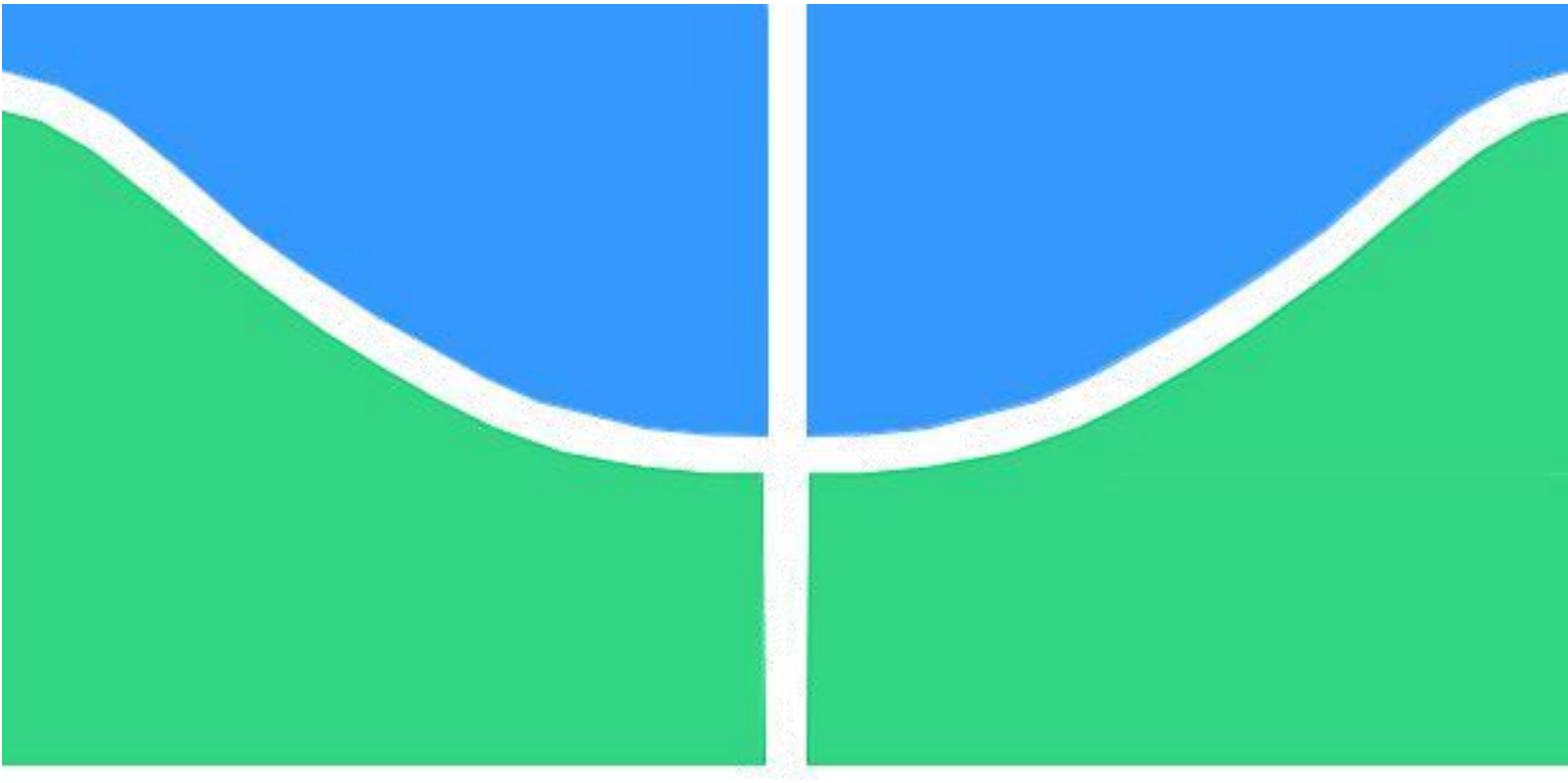


PROJETO DE GRADUAÇÃO

ANÁLISE DA EFICIÊNCIA DE MICRO E PEQUENAS EMPRESAS BRASILEIRAS NO SETOR INDUSTRIAL UTILIZANDO ANÁLISE ENVOLTÓRIA DE DADOS

Por,

Felipe Rosa Machado

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

UNIVERSIDADE DE BRASÍLIA

Faculdade de Tecnologia

Departamento de Engenharia de Produção

PROJETO DE GRADUAÇÃO

ANÁLISE DA EFICIÊNCIA DE MICRO E PEQUENAS EMPRESAS BRASILEIRAS NO SETOR INDUSTRIAL UTILIZANDO ANÁLISE ENVOLTÓRIA DE DADOS

Por,

Felipe Rosa Machado

Relatório submetido como requisito parcial para obtenção do grau em Engenharia de Produção

Banca Examinadora

Prof. Ph.D. Reinaldo Crispiniano Garcia, UnB/EPR (Orientador) _____

Prof. Ph.D. João Mello da Silva (Convidado) _____

Brasília, setembro de 2022

AGRADECIMENTOS

Agradeço primeiramente à minha família, em especial meus pais Ana e Érico e meu irmão Thiago, pessoas a quem devo tudo que tenho. Tudo que conquistei foi fruto das oportunidades que me foram proporcionadas por eles, seguido por apoio incondicional, conselhos, amizade e muito amor. Sou muito grato pela minha família e tenho a certeza de que, com eles, posso enfrentar qualquer coisa que a vida proporcionar.

Agradeço também a todos meus amigos e colegas, desde os amigos de infância quanto os conhecidos na universidade. Foram anos de muitos momentos inesquecíveis que tive o prazer de compartilhar com pessoas incríveis, em especial meus companheiros de curso Gabriel e Guilherme. Passamos pelos momentos mais difíceis e cansativos juntos, e os bons momentos se tornam ainda melhores quando compartilhados com as pessoas certas. Crescemos juntos e me tornei a pessoa que sou por conta deles.

Por último, agradeço aos meus professores de graduação pelos ensinamentos, em especial ao meu orientador, professor Reinaldo Garcia. Sinto que tive muita sorte de ter um orientador que é exemplo de professor e profissional, sempre muito dedicado e fazendo o seu melhor pelos seus alunos. Suas aulas e seus conselhos são incentivos para buscar sempre mais na vida profissional.

RESUMO

As medidas de eficiência e produtividade têm apresentado queda no ritmo de alta em muitos países no mundo, desenvolvidos ou em desenvolvimento. Esta tendência se torna preocupante porque o aumento contínuo dessas medidas leva ao crescimento e o desenvolvimento das nações no longo prazo. No Brasil, essas medidas têm desempenhado pior do que a média global, inclusive comparando com outros países com níveis de desenvolvimento parecido. Dessa forma, conseguir identificar pontos de ineficiência e propor soluções de melhoria para estas medidas se torna uma tarefa fundamental para retomar o ritmo de crescimento e produtividade. Um dos setores da economia brasileira que, apesar de ser de extrema importante econômica, apresenta níveis de produtividade e competitividade abaixo de outras economias mundiais, e até outros países latino-americanos, é a indústria.

Este trabalho apresenta uma proposta de quantificar a eficiência de um grupo de indústrias, utilizando um modelo matemático de análise de dados para investigar a ineficiência do setor no Brasil. Portanto, foi aplicado o modelo de Análise Envoltória de Dados (DEA) a dados públicos de micro e pequenas empresas da indústria de transformação, disponíveis no portal da Receita Federal. A análise dos resultados confirmou uma eficiência média baixa entre as empresas estudadas, sendo identificados grupos de empresas que podem ser analisadas para iniciar um processo de melhoria na eficiência do setor.

Palavras-chave: Eficiência; Indústria brasileira; Programação Linear; Análise Envoltória de Dados

ABSTRACT

Efficiency and productivity measures have shown a fall in the rate of increase in many countries in the world, developed or developing. This trend becomes worrisome because the continued increase in these measures leads to the growth and development of nations in the long run. In Brazil, these measures have performed worse than the global average, even compared to other countries with similar levels of development. Thus, being able to identify points of inefficiency and propose improvement solutions for these measures becomes a fundamental task to resume the pace of growth and productivity. One of the sectors of the Brazilian economy that, despite being extremely important economically, has levels of productivity and competitiveness below other world economies, and even other Latin American countries, is the industrial sector.

This paper presents a proposal to quantify the efficiency of a group of industries, using a mathematical model of data analysis to investigate the inefficiency of the sector in Brazil. Therefore, the Data Envelopment Analysis (DEA) model was applied to public data of micro and small companies in the manufacturing industry, available on the IRS portal. The analysis of the results confirmed a low average efficiency among the companies studied, being identified groups of companies that can be analyzed to start a process of improvement in the efficiency of the sector.

Keywords: Efficiency; Brazilian industry; Linear Programming; Data Envelopment Analysis

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Produtividade pode ser definida como uma medida da eficiência com que os recursos produtivos de uma empresa, setor ou país são utilizados (VELOSO; BONELLI; CASTELAR, 2017). Dessa forma, desde pequena organizações e empresas, a busca por maior eficiência e, consequentemente, maior produtividade deve ser priorizada, porque dela pode depender o crescimento e o desenvolvimento da organização. Essa lógica também vale para grandes empresas e setores da economia, chegando até mesmo a ser decisiva no desenvolvimento de uma nação.

Atualmente, tem-se observado sinais de clara diminuição no ritmo de crescimento da produtividade em muitos países, desenvolvidos ou em desenvolvimento. Para o Brasil, essa desaceleração tem sido maior ainda do que a média mundial, mesmo em comparação com outros países de estrutura econômica similar (VELOSO; BONELLI; CASTELAR, 2017).

O lento crescimento de um país é capaz de gerar uma grave crise econômica, e suas origens, que podem ser fatores externos ou internos, devem ser debatidas para propor soluções para o problema (VELOSO E BONELLI, 2016). Independentemente do tamanho e da origem dos fatores que levam à dificuldade de desenvolvimento, é importante analisar pontos de desequilíbrio estrutural e apontar estratégias para superar tais desequilíbrios para voltar à perspectiva de crescimento econômico.

A redução na eficiência é um dos fatores que pode resultar a um lento crescimento. Quando setores fundamentais para a economia de uma nação podem ser vistos como ineficientes, o país corre o risco de passar por uma crise e desacelerar seu desenvolvimento. E no caso de muitos países, um dos setores que tem grande influência na economia nacional é a indústria.

O setor industrial possui um papel muito importante para qualquer país, no crescimento econômico e geração de empregos, além de que o nível de industrialização de uma nação muitas vezes tem uma relação direta com o nível de desenvolvimento do país. Segundo dados da CNI (Confederação Nacional da Indústria), no Brasil, a indústria é responsável por 22,2% do PIB nacional, por 71,8% das exportações de bens e serviços e por 20,9% dos empregos formais no Brasil.

Estima-se que a cada R$ 1,00 produzido na indústria brasileira, R$ 2,43 são gerados na economia nacional, um retorno maior do que nos setores do Agropecuário e do Comércio e

Serviços, com R$ 1,75 e R$ 1,49 respectivamente. A indústria da transformação sozinha representa 11,3% do PIB nacional e 46,2% das exportações brasileiras, com um retorno por real produzido ainda maior, de R$ 2,67 (PORTAL DA INDÚSTRIA, 2022).

Em 2019, ano em que o Brasil era a 9ª maior economia do mundo, a indústria brasileira era apenas a 13 ª maior participação na produção industrial, com 1,48% do mercado global. Importante mencionar que as 7 economias mundiais no ano de 2019 também foram as nações com maior produção industrial. Na indústria da transformação, o Brasil cai uma posição no ranking mundial de produção, tendo uma parcela de 1,32%. Esses valores mostram que a indústria brasileira não está conseguindo acompanhar as indústrias de outros países com economias de tamanho similar.

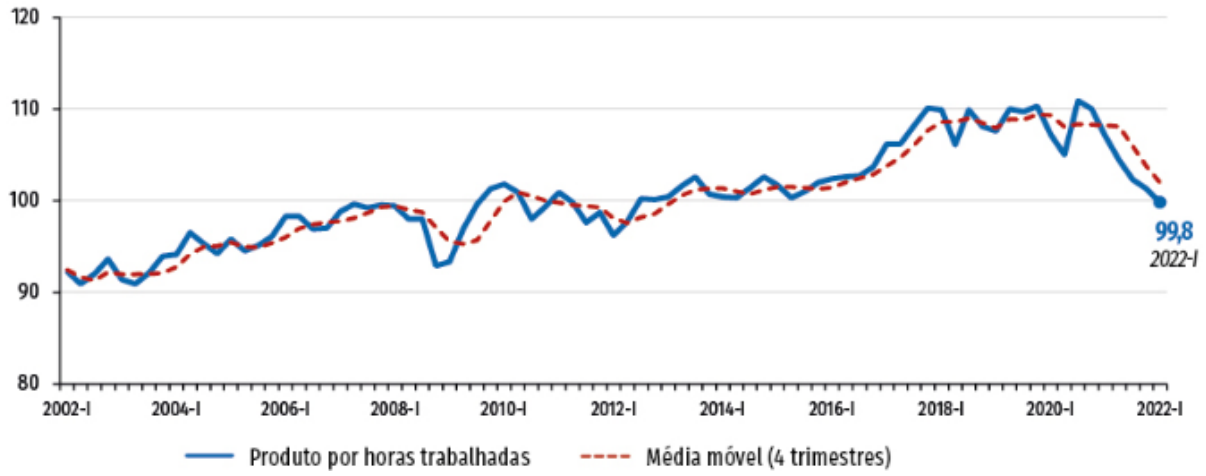
Um estudo de Competitividade da Indústria desenvolvido pela CNI em 2019 e 2020, aponta o Brasil como a penúltima nação mais competitiva, entre 18 países selecionados. O Brasil ficou a frente apenas da Argentina e atrás dos demais países latino-americanos que participaram do estudo, que são Chile, México, Colômbia e Peru.

Figura 1: Colocação dos países no estudo Competitividade Brasil

Fonte: Portal da Indústria

A indústria da transformação brasileira também não tem apresentado um bom resultado de produtividade. O setor apresentou, no primeiro trimestre de 2022, a sexta queda seguida no índice de produtividade calculado pela CNI, e não apresentava um resultado baixo assim desde 2015.

Figura 2: Produtividade do trabalho na indústria de transformação brasileira



Fonte: Portal da Indústria

Ampliando ainda mais um pouco o foco, um grupo de indústrias que costuma passar por maiores dificuldades dentro do cenário competitivo nacional e muitas vezes depende de incentivos para se manterem, é a Pequena Indústria, composta pelas micro e pequenas empresas. A pequena indústria, mesmo com muitas dificuldades, ainda gera 29,5% do PIB brasileiro e 54% dos empregos com carteira assinada no país, segundo dados do SEBRAE de 2021.

Em estudos de 2022 da CNI, a pequena indústria ainda sofre mais com mudanças econômicas no Brasil. Dentre as principais queixas dos empresários do grupo na indústria da transformação, estão a falta ou alto custo da matéria prima, a elevada carga tributária e a competição desleal (PORTAL DA INDÚSTRIA, 2022). Estes problemas tornam o crescimento de produção e desenvolvimento destas empresas um processo ainda mais difícil de ser atingido.

1.1 JUSTIFICATIVA

Diante do cenário exposto, fica evidente a importância do setor industrial brasileiro para a economia nacional, mesmo sendo considerado menos competitivo e menos produtivo do que de países com economias similares ao Brasil. Ficou claro também como a ineficiência de empresas, setores ou até países como um todo pode interferir no crescimento e no desenvolvimento de cada uma dessas instituições. Investigar e analisar áreas que não estão sendo eficientes, para propor novas estratégias que mudem essa realidade, se torna uma atividade fundamental para qualquer empresa e, no caso de países, para acelerar o desenvolvimento econômico.

Com isso, a motivação deste trabalho é exatamente estudar um grupo de empresas do setor industrial brasileiro para entender sobre a eficiência dessas instituições, além de buscar pontos em que a melhora na eficiência pode ser benéfica para o crescimento econômico nacional.

1.2 OBJETIVOS GERAIS

O presente trabalho tem como objetivo geral desenvolver uma metodologia para o cálculo da eficiência de indústrias brasileiras apoiado em dados de cadastro das empresas na base da Receita Federal, incluindo dados de capital social, diversificação de atuação e informações em relação aos sócios.

1.3 OBJETIVOS ESPECÍFICOS

A partir da definição do objetivo geral, é possível classificar os objetivos específicos, que são:

- Definir a melhor metodologia de cálculo de eficiência para o projeto em questão;
- Definir as variáveis que são relevantes para um cálculo de eficiência de indústrias;
- Aplicar a metodologia de cálculo;
- Analisar os resultados encontrados, fazendo comparativos por diferentes pontos de vista;
- Identificar as indústrias mais eficientes;
- Buscar identificar indústrias com baixa eficiência, mas com potencial para melhorar.

1.4 ESTRUTURA DO TRABALHO

Este trabalho é estruturado em capítulos, divididos da seguinte forma: no primeiro é apresentada a introdução, onde são expostos o contexto geral do trabalho, assim como a justificativa e os objetivos do mesmo; o segundo capítulo aborda o referencial teórico, que será a fundamentação necessária para a aplicação do projeto; o terceiro capítulo apresenta a metodologia, trazendo a classificação, desenvolvimento e aplicação em etapas da pesquisa

realizada; no quarto capítulo são apresentados a aplicação do método de cálculo de eficiência definido e seus resultados; e por último, no quinto capítulo estão as considerações finais e conclusão do trabalho.

2 REFERENCIAL TEÓRICO

No referencial teórico são abordados conceitos e fundamentos necessários para a realização do trabalho. Com o intuito de contextualização em relação ao que foi realizado neste projeto, aqui serão expostos conceitos de pesquisa operacional, programação linear, DEA (*Data Envelopment Analysis* – Análise Envoltória de Dados) e suas aplicações.

2.1 PESQUISA OPERACIONAL (P.O.)

Os primeiros trabalhos com o nome de pesquisa operacional surgiram nos primórdios da Segunda Guerra Mundial (HILLIER e LIEBERMAN, 2006), quando comandantes militares de nações que participavam da guerra convocaram cientistas que pudessem aplicar uma abordagem científica para resolver questões táticas e estratégicas durante os conflitos, principalmente em relação a alocação de recursos, que se tornavam escassos durante o período. A aplicação da pesquisa operacional foi decisiva em diversos momentos da guerra e, ao fim do conflito, iniciou-se um interesse em utilizar os mesmos métodos com outra finalidades que não fosse a militar. Com o crescimento industrial acelerado e tensões políticas surgindo no pós-guerra, os mesmos cientistas e consultores que trabalharam durante o conflito, começaram a introduzir a pesquisa operacional no ambiente comercial, industrial e governamental.

Dessa forma, a pesquisa operacional começou a encontrar espaço auxiliando empresas em diferentes departamentos, incluindo a minimização dos custos de produção, a maximização das vendas e ainda a minimização do uso de recursos para manter as operações (GUPTA, 2021). A P.O. ainda foi usada por empresas de seguros, que usavam técnicas para definir as melhores taxas nos seus produtos, de forma que fosse favorável para a empresa. Na agricultura, modelos de P.O. são usados para incluir variáveis climáticas e de logística, encontrando soluções para produção e distribuição de alimentos para uma crescente população mundial. Em instituições governamentais, pesquisa operacional é uma ferramenta importante no planejamento para projetos econômicos e de desenvolvimento de um país.

O objetivo da pesquisa operacional é fornecer embasamento científico para auxiliar no processo de tomada de decisões para problemas envolvendo operações de sistemas, proporcionando uma solução ótima que seja de grande interesse de uma organização (GUPTA, 2021).

Um estudo de PO pode ser sintetizado em 6 fases usuais (HILLIER e LIEBERMAN, 2006):

1. Definir o problema e coletar dados;
2. Formular um modelo matemático;
3. Desenvolver um programa computacional a partir do modelo matemático;
4. Testar e aprimorar;
5. Preparar para uma aplicação contínua do modelo;
6. Implementar.

Ao longo das últimas décadas, os estudos em pesquisa operacional estão se aprimorando, ao ponto que agora é categorizada em diferentes áreas. Uma das primeiras dessas áreas e ainda muito importante para muitas aplicações, é a programação linear.

2.2 PROGRAMAÇÃO LINEAR

O desenvolvimento da programação linear pode ser visto como um dos mais importantes avanços científicos desde a década de 50 (HILLIER e LIEBERMAN, 2006), sendo hoje considerada uma ferramenta capaz de solucionar diversos problemas e fazer empresas pouparem muito dinheiro.

A programação linear utiliza modelos matemáticos para descrever um problema, que necessariamente é composto por funções lineares. Assim, pode ser vista como um planejamento de atividades para alcançar o melhor resultado possível para o objetivo especificado pelo modelo matemático. A aplicação mais tradicional para a programação linear é para a alocação de recursos: quando uma organização tem uma quantidade limitada de recursos, os *inputs*, e precisa obter os melhores resultados possíveis de produção ou financeiro, por exemplo, os *outputs*, sendo assim uma questão de otimização. Um dos primeiros métodos dessa área, que ainda é muito utilizado e extremamente eficiente, é o Método Simplex, desenvolvido por um matemático norte-americano em 1947 (GUPTA, 2021).

A aplicação de métodos de programação linear em um ambiente corporativo tem o papel de auxiliar na tomada de decisão com base em dados, como por exemplo, um analista financeiro que precisa obter um portfólio ideal de investimentos para um cliente, levando em consideração o rendimento e os riscos que são esperados ou um agente de marketing que pretende distribuir o orçamento para anúncios entre diferentes opções de canais de comunicação (COOK e ZHU,

2013). Estes exemplos demonstram a versatilidade do uso da programação linear para empresas e organizações em diferentes áreas, justificando assim o porquê de ser uma ferramenta muito valorizada e utilizada tendo sido desenvolvida após a guerra.

Na programação linear, a formulação ou modelação do problema é o processo de transformar uma questão verbalizada em uma equação matemática (COOK e ZHU, 2013). Aprender a modelar um problema de programação linear acontece com prática e experiência, mas existem características comuns na maioria dos casos, sendo possível definir um método para guiar iniciantes a formular um problema. O primeiro passo é entender o problema e a situação de forma detalhada, para então conseguir verbalizar qual é o objetivo a ser alcançado e quais são as restrições que podem limitar o seu resultado.

Um exemplo, seria uma empresa que pretende maximizar a produção, mas tem como restrições o número de funcionários, quantidade de horas trabalhadas, matéria prima. Com essas informações em mãos, define-se quais são as variáveis de decisão, que são aquelas que podem ser controladas pelo tomador de decisão. A partir disso, é possível elaborar um conjunto de equações algébricas, que pode ser de maximização ou de minimização, enquanto é restrita pelas demais equações.

Este problema é um exemplo de situação muito comum para empresas, e que pode ser resolvido aplicando um método de programação linear. Um desses métodos que tem sido utilizado em diferentes áreas é o *Data Envelopment Analysis* (DEA), mais voltada para a análise de desempenho.

2.3 DATA ENVELOPMENT ANALYSIS (DEA) – ANÁLISE ENVOLTÓRIA DE DADOS

Toda empresa sente a necessidade de avaliar o desempenho de suas operações para entender se estão desempenhando de acordo com o que a própria organização e o mercado esperam dela. Em alguns casos, essa análise de desempenho é realizada através da implementação de requisitos a serem alcançados para diversas atividades realizadas dentro de uma organização (COOK e ZHU, 2013). Como exemplos, uma indústria consegue facilmente usar um cálculo de quantidade produzida de acordo com a quantidade de funcionários e horas trabalhadas, ou então o tempo de ociosidade de uma máquina, e caso estas medidas alcancem uma meta imposta, pode-se dizer que a empresa está sendo eficiente. Muitas vezes, esses

requisitos são definidos através da atividade de *benchmarking*, muito comum no mundo corporativo extremamente competitivo.

No entanto, existem casos em que essa atividade apenas não é suficiente para determinar a eficiência de uma empresa. Em casos de bancos, hospitais, escolas, e muitas organizações do setor de serviços, é muito difícil determinar tais metas levando em consideração apenas informações de mercado e, muitas vezes, não são capazes de medir o desempenho da melhor forma.

Uma possível abordagem para casos como estes é aplicar um método de programação linear, *Data Envelopment Analysis* (DEA). A teoria do DEA foi desenvolvida em 1978 (CHARNES; COOPER; RHODES, 1978), e até os anos 2000 era usada quase exclusivamente no mundo acadêmico para medir e gerenciar o desempenho de organizações. O método começou a ser introduzido no mundo corporativo com a chegada de novas ferramentas que facilitaram o uso do DEA sem a necessidade de possuir um extensivo conhecimento de programação linear. Os próprios desenvolvedores, Charnes, Cooper e Rhodes descrevem o DEA como uma técnica de programação matemática para determinar a eficiência relativa de Unidades de Tomada de Decisão (*Decision Making Units* - DMUs).

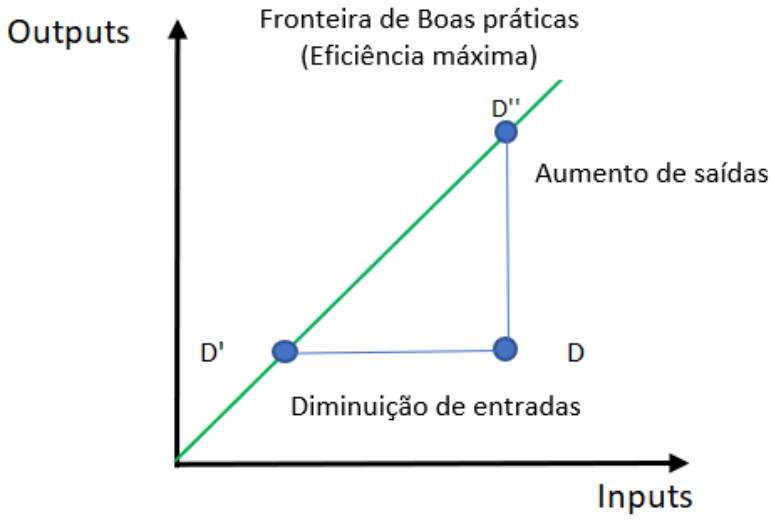
DEA pode ser então vista como uma visão orientada a dados para avaliar o desempenho de uma dada quantidade de entidades do mesmo formato, as DMUs (ZHU, 2014). As DMUs são os objetos do estudo na análise, podendo ser empresas, produtos, escolas, universidades, cidades, atletas e muitos outros. As aplicações desse método também podem ser encontradas em diferentes áreas, como mercado financeiro, marketing, transporte, esportes, contabilidade, energia, seguradora.

Uma característica do DEA que popularizou o seu uso dentre outros métodos de programação linear é o fato de não ser necessárias muitas suposições por parte de quem está fazendo a análise. Isso torna o método mais objetivo, e possibilita a aplicação para casos mais complexos, quando não se sabe exatamente sobre a relação entre as diferentes métricas de cada DMU.

A metodologia do DEA trabalha com o conceito de fronteiras de eficiência das DMUs. Todas as DMUs possuem entradas (*inputs*), que são seus recursos e saídas (*outputs*), que são seus resultados produtivos. Tendo como base a figura 3, a DMU D tem uma quantidade de *inputs* para gerar um valor de *outputs*, porém pelo gráfico, percebe-se que ela está abaixo da

fronteira de eficiência definida. Com isso, a DMU D tem duas opções: reduzir suas entradas mantendo as saídas até alcançar o ponto D', ou então aumentar suas saídas mantendo as entradas, até alcançar o ponto D''.

Figura 3: Gráfico da fronteira de eficiência DEA



Fonte: Adaptado de Cook e Zhu (2013)

Por conta desses diferentes caminhos que podem ser seguidos, o DEA possui duas orientações. Uma orientada para *inputs*, que busca minimizar os recursos utilizados para alcançar o mesmo resultado, e outra orientada para os *outputs*, que busca maximizar os resultados mantendo a quantidade de recursos usados. O modelo DEA representado na figura 4 é um modelo orientado para *inputs*.

Figura 4: Modelo matemático DEA

$$\vartheta^{\circ} = \min \vartheta$$

Sujeito a

$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{io} \qquad i = 1,2,\ldots,m;$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro} \qquad r = 1,2,\ldots,s;$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j \geq 0 \qquad j = 1,2,\ldots,n.$$

Fonte: Adaptado de Cook e Zhu (2013)

Neste modelo, θ é a equação a ser minimizada, respeitando um conjunto de restrições. As variáveis X e Y são, respectivamente, os *inputs* e os *outputs* para o modelo e λ são os multiplicadores que vão ser definidos através dos cálculos. A variável J corresponde a identificação de cada DMU do modelo.

2.4 PROBLEMA DE CÁLCULO DE EFICIÊNCIA UTILIZANDO MODELO DEA

Um exemplo comumente utilizado para exemplificar o método DEA é a análise comparativa entre laptops, os DMUs deste caso, tendo suas configurações como insumos. A figura 5 apresenta uma lista de laptops, com 5 configurações, que serão as variáveis para o modelo: preço, peso, duração da bateria, memória RAM e capacidade de armazenamento no HDD. Neste caso, é desejável um laptop que seja mais barato e mais leve, que tenha uma bateria mais duradoura, maior capacidade da memória RAM e maior armazenamento no HDD (ZHU, 2014).

Figura 5: Lista de laptops

	A	B	C	D	E	F
1	Modelo	Preço	Peso	Bateria	RAM	HDD
2	HP (11-E010nr)	414,99	3,4	5,00	4	500
3	Acer (V5-131)	339,98	3,3	6,50	4	500
4	Acer (AS1410)	449,99	3,1	6,00	2	250
5	Lenovo (X100e)	498,88	3,3	7,50	1	250
6	Acer (V5-171)	683,67	3,0	5,00	6	500
7	ASUS (X200CA)	299,99	2,6	5,00	2	500
8	ASUS (Q200E)	399,99	3,1	11,00	4	500
9	Lenovo (IdeaPad S210)	279,99	3,1	4,00	4	500
10	Dell (Inspiron 11)	299,99	3,1	8,00	2	500
11	Acer (S3-391-6046)	405,00	3,0	3,00	4	320
12	Lenovo (IdeaPad Yoga 13)	1.029,00	3,3	8,00	4	500
13	Gateway (LT41P04u)	338,45	2,4	5,00	2	320
14	Samsung (Series 5)	799,99	3,7	5,50	4	500
15	ASUS (1015E)	259,00	2,8	7,50	2	320
16	ASUS (X202E)	669,58	3,0	4,50	4	500
17	Toshiba (T115D)	300,00	3,5	9,00	2	500
18	Sony (YB Series)	340,00	3,2	5,45	2	320

Fonte: Adaptado de Zhu (2014)

Para cada uma dessas variáveis, é definido um multiplicador. Inicialmente, todas as variáveis são vistas como igualmente importantes, ou seja, todas elas possuem o mesmo impacto para a tomada de decisão, por isso o multiplicador para todas as variáveis é igual a 1. O objetivo vai ser determinar os valores destes multiplicadores para alcançar o valor máximo de eficiência para cada laptop.

Figura 6: Divisão das variáveis de laptops em categorias

	A	B	C	D	E	F	G
1	Modelo	Preço	Peso		Bateria	RAM	HDD
2	HP (11-E010nr)	414,99	3,4		5,00	4	500
3	Acer (V5-131)	339,98	3,3		6,50	4	500
4	Acer (AS1410)	449,99	3,1		6,00	2	250
5	Lenovo (X100e)	498,88	3,3		7,50	1	250
6	Acer (V5-171)	683,67	3,0		5,00	6	500
7	ASUS (X200CA)	299,99	2,6		5,00	2	500
8	ASUS (Q200E)	399,99	3,1		11,00	4	500
9	Lenovo (IdeaPad S210)	279,99	3,1		4,00	4	500
10	Dell (Inspiron 11)	299,99	3,1		8,00	2	500
11	Acer (S3-391-6046)	405,00	3,0		3,00	4	320
12	Lenovo (IdeaPad Yoga 13)	1.029,00	3,3		8,00	4	500
13	Gateway (LT41P04u)	338,45	2,4		5,00	2	320
14	Samsung (Series 5)	799,99	3,7		5,50	4	500
15	ASUS (1015E)	259,00	2,8		7,50	2	320
16	ASUS (X202E)	669,58	3,0		4,50	4	500
17	Toshiba (T115D)	300,00	3,5		9,00	2	500
18	Sony (YB Series)	340,00	3,2		5,45	2	320
19							
20	Multiplicadores	1	1		1	1	1

Fonte: Adaptado de Zhu (2014)

O próximo passo é dividir essas 5 variáveis em duas categorias, a categoria 1 de variáveis que é desejável minimizar e a categoria 2 de variáveis que é desejável maximizar. Percebe-se que segue a lógica *inputs* e *outputs*, sendo os inputs as variáveis que se desejam minimizar e os outputs as variáveis que se desejam maximizar. Nas duas categorias, é calculado a soma dos valores de cada variável, levando em consideração seus multiplicadores, que ainda são iguais a 1. Posteriormente, é feita razão entre as somas das categorias 2 e 1, chegando assim a um índice único para cada laptop, que enfim pode ser usado como medida para comparação entre eles, sendo que quanto maior este valor, melhor será considerado o laptop.

Figura 7: Resultado da eficiência dos laptops utilizando multiplicadores iguais

	A	B	C	D	E	F	G	H	I	J	K	L
1	Modelo	Preço	Peso		Bateria	RAM	HDD		Valor Categoria 1	Valor Categoria 2		Razão
2	HP (11-E010nr)	414,99	3,4		5,00	4	500		418,39	509,00		1,22
3	Acer (V5-131)	339,98	3,3		6,50	4	500		343,28	510,50		1,49
4	Acer (AS1410)	449,99	3,1		6,00	2	250		453,09	258,00		0,57
5	Lenovo (X100e)	498,88	3,3		7,50	1	250		502,18	258,50		0,51
6	Acer (V5-171)	683,67	3,0		5,00	6	500		686,67	511,00		0,74
7	ASUS (X200CA)	299,99	2,6		5,00	2	500		302,59	507,00		1,68
8	ASUS (Q200E)	399,99	3,1		11,00	4	500		403,09	515,00		1,28
9	Lenovo (IdeaPad S210)	279,99	3,1		4,00	4	500		283,09	508,00		1,79
10	Dell (Inspiron 11)	299,99	3,1		8,00	2	500		303,09	510,00		1,68
11	Acer (S3-391-6046)	405,00	3,0		3,00	4	320		408,00	327,00		0,80
12	Lenovo (IdeaPad Yoga 13)	1.029,00	3,3		8,00	4	500		1032,30	512,00		0,50
13	Gateway (LT41P04u)	338,45	2,4		5,00	2	320		340,85	327,00		0,96
14	Samsung (Series 5)	799,99	3,7		5,50	4	500		803,69	509,50		0,63
15	ASUS (1015E)	259,00	2,8		7,50	2	320		261,80	329,50		1,26
16	ASUS (X202E)	669,58	3,0		4,50	4	500		672,58	508,50		0,76
17	Toshiba (T115D)	300,00	3,5		9,00	2	500		303,50	511,00		1,68
18	Sony (YB Series)	340,00	3,2		5,45	2	320		343,20	327,45		0,95
19												
20	Multiplicadores	1	1		1	1	1					

Fonte: Adaptado de Zhu (2014)

Contudo, este cálculo não é o melhor a ser considerado, porque as variáveis possuem dimensões diferentes. Um computador que tenha uma bateria mais duradoura ou que tenha boa capacidade de memória RAM podem acabar prejudicados, porque o valor dessas variáveis é menor em comparação com a capacidade de armazenamento HDD, por exemplo. Por isso, é injusto que o multiplicador para todas as variáveis seja o mesmo.

Como dito antes, o objetivo do método para resolver o problema é identificar os valores ideais dos multiplicadores que maximizam o índice de eficiência calculado. Porém, dessa vez serão impostas restrições para o cálculo. A primeira restrição é que a razão entre o valor somado da categoria 2 (*outputs*) pelo da categoria 1 (*inputs*) é menor ou igual a 1 (Equação 6).

$$\frac{\text{Valor Categoria 1}}{\text{Valor Categoria 2}} \leq 1$$

(Eq. 1)

A segunda restrição é que o valor da soma na categoria 2 para um dos laptops é igual a 1. Isso porque este é um caso de DEA orientado para *inputs*, portanto o valor dos *outputs* tem que se manter o mesmo. Com essas restrições, o valor máximo que pode ser alcançado para o índice de eficiência é igual a 1.

Agora tem-se um modelo de DEA que tem como objetivo maximizar o índice de eficiência para cada um dos laptops, tendo duas restrições.

Finalmente, escolhe-se um dos laptops para que seja o primeiro objeto de estudo do modelo e calculado a sua eficiência. Aqui, o primeiro escolhido foi o laptop Acer (VS-171), e é possível observar que o resultado do modelo é igual a 1, ou seja, este laptop obteve nota máxima. Isso significa que foi possível obter um conjunto de multiplicadores que faz a medida da categoria 2 (*Weighted Category II*) ser igual a 1.

Figura 8: Resultado da eficiência do primeiro laptop

	A	B	C	D	E	F	G	H	I	J	K	L
1	Modelo	Preço	Peso		Bateria	RAM	HDD		Valor Categoria 1	Valor Categoria 2		Razão
2	HP (11-E010nr)	414,99	3,4		5,00	4	500		1,13333	0,66667		0,58824
3	Acer (V5-131)	339,98	3,3		6,50	4	500		1,09999	0,66667		0,60607
4	Acer (AS1410)	449,99	3,1		6,00	2	250		1,03333	0,33333		0,32258
5	Lenovo (X100e)	498,88	3,3		7,50	1	250		1,09999	0,16667		0,15152
6	**Acer (V5-171)**	**683,67**	**3,0**		**5,00**	**6**	**500**		**0,99999**	**1,00000**		**1**
7	ASUS (X200CA)	299,99	2,6		5,00	2	500		0,86667	0,33333		0,38461
8	ASUS (Q200E)	399,99	3,1		11,00	4	500		1,03333	0,66667		0,64517
9	Lenovo (IdeaPad S210)	279,99	3,1		4,00	4	500		1,03333	0,66667		0,64517
10	Dell (Inspiron 11)	299,99	3,1		8,00	2	500		1,03333	0,33333		0,32258
11	Acer (S3-391-6046)	405,00	3,0		3,00	4	320		0,99999	0,66667		0,66668
12	Lenovo (IdeaPad Yoga 13)	1.029,00	3,3		8,00	4	500		1,09999	0,66667		0,60607
13	Gateway (LT41P04u)	338,45	2,4		5,00	2	320		0,79999	0,33333		0,41667
14	Samsung (Series 5)	799,99	3,7		5,50	4	500		1,23333	0,66667		0,54054
15	ASUS (1015E)	259,00	2,8		7,50	2	320		0,93333	0,33333		0,35714
16	ASUS (X202E)	669,58	3,0		4,50	4	500		0,99999	0,66667		0,66668
17	Toshiba (T115D)	300,00	3,5		9,00	2	500		1,16667	0,33333		0,28571
18	Sony (YB Series)	340,00	3,2		5,45	2	320		1,06667	0,33333		0,31250
19												
20	Multiplicadores	0	0,33333		0	0,16667	0					

Fonte: Adaptado de Zhu (2014)

É importante ressaltar que os demais laptops não apresentaram o mesmo resultado porque os multiplicadores foram definidos para maximizar apenas o laptop escolhido. O último passo é exatamente aplicar o modelo para todos os demais laptops, definindo assim o valor otimizado para todos.

Figura 9: Resultado final da eficiência de todos os laptops

	A	B	C	D	E	F	G	H	I	J	K	L
1	Modelo	Preço	Peso		Bateria	RAM	HDD		Valor Categoria 1	Valor Categoria 2		Razão
2	HP (11-E010nr)	414,99	3,4		5,00	4	500		1,22	0,87		0,71
3	Acer (V5-131)	339,98	3,3		6,50	4	500		1,00	0,95		0,95
4	Acer (AS1410)	449,99	3,1		6,00	2	250		1,32	0,61		0,46
5	Lenovo (X100e)	498,88	3,3		7,50	1	250		1,47	1,02		0,69
6	Acer (V5-171)	683,67	3,0		5,00	6	500		2,01	0,73		0,36
7	ASUS (X200CA)	299,99	2,6		5,00	2	500		0,88	0,73		0,83
8	ASUS (Q200E)	399,99	3,1		11,00	4	500		1,18	1,18		1,00
9	Lenovo (IdeaPad S210)	279,99	3,1		4,00	4	500		0,82	0,82		1,00
10	Dell (Inspiron 11)	299,99	3,1		8,00	2	500		0,88	0,88		1,00
11	Acer (S3-391-6046)	405,00	3,0		3,00	4	320		1,19	0,65		0,55
12	Lenovo (IdeaPad Yoga 13)	1.029,00	3,3		8,00	4	500		3,03	1,03		1,03
13	Gateway (LT41P04u)	338,45	2,4		5,00	2	320		1,00	0,61		0,26
14	Samsung (Series 5)	799,99	3,7		5,50	4	500		2,35	0,90		1,18
15	ASUS (1015E)	259,00	2,8		7,50	2	320		0,76	0,74		0,38
16	ASUS (X202E)	669,58	3,0		4,50	4	500		1,97	0,85		0,43
17	Toshiba (T115D)	300,00	3,5		9,00	2	500		0,88	0,76		0,86
18	Sony (YB Series)	340,00	3,2		5,45	2	320		1,00	0,63		0,63

Fonte: Zhu (2014)

Após obter o resultado para todos os laptops, observa-se que apenas 7 laptops apresentaram o índice máximo de eficiência, podendo ser assim considerados como as melhores escolhas entre todos os que fizeram parte do problema.

2.5 APLICAÇÕES DO DEA EM DIFERENTES ÁREAS

Como foi comentado, o método DEA se tornou uma importante ferramenta no mundo empresarial e industrial, sendo aplicável em diversas áreas. Aqui estão alguns exemplos de aplicações do DEA em problemas reais.

Uma característica fundamental para qualquer empresa é ter um bom relacionamento com seus fornecedores, e a prática de troca de benefícios entre parceiros é comum no mundo empresarial. Em uma indústria do ramo petroleiro, é apresentada uma alternativa de avaliação de fornecedores, com o objetivo de identificar os principais parceiros para premiação (ROCHA E CAVALCANTI NETTO, 2012). A análise é feita utilizando o DEA em duas etapas, a primeira utilizando variáveis relativas às transações realizadas, e a segunda leva em consideração a opinião de pessoas com cargo de gestão, com o intuito de agregar a visão qualitativa para a aplicação.

O setor turístico pode ser muito também forte para a economia de algumas regiões, podendo até chegar a ser a atividade mais forte de um país. Um exemplo de região com turismo

muito forte é Algarve, em Portugal. Em um local com muita atividade turística, a indústria hoteleira também se torna muito importante, e a eficiência da rede hoteleira pode ser analisada usando o DEA. Aplicado o método para avaliar 28 hotéis na região de Algarve, foi possível identificar um elevado nível de ineficiência na região, além de quais modelos de gestão estão obtendo melhores resultados. (OLIVEIRA; PEDRO; MARQUES, 2015)

Conforme já mencionado, as origens da Pesquisa Operacional e da programação linear voltam até a época da segunda guerra mundial, servindo como auxílio para tomada de decisões estratégicas e de logística durante o período. Na Coréia do Sul, um projeto de aplicação de DEA pôde auxiliar comandos militares na avaliação de desempenho de esquadrões, ao definir um cenário otimizado (HAN E SOHN, 2011). O trabalho realizado foi aplicado para gerenciamento de eficiência em diferentes organizações militares.

O uso do DEA pode ser também acompanhado de outros modelos matemáticos, complementando um ao outro e fornecendo um bom resultado para o estudo. Em um estudo sobre a eficiência do sistema de saúde no estado do Rio de Janeiro, foi possível identificar municípios que estavam sendo ineficientes e, com o uso de regressão linear, identificar quais variáveis produziam um maior impacto no desempenho dos sistemas de saúde (MEDEIROS E MARCOLINO, 2018).

Devido à complexidade da Pesquisa Operacional e, mais especificamente, do DEA, faz-se necessário o uso de programação.

2.6 LINGUAGENS DE PROGRAMAÇÃO

Uma linguagem de programação é um conjunto de comandos e instruções escritas para gerar programas que são executadas por um computador. Existem muitas linguagens de programação, cada uma com suas regras e aplicações. Uma delas que tem se tornado cada vez mais popular por ser relativamente simples de aprender, além de ser muito versátil, é a *Python*. Os programas desenvolvidos nessa linguagem possuem aplicações em variadas áreas, como educação, desenvolvimento da web e de softwares, científico e empresarial.

As bibliotecas para linguagens de programação são conjuntos de módulos e funções que reduzem o uso de código no programa, o que simplifica a maneira de realizar comandos usando a linguagem sem que fique um texto muito longo. As bibliotecas *Python* estão disponíveis de

forma aberta para todos, sendo ferramentas muito potentes que ajudaram a popularizar a linguagem, principalmente quando se tem a finalidade de trabalhar com dados.

Duas das mais populares bibliotecas *Python* para análise e ciência de dados são *Numpy* e *Pandas*. *Numpy* é usada para fazer o processamento de matrizes e vetores, que são classificadas como *array* no código, e é muito usada para manipular informações deste tipo de forma rápida e eficiente, mesmo com grande volume de dados. Ela é muito importante porque permite integração com outras linguagens de programação, como a linguagem C, também muito popular para o tratamento de dados. Já a biblioteca *Pandas* se destaca muito por ser muito completa e relativamente fácil de usar, mesmo para programadores com pouca experiência. Com ela, é possível trabalhar com uma diversidade muito grande de dados, organizados em *dataframes*, além de fornecer suporte para vários tipos de arquivo, como excel, e possibilitar o trabalho com várias bases de dados simultaneamente.

Outra ferramenta entre as mais utilizadas no mundo é a biblioteca SQLite, em linguagem C de programação. A principal aplicação do SQLite é o armazenamento e manipulação de uma base de dados, sendo a ferramenta mais utilizada para essa finalidade no mundo. O próximo capítulo descreve a metodologia do presente trabalho.

3 METODOLOGIA

Neste capítulo, a metodologia empregada no estudo é classificada a partir de diferentes categorias. As etapas do projeto serão descritas, apresentando como o estudo foi realizado, sendo possível chegar aos resultados para análise e conclusão.

3.1 CLASSFICAÇÃO DA PESQUISA

A metodologia, no ambiente de pesquisa científica, pode ser classificada como um conjunto de métodos utilizados com o objetivo de resolver um problema, recorrendo a procedimentos científicos (SILVEIRA e CÓRDOVA, 2009). Uma pesquisa pode ser classificada a partir de algumas categorias, de acordo com sua abordagem e natureza, além de por objetivos e procedimentos técnicos (GIL, 2011).

Quanto à abordagem, uma pesquisa pode ser classificada como quantitativa, quando são buscados resultados quantitativos, sendo possível focar em uma análise mais objetiva e focada em fundamentos matemáticos ou, como qualitativa, quando não são utilizados dados numéricos para análise, tendo assim um foco maior em análises mais subjetivas.

Quanto à natureza, uma pesquisa pode ser classificada como básica, quando se tem como objetivo de pesquisa desenvolver novos conhecimentos sem aplicações práticas anteriores, ou aplicada, quando o objetivo é gerar conhecimento através de uma aplicação prática de soluções, para um problema específico.

Com relação aos objetivos, a pesquisa pode ser classificada como exploratória, que busca explorar melhor um assunto para tentar extrair hipóteses ou descritiva, que objetiva descrever uma realidade observada, ou ainda explicativa, que tenta trazer explicações para um fenômeno real.

Por sua vez, quanto aos procedimentos, uma pesquisa pode ser classificada de acordo com o conjunto de técnicas de pesquisa utilizado. Há 7 classificações para pesquisa de acordo com os procedimentos: bibliográfica, quando são utilizados livros e artigos; documental, quando são utilizados informações direto da fonte, sem nenhuma análise prévia; levantamento, quando são utilizados informações obtidas através de questionamentos para um público alvo; experimental, quando são selecionados dados que podem influenciar um objeto de estudo, sendo testada essa hipótese; estudo de campo, quando a pesquisa é feita através da observação

de um fenômeno, para buscar entendê-lo e explicá-lo; estudo de caso, quando é realizado um estudo aprofundado e detalhado sobre um assunto; e, por último, pesquisa-ação, quando o estudo é realizado em cooperação entre os participantes para a resolução de um problema coletivo.

Uma pesquisa também pode ser descrita no contexto de modelagens qualitativas, quantitativas ou informacionais, considerando o processo como realização de funções de um sistema sob uma estrutura e a tipologia de metodologias qualitativas (BERTRAND E FRANSOO, 2002). As pesquisas podem ser classificadas como Axiomática Quantitativa, dirigida a modelos com foco em soluções e que produzem conhecimento sobre o comportamento de variáveis, ou como Empírica Quantitativa, que busca assegurar que existe adesão entre uma observação da realidade e o modelo elaborado para entender esta realidade. Elas ainda vão ser classificadas em Normativa, quando desenvolve normas e estratégias para solucionar problemas de um sistema, ou Descritiva, quando desenvolve um modelo para descrever as relações de um processo real.

No estudo apresentado neste trabalho, a pesquisa é classificada como uma abordagem quantitativa, de natureza aplicada, com o objetivo descritivo e utilizando embasamento bibliográfico, quanto aos procedimentos. Quanto ao contexto de modelagens, a pesquisa é classificada como Empírica Quantitativa Descritiva.

3.2 ETAPAS DO TRABALHO

O presente trabalho foi dividido nas seguintes etapas: (i) Definição do problema; (ii) Definição dos objetivos gerais e específicos do estudo;(iii) Revisão da literatura referente a modelos de pesquisa operacional para análise comparativa das indústrias e ferramentas de aplicação; (iv) Busca e coleta dos dados que serão base para o modelo definido; (v) Manipulação dos dados encontrados utilizando linguagens de programação; (vi) Aplicação do modelo DEA utilizando a ferramenta PyDEA; (vii) Análise e discussão dos resultados encontrados; (viii) Conclusão e considerações finais.

As primeiras três etapas do trabalho já foram apresentadas, a seguir estão expostas desde a coleta dos dados até a conclusão do estudo.

3.3 COLETA DE DADOS

A coleta dos dados para a realização do projeto é o primeiro passo a ser dado na etapa de aplicação do método definido. O uso de dados confiáveis e relevantes, com volume suficiente grande para a aplicação do modelo matemático é de extrema importância e possui impacto direto na qualidade e relevância dos resultados obtidos. Por isso, neste projeto foram extraídas bases de dados públicas no portal da Receita Federal. Foram extraídas as bases que possuem informações públicas de empresas, do programa Simples Nacional, de estabelecimentos, de sócios, de qualificações de sócios e de CNAEs cadastrados na Receita Federal. A base foi extraída em maio de 2022, logo os dados que serão apresentados representam a situação das empresas neste período.

O Simples Nacional é um programa da Receita Federal de regime simplificado de arrecadação, cobrança e fiscalização de atributos, aplicável a micro e pequenas empresas. O objetivo do programa é facilitar a forma de pagamento para as empresas de menor porte, dando a opção de arrecadação de tributos através de uma única guia, além de uma possível redução da carga tributária para estas empresas.

Os dados são baixados no formato DB, banco de dados SQL, sendo organizados e armazenados por meio do software DB Browser, que utiliza a biblioteca SQLite da linguagem C de programação. Seguem abaixo as bases extraídas do site da Receita Federal.

Tabela 1: Dados da base Empresas da Receita Federal

BASE EMPRESAS	
CAMPO	**DESCRIÇÃO**
CNPJ BÁSICO	NÚMERO BASE DE INSCRIÇÃO NO CNPJ (OITO PRIMEIROS DÍGITOS DO CNPJ)
RAZÃO SOCIAL / NOME EMPRESARIAL	NOME EMPRESARIAL DA PESSOA JURÍDICA
NATUREZA JURÍDICA	CÓDIGO DA NATUREZA JURÍDICA
QUALIFICAÇÃO DO RESPONSÁVEL	QUALIFICAÇÃO DA PESSOA FÍSICA RESPONSÁVEL PELA EMPRES
CAPITAL SOCIAL DA EMPRESA	CAPITAL SOCIAL DA EMPRESA
PORTE DA EMPRESA	CÓDIGO DO PORTE DA EMPRESA: 00 – NÃO INFORMADO 01 - MICRO EMPRESA 03 - EMPRESA DE PEQUENO PORTE 05 - DEMAIS
ENTE FEDERATIVO RESPONSÁVEL	O ENTE FEDERATIVO RESPONSÁVEL É PREENCHIDO PARA OS CASOS DE ÓRGÃOS E ENTIDADES DO

Fonte: Receita Federal

A base apresentada na tabela 1 possui informações básicas de toda empresa cadastradas na Receita Federal, como CNPJ, razão social, capital social e também o porte da empresa.

Tabela 2: Dados da base Simples da Receita Federal

BASE SIMPLES	
CAMPO	**DESCRIÇÃO**
CNPJ BÁSICO	NÚMERO BASE DE INSCRIÇÃO NO CNPJ (OITO PRIMEIROS DÍGITOS DO CNPJ)
OPÇÃO PELO SIMPLES	INDICADOR DA EXISTÊNCIA DA OPÇÃO PELO SIMPLES: S - SIM N - NÃO EM BRANCO – OUTROS
DATA DE OPÇÃO PELO SIMPLES	DATA DE OPÇÃO PELO SIMPLES
DATA DE EXCLUSÃO DO SIMPLES	DATA DE EXCLUSÃO DO SIMPLE
OPÇÃO PELO MEI	INDICADOR DA EXISTÊNCIA DA OPÇÃO PELO MEI: S - SIM N - NÃO EM BRANCO - OUTROS
DATA DE OPÇÃO PELO MEI	DATA DE OPÇÃO PELO MEI
DATA DE EXCLUSÃO DO MEI	DATA DE EXCLUSÃO DO MEI

Fonte: Receita Federal

A base apresentada na tabela 2 é referente às informações do programa Simples Nacional, com as informações de empresas cadastradas e as datas de entrada e saída do programa.

Tabela 3: Dados da base de Estabelecimentos da Receita Federal

BASE ESTABELECIMENTOS	
CAMPO	**DESCRIÇÃO**
CNPJ BÁSICO	NÚMERO BASE DE INSCRIÇÃO NO CNPJ (OITO PRIMEIROS DÍGITOS DO CNPJ)
CNPJ ORDEM	NÚMERO DO ESTABELECIMENTO DE INSCRIÇÃO NO CNPJ (DO NONO ATÉ O DÉCIMO SEGUNDO DÍGITO DO CNP
CNPJ DV	DÍGITO VERIFICADOR DO NÚMERO DE INSCRIÇÃO NO CNPJ (DOIS ÚLTIMOS DÍGITOS DO CNPJ)
IDENTIFICADOR MATRIZ/FILIAL	CÓDIGO DO IDENTIFICADOR MATRIZ/FILIAL: 1 – MATRIZ 2 – FILIAL
NOME FANTASIA	CORRESPONDE AO NOME FANTASIA
SITUAÇÃO CADASTRAL	CÓDIGO DA SITUAÇÃO CADASTRAL: 01 – NULA 2 – ATIVA 3 – SUSPENSA 4 – INAPTA 08 – BAIXADA
DATA SITUAÇÃO CADASTRAL	DATA DO EVENTO DA SITUAÇÃO CADASTRAL
MOTIVO SITUAÇÃO CADASTRAL	CÓDIGO DO MOTIVO DA SITUAÇÃO CADASTRAL
NOME DA CIDADE NO EXTERIOR	NOME DA CIDADE NO EXTERIOR
PAIS	CÓDIGO DO PAIS
DATA DE INÍCIO ATIVIDADE	DATA DE INÍCIO DA ATIVIDADE
CNAE FISCAL PRINCIPAL	CÓDIGO DA ATIVIDADE ECONÔMICA PRINCIPAL DO ESTABELECIMENTO
CNAE FISCAL SECUNDÁRIA	CÓDIGO DA(S) ATIVIDADE(S) ECONÔMICA(S) SECUNDÁRIA(S) DO ESTABELECIMENTO
TIPO DE LOGRADOURO	DESCRIÇÃO DO TIPO DE LOGRADOURO
LOGRADOURO	NOME DO LOGRADOURO ONDE SE LOCALIZA O ESTABELECIMENTO
NÚMERO	NÚMERO ONDE SE LOCALIZA O ESTABELECIMENTO. QUANDO NÃO HOUVER PREENCHIMENTO DO NÚMERO HAVERÁ 'S/N'
COMPLEMENTO	COMPLEMENTO PARA O ENDEREÇO DE LOCALIZAÇÃO DO ESTABELECIMENTO
BAIRRO	BAIRRO ONDE SE LOCALIZA O ESTABELECIMENTO.
CEP	CÓDIGO DE ENDEREÇAMENTO POSTAL REFERENTE AO LOGRADOURO NO QUAL O ESTABELECIMENTO ESTA LOCALIZADO
UF	SIGLA DA UNIDADE DA FEDERAÇÃO EM QUE SE ENCONTRA O ESTABELECIMENTO
MUNICÍPIO	CÓDIGO DO MUNICÍPIO DE JURISDIÇÃO ONDE SE ENCONTRA O ESTABELECIMENTO
DDD 1	CONTÉM O DDD
TELEFONE 1	CONTÉM O NÚMERO DO TELEFONE 1
DDD 2	CONTÉM O DDD 2
TELEFONE 2	CONTÉM O NÚMERO DO TELEFONE 2
DDD DO FAX	CONTÉM O DDD DO FAX
FAX	CONTÉM O NÚMERO DO FAX
CORREIO ELETRÔNICO	CONTÉM O E-MAIL DO CONTRIBUINTE
SITUAÇÃO ESPECIAL	SITUAÇÃO ESPECIAL DA EMPRESA
DATA DA SITUAÇÃO ESPECIAL	DATA EM QUE A EMPRESA ENTROU EM SITUAÇÃO ESPECIAL

Fonte: Receita Federal

A base de Estabelecimentos (Tabela 3) possui dados de empresas por instalação, ou seja, pode incluir mais de um cadastro por empresa. Um exemplo seria o caso de uma empresa com várias filiais. As informações desta base serão complementares à base de empresas, como a situação cadastral, datas de início e encerramento das atividades, e códigos CNAE. O código CNAE é a Classificação Nacional de Atividades Econômicas, e consiste em um número que identifica a atividade econômica exercida por uma empresa.

Tabela 4: Dados da base Sócios da Receita Federal

BASE SÓCIOS	
CAMPO	**DESCRIÇÃO**
CNPJ BÁSICO	NÚMERO BASE DE INSCRIÇÃO NO CNPJ (CADASTRO NACIONAL DA PESSOA JURÍDICA)
IDENTIFICADOR DE SÓCIO	CÓDIGO DO IDENTIFICADOR DE SÓCIO: 1 – PESSOA JURÍDICA 2 – PESSOA FÍSICA 3 – ESTRANGEIRO
NOME DO SÓCIO (NO CASO PF) OU RAZÃO SOCIAL (NO CASO PJ)	NOME DO SÓCIO PESSOA FÍSICA OU A RAZÃO SOCIAL E/OU NOME EMPRESARIAL DA PESSOA JURÍDICA E/OU NOME DO SÓCIO/RAZÃO SOCIAL DO SÓCIO ESTRANGEIRO
CNPJ/CPF DO SÓCIO	CPF OU CNPJ DO SÓCIO (SÓCIO ESTRANGEIRO NÃO TEM ESTA INFORMAÇÃO)
QUALIFICAÇÃO DO SÓCIO	CÓDIGO DA QUALIFICAÇÃO DO SÓCIO
DATA DE ENTRADA SOCIEDADE	DATA DE ENTRADA NA SOCIEDADE
PAIS	CÓDIGO PAÍS DO SÓCIO ESTRANGEIRO
REPRESENTANTE LEGAL	NÚMERO DO CPF DO REPRESENTANTE LEGAL
NOME DO REPRESENTANTE	NOME DO REPRESENTANTE LEGAL
QUALIFICAÇÃO DO REPRESENTANTE LEGAL	CÓDIGO DA QUALIFICAÇÃO DO REPRESENTANTE LEGAL
FAIXA ETÁRIA	CÓDIGO CORRESPONDENTE À FAIXA ETÁRIA DO SÓCIO

Fonte: Receita Federal

A Base Sócios (Tabela 4) possui informações dos proprietários das empresas. Por ela, serão fornecidos dados de faixa etária, sendo possível determinar a quantidade de sócios de cada empresa.

Tabela 5: Dados da base CNAEs da Receita Federal

BASE CNAEs	
CAMPO	**DESCRIÇÃO**
CÓDIGO	CÓDIGO DA ATIVIDADE ECONÔMICA
DESCRIÇÃO	NOME DA ATIVIDADE ECONÔMICA

Fonte: Receita Federal

A Base CNAEs (Tabela 5) possui a identificação de cada atividade econômica por código. Agora, com os dados em mãos, é necessário criar ligações entre as bases para gerar um conjunto único de dados com todas as informações que serão necessárias para à aplicação do modelo no projeto. Chaves, que são os campos comuns em mais de uma base, vão ser usadas para fazer essa conexão entre elas, da seguinte forma:

(i) Base Empresas é conectada com a Base Estabelecimentos utilizando o campo CNPJ Básico como chave;

(ii) Base Estabelecimentos é conectada com a Base Simples utilizando também o campo CNPJ Básico como chave;

(iii) Base Estabelecimentos é conectada com a Base Sócios utilizando mais uma vez o campo CNPJ Básico como chave;

(iv) Base Estabelecimentos é conectada com a Base CNAEs utilizando o código CNAE dos campos CNAE Fiscal Principal e Código, respectivamente.

3.4 FILTRAGEM, MANIPULAÇÃO E LIMPEZA DOS DADOS

Para acessar os bancos de dados no SQL, é desenvolvido um código na linguagem *Python* de programação, por onde será possível manipular os dados e chegar a um conjunto de variáveis que vão ser utilizadas no modelo análise definido. O código foi desenvolvido

utilizando o *Jupyter Notebook*, plataforma gratuita para desenvolvimento de programas em diversas linguagens de programação.

O primeiro passo a ser tomado para escrever um código com o objetivo de realizar a análise dos dados é acessar as bibliotecas do *Python*. Com as bibliotecas, também será possível realizar um comando para acessar os bancos de dados do SQL.

A segunda etapa do código tem como objetivo conectar as bases utilizando chaves, que são os campos que estão presentes em mais de uma base. Neste caso, os campos CNPJ básico e CNAE Fiscal foram utilizados para conectar as bases. Nesta etapa também inicia-se o processo de filtrar os dados para tratar apenas as medidas que vão ser utilizadas. Os filtros aplicados foram:

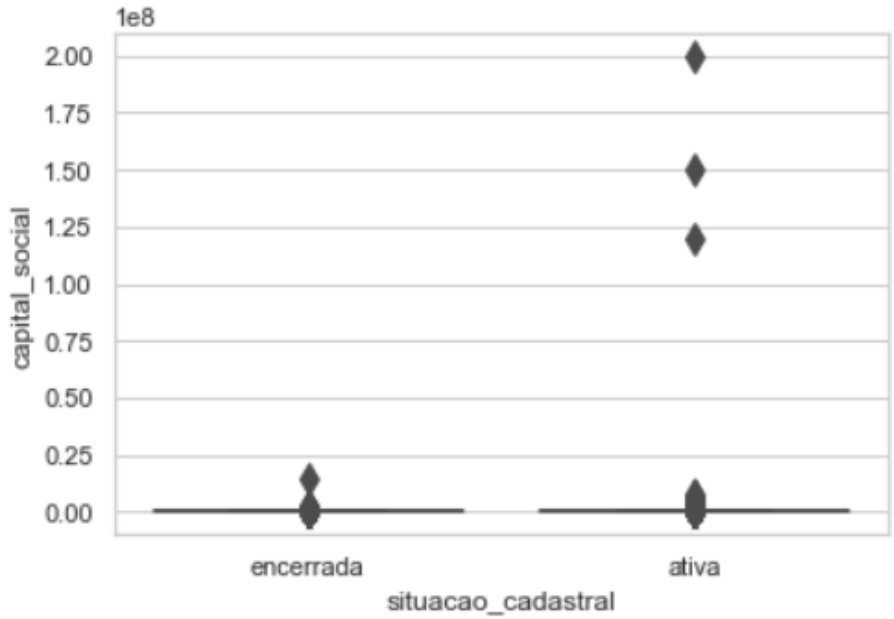
- Selecionar apenas empresas que possuem informações sobre atividade econômica;
- Selecionar apenas empresas que possuem informações sobre os sócios na base da Receita Federal;
- Selecionar apenas sócios que estão desde a abertura da empresa;
- Selecionar apenas empresas que possuem a situação cadastral como Ativa ou Baixada;
- Selecionar apenas empresas de porte classificadas como Micro Empresa ou Empresa de Pequeno Porte;
- Selecionar apenas empresas com o identificador de Matriz;
- Selecionar apenas empresas com data de início no ano de 2011.

Entende-se micros e pequenas empresas na indústria como organizações que empregam até 99 pessoas ou com faturamento anual de até R$ 4,8 milhões. A opção pela seleção apenas de empresas com data de início em 2011, ou seja, que foram fundadas neste ano, é com o intuito de considerar um período de tempo que seja suficiente para que as empresas amadurecessem e crescessem, ou então que acabassem não dando certo e fechassem. Segundo dados do IBGE, mais de 70% das empresas abertas no Brasil acabam fechando as portas em até dez anos de atividade. Cerca de 1 a cada 5 novas empresas fecham ainda no primeiro ano. Por isso, considerar um horizonte de tempo de 10 anos é importante para identificar as empresas que realmente se sobressaem da média nacional e ainda estão em atividade por mais de 10 anos.

Um comando de verificação da opção da empresa pelo programa Simples Nacional também foi incluído, fazendo uma análise entra a data de entrada no programa com a data de início das atividades da empresa. Um último comando executado tem como objetivo de remover

as empresas que podem ser classificadas como *outliers*, em relação ao capital social. Os *outliers* são empresas que apresentam um valor de capital social muito discrepante em comparação com a grande maioria, o que pode atrapalhar na análise dos dados.

Figura 10: Variação do valor de Capital Social



Fonte: Autor

A fim de evitar utilizar estes outliers no estudo, foi realizado um novo filtro para considerar apenas 99% das empresas que possuem o valor de capital social mais próximo da média. A partir desta lista, alguns comandos foram executados para manipular os dados da base com o objetivo de criar novos campos que são relevantes para o estudo. Os seguintes passos foram realizados:

- Criação de uma nova medida com o número de empresas ativas que cada sócio possuía ao abrir a empresa;
- Criação de uma nova medida com o número de empresas ativas por CNPJ, considerando todos os sócios;
- Criação de uma nova medida com o número de empresas encerradas que cada sócio possuía ao abrir a empresa;
- Criação de uma nova medida com o número de empresas encerradas por CNPJ, considerando todos os sócios;

- Criação de uma nova medida com o tempo de funcionamento de cada empresa, levando em consideração a data de início e data de encerramento das atividades;
- Criação de uma nova medida com o número de sócios por empresa;
- Criação de duas novas medidas, uma com a faixa etária do sócio mais velho e uma com a faixa etária do sócio mais novo;
- Formatação das medidas com datas, para adequar ao formato correto.

Com a manipulação e a filtragem dos dados realizadas, finalmente tem-se como resultado a lista de variáveis organizadas em um *dataframe*, e que serão utilizadas na aplicação do método definido e na análise dos resultados (tabela 6).

Tabela 6: Variáveis selecionadas para análise

VARIÁVEL	DESCRIÇÃO
CNPJ	Número base de inscrição do CNPJ
Situação cadastral	Ativa ou Encerrada
UF	UF da empresa
Região	Região geográfica da empresa
Capital Social	CAPITAL social da empresa
Opção pelo Simples Nacional	Sim ou Não
Faixa etária sócio mais velho	0 - não se aplica 1 - entre 0 a 12 anos 2 - entre 13 a 20 anos 3 - entre 21 a 30 anos 4 - entre 31 a 40 anos 5 - entre 41 a 50 anos 6 - entre 51 a 60 anos 7 - entre 61 a 70 anos 8 - entre 71 a 80 anos 9 - maiores de 80 anos
Faixa etária sócio mais novo	0 - não se aplica 1 - entre 0 a 12 anos 2 - entre 13 a 20 anos 3 - entre 21 a 30 anos 4 - entre 31 a 40 anos 5 - entre 41 a 50 anos 6 - entre 51 a 60 anos 7 - entre 61 a 70 anos 8 - entre 71 a 80 anos 9 - maiores de 80 anos
Tempo da empresa	Tempo em que a empresa ficou ou está ativa
Número de CNAEs secundários	Número de CNAEs em que a empresa atua além do CNAE principal
Número de sócios	Número de sócios da empresa ao iniciar as atividades
Número de outras empresas dos sócios ativas	Soma do número de empresas que cada sócio possui e que ainda estão ativas
Número de outras empresas dos sócios encerradas	Soma do número de empresas que cada sócio possui e que estão encerradas
Indústria	Setor da indústria correspondente ao CNAE principal da empresa

Fonte: Autor

Como pode ser observada, nem todas as variáveis da tabela 6 são numéricas, tendo algumas que são categóricas. Não é possível utilizar essas variáveis em um método de programação linear, sendo necessário definir as medidas que vão ser usadas no modelo DEA.

Entre as variáveis numéricas, que podem ser utilizadas no modelo, foi feita uma análise dos dados para determinar quais variáveis apresentavam um maior impacto para determinar se a empresa estava aberta ou encerrada. Esta análise, que agrega uma análise qualitativa sobre os dados obtidos, ajuda a determinar um grupo de variáveis que são mais relevantes para o estudo. Essas variáveis são:

- *Inputs*: Capital social, Faixa etária do sócio mais velho e Número de empresas dos sócios encerradas.
- *Outputs*: Número de CNAEs secundários e Número de empresas dos sócios ativas.

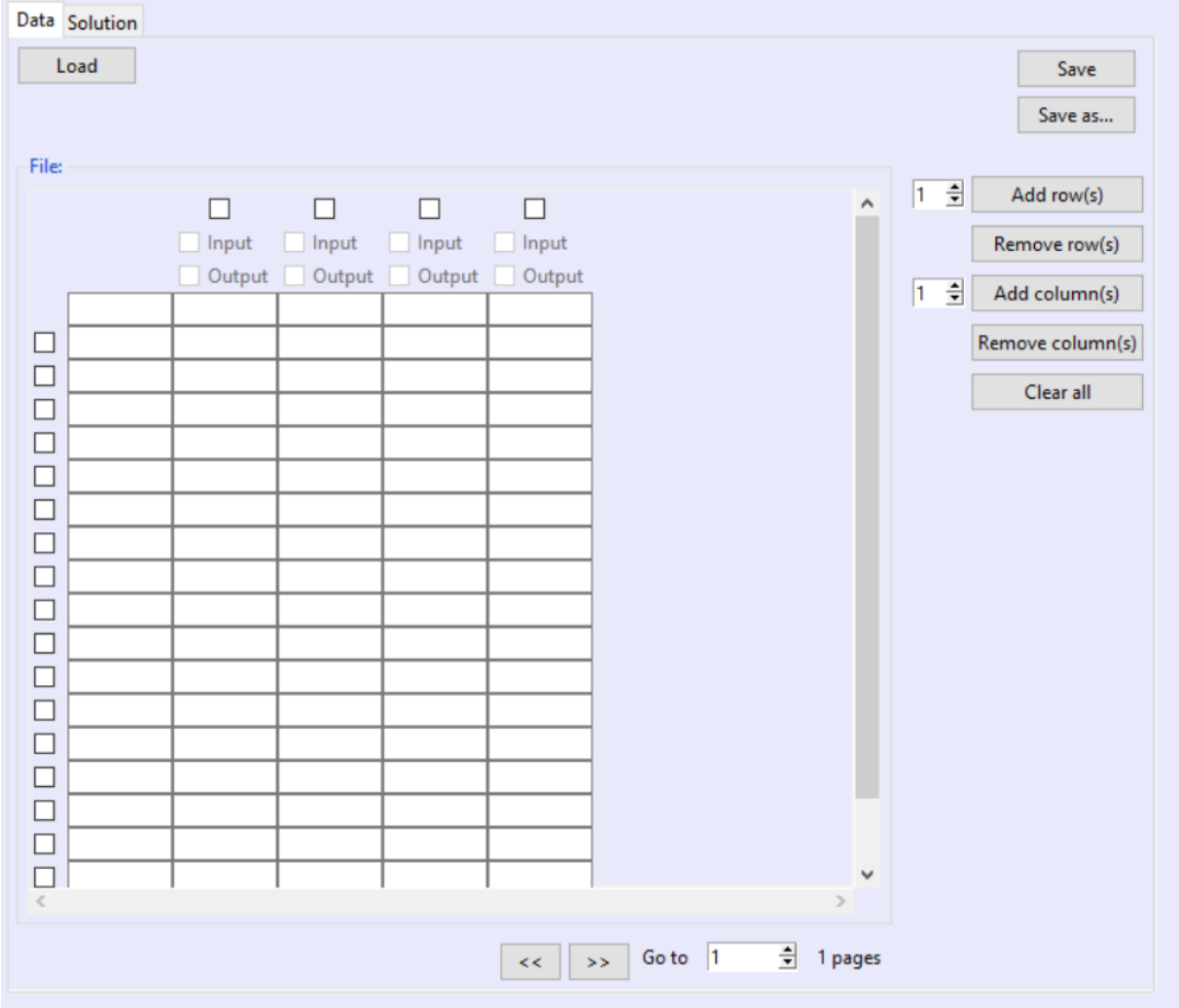
É desejável assim que as empresas tenham um número alto de CNAEs secundários, que seria resultado de uma empresa bem diversificada no mercado e, em geral, uma empresa que tenha sócios bem-sucedidos em outros empreendimentos. Por outro lado, seguindo a ideia de um modelo DEA orientado a *inputs*, é desejável que a empresa tenha boas saídas mesmo que tenha um menor investimento, possua sócios mais novos pensando na continuação do trabalho, além de haver um menor número de empreendimentos mal-sucedidos de seus sócios.

3.5 PyDEA

Desenvolver um modelo de *Data Envelopment Analysis* para o volume de dados presentes neste estudo seria uma tarefa extremamente complexa. Como foi dito no referencial, o uso do DEA era restrito a acadêmicos no estudo de eficiência e gestão de empresas porque era necessário possuir extenso conhecimento de programação linear. O método só começou a ser usado no ambiente corporativo quando novas ferramentas foram desenvolvidas e que fossem capazes de facilitar o uso desta metodologia. Neste projeto, foi usado um pacote de *software* em linguagem *Python* de programação.

PyDEA é um software escrito em linguagem *Python*, desenvolvido no departamento de engenharia da Universidade de Auckland, Nova Zelândia (RAITH; PEREDERIEIEVA; FAUZI; HARTON; LEE; PRIDDEY; ROUSE, 2016), desenhado especificamente para solucionar problemas de programação linear utilizando o método DEA.

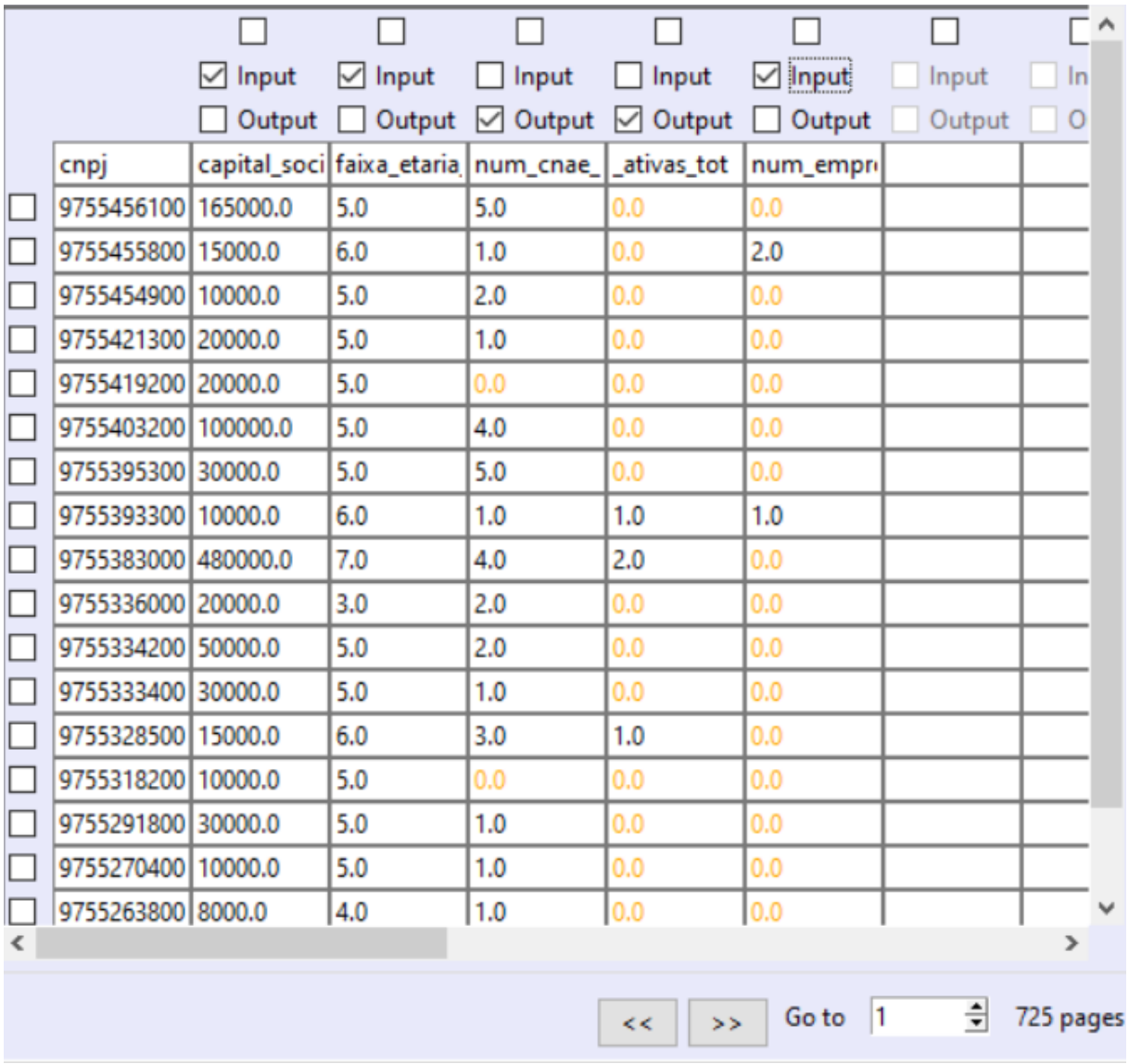
Figura 11: Interface Pydea



Fonte: Autor

O upload de dados para o PyDEA é feito por um arquivo excel, que foi criado após a aplicação dos filtros feitos em linguagem *Python*. Com os dados carregados no programa, são selecionadas quais variáveis serão *inputs* e quais serão *outputs* no modelo.

Figura 12: Inserção de dados no PyDEA

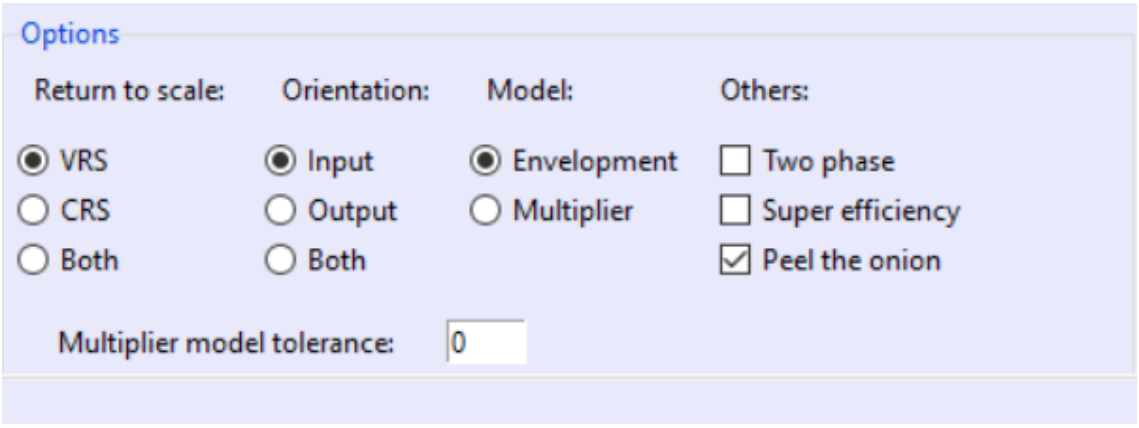


		Input	Input	Input	Input	Input	Input	
		Output	Output	Output	Output	Output	Output	
	cnpj	capital_soci	faixa_etaria	num_cnae_	_ativas_tot	num_empr		
	9755456100	165000.0	5.0	5.0	0.0	0.0		
	9755455800	15000.0	6.0	1.0	0.0	2.0		
	9755454900	10000.0	5.0	2.0	0.0	0.0		
	9755421300	20000.0	5.0	1.0	0.0	0.0		
	9755419200	20000.0	5.0	0.0	0.0	0.0		
	9755403200	100000.0	5.0	4.0	0.0	0.0		
	9755395300	30000.0	5.0	5.0	0.0	0.0		
	9755393300	10000.0	6.0	1.0	1.0	1.0		
	9755383000	480000.0	7.0	4.0	2.0	0.0		
	9755336000	20000.0	3.0	2.0	0.0	0.0		
	9755334200	50000.0	5.0	2.0	0.0	0.0		
	9755333400	30000.0	5.0	1.0	0.0	0.0		
	9755328500	15000.0	6.0	3.0	1.0	0.0		
	9755318200	10000.0	5.0	0.0	0.0	0.0		
	9755291800	30000.0	5.0	1.0	0.0	0.0		
	9755270400	10000.0	5.0	1.0	0.0	0.0		
	9755263800	8000.0	4.0	1.0	0.0	0.0		

Fonte: Autor

O último passo antes de rodar o programa, é definir as configurações que são mais adequadas para o objetivo do projeto. O modelo selecionado é envoltório, seguindo a proposta de uma análise envoltória dos dados e, a opção de escala é VRS (*Variable Returns to Scale*, em inglês) que cria um multiplicador extra para o modelo para garantir que o valor máximo do índice de eficiência seja igual a 1. Além disso, a orientação é para *inputs* e a opção de rodar é o modelo *Peel the Onion*.

Figura 13: Modelos selecionado no PyDEA



Fonte: Autor

O modelo *Peel the Onion* parte do princípio de que o DEA é uma análise comparativa entre os dados de cada DMU do modelo, chegando a um resultado. Neste modelo, o programa realiza os cálculos uma primeira vez, definindo quais são as DMUs com eficiência máxima, igual a 1 e são classificadas como *Tier* 1. A partir disso, o software roda o modelo novamente, porém sem considerar as DMUs que tiveram nota máxima na primeira rodagem, e como resultado, novas DMUs vão apresentar nota máxima e serão classificadas como *Tier* 2. O programa repete este procedimento até classificar todas as DMUs dos dados fornecidos (RAITH; PEREDERIEIEVA; FAUZI; HARTON; LEE; PRIDDEY; ROUSE, 2016). O modelo *Peel the Onion* segue o processo da figura 14.

Figura 14: Modelo matemático do modelo *Peel the Onion*

```
Input: Conjunto de DMUs, inputs, outputs
Output: notas de eficiência, Tier do modelo Peel the Onion
    1. S = conjunto de DMUs
    2. Tier Atual = 1
    3.    While S ≠ ∅
              Resolver DEA para cada DMU em S com conjunto de DMUs em S
              Para cada DMU ∈ S faça
                  se DMU é eficeinte (nota 1) então
                      Anota o número do Tier
                      Remover DMU de S
                  end
              end
              Tier Atual = Tier Atul + 1
         end
```

Fonte: Adaptado de Raith, Perederieieva e Fauzi (2016)

A aplicação deste modelo é interessante porque categoriza as DMUs em grupos, das mais eficientes para as menos eficientes, agregando mais um ponto de análise que pode auxiliar no processo de tomada de decisão. Após completar a execução do programa, os resultados de eficiência e de *Tier* de todas as DMUs são exportados em formato de planilha excel. Assim, um comando em Python é executado para anexar estes resultados ao *dataframe* original, para a análise dos resultados.

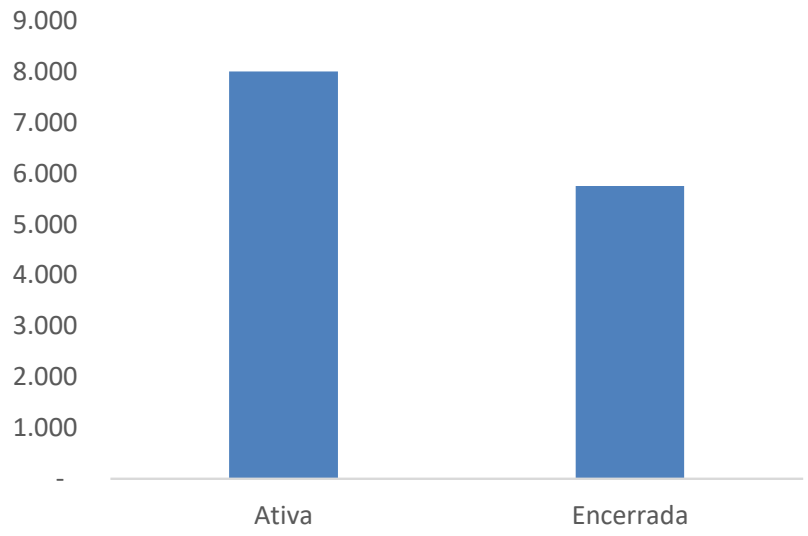
4 RESULTADOS E DISCUSSÕES

Nesta seção do trabalho, serão expostos os principais resultados do método DEA aplicado aos dados obtidos já mencionados no capítulo anterior. Análises e discussões para as soluções obtidas do modelo implementado também serão fornecidas.

4.1 RESULTADOS GERAIS

Após a filtragem dos dados mencionados na seção de Metodologia deste trabalho, foi definida uma lista de 13.769 empresas, sendo 8.009 ativas e 5.760 inativas.

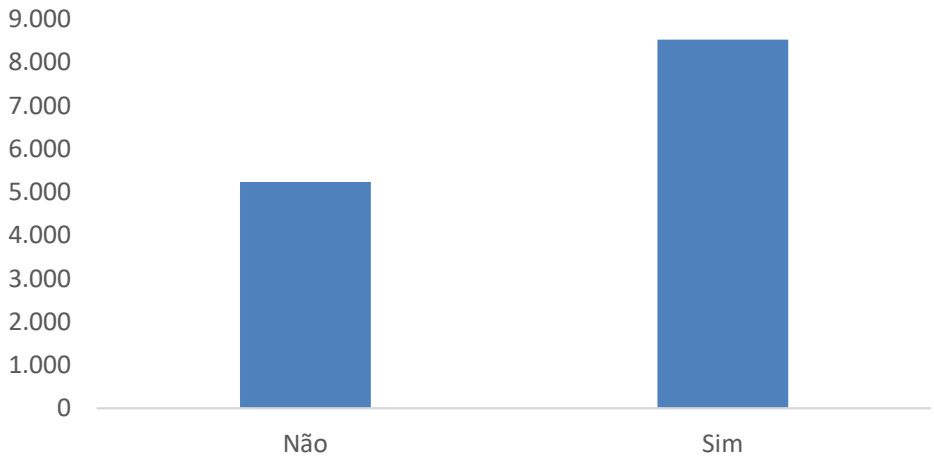
Figura 15: Distribuição das empresas por situação cadastral



Fonte: Autor

Dentre as empresas, 8.528 são optantes pelo Simples Nacional, que representa 61,9% das empresas que fazem parte do estudo. Vale relembrar que a alta carga tributária é um dos principais problemas relatados pelos empresários da Pequena Indústria, e mais da metade das empresas observadas buscam uma solução no programa da Receita Federal.

Figura 16: Distribuição das empresas por opção no programa Simples Nacional



Fonte: Autor

Na introdução deste trabalho, foi comentado sobre a ineficiência do setor industrial brasileiro, principalmente em comparação com os países mais industrializados e desenvolvidos no mundo. O cálculo realizado utilizando o modelo DEA corrobora essa afirmação, porque a média da eficiência como resultado foi de 0,232, sendo que o máximo é de 1,0. Entre as 13.769 empresas analisadas, apenas 26 obtiveram nota 1 e podem ser classificadas como eficientes.

Tabela 7: Distribuição de empresas entre escalas de eficiência

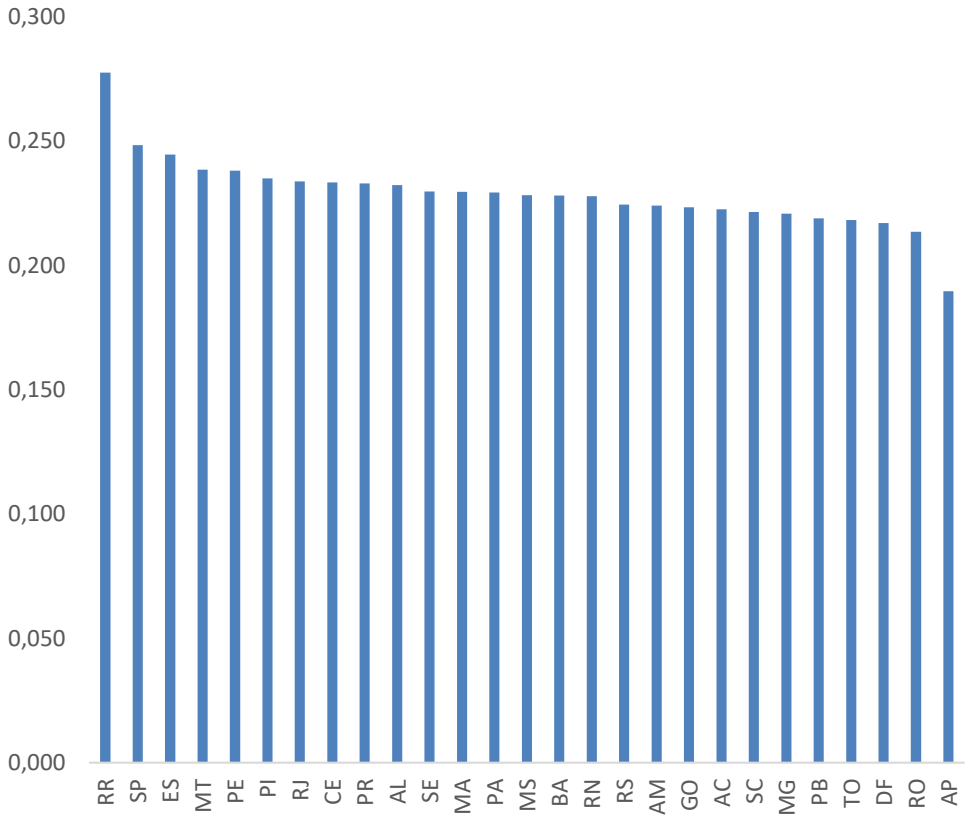
Eficiência	Número de empresas
Entre 1,00 e 0,90	31
Entre 0,89 e 0,80	15
Entre 0,79 e 0,70	28
Entre 0,69 e 0,60	71
Entre 0,59 e 0,50	133
Entre 0,49 e 0,40	343
Entre 0,39 e 0,30	1.230
Entre 0,29 e 0,20	7.188
Entre 0,19 e 0,10	4.730

Fonte: Autor

Como mostrado na tabela 7, a grande maioria das empresas, equivalente a 95%, tiveram um resultado abaixo de 0,40. Este resultado traduz muito bem sobre a ineficiência das micro e pequenas indústrias brasileiras.

Segregando as empresas por UF, é observado que nenhum estado tem a média de eficiência das suas empresas acima de 0,3, sendo o estado de Roraima classificado como o mais eficiente. Contudo, das mais de 13 mil empresas observadas, apenas 8 estão localizadas no estado de Roraima, sendo o estado com menor representação, inferior a 0,1%. Já o estado de São Paulo, o mais industrializado do país e principal economia, ocupa a segunda posição na média de eficiência de 3.316 empresas, representando 24,1% de todas as empresas observadas.

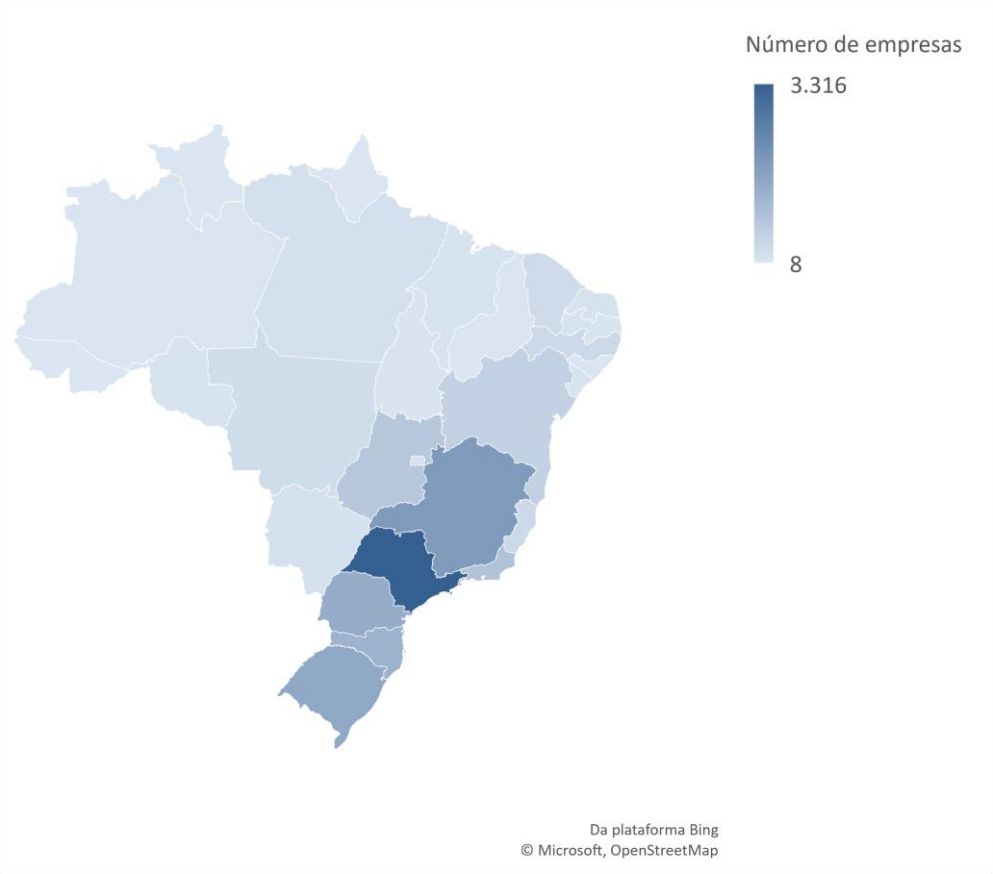
Figura 17: Média de eficiência das empresas por UF



Fonte: Autor

A figura 18 representa a distribuição das empresas entre os estados brasileiros. É nítida a concentração das empresas no sudeste e no sul do país, que somam mais de 10.500 de todas as empresas deste projeto, enquanto a região norte tem uma ocupação muito menor, com muitos estados representando menos de 1% das empresas analisadas.

Figura 18: Distribuição das empresas pelo território brasileiro



Fonte: Autor

Buscando uma análise dos resultados por setor da indústria, a tabela 8 traz um panorama da quantidade de empresas de cada setor que fizeram parte do estudo. Os setores de Artigos do vestuário e acessórios; Manutenção, reparação e instalação de máquinas e equipamentos; Produtos de metal, exceto máquinas e equipamentos; Produtos alimentícios; e Produtos minerais não-metálicos são os setores que possuem mais de 1.000 empresas entre as analisadas. Na outra ponta, Produtos do fumo; Coque, de produtos derivados do petróleo e biocombustíveis; e Produtos farmoquímicos e farmacêuticos são os setores menos representados.

Tabela 8: Número de empresas por setores da indústria

Setor da Indústria	N° de empresas	% do Total
ARTIGOS DO VESTUÁRIO E ACESSÓRIOS	2.408	17,49
MANUTENÇÃO, REPARAÇÃO E INSTALAÇÃO DE MÁQUINAS E EQUIPAMENTOS	1.793	13,02
PRODUTOS DE METAL, EXCETO MÁQUINAS E EQUIPAMENTOS	1.587	11,53
PRODUTOS ALIMENTÍCIOS	1.534	11,14
PRODUTOS DE MINERAIS NÃO-METÁLICOS	1.080	7,84
MÓVEIS	836	6,07
REPRODUÇÃO DE GRAVAÇÕES	784	5,69
PRODUTOS DIVERSOS	549	3,99
COUROS E ARTEFATOS DE COURO, ARTIGOS PARA VIAGEM E CALÇADOS	476	3,46
MÁQUINAS E EQUIPAMENTOS	439	3,19
PRODUTOS DE BORRACHA E DE MATERIAL PLÁSTICO	409	2,97
PRODUTOS DE MADEIRA	393	2,85
PRODUTOS TÊXTEIS	363	2,64
PRODUTOS QUÍMICOS	213	1,55
VEÍCULOS AUTOMOTORES, REBOQUES E CARROCERIAS	209	1,52
MÁQUINAS, APARELHOS E MATERIAIS ELÉTRICOS	159	1,15
CELULOSE, PAPEL E PRODUTOS DE PAPEL	147	1,07
EQUIPAMENTOS DE INFORMÁTICA, PRODUTOS ELETRÔNICOS E ÓPTICOS	147	1,07
BEBIDAS	99	0,72
METALURGIA	65	0,47
OUTROS EQUIPAMENTOS DE TRANSPORTE, EXCETO VEÍCULOS AUTOMOTORES	60	0,44
PRODUTOS FARMOQUÍMICOS E FARMACÊUTICOS	12	0,09
COQUE, DE PRODUTOS DERIVADOS DO PETRÓLEO E DE BIOCOMBUSTÍVEIS	4	0,03
PRODUTOS DO FUMO	3	0,02
TOTAL	**13.769**	**100**

Fonte: Autor

Nas notas de eficiência, os setores de Produtos Químicos e de Máquinas, aparelhos e materiais elétricos são os mais eficientes, com uma média do índice calculado para todas as empresas estando acima de 0,250. Ainda, quatro dos cincos setores mencionados com os maiores números de empresas tiveram a nota média do setor tendendo a média total, de 0,232. Assim, os setores que podem apresentar as maiores capacidades produtivas não estão conseguindo ser os mais eficientes, o que pode ser visto como uma grande oportunidade para melhorar a competitividade das indústrias brasileiras. Com relação aos setores menos representados, Produtos do fumo também foi o setor com a menor média de eficiência, mesmo sendo representado por apenas 3 empresas. O destaque negativo com o maior impacto fica com o setor de Artigos do vestuário e acessórios, o mais representado, mas que fica colocado como o quinto setor menos eficiente.

Tabela 9: Eficiência média das empresas por setor da indústria

Setor da Indústria	Eficiência média
PRODUTOS QUÍMICOS	0,258
MÁQUINAS, APARELHOS E MATERIAIS ELÉTRICOS	0,257
COQUE, DE PRODUTOS DERIVADOS DO PETRÓLEO E DE BIOCOMBUSTÍVEIS	0,249
BEBIDAS	0,243
MÁQUINAS E EQUIPAMENTOS	0,241
CELULOSE, PAPEL E PRODUTOS DE PAPEL	0,240
PRODUTOS DE BORRACHA E DE MATERIAL PLÁSTICO	0,237
PRODUTOS ALIMENTÍCIOS	0,235
PRODUTOS DE METAL, EXCETO MÁQUINAS E EQUIPAMENTOS	0,234
PRODUTOS DE MADEIRA	0,234
PRODUTOS DIVERSOS	0,233
MANUTENÇÃO, REPARAÇÃO E INSTALAÇÃO DE MÁQUINAS E EQUIPAMENTOS	0,233
MÓVEIS	0,232
PRODUTOS DE MINERAIS NÃO-METÁLICOS	0,231
VEÍCULOS AUTOMOTORES, REBOQUES E CARROCERIAS	0,231
PRODUTOS TÊXTEIS	0,231
COUROS E ARTEFATOS DE COURO, ARTIGOS PARA VIAGEM E CALÇADOS	0,230
REPRODUÇÃO DE GRAVAÇÕES	0,229
EQUIPAMENTOS DE INFORMÁTICA, PRODUTOS ELETRÔNICOS E ÓPTICOS	0,227
ARTIGOS DO VESTUÁRIO E ACESSÓRIOS	0,225
PRODUTOS FARMOQUÍMICOS E FARMACÊUTICOS	0,224
METALURGIA	0,224
OUTROS EQUIPAMENTOS DE TRANSPORTE, EXCETO VEÍCULOS AUTOMOTORES	0,219
PRODUTOS DO FUMO	0,191
TOTAL	**0,232**

Fonte: Autor

Olhando para as categorias, os *Tiers* definidos usando o modelo *Peel the Onion*, são observadas as mesmas 26 empresas que obtiveram a nota máxima de eficiência ocupando o *Tier* 1. Interessante observar para o *Tier* 2, pois outras 60 empresas seriam consideradas as mais eficientes, caso as empresas da primeira categoria não fizessem parte do estudo. Dessa forma, estas 60 empresas, seguindo a metodologia do modelo *Peel the Onion*, podem ser classificadas como próximas de serem eficientes, seguidas pelas mais 373 empresas que foram classificadas no *Tier* 3. É possível observar também a concentração de empresas nos *Tiers* 4, 5, 6 e um pouco menos na 7, que são as empresas que ficaram mais próximas da média de 0,23, somando um total de 12.462 empresas nestas categorias médias. Por fim estão as empresas nas *Tiers* 8 e 9, as menos eficientes de todas.

Tabela 10: Número de empresas por *Tier*

Tier	Número de empresas
1	26
2	60
3	373
4	2.804
5	4.234
6	3.488
7	1.936
8	667
9	181

Fonte: Autor

Após olhar para os resultados tendo uma visão geral, é importante explorar um pouco mais as informações olhando por diferentes pontos de vista. Entender, por exemplo, onde estão localizadas as empresas mais eficientes e em quais setores elas atuam. É interessante também analisar as empresas que foram identificadas como as mais ineficientes mesmo tendo características iniciais que poderiam apontar para um resultado melhor.

4.2 EMPRESAS MAIS EFICIENTES

O primeiro ponto de vista a ser observado é o das empresas mais eficientes. Como mostrado anteriormente, apenas 26 empresas, em um total de 13.769, apresentaram nota máxima de eficiência. Como apresentado na tabela 11, essas empresas estão divididas entre 8 estados, sendo quase todas nas regiões sul, sudeste e nordeste do Brasil, além de apenas uma empresa no centro-oeste e nenhuma no norte do país.

Tabela 11: Número de empresas com eficiência máxima por UF

UF	N° de empresas
SC	7
SP	5
RS	4
BA	3
MG	3
RJ	2
CE	1
MT	1

Fonte: Autor

Entre os setores da indústria, o setor de Artigos do vestuário e acessórios é um dos que mais possui empresas consideradas eficientes. Lembrando que este setor também é o que totaliza o maior número de empresas no estudo e um dos que tiveram a menor média. Evidentemente, cinco empresas entre mais de 2.400 não são suficientes para fazer a média subir significativamente, o que leva a entender que a maioria das empresas no setor estão um pouco abaixo da média geral. O outro setor com mais representantes com nota máxima é de Couros e artefatos de couro, artigos para viagem e calçados. Este setor teve uma média de 0,230, muito próximo da média geral, e um total de 476 empresas, o que o deixa com uma proporção bem melhor de empresas eficientes em comparação com o citado anteriormente. Os outros setores que tiveram mais de uma empresa com nota máxima de eficiência foram Produtos de metal, exceto máquinas e equipamentos; Manutenção, reparação e instalação de máquinas e equipamentos e Máquinas e equipamentos.

Tabela 12: Número de empresas com eficiência máxima por setor da indústria

Setor da Indústria	N° de empresas
ARTIGOS DO VESTUÁRIO E ACESSÓRIOS	5
COUROS E ARTEFATOS DE COURO, ARTIGOS PARA VIAGEM E CALÇADOS	5
PRODUTOS DE METAL, EXCETO MÁQUINAS E EQUIPAMENTOS	4
MANUTENÇÃO, REPARAÇÃO E INSTALAÇÃO DE MÁQUINAS E EQUIPAMENTOS	3
MÁQUINAS E EQUIPAMENTOS	3
MÁQUINAS, APARELHOS E MATERIAIS ELÉTRICOS	1
PRODUTOS ALIMENTÍCIOS	1
PRODUTOS DE BORRACHA E DE MATERIAL PLÁSTICO	1
PRODUTOS DE MINERAIS NÃO-METÁLICOS	1
PRODUTOS DIVERSOS	1
PRODUTOS QUÍMICOS	1

Fonte: Autor

Visando observar uma parcela maior de empresas que estavam entre as mais bem classificadas, para entender melhor sobre a distribuição no território nacional e entre atividades industriais, foi necessário expandir um pouco a seleção. Com isso, decidiu-se selecionar empresas com nota de eficiência bem acima da média, mais especificamente, acima da média mais o desvio do padrão dos valores calculados. O desvio padrão mede o grau de dispersão de uma amostra de dados, o que pode ser entendido como quanto os dados se desviam da média. Neste caso, ao selecionar as empresas com notas maiores do que a soma entre a média e o desvio padrão, serão selecionadas certamente empresas com eficiência acima da média, comparativamente com todas as empresas observadas. O desvio padrão á calculado segundo a fórmula:

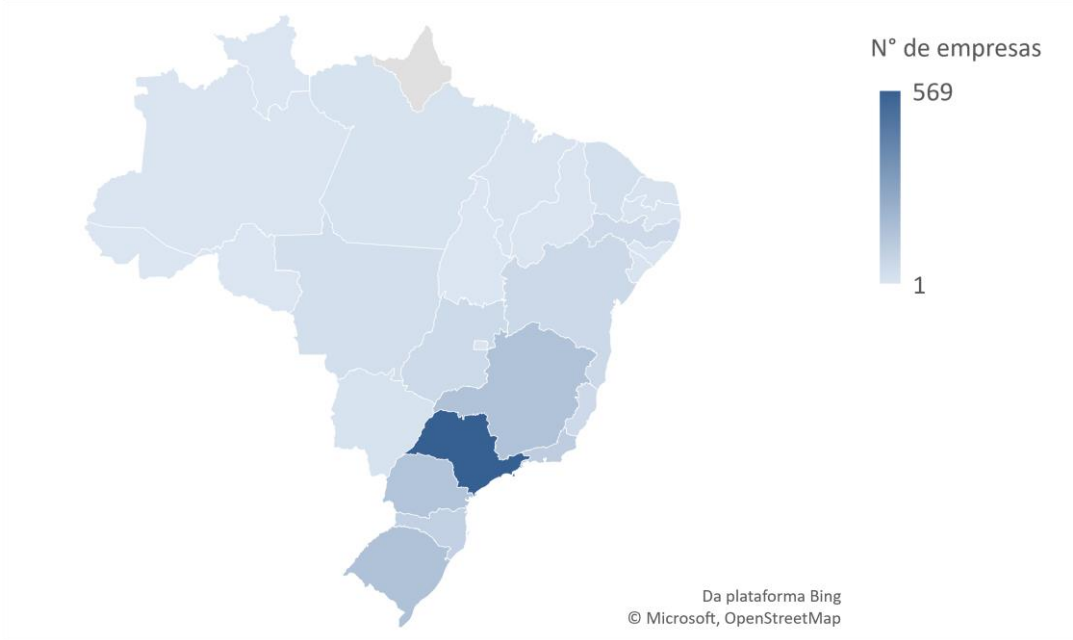
$$s = \sqrt{\frac{\sum \left(x_i - \bar{x}\right)^2}{n-1}}$$

(Eq. 2)

Seguindo este cálculo, a nota corte foi de 0,324, sendo selecionadas apenas as empresas com eficiência maior do que este valor.

Esta nova lista de empresas selecionadas está dividida entre os estados segundo a figura 20. Para esta seleção, a concentração de empresas é ainda mais favorável para as regiões sul e sudeste em comparação com a base original. Portanto, São Paulo, Rio Grande do Sul, Minas Gerais, Paraná, Rio de Janeiro e Santa Catarina são os estados com maior número de empresas, somando 1.206 de 1.567, que representa 77% do total. São Paulo sozinho possui 569 dessas empresas, e Amapá foi o único estado sem nenhuma empresa entre as mais eficientes.

Figura 19: Distribuição das empresas mais eficientes pelo território brasileiro



Fonte: Autor

Na média das notas, Roraima novamente foi o estado que apresentou melhor média, com média de 0,648, incluindo, entretanto, apenas uma empresa e este seria um resultado muito acima dos demais estados. Entretanto, o fato de ter apenas uma empresa com eficiência muito acima da média não é suficiente para classificar o estado entre os mais eficientes. Dentre os estados com participação no número de empresas maior que 5%, o melhor classificado é o de Santa Catarina, com média de 0,450 entre 91 empresas.

Tabela 13: Eficiência média das empresas mais eficientes por UF

UF	Eficiência média
RR	0,648
RO	0,497
PI	0,457
MT	0,456
AL	0,453
SC	0,450
GO	0,446
PR	0,444
ES	0,442
SE	0,441
BA	0,430
RJ	0,424
MA	0,420
TO	0,417
SP	0,414
PE	0,413
CE	0,412
PA	0,412
MG	0,409
RS	0,408
MS	0,396
DF	0,388
RN	0,370
PB	0,365
AC	0,361
AM	0,331

Fonte: Autor

Olhando para os setores da indústria, novamente Artigos do vestuário e acessórios é o setor com maior número de empresas, seguido por Manutenção, reparação e instalação de máquinas e equipamentos; Produtos alimentícios; Produtos de metal, exceto máquinas e equipamentos; e Produtos de minerais não metálicos. Estes 5 setores são exatamente os mesmos 5 setores com mais empresas na base original, o que leva a entender que mesmo não estando entre as melhores médias entre todos os setores, ainda possuem uma boa quantidade de empresas entre as mais eficientes.

Tabela 14: Número de empresas mais eficientes por setor da indústria

Setor da Indústria	N° de empresas
ARTIGOS DO VESTUÁRIO E ACESSÓRIOS	233
MANUTENÇÃO, REPARAÇÃO E INSTALAÇÃO DE MÁQUINAS E EQUIPAMENTOS	199
PRODUTOS ALIMENTÍCIOS	182
PRODUTOS DE METAL, EXCETO MÁQUINAS E EQUIPAMENTOS	179
PRODUTOS DE MINERAIS NÃO-METÁLICOS	119
REPRODUÇÃO DE GRAVAÇÕES	84
MÓVEIS	82
PRODUTOS DIVERSOS	69
MÁQUINAS E EQUIPAMENTOS	66
PRODUTOS DE BORRACHA E DE MATERIAL PLÁSTICO	56
PRODUTOS DE MADEIRA	53
COUROS E ARTEFATOS DE COURO, ARTIGOS PARA VIAGEM E CALÇADOS	52
PRODUTOS TÊXTEIS	39
PRODUTOS QUÍMICOS	37
MÁQUINAS, APARELHOS E MATERIAIS ELÉTRICOS	29
VEÍCULOS AUTOMOTORES, REBOQUES E CARROCERIAS	23
CELULOSE, PAPEL E PRODUTOS DE PAPEL	20
EQUIPAMENTOS DE INFORMÁTICA, PRODUTOS ELETRÔNICOS E ÓPTICOS	18
BEBIDAS	14
METALURGIA	7
OUTROS EQUIPAMENTOS DE TRANSPORTE, EXCETO VEÍCULOS AUTOMOTORES	3
PRODUTOS FARMOQUÍMICOS E FARMACÊUTICOS	2
COQUE, DE PRODUTOS DERIVADOS DO PETRÓLEO E DE BIOCOMBUSTÍVEIS	1

Fonte: Autor

Novamente, nenhum dos cinco setores com maior número de empresas fica entre as melhores médias. Máquinas, aparelhos e materiais elétricos e Produtos químicos, assim como na base original, estão entre as melhores médias. A principal diferença na parte de cima é o setor de Couros e artefatos de couro, artigos para viagem e calçados, um dos setores com maior número de empresas com eficiência igual a 1, mas que na seleção na relação de mais de 13 mil empresas, tinha a oitava pior média. Isso leva a entender que este setor possui uma relação boa de empresas entre as mais eficientes ao mesmo tempo que tem muitas empresas com eficiência muito baixa, o que faz a média do setor ser baixa na base original.

Tabela 15: Eficiência média das empresas mais eficientes por setor da indústria

Setor da Indústria	Eficiência média
MÁQUINAS, APARELHOS E MATERIAIS ELÉTRICOS	0,470
COUROS E ARTEFATOS DE COURO, ARTIGOS PARA VIAGEM E CALÇADOS	0,463
PRODUTOS QUÍMICOS	0,455
OUTROS EQUIPAMENTOS DE TRANSPORTE, EXCETO VEÍCULOS AUTOMOTORES	0,449
CELULOSE, PAPEL E PRODUTOS DE PAPEL	0,435
MÁQUINAS E EQUIPAMENTOS	0,429
PRODUTOS TÊXTEIS	0,427
PRODUTOS DE MADEIRA	0,427
PRODUTOS DE METAL, EXCETO MÁQUINAS E EQUIPAMENTOS	0,427
PRODUTOS ALIMENTÍCIOS	0,426
MÓVEIS	0,425
VEÍCULOS AUTOMOTORES, REBOQUES E CARROCERIAS	0,421
COQUE, DE PRODUTOS DERIVADOS DO PETRÓLEO E DE BIOCOMBUSTÍVEIS	0,418
PRODUTOS DE BORRACHA E DE MATERIAL PLÁSTICO	0,417
MANUTENÇÃO, REPARAÇÃO E INSTALAÇÃO DE MÁQUINAS E EQUIPAMENTOS	0,416
PRODUTOS DIVERSOS	0,416
ARTIGOS DO VESTUÁRIO E ACESSÓRIOS	0,415
BEBIDAS	0,414
PRODUTOS DE MINERAIS NÃO-METÁLICOS	0,411
METALURGIA	0,402
REPRODUÇÃO DE GRAVAÇÕES	0,390
EQUIPAMENTOS DE INFORMÁTICA, PRODUTOS ELETRÔNICOS E ÓPTICOS	0,382
PRODUTOS FARMOQUÍMICOS E FARMACÊUTICOS	0,332

Fonte: Autor

Muito importante agora analisar empresas com baixa eficiência. Portanto, analisa-se a seguir empresas com eficiência abaixo da média, mas que possuem características favoráveis e poderiam aumentar os seus níveis de eficiência.

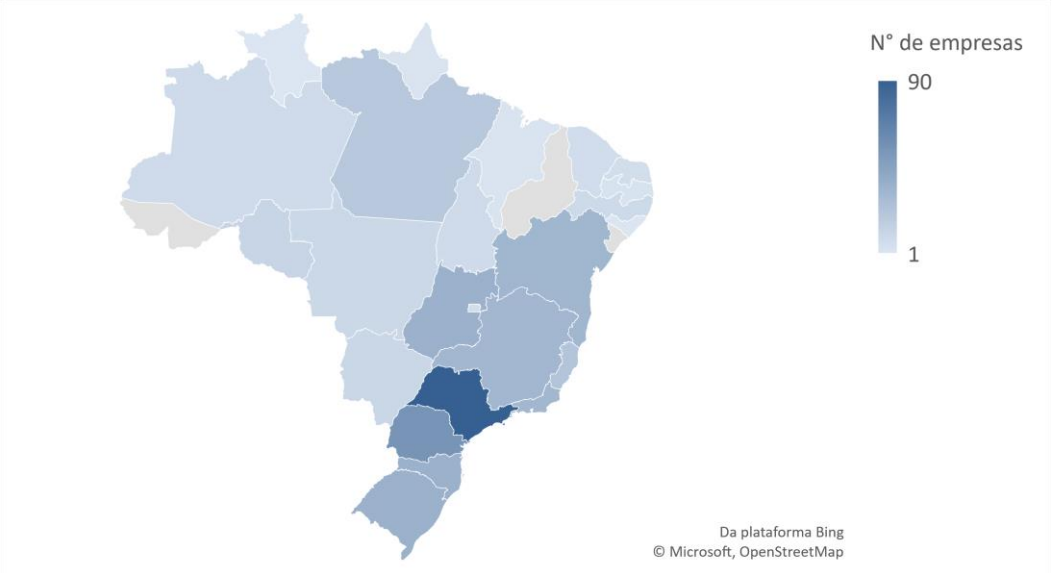
4.3 EMPRESAS INEFICIENTES COM ALTO CAPITAL SOCIAL

O capital social é descrito como o valor investido pelos sócios em uma empresa. Quanto maior este valor investido, mais recursos a empresa tem para trabalhar e, assim, mais chances de ser bem-sucedida. Contudo, foi possível observar que algumas empresas tinham um capital social muito acima da média, seguindo a mesma lógica anterior da soma da média com o desvio padrão, e ainda foram classificadas com eficiência abaixo da média.

É interessante identificar estas empresas porque podem ser vistas, mesmo sendo ineficientes, com potencial de crescimento e desenvolvimento. As empresas que seguem esses

requisitos, dentre todas observadas neste estudo, somam um total de 479, distribuídas no território nacional segundo a figura 20. A concentração destas empresas está localizada nas regiões sul e sudeste, com São Paulo tendo a maior quantidade, um total de 90 empresas. Contudo, os estados de Goiás e Bahia também estão entre os mais numerosos, sendo terceiro e sexto, respectivamente. Seria interessante analisar as empresas destes estados para ajudar na diversificação e desenvolvimento de outras regiões do país.

Figura 20: Distribuição de empresas ineficientes com alto capital social pelo território brasileiro



Fonte: Autor

Observando para os setores dessas empresas, não há muitas mudanças entre os mais numerosos. Os mesmos cinco setores, apesar de em ordem diferente, que tinham mais empresas na base original, são os com maiores empresas que podem ser observadas.

Contudo, para este caso, vale a pena dar mais atenção para os setores que não tinham muitas empresas entre as mais eficientes, como o setor de Outros equipamentos de transporte, exceto veículos automotores; Equipamentos de informática, produtos eletrônicos e ópticos; e Bebidas. Estudar para diferentes setores daqueles que já estão entre os principais fortalece a diversificação da indústria e ajuda em um desenvolvimento mais saudável, menos dependente de poucas atividades.

Tabela 16: Número de empresas ineficientes com alto capital social por setor da indústria

Setor	N° de empresas
PRODUTOS DE MINERAIS NÃO-METÁLICOS	70
PRODUTOS ALIMENTÍCIOS	57
MANUTENÇÃO, REPARAÇÃO E INSTALAÇÃO DE MÁQUINAS E EQUIPAMENTOS	52
PRODUTOS DE METAL, EXCETO MÁQUINAS E EQUIPAMENTOS	50
ARTIGOS DO VESTUÁRIO E ACESSÓRIOS	42
MÁQUINAS E EQUIPAMENTOS	30
MÓVEIS	26
PRODUTOS DE BORRACHA E DE MATERIAL PLÁSTICO	26
PRODUTOS DE MADEIRA	23
REPRODUÇÃO DE GRAVAÇÕES	19
MÁQUINAS, APARELHOS E MATERIAIS ELÉTRICOS	13
PRODUTOS DIVERSOS	12
BEBIDAS	11
PRODUTOS QUÍMICOS	10
PRODUTOS TÊXTEIS	10
COUROS E ARTEFATOS DE COURO, ARTIGOS PARA VIAGEM E CALÇADOS	7
VEÍCULOS AUTOMOTORES, REBOQUES E CARROCERIAS	7
CELULOSE, PAPEL E PRODUTOS DE PAPEL	6
EQUIPAMENTOS DE INFORMÁTICA, PRODUTOS ELETRÔNICOS E ÓPTICOS	4
OUTROS EQUIPAMENTOS DE TRANSPORTE, EXCETO VEÍCULOS AUTOMOTORES	4

Fonte: Autor

A questão de diversificação da indústria também pode ser aplicada para cada umas das empresas individualmente, e é muito por conta disso que diversas empresas possuem não só um código CNAE principal, mas alguns CNAEs secundários também.

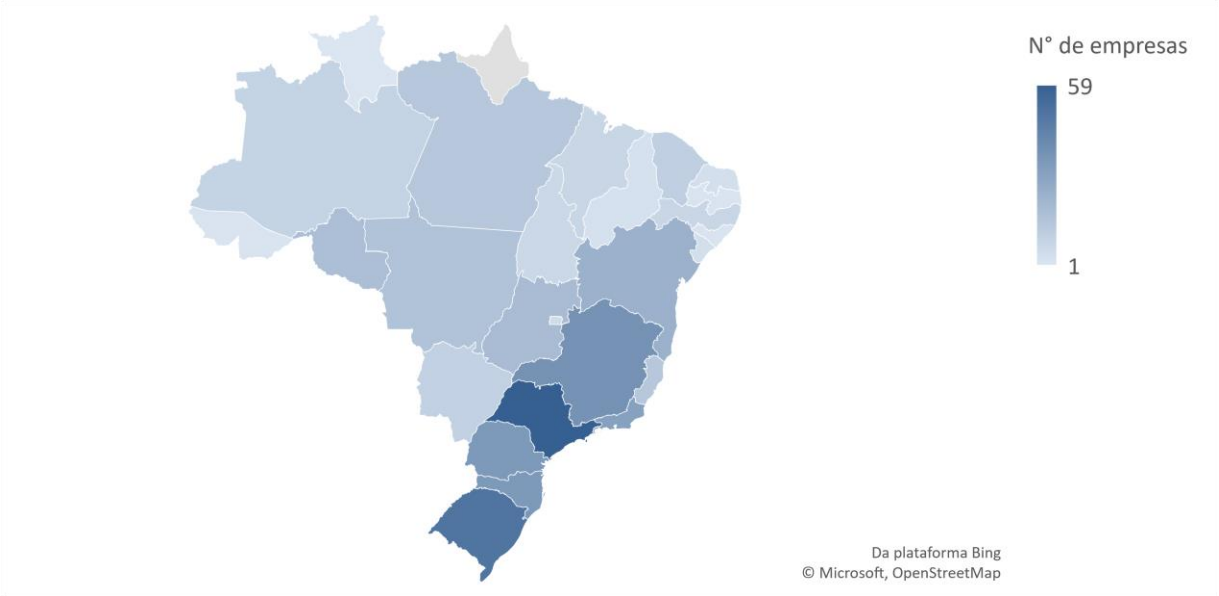
4.4 EMPRESAS INEFICIENTES BEM DIVERSIFICADAS

Neste ponto de vista, serão observadas as empresas que podem ser classificadas como mais diversificadas do que a média das empresas que estavam no modelo de programação linear aplicado. Para isso, serão considerados o número de CNAEs secundários, ou seja, as atividades em que uma empresa atua, mas que não são a sua atividade principal. Seguindo a mesma lógica

da seção anterior, foram selecionadas as 426 empresas que possuem o número de CNAEs secundários maior do que a soma da média geral mais o desvio padrão, além de possuírem a nota de eficiência abaixo da média. O número de corte deste cálculo são 8 atividades secundárias.

Estas empresas estão distribuídas segundo a figura 21. Mais uma vez, Amapá foi o único estado sem nenhum representante e os estados de São Paulo, Rio Grande do Sul, Minas Gerais, Paraná, Santa Catarina e Rio de Janeiro possuem mais de 30 empresas com eficiência abaixo da média e com o número de CNAEs secundários acima da média.

Figura 21: Distribuição de empresas ineficientes bem diversificadas pelo território brasileiro



Fonte: Autor

Já entre os setores da indústria, um novo setor aparece entre os mais representados. Reproduções de gravações é o terceiro, com 45 empresas bem diversificadas, mas ineficientes. Pensando nos setores com poucos representantes entre os mais eficientes, as empresas Metalurgia e Produtos Farmoquímicos e Farmacêuticos se destacam.

Tabela 17: Número de empresas ineficientes bem diversificadas por setor da indústria

Setor da indústria	N° de empresas
MANUTENÇÃO, REPARAÇÃO E INSTALAÇÃO DE MÁQUINAS E EQUIPAMENTOS	75
PRODUTOS DE METAL, EXCETO MÁQUINAS E EQUIPAMENTOS	60
REPRODUÇÃO DE GRAVAÇÕES	45
ARTIGOS DO VESTUÁRIO E ACESSÓRIOS	42
PRODUTOS DE MINERAIS NÃO-METÁLICOS	28
MÓVEIS	25
PRODUTOS ALIMENTÍCIOS	25
MÁQUINAS E EQUIPAMENTOS	17
MÁQUINAS, APARELHOS E MATERIAIS ELÉTRICOS	17
EQUIPAMENTOS DE INFORMÁTICA, PRODUTOS ELETRÔNICOS E ÓPTICOS	14
PRODUTOS DE MADEIRA	14
PRODUTOS DIVERSOS	14
PRODUTOS DE BORRACHA E DE MATERIAL PLÁSTICO	10
COUROS E ARTEFATOS DE COURO, ARTIGOS PARA VIAGEM E CALÇADOS	7
PRODUTOS QUÍMICOS	7
VEÍCULOS AUTOMOTORES, REBOQUES E CARROCERIAS	7
PRODUTOS TÊXTEIS	6
CELULOSE, PAPEL E PRODUTOS DE PAPEL	5
BEBIDAS	3
OUTROS EQUIPAMENTOS DE TRANSPORTE, EXCETO VEÍCULOS AUTOMOTORES	3
METALURGIA	1
PRODUTOS FARMOQUÍMICOS E FARMACÊUTICOS	1

Fonte: Autor

Por fim, imaginar empresas que possuem a combinação das características comentadas é pensar naquelas de maior potencial para crescimento e desenvolvimento, caso consigam se tornar mais eficientes.

4.5 EMPRESAS COM MELHOR POTENCIAL PARA OBSERVAÇÃO

No último ponto de vista, foram selecionadas empresas com capital social e número de CNAEs secundários muito acimas da média, enquanto tiveram um resultado de eficiência abaixo da média. Combinando estes requisitos, foi definida uma lista de 104 empresas, distribuídas pelo país segundo a figura 22. Desta vez, foi possível observar uma distribuição melhor entre as regiões, com São Paulo novamente tendo o maior número de empresas, mas seguido por Pará, Santa Catarina, Bahia, Rio Grande do Sul e Amazonas. Dois estados sozinhos da região norte com 16 empresas somadas dentre 104 observadas. O estado da região centro-oeste com maior número de empresas é Goiás, com 5.

Figura 22: Distribuição de empresas com melhor potencial para observação pelo território brasileiro

Fonte: Autor

Nos setores, Manutenção, reparação e instalação de máquinas e equipamentos; Produtos de metal, exceto máquinas e equipamentos; e Produtos de mineiras não metálicos são os setores com maiores quantidades de empresas. Lembrando que na base original com mais de 13 mil empresas havia 24 setores, enquanto nesta lista reduzida, ainda são 18 setores com pelo menos uma empresa com potencial muito bom. Pode-se concluir que quase todos os setores da indústria possuem empresas que ainda podem ser bem-sucedidas mesmo sendo ineficientes em um primeiro momento.

Tabela 18: Número de empresas com melhor potencial para observação por setor da indústria

Setor da indústria	N° de empresas
MANUTENÇÃO, REPARAÇÃO E INSTALAÇÃO DE MÁQUINAS E EQUIPAMENTOS	22
PRODUTOS DE METAL, EXCETO MÁQUINAS E EQUIPAMENTOS	14
PRODUTOS DE MINERAIS NÃO-METÁLICOS	12
ARTIGOS DO VESTUÁRIO E ACESSÓRIOS	9
MÁQUINAS E EQUIPAMENTOS	7
REPRODUÇÃO DE GRAVAÇÕES	7
PRODUTOS ALIMENTÍCIOS	6
MÓVEIS	5
MÁQUINAS, APARELHOS E MATERIAIS ELÉTRICOS	4
PRODUTOS DE BORRACHA E DE MATERIAL PLÁSTICO	4
COUROS E ARTEFATOS DE COURO, ARTIGOS PARA VIAGEM E CALÇADOS	3
CELULOSE, PAPEL E PRODUTOS DE PAPEL	2
EQUIPAMENTOS DE INFORMÁTICA, PRODUTOS ELETRÔNICOS E ÓPTICOS	2
OUTROS EQUIPAMENTOS DE TRANSPORTE, EXCETO VEÍCULOS AUTOMOTORES	2
PRODUTOS DE MADEIRA	2
BEBIDAS	1
PRODUTOS QUÍMICOS	1
PRODUTOS TÊXTEIS	1

Fonte: Autor

Os resultados apresentados nas seções (4.1 a 4.5) podem ser resumidas do seguinte modo:

(i) Nos resultados gerais, considerando a base completa com 13.769 empresas, a eficiência média das empresas foi relativamente baixa, de 0,232. Há uma concentração de empresas nas regiões sudeste e sul do país, mas o estado de Roraima é o estado com a maior média. Entre os setores da indústria, o setor de Produtos Químicos e de Máquinas, Aparelhos e Materiais Elétricos são os setores com maiores médias de eficiência das empresas;

(ii) Entre as empresas com eficiência máxima, os estados de Santa Catarina, São Paulo e Rio grande do Sul são os estados com maior número de empresas, e os setores de

Artigos do Vestuário e Acessórios e de Couros e Artefatos de Couro, Artigos para Viagem e Calçados sãos os que mais possuem empresas;

(iii) Nas empresas ineficientes, mas com alto capital social, as empresas estão concentradas nas regiões sudeste e sul do país e nos setores de Produtos de Minerais Não Metálicos, de Produtos Alimentícios, de Manutenção, Reparação e Instalação de Máquinas e Equipamentos e de Produtos de Metal, exceto Máquinas e Equipamentos;

(iv) Em empresas ineficientes bem diversificadas, as empresas estão concentradas em seis diferentes estados das regiões sul e sudeste e os setores com mais empresas são os de Manutenção, Reparação e Instalação de Máquinas e Equipamentos e de Produtos de Metal, exceto Máquinas e Equipamentos;

(v) Por último, as empresas com maior potencial para observação, aquelas com eficiência baixa, mas com alto capital social e bem diversificadas, estão distribuídas entre todas as regiões, sendo os estados de São Paulo, Pará, Santa Catarina, Bahia, Rio grande do Sul e Amazonas os estados com mais empresas. Entre os setores da indústria, os setores de Manutenção, Reparação e Instalação de Máquinas e Equipamentos, de Produtos de Metal, exceto Máquinas e Equipamentos e de Produtos de Mineiras não Metálicos possuem as maiores quantidades de empresas.

5 CONCLUSÃO E CONSIDERAÇÕES FINAIS

A queda no ritmo de crescimento da eficiência e produtividade que muitos países têm apresentado pode ser vista como uma causa na desaceleração também do ritmo de desenvolvimento do país. O Brasil apresenta uma queda ainda maior desse ritmo em comparação com a média global e de concorrentes, e o setor industrial é uma das origens da ineficiência que pode atrasar no crescimento econômico brasileiro. As micro e pequenas empresas da indústria da transformação, um grupo de extrema importância e participação na economia brasileira, são o objeto de estudo deste trabalho para analisar a eficiência do setor no Brasil.

Este projeto propôs (i) definir uma metodologia de cálculo de eficiência; (ii) definir variáveis relevantes para uma análise da eficiência a partir da base de dados extraída; (iii) aplicar o método com os dados obtidos de micro e pequenas empresas da indústria de transformação; (iv) analisar e discutir os resultados de eficiência encontrados; (v) identificar quais são as industrias mais eficientes e como são caracterizadas; e (vi) identificar empresas que são ineficientes mas possuem potencial para reverter a situação atual.

Quanto ao primeiro ponto, foi definida o modelo DEA de programação linear como o ideal para realizar cálculos de eficiência com os dados disponíveis, assim como já vai sendo aplicado com essa finalidade no mundo corporativo em alto nível. Foi possível também determinar quais eram as variáveis que seriam consideradas desejáveis e quais seriam indesejáveis para cada empresa, informações importantes para sustentar no que se baseava a eficiência que era calculada.

Aplicado o modelo definido aos dados encontrados e utilizando as variáveis mais relevantes para o estudo, foi possível chegar a resultados de eficiência para as empresas observadas. Na análise destes resultados, foi confirmada a ideia de que os setores da indústria brasileira realmente poderiam ser classificados como ineficientes, ao apresentar uma média de eficiência baixa. A quantidade de empresas que conseguiam se destacar em relação às demais, ao ponto de apresentar eficiência máxima, correspondia a uma parcela muito pequena do total analisado.

A partir dos resultados, analisou-se ainda as empresas mais ineficientes. Observando as empresas que estão apresentando uma produtividade abaixo do esperado e, não apenas considerando como elas estão distribuídas geograficamente, mas também em quais atividades

estão atuando, auxilia o entendimento sobre as áreas que podem ser responsáveis pela queda do ritmo de crescimento do setor.

Analisar essas empresas, identificando principalmente aquelas com potencial de melhoria e crescimento, pode ser um começo para reverter a situação e fazer o setor industrial voltar ao ritmo de desenvolvimento. Alguns caminhos para chegar a esse objetivo seriam uma qualificação da gestão e no desenvolvimento de uma estratégia para essas empresas.

Durante o desenvolvimento do trabalho ficou evidente a importância e a aplicação de ferramentas, de programação neste caso, para resolver problemas de forma rápida e simplificada. Como foi comentado neste trabalho, o modelo DEA passou décadas ainda sendo utilizado apenas no mundo acadêmico porque exigia conhecimento extensivo de programação linear, sendo de difícil aplicação no ambiente corporativo, por exemplo. Após desenvolvimento e ferramentas de programação, o modelo se tornou mais acessível e versátil, sendo este trabalho um exemplo disso.

Em complemento ao estudo realizado, realizar uma análise de resultados de eficiência utilizando variáveis diferentes, capazes de retratar de forma mais aprofundada o funcionamento de uma empresa auxilia as análises realizadas neste trabalho. Dados que conseguem traduzir de forma precisa o desempenho de cada empresa, como volume produzido, receita, número de empregados, podem ser utilizados em conjunto com os apresentados para qualificar ainda mais os resultados.

Em conclusão, o modelo DEA de programação linear mostrou-se uma alternativa viável para realizar um estudo de eficiência de empresas, que pode ser aplicado em outros ambientes, gerando resultados para embasar análises aprofundadas no assunto e que servem como o ponto de início para propostas de melhoria da produtividade das indústrias brasileiras.

REFERÊNCIAS BIBLIOGRÁFICAS

BERTRAND, J. W. M.; FRANSOO, J. C., **Operations management research methodologies using quantitative modeling**. International Journal of Operations & Production Management, 2002

CHARNES, A.; COOPER, W.W.; RHODES, E., **Measuring the efficiencey of decision making units**. European Journal of Operational Research, 2, 1978

CNN BRASIL, **Em 13° entre maiores economias, PIB do Brasil fica abaixo da média global**. Disponível em <https://www.cnnbrasil.com.br/business/em-13o-entre-maiores-economias-pib-do-brasil-fica-abaixo-de-media-global/>

CNN BRASIL, **Pequenas empresas no Brasil beneficiam 40% da população, aponta Sebrae**. Disponível em <https://www.cnnbrasil.com.br/business/pequenas-empresas-no-brasil-beneficiam-40-da-populacao-aponta-sebrae/>

COOK, W. D.; ZHU, J., **Data Envelopment Analysis: Balanced Benchmarking**. CreateSpace Independent Publishing Platform, 2013

DB BROWSER, **DB Browser for SQLite**. Disponível em < https://sqlitebrowser.org/>

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2007

GUPTA, R. K., **Operations Research**. Krishna Prakashan Media, 2021

HAN, H. K.; SOHN, S. Y., **DEA Apolication to Grouping Military Airbases**. Military Operations Research SOC, 2011

HILLIER, F. S.; LIEBERMAN, G. J., **Introdução à pesquisa operacional**. McGraw Hill Brasil, 2006

MEDEIROS, R. V. V.; MARCOLINO, V. A., **A Eficiência dos Municípios do Rio de Janeiro no Setor de Saúde: Uma análise através da DEA e regressão logística**. Fundação Cesgranrio, 2018

OLIVEIRA, R. S. L, P.; PEDRO, M. I. C.; MARQUES, R. D. R. C., **Efficiency Evaluation of Portuguese Hotels in the Algarve using Data Envelopment Analysis (DEA)**. Revista Brasileira de Gestão de Negócios, 2015

PORTAL DA INDÚSTRIA, **Produtividade na Indústria**. Disponível em <https://static.portaldaindustria.com.br/media/filer_public/e6/b2/e6b265af-e4a4-4f90-9f26-f90cb28dfe09/produtividade_na_industria_janeiro-marco_2022.pdf>

PORTAL DA INDÚSTRIA, **Competitividade Brasil**. Disponível em <https://static.portaldaindustria.com.br/media/filer_public/ca/fc/cafc2274-9785-40db-934d-d1248a64dd94/competitividadebrasil_2019-2020_v1.pdf>

PORTAL DA INDÚSTRIA, **Indústria Brasileira no Mundo**. Disponível em <https://industriabrasileira.portaldaindustria.com.br/grafico/transformacao/mundo/#/industria-total>

PORTAL DA INDÚSTRIA, **A Importância da Indústria para o Brasil**. Disponível em <https://static.portaldaindustria.com.br/media/filer_public/1c/7e/1c7e271f-687e-46f9-81da-ebc99a88ccb6/flyer_a_importancia_da_industria_no_brasil_marco2022.pdf>

PORTAL DA INDÚSTRIA, **A Importância da Indústria de Transformação para o Brasil**. Disponível em <https://static.portaldaindustria.com.br/media/filer_public/6d/44/6d44a3a1-a017-4094-ae7c-78f469393c57/flyer_a_importancia_da_industria_no_brasil_transformacao_marco2022.pdf>

PORTAL DA INDÚSTRIA, **Panorama da Pequena Indústria**. Disponível em <https://static.portaldaindustria.com.br/media/filer_public/cf/94/cf948bb1-773c-48ca-8ba5-3587b50fe923/panorana_da_pequena_industria_abr-jun2022.pdf>

PYTHON SOFTWARE FOUNDATION, **Applications for Python**. Disponível em <https://www.python.org/about/apps/>

RAITH, A.; PEREDERIEIEVA, O.; FAUZI, F., **PyDEA** Disponível em <https://araith.github.io/pyDEA/index.html>

RECEITA FEDERAL, **Dados Públicos CNPJ**. Disponível em <https://www.gov.br/receitafederal/pt-br/assuntos/orientacao-tributaria/cadastros/consultas/dados-publicos-cnpj>

ROCHA, R. B.; CAVALCANTI NETTO, M. A., **A data envelopment analysis model for rank ordering suppliers in the oil industry**. Sociedade Brasileira de Pesquisa Operacional, 2002

SEBRAE, **Entenda as diferenças entre micro empresa, pequena empresa e MEI**. Disponível em <https://www.sebrae.com.br/sites/PortalSebrae/artigos/entenda-as-diferencas-entre-microempresa-pequena-empresa-e-mei,03f5438af1c92410VgnVCM100000b272010aRCRD>

SEBRAE, **Pequenos negócios em números**. Disponível em <https://www.sebrae.com.br/sites/PortalSebrae/ufs/sp/sebraeaz/pequenos-negocios-em-numeros,12e8794363447510VgnVCM1000004c00210aRCRD>

SEBRAE, **Micro e pequenas empresas geram 27% do PIB no Brasil**. Disponível em <https://www.sebrae.com.br/sites/PortalSebrae/ufs/mt/noticias/micro-e-pequenas-empresas-geram-27-do-pib-do-brasil,ad0fc70646467410VgnVCM2000003c74010aRCRD>

SILVEIRA, D. T.; CÓRDOVA, F. P. **Métodos de pesquisa**. Universidade Federal do Rio Grande do Sul, 2009.

SIMPLES NACIONAL, **O que é o Simples Nacional?** Disponível em <http://www8.receita.fazenda.gov.br/SimplesNacional/Documentos/Pagina.aspx?id=3>

SQLITE CONSORTIUM, **What is SQLite?** Disponível em < https://www.sqlite.org/index.html>

VALOR ECONÔMICO, **Maioria das empresas no país não dura 10 anos, e 1 de 5 fecha após 1 ano**. Disponível em <https://valor.globo.com/brasil/noticia/2020/10/22/maioria-das-empresas-no-pais-nao-dura-10-anos-e-1-de-5-fecha-apos-1-ano.ghtml>

VELOSO, F.; BONELLI, R.; CASTELAR, A. **Anatomia da Produtividade no Brasil**. Elsevier Editora Ltda e FGV IBRE, 2017

VELOSO, F.; BONELLI, R. **A Crise do Crescimento no Brasil**. Elsevier Editora Ltda e FGV IBRE, 2016

ZHU, J., **Data Envelopment Analysis: Let the Data Speak for Themselves**. CreateSpace Independent Publishing Platform, 2014

APÊNDICE

Os códigos em linguagem Pythona presentados a seguir foram desenvolvidos no software Jupyter Notebook, para realizar a filtragem, manipulação e limpeza dos dados, como foi mencionado no capítulo 3 deste trabalho.

APÊNDICE 1 – ADIÇÃO DE UMA COLUNA COM O NOME DO SETOR DA INDÚSTRIA RELACIONADO AO CÓDIGO CNAE PARA A BASE DE DADOS

```
import sqlite3
import pandas as pd
import numpy as np

df = pd.read_excel('cnae_names_final.xlsx')
df['codigo'] = df['codigo'].apply(lambda x: str(x)+'%')

def create_column(table_name, column_name, type='TEXT'):
    conn = sqlite3.connect('cnpj.db')
    cur = conn.cursor()
    cur.execute (f'ALTER TABLE {table_name} ADD {column_name} {type}')
    conn.commit()
    cur.close

create_column('cnae', 'industria')

def insert_data(df):
    conn = sqlite3.connect('cnpj.db')
    cur = conn.cursor()
    for codigo, nome in zip(df.codigo, df.nome):
        cur.execute(f'UPDATE cnae SET industria = "{nome}"WHERE codigo LIKE
"{codigo}";')
    conn.commit()
    cur.close()
```

```
    conn.close()

insert_data(df)

conn = sqlite3.connect('cnpj.db')
cur = conn.cursor()
cur.execute('PRAGMA table_info(empresas)')
result = cur.fetchall()
for a,b in zip(pd.DataFrame(result)[1], pd.DataFrame(result)[2]):
    print(a,(b.lower()))

table_list = [ ]
for table in result:
    table_list.append(table[0])
table_list[0]

def columns_info(db):
    table_list = []
    conn = sqlite3.connect(db)
    cur = conn.cursor()
    cur.execute('''SELECT name from sqlite_schema WHERE type = 'table' ''')
    result = cur.fetchall()
    for table in result:
        table_list.append(table[0])
    for table in table_list:
        cur.execute(f'PRAGMA table_info({table})')
        table_data = cur.fetchall()
        print('')
        print('TABLE ' + table + ' {')
        print('')
        for column_name, column_type in zip(pd.DataFrame(table_data)[1],
        pd.DataFrame(table_data)[2]):
            print(column_name, column_type.lower())
        print('}')

def create_view(view_name, query, db):
```

```
    conn = sqlite3.connect(db)
    cur = conn.cursor()
    cur.execute(f'CREATE VIEW IF NOT EXISTS v_socios AS {query}')
    print('view criada com sucesso!')
    conn.commit()
    print('commit realizado com sucesso!')
    cur.close()
    conn.close()

query = '''
        with empresas_ativas as (
            select
                socios.cnpj_cpf_socio,
                count(distinct socios.cnpj_basico) as num_empresas_ativas
            from
                socios
            left join
                estabelecimento on socios.cnpj_basico =
                estabelecimento.cnpj_basico
            where
                estabelecimento.situacao_cadastral = '02'
            group by 1
            ),

            empresas_inativas as (
            select
                socios.cnpj_cpf_socio,
                count(distinct socios.cnpj_basico) as num_empresas_inativas
            from
                socios
            left join
                estabelecimento on socios.cnpj_basico =
                estabelecimento.cnpj_basico
            where
                estabelecimento.situacao_cadastral != '02'
            group by 1
```

```
            )

            select
                socios.qualificacao_socio,
                socios.data_entrada_sociedade,
                socios.representante_legal,
                socios.faixa_etaria,
                empresas_ativas.num_empresas_ativas,
                empresas_inativas.num_empresas_inativas
            from socios
            left join empresas_ativas on socios.cnpj_cpf_socio =
            empresas_ativas.cnpj_cpf_socio
            left join empresas_inativas on socios.cnpj_cpf_socio =
            empresas_inativas.cnpj_cpf_socio

        '''

create_view('v_socios',query, 'cnpj.db')

conn = sqlite3.connect('cnpj.db')
cur = conn.cursor()
cur.execute('CREATE INDEX idx_estabelecimento_cep ON estabelecimento (cep)')
conn.commit()
cur.close()
conn.close()

conn = sqlite3.connect('cnpj.db')
cur = conn.cursor()
cur.execute('CREATE INDEX idx_estabelecimento_data_inicio_atividades ON
estabelecimento (data_inicio_atividades)')
conn.commit()
cur.close()
conn.close()

conn = sqlite3.connect('cnpj.db')
cur = conn.cursor()
```

```
cur.execute('CREATE INDEX idx_estabelecimento_data_situacao_cadastral ON
estabelecimento (data_situacao_cadastral)')

conn.commit()

cur.close()

conn.close()
```

APÊNDICE 2 – MANIPULAÇÃO E FILTRAGEM DOS DADOS ACESSADOS NO DATABASE

```
import sqlite3
import pandas as pd
import numpy as np
from datetime import datetime
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")

def df_query_result(query):
    conn = sqlite3.connect('cnpj.db') # conecta com o banco de dados utilizado no estudo
    conn.row_factory = sqlite3.Row # método utilizado para que o o cur.fetchall retorne o nome das colunas consultadas
    cur = conn.cursor()
    cur.execute(query) # executa a query
    rows = cur.fetchall() #retorna o resultado da query, mas é necessário formatar
    data=[]
    try:
        for row in rows:
            data.append(dict(row))
        return pd.DataFrame(data) # cria um DataFrame com os resultados da query
    except:
        return rows
    cur.close_connection() # fecha a conexão com o banco de dados

query = '''
        select
            -- estabelecimento
            estabelecimento.cnpj,
            estabelecimento.cnpj_basico,
            industria,
```

```
            case when estabelecimento.situacao_cadastral = '02' then 'ativa'
            else 'encerrada' end as situacao_cadastral,
            situacao_cadastral as situacao_cadastral_raw,
            estabelecimento.data_situacao_cadastral,
            estabelecimento.data_inicio_atividades,
            estabelecimento.cnae_fiscal,
            estabelecimento.cnae_fiscal_secundaria,
            estabelecimento.uf,
            estabelecimento.tipo_logradouro,
            estabelecimento.nome_fantasia,
            -- empresas
            empresas.capital_social,
            empresas.natureza_juridica,
            empresas.qualificacao_responsavel,
            -- simples
            case
                when
                    simples.data_opcao_simples                              <=
estabelecimento.data_inicio_atividades
                    and simples.data_opcao_simples is not null then 'S'
                else
                    'N' end as opcao_simples,
            simples.data_opcao_simples,
            simples.data_exclusao_simples,
            --socios
            socios.qualificacao_socio,
            socios.data_entrada_sociedade,
            socios.representante_legal,
            socios.faixa_etaria,
            socios.nome_socio,
            socios.cnpj_cpf_socio
        from estabelecimento
        left    join    empresas    on    estabelecimento.cnpj_basico    =
empresas.cnpj_basico
        inner join socios on estabelecimento.cnpj_basico = socios.cnpj_basico
        left    join    simples    on    estabelecimento.cnpj_basico    =
simples.cnpj_basico
```

```
        inner join cnae on estabelecimento.cnae_fiscal = cnae.codigo
        where
            industria is not null
            and situacao_cadastral_raw in ('02', '08')
            and porte_empresa in ('01', '03')
            and cast(data_inicio_atividades as int) >= 20110101
            and cast(data_inicio_atividades as int) < 20120101
            and socios.nome_socio is not null
            and socios.data_entrada_sociedade = data_inicio_atividades
            and estabelecimento.matriz_filial = '1'
            and estabelecimento.cnpj_ordem = '0001'
            '''
df_base = df_query_result(query)


data=df_base.drop_duplicates('cnpj')[['situacao_cadastral']].value_counts()
.reset_index().rename(columns={0:'value'})


plt.bar(data=data, x='situacao_cadastral',height='value')


plt.show()


data


# Função para acrescentar uma coluna com o número de empresas ativas que o
sócio possuia na época em que abriu a empresa


def num_empresas_ativas(df):
    conn = sqlite3.connect('cnpj.db')
    cur = conn.cursor()
    num_empresas_ativas = []
    for nome_socio, cnpj_cpf_socio, data_inicio_atividades in
zip(df['nome_socio'], df['cnpj_cpf_socio'], df['data_inicio_atividades']):
        query = f'''
                select
                    socios.cnpj_cpf_socio || socios.nome_socio,
                    count(distinct socios.cnpj)
                from
```

```
                    socios
                left join
                    estabelecimento on socios.cnpj = estabelecimento.cnpj
                where
                    (socios.cnpj_cpf_socio = '{cnpj_cpf_socio}'
                    and socios.nome_socio = '{nome_socio}'
                    and       estabelecimento.data_inicio_atividades       <
{data_inicio_atividades}
                    and estabelecimento.matriz_filial = '1')
                    and (estabelecimento.situacao_cadastral = '02'
                    or estabelecimento.situacao_cadastral != '02'
                    and      estabelecimento.data_situacao_cadastral      <=
{data_inicio_atividades})
                group by 1
        '''
        cur.execute(query)
        row = cur.fetchall()
        if len(row)==0:
            num_empresas_ativas.append(0)
        else:
            num_empresas_ativas.append(row[0][-1])


    cur.close()
    conn.close()


    df.loc[:,'num_empresas_ativas'] = num_empresas_ativas


    return df


df_base = num_empresas_ativas(df_base)


df_ativa                                                                =
df_base.groupby('cnpj')['num_empresas_ativas'].sum().reset_index().rename(c
olumns={'num_empresas_ativas':'num_empresas_ativas_tot'})
df_base = df_base.merge(df_ativa, on='cnpj')


# Função para acrescentar uma coluna com o número de empresas encerradas que
o sócio possui
```

```
def num_empresas_encerradas(df):
    conn = sqlite3.connect('cnpj.db')
    cur = conn.cursor()
    num_empresas_encerradas = []
    for nome_socio, cnpj_cpf_socio, data_inicio_atividades in
zip(df['nome_socio'], df['cnpj_cpf_socio'], df['data_inicio_atividades']):
        query = f'''
                select
                    socios.cnpj_cpf_socio || socios.nome_socio,
                    count(distinct socios.cnpj)
                from
                    socios
                left join
                    estabelecimento on socios.cnpj = estabelecimento.cnpj
                where
                    socios.cnpj_cpf_socio = '{cnpj_cpf_socio}'
                    and socios.nome_socio = '{nome_socio}'
                    and estabelecimento.data_inicio_atividades <
{data_inicio_atividades}
                    and estabelecimento.matriz_filial = '1'
                    and estabelecimento.situacao_cadastral != '02'
                group by 1
        '''
        cur.execute(query)
        row = cur.fetchall()
        if len(row)==0:
            num_empresas_encerradas.append(0)
        else:
            num_empresas_encerradas.append(row[0][-1])

    cur.close()
    conn.close()

    df.loc[:,'num_empresas_encerradas'] = num_empresas_encerradas

    return df
```

```
df_base = num_empresas_encerradas(df_base)


df_encerrada                                                                    =
df_base.groupby('cnpj')['num_empresas_encerradas'].sum().reset_index().rena
me(columns={'num_empresas_encerradas':'num_empresas_encerradas_tot'})

df_base = df_base.merge(df_encerrada, on='cnpj')


# Função para acrescentar uma coluna com o número de empresas encerradas que
o sócio possui


def num_filiais(df):

    conn = sqlite3.connect('cnpj.db')

    cur = conn.cursor()

    num_filiais = []

    for   cnpj_basico,   data_inicio_atividades   in   zip(df['cnpj_basico'],
df['data_inicio_atividades']):

        query = f'''

                select

                    count(distinct estabelecimento.cnpj)

                from

                    estabelecimento

                where

                    estabelecimento.cnpj_basico = '{cnpj_basico}'

                    and       estabelecimento.data_inicio_atividades       =
{data_inicio_atividades}

                    and estabelecimento.matriz_filial = '2'

        '''

        cur.execute(query)

        row = cur.fetchall()

        num_filiais.append(row[0][0])


    cur.close()


    conn.close()


    df.loc[:,'num_filiais'] = num_filiais
```

```
    return df


df_base = num_filiais(df_base)


df_base.loc[:,'cnpj_basico'] = df_base['cnpj'].apply(lambda x: x[:8])


df = df_base.copy()


df.loc[:,'data_inicio_atividades']                                    =
df['data_inicio_atividades'].apply(lambda      x:      datetime.strptime(x,
"%Y%m%d").date())


df.loc[:,'data_situacao_cadastral']                                   =
np.where(df['data_situacao_cadastral']=='0',

df['data_inicio_atividades'],

df['data_situacao_cadastral'])


df.loc[:,'data_situacao_cadastral']                                   =
df['data_situacao_cadastral'].apply(lambda      x:      datetime.strptime(x,
"%Y%m%d").date() if type(x) is str else x)


data_bd = datetime(2022,2,12).date()


df.loc[:, 'tempo_empresa'] = np.where(df['situacao_cadastral'] == 'ativa',
                                      data_bd                          -
df['data_inicio_atividades'],
                                      df['data_situacao_cadastral']    -
df['data_inicio_atividades'])


df.loc[:,'num_cnae_sec'] = np.where(df['cnae_fiscal_secundaria']=='',
                         0,
                         df['cnae_fiscal_secundaria'].apply(lambda        x:
len(x.split(','))))


num_socios                                                            =
df.groupby('cnpj')[['qualificacao_socio']].count().rename(columns={'qualifi
cacao_socio':'num_socios'}).reset_index()


df = df.merge(num_socios, how='left')
```

```
df_y = df.sort_values(['cnpj','faixa_etaria'], ascending=True).drop_duplicates('cnpj',keep='first').reset_index(drop=True)

df_y.rename(columns={'faixa_etaria':'faixa_etaria_y'},inplace=True)


df_o = df.sort_values(['cnpj','faixa_etaria'], ascending=False).drop_duplicates('cnpj',keep='first').reset_index(drop=True)

df_o.rename(columns={'faixa_etaria':'faixa_etaria_o'},inplace=True)


df = df_o.merge(df_y[['cnpj','faixa_etaria_y']], on='cnpj')


pd.read_html('https://inanyplace.blogspot.com/2017/01/lista-de-estados-brasileiros-sigla-estado-capital-e-regiao.html')[0].head()


# Utilizei uma tabela disponibilizada online, para fazer essa relação entre estado e região


tabela_estados = pd.read_html('https://inanyplace.blogspot.com/2017/01/lista-de-estados-brasileiros-sigla-estado-capital-e-regiao.html')[0][['Sigla','Região','Estado']]


df = pd.merge(df,tabela_estados, how='left', left_on='uf', right_on='Sigla')


df.rename(columns={'Região':'regiao',    'Sigla':'sigla','Estado':'estado'}, inplace=True)


df.loc[:,'regiao'] = df['regiao'].apply(lambda x: str(x).lower())


sns.set_theme(style="whitegrid")

sns.boxplot(x='situacao_cadastral',       y='capital_social',       data=df, fliersize=10)

plt.show()


# Serão considerados 99% dos registros -- ainda necessita de validação


df = df[df['capital_social'] <= df['capital_social'].quantile(.99)].reset_index(drop = True)
```

```
df.columns


colunas_modelo = ['cnpj', 'situacao_cadastral', 'uf', 'capital_social',
                  'opcao_simples', 'faixa_etaria_o', 'faixa_etaria_y',
                  'tempo_empresa','num_cnae_sec', 'qualificacao_socio',
                  'regiao','num_socios', 'num_empresas_ativas_tot',
                  'num_empresas_encerradas_tot','natureza_juridica',
                  'num_filiais','tipo_logradouro',
                  'qualificacao_responsavel', 'industria']


df_model = df[colunas_modelo]


df_model.head()


file_name = 'df_model_v1.xlsx'


df_model.to_excel(file_name)

df_dea = pd.read_excel('dea_solution_v1.xlsx', sheet_name = 'onion_rank',
header = 2)


df_dea.head()
```