

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Empresas DNA USP

*Análise e monitoramento de
dados sobre o empreendedorismo
na Universidade de São Paulo*

Daniel Silva Nunes
Lucas Toshio Loschner Fujiwara

MONOGRAFIA FINAL

MAC 499 — TRABALHO DE
FORMATURA SUPERVISIONADO

Supervisor: Prof. Dr. Alfredo Goldman
Cossupervisora: Prof^ª Geciane Silveira Porto

20 de Dezembro de 2021

Agradeço a toda a minha família, a todos os professores que fizeram parte dessa jornada — aqueles do SESI, do SENAI, do Cursinho da Poli e da USP — e a todas as pessoas que de alguma forma fizeram parte de tudo o que aconteceu até aqui. Muito obrigado.

(Daniel Silva Nunes)

A meus pais e professores, que tornaram tudo isso possível, e ao meu irmão, André, quem espero que esteja dedicando uma monografia para mim daqui alguns anos.
(Lucas Toshio Loschner Fujiwara)

Resumo

Daniel Silva Nunes, Lucas Toshio Loschner Fujiwara. **Empresas DNA USP: Análise e monitoramento de dados sobre o empreendedorismo na Universidade de São Paulo.** Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2021.

Empresas com DNA USP são empreendimentos que estão de alguma forma relacionados com a Universidade de São Paulo (USP), seja por serem idealizados, fundados ou mantidos por pessoas que fazem ou fizeram parte da comunidade, ou por serem organizações que tenham recebido alguma categoria de incentivo da instituição. De modo a identificar e promover empresas com tais características, a Agência USP de Inovação (AUSPIN) criou o selo DNA USP e disponibilizou uma interface de cadastro com o intuito de coletar dados dessas companhias. Neste trabalho dados coletados pela AUSPIN foram tratados e analisados para trazer reflexões acerca da prática empreendedora no contexto da universidade. A partir das reflexões formadas, foi desenvolvido um sistema para centralizar operações de validação e padronização dos dados e oferecer acesso a relatórios interativos gerados sob demanda por meio de uma aplicação *web*.

Palavras-chave: empreendedorismo. inovação. extensão. sociedade. ciência de dados. engenharia de software. DNA USP.

Abstract

Daniel Silva Nunes, Lucas Toshio Loschner Fujiwara. **DNA USP Companies: *Analysis and monitoring of data on entrepreneurship at the University of São Paulo***. Capstone Project Report (Bachelor). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2021.

Companies with USP DNA are enterprises that are somehow related to the University of São Paulo (USP), either because they are idealized, founded or maintained by people who are or were part of the university, or because they are organizations that have received some category of incentive of the institution. In order to identify and promote companies with such characteristics, the USP Innovation Agency (AUSPIN) created the USP DNA badge and made available a registration interface in order to collect data from these companies. In this work, data collected by AUSPIN were treated and analyzed to yield insights on entrepreneurial practice in the context of the university. From the insights found, a software was developed to centralize data validation and standardization operations and provide access to interactive reports generated on demand through a web application.

Keywords: entrepreneurship. innovation. extension. society. data science. software engineering. USP DNA.

Lista de Abreviaturas

AUSPIN	Agência USP de Inovação
IBGE	Instituto Brasileiro de Geografia e Estatística
FN	Forma Normal
ICMS	Imposto sobre Circulação de Mercadorias e Serviços
NEU	Núcleo de Empreendedorismo da USP
CSV	<i>Comma-separated values</i>
CNAE	Classificação Nacional de Atividades Econômicas
DER	Diagrama Entidade-Relacionamento
CONCLA	Comissão Nacional de Classificação
API	<i>Application Programming Interface</i>
JSON	<i>JavaScript Object Notation</i>
INPI	Instituto Nacional da Propriedade Industrial
CQRS	<i>Command Query Responsibility Segregation</i>
CI	<i>Continuous Integration</i>
CD	<i>Continuous Delivery</i>
NPM	<i>Node Package Manager</i>

Lista de Figuras

2.1	Taxas percentuais de empreendedorismo segundo estágio do empreendimento (inicial ou estabelecido)	6
3.1	Exemplos de passos executados no fluxo de Continuous Integration do projeto dnausp-core	12
4.1	Formulário de cadastro de empresas - HUB USP Inovação	16
4.2	Adicionando múltiplos sócios - HUB USP Inovação	18
4.3	Diagrama entidade-relacionamento do novo modelo de dados proposto	19
5.1	Gráfico com a proporção de áreas de atuação dentre as empresas DNA USP.	22
5.2	Distribuição de empresas de 1960 a 2000	23
5.3	Distribuição de empresas de 2001 a 2021	23
5.4	Seção do formulário de empresas DNA USP com seleção de institutos	25
5.5	Percentual de empresas em cada instituto categorizadas por área de atuação	26
5.6	Gráfico de Setores — Área de Atuação por Instituto — Atividades Profissionais, Científicas e Técnicas	26
5.7	Gráfico de Setores — Área de Atuação por Instituto — Comércio e Serviços	27
5.8	Gráfico de Setores — Área de Atuação por Instituto — Educação, Artes e Esportes	28
5.9	Gráfico de Setores — Área de Atuação por Instituto — Saúde e Serviços Sociais	28
5.10	Gráfico de Setores — Distribuição de empresas DNA USP com mulheres e homens no quadro de sócios	31
5.11	Distribuição de empresas DNA USP com mulheres e homens no quadro de sócios por instituto	32
5.12	Distribuição percentual de mulheres e homens nas empresas agrupadas por unidade	34

5.13	Distribuição de homens e mulheres em empresas que exercem Atividades Profissionais, Científicas e Técnicas.	35
5.14	Distribuição feminina e masculina por CNAE	36
5.15	Mapa interativo exibindo a distribuição de empresas no território brasileiro	37
5.16	Empresas na região sudeste	38
5.17	Empresas incubadas ou graduadas nas demais regiões	38
5.18	Empresas que foram direto para o mercado	38
5.19	Proporção de empresas com sócios vinculados a programas de pós-graduação	39
5.20	Distribuição de CNAE de empresas com sócios vinculados a pós-graduação	40
5.21	Percentual de sócios com pós-graduação por unidade da USP vinculada .	40
5.22	Proporção de empresas com patentes registradas por CNAE	41
5.23	Proporção de empresas com patentes registradas por instituto	42
6.1	Exemplo de visualização da versão inicial.	44
6.2	Pacote <i>core</i> publicado no repositório NPM	45
6.3	Visualização dos testes realizados com Jest	46
6.4	Visualização do relatório de cobertura do módulo <i>core</i> produzido pela ferramenta Jest	47
6.5	Visualização da execução de testes, checagem de tipos e qualidade de código através do painel <i>GitHub Actions</i> do <i>GitHub</i>	47
6.6	Diagrama de classes gerado pelo ambiente de desenvolvimento integrado WebStorm	49
6.7	Diagrama das classes que agregam fábricas gerado pelo ambiente de desenvolvimento integrado WebStorm	50
6.8	Descrição da Arquitura Limpa em imagem, por Robert C. Martin, em ROBERT MARTIN, 2012	51
6.9	Visualização da execução da construção da imagem <i>OCI</i> do módulo <i>WebAPI</i> no painel <i>GitHub Actions</i> do <i>GitHub</i>	53
6.10	Diagrama do fluxo de execução de uma requisição HTTP	54
6.11	Painel de controle da plataforma <i>Vercel</i> , exibindo a última versão disponibilizada ao cliente.	56
6.12	Tela inicial com <i>menu</i> de navegação.	57
6.13	Tela para mapeamento dos dados das planilhas e envio para o servidor. .	58
6.14	Tela para mapeamento dos dados das planilhas e envio para o servidor, aba de listagem erros de mapeamento para direcionar o usuário para tratá-los.	59
6.15	Tela de exibição de dados, listagem de gráficos.	60
6.16	Tela de exibição de dados, gráfico de distribuição de atividades econômicas das empresas DNA USP.	61

6.17	Tela de exibição de dados, gráfico de distribuição de atividades principais das empresas com maior granularidade.	62
6.18	Tela de exibição de dados, seção de filtros.	63
6.19	Página para consulta de dados de um dado CNPJ, implementada por demanda do cliente.	64

Lista de Tabelas

2.1	Percentual dos empreendedores segundo motivação para iniciar um novo negócio - Brasil, 2019. Fonte: GEM Brasil 2019	7
5.1	Total de mulheres e homens nos cinco institutos com mais empresas DNA USP cadastradas	32
5.2	Dez principais depositantes de patentes de invenção, segundo relatório do INPI em 2020.	41

Sumário

1	Introdução	1
1.1	Contextualização	1
1.2	Objetivos	2
1.3	Metodologia	2
1.4	Organização do texto	2
2	Empreendedorismo	5
2.1	Definição	5
2.2	Empreendedorismo no Brasil	5
2.3	Empreendedorismo nas Universidades	7
3	Conceitos aplicados	9
3.1	Análise exploratória de dados	9
3.2	Normalização dos dados	9
3.3	Distância de Levenshtein	10
3.4	Clustering	10
3.5	Arquitetura limpa	11
3.6	Padrões de projeto	11
3.7	Integração Contínua	12
3.8	Entrega contínua	13
3.9	Visão geral	13
4	Tratamento inicial dos dados	15
4.1	Coleta e exploração	15
4.2	Estrutura inicial e tratamento	17
4.3	Enriquecimento do <i>dataset</i>	18
4.4	Modelagem de dados	18
5	Análises	21

5.1	Área de atuação	21
5.2	Institutos	23
5.3	Participação feminina	29
5.3.1	Propostas de incentivo	35
5.4	Localização das empresas	37
5.5	Spin-offs	39
5.5.1	Spin-offs por pós-graduação	39
5.5.2	Spin-offs por patentes	40
6	Desenvolvimento	43
6.1	Objetivo do sistema	43
6.2	Versão inicial	43
6.3	Módulo <i>core</i>	44
6.3.1	Tecnologias utilizadas	44
6.3.2	Visão geral	47
6.4	Módulo <i>WebAPI</i>	51
6.4.1	Tecnologias utilizadas	52
6.4.2	Visão geral	53
6.4.3	Disponibilização	54
6.5	Módulo <i>Web</i>	55
6.5.1	Tecnologias utilizadas	55
6.5.2	Visão geral	55
6.5.3	Disponibilização	56
6.5.4	Telas	57
7	Conclusões	67
7.1	Análises realizadas	67
7.2	<i>Software</i> desenvolvido	68
7.3	Próximos passos	69
8	Apreciação Pessoal	71
8.1	Lucas	71
8.2	Daniel	72
	Referências	75

Capítulo 1

Introdução

1.1 Contextualização

Quando se pensa em São Paulo, prontamente associa-se ao Estado densas regiões urbanas e um enorme aglomerado de atividades comerciais. A partir do declínio da economia cafeeira e de uma necessidade de diversificação das atividades produtivas, a Região Sudeste foi palco de iniciativas comerciais e industriais pioneiras no Brasil, que impactaram nas dinâmicas ambientais e sociais de diversas formas. Com isso, essa região brasileira, e sobretudo o Estado de SP, se mostra o lugar com o mais denso polo de indústrias e serviços do Brasil na atualidade.

De acordo com estimativa do [MINISTÉRIO DA ECONOMIA \(2021\)](#), mais de três milhões de empresas foram abertas no Brasil e dessas, cerca de novecentas mil, ou 28% do total, originaram-se do Estado de São Paulo. De acordo com o Relatório da Receita Tributária do Estado de São Paulo, disponibilizado pela [SECRETARIA DA FAZENDA E PLANEJAMENTO \(2021\)](#), valores provenientes do Imposto sobre Circulação de Mercadorias e Serviços (ICMS) contribuíram em mais de 83% para o valor total coletado pelo estado no ano de 2020.

A Universidade de São Paulo (USP) contribui diretamente para esse aspecto da Região Sudeste. Como espaço de referência de ensino, pesquisa, cultura e extensão, a USP mostra-se como instituição de suma importância para o panorama econômico regional e nacional por meio da inovação e do empreendedorismo. A partir de iniciativas como a Agência USP de Inovação (AUSPIN), o Núcleo de Empreendedorismo da USP (NEU), incubadoras e programas de aceleração, muitas das empresas abertas no país têm em sua origem incentivos possibilitados pela USP.

O Hub USP Inovação, iniciativa da AUSPIN, reconhece essas empresas como empresas com DNA USP, ou simplesmente empresas DNA USP. Esse conceito descreve empresas que estão associadas à Universidade por serem fundadas por alunos, por terem recebido incentivos, por serem gerenciadas por membros do corpo docente ou por terem passado pelo processo de incubação em alguma das incubadoras da USP, por exemplo. A iniciativa promove o cadastro dessas empresas nos sistemas do Hub para a produção de análises e exposição das empresas no portal DNA USP.

1.2 Objetivos

Este trabalho apresenta um estudo sobre tais companhias, com a proposta de analisá-las sobre diferentes aspectos. A partir dos dados coletados para o selo DNA USP, buscou-se responder perguntas pertinentes para o empreendedorismo na Universidade e no país, tais como a relação entre programas de incentivo (como bolsas de pesquisa, programas de incubação e habitats de inovação) e a longevidade de um empreendimento, como se dão a disparidade entre mulheres e homens no corpo executivo das companhias e como se distribuem as áreas de atuação das empresas que nascem ou se viabilizam na USP. Para isso, foram aplicadas estratégias para limpar, padronizar e gerar inferências sobre os dados disponíveis. Foram feitas inicialmente análises exploratórias sobre o *dataset*, que está estruturado em uma planilha única no serviço Google Planilhas. Em seguida, a partir das análises realizadas, produzimos gráficos e interfaces computacionais para responder de forma intuitiva e direta perguntas formuladas sobre as empresas e a universidade e evidenciar constatações obtidas a partir das mesmas.

1.3 Metodologia

O trabalho foi realizado em parceria com membros do Hub USP Inovação a partir de reuniões periódicas ao longo do ano de 2021. Nessas reuniões, foram solicitadas análises e filtragens dos dados, bem como foram realizadas discussões sobre caminhos interessantes a serem tomados no trabalho. Cada reunião tinha cerca de quarenta minutos de duração e foram realizadas quinzenalmente.

As principais ferramentas utilizadas neste trabalho foram a biblioteca *Pandas* e a plataforma *Jupyter Notebook*. Para visualização de dados, os pacotes *seaborn* e *matplotlib* e para a construção da plataforma DNA USP Web, ambiente para visualização interativa das análises obtidas, o arcabouço em *Javascript Next.js* para construção da interface gráfica, *Typescript* para a construção de um domínio que contém validação e normalização de dados, e o arcabouço em *Javascript NestJS* para a construção de um *back-end* para suportar a persistência e o gerenciamento desses dados.

1.4 Organização do texto

O texto desta monografia está organizado em capítulos que podem apresentar seções e subseções. No capítulo 2, é apresentada uma noção geral do conceito de empreendedorismo e se discute a situação atual dessa prática no Brasil e nas universidades brasileiras.

No capítulo 3, discorre-se sobre alguns dos principais conceitos aplicados nesse trabalho são exibidos exemplos e ilustrações para tal. Os conceitos apresentados podem não estar relacionados aos adjacentes no texto, mas serão referenciados posteriormente ao longo do texto.

Já no capítulo 4, são apresentados os dados iniciais utilizados nas análises apresentadas neste trabalho. São discutidos os problemas de estruturação e as dificuldades trazidas em consequência deles. Também é apresentando o modo como os dados foram coletados e

algumas das estratégias para padronizá-los.

No capítulo 5 há o compilado de análises, representadas por gráficos e tabelas, e as discussões a partir desse material. São exibidas análises e discutido sobre a área de atuação das empresas DNA USP, a frequência das diferentes unidades da USP no vínculo com as empresas cadastradas, a participação feminina no empreendedorismo desenvolvido na universidade e a distribuição geográfica de tais empreendimentos. Por fim, é apresentado o conceito de empresas *spin-offs* e como elas estão representadas no corpo de empresas vinculadas à Universidade de São Paulo.

No capítulo 6 é apresentada a plataforma DNA USP Web, bem como os módulos responsáveis pela API e pelo processamento de dados do site. Além da apresentação do projeto, é feito também um panorama das tecnologias utilizadas e retomam-se alguns dos conceitos discutidos no capítulo 3.

Capítulo 2

Empreendedorismo

Neste capítulo será dada uma visão geral sobre o empreendedorismo no contexto do Brasil e das universidades brasileiras, incluindo reflexões sobre o perfil do empreendedor brasileiro e do empreendedor brasileiro que aderiu à prática no contexto da universidade.

2.1 Definição

O desenvolvimento de um país se dá majoritariamente pelo total de riquezas que ele produz. Um importante componente que contribui para a produção de bens e serviços é a prática empreendedora. De modo geral, empreender significa iniciar novos negócios ou aprimorar empreendimentos já existentes. É uma prática fortemente associada ao setor econômico, sendo inerente ao meio de produção capitalista.

O termo empreendedorismo foi introduzido pelo economista austríaco Joseph Schumpeter ainda na década de 1940, como parte de sua conhecida teoria de Destruição Criativa. Para [SCHUMPETER \(1942\)](#), empreender significa introduzir qualquer forma de inovação ao sistema econômico. O agente que realiza tal introdução é nomeado por Schumpeter empresário empreendedor. É alguém que exerce uma postura de versatilidade, que domina a técnica do negócio e que possui sólidas noções administrativas e econômicas para gerenciar um negócio com êxito.

Também existe a noção de intra-empendedor, que classifica a pessoa que demonstra postura de empresário empreendedor, porém inserida em uma corporação não necessariamente fundada ou idealizada por ela.

2.2 Empreendedorismo no Brasil

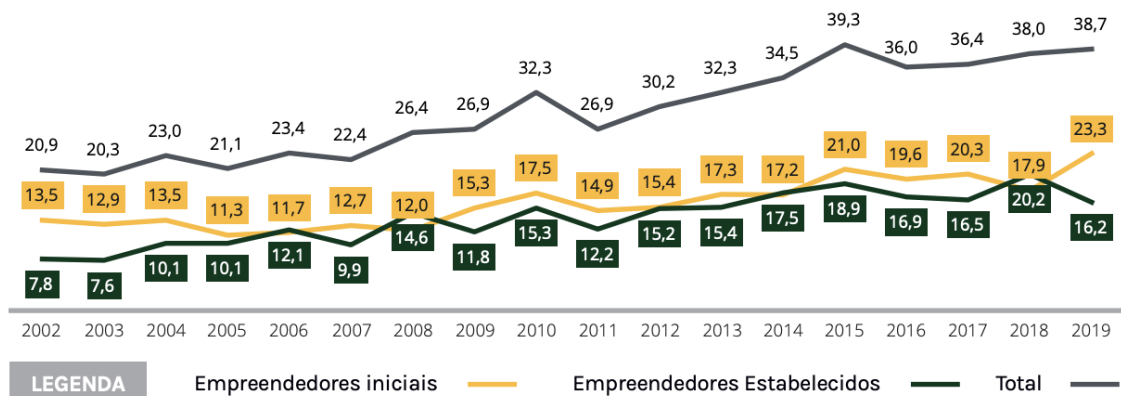
O setor de serviços é o de maior participação na geração de renda e empregos no Brasil. Nele estão incluídas atividades comerciais, grandes corporações, o setor de transportes, a administração pública, entre outros.

De acordo com o último Relatório Executivo da pesquisa *Global Entrepreneurship*

Monitor (GEM), coordenada no Brasil pelo **INSTITUTO BRASILEIRO DE QUALIDADE E PRODUTIVIDADE (2019)**, a taxa total de empreendedorismo no Brasil foi de 38,7% entre a população adulta, que corresponde a faixa de brasileiros que tem entre 18 e 64 anos de idade. Esse valor coloca o país no topo do *ranking* mundial de empreendedorismo. A métrica abrange qualquer pessoa que participou de alguma atividade empreendedora, seja criando um novo negócio ou mantendo um empreendimento já existente.

A pesquisa GEM trabalha com duas classes de empreendedores, os iniciais e os estabelecidos. Empreendedores iniciais podem ser classificados entre empreendedores nascentes - aqueles que são proprietários de um negócio que ainda não gerou remuneração por mais de três meses - e empreendedores novos, que são donos de iniciativas que já renderam algum retorno aos participantes. Já os empreendedores consolidados são empresários que administram negócios duradouros, que apresentaram retorno financeiro por um período superior a 42 meses.

O relatório suprarreferido mostra que, em 2019, o empreendedorismo inicial superou o empreendedorismo estabelecido, indicando que há um crescimento de pessoas iniciando um negócio próprio. O empreendedorismo inicial, nesse ano, correspondeu a 23,3% do total.



Fonte: GEM Brasil 2019

¹ Percentual da população de 18 a 64 anos.

Figura 2.1: Taxas percentuais de empreendedorismo segundo estágio do empreendimento (inicial ou estabelecido)

De acordo com a pesquisa, esse aumento ocorreu por conta de um leve aquecimento da economia, ocorrido no ano anterior, associado a uma alta taxa de desemprego, a qual aumenta a incerteza da população sobre sua ocupação e dá espaço para que pessoas tentem empreender em negócios próprios para manter uma renda mínima.

Embora a atividade empreendedora seja comumente associada à disposição de assumir riscos e inovar, sendo muitas vezes confundida com um estilo de vida e uma escolha pessoal, observa-se que a realidade socioeconômica do país mostra-se um fator mais decisivo para que as pessoas pratiquem o empreendedorismo, como mostra a tabela 2.1. A maior motivação para empreender, segundo a pesquisa GEM, é buscar uma alternativa a escassez de empregos. Assim, podemos dizer que no Brasil, o “empresário empreendedor”

definido por Schumpeter é, acima de tudo e na maioria das vezes, alguém que busca meios de atenuar os efeitos da turbulenta situação econômica e social do país.

Motivação	Percentual
Para ganhar a vida porque os empregos são escassos	88,4
Para fazer a diferença no mundo	51,4
Para construir uma grande riqueza ou uma renda muito alta	36,9
Para continuar uma tradição familiar	26,6

Tabela 2.1: Percentual dos empreendedores segundo motivação para iniciar um novo negócio - Brasil, 2019. Fonte: GEM Brasil 2019

2.3 Empreendedorismo nas Universidades

A partir da segunda metade do século XX foi iniciado um processo de aproximação entre o ambiente acadêmico e as atividades econômicas em geral. Anteriormente focada somente em ensino e pesquisa, o papel da universidade até então era limitado em produzir e transmitir conhecimento entre aqueles que a frequentavam. Com a segunda revolução acadêmica, uma nova missão é atribuída à universidade como instituição. Esta nova missão está associada ao desenvolvimento econômico e social e busca conciliar a produção acadêmica com as necessidades da sociedade na qual os centros de ensino se inserem. Com essa mudança, surge o conceito de Universidade Empreendedora, que introduz a inovação no rol de princípios atrelados as universidades e traz a pesquisa como um meio para alcançar tal princípio, como explicam [ALMEIDA E CRUZ \(2010\)](#).

No Brasil, embora a pesquisa GEM demonstre alta taxa de empreendedorismo, ainda são poucos os cursos de ensino de superior que oferecem disciplinas de empreendedorismo. De acordo com o relatório Empreendedorismo nas Universidades Brasileiras em sua última edição, publicada por [SEBRAE E ENDEAVOR BRASIL \(2016\)](#), os cursos de empreendedorismo se concentram em cursos de engenharia e de ciências sociais aplicadas, como economia e administração. Nas demais áreas do conhecimento, a oferta de cursos dessa natureza representa menos de 30% do total. Com isso, forma-se uma segregação no acesso a conhecimentos associados a empreendedorismo para áreas eminentemente tecnológicas.

Outra problemática na dinâmica empreendedora no ambiente acadêmico é a democratização do ensino superior no Brasil. Segundo dados levantados pela [REVISTA PIAUÍ \(2021\)](#) apenas 21% dos adultos brasileiros de 25 a 34 anos de idade concluíram o ensino superior. Esse percentual representa a taxa mais baixa dentre os países da América Latina e é significativamente inferior à média dos países que compõem a Organização para Cooperação e Desenvolvimento Econômico (OCDE).

De acordo com pesquisa realizada pelo [INSTITUTO BRASILEIRO DE QUALIDADE E PRODUTIVIDADE \(2019\)](#) sobre o perfil do empreendedor brasileiro, foi constatado que a faixa etária de 18 a 30 anos - mesma faixa etária que representa os universitários brasileiros - é aquela que mais empreende no país. Essa comparação indica o quanto o ambiente acadêmico pode ser campo fértil para iniciativas de ensino sobre empreendedorismo.

Perante ao baixo acesso a educação superior, o levantamento feito por **SEBRAE E ENDEAVOR BRASIL (2016)** demonstrou que somente 35.5% das instituições brasileiras possuem disciplinas abertas que podem ser frequentadas por alunos de diferentes cursos. Já as disciplinas transversais, que acolhem alunos de diferentes cursos nas mesmas turmas, são adotadas por apenas 41.1% das universidades. Essa constatação representa um aspecto negativo a respeito do modo como tem sido promovido o empreendedorismo nas universidades, já que espaços que contemplem somente grupos específicos e pouco diversos promovem poucas chances de interações, que podem ser o começo de muitas parcerias empreendedoras.

Ainda que relativamente tímido, o incentivo ao empreendedorismo nas universidades brasileiras produz resultados, evidenciados pelas próprias empresas DNA USP, introduzidas no capítulo anterior. Considerando a prática empreendedora como componente chave para a produção de riqueza de um país, e no contexto do Brasil, também como alternativa à caótica situação laboral, é necessário avaliar tais resultados com minúcia, a fim de promovê-la na universidade e no país. Tal avaliação é realizada neste trabalho utilizando-se largamente dos conceitos introduzidos no próximo capítulo.

Capítulo 3

Conceitos aplicados

3.1 Análise exploratória de dados

Realizar análise exploratória sobre determinado conjunto de dados significa, antes de tudo, compreender cada variável disponível e as relações entre elas. Essa compreensão passa por preparar os dados, visando padronizá-los e mapeá-los, detectando e corrigindo inconsistências, se possível; gerar visualizações simples sobre os recursos envolvidos, como histogramas e diagramas circulares, para extrair as primeiras informações relevantes para as perguntas a serem respondidas; avaliar a presença de dados ausentes e realizar algumas suposições mais simples, como relações lineares entre as variáveis presentes. Neste trabalho, podemos pensar nas variáveis ano de fundação e faturamento das empresas, por exemplo.

3.2 Normalização dos dados

O matemático britânico Edgar Frank Codd (1923-2003) foi quem propôs inicialmente o processo de normalização de banco de dados, em 1972. A normalização dos dados é uma maneira de garantir a semântica dos atributos, de reduzir a quantidade de valores redundantes ou nulos e suprimir eventuais inconsistências. Formalmente, são definidas três formas normais que representam diretrizes para atingir tais objetivos.

A Primeira Forma Normal (1FN), suprime o uso de atributos multivalorados e compostos, bem como suas combinações. O objetivo dessa normalização é promover flexibilidade e independência aos dados, além de simplificar a representação de atributos e relações. A 1FN também facilita as demais normalizações diminuindo a incidência de redundâncias e anomalias.

A maioria dos Sistemas Gerenciadores de Banco de Dados (SGBDs) não oferecem suporte para registros multivalorados, de modo que as relações modeladas nesses sistemas já satisfazem essa primeira normalização. Dessa forma, ao migrar dados para um banco de dados relacional, é preciso antes normalizá-los sobre as diretrizes da Primeira Forma Normal.

A Segunda Forma Normal (2FN) define que todo atributo em uma relação da base de dados seja dependente da chave primária daquela relação. Para uma relação estar na 2FN, precisa também estar na 1FN.

Já a Terceira Forma Normal (3FN) indica que todo atributo em uma relação que seja derivado de outros atributos, isto é, que sejam funcionalmente dependentes, deve ser removido. Tais atributos devem ser acessados de outras formas, como a partir de *views*.

3.3 Distância de Levenshtein

Quando pretende-se trabalhar com conjuntos de dados, é comum encontrá-los sem índices que permitam encontrar correspondências rapidamente entre outras tabelas para realizar análises pertinentes. Por exemplo, considerando uma tabela com o nome de inventores e outra contendo patentes por inventor, seria conveniente que a primeira tabela identificasse cada inventor com um inteiro que seria simplesmente referenciado pela segunda, compondo uma associação direta entre as duas relações de dados.

id	Inventor	id	Invento	inventor_id
1	Santos Dummont	1	Avião	1
2	Francisco Canho	2	Chuveiro Elétrico	2
3	Nelson Guilherme Bardini	3	Cartão telefônico	3

Porém, em determinados conjuntos de dados, a única forma possível de associar os dados pode ser a partir de colunas contendo somente cadeias de caracteres, adicionando significativa complexidade ao processo de intercalação entre múltiplos registros.

id	Invento	inventor_nome
1	Avião	santos dummont
2	Chuveiro Elétrico	Canho
3	Cartão telefônico	Nelson G Bardini

Uma técnica para tratar esse tipo de problema foi proposta por Vladimir Levenshtein (1965). A Distância Levenshtein, ou distância de edição entre duas cadeias de caracteres, é uma precisa métrica útil para determinar o quão similares dois textos são. A distância de Levenshtein entre duas palavras A e B é dada por um número que representa o mínimo de operações para transformar A em B, de modo que, quanto maior for esse número, mais diferentes as palavras são. As operações contabilizadas incluem exclusões, substituições ou inserções de novas letras. Além de ser útil para buscar correspondências entre palavras, a técnica também é usada em verificadores ortográficos.

3.4 Clustering

Uma tradução precisa para o termo *clustering* é agrupamento, essa técnica em ciência de dados é exatamente sobre isso: agrupar dados a partir de características similares. Essa técnica é utilizada com o propósito de extrair informações de grandes conjuntos de dados estruturados ou não, fornecendo uma visão geral de como se distribuem e quais são as características mais proeminentes na coleção estudada.

Há diversos algoritmos para agrupar conjuntos de dados. Pode-se, por exemplo, utilizar o algoritmo *K-means* para tal. Essa abordagem define K centróides para criar K agrupamentos, e cada ponto da coleção será então associado ao centróide mais próximo. Quando todos os pontos estiverem associados ao respectivo grupo, calcula-se os pontos centrais de cada cluster e os elementos são reposicionados em relação ao novo centro. O processo é repetido até que se encontre um centróide definitivo, tal que uma nova interação do processo não altere a posição dos pontos presentes no *dataset*.

3.5 Arquitetura limpa

Introduzida no célebre livro de [ROBERT MARTIN, 2017](#), a arquitetura limpa consiste em um conjunto de técnicas e padrões utilizados em desenvolvimento de software de modo a maximizar a coesão e reduzir o acoplamento entre as partes. Um princípio fundamental dessa arquitetura é a separação de responsabilidades, alcançada dividindo o software em camadas que isolam regras de negócio de implementações e dependências vinculadas a arcabouços específicos. Dessa maneira, obtém-se um sistema com uma regra de negócio encapsulada e facilmente testável, que passa a não depender de detalhes de implementação, como o SGBD escolhido ou as bibliotecas empregadas para tratar requisições HTTP.

Normalmente, segrega-se um projeto seguindo essa arquitetura em duas grandes partes, a de domínio e a de aplicação. No domínio reunimos as entidades, que representam modelos e entidades envolvidos na lógica da aplicação, bem como métodos ou estruturas de dados fundamentais para os requisitos que se pretende atingir. Há também os casos de uso, que descrevem a lógica por trás de cada funcionalidade da aplicação, orquestrando os fluxos e tratando diferentes cenários. Na camada de domínio temos também as principais interfaces de comunicação, estabelecendo os contratos fundamentais para as entradas e saídas de cada caso de uso. Na camada de aplicação, por outro lado, concentram-se os detalhes de implementação, modelados a partir dos contratos definidos nos módulos presentes no domínio. Dessa forma, alterações podem ser feitas na camada de aplicação sem ocorrerem interferências drásticas nas regras de negócio.

Aplicações desse conceito são apresentadas no capítulo 5.

3.6 Padrões de projeto

Ao longo do tempo, desenvolvedoras e desenvolvedores elaboraram soluções eficientes para diversos problemas com os quais a maioria das pessoas se depara quando estão construindo um software. Essas soluções são chamadas padrões de projeto e podem ser definidas como estratégias e conceitos bem definidos, que se dispõem como artifícios para a construção de softwares coesos, consistentes e pouco acoplados. São ferramentas úteis e eficientes para resolver problemas conhecidas de maneira otimizada e organizada.

Vale ressaltar, porém, que a presença de padrões de projeto em um sistema não o torna necessariamente bom. O aspecto mais importante é aplicar os padrões em situações que sejam coerentes com as capacidades viabilizadas.

O principal livro difundido pela comunidade sobre o assunto é o [GANG OF FOUR \(1994\)](#),

que define três principais categorias de padrões de projetos, os criacionais, estruturais e comportamentais. Os padrões criacionais postulam estratégias de criação de objetos flexíveis e reutilizáveis. Os padrões estruturais, por sua vez, podem ser utilizados para aprimorar a interação entre diferentes classes e paradigmas sem perder a flexibilidade e o isolamento. Já os comportamentais indicam formas de entender e delegar responsabilidades e capacidades entre diferentes entidades.

3.7 Integração Contínua

Na metodologia de desenvolvimento ágil, alterações são feitas frequentemente em repositórios de código por diversas pessoas. A prática de integração contínua, conhecida também como *Continuous Integration (CI)* consiste em testar antecipadamente e com frequência qualquer alteração realizada em relação às regras definidas anteriormente, avaliando regras de negócio, desempenho e a introdução de defeitos a partir da execução de testes de unidade e de integração, compilando a aplicação de modo contínuo e com isso, se antecipando a defeitos de diversas naturezas. Desse modo, a escrita de testes é parte fundamental do processo.

De modo prático, os diversos serviços de integração contínua apresentam rapidamente relatórios em relação a cada nova alteração enviada para um repositório de código. Informações como cobertura de código, resultados dos testes, logs de compilação, verificação de qualidade de código e vulnerabilidades podem ser exibidos para cada mudança submetida, como mostrado na figura 3.1.

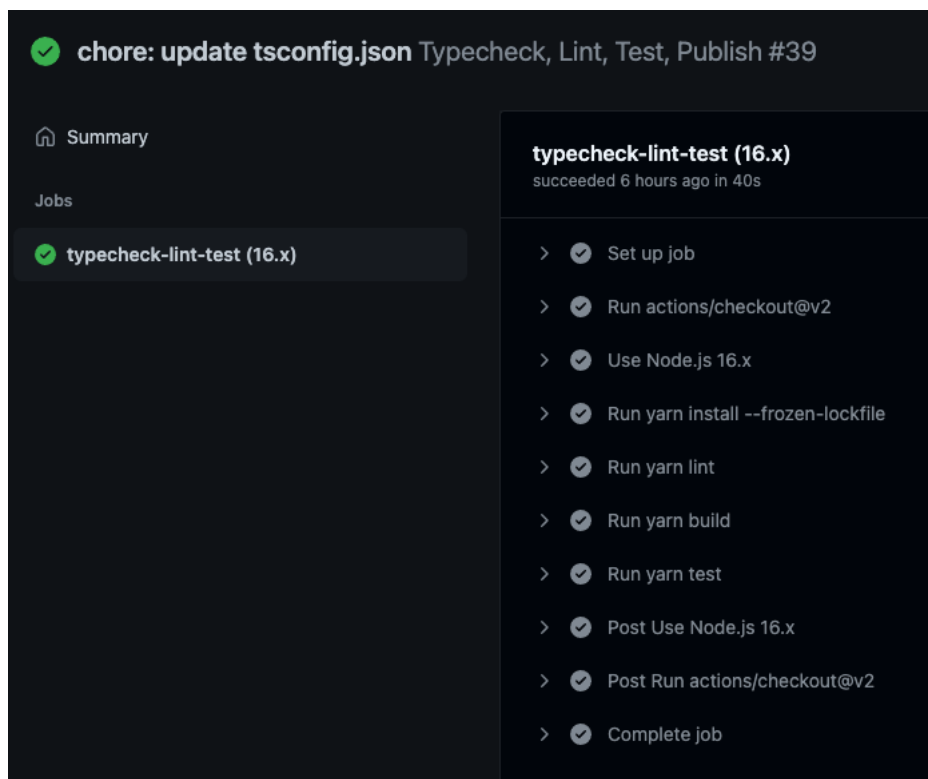


Figura 3.1: Exemplos de passos executados no fluxo de Continuous Integration do projeto dnausp-core

O desenvolvimento de software atual depende fortemente das práticas de integração contínua. Uma gama de ferramentas com propósitos e funcionalidades variados está disponível no mercado, muitas vezes de gratuitamente. É um processo fundamental para o desenvolvimento com qualidade e consistência.

3.8 Entrega contínua

Se por um lado a integração contínua introduz estratégias para validar a consistência e validade de cada nova alteração submetida a um código-fonte, a prática de entrega contínua, ou *Continuous Delivery (CD)*, busca agilizar o processo de publicação e versionamento de um código após o fluxo de desenvolvimento. Com essa prática, alterações feitas são automaticamente testadas, carregadas e implantadas em um ambiente de produção ou de testes, tornando possível a utilização por uma equipe de controle de qualidade ou mesmo por usuários finais.

As práticas de CI e CD se relacionam diretamente ao conceito de *DevOps*, que busca mesclar as práticas de implementação e de gerenciamento de infraestrutura e disponibilidade de serviços. A abordagem *DevOps* envolve a estruturação de uma *pipeline* de entrega contínua.

3.9 Visão geral

Os conceitos relacionados à análise de dados permeiam os métodos utilizados para o tratamento e extração de métricas e indicadores a partir dos dados coletados pela AUSPIN. As concepções acerca do processo de desenvolvimento de software foram utilizadas para desenvolver um sistema capaz de centralizar e serializar as operações de validação e padronização dos dados, assim como oferecer acesso a uma *interface web* para que as métricas e indicadores produzidos sejam utilizados para produzir relatórios interativos sob demanda.

Capítulo 4

Tratamento inicial dos dados

Neste capítulo é apresentado o conjunto de dados coletados pela AUSPIN sobre as empresas DNA USP, objeto do estudo deste trabalho. São descritos os processos de coleta e aspectos da exploração dos dados e da maneira com que os mesmos foram adquiridos; os passos iniciais para exploração e ferramentas utilizadas; a estruturação original dos dados, bem como estratégias utilizadas para o tratamento dos dados a fim de chegar em uma estrutura apropriada para realizar análises mais profundas; técnicas utilizadas para o enriquecimento do conjunto de dados a partir de fontes externas e a modelagem proposta para realização da análise exploratória.

4.1 Coleta e exploração

A fim de agregar empresas com DNA USP, a AUSPIN disponibilizou, através do **Hub USP Inovação**, uma interface que permite que empreendedores ou representantes de companhias cadastrem dados de companhias com esse perfil. A *interface* foi projetada a partir da ferramenta Google Forms, conforme mostrado na figura 4.1. Os dados inseridos nela completam linhas de uma tabela disponível no Google Planilhas, usada como base de dados pelos membros da equipe responsável pela gerência e manutenção da marca DNA USP.

A planilha na qual os dados cadastrados são depositados contém 82 colunas descrevendo atributos das empresas, das pessoas empreendedoras envolvidas, dos vínculos com a USP, sobre incentivos recebidos, dados a respeito do faturamento e também metadados sobre o cadastro, por exemplo, datas de atualizações efetuadas em cada linha e se determinada relação foi validada.

Para a análise exploratória dos dados, a planilha foi baixada periodicamente em formato CSV e manipulada utilizando a linguagem *Python*, a partir da biblioteca de ciência de dados *Pandas*. Para organizar e apresentar as análises e as implementações feitas, foi utilizada a plataforma *Jupyter Notebook*.



Empresas com DNA USP

Esse formulário tem como objetivo prospectar empresas criadas por integrantes da comunidade USP, a fim de dar visibilidade para as mesmas no Hub USPInovação (<https://hubusp.inovacao.usp.br/empresas>).

Para preencher este formulário, a empresa necessita ter sido criada por alunos(as) que cursaram ou que ainda estejam cursando a graduação ou pós-graduação na USP, ou então por pesquisadores e docentes da USP. Também podem se cadastrar as empresas que sejam resultantes de processos de incubação ou de aceleração em alguma das incubadoras associadas à Universidade de São Paulo (CIETEC, ESALQTEC, HABITs e Supera).

Somente empresas formalmente constituídas podem preencher este formulário.

Os dados sensíveis não serão divulgados. Somente as informações públicas serão disponibilizadas no Hub USPInovação.

Em caso de dúvidas, entre em contato pelo email: hubusp.inovacao@usp.br

A foto e o nome associados à sua Conta do Google serão registrados quando você fizer upload de arquivos e enviar este formulário.. Seu e-mail não faz parte da resposta.

[Próxima](#) [Limpar formulário](#)

Figura 4.1: Formulário de cadastro de empresas - HUB USP Inovação

4.2 Estrutura inicial e tratamento

O elevado número de colunas e a não diferenciação das entidades envolvidas no cadastro - como empresa, sócios e institutos - foi o primeiro grande desafio para analisar atributos específicos de cada registro. Como exemplo, temos a situação em que o uso de uma coluna contendo nome de institutos para caracterizar o vínculo com a USP de cada sócio, ao invés de um índice no formato chave-estrangeira que relacionasse cada entrada com uma linha em outra coluna, adicionou complexidade na separação dos dados para análises que seriam naturalmente simples de serem feitas a partir de uma operação JOIN. Dessa forma, para responder quais são os institutos da USP nos quais mais empresas se originaram, foi necessário extrair todos os valores presentes na coluna correspondente ao instituto de cada sócio e eliminar inúmeras duplicações encontradas, que não puderam ser facilmente tratadas computacionalmente, por representarem o instituto mas com textos diferentes. É o caso, por exemplo, de linhas contendo a palavra "IME" e outras contendo "Instituto de Matemática e Estatística". Muitos desses cenários precisaram ser tratados individualmente a partir de filtros e processos de busca de similaridades de palavra, utilizando para isso a distância de Levenshtein implementada em bibliotecas como a *fuzzywuzzy*. Como os dados cadastrados não passam por critérios de validação no formulário, por limitação da própria plataforma Google Forms utilizada, situações similares se repetiram em muitas outras colunas, a biblioteca desenvolvida ao longo desse projeto, apresentada no capítulo 6, pode servir de base para o desenvolvimento de um novo formulário em plataforma própria.

Outro aspecto que trouxe complexidade para a exploração dos dados é a presença de atributos multivalorados na tabela. O formulário disponibilizado pela AUSPIN permite que o usuário cadastre no mínimo um e no máximo cinco sócios para uma mesma empresa, como mostrado na figura 4.2. Para isso, na planilha foram criadas, cinco vezes cada, colunas de nome, instituto e vínculo com a USP, aumentando drasticamente a incidência de campos nulos, uma vez que a maioria dos usuários registraram no máximo dois sócios. Além disso, na situação em que se almeja contar o total de empresas por instituto, foi necessário, para cada linha, analisar cada uma das cinco colunas para capturar as entradas não vazias e testá-las usando o processo de similaridade de textos para validar o instituto cadastrado.

Dos campos mais importantes disponibilizados na planilha, temos, sobre empresas: CNPJ, nome fantasia, razão social, ano de fundação, dados de endereço, faturamento por ano, porte, status da empresa (se está ativa para a Receita Federal), CNAE (Classificação Nacional de Atividades Econômicas) e um campo indicando se empresa é um unicórnio - isto é, startups que atingiram a marca de valor de mercado igual a ou superior a US\$1 bilhão - e o tipo da empresa.

Sobre os sócios, temos o nome, o vínculo mais recente com a USP e com qual instituto esse vínculo foi estabelecido, além de informações de contato. Em relação aos faturamentos, temos o valor anual e o ano fiscal correspondente. Já sobre investimentos e incentivos, temos campos que indicam quais processos de incubação ou aceleração cada empresa participou, se recebeu investimentos e de quais tipos.

HUB USP INOVAÇÃO

Empresas com DNA USP

A foto e o nome associados à sua Conta do Google serão registrados quando você fizer upload de arquivos e enviar este formulário.. Seu e-mail não faz parte da resposta.

***Obrigatório**

Sócios

Gostaria de adicionar os dados dos demais sócios? *

Sim

Não

Voltar Próxima Limpar formulário

Nunca envie senhas pelo Formulários Google.

Este formulário foi criado em Universidade de São Paulo. [Denunciar abuso](#)

Google Formulários

Figura 4.2: Adicionando múltiplos sócios - HUB USP Inovação

4.3 Enriquecimento do *dataset*

Alguns dos dados utilizados nas análises foram inferidos dos campos existentes, como a área de atuação da empresa, determinada a partir dos dígitos do CNAE, da geolocalização de cada empreendimento, extraída a partir dos dados de endereço informados e cruzando-os com dados disponibilizados pelo IBGE e também os dados sobre a distribuição de gênero, inferida a partir do primeiro nome de cada sócio cadastrado e observando a proporção desses nomes entre homens e mulheres numa perspectiva binária. Apesar de não ser um indicador totalmente preciso devido às inúmeras nuances sobre a identidade de gênero de cada pessoa, as proporções extraídas dessa forma abriram espaço para reflexões acerca das diferenças da participação de homens e mulheres nas iniciativas de empreendedorismo e foram compatíveis com análises dessa natureza realizadas no mercado de trabalho.

4.4 Modelagem de dados

Após análise preliminar dos dados disponíveis, foi projetado um modelo de banco de dados adequado para depositar de maneira consistente e mais segura os dados disponibilizados na planilha. Utilizando o modelo de banco de dados relacional, foi mapeado o domínio dos dados e construído diferentes Diagramas Entidade-Relacionamento (DER), como o da figura 4.3.

A partir do modelo projetado e das estratégias desenvolvidas para normalização e enriquecimento do conjunto de dados, foi desenvolvida a plataforma DNA USP Web e seus módulos constituintes, que serão apresentados em mais detalhes no capítulo 6, que aborda

4.4 | MODELAGEM DE DADOS

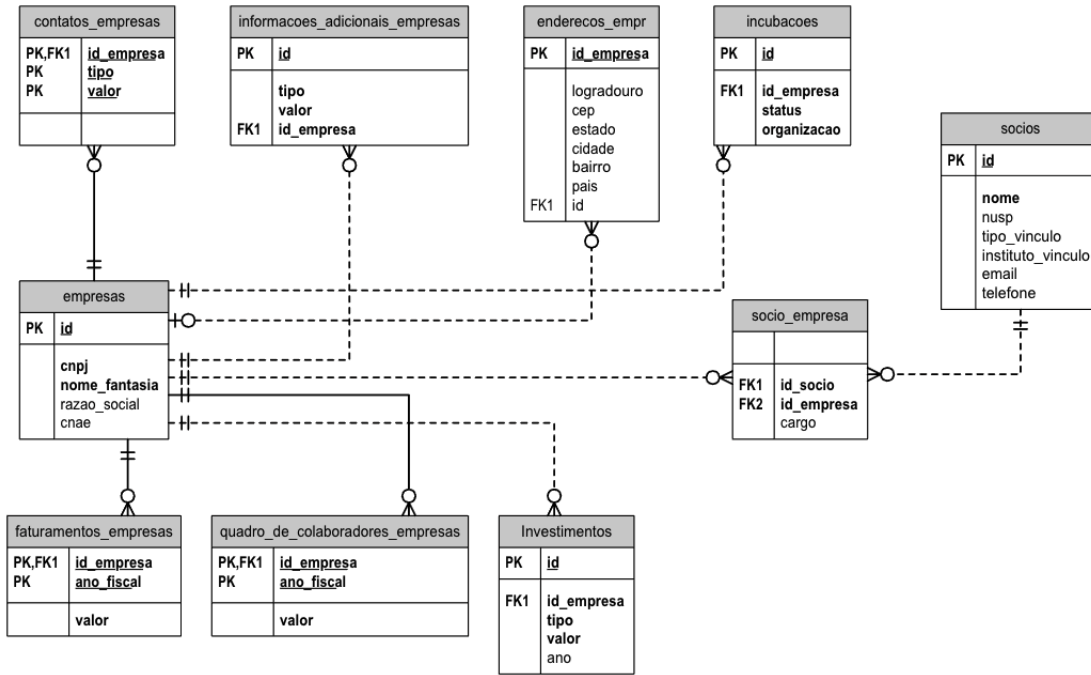


Figura 4.3: Diagrama entidade-relacionamento do novo modelo de dados proposto

o desenvolvimento de *software* no projeto.

Capítulo 5

Análises

A partir dos dados das empresas DNA USP, buscou-se responder questões relevantes para compreensão de aspectos atuais das práticas de inovação e empreendedorismo exercidas pela Universidade de São Paulo, neste capítulo são introduzidas reflexões e possíveis respostas para essas questões. Os dados disponíveis contemplam empresas nascidas, em sua maioria, entre meados da década de 1990 até o ano atual. Existe, ainda, uma pequena parte dos registros cuja data de fundação está fora desse intervalo.

Os aspectos observados dizem respeito à área de atuação das empresas relacionadas à Universidade; institutos de origem mais frequentes dentre as empresas cadastradas; regiões geográficas em que as empresas DNA USP têm sua sede; distribuição de gênero no corpo de sócios de cada empreendimento; empresas que são spin-offs, isto é, que possuem relação com alguma patente registrada ou possuem sócios que foram ou são estudantes de pós-graduação; e iniciativas da USP de incentivo à formação de novos negócios.

5.1 Área de atuação

Para analisar a natureza das atividades exercidas por empresas DNA USP, foi utilizado o campo contendo o código CNAE (Classificação Nacional de Atividades Econômicas) cadastrado para cada empresa. O CNAE é o sistema padrão para classificar atividades econômicas no Brasil e é usado por diversos órgãos tributários do país. Desse modo, o CNAE está associado somente a empresas que possuem CNPJ, isto é, que estão registradas no Brasil.

Conforme esquematização do [CONCLA \(COMITÊ NACIONAL DE CLASSIFICAÇÃO \(2021\)\)](#), associado ao IBGE, o CNAE se caracteriza como um código contendo 7 dígitos. O primeiro número do CNAE representa a seção, e existem nesse campo 21 valores diferentes. Em seguida, há o dígito para designar as divisões, e estas possuem 87 valores. Depois, há ao todo 285 grupos no domínio do terceiro dígito. O quarto dígito é usado para designar uma classe, sendo 672 no total, e por fim há a subclasse, com 1318 valores no total. No CNAE, cada um dos dígitos compõe uma hierarquia usada para descrever uma área de atuação econômica.

As empresas DNA USP foram analisadas, em um primeiro momento, a partir do primeiro

dígito do CNAE, isto é, descrevendo a seção, de modo a extrair um agrupamento menos definido de atuação a fim obter uma visão mais ampla das atividades de cada empresa. Da proporção de cada empresa segundo essa divisão, obtemos o gráfico exibido na figura 5.1, gerado no sistema [DNA USP Web](#), temos que na série histórica, dentre todos os institutos, o grupo de empresas mais numeroso foi aquele cujas empresas atuavam em atividades profissionais, científicas e técnicas, representando 548 das empresas. Em seguida, com 398 empreendimentos nessa categoria, o grupo informação e comunicação se destaca em segundo lugar. Outros grupos de destaque são educação, indústrias de transformação, saúde humana e serviços sociais, comércio, reparação de veículos automotores e motocicletas e atividades administrativas e serviços complementares. O grupo com menos exemplares foi transporte, armazenagem e correio.

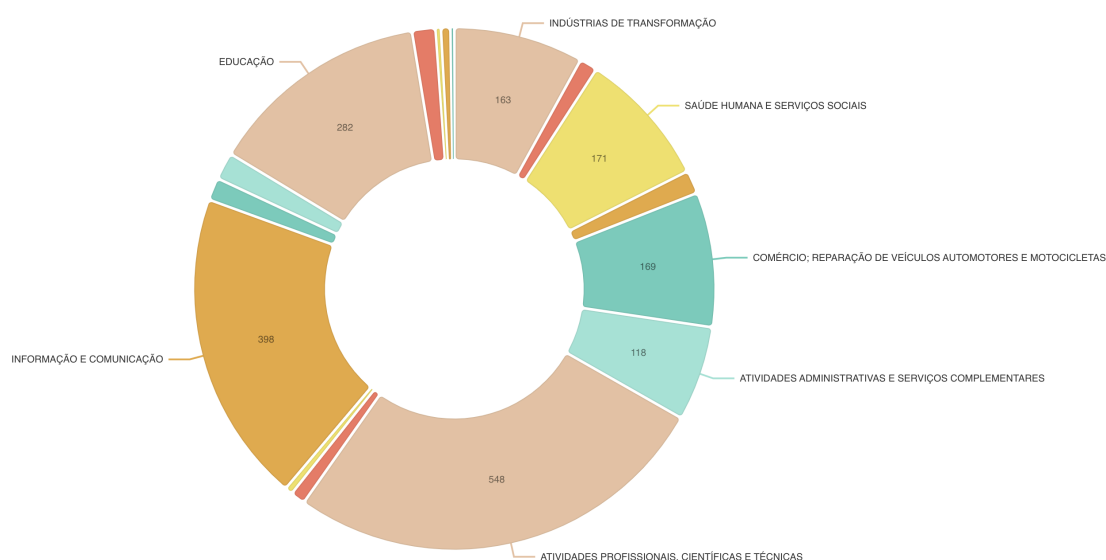


Figura 5.1: Gráfico com a proporção de áreas de atuação dentre as empresas DNA USP.

Observando a evolução histórica das áreas de atuação das empresas vinculadas à universidade, observa-se na figura 5.2 que, nas empresas fundadas de 1960 a 2000, as áreas mais notáveis são informação e comunicação, indústrias de transformação e atividades profissionais, científicas e técnicas. Havia poucas empresas vinculadas a educação ou ao comércio.

Já nas empresas surgidas entre 2001 e 2021, como mostra a figura 5.3, é perceptível um significativo aumento do total de empresas voltadas para educação, comércio e serviços e saúde, assim como uma expressiva queda de empresas voltadas para indústria de transformação.

5.2 | INSTITUTOS

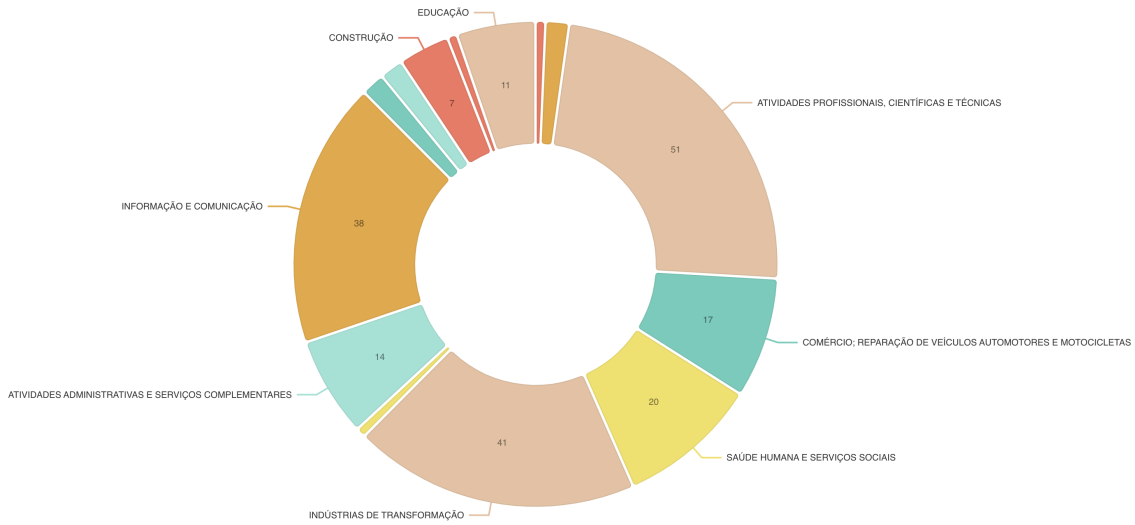


Figura 5.2: Distribuição de empresas de 1960 a 2000

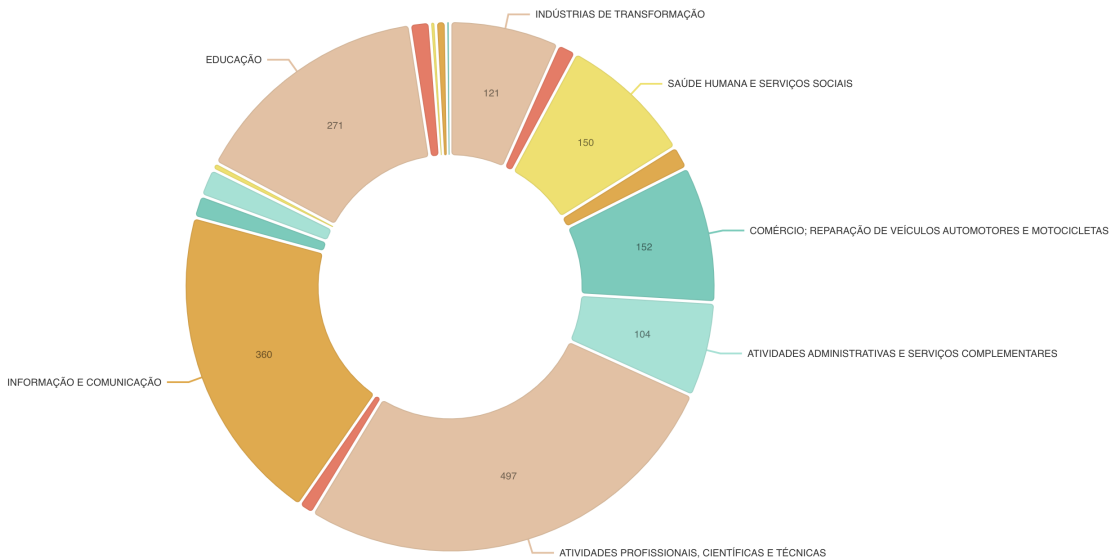


Figura 5.3: Distribuição de empresas de 2001 a 2021

5.2 Institutos

No cadastro de empresas DNA USP, o HUB USP Inovação solicita que sejam informados alguns dados a respeito das pessoas associadas aos sócios das companhias. É permitido que se cadastrem até cinco sócios e dentre os dados requisitados estão o nome, o número USP e o instituto com o qual a pessoa tem ou teve seu vínculo mais recente com a universidade. Desse modo, para cada linha da tabela em que os dados informados são depositados, existem até cinco colunas preenchidas contendo nomes de institutos da USP associando-os à empresa. No formulário, esses dados são recebidos a partir de uma caixa de seleção contendo o nome dos institutos já definido, diminuindo assim a entropia desse dado ao

não permitir que o usuário digite o nome do instituto em formatos arbitrários, no entanto, devido à existência de dados antigos não adequados a esse formato, ainda foi necessário tratar essa informação utilizando bibliotecas como *fuzzywuzzy* e *fuzzy-set*.


Com isso, foi possível analisar, dentre as mais de duas mil empresas cadastradas no portal DNA USP, quais são os institutos da USP que mais originam novas empresas. O vínculo com os institutos é definido a partir de alunos de graduação ou pós-graduação, docentes, servidores ou pesquisadores, uma empresa pode ter vínculo mais de um instituto. O cadastro ainda permite que o vínculo com a USP seja determinado pelo fato de a empresa estar ou ter sido incubada ou graduada em incubadora associada à instituição, mas nesse caso nenhum instituto pode ser selecionado.

Para realizar as análises, foi efetuada contagem do número de ocorrências de cada instituto dentre as empresas presentes na planilha e associada cada uma ao seu setor de atuação, uma ocorrência é definida pelo relacionamento indireto da empresa ao setor de atuação por meio de um ou mais sócios. A contagem foi colocada em termos percentuais e representada em um gráfico de barras que apresenta o total de empresas por área de atuação em cada instituto da USP, apresentado na figura 5.5. A partir dos cinco institutos com mais empresas, que são Escola Politécnica (EP), Escola de Engenharia de São Carlos (EESC), Escola Superior de Agricultura "Luiz de Queiroz"(ESALQ), Faculdade de Medicina (FM) e Faculdade de Economia, Administração e Contabilidade (FEA), observa-se que as companhias cujas áreas de atuação são voltadas para mais técnicas ou administrativas tendem a apresentar mais atividades associadas ao empreendedorismo e é cabível afirmar que uma das razões é a distribuição desigual de acesso a disciplinas e atividades sobre o assunto, conforme apresentado na seção 2.3. Por outro lado, é interessante notar que a Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH), mostrou-se com um número de empresas equiparável ao da FEA, mesmo focando seu ensino e pesquisa em áreas mais afastadas da engenharia e de atividades tecnológicas. Na FFLCH, a maioria das empresas tem como foco a área de Comércio de Serviços.


É detalhada a distribuição de empresas associadas a cada instituto agrupando-as por sua área de atuação e construindo a visualização dos dados a partir da utilização de gráficos de setores. Para a seção do CNAE correspondente a atividades científicas e técnicas, temos como maiores destaques a Escola Politécnica e a ESALQ, dois importantes polos de engenharia da USP, como mostrado na figura 5.6. O total de empresas vinculadas à Escola Politécnica para este setor corresponde a cerca de 25% de todas as empresas cadastradas com CNAEs similares. A fim de não poluir o gráfico com setores pouco significativos, institutos com menos empresas representantes foram agrupados em um único setor nomeado "Outros institutos".

Em relação ao setor de Comércio e Serviços, representado na figura 5.7, a Escola Politécnica também destaca-se no percentual de empresas as quais está vinculada, com 23% de todas as empresas. A Escola de Engenharia de São Carlos (EESC) aparece em segundo lugar com cerca de 12.5% de todas as empresas com essa área de atuação. Em terceiro lugar, temos o Instituto de Matemática e Estatística (IME).

Já no setor de Educação, Artes e Esportes, conforme apresentado na figura 5.8, protagonizam a Escola de Educação Física e Esportes (EEFE), a Escola de Artes, Ciências e Humanidades (EACH) e a Escola de Engenharia de São Carlos. Uma observação interessante



Empresas com DNA USP

 Rascunho salvo.

A foto e o nome associados à sua Conta do Google serão registrados quando você fizer upload de arquivos e enviar este formulário.. Seu e-mail não faz parte da resposta.

***Obrigatório**

Vínculo com a Universidade de São Paulo (USP)

Com qual instituto, escola ou centro é o vínculo atual ou mais recente? *

Instituto de Matemática e Estatística - IME

[Voltar](#) [Próxima](#) [Limpar formulário](#)

Nunca envie senhas pelo Formulários Google.

Este formulário foi criado em Universidade de São Paulo. [Denunciar abuso](#)

Google Formulários

Figura 5.4: Seção do formulário de empresas DNA USP com seleção de institutos

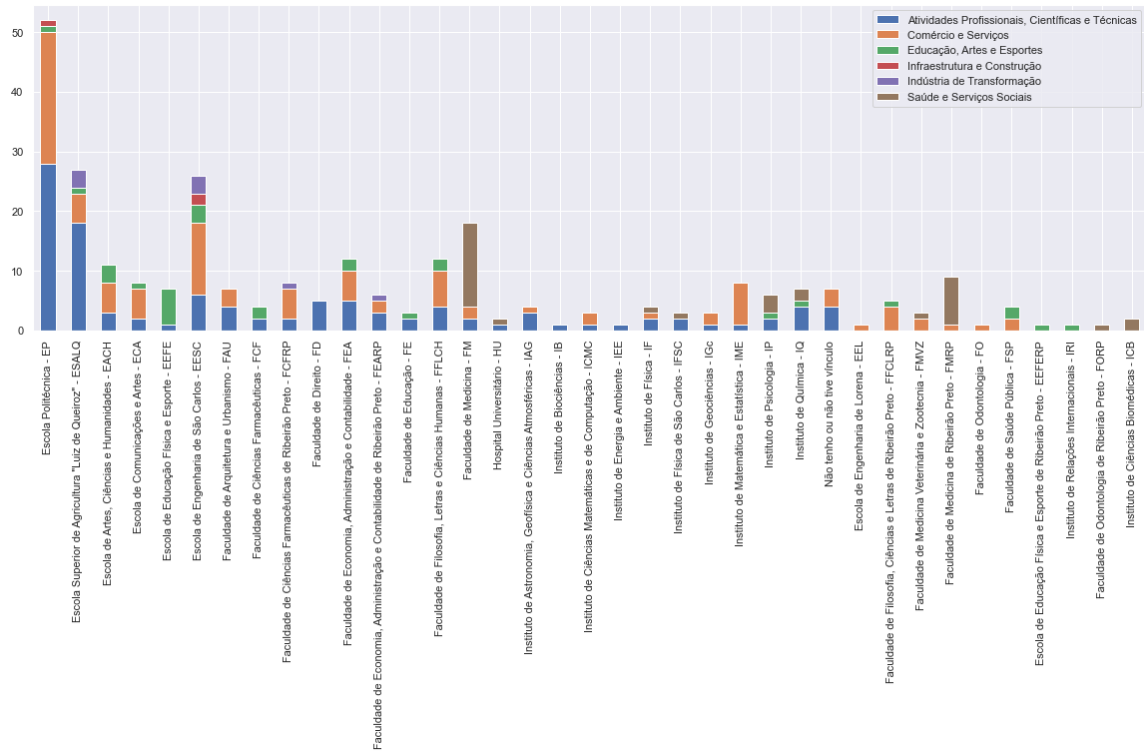


Figura 5.5: Percentual de empresas em cada instituto categorizadas por área de atuação

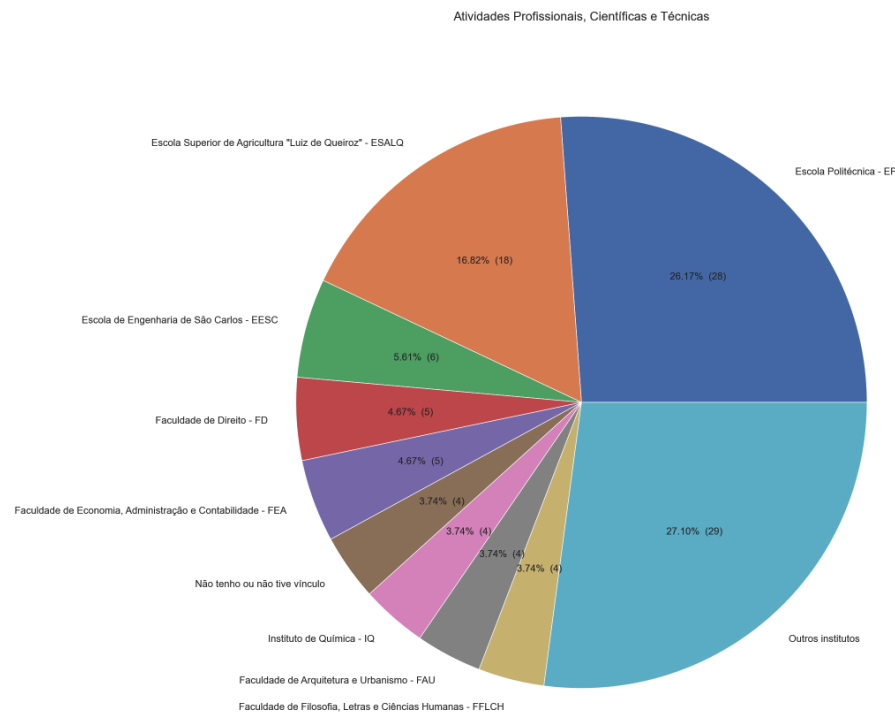


Figura 5.6: Gráfico de Setores — Área de Atuação por Instituto — Atividades Profissionais, Científicas e Técnicas

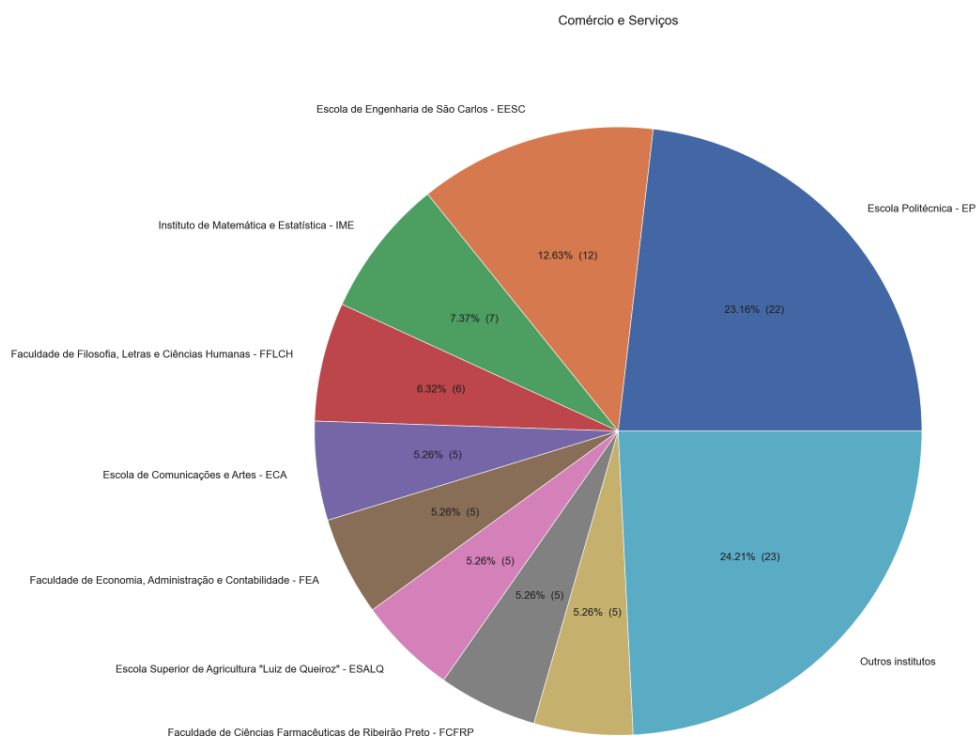


Figura 5.7: Gráfico de Setores — Área de Atuação por Instituto — Comércio e Serviços

sobre esse recorte do conjunto de dados é que estão cadastradas mais empresas vinculadas a Escola de Comunicação de Artes (ECA) no setor de Comércio e Serviços do que nas empresas cuja área de atuação é Educação, Artes e Esportes.

Nas empresas cujo foco é Saúde e Serviços Sociais, cuja distribuição é apresentada na figura 5.9, temos os dois polos de medicina da USP em destaque. A maioria das empresas representadas por esse grupo possuem vínculos com a Faculdade de Medicina, sendo 41,18% das empresas ao todo, enquanto em segundo lugar temos a Faculdade de Medicina de Ribeirão Preto com 23,5% do total. Os demais setores são compostos pelo Instituto de Psicologia (IP), Instituto de Química (IQ), Instituto de Ciências Biomédicas (ICB) e Faculdade de Odontologia de Ribeirão Preto (FORP). É curioso notar que a Faculdade de Odontologia localizada no Campus da Capital não possui nenhum vínculo informado até então. As únicas instituições não relacionadas diretamente com estudos na área de saúde e que aparecem entre as escolas que originaram empresas na área da saúde é o Instituto de Física (IF), tanto o de São Paulo como o localizado em São Carlos (IFSC).

Por fim, destaca-se ainda, com menor participação, empresas cuja área de atuação é Infraestrutura e Construção e Indústria de Transformação. Para o primeiro grupo, existem somente três empresas representativas distribuídas entre duas faculdades da USP: A Escola de Engenharia de São Carlos e a Escola Politécnica. Para o segundo grupo, associado a Indústria de Transformação, temos como mais representativas a ESALQ e novamente a Escola de Engenharia de São Carlos.

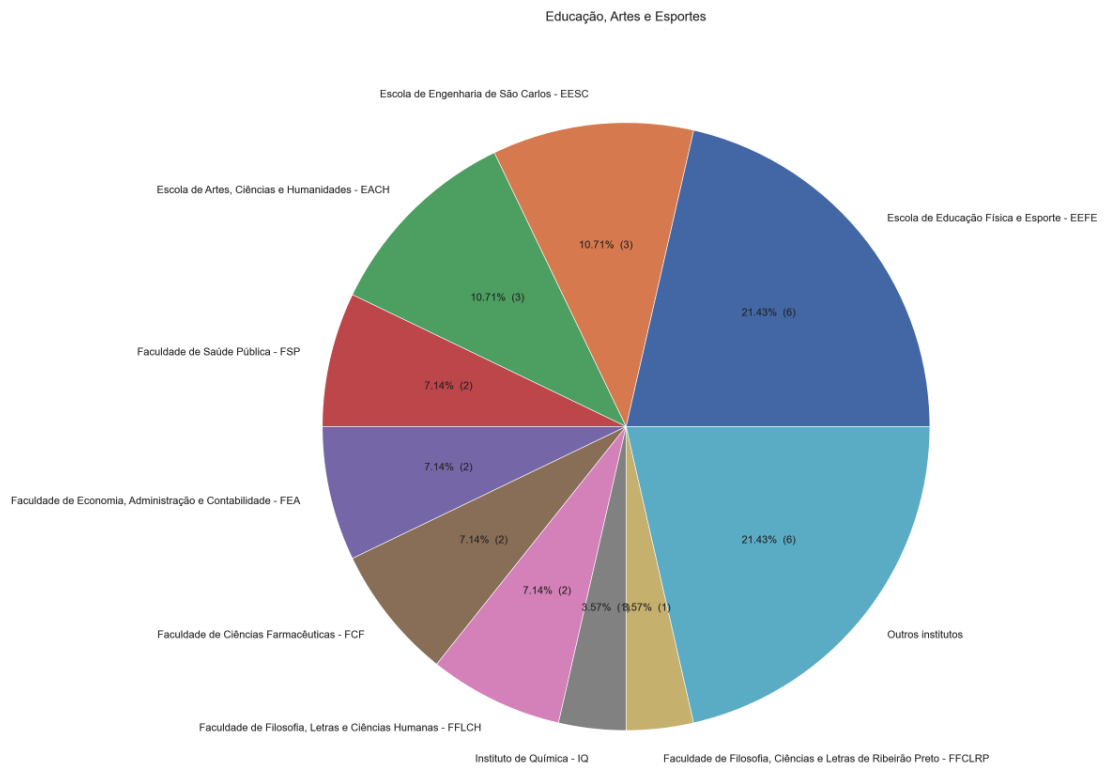


Figura 5.8: Gráfico de Setores — Área de Atuação por Instituto — Educação, Artes e Esportes

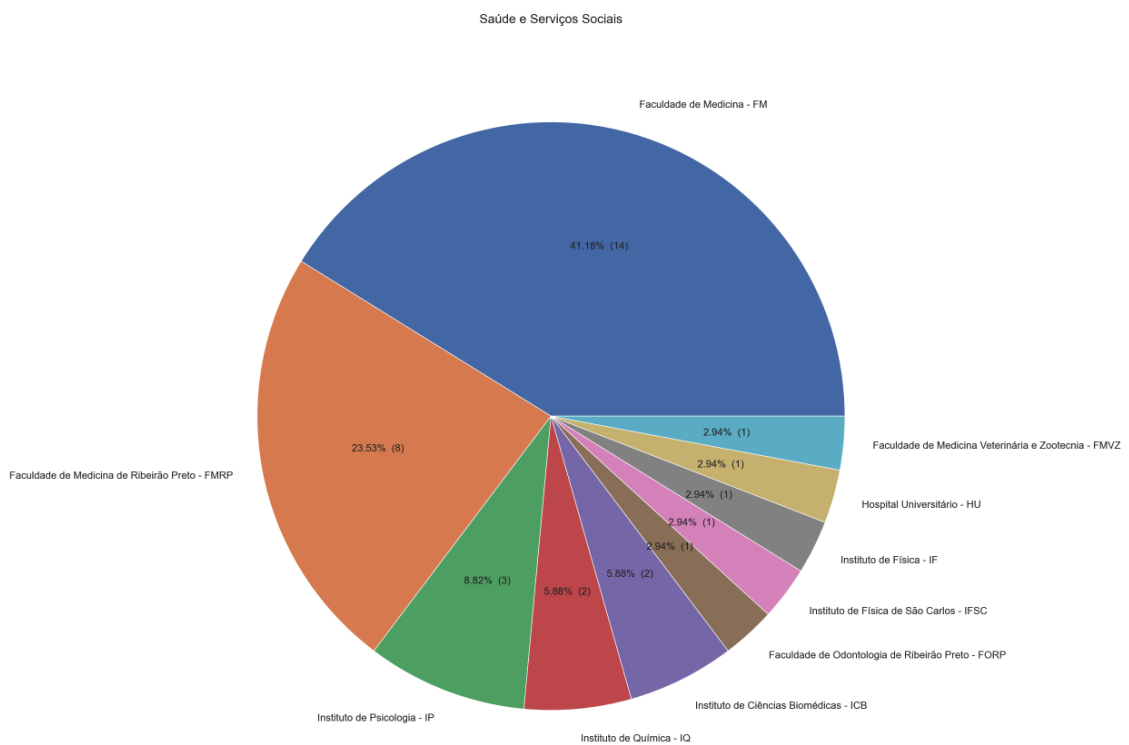


Figura 5.9: Gráfico de Setores — Área de Atuação por Instituto — Saúde e Serviços Sociais

5.3 Participação feminina

De acordo com o relatório da pesquisa GEM (**INSTITUTO BRASILEIRO DE QUALIDADE E PRODUTIVIDADE (2019)**), ocorreu uma diminuição da participação de mulheres em atividades empreendedoras no Brasil, seguindo uma tendência de queda no empreendedorismo brasileiro, que no último ano foi de 18%, também em virtude da pandemia de Sars-Cov-2 e da situação econômica geral do país. Peça fundamental para a obtenção da equidade entre mulheres e homens nas atividades econômicas, políticas e sociais, o empreendedorismo feminino encontra ainda muitos obstáculos. De acordo com o relatório global da GEM, na maioria das economias as atividades empreendedoras tendem a ser iniciadas por homens.

No Brasil, 46.2% das empresas tradicionais foram fundadas por mulheres, mas quando se coloca em foco a distribuição de gênero nas startups, o quadro é significativamente mais desequilibrado, como mostra o **FEMALE FOUNDERS REPORT (2021)**. De acordo com essa pesquisa, apenas 4.7% das startups brasileiras são fundadas por mulheres e 5.1% são fundadas por homens e mulheres. Portanto, cerca de 90% das empresas no ecossistema de inovação tem como fundadores somente homens. De acordo com o relatório, a notável diferença entre homens e mulheres pode ser justificada pelo fato de que iniciativas focadas no empreendedorismo feminino ainda são muito recentes no Brasil, de modo que 67% das startups com fundadoras foram fundadas somente nos últimos cinco anos. O desenvolvimento tardio e o histórico de vieses e tendências sexistas no mercado culminam no resultado ainda muito desigual que se observa atualmente.

De forma similar, na USP, a participação feminina ainda não está em patamar de igualdade. De acordo com dados do **ANUÁRIO DE ESTATÍSTICA DA USP (2020)**, o número de alunas matriculadas em 2020 totalizou 44835, ou 47% do total. Ainda assim, em algumas unidades da universidade a disparidade é notável. No Instituto de Matemática e Estatística, por exemplo, foram contabilizadas 600 pessoas do sexo feminino matriculadas em 2020, em contraste com 2042 do sexo masculino. Na Escola Politécnica, estão matriculadas somente 1756 alunas em relação a 6638 alunos.

A partir dessa realidade de desequilíbrio, buscou-se analisar os efeitos dela também nas atividades empreendedoras exercidas nas atividades a partir dos dados das empresas com DNA USP. No formulário de cadastro, não são solicitados dados sobre o sexo com o qual os sócios se identificam, de modo que um dado nessa direção teria que ser inferido a partir dos demais campos fornecidos. Para isso, foi utilizado o campo nome, extraíndo o primeiro nome informado por cada empreendedor e buscando a distribuição desse nome no Brasil utilizando para isso a API Nomes, disponibilizada pelo IBGE, que indica, em termos de frequência, a proporção de homens e mulheres com determinado nome a partir de dados do censo.

Por exemplo, podemos consultar na API Nomes a frequência do nome Francis em pessoas do sexo masculino nos dados do censo, a partir da chamada <https://servicodados.ibge.gov.br/api/v2/censos/nomes/Francis?sexo=M> que retorna o seguinte resultado, em formato JSON.

```
[
{
```

```

"nome": "FRANCIS",
"sexo": "M",
"localidade": "BR",
"res": [
  {
    "periodo": "[1930,1940[",
    "frequencia": 37
  },
  {
    "periodo": "[1940,1950[",
    "frequencia": 75
  },
  {
    "periodo": "[1950,1960[",
    "frequencia": 144
  },
  {
    "periodo": "[1960,1970[",
    "frequencia": 265
  },
  {
    "periodo": "[1970,1980[",
    "frequencia": 1034
  },
  {
    "periodo": "[1980,1990[",
    "frequencia": 3514
  },
  {
    "periodo": "[1990,2000[",
    "frequencia": 1927
  },
  {
    "periodo": "[2000,2010[",
    "frequencia": 499
  }
]
}
]

```

A mesma chamada pode ser feita mudando o parâmetro sexo para F, obtendo, portanto, o total de pessoas do sexo feminino em cada período que se chamam Francis. Para decidir qual sexo utilizar, a partir dessa perspectiva binária, foi optado por aquele de maior frequência.

Com isso, foi possível atribuir um provável sexo aos sócios cadastrados a fim de quantificar a distribuição de homens e mulheres no quadro societário das empresas com

DNA USP. É importante ressaltar que, uma vez que o dado indicador de gênero foi inferido, essa métrica não se expressa em termos exatos e é aberta a interpretações e correções.

A primeira pergunta que se buscou responder foi a respeito do modo como se distribuem homens e mulheres nas empresas cadastradas. O primeiro resultado confirmou a tendência de desigualdade observada nas atividades empreendedoras do ecossistema de inovação, conforme mostra a figura 5.10.

Percentual geral de pessoas fundadoras/sócias por sexo

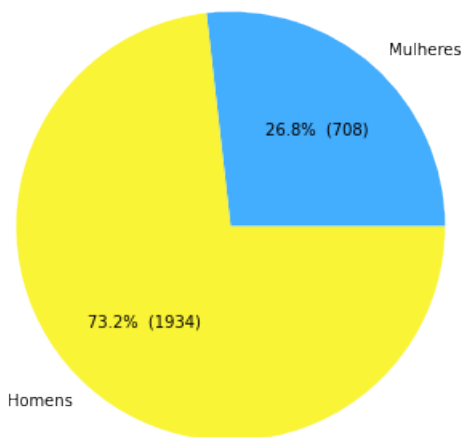


Figura 5.10: Gráfico de Setores — Distribuição de empresas DNA USP com mulheres e homens no quadro de sócios

Observa-se, a partir da métrica analisada, que o total de empresas que possuem mulheres é de apenas pouco mais de 26% do total, fato que pode ser correlacionado com o fato dos institutos com mais empresas terem pouca participação feminina, como mostrado na tabela 5.1, montada com dados extraídos do ANUÁRIO DE ESTATÍSTICA DA USP (2020).

Unidade	Mulheres	Homens
Escola Politécnica	1756	6638
Escola de Engenharia de São Carlos	1086	3215
Escola Superior de Agricultura "Luiz de Queiroz"	1603	1830
Faculdade de Medicina	2447	1830
Faculdade de Economia, Administração e Contabilidade	1250	2856

Tabela 5.1: Total de mulheres e homens nos cinco institutos com mais empresas DNA USP cadastradas

Os gráficos abaixo mostram o percentual de sócios agrupados pelo gênero identificado, para cada instituto cadastrado em empresas DNA USP.

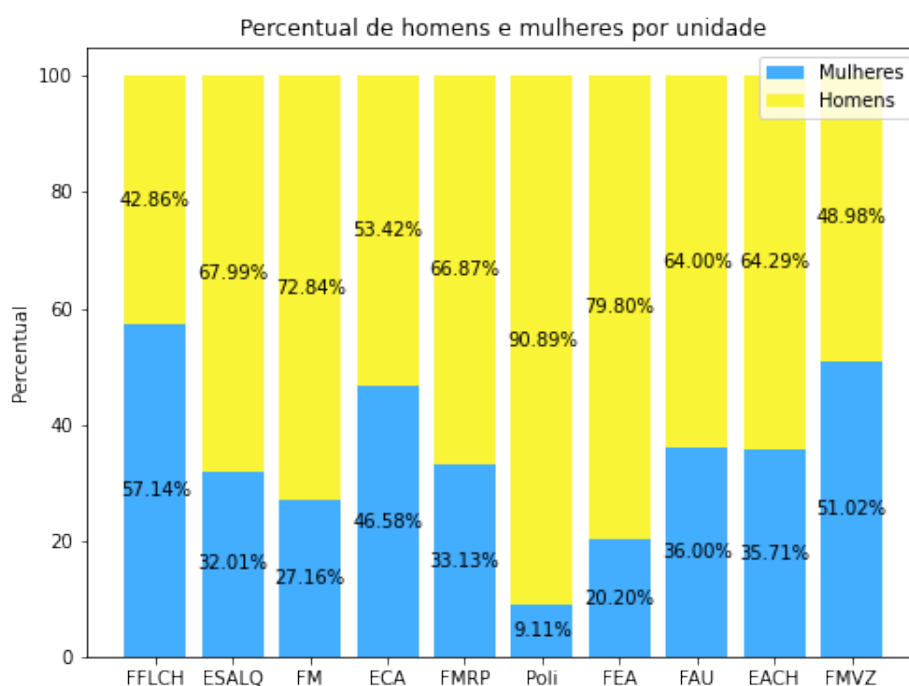


Figura 5.11: Distribuição de empresas DNA USP com mulheres e homens no quadro de sócios por instituto

O gráfico exibido na figura 5.11 evidencia que na unidade que possui mais atividades empreendedoras registradas, o percentual de mulheres que compuseram o corpo de sócios é de somente 9.11%. A taxa mais alta foi observada na FFLCH, que se destaca com empresas que atuam no setor de comércio e serviços e também no setor de educação, artes e esportes, conforme mostrado na seção 5.2.

A partir da mesma visualização com as demais unidades, na figura 5.12, observa-se que a maioria delas concentram mais homens do que mulheres, ainda que em algumas delas isso ocorra principalmente pela pequena quantidade de empreendedores.

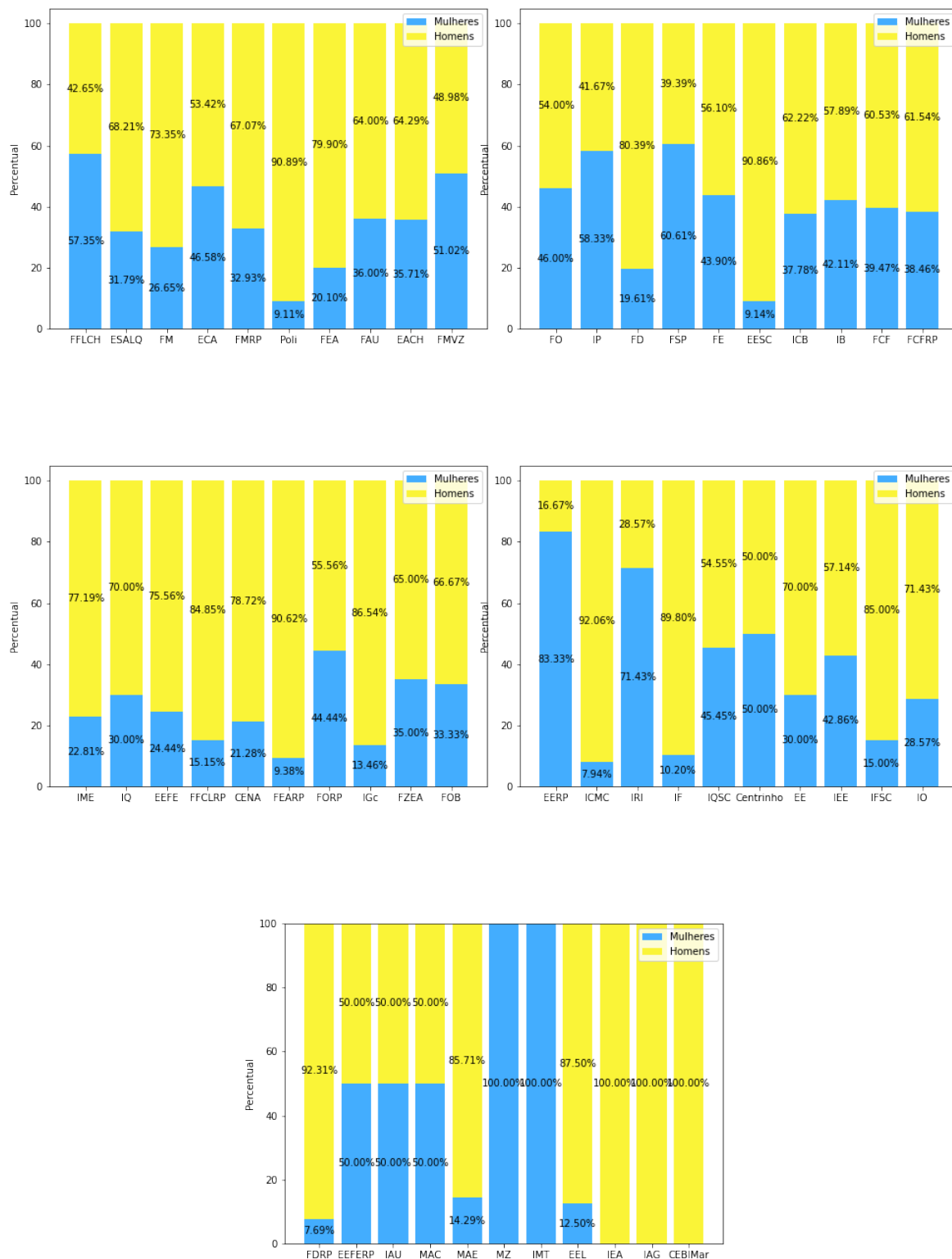


Figura 5.12: Distribuição percentual de mulheres e homens nas empresas agrupadas por unidade

Foi feita também a análise da quantidade de homens e mulheres categorizados pelas áreas de atuação das empresas em que trabalham, seguindo os agrupamentos utilizados na Classificação Nacional de Atividades Econômicas (CNAE). Para todos os grupos, observou-se uma maioria masculina em relação à quantidade de sócios. Por conta de alguns dos setores possuírem um baixo número de empresas cadastradas, algumas das proporções observadas foram significativamente desiguais, por isso, foi apontado o total de homens e mulheres também em números absolutos.

No grupo que contempla atividades científicas e técnicas, o que mais possui empresas vinculadas, somente 27.4% das pessoas cadastradas como associadas as empresas são do sexo feminino, como mostrado na figura 5.13.

Atividades Profissionais, Científicas e Técnicas

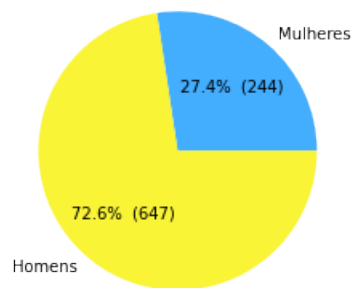


Figura 5.13: Distribuição de homens e mulheres em empresas que exercem Atividades Profissionais, Científicas e Técnicas.

No setor de comércio e serviços, mais de 80% dos associados são homens, em um total de 860 pessoas sócias dessas empresas. Para empresas da área de Educação, Artes e Esportes, a diferença diminui mas ainda existe a maioria masculina, representando 61.5% do total de sócios. Nas demais áreas, a diferença se repete, sendo que em algumas a participação feminina é timidamente maior, conforme mostra a figura 5.14.

5.3.1 Propostas de incentivo

Apesar da situação observada na USP refletir a situação de inequidade entre homens e mulheres no mercado de trabalho e sobretudo em atividades de incentivo ao empreendedorismo, é importante ressaltar que a USP tem incentivado esforços que visam reverter esse cenário e proporcionar condições que possibilitam uma participação mais abrangente do público feminino.

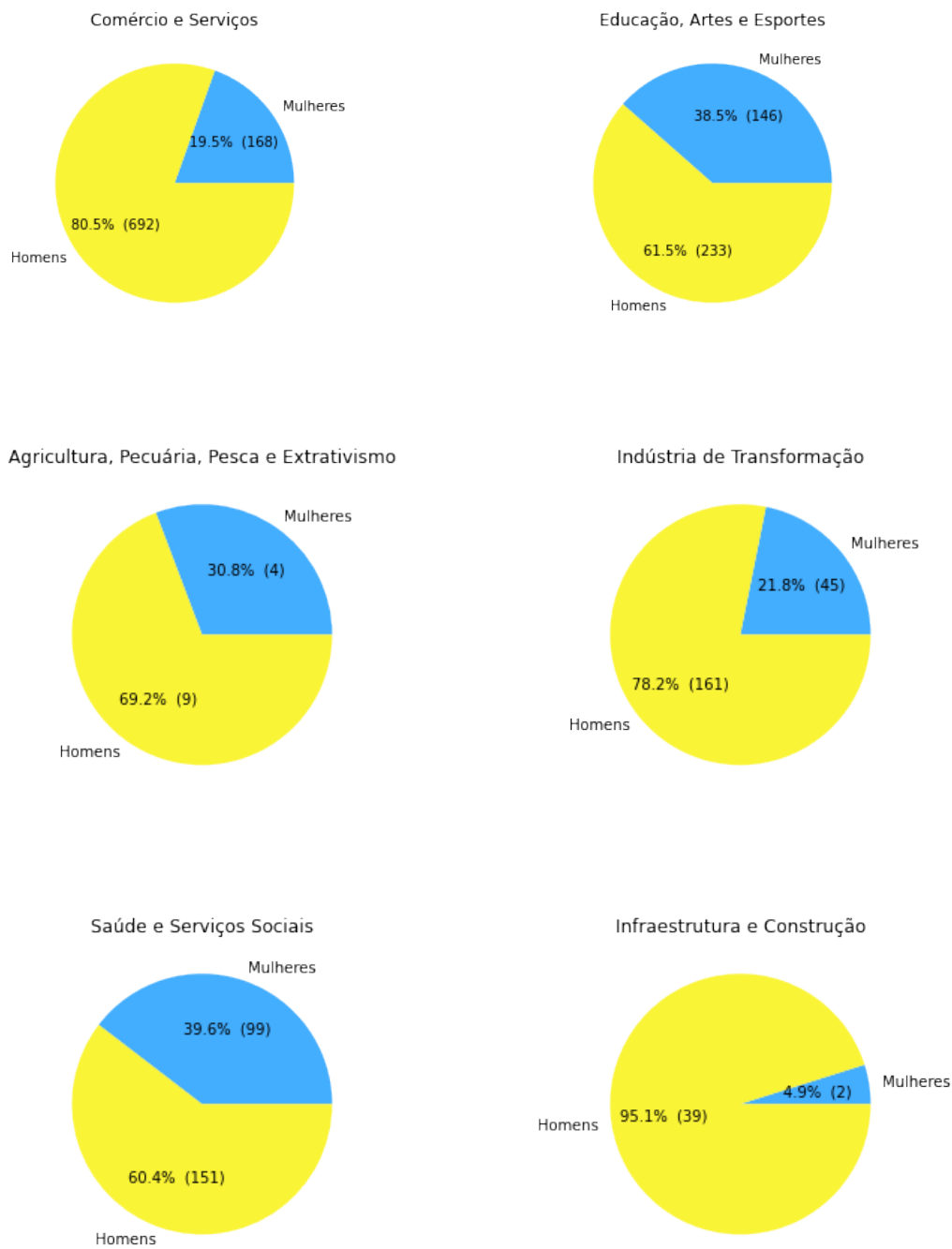


Figura 5.14: Distribuição feminina e masculina por CNAE

5.4 Localização das empresas

A contribuição positiva da USP para a economia do Estado de São Paulo é incontestável. O incentivo ao empreendedorismo no ambiente universitário foi um fator-chave para isso, dado o notável número de empresas fundadas a partir da universidade. Contudo, os dados disponibilizados permitem afirmar que a contribuição direta do empreendedorismo difundido na USP extrapola os limites estaduais, uma vez que existem empreendimentos com DNA USP surgindo em todo o país. De modo a visualizar a distribuição de tais companhias pelo território nacional, foram elaboradas visualizações no formato de mapa interativo, conforme a figura 5.15. O mapa pode ser acessado em <https://dnausp-web.vercel.app/mapa>.



Figura 5.15: Mapa interativo exibindo a distribuição de empresas no território brasileiro

Os dados utilizados para obter a localização de cada companhia foram obtidos a partir do cruzamento dos dados de endereço informados no momento do cadastro e de um conjunto de dados contendo para cada município brasileiro um par de coordenadas representando latitude e longitude. Dessa forma, foram incluídas somente aquelas empresas que possuem municípios válidos informados.

Região sudeste

Destaca-se a região sudeste por ser a região com o maior número de empresas DNA USP, representando cerca de 74% do total de empresas cadastradas. A figura 5.16 mostra que, sobretudo em São Paulo, temos mais de 1300 empresas cadastradas, de um total de cerca de 2500.

Demais regiões

Todas as regiões brasileiras possuem ao menos uma empresa DNA USP cadastrada. Exceto pela região sudeste, nas demais regiões o vínculo da USP com as companhias se dá

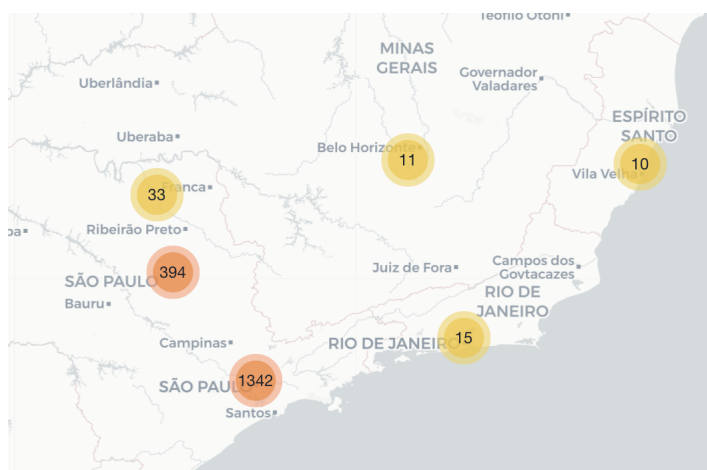


Figura 5.16: Empresas na região sudeste

principalmente pelo vínculo que os sócios e sócias tem ou já tiveram com a instituição, dado o baixo número de empresas incubadas ou graduadas que pertencem a essas regiões, como mostrado nas figuras 5.17 e 5.18.



Figura 5.17: Empresas incubadas ou graduadas nas demais regiões



Figura 5.18: Empresas que foram direto para o mercado

5.5 Spin-offs

Além de empresas que são classificadas como startups, na base de dados do HUB USP Inovação temos empresas que são caracterizadas pelo termo *spin-offs*. Uma empresa spin-off é uma empresa fundada como derivação de outra instituição. As empresas spin-offs DNA USP são majoritariamente spin-offs acadêmicas, pois surgem em espaços da universidade como produto de ensino e pesquisa. As spin-offs acadêmicas são um meio de introduzir no mercado os produtos desenvolvidos em pesquisas acadêmicas.

A partir dos dados, foram caracterizadas como spin-offs as empresas que possuem ao menos um sócio com vínculo de pós-graduando, pós-graduado ou docente. Também foram consideradas aquelas cujos sócios estejam associados a alguma patente registrada, inferindo que a empresa vinculada a esse sócio surgiu a partir da inovação patenteadada.

5.5.1 Spin-offs por pós-graduação

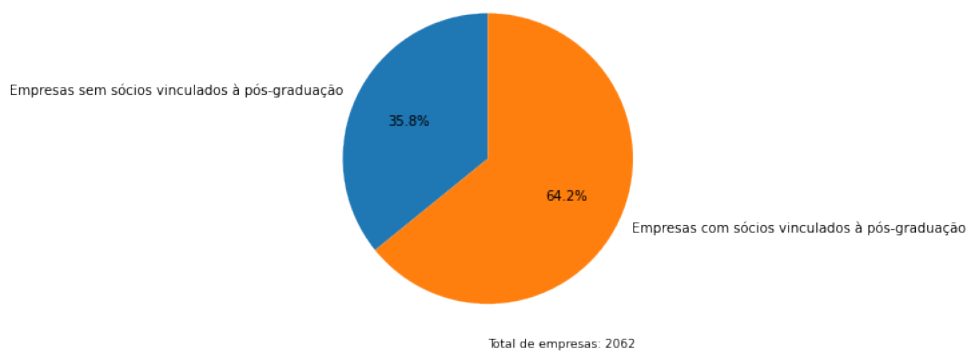


Figura 5.19: *Proporção de empresas com sócios vinculados a programas de pós-graduação*

O gráfico na figura 5.19 mostra que a maioria das empresas DNA USP tem pelo menos um sócio cadastrado que possui vínculos com programas de pós-graduação da Universidade de São Paulo, indicando uma correlação entre a adesão a tais cursos e a concretização da atividade empreendedora, uma vez que as atividades de pesquisa realizadas em cursos de pós-graduação culminam em significativos aumentos nos índices de inovação. O vínculo de sócios fundadores à pós-graduação é um forte argumento contra a falácia que afirma que as práticas de pesquisa e inovação exercidas pelas universidades brasileiras estão afastadas e alheias aos movimentos do mercado.

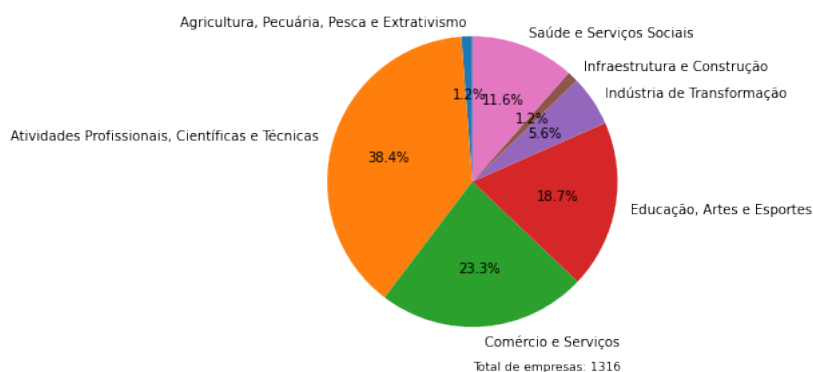


Figura 5.20: Distribuição de CNAE de empresas com sócios vinculados a pós-graduação

Ao analisar pelo CNAE correspondente, nota-se que as empresas com maior número de sócios com vínculos na pós-graduação são aquelas do setor de Atividades Profissionais, Científicas e Técnicas, como mostrado na figura 5.20. Comércio e Serviços aparece em segundo lugar, logo na frente de Educação, Artes e Esportes. Os setores de Infraestrutura e Construção e o de Agricultura, Pecuária, Pesca e Extrativismo são os que menos apresentam spin-offs por vínculos com pós, também por serem os setores com menor amostragem de empresas, conforme mostrado na figura 5.1.

Em relação aos institutos que mais possuem empreendedores alunos ou ex-launos de pós-graduação, temos em destaque a Escola Politécnica, a ESALQ e a Faculdade de Medicina. A figura 5.21 mostra as 20 unidades com as maiores proporções de spin-offs por vínculo de pós-graduação.

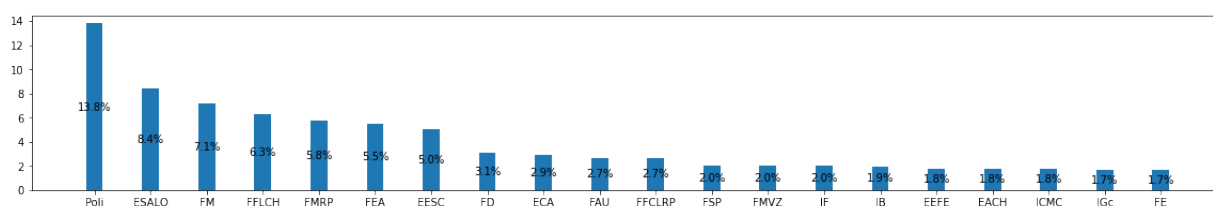


Figura 5.21: Percentual de sócios com pós-graduação por unidade da USP vinculada

5.5.2 Spin-offs por patentes

O registro de patentes é um importante indicador do índice de inovação. De acordo com dados divulgados pelo INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL (2020), dos dez principais depositantes de patentes de invenções no Brasil em 2020, nove deles são universidades federais ou estaduais, conforme mostra a tabela 5.2 extraída dos dados citados.

Depositante	Total de depósitos
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE	96
PETROBRAS	79
UNIVERSIDADE FEDERAL DA PARAIBA	74
UNIVERSIDADE FEDERAL DE MINAS GERAIS	63
UNIVERSIDADE ESTADUAL PAULISTA JULIO DE MESQUITA FILHO	55
UNIVERSIDADE FEDERAL DE PERNAMBUCO	55
UNIVERSIDADE DE SÃO PAULO	51
UNIVERSIDADE ESTADUAL DE CAMPINAS	50
UNIVERSIDADE FEDERAL DE PELOTAS	38
UNIVERSIDADE FEDERAL DE UBERLÂNDIA	38

Tabela 5.2: Dez principais depositantes de patentes de invenção, segundo relatório do INPI em 2020.

Dentre patentes depositadas pela Universidade de São Paulo, que em 2020 ficou na sétima posição do *ranking* de patentes de invenção no país, temos algumas delas vinculadas às empresas DNA USP, o que caracteriza tais companhias também como spin-offs, por serem originadas ou derivadas da concretização e registro dessas invenções.

Para analisar esse aspecto das empresas DNA USP, utilizou-se o conjunto de dados de registros de patentes disponibilizado por membros do Hub USP Inovação, em que um registro de CIP (Classificação Internacional de Patente) é vinculada a um nome de inventor associado a USP. Para correlacionar com os dados de empresas DNA USP, foram correlacionados os nomes dos inventores com os dos sócios cadastrados utilizando para isso a distância de Levenshtein. Essa técnica foi aplicada uma vez que não haviam outros dados passíveis de comparação entre os dois conjuntos. Constatou-se que 84 das empresas cadastradas tem algum vínculo com inventores de patentes.

Filtrando as empresas com patentes registradas por CNAE, obtemos as maiores proporções para os setores de Atividades Profissionais, Científicas e Técnicas e de Comércio e Serviços, como mostrado na figura 5.22.

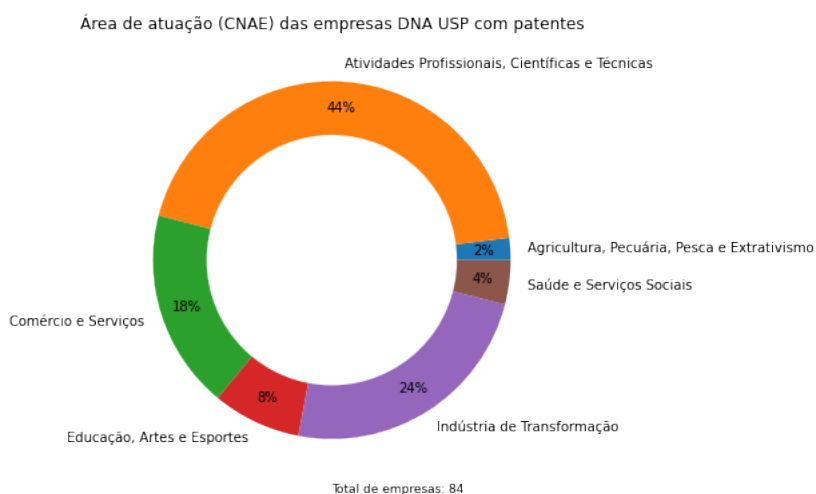


Figura 5.22: Proporção de empresas com patentes registradas por CNAE

Dentre as empresas que registraram patentes, seu institutos de origem são elencados no gráfico contido na figura 5.23.

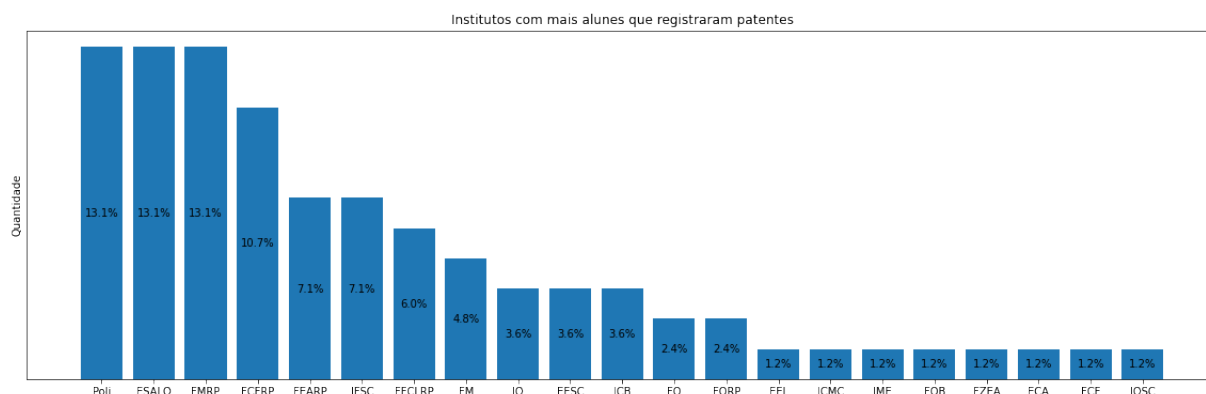


Figura 5.23: *Proporção de empresas com patentes registradas por instituto*

Além dos dados exibidos, o vínculo da universidade com o registro de invenções pode também se dar de forma indireta, a partir de parcerias com grandes empresas e com iniciativas externas à instituição. Por exemplo, em 2019, a USP (SÃO PAULO, 2019) regulamentou o compartilhamento e a permissão de uso de seus equipamentos, infraestrutura, materiais e instalações em ações voltadas a desenvolvimento e inovação tecnológica, o que viabiliza e potencializa o desenvolvimento de novas patentes.

Também se pode citar a parceria da Escola Politécnica com a Shell e a FAPESP, formando o *Research Centre for Greenhouse Gas Innovation* (SÃO PAULO FAPESP (2016)), que realiza pesquisas inovadoras no setor de gás, propondo soluções sustentáveis e eficientes e criando um campo fértil para inovação.

Para aprimorar as análises apresentadas, fazendo com que elas fossem facilmente personalizadas e repetidas sob demanda para novos dados, foi implementado um sistema computacional que será apresentado no próximo capítulo.

Capítulo 6

Desenvolvimento

Nesse capítulo será apresentado o processo de desenvolvimento de software no contexto deste trabalho de maneira detalhada, a fim de descrever os aspectos mais importantes e justificar as escolhas feitas ao longo da realização do trabalho. O código está dividido em três pacotes, sendo um deles um cliente *web* para utilização pelo usuário final, outro para servir esse cliente *web* por meio de chamadas *HTTP*, e um pacote comum, que é utilizado por ambos os outros pacotes.

6.1 Objetivo do sistema

A partir das análises realizadas, buscou-se implementar um processo para que se tornassem reproduzíveis com conjuntos de dados novos ou atualizados de maneira simplificada e tão automatizada quanto possível, além de forma compatível com o sistema de ingestão de dados atualmente utilizado pelos clientes. Dessa forma, as análises apresentadas deixam de ser apenas relatórios pontuais, mas informações disponíveis sob demanda.

6.2 Versão inicial

A versão inicial do projeto foi feita utilizando uma interface web e foi implementada utilizando o arcabouço em Javascript *NextJS*, que possui suporte para renderização do lado do servidor (*Server-Side Rendering*), dessa maneira, foram utilizados os dados processados pelas ferramentas *Python* e *Pandas* para exibir os gráficos.

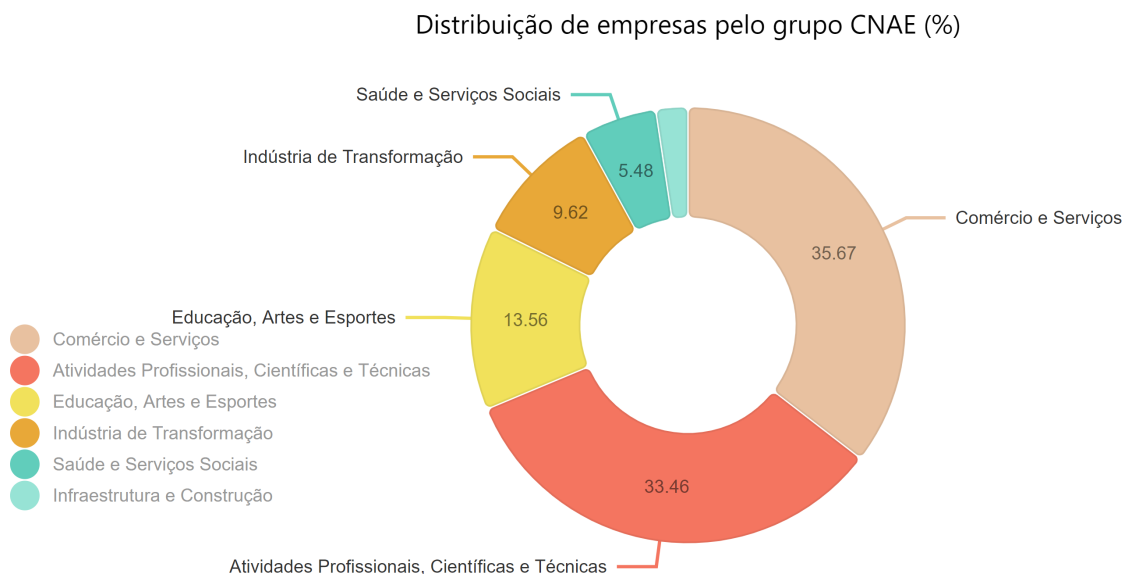


Figura 6.1: Exemplo de visualização da versão inicial.

6.3 Módulo core

Para tornar o processamento dos dados, como feito na versão inicial, escalável e simples de se repetir, mostrada na figura 6.1, foi desenvolvido um módulo *core* (núcleo), contendo as lógicas de validação, e representando as entidades derivadas do modelo de dados (contida na figura 4.3).

Este módulo foi publicado como um pacote no NPM, repositório para pacotes *Javascript*, para que seja possível utilizá-lo em qualquer plataforma, seja do lado do cliente ou do servidor. Dessa maneira, a coesão entre os objetos compartilhados dentre as plataformas é maior, pois passaram pelo mesmo processo de validação e instanciação. O pacote, cuja página está representada na figura 6.2, pode ser visualizado em `@dnausp/core` e acessado em <https://github.com/lfujiwara/dnausp-core>.

6.3.1 Tecnologias utilizadas

Typescript

Typescript é uma linguagem de programação criada pela *Microsoft* e um superconjunto da linguagem *Javascript*, oferecendo funcionalidades como tipagem (estática), *namespaces* e *enums*, características que aproximam *Typescript* de linguagens orientadas a objetos já estabelecidas, como *C#* em questão de familiaridade de sintaxe. Além disso, *Typescript* é uma linguagem transpilada para *Javascript*, linguagem adotada para o ambiente web e pelos navegadores modernos, logo, é possível utilizar o código *Typescript* tanto em aplicações para o lado do cliente (navegador) quanto do lado do servidor (*runtimes* para *Javascript*, como *NodeJS* e *Deno*). A versão alvo *Javascript* para o resultado da transpilação é definida pelo usuário, que deve escolhê-la de maneira apropriada, pois o conjunto de

@dnausp/core TS

0.0.1-alpha.16 • Public • Published 2 days ago

[Readme](#) [Explore](#) BETA [2 Dependencies](#) [0 Dependents](#) [17 Versions](#) [Settings](#)

@dnausp/core

Pacote contendo código de domínio e aplicação.

Índice

- `domain` : Contém as entidades.
- `app` : Contém queries/mutations `queries/mutations` e especificações para portas/adaptadores (`ports`) em classes abstratas.

Trabalho realizado para a disciplina Trabalho de Formatura Supervisionado (MAC0499) do IME - USP

Install

```
> npm i @dnausp/core
```

Repository

[github.com/lfujiwara/dnausp-core](#)

Homepage

[github.com/lfujiwara/dnausp-co...](#)

Weekly Downloads

541

Version	License
0.0.1-alpha.16	MIT

Figura 6.2: Pacote core publicado no repositório NPM

funcionalidades pode variar dependendo da versão escolhida. A mais antiga disponível como alvo é a *ECMAScript 3*, também referida como *es3*, lançada em 1999. Neste projeto, foi utilizada como versão alvo a *ECMAScript 2017* (*es2017*).

Yarn

Yarn é um gerenciador de pacotes e de projeto para *NodeJS*. Essa ferramenta foi utilizada para instalar pacotes utilizados no projeto, definir scripts para executar testes, verificação de tipos e qualidade de código, além de utilizá-la para publicar o código no *NPM*.

Jest

Para oferecer garantias de qualidade e simplificar o trabalho de uma futura extensão desse módulo, foi utilizada a ferramenta *Jest*, mostrada na figura 6.3. *Jest* é uma suíte de testes mantida pelo *Facebook* focada no ecossistema *Javascript*, ao executá-la, ela procura por testes definidos em arquivos com extensão `.spec.ts` ou `.test.ts`, os executa e gera relatórios indicando aqueles que falharam, além de produzir relatórios de cobertura como os mostrados na figura 6.4, que indicam quais linhas de código, expressões, bifurcações (*branches*), funções e classes tiveram sua execução testada, inclusive exibindo a proporção. A partir desse relatório, foi determinado um alvo de 90% de cobertura para as linhas de código, isto é, que 90% das linhas de código devem ter sido executadas durante os testes.



The image shows a screenshot of the Jest Test Results interface. The root node is 'Test Results' with a total execution time of 150 ms. It is expanded to show a list of test files and their sub-items, all of which passed. The execution times for each item are listed on the right side of the table.

Test File / Item	Execution Time
Test Results	150 ms
valor-inteiro-anual.spec.ts	15 ms
Valor inteiro anual	15 ms
Factory method (create)	1 ms
Constructor	14 ms
agregado-anual.spec.ts	17 ms
Agregado anual	17 ms
Stores the values correctly	8 ms
Adds a value correctly	1 ms
Fails to add a value that already exists	0 ms
Throws an error if values are set with an invalid array	8 ms
upsert-empresa.mutation.spec.ts	13 ms
investimento.spec.ts	6 ms
add-faturamento.mutation.spec.ts	4 ms
cnae.spec.ts	26 ms
cnpj.spec.ts	12 ms
remove-faturamento.mutation.spec.ts	6 ms
empresa.factory.spec.ts	3 ms
investimento.spec.ts	10 ms
registro-anual.spec.ts	8 ms
vinculo-universidade.spec.ts	12 ms
incubacao.spec.ts	12 ms
valor-inteiro-positivo-anual.spec.ts	6 ms

Figura 6.3: Visualização dos testes realizados com Jest

All files

86.53% Statements 540/624 56.32% Branches 98/174 89.65% Functions 104/116 90.07% Lines 508/564

Press *n* or *j* to go to the next uncovered block, *b*, *p* or *k* for the previous block.

Filter:

File	Statements	Branches	Functions	Lines
app/mutations	93.87%	46/49	75%	9/12
domain	80.71%	159/197	64.1%	50/78
domain/agregados-anuais	90%	27/30	60%	3/5
domain/enums	100%	117/117	100%	6/6
domain/factories	77.18%	115/149	24.48%	12/49
domain/valores-anuais	92.68%	76/82	75%	18/24

Figura 6.4: Visualização do relatório de cobertura do módulo core produzido pela ferramenta Jest

GitHub

O código desse módulo foi armazenado no *GitHub*, que oferece suporte a repositórios Git, *Pull Requests* e integração e entrega contínua por meio da ferramenta *GitHub Actions*, que utiliza configuração por meio de arquivos *yaml* para executar tarefas quando o repositório recebe novos *commits* ou *tags*. Entre essas tarefas, podem ser executados testes ou compilação e publicação de código em repositórios como o *NPM*. Para o escopo do projeto, foi utilizado o *GitHub Actions* para realizar testes, checagem de tipos e qualidade de código a cada novo *commit* na *branch* principal (*main*). A interface de uma *pipeline* do *GitHub Actions* pode ser vista na figura 6.5.

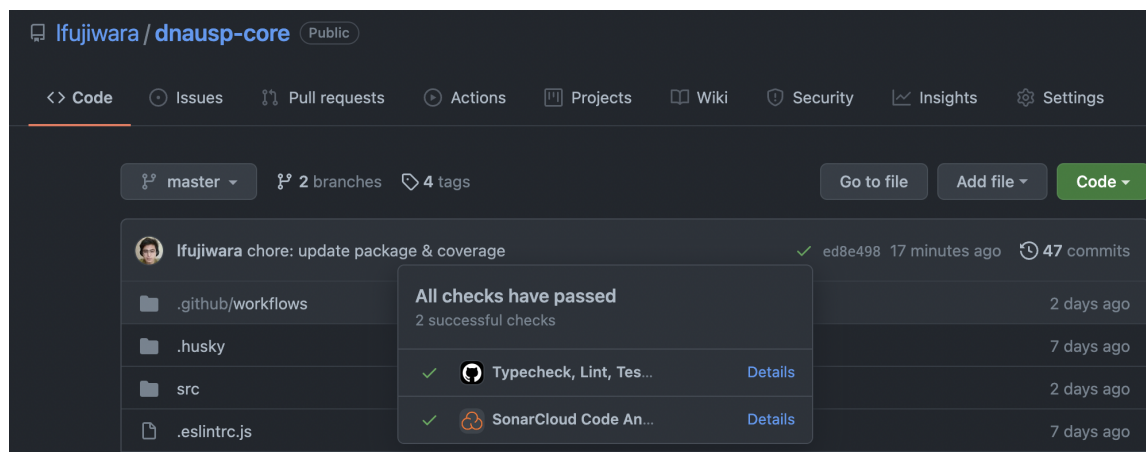


Figura 6.5: Visualização da execução de testes, checagem de tipos e qualidade de código através do painel *GitHub Actions* do *GitHub*

6.3.2 Visão geral

O código foi dividido em duas seções, uma delas contendo o domínio, onde se encontram as entidades do projeto definidas em classes e relacionamentos, em outra seção, contendo o

código de aplicação, onde se encontram casos de uso e especificações de suas dependências (objetos consumidos pelos casos de uso), além de especificações para consultas (*queries*). Dessa forma, o projeto se aproxima da definição de Arquitetura Limpa, apresentada em detalhes em [ROBERT MARTIN, 2017](#).

Domínio

O domínio da aplicação é concentrado em torno da classe *Empresa*, que possui alguns atributos escalares (incluindo identificadores) e *arrays* para atributos sem lógica complexa definida e agregados (como definido em [FOWLER, 2013](#)) como histórico de faturamentos e quadro de colaboradores para atributos que envolvem conjuntos de objetos que devem ser tratados como apenas um, onde devem ser garantidos invariantes, como a unicidade de apenas uma entrada de faturamento por ano para cada empresa. Outros atributos com lógicas complexas, como CNPJ e CNAE também foram separados em classes diferentes e referenciadas como atributos dentro da classe *Empresa*, tendo em vista o princípio *Composition over inheritance* (traduzido para o português como "preferir composição ao invés de herança") apresentado em [GANG OF FOUR, 1994](#). O diagrama de classes desse domínio pode ser visto na figura 6.6.

Fábricas

A instanciação de objetos do domínio ocorre por meio de fábricas, contidas no domínio, que são métodos da própria classe instanciada, ou de outras classes que possuem essa finalidade. As fábricas são responsáveis por receber informações como argumentos de funções e realizar a validação desses dados e a instanciação da classe desejada. Também foram criados métodos de validação para serem consumidos pelas fábricas.

No projeto, as classes que representam as entidades possuem métodos estáticos como fábricas, para objetos mais complexos, foram criadas classes abstratas auxiliares que contêm apenas fábricas, como exibido na figura 6.7.

O fluxo de execução das fábricas se dá por meio de mônadas, que funcionam como tipos invólucros para outros tipos e possuem atributos para indicar qual tipo é válido. Por exemplo, no projeto são utilizadas mônadas com um tipo representando sucesso, e outro tipo representando erro, a mônada possui um atributo `isOk` ou equivalente para indicar se o valor é um sucesso ou uma falha, além de método `unwrap` e `unwrapFail` para retornar o valor de sucesso ou erro. Dessa maneira, é possível lidar com erros de validação e outros possíveis problemas sem lançar exceções e alterar o fluxo de execução de maneira brusca.

No entanto, nos construtores, se quaisquer asserções que garantam a integridade dos objetos falharem, teremos como resultado uma exceção. A finalidade desse aspecto do código é garantir a integridade de qualquer objeto instanciado.

Aplicação

A parte de aplicação desse módulo contém a definição de casos de uso (referenciados como *mutations* dentro do código), especificação das dependências dos casos de uso (*ports*) e especificação de consultas *queries*.

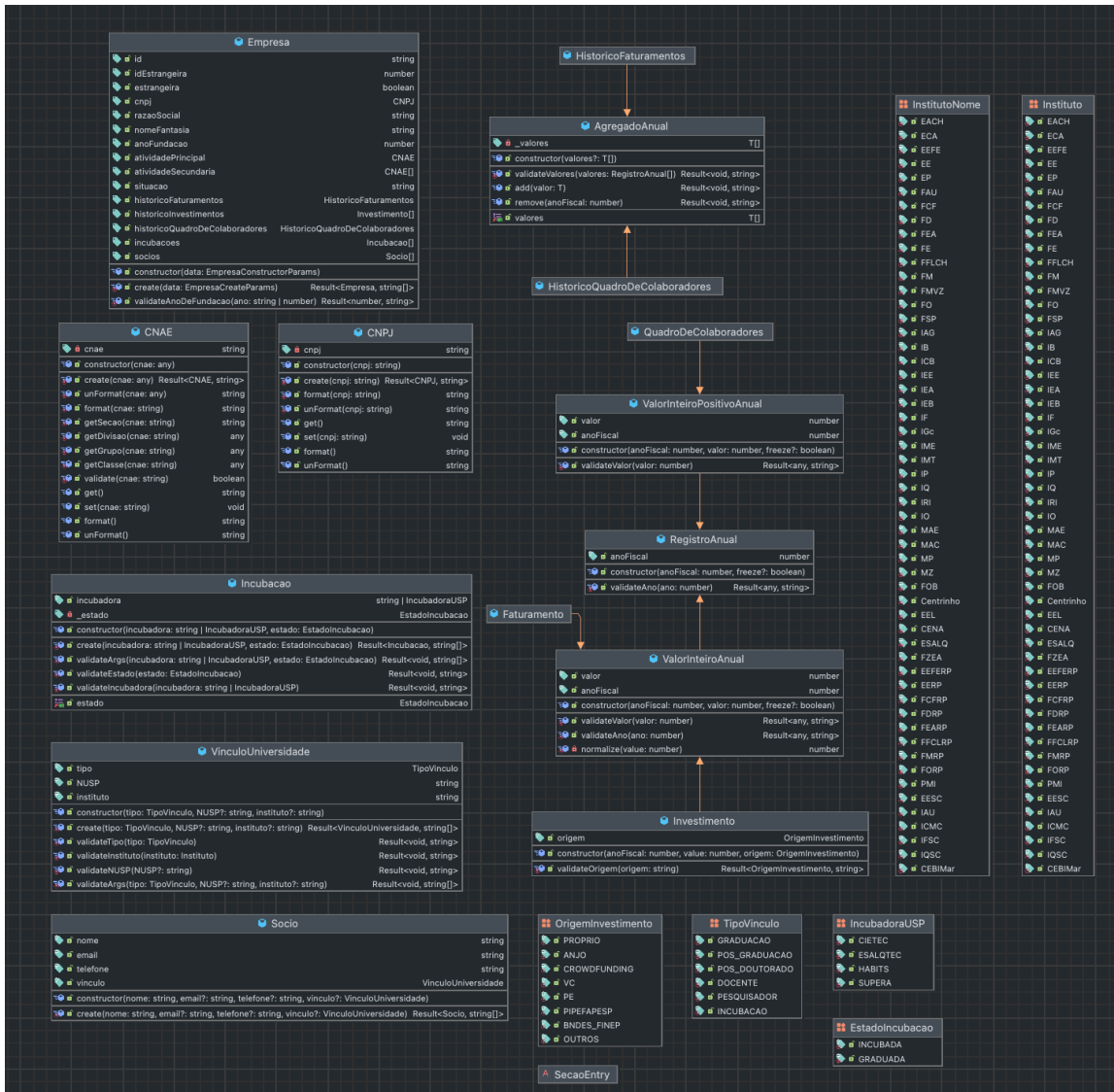


Figura 6.6: Diagrama de classes gerado pelo ambiente de desenvolvimento integrado WebStorm

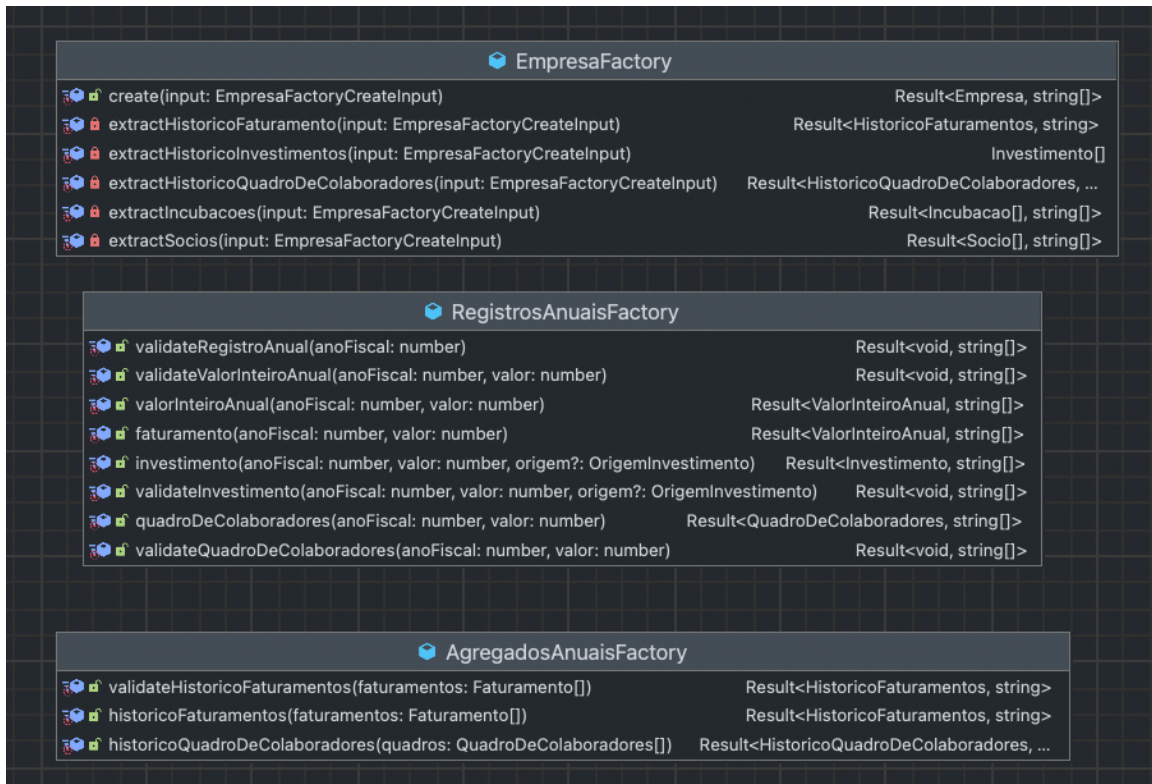


Figura 6.7: Diagrama das classes que agregam fábricas gerado pelo ambiente de desenvolvimento integrado WebStorm

Os casos de uso são operações que realizam modificações na base de dados, estão representadas como classes que possuem métodos execute para realizar a operação desejada, esse métodos realizam as consultas em bases de dados, instanciação, validação e operações sobre os objetos. Mas não realizam de maneira direta as consultas e modificações na base de dados, pois se considera que as interações com bases dados, sejam elas bancos de dados relacionais, documentos ou até mesmo arquivos de texto, são apenas detalhes de implementação, portanto, os casos de uso recebem em seu construtor estruturas que realizam essas operações, que chamaremos de adaptadores para bases de dados. Essa técnica é chamada de injeção de dependência por meio de construtor, serve para atender ao princípio de inversão de dependência, postulado em [ROBERT MARTIN, 1994](#).

As dependências dos casos de uso são definidas como classes abstratas, as assinaturas das funções definem o que é esperado que a implementação concreta dessas classes deve receber e retornar. O uso de interfaces foi descartado, pois a linguagem *Typescript* é sempre transpilada para *Javascript*, onde essas estruturas não existem em tempo de execução. Dessa maneira, os casos de uso possuem como dependência apenas o domínio e a especificação das dependências, que podemos chamar de *Gateways* para o contexto desse projeto, atendendo aos requisitos da Arquitetura Limpa, citada anteriormente. Um diagrama esquematizando os principais componentes de um sistema com arquitetura limpa é mostrado em 6.8.

Por fim, as especificações de consultas, assim como os *Gateways*, são definidas como classes abstratas e não possuem dependências definidas pelo próprio pacote. Além disso, o modelo utilizado pelas consultas não envolve valores do domínio, apenas objetos simples

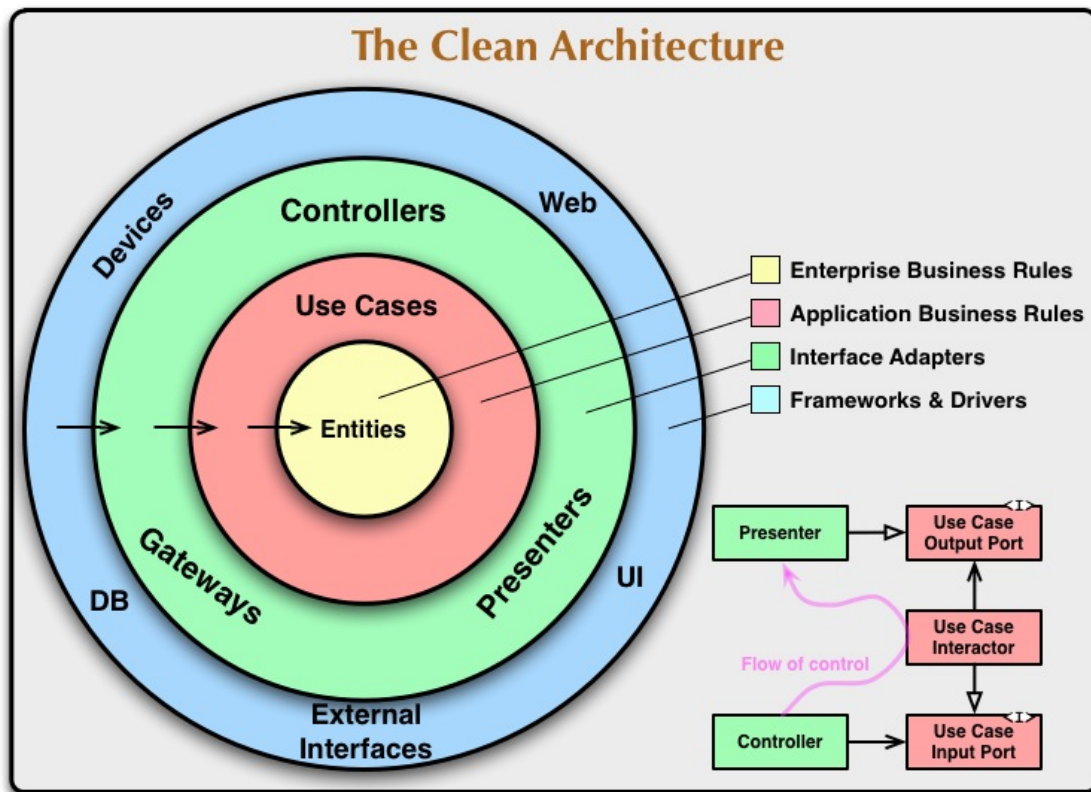


Figura 6.8: Descrição da Arquitetura Limpa em imagem, por Robert C. Martin, em *ROBERT MARTIN, 2012*

que podem ser representados como JSON, pois essa consulta deve ser utilizada pelos clientes. Os clientes devem utilizar os resultados dessas consultas para produzir interfaces gráficas, e possivelmente executar casos de uso. Essa separação entre o modelo utilizado para consultas e para os casos de uso é chamado de CQRS, descrita em *FOWLER, 2011*.

6.4 Módulo WebAPI

Para que o processamento e persistência dos dados ocorresse de maneira centralizada, foi desenvolvido um módulo *WebAPI*, responsável por receber os dados por meio de chamadas *HTTP*, realizar a validação com as utilidades fornecidas pelo pacote *core* e persisti-la em uma base de dados. Além disso, o módulo deve devolver ao cliente, também por meio de chamadas *HTTP*, informações suficientes para que o cliente possa exibir gráficos e relatórios para o usuário. Tudo isso deve ocorrer de maneira segura. O repositório contendo o código-fonte desse módulo pode ser acessado em <https://github.com/lfujiwara/dnausp-webapi>.

6.4.1 Tecnologias utilizadas

Node.js

Node.js é um ambiente de execução de código aberto, baseado no interpretador V8, desenvolvido pela *Google*, que permite a execução de código *Javascript* fora de um navegador *web*. A decisão de utilizar esse ambiente de execução facilitou a integração com o pacote *core*, escrito em *Typescript* e transpilado para *Javascript*.

NestJS

Por se tratar de uma aplicação *web* moderna, é preciso que seja utilizado um ferramental adequado para lidar com requisições HTTP, autenticação e autorização, além da integração com o domínio da aplicação, portas e adaptadores para outras camadas da aplicação ou sistemas externos, como bancos de dados e outros módulos. Para essa função, foi escolhido o arcabouço *NestJS*, de código aberto, feito para desenvolver aplicações *web* eficientes e escaláveis. O framework combina elementos de programação orientada a objetos, programação funcional e programação funcional reativa, simplificando o processo de escrita de código elegante e fácil de entender. Além disso, oferece classes e decoradores aplicados de modo a realizar diversas tarefas, dentre elas o manuseio de chamadas *HTTP*, implementação de políticas de autenticação e controle de acesso, além do uso de técnicas como injeção de dependências.

PostgreSQL

A necessidade de persistir os dados e realizar consultas complexas e bem estruturadas tornou necessária a utilização de sistema gerenciador de bancos de dados (SGBD), que realiza o gerenciamento de acesso, manipulação e organização das informações, e oferece ao cliente funções simples e mais expressivas para a interação com os dados. Para o projeto, foi escolhido o SGBD relacional *PostgreSQL*, popular e robusto e que oferece funcionalidades para o ambiente *web* como funções de agregação *JSON* e suporte a colunas de documentos também em formato *JSON*.

Prisma

Para simplificar a interação com o banco de dados, foi utilizado um mapeador objeto-relacional, software que realiza a serialização dos objetos na memória em entradas em um banco de dados relacional. Para esse papel, foi escolhido o *Prisma*, um conjunto de ferramentas para interação com bancos de dados que inclui um mapeador expressivo e minimalista.

Docker

A padronização do ambiente de execução é essencial para evitar surpresas entre o ambiente de desenvolvimento e de produção. Nesse sentido, foram utilizados containers para execução desse módulo.

Contêineres são ambientes padronizados para a execução de aplicações, que oferecem um conjunto de bibliotecas e programas definidos pelo usuário. Um contêiner é criado

a partir de uma imagem, que possui todas as informações para capacitar o acesso às ferramentas e serviços pré-definidas. O padrão de contêineres utilizados no projeto é o padrão *OCI*, definido pela *Open Container Initiative*.

Para realizar a construção das imagens e execução dos contêineres, foi escolhida a ferramenta *Docker*, que oferece ferramental para ambas as funções, de acordo com o padrão supracitado.

GitHub

O *GitHub* foi utilizado assim como o pacote *core*, no entanto, no contexto do pacote *WebAPI*, a ferramenta *GitHub Actions* foi utilizada para a construção e publicação de imagens *OCI*, parte do processo de integração e entrega contínua, como mostrado na figura 6.9.

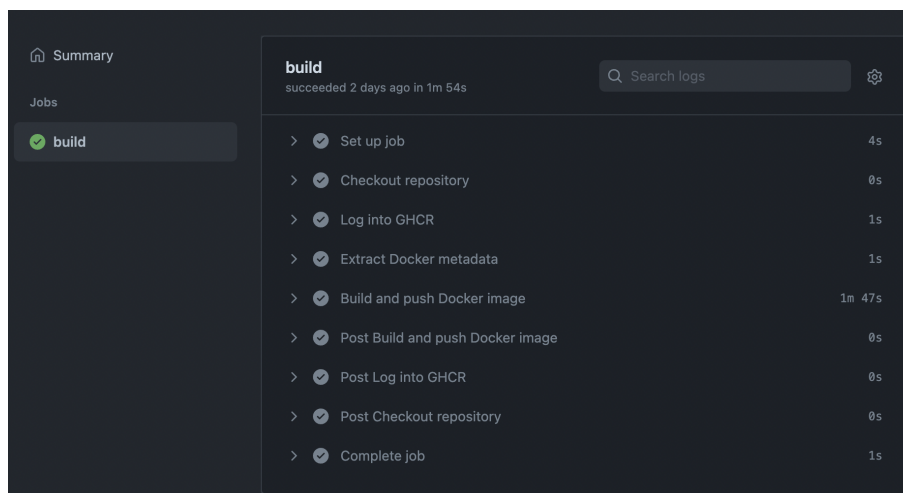


Figura 6.9: Visualização da execução da construção da imagem *OCI* do módulo *WebAPI* no painel *GitHub Actions* do *GitHub*

6.4.2 Visão geral

Dada a existência de um pacote *core* dedicado ao domínio, casos de uso e especificações de consulta da aplicação, coube a esse módulo tão somente conter versões concretas das classes abstratas definidas pelo *core* e conectá-las aos casos de uso por meio dos construtores dos casos de uso lá implementados, além de quaisquer outras adições externas ao escopo do domínio, como o tratamento de requisições *HTTP* e conexões com serviços externos (como *APIs* externas).

Fluxo de execução

A partir de uma requisição *HTTP*, o sistema passa o conteúdo da requisição por uma série de *middlewares*, que são objetos responsáveis por realizar validações e outras verificações intermediárias. Nesse contexto, essas validações são a construção de um objeto de requisição, como a representação de uma empresa *DNA USP*, e são também orquestradoras dos mecanismos de autenticação e autorização de usuários, que devem

possuir um *token* ligado a seu *email* USP e estarem na lista de usuários autorizados, definida como variável de ambiente da aplicação.

Quando as validações de responsabilidade dos *middlewares* são feitas com sucesso, a requisição é passada ao método responsável por lidar com o recurso pedido pelo usuário, este, por sua vez, realiza uma chamada ao caso de uso ou consulta, dependendo do tipo de recurso especificado.

Caso o recurso seja um caso de uso, o controlador chamará uma instância do caso de uso, que foi instanciada pelo arcabouço *NestJS* (utilizando as funcionalidade de injeção de dependência citadas). O caso de uso, por sua vez, chamará as funções do adaptador de base de dados especificado no pacote *core* que foi fornecido por meio de uma implementação do pacote *WebAPI* para realizar as operações necessárias, incluindo persistência de dados e comunicação com serviços externos.

Se o recurso for uma consulta, será chamado objeto de consulta, que, assim como o adaptador para comunicação com a base dados, é definido como uma classe abstrata no pacote *core* e fornecido por meio de uma implementação desse pacote (*WebAPI*). Por fim, o controlador recebe os resultados dessas operações e devolve ao usuário em formato *JSON*. O fluxo completo é ilustrado na figura 6.10.

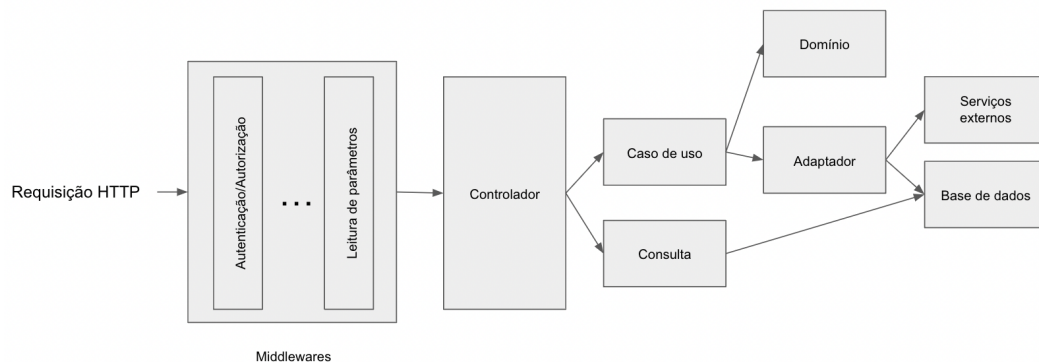


Figura 6.10: Diagrama do fluxo de execução de uma requisição HTTP

6.4.3 Disponibilização

O módulo *WebAPI* foi disponibilizado via internet por meio de uma máquina virtual hospedada pela *Amazon Web Services* (AWS), chamada de *EC2* (*Elastic Compute Cloud*). Foi realizado o provisionamento da máquina por meio do *AWS Console* e a configuração de certificados para conexão via *SSH* (*Secure Shell*).

Dentro da máquina virtual foi instalado o *Docker*, utilizado para executar a imagem do módulo *WebAPI* gerada via *GitHub Actions*. Para servir o container executando o módulo *WebAPI*, foi executada uma imagem contendo o *PostgreSQL*. Para rotear as requisições que chegam ao servidor para o cliente de maneira apropriada e gerenciar a criptografia das requisições, foi utilizado o *proxy* reverso *Traefik*.

6.5 Módulo Web

A fim de oferecer uma interface amigável para o usuário visualizar os dados em gráficos, enviar dados ao servidor e realizar outras operações definidas pelos clientes do projeto, foi desenvolvido um cliente *web*, para que tudo isso seja realizado por meio de um navegador. O código fonte do cliente implementado pode ser acessado em <https://github.com/lfujiwara/dnausp-web>.

Devido à natureza desestruturada dos dados, que estão servidos por meio de planilhas no serviço *Google Sheets*, atribuímos ao cliente *web* a responsabilidade de mapear e normalizar as linhas das planilhas em objetos *JSON* devidamente formatados e enviá-los ao servidor para que sejam persistidos.

6.5.1 Tecnologias utilizadas

NextJS

A tecnologia predominante para o desenvolvimento do sistema foi o arcabouço *NextJS*, baseado na popular biblioteca *React*. O arcabouço oferece as funcionalidades necessárias para construir páginas *web* por meio de componentes visuais reutilizáveis, cabendo aos desenvolvedores implementar as funcionalidades desejadas e então referenciá-las nos componentes visuais.

Nivo

Para a exibição dos dados foi utilizada a biblioteca *Nivo*, que oferece componentes para exibição de diversos tipos de gráficos.

GitHub

Assim como os outros módulos, o módulo *Web* também utiliza repositórios hospedados no *GitHub* e técnicas de integração e entrega contínua com *GitHub Actions*. Nesse caso, a ferramenta *GitHub Actions* está integrada com a plataforma *Vercel* para que o código seja transpilado e processado para ser usado pelo cliente final em um navegador *web*.

6.5.2 Visão geral

O código desse módulo está dividido entre biblioteca, componentes visuais, e páginas. As páginas agregam os componentes visuais e funções da biblioteca para prover uma interface de usuário elegante e organizada. Os componentes visuais, por sua vez, utilizam funções da biblioteca para refinar seu comportamento de acordo com as preferências do usuário.

A biblioteca, é a parte mais complexa desse módulo e também sua última peça. A alta complexidade se deve ao fato desse módulo precisar lidar com diversas responsabilidades, como realizar as chamadas para os serviços *Google* que fornecem os dados das planilhas previamente preenchidas pelos usuários (por meio do formulário apresentado); normalizar, validar e mapear os dados para o formato aceito pelo servidor; devolver informações sobre

registros inválidos para serem exibidas para o usuário; realizar chamadas para o servidor para enviar os registros mapeados e receber os dados para serem exibidos ao usuário e, por fim, combinar as informações recebidos para gerar visualizações diferentes sob demanda.

O procedimento de normalização e mapeamento foi um dos mais trabalhosos do projeto, pois teve de lidar com alterações do modelo das planilhas utilizadas pelos clientes, que contam com dezenas de colunas, que devem ser devidamente mapeadas e tratadas como atributos dos objetos definidos no domínio. Além disso, algumas informações não possuíam dados bem definidos, tornando necessária a utilização de técnicas como correspondência difusa de textos, para mapear dados mal formados, como nomes de institutos (exemplo: "EP - Escola Politécnica" e "Poli - Escola Politécnica") e tipos de vínculo com a universidade ("Aluno de graduação" e "Aluno ou ex-aluno de graduação").

6.5.3 Disponibilização

O cliente *web* foi disponibilizado por meio da plataforma *Vercel* integrada ao *GitHub*. A plataforma é responsável por realizar o processo de *build* (empacotamento e construção) da aplicação para o usuário final e hospedar o conteúdo por trás de uma conexão criptografada (*HTTPS*). A cada novo *commit*, a plataforma realiza um novo empacotamento e construção e atualiza a versão disponibilizada ao usuário, de acordo com as práticas de integração contínua e entrega contínua. O painel de controle com o acesso às funcionalidades oferecidas pela plataforma *Vercel* é mostrado na figura 6.11.

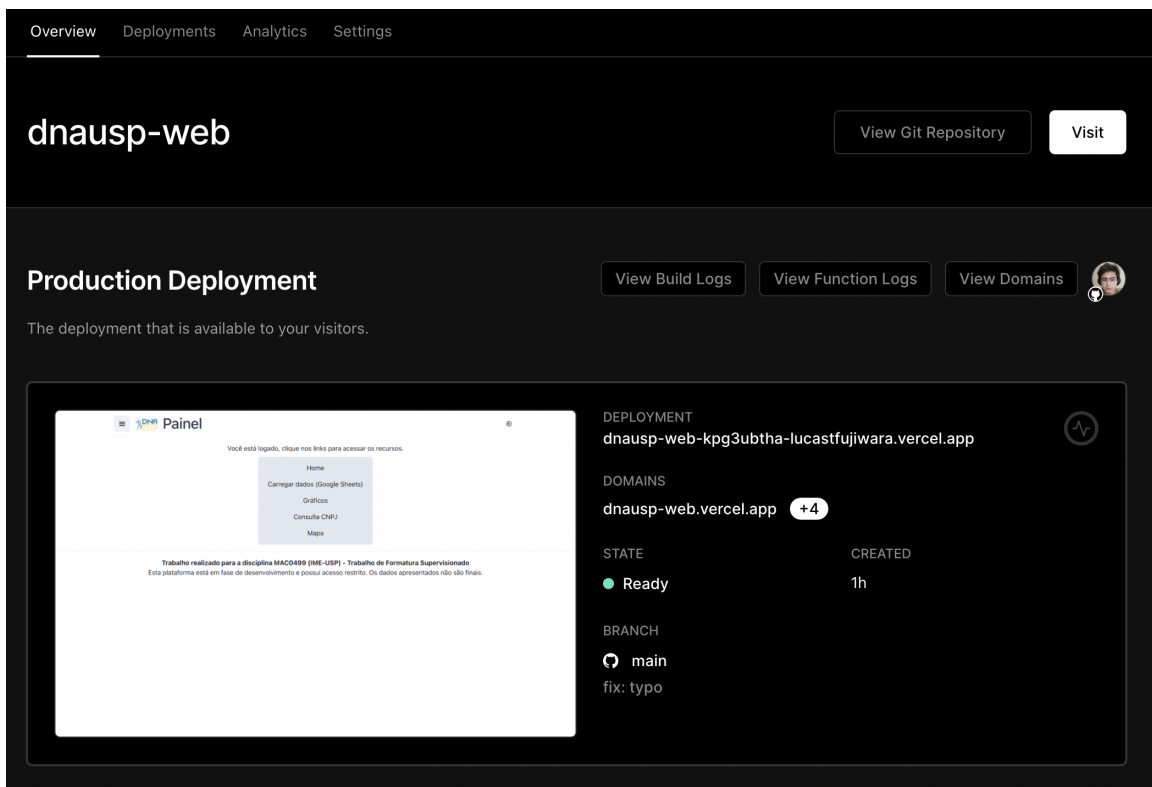


Figura 6.11: Painel de controle da plataforma *Vercel*, exibindo a última versão disponibilizada ao cliente.

6.5.4 Telas

Tela inicial

Nessa tela, representada na figura 6.12, o usuário é recebido com uma lista de funcionalidades do sistema em um menu vertical.

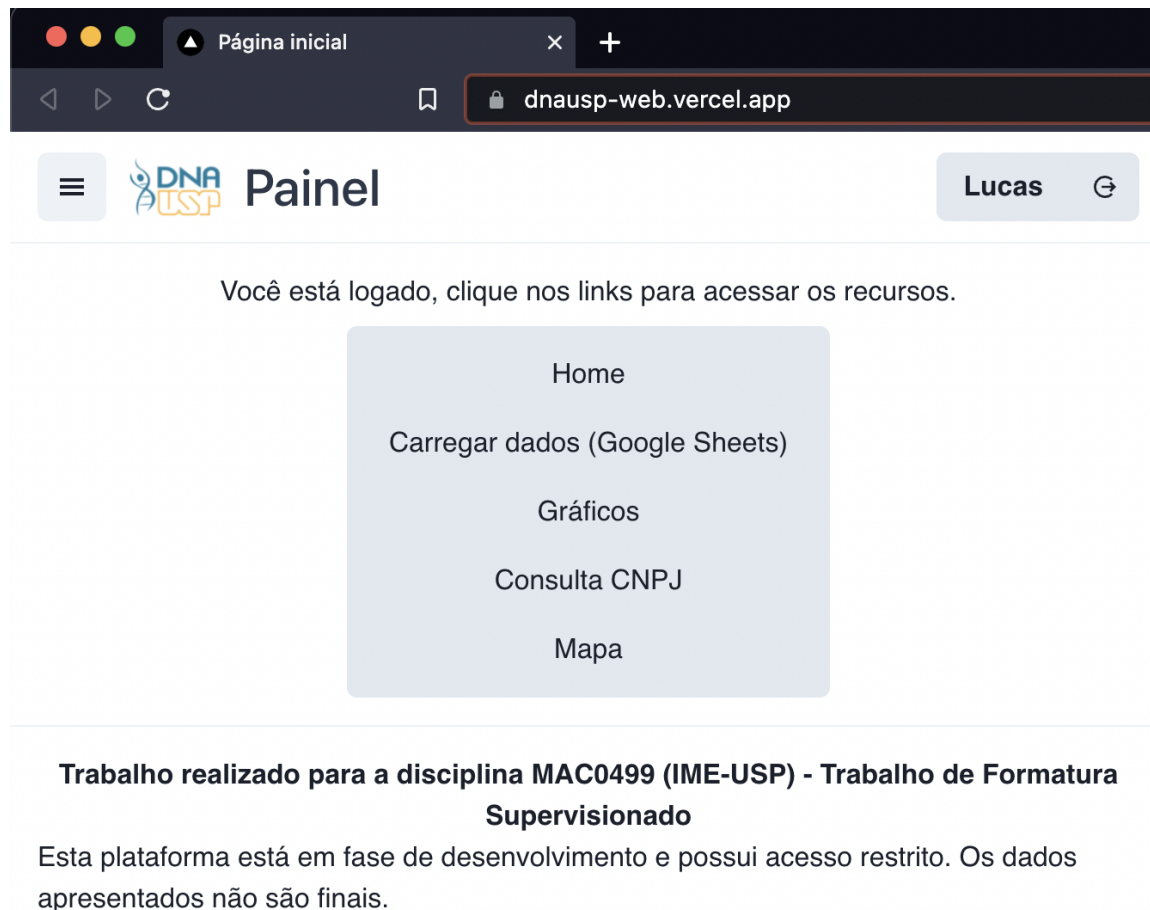


Figura 6.12: Tela inicial com menu de navegação.

Tela de mapeamento

A tela ou página de mapeamento, ilustrada na figura 6.13, serve para que o usuário possa incorporar os dados da planilha desejada na base dados, para que seja possível visualizar esses dados em gráficos posteriormente.

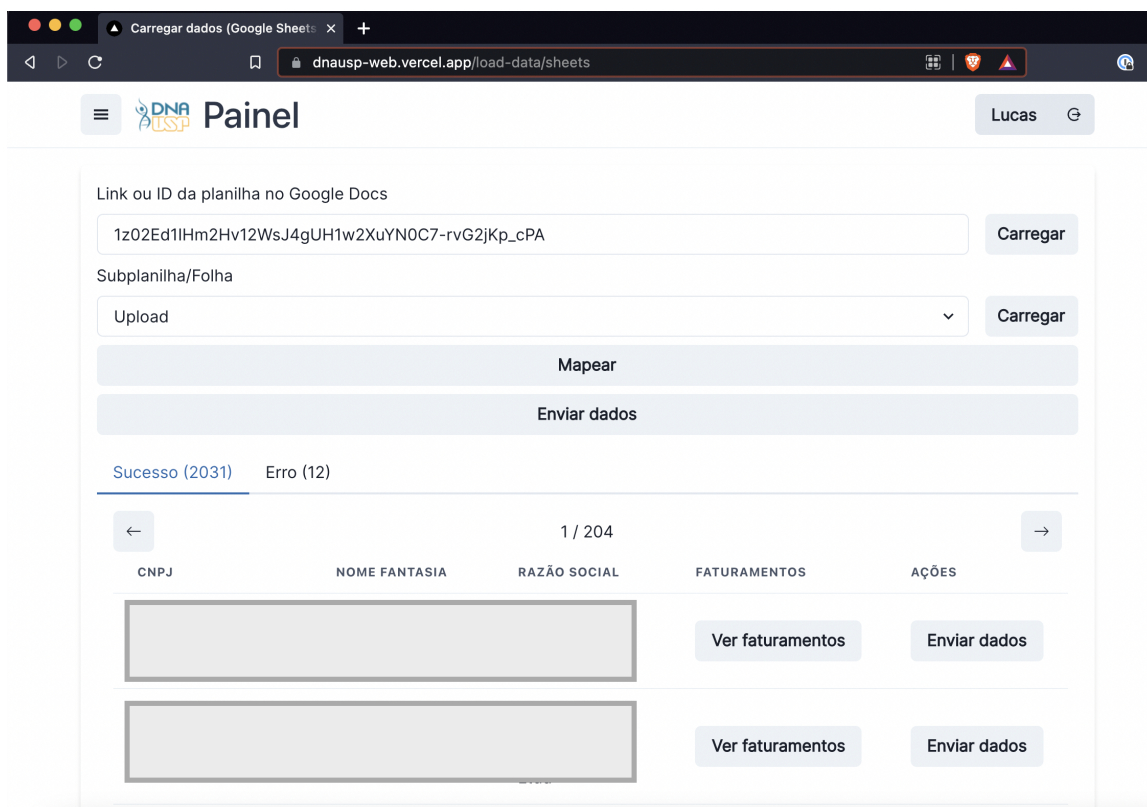


Figura 6.13: Tela para mapeamento dos dados das planilhas e envio para o servidor.

Além disso, a página oferece uma aba com os erros de validação das empresas que não puderam ser mapeadas, incluindo o índice da empresa na planilha, para auxiliar o usuário na correção dos erros, conforme exibido na figura 6.14.

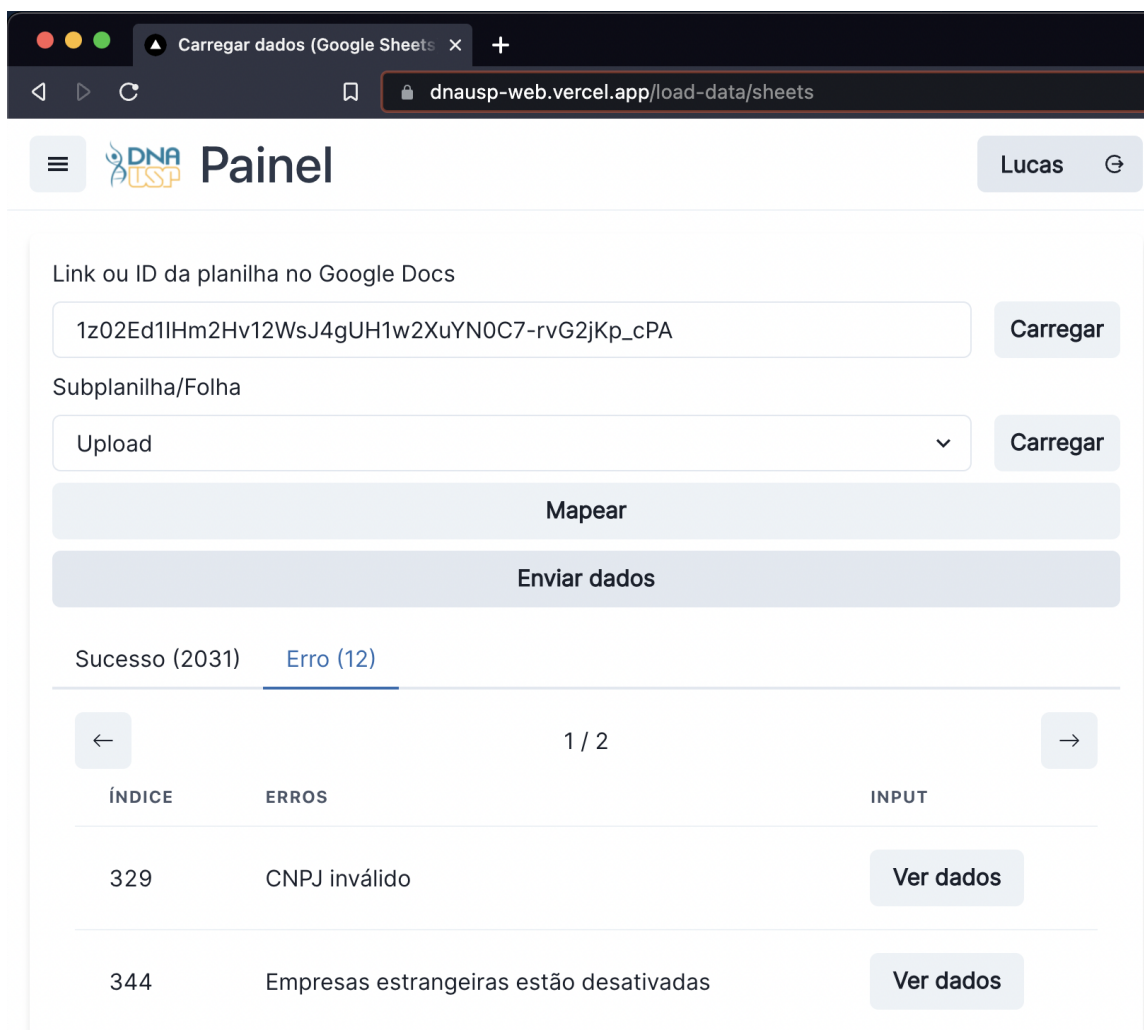
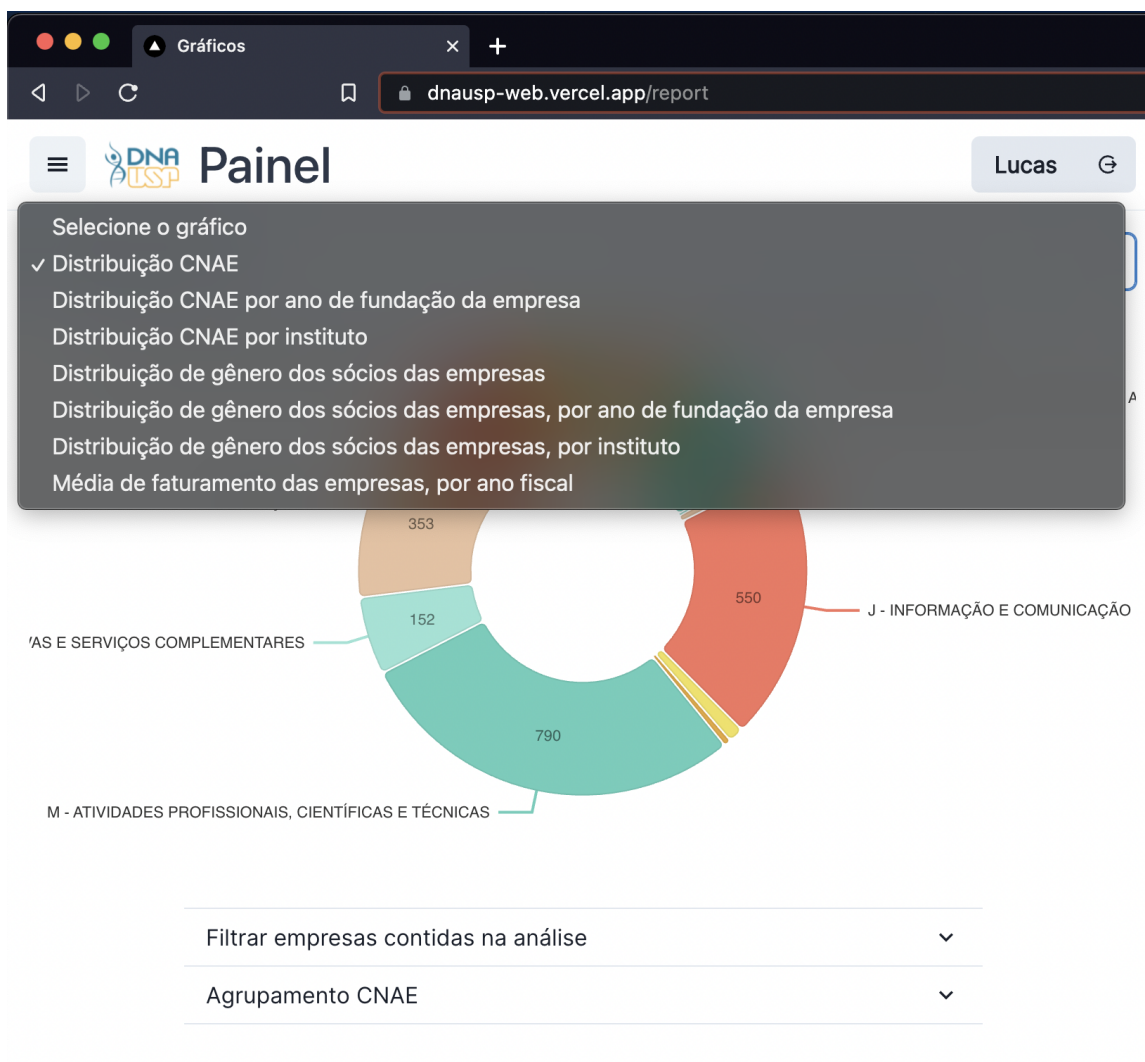


Figura 6.14: Tela para mapeamento dos dados das planilhas e envio para o servidor, aba de listagem erros de mapeamento para direcionar o usuário para tratá-los.

Tela de exibição de gráficos

A tela de exibição de gráficos, na figura 6.15 possui uma lista de todos os gráficos disponíveis para visualização, basta o usuário clicar na caixa e selecionar o gráfico desejado para visualizá-lo.



Trabalho realizado para a disciplina MAC0499 (IME-USP) - Trabalho de Formatura Supervisionado

Figura 6.15: Tela de exibição de dados, listagem de gráficos.

São oferecidos diversos controles para visualizar os dados de acordo com a necessidade do usuário. Um exemplo é o ano de referência e opção para visualizar os dados em sua forma cumulativa, isto é, não apenas para o ano de referência, mas em conjunto com todos os anos anteriores, como mostra a figura 6.16.

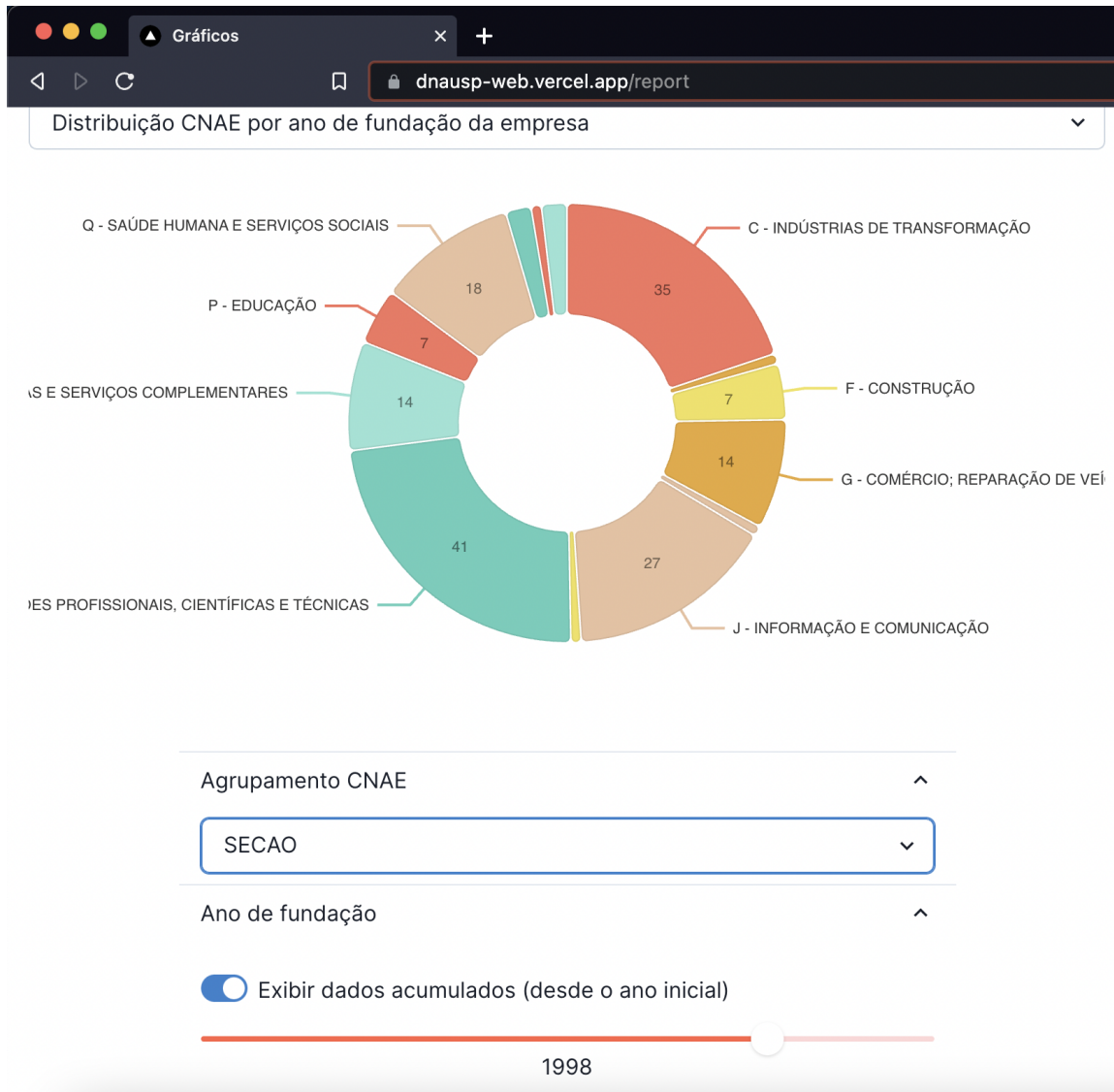


Figura 6.16: Tela de exibição de dados, gráfico de distribuição de atividades econômicas das empresas DNA USP.

Através de um seletor se altera o nível de agrupamento da visualização de atividades econômicas das empresas (CNAE), de acordo com a definição do Concla, que descreve o agrupamento em níveis como seção, divisão, grupo, classe e subclasse, todos disponíveis através desse seletor, como mostrado na figura 6.17.

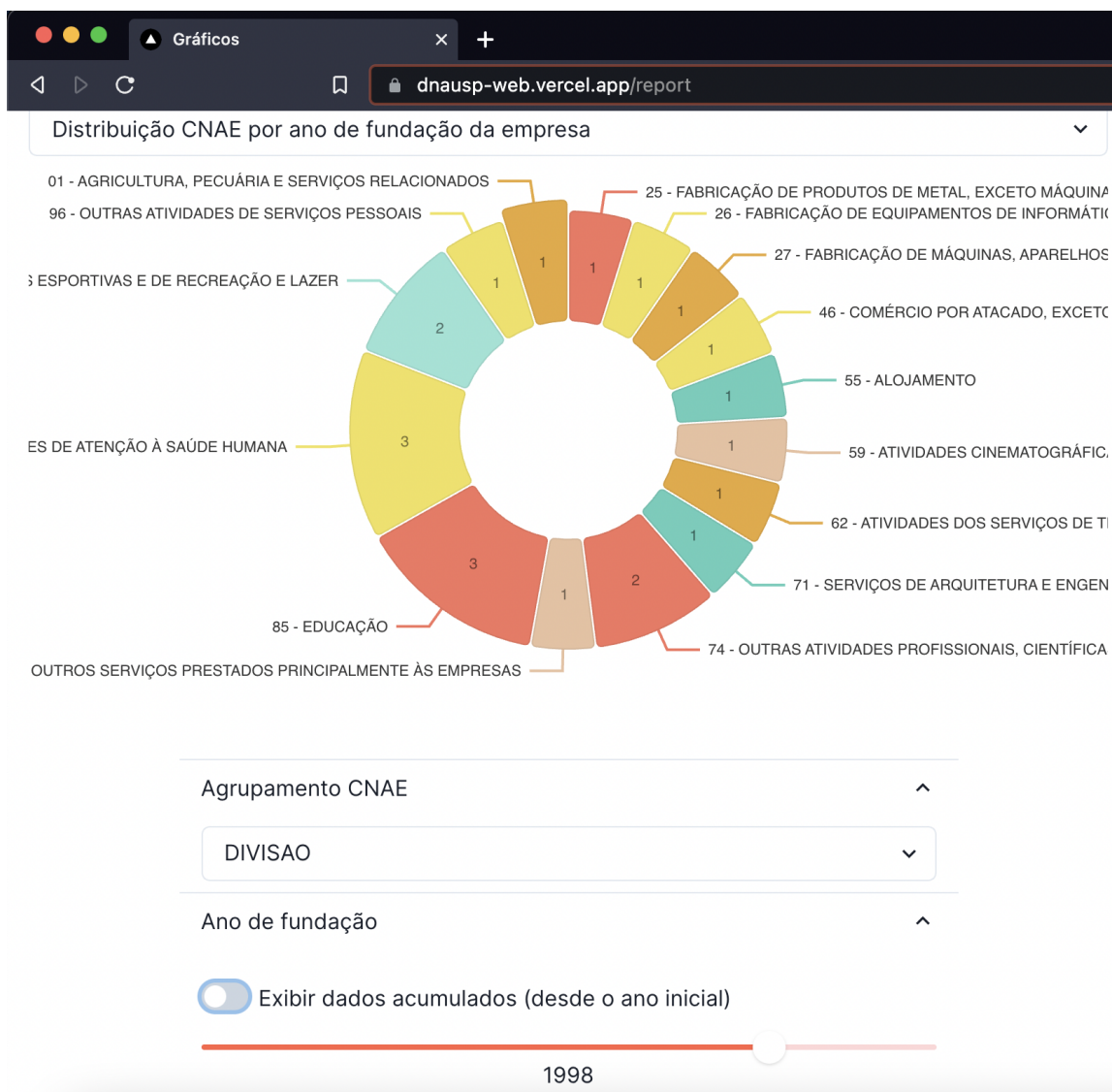
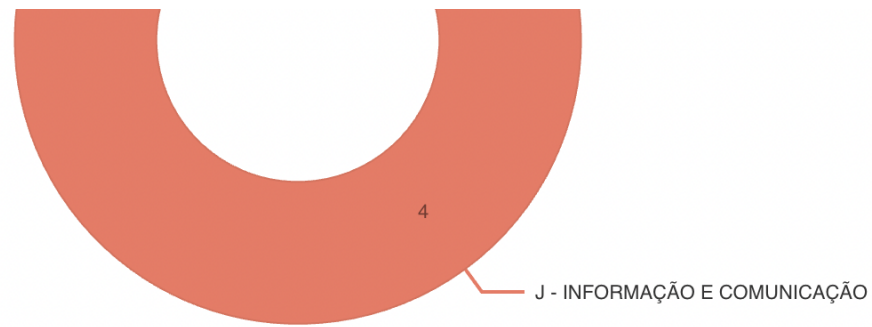


Figura 6.17: Tela de exibição de dados, gráfico de distribuição de atividades principais das empresas com maior granularidade.

Em alguns gráficos está disponível a opção de filtrar de maneira bastante granular as empresas incluídas na análise, desde o instituto vinculado à empresa por meio de seus sócios até a origem dos investimentos na empresa. O painel de seleção de filtros é exibido na figura 6.18.



Filtrar empresas contidas na análise ^

Instituto de Matemática e Estatística - IME x | v

Aluno/Ex-aluno de graduação x | v

Aluno/Ex-aluno de pós-graduação x | v

PIPE-FAPESP x | v

Atividade principal (CNAE) | v

Incubadora | v

2010 | Ano de fundação (Máximo)

Aplicar

Agrupamento CNAE v

Figura 6.18: Tela de exibição de dados, seção de filtros.

Tela de consulta de CNPJ

A pedido dos clientes, foi implementada uma tela, mostrada na figura 6.19, para realização de consulta às informações de um dado CNPJ, para auxiliá-los a preencher ou corrigir erros na base dados que precisem de intervenção manual.



Figura 6.19: Página para consulta de dados de um dado CNPJ, implementada por demanda do cliente.

Por meio das interfaces interativas e com filtros de alta granularidade, onde aplicável, demonstradas ao longo deste capítulo, o sistema desenvolvido atende aos requisitos definidos, sendo a última peça a ser entregue no escopo deste trabalho. A biblioteca comum aos módulos de implementação (*Web* e *WebAPI*) pode ser facilmente reutilizada em outros contextos e possíveis extensões desse projeto, ou mesmo da própria biblioteca, pois conta com testes automatizados para garantir que novas extensões não afetam o funcionamento do conteúdo existente. Quanto aos módulos de implementação, ambos contam com recursos para integração e entrega contínua, de maneira que podem ser executados em diferentes

contextos - possivelmente por outra universidade - com facilidade.

Capítulo 7

Conclusões

7.1 Análises realizadas

O objetivo deste trabalho foi analisar e refletir a importância da prática empreendedora a partir de iniciativas incentivadas e viabilizadas pela Universidade de São Paulo — que podem ser estendidas para as demais instituições de ensino do país — e demonstrar que a USP, como polo de pesquisa e inovação financiado por recursos públicos cumpre um papel de extrema importância ao oferecer retornos de tais investimentos para a sociedade em sua totalidade, extrapolando os limites de seus campi e de seu Estado de origem.

A disponibilização dos dados coletados pela equipe do Hub USP Inovação foi fundamental para que as primeiras investigações e questionamentos fossem feitos, utilizando para isso conjuntos de dados auxiliares, como a relação de inventores e patentes associados à USP. A partir das perguntas respondidas usando estratégias de análise exploratória de dados, foi possível traçar um perfil do empreendedorismo na Universidade e traçar paralelos com a realidade brasileira atual. Com as análises iniciais, constatou-se que empreendimentos focados em atividades científicas, em informação e comunicação e em educação são os mais frequentes atualmente, e que recentemente tivemos aumentos na proporção de empresas associadas a comércio e serviços e também à saúde.

Se observa que as unidades de ensino que abrangem grandes áreas mais voltadas à tecnologia são aquelas que mais participam de iniciativas empreendedoras na Universidade. Essa taxa tende a demonstrar variações quando ao analisar em diferentes áreas de atuação, embora seja perceptível uma concentração de empresas com DNA USP em um conjunto ainda pequeno de unidades. Esse fato reforça a necessidade de difundir cada vez mais cursos voltados ao empreendedorismo entre as diversas áreas do conhecimento, mesclando ideias e conceitos de forma transversal e assim dando margem para a produção de inovação e conhecimento.

Sobre a participação feminina, constata-se que as circunstâncias que se fazem presentes no mercado de trabalho brasileiro e mundial refletem no ecossistema empreendedor da USP. Por exemplo, nos dados cadastrados apenas 26.8% das pessoas fundadoras são do sexo feminino, revelando uma disparidade que tem sido combatida e que ainda precisa ser constantemente trabalhada e refletida para se poder atingir realidades mais igualitárias.

Descobrimos também que essa desigualdade se dá no conjunto de estudantes matriculados nas unidades da USP, sendo que em alguns casos a diferença de pessoas do sexo masculino chega a ser quase três vezes maior do que aquelas do sexo feminino.

Nota-se que a USP, como referência nacional de ensino, pesquisa e inovação, produz impactos positivos a partir da prática empreendedora por todo o território nacional. A análise de dados geográficos mostrou que a instituição, embora esteja localizada em São Paulo, apresenta vínculos com empresas fundadas por todo o país, evidenciando a importância e a contribuição da USP para o desenvolvimento econômico regional.

Por fim, percebe-se forte correlação entre as empresas DNA USP e pessoas que cursam ou cursaram pós-graduação. Ao estudar as empresas chamadas *spin-offs*, nota-se que a maior parte das empresas DNA USP possuem vínculos com a pós-graduação, evidenciando a importância das atividades de pesquisa para o mercado tecnológico brasileiro e para o setor econômico do país. Em tempos em que a ciência sofre golpes frequentes por membros do poder público, essa análise enfatiza a importância do ensino e da pesquisa para o desenvolvimento econômico e social do país e torna evidente a conexão entre o que se produz em universidades e o que é oferecido à sociedade.

7.2 *Software desenvolvido*

O conhecimento obtido a partir do convívio com os dados durante o processo de análise exploratória foi o ponto de inflexão no desenvolvimento do *software* a ser entregue ao cliente. Com o cliente satisfeito com as análises realizadas, foi determinado um objetivo concreto: tornar as análises interativas e escaláveis, por meio de um *software* capaz de produzir os relatórios desejados a partir do conjunto de dados já utilizado pelo cliente (*Google Sheets*).

Além do objetivo concreto definido, também foram mantidas as reuniões frequentes com os clientes para que o produto pudesse ser testado e ajustado aos moldes desejados, além da implementação de partes periféricas ao projeto com a finalidade de atender demandas pontuais. Os clientes ficaram satisfeitos com as capacidades oferecidas pelo sistema e deram o seguinte depoimento:

A plataforma desenvolvida facilita e otimiza a análise do grande volume de dados das empresas com DNA USP cadastradas. Além disso, permite a análise de dados que não poderiam ser avaliados manualmente ou utilizando-se recursos básicos. A partir da plataforma, é possível obtermos informações relevantes sobre as empresas DNA USP, a composição dessas e os impactos gerados, as quais são facilmente filtradas e visualizadas por meio dos gráficos dinâmicos. As funcionalidades de validação de dados e consulta de CNPJ ainda permitem que o banco de dados de empresas seja constantemente ajustado e padronizado, mantendo apenas informações válidas e coerentes. Portanto, a plataforma atende as nossas necessidades e apresenta grande valor para o nosso trabalho diário de manutenção e análise de dados.

Em questões técnicas, as decisões tomadas acerca das tecnologias utilizadas e disposição do código entre múltiplos repositórios mostraram-se fundamentais, em especial a decisão de

utilizar um pacote comum entre cliente e servidor, que poupou o esforço de escrever códigos análogos duas vezes, auxiliou na homogeneidade do fluxo de execução e tornou a interação cliente-servidor bastante previsível. O processo de pesquisa e estudo das alternativas possíveis, bem como de gerenciamento de projetos, ambiente *web* e seu ferramental foi essencial para que houvesse bagagem de conhecimento para tomar tais decisões.

Por fim, foi possível entregar um sistema com uma biblioteca comum bem testada e disponibilizada como um pacote público, que pode ser reutilizada em extensões desse projeto, além de módulos de cliente e servidor empacotados de acordo com as técnicas mais modernas em desenvolvimento *web* e disponibilizados em infraestrutura adequada para que os clientes possam acessá-los a um clique em seu navegador *web*.

7.3 Próximos passos

Ao longo do trabalho, foi fácil perceber como os dados sobre as empresas DNA USP podem ajudar a responder questões pertinentes sobre a relação entre a Universidade e a sociedade. Para isso ser feito com mais agilidade e eficácia, um primeiro passo pode ser o aprimoramento das técnicas de coleta e limpeza dos dados acerca do empreendedorismo na USP, a partir do desenvolvimento de formulários com validações já implementadas que facilitem o desenvolvimento de análises desambiguando os dados inseridos. Para essa finalidade pode ser utilizada a própria biblioteca *core* desenvolvida nesse trabalho. Pode-se, por exemplo, diversificar os mecanismos de entrada dos dados, desenvolvendo sistemas móveis e serviços que busquem atualizar as informações de forma automatizada.

Além disso, podem ser propostas outras perguntas para serem respondidas com os dados coletados, mesclando para isso os resultados das consultas com outros registros e produzindo análises interessantes. Por exemplo, pode-se buscar comparar o desempenho do empreendedorismo da USP com o de outras universidades do país e do mundo, assim como comparar os indicadores econômicos dos locais em que essas instituições estão inseridas.

Capítulo 8

Apreciação Pessoal

8.1 Lucas

O processo de desenvolvimento desse trabalho foi uma das tarefas mais difíceis que já realizei, não só pela conciliação com semestres bastante carregados e a vida profissional, mas também pelo porte avantajado e responsabilidade de gerir um projeto por conta própria, ainda que com boa companhia para isso.

Em linha crítica, poderia ter sido mais assertivo no trabalho, definido metas e objetivos antes e realizado o desenvolvimento de forma mais uniforme em questão de alocação de tempo. Também poderia ter sido mais proativo no início do projeto, teria feito toda a diferença. Sinto que exagerei na carga de tarefas durante esse semestre, englobando diversas matérias, vida profissional e esse trabalho.

Por outro lado, me sinto satisfeito com o resultado do trabalho, pude ver o impacto da universidade na formação de empreendimentos que contribuem direta e indiretamente de maneira positiva na vida de tantas pessoas no Brasil e no mundo, na produção científica e inovação e descrevê-lo com gráficos e relatórios interativos, utilizando tudo aquilo que a própria universidade me ensinou. De certo, esse é o *loop* de *feedback* que gosto de presenciar.

Foi uma verdadeira viagem em questão de conceitos, tecnologias e técnicas utilizadas. Trabalhamos com análise e visualização de dados, conceitos de programação funcional e orientada a objetos, padrões de código e projeto, infraestrutura em nuvem, segurança, integração e entrega contínua e muito mais. Também aprendemos a trabalhar com clientes, recebendo *feedback* de maneira contínua e guiando o trabalho a fim de oferecer mais valor.

Como o final de uma jornada onde se retoma todos eventos relevantes da mesma, esse trabalho final foi uma oportunidade de revisitar todos os conceitos abordados ao longo da graduação, pois não há como citar uma disciplina que não tenha contribuído de maneira relevante para a execução desse projeto, o caminho de MAC0101 a MAC0499 fez todo o sentido em sua reta final.

Deixo meus agradecimentos ao Alfredo Goldman e à Geciane Porto, que supervisião-

naram nosso trabalho e nos ajudaram a dar direção ao projeto, a toda equipe Hub USP Inovação, André Luís Balico da Silva, Lara Mendes Ferreira Guimarães que colaboraram com o projeto com *feedback* constante e instruções sobre os dados coletados e toda a plataforma do Hub USP Inovação, ao João Daniel, que nos deu apoio técnico ao longo do processo e ao Daniel, que colaborou muito, principalmente com a análise de dados, para que conseguíssemos entregar esse projeto, aprendi muito com as abordagens que ele utilizou para produzir as análises, inferir informações e unir conjuntos de dados para enriquecer nosso *dataset*.

Agradeço todos os professores com quem tive o privilégio de aprender, seja de maneira direta ou indireta, desde o ensino básico até aqui, isso certamente fez toda diferença, cada um foi responsável não só por me ajudar a inserir um pedaço de conhecimento a mais em minha caixinha de ferramentas que costumo chamar de "mente", mas também por contribuir para que eu tivesse gosto por aprender matemática, computação e todo o mais que existe por aí para se saber.

Agradeço todos os colegas de curso, que me ajudaram por várias vezes ao longo da graduação, seja por uma explicação de alguns minutos no CCSL ou através de suas atividades nos grupos de extensão que serviram de pontapé inicial para que eu me aprofundasse em muita coisa.

Por fim, agradeço à Elisa, com quem compartilhei várias entregas de trabalhos, desesperos e esperanças nesses últimos anos, e à minha família, que me apoiou desde o início de (literalmente) tudo.

8.2 Daniel

Desde 2017, quando ingressei no Bacharelado em Ciência da Computação no IME-USP, presenciei muitas afirmações imprecisas de que as universidades, sobretudo as públicas, são demasiadamente afastadas da sociedade e do que é tido como 'vida real', seja por não serem acessíveis pela maioria das pessoas ou pelo fato de que boa parte de seus alunos e alunas ingressam mais tardiamente no mercado de trabalho, diferentemente da maioria dos brasileiros. Ao realizar esse trabalho, percebi que embora o acesso à universidade precise ainda ser muito mais democrático e igualitário, as contribuições que esta oferece à sociedade são amplas e diversificadas, merecendo destaque na análise do papel das instituições de ensino superior no Brasil. As empresas com o conceito DNA USP são demonstrações de que a produção científica e a inovação almejadas diariamente por milhares de pessoas na USP extrapola as paredes de seus laboratórios e salas de aula.

Pessoalmente, acredito que eu poderia ter tido mais assiduidade na produção do trabalho em alguns momentos do ano, quando estive concentrado em outras atividades que também demandam meu tempo e dedicação. Acredito ainda dinâmica de trabalho remoto foi um fator que dificultou o entrosamento entre mim e Lucas, que acabei conhecendo melhor somente após o início do trabalho e mesmo assim nos reunimos esporadicamente somente para discutir assuntos relacionados ao TCC e as demandas do HUB USP Inovação. De qualquer forma, agradeço ao Lucas pelo trabalho que realizamos juntos e por todo o esforço e empenho que ele demonstrou ao longo do ano. Aprendi muito com ele, ainda que de forma distante.

Também agradeço enormemente o tempo, a paciência e a colaboração dos membros do Hub USP Inovação, André Luís Balico da Silva, Lara Mendes Ferreira Guimarães, a professora, co-supervisora do trabalho e também membra do Hub Geciane Silveira Porto, ao nosso orientador, Alfredo Goldman e aos membros do USP Code Lab. As críticas e sugestões feitas certamente nos direcionaram para caminhos positivos. Agradeço também a todas as professoras, professores e colegas da Universidade de São Paulo que participaram dessa jornada para chegarmos até aqui, do começo ao fim e além!

Fico feliz por conseguir vislumbrar a importância de que o estudo que realizamos tem na atualidade, em que a ciência sofre frequentes ataques de autoridades que deveriam incentivá-la conforme estabelece o artigo 218 da Constituição Federal: *O Estado promoverá e incentivará o desenvolvimento científico, a pesquisa, a capacitação científica e tecnológica e a inovação.* Espero que o trabalho que produzimos possa servir como um forte argumento para fundamentar a importância do ensino, pesquisa e extensão nas universidades.

Referências

- [SCHUMPETER 1942] Joseph SCHUMPETER. “Capitalismo, socialismo e democracia”. Em: 1942 (citado na pg. 5).
- [GANG OF FOUR 1994] Ralph Johnson John Vlissides ERICH GAMMA Richard Helm. “Design patterns: elements of reusable object-oriented software”. Em: 1994 (citado nas pgs. 11, 48).
- [ROBERT MARTIN 1994] Robert C. MARTIN. “Oo design quality metrics: an analysis of dependencies”. Em: 1994 (citado na pg. 50).
- [ALMEIDA E CRUZ 2010] A. D. A ALMEIDA D. R.; CRUZ. “O brasil e a segunda revolução acadêmica. interfaces da educação”. Em: 2010 (citado na pg. 7).
- [FOWLER 2011] Martin FOWLER. *CQRS*. 2011. URL: <https://martinfowler.com/bliki/CQRS.html> (citado na pg. 51).
- [ROBERT MARTIN 2012] Robert C. MARTIN. *The Clean Architecture*. 2012. URL: <https://blog.cleancoder.com/uncle-bob/2012/08/13/the-clean-architecture.html> (citado na pg. 51).
- [FOWLER 2013] Martin FOWLER. *DDD Aggregate*. 2013. URL: https://martinfowler.com/bliki/DDD_Aggregate.html (citado na pg. 48).
- [SEBRAE E ENDEAVOR BRASIL 2016] SEBRAE E ENDEAVOR BRASIL. *Empreendedorismo nas Universidades Brasileiras*. 2016. URL: <https://www.sebrae.com.br/Sebrae/Portal%20Sebrae/Anexos/Relatorio%20Endeavor%20impresao.pdf> (acesso em 21/11/2019) (citado nas pgs. 7, 8).
- [SÃO PAULO FAPESP 2016] Fundação de Amparo à Pesquisa do Estado de SÃO PAULO FAPESP. *RCGI – Research Centre for Greenhouse Gas Innovation*. 2016. URL: https://fapesp.br/cpe/rcgi_%E2%80%93_research_centre_for_greenhouse_gas_innovation/22 (citado na pg. 42).
- [ROBERT MARTIN 2017] Robert C. MARTIN. “Clean architecture”. Em: 2017 (citado nas pgs. 11, 48).

- [INSTITUTO BRASILEIRO DE QUALIDADE E PRODUTIVIDADE 2019] INSTITUTO BRASILEIRO DE QUALIDADE E PRODUTIVIDADE. *Empreendedorismo no Brasil - 2019. Relatório Executivo (GEM)*. Dez. de 2019. URL: <https://ibqp.org.br/wp-content/uploads/2021/02/Empreendedorismo-no-Brasil-GEM-2019.pdf> (acesso em 20/12/2021) (citado nas pgs. 6, 7, 29).
- [SÃO PAULO 2019] Universidade de SÃO PAULO. *RESOLUÇÃO Nº 7661, DE 22 DE MAIO DE 2019*. 2019. URL: <http://www.leginf.usp.br/?resolucao=resolucao-no-7661-de-22-de-maio-de-2019> (citado na pg. 42).
- [ANUÁRIO DE ESTATÍSTICA DA USP 2020] ANUÁRIO DE ESTATÍSTICA DA USP. *Anuário de Estatística da USP*. 2020. URL: (<https://uspdigital.usp.br/anuario/AnuarioControle> (citado nas pgs. 29, 32).
- [INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL 2020] INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL. *Ranking nacional de depositantes 2020*. 2020. URL: <https://www.gov.br/inpi/pt-br/central-de-conteudo/noticias/inpi-divulga-rankings-dos-maiores-depositantes-em-2020> (citado na pg. 40).
- [SECRETARIA DA FAZENDA E PLANEJAMENTO 2021] Governo do Estado de São Paulo Secretaria da Fazenda e Planejamento Coordenadoria da ADMINISTRAÇÃO TRIBUTÁRIA. *Relatório da Receita Tributária do Estado de São Paulo*. Out. de 2021. URL: https://portal.fazenda.sp.gov.br/acessoinformacao/Downloads/Relatorios-da-Receita-Tributaria/2021/janeiro/INTERNET_janeiro21.pdf (acesso em 25/10/2021) (citado na pg. 1).
- [CONCLA (COMITÊ NACIONAL DE CLASSIFICAÇÃO 2021)] CONCLA (COMITÊ NACIONAL DE CLASSIFICAÇÃO. *Busca por CNAE*. 2021. URL: <https://concla.ibge.gov.br/busca-online-cnae.html> (citado na pg. 21).
- [MINISTÉRIO DA ECONOMIA 2021] Ministério da ECONOMIA. *Painel Tempos de Abertura de Empresas*. Out. de 2021. URL: <https://www.gov.br/governodigital/pt-br/mapa-de-empresas/painel-mapa-de-empresas> (acesso em 22/10/2021) (citado na pg. 1).
- [FEMALE FOUNDERS REPORT 2021] FEMALE FOUNDERS REPORT. *Female Founders Report*. 2021. URL: <https://materiais.districto.me/dataminer-female-founders-report> (citado na pg. 29).
- [REVISTA PIAUÍ 2021] Renata Buono LIANNE CEARÁ Marcos Amorozo. *Diploma, Acesso e Retrocesso*. 2021. URL: <https://piaui.folha.uol.com.br/diploma-acesso-e-retrocesso/> (acesso em 21/11/2021) (citado na pg. 7).