Segmentação de Carteira de Clientes com base no Cadastro Nacional de Empresas Exportadoras Brasileiras

Bonecker. Renata M. M., Moreira. Cesar A. C., e Paula. Dayane E.C.

Esse trabalho foi apresentado como trabalho de conclusão do Curso de Especialização em *Data Science* (CEDS) do Instituto de Aeronáutica (ITA), com o patrocínio da Agência Brasileira de Promoção de Exportações e Investimentos (Apex-Brasil).

RESUMO Este estudo realiza a identificação do perfil das empresas brasileiras exportadoras, a partir da lista de exportadoras e importadoras divulgada pelo governo brasileiro, dados fornecidos pela SECEX-MDIC (Secretaria de Comércio Exterior do Ministério do Desenvolvimento, Indústria, Comércio e Serviços) e pela Receita Federal do Brasil. Realizada a análise exploratória dos dados, técnicas de análise de componentes principais, de agrupamento (k-médias) e de mineração de regras de associação foram aplicadas. Por meio da análise de componentes principais, considerando dez atributos quantitativos, foi possível identificar uma quantidade de grupos para iniciar o agrupamento via k-médias. Por conseguinte, por meio de k-médias, foi possível validar a quantidade de grupos e analisar as características de cada grupo para os atributos quantitativos selecionados. E, por fim, com a utilização da técnica de mineração de regras de associação, foi possível extrair características interpretáveis de apenas um grupo. Como resultado, há evidências que os dados existentes não são adequados para identificar o perfil das empresas e que, portanto, para atingir o objetivo será necessário enriquecer a base com informações, tais como valor exportado pela empresa, produto exportado e quantidade de países para os quais a empresa exportou, entre outras.

PALAVRAS-CHAVES Apex-Brasil, ciência de dados, empresas brasileiras, exportadoras, importadoras, análise de componentes principais, agrupamento, método k-médias, mineração de regras de associação.

I. INTRODUÇÃO

A Apex-Brasil apoiou em 2021, o total de 3.660 empresas exportadoras [1], em um universo de 28.147 exportadoras brasileiras [2]. Isso corresponde a cerca de 13% do universo de exportadoras, evidenciando que 87% deste grupo não foi atingido pelo atendimento realizado pela Apex-Brasil. Nesse contexto, entende-se que é premissa conhecer melhor as características das empresas exportadoras brasileiras, e não apenas as das empresas já apoiadas pela Agência. Dessa forma, será possível construir estratégias assertivas de prospecção e retenção de clientes e adequar o portfólio de produtos e serviços às necessidades desse público-alvo em seu ciclo de internacionalização, que pode ser entendido como "um desses fenômenos emergentes, que assume nas últimas décadas um caráter basilar nas estratégias organizacionais. Dito de outro modo, a internacionalização não é mais vista como um processo secundário vislumbrado em um futuro estratégico das empresas e sim como um pressuposto da natureza estratégica das empresas globais com atuação no mercado nacional e/ou internacional" [3].

A motivação deste trabalho é justificada pelos seguintes argumentos:

i. A Apex-Brasil iniciou em 2022 o trabalho de reestruturação do portfólio de produtos e serviços, porém, sem o conhecimento das empresas que ainda não são clientes. Direcionar corretamente as ofertas atualmente é um trabalho instintivo;

ii. Não existe uma jornada do cliente formalmente definida na Apex-Brasil. É o cliente que constrói sua jornada, conforme conhece a oferta de produtos e serviços vigentes. Já foi realizado de forma descritiva um estudo para identificar o comportamento de consumo dos clientes da Apex-Brasil ao longo dos anos. Com a identificação de grupos de empresas a partir de determinadas características, será possível analisar se há uma jornada comum para esses grupos, adequando o portfólio para essas necessidades e realizando uma oferta mais assertiva aos clientes tanto da base da Apex-Brasil, como também para potenciais clientes.

iii. A partir desse estudo, espera-se que a Apex-Brasil possa expandir sua atuação, trazendo novas empresas para o seu portfólio e mantendo-as com interesse contínuo nos produtos e serviços oferecidos.

iv. Ajudará a estabelecer estratégias mais assertivas de segmentação e relacionamento com o cliente.

II. OBJETIVOS

O objetivo geral desta pesquisa é buscar, por meio de aprendizado de máquina não supervisionado, agrupar as empresas exportadoras brasileiras, modelando grandes grupos que serão trabalhados pela Apex-Brasil por suas diversas áreas que realizam atendimento às empresas. Com esses resultados, será possível propor o aprofundamento e complementação de estudos futuros para suportar as operações de segmentação de clientes da Agência.

São objetivos específicos desta pesquisa:

- Obter as bases de dados do cadastro de empresas exportadoras brasileiras entre os anos de 2017 e 2021, disponibilizadas até então pelo governo brasileiro;
- Enriquecimento dos dados a partir de outras fontes, como a Receita Federal do Brasil ou outra fonte pública;
- Realizar o agrupamento das empresas, buscando regras de associação para cada grupo e procurando por padrões, por meio de modelo descritivo de aprendizado de máquina;
- Identificação dos grandes grupos e suas características básicas.

III. EMPRESAS BRASILEIRAS E O PROCESSO DE INTERNACIONALIZAÇÃO

Segundo a Apex-Brasil [4], a globalização e a internacionalização das empresas são fatos que impactam as economias, com reflexos nas empresas por todo o mundo. A realidade impõe-se à competitividade dos setores econômicos, não restando alternativa a não ser projetar-se para fora do mercado doméstico e se inserir nas Cadeias Globais de Valores, com a internacionalização da produção. Apenas para a compreensão da importância da internacionalização, destaca-se que o estudo do Comércio Internacional deve-se dar em duas abordagens: a Econômica e a de Internacionalização.

Em linhas gerais, a abordagem Econômica ou da Teoria Econômica estuda a relação de trocas comerciais entre Estados, busca explicar a necessidade da dinâmica comercial entre eles [5] e:

(i) os principais fatores e a competitividade do mercado internacional. As principais teorias dessa abordagem são 5 (cinco): das Vantagens Absolutas; das Vantagens Comparativas; da Dotação Relativa de Fatores; de Comércio de Produtos Industrializados; e da Vantagem Competitiva.

(ii) No que tange às Teorias Econômicas de Internacionalização, as quais tentam explicar do ponto de vista das empresas a sua necessidade de internacionalização, temos 3 (três) teorias: a do Poder de Mercado; a da Internalização; e a do Paradigma Eclético.

(iii) conceitualmente, na Internacionalização [6], busca-se ilustrar o comportamento das empresas, seus fatores críticos de competitividade e sua inserção estratégica no mercado internacional. As principais Teorias Comportamentais de Internacionalização são 4 (quatro): O Modelo de Internacionalização de Uppsala; *Networks* ou Redes de Negócios; Empreendedorismo Internacional; e *The Born Global Firm*.

Não é escopo deste trabalho esmiuçar exaustivamente nenhum desses conceitos, mas tão somente buscar padrões por intermédio do estudo de ciência de dados para identificar perfis das empresas brasileiras exportadoras, de modo a contribuir, em pesquisas futuras, em referenciar os achados às teorias relacionadas.

IV. MATERIAIS E MÉTODOS

Para o referido estudo, buscou-se os dados de empresas exportadoras e importadoras disponibilizado anualmente pela Secretaria de Comércio Exterior (SECEX), órgão específico ligado ao Ministério do Desenvolvimento, Indústria e Comércio Exterior (MDIC), que lida com a definição de políticas públicas para o comércio exterior e gestão do comércio. Dentre suas atribuições estão a normatização, supervisão, orientação, planejamento, controle e avaliação do comércio exterior, sendo essa Secretaria a responsável pela coleta e disponibilização dos dados das empresas exportadoras e importadoras.

Os dados trabalhados nesse estudo compreendem o período de 2017 a 2021. Não foi possível incluir os dados referentes ao ano de 2022 no estudo em decorrência da descontinuidade dessa publicação, conforme nota informativa publicada pela SECEX em 08/03/2023 [7]. Portanto, além de poucos dados, o estado ainda suspendeu o acesso aos dados de empresas, dificultando a pesquisa com este escopo.

Adicionalmente, os dados obtidos junto à SECEX foram enriquecidos com os dados do Cadastro Nacional da Pessoa Jurídica (CNPJ), disponibilizados pela Receita Federal do Brasil [8], órgão responsável pelas informações cadastrais das pessoas jurídicas. Como esses dados são referentes ao presente, com um ano de defasagem por natureza do banco de dados, decidiu-se limitar os dados de exportação e importação disponibilizados pelo MDIC ao período de cinco anos, a fim de não enviesar a identificação do perfil das empresas.

Assim, em relação à estruturação da base, coletou-se os dados das empresas brasileiras exportadoras de 2017 a 2021 e identificou-se quais delas haviam importado entre os anos avaliados. Desta forma, para esses atributos se exportou/importou ou não entre 2017 e 2021, utilizou-se a técnica de *one hot-encoding* [9], uma técnica de conversão simbólica-numérica em que cada um dos atributos sobre se a empresa exportou entre 2017 e 2021 e se importou entre 2017 e 2021 foi transformado em coluna, e preenchida por 0

(zero) – indicando a ausência da característica, ou 1 (um) – indicando a presença da característica, para cada empresa.

Portanto, o banco de dados utilizado para o estudo é composto pelas seguintes variáveis e tipologias:

TABELA I
DESCRIÇÃO DOS ATRIBUTOS DO CONJUNTO DE DADOS.

Atributo	Tipologia	Descrição
CNPJ	qualitativa nominal	Código identificador da empresa no Cadastro Nacional de Pessoas Jurídicas (CNPJ 14 dígitos)
Empresa	qualitativa nominal	Nome informado pela Empresa
Endereço	qualitativa nominal	Endereço informado pela empresa
Número	quantitativa racional	Nº do local em que a empresa se localiza
Bairro	qualitativa nominal	Bairro onde a empresa está localizada
CEP	qualitativa nominal	Código de Endereçamento Postal (CEP) do local da empresa
Município	qualitativa nominal	Município em que a empresa está localizada
UF	qualitativa nominal	Unidade Federativa (UF) onde a empresa está localizada
CNAE Primária	qualitativa nominal	Descrição do nível de menor granularidade da Classificação Nacional de Atividades Econômicas (CNAE) informado pela empresa
Seção CNAE Código	qualitativa nominal	Código do primeiro nível da estrutura da CNAE
Seção CNAE Descrição	qualitativa nominal	Descrição do primeiro nível da estrutura da CNAE
Divisão CNAE 2.0 Código	qualitativa nominal	Código do segundo nível da estrutura da CNAE
Divisão CNAE 2.0 Descrição	qualitativa nominal	Descrição do segundo nível da estrutura da CNAE
Natureza Jurídica	qualitativa nominal	Código de Natureza Jurídica informado pela empresa
Porte	qualitativa nominal	Classificação quanto ao porte da empresa cadastrado na RFB
Região	qualitativa nominal	Região geográfica, dada de acordo com a UF
Exportou em 2017	quantitativa racional	Verificação de exportação realizada pela empresa: 1 (sim) e 0 (não) no ano analisado
Importou em 2017	quantitativa racional	Verificação de importação realizada pela empresa: 1 (sim) e 0 (não) no ano analisado
Exportou em 2018	quantitativa racional	Verificação de exportação realizada pela empresa: 1 (sim) e 0 (não) no ano analisado
Importou em 2018	quantitativa racional	Verificação de importação realizada pela empresa: 1 (sim) e 0 (não) no ano analisado
Exportou em 2019	quantitativa racional	Verificação de exportação realizada pela empresa: 1 (sim) e 0 (não) no ano analisado
Importou em 2019	quantitativa racional	Verificação de importação realizada pela empresa: 1 (sim) e 0 (não) no ano analisado
Exportou em 2020	quantitativa racional	Verificação de exportação realizada pela empresa: 1 (sim) e 0 (não) no ano analisado
Importou em 2020	quantitativa racional	Verificação de importação realizada pela empresa: 1 (sim) e 0 (não) no ano analisado
Exportou em 2021	quantitativa racional	Verificação de exportação realizada pela empresa: 1 (sim) e 0 (não) no ano analisado
Importou em 2021	quantitativa racional	Verificação de importação realizada pela empresa: 1 (sim) e 0 (não) no ano analisado
Regularidade Exportadora	quantitativa racional	Soma dos períodos em que a empresa exportou, entre 2017 e 2021 (1: Iniciante; 2: Iniciante; 3: Intermediária; 4:Experiente; e 5:Experiente)
Regularidade Exportadora Classificação	qualitativa nominal	Classificação da empresa quanto à soma dos períodos em que a empresa exportou (Iniciante, Intermediária; e Experiente)
Regularidade Importadora	quantitativa racional	Soma dos períodos em que a empresa importou, entre 2017 e 2021(0:Não importadora; 1:Iniciante; 2:Iniciante; 3: Intermediária; 4:Experiente; e 5: Experiente)

Regularidade Importadora Classificação	qualitativa nominal	Classificação da empresa quanto à soma dos períodos em que a empresa importou (Não Importadora; Iniciante, Intermediária; e Experiente)

As variáveis da base de dados criada podem se dividir em seis grupos, quanto ao objetivo de cada:
(i) Identificação da empresa: CNPJ (14 dígitos), Empresa;
(ii) Localização geográfica: Endereço, Número, Bairro, CEP, Município, UF e Região;
(iii) Principal atividade econômica exercida: CNAE PRIMÁRIA, Secão CNAE Código, Seção CNAE Descrição, Divisão CNAE 2.0 Código e Divisão CNAE 2.0 Descrição;
(iv) Estrutura da empresa: Natureza Jurídica e Porte;
(v) Quando a empresa exportou, entre 2017 e 2021: Exportou em 2017, Exportou em 2018, Exportou em 2019, Exportou em 2020, Exportou em 2021; e sua Regularidade Exportadora e Classificação da Regularidade Exportadora.
(vi) Quando a empresa exportadora também importou, entre 2017 e 2021: Importou em 2017, Importou em 2018, Importou em 2019, Importou em 2020, Importou em 2021 e sua Regularidade Importadora e Classificação da Regularidade Importadora.

Assim, a base de dados foi criada para se ter apenas uma observação para cada CNPJ 14.

Destaca-se que toda a pesquisa foi executada utilizando a linguagem e algoritmos em R.

Em um primeiro momento, fez-se a análise exploratória dos dados e o pré-processamento das informações. Foram identificadas as variáveis com valores ausentes e quais deveriam ter a classificação com maior granularidade para apoiar no posterior agrupamento.

Conhecidos e preparados os dados, seguiu-se com o método de agrupamento k-médias, a fim de auxiliar na identificação de grupos das empresas brasileiras exportadoras.
No uso do método do k-médias [10], utiliza-se a seguinte rotina:
1. Escolher o valor de K e as K estimativas iniciais para os centroides;
2. Calcular a distância de cada ponto de dados (x, y) para cada centroide. Atribuir cada ponto ao centroide mais próximo. Essa associação define os primeiros K grupos.
3. Calcular o centroide, o centro de massa, de cada grupo recém-definido da Etapa 2.

O processo é repetido até que os centroides não sejam mais alterados ou quando o algoritmo atinge um critério de parada, como um número predeterminado de iterações [11].
As técnicas de agrupamento possuem aprendizado não supervisionado, ou seja, usam conjunto de dados não rotulados para criar modelos. Os objetos são comparados em medida de distância [12].

De acordo com Jain [13], são 5 passos que devem ser observados para a utilização de técnica de agrupamento, a saber: 1. Definição da forma de representação dos dados de entrada a serem agrupados;
2. Definição de uma medida adequada de aproximação entre os dados;
3. Definir qual a técnica de agrupamento a ser aplicada para a construção dos grupos;
4. Definição de uma abstração dos dados. Este passo envolve a definição de uma forma de abstração dos dados que permita uma representação simples e compacta do conjunto de dados. A representação deve ser intuitiva e fácil de ser compreendida, bem como a abstração dos dados deve permitir um processamento posterior eficiente; e
5. Avaliação do resultado do agrupamento. O que se procura neste passo é, basicamente, avaliar quão bom foi o agrupamento dos dados conseguido pela técnica de agrupamento aplicada.

Para isso, primeiro foram selecionadas as variáveis quantitativas da base, as quais indicam se a empresa exportou e importou entre 2017 e 2021. Não foram consideradas as variáveis “Regularidade Exportadora” e “Regularidade Importadora”, pois traziam redundância de informação.

A motivação para a escolha apenas destas variáveis para a identificação dos grupos é que se quis segmentar as empresas quanto ao seu comportamento (padrão de importação e exportação) e, posteriormente, explicar os grupos pelos atributos intrínsecos às empresas (porte, atividade econômica, localização).

Em seguida, aplicou-se a análise de componentes principais, a fim de apoiar na escolha inicial do número de grupos (K) para iniciar o agrupamento.

A análise dos componentes principais (ACP ou PCA - do inglês), que faz parte da estatística multivariada, é um método que tem por finalidade básica a análise dos dados usados visando sua redução, eliminação de sobreposições e a escolha das formas mais representativas de dados a partir das combinações lineares das variáveis originais [14].

Definido o número inicial de grupos, seguiu-se com a aplicação do k-médias, que possibilitou:
- Identificação do grupo em que cada observação ficou alocada;
- Análise dos grupos encontrados, de acordo com o centroide encontrado em cada atributo, para cada grupo definido.

Posteriormente, foi realizada a análise da estabilidade do algoritmo para diferentes quantidades de grupos e avaliada a qualidade do agrupamento para o número K de grupos definido. Nesta etapa, foi utilizado gráfico boxplot com K

assumindo valores de 2 a 9 no eixo x, e no eixo y assumindo o valor da soma total dos quadrados das distâncias entre todas as observações e seus respectivos centros de grupo. Além disso, foi utilizado o método de reamostragem *bootstrap* no conjunto de dados para analisar a variabilidade dos grupos para 10 diferentes amostras.

A técnica de *Bootstrap* realiza repetidamente a substituição com reposição de amostras de um conjunto de dados de acordo com uma distribuição de probabilidade uniforme, mantendo o total de exemplos na base [15].

Considerando que o algoritmo k-médias é sensível à escolha inicial dos centroides e da sua forma de atualização [16], somente seguiu-se com a análise dos resultados da próxima etapa da pesquisa, após executado o algoritmo de agrupamento diversas vezes.

Portanto, após identificação de grupos pelas variáveis quantitativas da base, seguiu-se com a mineração de regras de associação, considerando agora, da base de dados, os dados qualitativos e a informação do grupo obtido com o método de agrupamento k-médias.

A mineração de regras de associação é uma técnica de aprendizado não supervisionado, que é um tratamento descritivo que busca identificar relações em um determinado conjunto de dados [17].

Assim, para esta etapa da pesquisa, foram utilizadas as seguintes variáveis: “Seção CNAE Descrição", "Natureza Jurídica", "Porte", "Região", "Regularidade Exportadora Classificação", "Regularidade Importadora Classificação" e grupo identificado na etapa anterior.

Com esses dados, foram criados objetos de transações, que é a representação das transações de associação entre os itens (atributos), em que cada linha representa uma transação, e cada coluna representa um item. Cada célula contém um valor booleano indicando se o item está presente ou não naquela transação. Cada transação, portanto, indica um conjunto de itens que aparecem juntos com frequência [18].

O próximo passo foi, portanto, encontrar os conjuntos frequentes de itens da base de dados transacional criada, por meio do algoritmo *eclat (Equivalence Class Grupoing and Bottom-Up Lattice Traversal)*, com suporte mínimo de 0,1 (para encontrar conjuntos que aparecem pelo menos em 10% das transações) e tamanho máximo de 2 itens por conjunto.

Como a partir de conjuntos frequentes é possível derivar regras de associação, seguiu-se com a definição de regras, com confiança mínima de 0,9. A confiança consiste na probabilidade de ocorrer um conjunto de termos dado que ocorreu um outro conjunto e o algoritmo seleciona as regras cuja confiança é superior à confiança mínima definida [19].

Dada a definição das regras, foram identificadas e analisadas as regras de associação encontradas, conforme valores obtidos de antecedente, consequente, suporte, confiança e *lift*. O *lift*, ou coeficiente de interesse, reflete a noção estatística de independência entre duas variáveis aleatórias e auxilia na seleção das regras de associação [20].

Por fim, por meio de agrupamento seguido da identificação de regras de associação, buscou-se alcançar os objetivos desta pesquisa, precipuamente, a definição de grupos e perfis de empresas brasileiras exportadoras para a base de dados acessível à Apex-Brasil.

V. RESULTADOS

A. Análise Exploratória dos Dados

Foi realizada a análise exploratória com o objetivo de conhecer, gerar insights e hipóteses sobre os dados, e preparar a base para o uso de algoritmos capazes de identificar os grandes grupos de empresas brasileiras exportadoras, objeto desse estudo.
Foi realizada a análise de variáveis ausentes, com o seguinte resultado:

TABELA II
ANÁLISE DE VARIÁVEIS AUSENTES

Nome da Variável	Contagem de Ausências
NÚMERO	20560
BAIRRO	936
ENDEREÇO	808
DIVISÃO CNAE 2.0 – DESCRIÇÃO	62
CEP	6
REGIÃO	6
MUNICÍPIO	1
UF	1

Embora a variável “Número” tenha 20.560 linhas sem informação, corresponde a um complemento específico do endereço e não acrescenta informação a essa análise. Essas linhas não foram desconsideradas, apenas a coluna “Número” não foi utilizada.

Portanto, dado os valores ausentes e a informação apresentada, as variáveis que não foram utilizadas para as etapas de agrupamento e regras de associação foram: “CNPJ”, “Empresa”, “Número”, “Bairro”, “Endereço”, “Município”, “UF”, “Seção CNAE Código”, “Divisão CNAE 2.0 Código” e “Divisão CNAE 2.0 Descrição”.

A análise das variáveis numéricas não apresentou valores inválidos ou outliers no banco de dados.

A base de dados contém 51.620 empresas com CNPJs, considerando a quantidade de 14 dígitos do Código CNPJ.
Observa-se que em casos especiais pode se utilizar o código com 8 dígitos, para que se identifiquem empresas de um

mesmo grupo econômico (matriz e filiais). Contudo, para efeitos do estudo, foram utilizados os 14 dígitos originais do Banco de Dados, a fim de não perder informações, tais como da região geográfica das empresas brasileiras exportadoras.

Por meio dos dados do CEP, relacionados ao endereço, foi possível identificar a geolocalização das empresas brasileiras exportadoras ao longo do período de 2017 a 2021, conforme figura 1 a seguir:

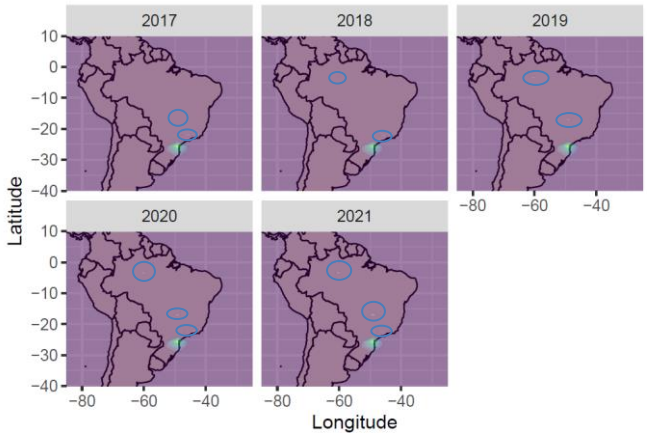


Fig.1. Geolocalização das empresas brasileiras exportadoras, entre 2017 e 2021. Cores representam a densidade estimada das empresas – tons rosas indicam pouca concentração e tons amarelos muita concentração. Círculos em azul são destaques para facilitar a visualização.

Observa-se que durante os cinco anos avaliados a região Sudeste e Sul mantiveram a maior concentração de empresas brasileiras exportadoras em relação às outras regiões. Além disso, observa-se que a partir de 2019 há um certo crescimento de empresas brasileiras exportadoras do Centro-Oeste e do Norte, conforme destacado pelos círculos azuis na figura 1.

A base de dados apresenta uma concentração de empresas na Unidade Federativa de São Paulo, contendo 42,83% das empresas exportadoras, seguida do Rio Grande do Sul, com 9,97%.

Pela análise da variável "CEP", é possível identificar polos empresariais nos estados, tais como o CEP 9370-0000, que indica a localização de 132 empresas na cidade de Campo Bom, no Rio Grande do Sul, e o CEP 1329-5000, que indica a localização de 127 empresas no município de Itupeva, em São Paulo.

Em relação à atividade econômica, os 7 primeiros CNAES com maior número de frequência nos dados, representam cerca de 12,07% dos registros, o que indicou a necessidade de considerar a atividade econômica da empresa pela variável "Seção CNAE Descrição", com maior sumarização, do que pela variável "CNAE Primária". São estes os 7 CNAES: comércio varejista de outros produtos novos não especificados anteriormente; confecção de peças de vestuário, exceto roupas íntimas; fabricação de artefatos de material plástico não especificados anteriormente; comércio atacadista de máquinas e equipamentos para uso industrial, partes e peças; comércio varejista de artigos de vestuário e acessórios; comércio atacadista de equipamentos e artigos de uso pessoal e doméstico não especificados anteriormente.

Também é possível observar uma predominância de empresas exportadoras no setor "Indústrias de Transformação", que corresponde a 50,8% da base (26.230 empresas), seguida pelo setor "Comércio; Reparação de Veículos Automotores e Motocicletas", com 36,9% da base (19.077 empresas). O setor de "Agricultura, Pecuária, Produção Florestal, Pesca e Aqüicultura" detém apenas 0,01% da base, com 745 empresas.

Em relação à natureza jurídica, 69,15% são de sociedade por cotas de responsabilidade limitada - empresa privada, seguida por 13,1% da natureza jurídica do tipo empresa individual de responsabilidade limitada (natureza empresária).

Ao analisarmos o porte dessas empresas, observa-se a seguinte distribuição: 20,5% delas são microempresas (10.594 empresas), 18,4% são pequenas empresas (9.512 empresas) e 60,6% são médias e grandes empresas (31.275), conforme classificação da RFB. Existe ainda um universo de 239 empresas (0,5%) com o porte não informado.

Dentre o período analisado, o ano de 2021 é o que apresenta o maior número de exportadoras, seguido pelo ano de 2020, 2019, 2018 e 2017.

Fazendo a análise das variáveis e de relacionamento entre elas, observa-se que entre 2017 e 2021 o número de empresas brasileiras que exportaram continuamente, entre 2017 e 2021, foi de 13.018, o que corresponde a 25,22% de toda a base em análise.

Em relação à localização geográfica dessas empresas que exportaram continuamente por 5 anos, observou-se que 55,54% são da região Sudeste, 31,25% da Região Sul, 5,64% da Região Nordeste, 4,19% da Região Centro-Oeste e 3,39% da Região Norte. Ou seja, juntos as regiões Nordeste, Centro-Oeste e Norte correspondem a menos de 1/4 de toda exportação para a amostra analisada, conforme o gráfico da figura 2 abaixo:

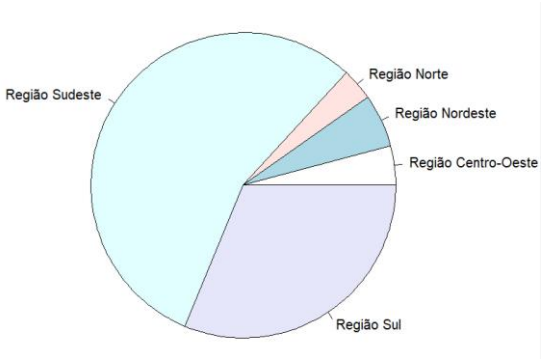


Fig. 2. Distribuição das empresas por Região Econômica.

Em relação à natureza jurídica, das empresas brasileiras que exportaram continuamente entre 2017 e 2021, 74,07% são de sociedade por cotas de responsabilidade limitada -empresa privada.

Em relação à atividade econômica primária das empresas exportadoras que exportaram continuamente entre 2017 e 2021, o CNAE de maior frequência nos dados é o de fabricação de artefatos de material plástico não especificados anteriormente, representando 2,71%, seguido por fabricação de peças e acessórios para veículos automotores não especificados anteriormente (2,6%) e de fabricação de máquinas e equipamentos de uso geral não especificados anteriormente (1,94%). Não é a mesma ordem de frequência dos CNAES para quando se olha para toda base (as 51.620 empresas).

Na análise das que exportaram e importaram no mesmo ano, apresenta-se:

TABELA III
% DE EXPORTAÇÃO E IMPORTAÇÃO POR ANO.

Ano	Nº de Empresas que exportaram e importaram (A)	Nº Total de Empresas Exportadoras (B)	% (A/B)
2017	12.416	25.057	49,55
2018	12.623	25.266	49,96
2019	13.460	27.857	48,31
2020	13.305	28.736	46,30
2021	14.043	30.959	45,36

Observa-se o crescimento do número de exportadoras entre o período de 2017 a 2021, mas há uma redução do percentual das empresas que exportaram e importaram no mesmo ano, induzindo que houve o crescimento de número de exportadoras que não importaram bens no período.

B. Análise de Agrupamento

Iniciada a fase de Análise de Agrupamento, foi realizada uma redução de dimensionalidade pela análise de componentes principais, a qual considerou os dados de dez atributos quantitativos, sobre se a empresa brasileira exportou e importou entre 2017 e 2021. O gráfico de dispersão (figura 3) gerado com os dois componentes principais permitiu identificar 6 grupos, entre eles 4 mais definidos e outros 2 nas extremidades com menos pontos. Todos os grupos aparentam ter uma estrutura hierárquica dentro deles.

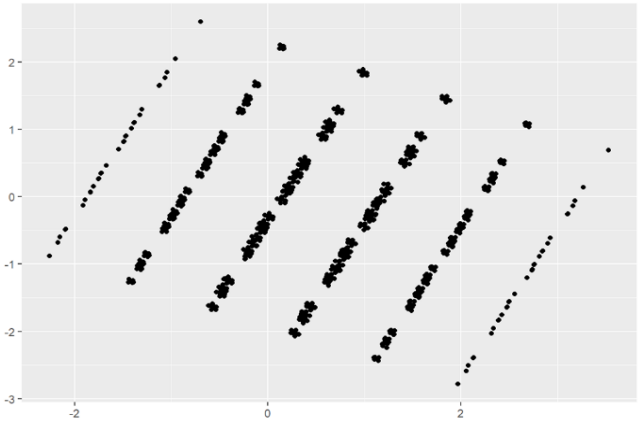


Fig. 3. Análise de Componentes Principais de dez variáveis quantitativas da base estudada. Os eixos representam os dois principais componentes.

Seguida a execução do algoritmo k-médias, foi possível identificar o centroide de cada grupo (1 a 6) para cada atributo, que, para esta pesquisa, pode variar entre 0 e 1 (enquanto mais próximo de 1, indica que há mais empresas com aquela característica). Assim, considerando que na base de dados não haverá empresas que apenas importaram entre 2017 e 2021, foram identificadas as seguintes características em cada grupo:

(i) Grupo 1: possui empresas que importaram e exportaram em todos os anos avaliados (centroides entre 0,89 a 0,98).

(ii) Grupo 2: possui número expressivo de empresas que importaram e exportaram apenas em 2020 (centroide de 0,9 e de 0,99, respectivamente).

(iii) Grupo 3: apresenta concentração de empresas que exportaram apenas em 2017 (centroide de 0,76).

(iv) Grupo 4: apresenta concentração de empresas que exportaram apenas em 2021 (centroide de 0,76).

(v) Grupo 5: possui empresas que importaram em todos os anos (com mais expressividade em 2019,2020 e 2021 - centroides maiores que 0,91), e com pouca expressividade em exportação, para qual se destacam apenas os anos de 2019 e 2021 (centroide de 0,64 e 0,67, respectivamente). –

(vi) Grupo 6: apresenta concentração de empresas que importaram em 2017 e 2018 (centroides maiores que 0,91) e com alguma expressividade na importação em 2019 (centroides de 0,77). Também apresenta empresas que exportaram em 2017 e em 2018, mas com centroides de 0,7 e 0,64 respectivamente.

Assim, foi gerado código para identificar os grupos com centroides maiores que 0,9 – em que há um número significativo de empresas brasileiras para aquela variável. O resultado identificou que apenas quatro grupos possuem características fortes: os grupos 1,2, 5 e 6, sendo o grupo 1 o que apresenta presença em todas as variáveis quantitativas selecionadas para o agrupamento. Além disso, os grupos 5 e 6 indicam apenas características significativas de importação, o que implica dizer que a maioria das empresas exportadoras do grupo 5 importaram em todos os três anos (2019 a 2021) e a maioria do grupo 6 importaram em todos os dois anos (2017 e 2018).

O resultado corrobora o que foi encontrado na análise PCA, em que, dos seis grupos encontrados, quatro grupos apresentam características mais fortes.

Ao se analisar a estabilidade do algoritmo utilizado, obteve-se o gráfico da figura 4, que exibe a distribuição da variável *tot.withinss* (que é a soma dos quadrados das distâncias entre cada observação e seu respectivo centro de grupo) para cada valor de K. Cada caixa representa a mediana, quartis e possíveis valores discrepantes (outliers) da distribuição da variável tot.withinss para cada K de 2 a 9. Portanto, o gráfico abaixo indica que 6 é o número ideal de grupos, pois os grupos apresentam melhor comportamento quando K igual a 6, apresentando instabilidade a partir de 7 novamente. Isso pode ocorrer pela característica hierárquica dos grupos.

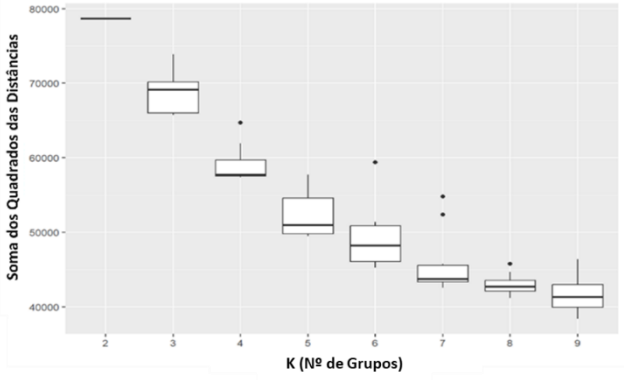


Fig. 4. Distribuição da variável Soma dos Quadrados das Distâncias (*tot.withinss)* para K (número de grupos) de 2 a 9

Ao avaliar a qualidade do agrupamento para K igual a 6, foi aplicada a técnica de *bootstrap*, em que foram geradas 10 amostras de *bootstrap*, considerando as variáveis se a empresa exportou e importou em 2021. O ano de 2021 foi escolhido por ser o mais próximo de 2023, dada as variações que possam ter ocorrido devido à pandemia iniciada em 2019, e por apresentar o maior número de exportadoras dentre os anos avaliados. O resultado está apresentado na figura 5, que contém informações detalhadas sobre os modelos de k-médias ajustados em cada uma das amostras *bootstrap*.

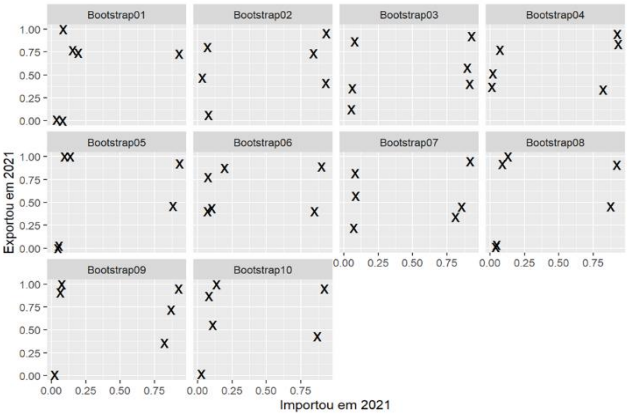


Fig. 5. Técnica *Bootstrap* - variabilidade nos grupos encontrados em diferentes amostras

O gráfico da figura 5 permite observar a variabilidade nos grupos encontrados em diferentes amostras. Os mesmos centros não são encontrados nas diferentes amostras, o que indica que não são grupos bem discriminados. Para a seleção das variáveis "Exportou em 2021" e "Importou em 2021", em sete amostras foi possível identificar grupos que não exportaram nem importaram neste ano, que, portanto, pertencem aos outros anos avaliados (que seriam os grupos 2,3 e 6 identificados anteriormente). Em todas as amostras identifica-se grupos que exportaram em 2021, mas não importaram no mesmo ano (característica do grupo 4), e grupos que importaram e exportaram no mesmo ano (o que engloba os grupos 1 e 5). Além disso, observa-se que em oito das dez amostras, há pelo menos dois grupos muito próximos ou se sobrepondo, corroborando as análises anteriores sobre a predominância de quatro grupos, dentro os seis identificados.

Dada a identificação dos grupos para cada observação e as análises sobre os resultados encontrados, passou-se para a identificação de regras de associação, considerando as variáveis qualitativas e os grupos encontrados. Para suporte mínimo de 0,1, 2 itens por conjunto de itens frequentes – para facilitar a interpretação - e confiança mínima de 0,9, foram encontradas 2 regras de associação. As regras encontradas estão apresentadas na figura 6:

```
##     lhs             rhs                                                    support confidence     lift itemset
## [1] {.cluster=1} => {Regularidade Exportadora Classificação=EXPERIENTE} 0.1975010  0.9817044 2.876516      13
## [2] {.cluster=1} => {PORTE=DEMAIS(RFB)}                                 0.1812863  0.9011074 1.487295      10
```

Fig. 6. Regras de associação encontradas pela metodologia descrita neste trabalho.

Assim a figura 6 traz a seguinte interpretação: considerando que um valor de *lift* superior a 1 indica que o antecedente (coluna lhs) e consequente (coluna rhs) aparecem mais frequentemente juntos do que o esperado, isso significa que a ocorrência de grupo igual a 1 tem um efeito positivo sobre a ocorrência de regularidade exportadora experiente (exportou em todos os anos avaliados) e porte diferente de micro e pequena empresa (que é a interpretação dada pela Receita Federal quanto ao porte "demais"). Portanto, as regras acima indicam que apenas o grupo 1 é coerentemente explicado pelos outros atributos. E conforme verificado na

parte de agrupamento, é o grupo que apresenta as empresas que importaram e exportaram em todos os anos avaliados, o que valida o resultado encontrado sobre a regularidade exportadora. Assim, tem-se, como conclusão em relação aos perfis das empresas brasileiras exportadoras, que as que exportaram e importaram no período de 5 anos – de 2017 a 2021 – atualmente são empresas de médio e grande porte (base de dados atualizada anualmente pela Receita Federal).

Como não foi possível identificar mais características do grupo 1 e não foi possível identificar características dos outros grupos, têm-se a evidência de que as informações atualmente disponíveis à Apex-Brasil sobre as empresas exportadoras não são suficientes para definir o perfil dessas. O que implica na necessidade de obter mais variáveis.

VII. CONCLUSÃO

O estudo buscou realizar a análise de ciência de dados das empresas brasileiras exportadoras no período de 2017 a 2021, por meio de segmentação e identificação de regras de associação. Conforme apresentado nos resultados, foi possível identificar o perfil de apenas um grupo - das empresas que exportaram e importaram continuamente no período avaliado - mas, ainda assim, com pouca informação sobre elas. Conclui-se, assim, a necessidade de enriquecimento de dados para este fim. Desse modo, recomenda-se:

(i) Obter mais atributos de "comportamento das empresas" relacionados a importação e exportação para aumentar a capacidade dos algoritmos de separar os dados no espaço. Pois com n atributos binários, tem-se no máximo \(2^n\) possíveis exemplos diferentes.

(ii) Obter mais atributos para entender e explicar cada grupo definido pelo padrão de comportamento, tais como: número de funcionários, tempo de existência da empresa, se foi apoiada pela Apex, valor exportado por ano, o que foi exportado, quantidade de países para os quais exportou, entre outros.

(iii) Aumentar o número de anos avaliados, com o cuidado de não enviesar os dados do passado sobre exportação, com os dados atuais (tais como porte e número de funcionários).

É indicada a continuação da pesquisa no âmbito do trabalho de segmentação de clientes da Apex-Brasil, para que se possa melhor entender o comportamento e os padrões de exportação das empresas brasileiras e a sua relação com as teorias de internacionalização atuais.

RECONHECIMENTO

São imperiosos o agradecimento e reconhecimento ao Professor Filipe Verri do Curso de Especialização em *Data Science* (CEDS) do Instituto Tecnológico da Aeronáutica (ITA), por sua liderança e parceria na elaboração e análise técnica desse artigo.

REFERÊNCIAS

[1] Apex-Brasil. Relatório de Gestão 2021, página 56. Disponível em: Transparência - Relatório de Gestão Apex-Brasil 2021.pdf - Todos os Documentos (sharepoint.com). Acessado em: 10 de mai. 2023.

[2] Brasil. Ministério do Desenvolvimento, Indústria, Comércio e Serviços (MDIC). Secretaria de Comércio Exterior (SECEX). Lista de Exportadoras e Importadoras. Disponível em: Outras Estatísticas de Comércio Exterior — Ministério da Economia (www.gov.br). Disponível em: 31 de dez. 2022.

[3] Costa, L. F. L. G. et al. Escolas teóricas do processo de internacionalização: uma visão epistemológica. Página 961. Cad. EBAPE.BR, Rio de Janeiro, v. 15, n. 4, art. 12, out./dez., 2017.

[4] Apex-Brasil. As Cadeias Globais de Valores e os desafios para os países que desejam fazer parte de um modelo de fragmentação da produção. Apex-Brasil, Brasília, 2019. Disponível em: https://portal.Apex-Brasil.com.br/as-cadeias-globais-de-valores-e-os-desafios-para-os-paises-que-desejam-fazer-parte-de-um-modelo-de-fragmentacao-da-producao. Acesso em: 17 jun. 2020.

[5] MOREIRA, U. Teorias do comércio internacional: um debate sobre a relação entre crescimento econômico e inserção externa. Revista de Economia Política, São Paulo, v. 32, n. 2, p. 213-228, abr./jun., 2012.

[6] CARNEIRO, J.; DIB, L. A. Avaliação comparativa do escopo descritivo e explanatório dos principais modelos de internacionalização de empresas. Internext: Revista Eletrônica de Negócios Internacionais da ESPM, São Paulo, v. 2, n. 1, p. 1-25, jan./jun. 2007.

[7] Brasil. Ministério do Desenvolvimento, Indústria, Comércio e Serviços (MDIC). Secretaria de Comércio Exterior (SECEX). Informativo nº 10/2023 sobre Nota informativa sobre a divulgação da lista de exportadores e importadores. Disponível em: https://www.gov.br/produtividade-e-comercio-exterior/pt-br/assuntos/comercio-exterior/estatisticas/informativos/informativo-2023-010. Disponível em: 08 de mar. 2022.

[8] Brasil. Ministério da Fazenda. Secretaria da Receita Federal do Brasil (RFB). Dados abertos – Cadastro Nacional da Pessoa Jurídica. Disponível em: Portal de Dados Abertos. Disponível em: 10 de mai. 2023.

[9] Fawcett. A. What is One Hot Encoding?. Disponível em: Data Science in 5 Minutes: What is One Hot Encoding? (educative.io). Disponível em: 11 de mai. 2023.

[10] DIETRICH, David et al. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. pp. 118-134. Indianapolis: John Wiley & Sons, Inc, 2015.

[11] FACELI, Katti et al. Inteligência Artificial: Uma abordagem de aprendizado de máquina. 2. ed. p. 196. Rio de Janeiro: LTC, 2022.

[12] DIETRICH, David et al. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and

Presenting Data. p. 118. Indianapolis: John Wiley & Sons, Inc, 2015.
[13] JAIN, A. K., DUBES, R. C. Algorithms for Grupoing Data. Prentice-Hall advanced reference series. pp. 55-135. PrenticeHall, Inc., Upper Saddle River, NJ, 1988.
[14] VARELLA, C.A.A. Análise de componentes principais: análise multivariada aplicada as ciências agrárias. p. 3. Seropédica: Universidade Federal Rural do Rio de Janeiro, 2008.
[15] DIETRICH, David et al. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. p. 228. Indianapolis: John Wiley & Sons, Inc, 2015.
[16] FACELI, Katti et al. Inteligência Artificial: Uma abordagem de aprendizado de máquina. 2. ed. p. 197. Rio de Janeiro: LTC, 2022.
[17] DIETRICH, David et al. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. pp. 138-158. Indianapolis: John Wiley & Sons, Inc, 2015.
[18] FACELI, Katti et al. Inteligência Artificial: Uma abordagem de aprendizado de máquina. 2. ed. p. 167. Rio de Janeiro: LTC, 2022.
[19] FACELI, Katti et al. Inteligência Artificial: Uma abordagem de aprendizado de máquina. 2. ed. p. 170. Rio de Janeiro: LTC, 2022.
[20] FACELI, Katti et al. Inteligência Artificial: Uma abordagem de aprendizado de máquina. 2. ed. p. 175. Rio de Janeiro: LTC, 2022.