Mapeamento do Ecossistema de Inovação em Minas Gerais

Renato Silva Santos

29 de junho de 2025

Sumário

1 Introdução

O estado de Minas Gerais possui um ecossistema de inovação dinâmico e de grande relevância para o desenvolvimento tecnológico e econômico do Brasil. Contudo, a articulação eficiente entre os diversos atores, como empresas, startups, pesquisadores, Instituições de Ciência e Tecnologia (ICTs) e órgãos de fomento, é frequentemente dificultada pela ausência de uma fonte de dados centralizada, estruturada e classificada.

Atualmente, a identificação de organizações e pesquisadores para iniciativas, como eventos de prospecção tecnológica e formação de parcerias, depende de um processo predominantemente manual e ineficiente. Isso acarreta consequências significativas, como a perda de oportunidades de colaboração com agentes de ponta que não possuem visibilidade, a alocação de recursos em tarefas de pesquisa manual e a dificuldade em alinhar as competências existentes no estado com as missões prioritárias de políticas públicas, a exemplo da "Nova Indústria Brasil"(NIB). A carência de dados estruturados limita, ainda, a capacidade de realizar análises macro sobre as vocações tecnológicas do estado, prejudicando a formulação de estratégias de desenvolvimento e a atração de novos investimentos.

Este projeto visa solucionar tal problema por meio da construção de um pipeline de coleta, tratamento e classificação de dados sobre o ecossistema de inovação mineiro. O objetivo central é transformar o processo de prospecção e análise, de um modelo manual, para uma abordagem mais automatizada e inteligente, gerando valor não apenas para instituições específicas, como o BH-TEC, mas para toda a comunidade de inovação em Minas Gerais.

1.1 Objetivos

Para atender à necessidade exposta, este trabalho define os seguintes objetivos principais:

- **Coleta e Estruturação de Dados:** Realizar a coleta de dados públicos sobre organizações de Pesquisa, Desenvolvimento e Inovação (PD&I) a partir de fontes heterogêneas, com destaque para o Sistema Mineiro de Inovação (SIMI) e a API da *Semantic Scholar*. O trabalho inclui a estruturação e limpeza de dados originalmente não padronizados.

- **Desenvolvimento de Modelo de Classificação Inteligente:** Construir e avaliar um modelo de Aprendizado de Máquina, utilizando técnicas de Processamento de Linguagem Natural (PLN), para classificar automaticamente as organizações com base em suas descrições textuais. A classificação será alinhada às 6 missões estratégicas da política "Nova Indústria Brasil", permitindo a identificação de competências e potenciais sinergias no ecossistema.

2 Coleta e Estruturação de Dados

A primeira etapa do projeto consistiu na construção de um dataset. Este processo envolveu a extração, a normalização inicial, um pré-processamento e a segmentação dos dados para garantir a qualidade do modelo de classificação.

2.1 Extração e Normalização Inicial dos Dados

A principal fonte de dados para obtenção das organizações foi o **Sistema Mineiro de Inovação (SIMI)**. Por meio da análise de requisições de rede do portal, foi possível extrair um conjunto de dados brutos contendo **2.448 registros** de organizações. O formato original dos dados se apresentava como um array de objetos JavaScript, em que os atributos eram perguntas longas vindas de um formulário encontrado no site.

Para facilitar a manipulação computacional em Python, foi realizada uma normalização inicial. Os nomes dos atributos foram convertidos para um padrão conciso e padronizado (*snake_case*). Foi criado um dicionário de dados para mapear os novos nomes às suas descrições originais. Este dicionário está detalhado no **Apêndice A**.

2.2 Pré-processamento e Limpeza

A análise exploratória revelou que os dados coletados via formulários apresentavam um alto grau de desestruturação, com 33 colunas. Para mitigar as inconsistências, as seguintes tarefas de limpeza foram executadas:

- **Tratamento de Valores Nulos:** As entradas que representavam ausência de informação eram strings vazias ou com espaços, que foram convertidas para `NaN`.
- **Remoção de Registros Críticos:** Foram descartados 17 registros que possuíam o campo `descricao_organizacao` nulo. A remoção foi justificada pela criticidade deste campo, que serve como principal fonte de informação textual para o modelo de PLN. Registros com valores nulos em colunas não essenciais para a classificação foram mantidos.
- **Seleção de Atributos:** Colunas que continham metadados, informações de contato ou dados não relevantes para a classificação de competências tecnológicas foram removidas. A lista de colunas descartadas inclui:
  - `email_contato`
  - `site_organizacao`
  - `logo_url`
  - `linkedin_url`
  - `modelo_negocio`
  - `cadastro_aprovado`
  - `autorizacao_divulgacao`
  - `numero_funcionarios`
  - `uf_sede`
  - `cidade_sede`

- **Padronização de Variáveis Categóricas:** Foi realizado um agrupamento de valores em campos de texto livre com alta cardinalidade. Por exemplo, na coluna `tipo_organizacao` entradas como "Capacitação e qualificação profissional" e "Capacitação e Qualificação Profissional" foram normalizadas para uma única categoria.

2.3 Segmentação Estratégica do Ecossistema

Para permitir uma análise mais focada, como a base de dados incluia atributos que faziam sentido apenas para cada categoria de organização, as organizações foram agrupadas em segmentos estratégicos com base no campo `categoria_organizacao`. O mapeamento realizado foi o seguinte:

- **Executores Diretos das Missões:** Startups, Empresas de Base Tecnológica (EBTs) e Médias/Grandes Empresas.
- **Base de Conhecimento e P&D:** ICTs, IES e Centros de P&D.
- **Estruturadores e Multiplicadores:** Incubadoras, Aceleradoras, Pré-aceleradoras e Venture Builders.
- **Financiadores do Ecossistema:** Fundos de Investimento e Investidores Anjo.
- **Articuladores e Fomentadores Públicos:** Governos Municipais.
- **Demais Organizações:** Categoria residual.

Essa segmentação resultou na distribuição quantitativa apresentada na Tabela 1. Dessa forma, dividimos a base de dados em outras 6 bases que podem ser classificadas cada uma com um modelo ou rotuladas manualmente.

Tabela 1: Distribuição das organizações por papel no ecossistema.

Segmento	**Nº de Organizações**
Executores Diretos das Missões	1530
Base de Conhecimento e P&D	489
Demais Organizações	298
Estruturadores e Multiplicadores do Ecossistema	70
Articuladores e Fomentadores Públicos	32
Financiadores do Ecossistema	12
Total	**2431**

3 Metodologia do Modelo de Classificação

Com o dataset tratado e segmentado, a etapa seguinte focou no desenvolvimento do modelo de classificação multiclasse. A metodologia foi concentrada no subconjunto "Executores Diretos das Missões", por ser o maior e mais relevante grupo para o objetivo proposto, afim de validar a metodologia do projeto.

3.1 Definição do Conjunto de Dados de Foco

O dataset "Executores Diretos das Missões", composto inicialmente por 1.530 registros, passou por um processo adicional de limpeza. Para a tarefa de classificação, foram mantidos apenas os seguintes atributos: `tecnologias_disruptivas`, `fase_negocio`, `nome_organizacao`, `descricao_organizacao`, `categoria_organizacao` e `segmento_atuacao`. Adicionalmente, foram removidas 145 empresas cujo campo `segmento_atuacao` indicava um nicho de atuação excessivamente específico, além de que essas instâncias apresentavam um preenchimento inadequado do formulário. Após estas etapas, o conjunto de dados final para a modelagem foi consolidado em **1.385 organizações**.

3.2 Estratégia de Rotulação de Dados

A criação de um conjunto de dados rotulado para treinar e avaliar o modelo supervisionado foi realizada em duas fases.

3.2.1 Fase 1: Heurística e Amostragem

Inicialmente, aplicou-se uma heurística baseada em palavras-chave para gerar rótulos preliminares. A heurística consistia em classificar cada organização de acordo com palavras-chave encontradas nos atributos `tecnologias_disruptivas`, `descricao_organizacao` e `segmento_atuacao`. Essa abordagem permitiu uma primeira análise da distribuição das organizações entre as missões da NIB (Tabela 2), que serviu como base para a etapa de amostragem estratificada para a posterior rotulação manual.

Tabela 2: Distribuição heurística das organizações por missão.

Missão (Rótulo Heurístico)	**Nº de Organizações**
M4 Transformação Digital	464
M1 Agro	262
Não Classificado	215
M2 Saúde	203
M3 Infraestrutura e Mobilidade	203
M5 Bioeconomia e Energia	22
M6 Defesa e Soberania	16

3.2.2 Fase 2: Rotulação Manual e Aumento de Dados

A partir da estratificação, uma amostra inicial de 300 registros foi extraída para rotulação manual, baseando-se na distribuição dos rótulos encontrados pela heurística. Como os primeiros experimentos indicaram que o desbalanceamento das classes prejudicava o desempenho, o conjunto de dados foi enriquecido com mais 75 exemplos de classes estratégicas:

27 da missão M1 (Agro), 30 da M2 (Saúde) e 18 da M5 (Bioeconomia e Energia). Utilizando a ferramenta de anotação **Doccano**, este processo resultou em um dataset com **375 registros** rotulados manualmente, que serviu como base para o treinamento e a avaliação dos modelos. A distribuição final das amostras rotuladas é apresentada na Tabela 3. Devido à ausência de amostras, a Missão 6 (Defesa e Soberania) foi temporariamente ignorada na etapa de modelagem.

Tabela 3: Distribuição final do conjunto de dados rotulado manualmente.

Missão (Rótulo Manual)	**Nº de Amostras**
M4 Transformação Digital	101
M1 Agro	84
M2 Saúde	74
Não Classificado	47
M3 Infraestrutura e Mobilidade	44
M5 Bioeconomia e Energia	25
M6 Defesa e Soberania	0

3.3 Arquitetura dos Modelos

Foram exploradas duas abordagens distintas para a tarefa de classificação textual multilabel, ambas implementadas como um pipeline da biblioteca Scikit-learn. Os parâmetros de cada modelo, como os hiperparâmetros do vetorizador e do classificador, foram definidos por meio de experimentação visando o melhor equilíbrio entre as métricas de avaliação. A primeira abordagem foi implementada como um *pipeline* integrado da biblioteca Scikit-learn, enquanto a segunda envolveu um processo sequencial de geração de *embeddings* seguido pela aplicação de um classificador.

3.3.1 Pré-processamento Textual para Modelagem

Antes da vetorização, os principais atributos textuais de cada organização (`nome_organizacao`, `descricao_organizacao`, `segmento_atuacao` e `tecnologias_disruptivas`) foram concatenados em um único documento. Essa unificação criou um campo de texto para servir como entrada para os modelos.

3.3.2 Modelo 1 (Baseline): TF-IDF com Regressão Logística

Como primeira abordagem que posteriormente se transformou em baseline, foi implementado um modelo utilizando a representação **TF-IDF**. O vetorizador foi configurado com os seguintes parâmetros:

- **Stop words:** Utilização de uma lista de stop words do português (via NLTK) para remover termos comuns sem valor semântico.
- **N-gramas:** Uso de unigramas e bigramas (`ngram_range=(1, 2)`) para capturar não apenas palavras isoladas, mas também pequenas frases e termos compostos.
- **Tamanho do Vocabulário:** Limitação a um máximo de 3.000 características (*features*) para controlar a dimensionalidade do problema.

Sobre essa representação vetorial, foi treinado um classificador de **Regressão Logística** encapsulado por um `MultiOutputClassifier`, permitindo uma abordagem de um classificador por rótulo. Crucialmente, o parâmetro `class_weight='balanced'` foi utilizado para que o algoritmo de treinamento penalizasse mais os erros nas classes minoritárias, combatendo o desbalanceamento de dados observado.

3.3.3 Modelo 2 (Principal): Embeddings Semânticos com Regressão Logística

A abordagem principal visou capturar o significado contextual dos textos. Diferente do modelo de *baseline*, o processo foi executado em duas etapas sequenciais:

1. **Geração de Embeddings:** Inicialmente, todos os textos do conjunto de dados de 375 amostras foram processados pelo modelo pré-treinado `paraphrase-multilingual-mpnet-base-v2`, da biblioteca *Sentence Transformers*. Cada documento foi transformado em um vetor de características (embedding), resultando em uma matriz numérica de representações semânticas.

2. **Treinamento do Classificador:** Esta matriz de *embeddings* foi então utilizada como a matriz de *features* de entrada (X). Apenas nesta etapa os dados (embeddings e rótulos) foram divididos em conjuntos de treino e teste. O mesmo classificador de **Regressão Logística** (`MultiOutputClassifier` com `class_weight='balanced'`) foi treinado sobre os *embeddings*.

4 Resultados Parciais

Os modelos foram avaliados no conjunto de teste (20% do dataset rotulado, ou 75 amostras), utilizando as métricas de Precisão, Revocação (*Recall*) e F1-Score por classe, além das médias gerais (macro e ponderada).

4.1 Métricas de Avaliação dos Modelos

Os resultados detalhados para o modelo de *baseline* (TF-IDF) e para o modelo principal (Embeddings) são apresentados nas Tabelas 4 e 5, respectivamente.

Tabela 4: Relatório de Classificação do Modelo 1 (TF-IDF).

Missão	**Precision**	**Recall**	**F1-Score**	**Support**
M1 Agro	1.00	0.61	0.76	23
M2 Saúde	1.00	0.92	0.96	13
M3 Infra. e Mobilidade	1.00	0.50	0.67	12
M4 Transf. Digital	0.76	0.93	0.83	40
M5 Bioeconomia e Energia	0.80	0.50	0.62	16
M6 Defesa e Soberania	0.00	0.00	0.00	1
Macro Avg	**0.76**	**0.58**	**0.64**	**105**
Weighted Avg	**0.87**	**0.73**	**0.77**	**105**

Tabela 5: Relatório de Classificação do Modelo 2 (Embeddings).

Missão	**Precision**	**Recall**	**F1-Score**	**Support**
M1 Agro	0.86	0.78	0.82	23
M2 Saúde	0.77	0.77	0.77	13
M3 Infra. e Mobilidade	0.53	0.67	0.59	12
M4 Transf. Digital	0.85	0.82	0.84	40
M5 Bioeconomia e Energia	0.57	0.75	0.65	16
M6 Defesa e Soberania	0.20	1.00	0.33	1
Macro Avg	**0.63**	**0.80**	**0.67**	**105**
Weighted Avg	**0.76**	**0.78**	**0.76**	**105**

4.2 Análise Comparativa

A análise dos resultados revela um *trade-off* claro entre as duas abordagens. O **modelo TF-IDF** apresentou alta precisão em diversas classes (1,00 para Agro, Saúde e Infraestrutura), indicando que, quando realiza uma previsão positiva, ela tende a ser correta. No entanto, sua revocação média macro foi baixa (0,58), o que revela que o modelo deixou de identificar uma quantidade significativa de organizações pertencentes a essas categorias. Isso sugere uma baixa capacidade de generalização, possivelmente devido à memorização de padrões muito específicos do conjunto de treinamento. Dessa forma, o desempenho

do modelo se aproxima ao da heurística baseada em palavras-chave, o que limita sua aplicação em cenários mais diversos.

Por outro lado, o **modelo de Embeddings** apresentou um comportamento muito mais equilibrado e desejável para o problema em questão. Embora sua precisão média tenha sido inferior, a revocação foi significativamente superior (0,80 na média macro). Isso demonstra uma capacidade muito maior de identificar corretamente as organizações de todas as classes, incluindo as minoritárias.

Uma análise detalhada mostra que o modelo teve desempenho sólido em classes com maior quantidade de dados, como *Transformação Digital* (F1 = 0,84) e *Agro* (F1 = 0,82). Além disso, mesmo classes intermediárias, como *Bioeconomia e Energia*, apresentaram revocação satisfatória (0,75), o que indica que o modelo está conseguindo capturar padrões sem depender apenas de volume de dados. No entanto, observou-se uma discrepância entre precisão e revocação (0,63 vs. 0,80 na média macro), sugerindo que o modelo tende a fazer mais previsões positivas do que o ideal, o que leva a um número maior de falsos positivos. Ainda assim, para o objetivo do projeto, essa característica não é tão crítica, já que o custo de uma omissão (falso negativo) é mais alto do que o de uma classificação excessiva (falso positivo).

5 Conclusão e Trabalhos Futuros

Este relatório documentou as etapas iniciais do projeto de mapeamento do ecossistema de inovação de Minas Gerais, desde a coleta e tratamento dos dados até o desenvolvimento e avaliação de modelos de classificação.

5.1 Conclusões Parciais

O trabalho resultou na construção de um pipeline de dados funcional, capaz de extrair, limpar, segmentar e, finalmente, classificar organizações do ecossistema de inovação. A principal conclusão técnica é a superioridade da abordagem baseada em *embeddings* semânticos para esta tarefa. O modelo demonstrou uma capacidade de generalização significativamente maior em comparação com o *baseline* TF-IDF, alcançando um recall médio macro de 0,80, o que é crucial para o objetivo de minimizar a omissão de atores relevantes. A metodologia de rotulação, combinando heurísticas, amostragem estratificada e aumento de dados, mostrou-se eficaz para criar um dataset de treinamento a partir de dados brutos e desestruturados.

5.2 Limitações e Trabalhos Futuros

Apesar dos resultados promissores, o trabalho atual possui limitações. A modelagem focou-se exclusivamente no segmento de "Executores Diretos das Missões", e a classe "Defesa e Soberania"(M6) ainda carece de dados suficientes para um treinamento eficaz.

Dessa forma, os próximos passos do projeto seguirão duas frentes principais:

- **Expansão da Análise:** Aplicar a metodologia de classificação desenvolvida para outros dois *datasets*: "Base de Conhecimento e P&D"e "Estruturadores e Multiplicadores". Os demais conjuntos de dados serão rotulados manualmente por terem poucas instâncias.

- **Mapeamento de Pesquisadores:** Iniciar a segunda grande fase do projeto, que consiste em dar continuidade à coleta e análise de dados de pesquisadores via API do *Semantic Scholar*. O objetivo será filtrar pesquisadores relevantes das principais universidades de Minas Gerais e, eventualmente, conectá-los a organizações que façam sentido de acordo com sua área de pesquisa.
- **Desenvolvimento de Ferramenta de Visualização:** Como objetivo final, planeja-se a construção de uma aplicação ou *dashboard* interativo para consultar e visualizar os dados mapeados, tornando os resultados acessíveis para os stakeholders do ecossistema de inovação.

A Apêndice: Dicionário de Dados

Tabela 6: Mapeamento e descrição dos atributos normalizados.

Coluna (snake_case)	Descrição Original
cadastro_aprovado	Indica se o cadastro foi aprovado pela equipe do SIMI.
autorizacao_divulgacao	Confirma se a organização autorizou a divulgação pública de suas informações.
nome_organizacao	O nome oficial ou fantasia da organização.
cidade_sede	A cidade onde a sede principal da organização está localizada.
uf_sede	A Unidade Federativa (Estado) da sede da organização.
logo_url	O link (URL) para a imagem do logotipo da organização.
descricao_organizacao	Texto descritivo sobre as atividades, missão e propósito da organização.
site_organizacao	O endereço do site oficial da organização.
email_contato	O principal e-mail de contato fornecido pela organização.
categoria_organizacao	A principal categoria em que a organização se enquadra (ex: Startup, Governo, ICT).
fase_investimento_busca	Qual estágio de desenvolvimento das startups/empresas que o programa trabalha? (válido para Pré-aceleradora — Aceleradora — Venture Builder)
incentivos_fiscais_municipio	Descreve os incentivos fiscais que um município oferece para empresas de tecnologia.
orgao_responsavel_cti	O departamento responsável pela pauta de Ciência, Tecnologia e Inovação (CTI).
tipo_organizacao	Uma sub-categoria ou tipo mais específico da organização.
contato_divulgacao_cientifica	O contato do setor ou pessoa responsável pela divulgação científica.
possui_nit	Informa se a organização possui um NIT (Núcleo de Inovação Tecnológica).
areas_expertise	As principais áreas do conhecimento ou competências técnicas da organização.
tese_investimento	Descreve os critérios e o foco de um investidor para realizar investimentos.
num_startups_investidas	O número de startups no portfólio de um investidor ou aceleradora.
segmento_atuacao	O principal mercado ou setor de atuação da organização (ex: Saúde, Agro).
programas_desenvolvimento	Descreve os programas (de aceleração, fomento) oferecidos pela organização.
modelo_negocio	Como a organização gera receita ou entrega valor (ex: SaaS, Marketplace).
tecnologias_disruptivas	As principais tecnologias que a organização desenvolve ou utiliza (ex: IA, IoT).
segmentacao_clientes	O público-alvo principal da organização (ex: B2B, B2C).

Tabela 6 – continuação

Coluna (snake_case)	Descrição Original
fase_negocio	O estágio atual de desenvolvimento do negócio (ex: Ideação, Tração).
numero_funcionarios	O tamanho da equipe da organização, geralmente apresentado em faixas.
fase_investimento_captada	O estágio de investimento mais recente que a organização já recebeu.
busca_investimento	Indica se a organização está ativamente procurando por investimento (”Sim”ou ”Não”).
linkedin_url	O link para a página oficial da organização no LinkedIn.
preferencia_segmento	Indica a preferência de um investidor ou programa por um segmento de atuação.
beneficios_programa	Os benefícios específicos oferecidos por um programa de desenvolvimento.
investimento_medio_startup	O valor médio (check-size) que um investidor aporta por startup.
exige_equity	Informa se um programa exige participação acionária (equity) da startup.