FORMAÇÃO EM DADOS - DNC

*Categorização de todas as empresas do Brasil para classificação de Good Places to work*

  
TRUST WORKS

GRUPO 02

**Sumário**

[1 Business Understanding 1](#_Toc189776265)

[1.1 Objetivo do Negócio 1](#_Toc189776266)

[1.2 Avaliação da Situação 2](#_Toc189776267)

[1.3 Metas de Mineração de Dados 3](#_Toc189776268)

[1.4 Plano do Projeto 3](#_Toc189776269)

[1.5 Tarefas Principais: 4](#_Toc189776270)

[1.6 Ferramentas e Tecnologias: 4](#_Toc189776271)

[1.7 Cronograma Resumido: 4](#_Toc189776272)

[2 Data Understanding 6](#_Toc189776273)

[2.1 Coleta Inicial dos Dados 6](#_Toc189776274)

[2.2 Descrição dos Dados 6](#_Toc189776275)

[2.3 Explorando os Dados 10](#_Toc189776276)

[2.4 Qualidade dos Dados 11](#_Toc189776277)

# 1 Business Understanding

## **1.1 Objetivo do Negócio**

A Trust Works é uma plataforma dedicada a fornecer um canal seguro e anônimo para que funcionários e ex-funcionários relatem práticas abusivas em seus locais de trabalho. O objetivo central da empresa é promover a transparência corporativa e atuar como facilitador na resolução dessas questões, contribuindo para um ambiente de trabalho mais justo e saudável.

O projeto em questão visa implementar um pipeline de ETL (Extração, Transformação e Carga) para extrair, limpar e categorizar dados provenientes da Receita Federal. A estruturação desses dados permitirá à Trust Works aprimorar sua capacidade de monitoramento das empresas, proporcionando análises mais precisas e informações valiosas para stakeholders e gestores.

**Problema**

A Trust Works enfrenta o desafio de monitorar e analisar eficazmente a conduta das empresas em relação à transparência corporativa e práticas éticas. A falta de um pipeline automatizado para a coleta e tratamento de dados da Receita Federal limita a capacidade da empresa em realizar análises detalhadas e informar os interessados sobre possíveis irregularidades. Em particular, os principais problemas incluem:

1. **Inconsistência de Dados**: Dados extraídos manualmente podem conter erros e não estar formatados adequadamente.
2. **Dados Fragmentados**: Informações sobre as empresas estão dispersas em diferentes fontes, dificultando a análise consolidada.
3. **Oportunidade de Análise Atrasada**: Sem um sistema automatizado, o tempo entre a coleta de dados e a geração de relatórios pode ser prolongado, reduzindo a capacidade da Trust Works de responder rapidamente a denúncias e tendências de comportamento empresarial.

**Indicadores e Métricas de Sucesso**

Para medir o sucesso do projeto e a eficácia do pipeline de ETL, os seguintes indicadores e métricas serão utilizados:

1. **Taxa de Integração de Dados**: Percentual de dados extraídos com sucesso e incorporados à base de dados. A meta é ter uma taxa superior a 90%.
2. **Tempo de Processamento**: Tempo médio necessário para processar e limpar os dados desde a coleta até a análise. Idealmente, deve ser reduzido a menos de 24 horas.
3. **Taxa de Erros**: Número de inconsistências encontradas nos dados após o processamento em relação ao total de registros analisados. O objetivo é manter essa taxa abaixo de 5%.
4. **Acuracidade da Classificação**: Percentual de classificações de empresas que correspondem à realidade, utilizando dados de referência. Deve ser maior que 85%.
5. **Satisfação do Usuário Final**: Medida qualitativa coletada através de feedback de usuários sobre a usabilidade e eficácia dos relatórios gerados.
6. **Relatórios Gerados**: Número de relatórios e análises gerados e distribuídos para stakeholders ao longo do projeto. A meta é ter um aumento de 50% no número de relatórios gerados em comparação com o processo anterior.

**Contexto de Negócio**

A Trust Works atua em um ambiente em que transparência e ética são cada vez mais valorizadas por organizações e usuários. No cenário atual, as empresas estão sob crescente pressão para demonstrar responsabilidade social, o que inclui a gestão de denúncias sobre práticas inadequadas no local de trabalho.

A capacidade da Trust Works em oferecer dados precisos e analisados pode ser um diferencial competitivo, permitindo que a empresa:

1. **Aumente a Confiança**: Proporcione informações confiáveis para stakeholders e parceiros, melhorando a imagem da empresa no mercado.
2. **Identifique Tendências**: Utilize análises de dados para identificar padrões de comportamento problemáticos e fornecer insights práticos a empresas sobre como melhorar.
3. **Contribua para a Sociedade**: Apoie a construção de um ambiente de trabalho mais saudável e ético, contribuindo para a responsabilização corporativa.
4. **Atraia mais Clientes**: Com dados confiáveis e análises significativas, a Trust Works pode atrair mais clientes, aumentando sua atuação no mercado.

Esse contexto evidencia a necessidade de um sistema estruturado e eficiente para a coleta e análise de dados, que pode não só resolver problemas internos, mas também contribuir significativamente para um ambiente empresarial mais ético.

## **1.2 Avaliação da Situação**

Este projeto apresenta diversos desafios fundamentais, entre os quais se destacam:

* **Extração de Dados**: Obter grandes volumes de dados da Receita Federal implica em definir métodos eficientes de extração, considerando a complexidade e o tamanho dos arquivos.
* **Tratamento de Dados**: Manter a integridade dos dados, corrigindo inconsistências e lidando com valores ausentes, é crucial para garantir a confiabilidade dos resultados.
* **Classificação de Empresas**: A correta classificação de empresas por segmento econômico, utilizando códigos CNAE, é essencial para análises posteriores.

Os principais aspectos a serem considerados incluem:

* **Recursos e Ferramentas**: O projeto utilizará tecnologias como Python para automação de ETL, Pandas para manipulação de dados, Pyspark para armazenamento e consultas, e Matplotlib para visualização analítica.
* **Restrições e Desafios**: Espera-se que os dados contenham erros e valores ausentes, exigindo um processo robusto de limpeza e validação.
* **Stakeholders Envolvidos**: O projeto envolverá desenvolvedores, cientistas de dados, gestores da Trust Works e especialistas em regulação empresarial, promovendo a colaboração interfuncional.
* **Cronograma e Riscos**: Com um prazo estipulado de seis semanas e entregas parciais programadas semanalmente, os riscos identificados incluem atrasos na extração e problemas de integração entre sistemas de dados.

Os resultados esperados incluem um aumento considerável na precisão das análises realizadas e a possibilidade de fornecer dados confiáveis sobre as empresas monitoradas pela Trust Works.

## **1.3 Metas de Mineração de Dados**

As metas de mineração de dados visam transformar o fluxo de dados, agregando valor às operações da Trust Works. Dentre as metas, destacam-se:

* **Padronização e Limpeza**: Implementar processos para remover valores nulos, corrigir inconsistências e validar informações empresariais, assegurando a integridade dos dados.
* **Classificação de Empresas**: Utilizar o Código Nacional de Atividades Econômicas (CNAE) para uma segmentação eficaz das empresas.
* **Validação de Qualidade**: Estabelecer critérios rigorosos para garantir que os dados utilizados sejam precisos e de qualidade.
* **Geração de Relatórios**: Criar dashboards e visualizações em Power BI, facilitando a análise e interpretação das informações coletadas.

As métricas de sucesso incluirão a taxa de acurácia da classificação de empresas, a redução no percentual de dados inconsistentes e o tempo médio de processamento do pipeline ETL.

## **1.4 Plano do Projeto**

O desenvolvimento do projeto seguirá as etapas da metodologia CRISP-DM:

1. **Compreensão do Negócio**: Identificação e clareza sobre os objetivos da Trust Works.
2. **Compreensão dos Dados**: Coleta e análise inicial dos dados da Receita Federal.
3. **Preparação dos Dados**: Tratamento, padronização e integração dos dados.
4. **Modelagem**: Desenvolvimento de técnicas de classificação e agrupamento das empresas com base nos dados processados.
5. **Avaliação**: Validação contínua da qualidade dos dados e revisão do pipeline para garantir conformidade e eficiência.
6. **Implantação**: Entrega final do projeto com documentação completa e treinamento para os usuários finais.

## 1.5 Tarefas Principais:

As principais tarefas do projeto incluem:

* **Coleta de Dados**: Criar scripts para a extração dos dados da Receita Federal.
* **Limpeza dos Dados**: Implementar processos para remover inconsistências, lidar com valores ausentes e normalizar os dados.
* **Classificação e Categorização**: Aplicar técnicas para classificar as empresas utilizando códigos CNAE, facilitando a segmentação por setores econômicos.
* **Integração de Dados**: Consolidar dados extraídos em uma base que permite análises mais eficientes.
* **Validação de Dados**: Realizar verificações regulares para assegurar a qualidade e integridade das informações.
* **Documentação e Apresentação**: Elaborar documentação técnica e uma apresentação final para apresentar os resultados e processos para a Trust Works.

## 1.6 Ferramentas e Tecnologias:

As principais ferramentas e tecnologias que serão utilizadas no projeto incluem:

* **Python**: Para a automação do processo de ETL, utilizando bibliotecas como Pandas e NumPy.
* **GitHub**: Para controle de versão e colaboração entre os membros da equipe.
* **Jupyter Notebook**: Para desenvolvimento e teste de scripts de maneira interativa.

## 1.7 Cronograma Resumido:

**Fase 1 – Coleta e Compreensão do Problema (01/02 a 07/02):**

* Coleta de dados da Receita Federal, incluindo CNPJ, razão social, natureza jurídica e capital social.
* Reunião com stakeholders para alinhamento sobre os objetivos do projeto e entendimento das necessidades específicas da Trust Works.

**Fase 2 – Preparação dos Dados (08/02 a 21/02):**

* Limpeza dos dados, removendo inconsistências e realizando a normalização.
* Classificação das empresas utilizando Código Nacional de Atividades Econômicas (CNAE).
* Integração dos dados extraídos em uma base consolidada que facilite a análise posterior.

**Fase 3 – Modelagem e Validação (22/02 a 07/03):**

* Desenvolvimento de modelos analíticos para classificar empresas e identificar padrões relevantes.
* Validação dos modelos contra dados de referência para garantir a acuracidade dos resultados.
* Testes e simulações para avaliar a eficácia das abordagens de modelagem adotadas.

**Fase 4 – Implementação e Monitoramento (08/03 a 15/03):**

* Implantação do pipeline de ETL.
* Monitoramento contínuo da qualidade e integridade dos dados processados.

**Fase 5 – Geração de Relatórios e Apresentação Final (16/03 a 26/03):**

* Criação de um relatório detalhado com análises, insights e recomendações práticas para a Trust Works.
* Apresentação dos resultados na banca avaliadora.

# 2 Data Understanding

## **2.1 Coleta Inicial dos Dados**

Nesta fase, serão realizados esforços para garantir que todos os dados necessários estejam disponíveis e acessíveis para análise.

| **Atividades** | **Detalhes** |
| --- | --- |
| **Fontes de Dados** | Os dados foram extraídos da Receita Federal, contendo informações críticas, como CNPJ, razão social, natureza jurídica, capital social e porte da empresa, obtidas por meio da API da Receita. |
| **Métodos de Coleta** | A coleta de dados será facilitada por scripts em Python, utilizando bibliotecas como pandas e requests para automatizar a extração, bem como garantir a normalização e limpeza inicial. |
| **Critérios de Seleção** | Os dados selecionados abrangem apenas informações pertinentes às análises propostas e devem ser atualizados, respeitando um período máximo de um ano desde a última atualização. |
| **Dificuldades Encontradas** | Antecipamos desafios relacionados ao acesso a dados, como restrições de API e formatos inconsistentes. Soluções como a utilização de proxies e ajustes na formatação foram planejadas. |
| **Alternativas** | Se dados requeridos não estiverem disponíveis, consideraremos fontes alternativas, como registros estaduais, ou ajustaremos o escopo do projeto para focar exclusivamente em categorias de empresas com dados disponíveis. |

## **2.2 Descrição dos Dados**

A exploração inicial incluirá estatísticas descritivas, distribuição de CNAEs por setor, verificação de inconsistências e duplicidades.

**Dados Abertos do CNPJ**

**Empresas**

| **Campo** | **Descrição** |
| --- | --- |
| **CNPJ Básico** | Número base de inscrição no CNPJ (oito primeiros dígitos do CNPJ). |
| **Razão Social** | Nome empresarial da pessoa jurídica. |
| **Natureza Jurídica** | Código da natureza jurídica. |
| **Qualificação do Responsável** | Qualificação da pessoa física responsável pela empresa. |
| **Capital Social da Empresa** | Capital social da empresa. |
| **Porte da Empresa** | Código do porte da empresa: <br> 00 – Não informado <br> 01 - Micro empresa <br> 03 - Empresa de pequeno porte <br> 05 - Demais |
| **Ente Federativo Responsável** | O ente federativo responsável é preenchido para os casos de órgãos e entidades do grupo de natureza jurídica 1XXX; demais naturezas ficam em branco. |

**Estabelecimentos**

| **Campo** | **Descrição** |
| --- | --- |
| **CNPJ Básico** | Número base de inscrição no CNPJ (oito primeiros dígitos do CNPJ). |
| **CNPJ Ordem** | Número do estabelecimento de inscrição no CNPJ (do nono até o décimo segundo dígito do CNPJ). |
| **CNPJ DV** | Dígito verificador do número de inscrição no CNPJ (dois últimos dígitos do CNPJ). |
| **Identificador Matriz/Filial** | Código do identificador matriz/filial: <br> 1 – Matriz <br> 2 – Filial |
| **Nome Fantasia** | Nome fantasia da empresa. |
| **Situação Cadastral** | Código da situação cadastral: <br> 01 – Nula <br> 02 – Ativa <br> 03 – Suspensa <br> 04 – Inapta <br> 08 – Baixada |
| **Data Situação Cadastral** | Data do evento da situação cadastral. |
| **Motivo Situação Cadastral** | Código do motivo da situação cadastral. |
| **Nome da Cidade no Exterior** | Nome da cidade no exterior, se aplicável. |
| **País** | Código do país. |
| **Data de Início Atividade** | Data de início da atividade da empresa. |
| **CNAE Fiscal Principal** | Código da atividade econômica principal do estabelecimento. |
| **CNAE Fiscal Secundária** | Código da(s) atividade(s) econômica(s) secundária(s) do estabelecimento. |
| **Tipo de Logradouro** | Descrição do tipo de logradouro onde se localiza o estabelecimento. |
| **Logradouro** | Nome do logradouro onde se localiza o estabelecimento. |
| **Número** | Número do local onde se encontra o estabelecimento. Com um preenchimento "S/N" se não houver número. |
| **Complemento** | Complemento para o endereço de localização do estabelecimento. |
| **Bairro** | Bairro onde se localiza o estabelecimento. |
| **CEP** | Código de endereçamento postal referente ao logradouro onde está localizado o estabelecimento. |
| **UF** | Sigla da unidade da federação em que se encontra o estabelecimento. |
| **Município** | Código do município de jurisdição onde se encontra o estabelecimento. |
| **DDD 1** | Contém o DDD 1 do telefone. |
| **Telefone 1** | Contém o número do telefone 1. |
| **DDD 2** | Contém o DDD 2 do telefone, se aplicável. |
| **Telefone 2** | Contém o número do telefone 2, se aplicável. |
| **DDD do Fax** | Contém o DDD do fax, se aplicável. |
| **Fax** | Contém o número do fax, se aplicável. |
| **Correio Eletrônico** | Contém o e-mail do contribuinte. |
| **Situação Especial** | Situação especial da empresa, se aplicável. |
| **Data da Situação Especial** | Data em que a empresa entrou em situação especial. |

**Dados do Simples**

| **Campo** | **Descrição** |
| --- | --- |
| **CNPJ Básico** | Número base de inscrição no CNPJ (oito primeiros dígitos do CNPJ). |
| **Opção pelo Simples** | Indicador da existência da opção pelo Simples: <br> S - SIM <br> N - NÃO <br> Em branco – OUTROS |
| **Data de Opção pelo Simples** | Data de opção pelo Simples. |
| **Data de Exclusão do Simples** | Data de exclusão do Simples, se aplicável. |
| **Opção pelo MEI** | Indicador da existência da opção pelo MEI: <br> S - SIM <br> N - NÃO <br> Em branco - OUTROS |
| **Data de Opção pelo MEI** | Data de opção pelo MEI, se aplicável. |
| **Data de Exclusão do MEI** | Data de exclusão do MEI, se aplicável. |

**Sócios**

| **Campo** | **Descrição** |
| --- | --- |
| **CNPJ Básico** | Número base de inscrição no CNPJ (cadastro pessoa jurídica). |
| **Identificador de Sócio** | Código do identificador de sócio: <br> 1 – Pessoa Jurídica <br> 2 – Pessoa Física <br> 3 – Estrangeiro |
| **Nome do Sócio** | Nome do sócio pessoa física ou a razão social e/ou nome. |
| **CNPJ/CPF do Sócio** | CPF ou CNPJ do sócio (sócio estrangeiro não tem esta informação). |
| **Qualificação do Sócio** | Código da qualificação do sócio. |
| **Data de Entrada** | Data de entrada na sociedade. |
| **País** | Código do país do sócio estrangeiro, se aplicável. |
| **Representante Legal** | Número do CPF do representante legal. |
| **Nome do Representante** | Nome do representante legal. |
| **Qualificação do Representante Legal** | Código da qualificação do representante legal. |
| **Faixa Etária** | Código correspondente à faixa etária do sócio. |

## **2.3 Explorando os Dados**

A lista de ações para tratamento de dados inclui a imputação de dados faltantes, exclusão de outliers e padronização de formatos. Uma estratégia clara foi desenvolvida para abordar esses problemas:

**Revisão Contínua**: Um protocolo de monitoramento da qualidade dos dados será estabelecido para garantir a integridade dos dados durante todo o projeto.

**Formato do Arquivo e Regras de Manipulação**

1. **Formato do Arquivo**:
   * O arquivo deve seguir o padrão de carga automática em RDBMS. Utilizar ponto e vírgula (;) como separador de atributos.
2. **Descaracterização de Dados**:
   * Os campos **169 (CNPJ/CPF DO SÓCIO)** e **271 (CNPJ/CPF DO REPRESENTANTE)** devem ser descaracterizados:
     + Ocultar os três primeiros dígitos e os dois dígitos verificadores do CPF, conforme o art. 129 § 2° da Lei no 13.473/2017 (LDO 2018).
3. **Campo Ente Federativo Responsável - EFR**:
   * Preenchido para órgãos e entidades do grupo de natureza jurídica **1XX**. Para as demais naturezas, deve ficar em branco.
   * Exemplos para o arquivo final: **UNIÃO**, **DISTRITO FEDERAL**, **BAHIA**. Para municípios, incluir a sigla da UF: **SÃO PAULO – SP**, **BELO HORIZONTE – MG**.
4. **Campo CNAE Fiscal Secundária**:
   * Preencher com cada ocorrência, separadas por vírgula, em caso de múltiplas atividades.

## **2.4 Qualidade dos Dados**

A qualidade dos dados é crucial para garantir a precisão e a confiabilidade das análises, sendo fundamental para o sucesso do projeto. Durante o processo de exploração, foram identificadas algumas lacunas e inconsistências nos dados, que devem ser tratadas para proporcionar uma análise mais robusta.

Várias medidas serão implementadas para assegurar que os dados mantenham altos padrões de qualidade. As principais ações incluem:

1. **Validação de Dados**:
   * Implementar regras de validação durante a entrada e processamento de dados para garantir que os dados atendam a critérios específicos de precisão e formato.
2. **Tratamento de Dados Faltantes**:
   * Para entradas com dados ausentes, usar métodos estatísticos para imputação, quando apropriado, ou remoção de registros incompletos se a falta de dados for crítica para a análise.
3. **Identificação de Outliers**:
   * Utilizar técnicas estatísticas para detectar e revisar outliers que possam indicar erros de entrada ou eventos incomuns nos dados.
4. **Padronização de Formatos**:
   * Garantir que todos os dados estejam em formatos consistentes, especialmente para campos críticos como datas e identificadores, para facilitar a integração e análise.
5. **Revisão Contínua**:
   * Estabelecer um protocolo de monitoramento da qualidade dos dados, permitindo ajustes e correções contínuas ao longo do projeto.
6. **Treinamento de Equipe**:
   * Oferecer treinamento contínuo para a equipe envolvida na entrada e manipulação de dados, para aumentar a conscientização sobre a importância da qualidade dos dados e as melhores práticas.