal

valores  
melhoria contínua  
engajamento  
motivação  
liderança  
resolução de problemas  
curiosidade  
autoconhecimento  
comunicação  
valores  
melhoria contínua  
engajamento  
motivação  
liderança  
resolução de problemas  
motivação  
liderança  
resolução de problemas

curiosidade  
autoconhecimento  
comunicação  
desenvolvimento  
proatividade  
pertencimento  
colaboração  
desafio  
feedback  
responsabilidade  
valores  
melhoria contínua  
engajamento  
curiosidade





# Proposta para Decision

Ranking de candidatos compatíveis com as vagas

# Como Funciona o Ranking de Candidatos por Vaga

O objetivo do ranking é encontrar os candidatos mais aderentes ao perfil da vaga, comparando o texto da descrição da vaga com as informações do perfil do candidato. Isso é feito usando técnicas de Processamento de Linguagem Natural (NLP), principalmente o método de similaridade textual com TF-IDF e cosseno.



# Etapas do Processo

## 1. Consolidação das Informações

- Cada vaga tem uma descrição consolidada (**texto\_vaga**), combinando:
  - Título da vaga
  - Área de atuação
  - Atividades esperadas
  - Competências e habilidades
- Cada candidato também tem um texto (**perfil\_texto**), combinando:
  - Objetivo profissional
  - Título profissional
  - Nível acadêmico
  - Idiomas
  - Cursos e área de atuação


## 2. Transformação de Texto com TF-IDF

- TF-IDF (Term Frequency-Inverse Document Frequency) transforma os textos em **vetores numéricos**.
- Esse vetor representa a **importância de cada palavra** no documento em relação ao conjunto total.

## 3. Cálculo de Similaridade Cosseno

- Com os vetores criados, usamos a **similaridade do cosseno** para medir a "distância" entre o texto da vaga e o texto de cada candidato.
- Quanto mais próximo de 1, **mais similar é o candidato à vaga**.

## 4. Geração do Ranking

- Calculamos a similaridade de todos os candidatos para uma vaga.
  - Ordenamos os candidatos com **maior pontuação de similaridade no topo**.
  - Exibimos os **5 melhores candidatos** com suas respectivas **pontuações de aderência**.
- 

# Sugestões

Como melhorar o rankeamento

# Como melhorar a assertividade do modelo

A **qualidade das bases e a padronização dos dados** são cruciais para o sucesso de um modelo de ranking baseado em NLP. Abaixo estão sugestões de **melhoria nas bases de candidatos e vagas**, com foco em:

- Melhorar a **precisão do ranking**
- Reduzir **ruído textual**
- Tornar os dados mais **padronizados e comparáveis**



# Priorização na padronização

A padronização de variáveis como título do cargo, regime de trabalho, nível de escolaridade, conhecimentos em linguagens de programação, formação acadêmica e certificações, bem como a uniformização das descrições das vagas e dos requisitos técnicos e comportamentais, são práticas fundamentais para garantir a consistência e integridade dos dados. Essa consistência potencializa a acurácia dos modelos de ranqueamento e otimiza a capacidade preditiva na correspondência entre candidatos e vagas.



# Aplicação da Proposta

Streamlit



# Link Streamlit

**Introdução**

**Pesquisa padronizada**

**Ranking de candidatos**

<https://fiaptechv-j4laqmgkxh53fa9d5a924c.streamlit.app/>

